Random Cluster Model on Phylogenetic Trees

MATH 833 - Fall 2012 *Presenter: Gautam Dasarathy*

1 Introduction

In this presentation we will discuss parts of [1]. Recall that the central question that we were interested in in the first half of the course was the following: how much about the evolutionary ancestry can we infer from the genetic information that is carried by the extant species. We will explore this issue using a simple model called the *random cluster model (RCM)* and demonstrate that, much like the Markov model on Trees, the relationship between the parameters of the RCM and the amount of data needed to infer the correct tree undergoes a sharp phase transition.

Throughout this writeup and the attendant presentation, X will denote a finite set with cardinality n and a phylogenetic X-Tree is a tree $\mathcal{T}$ having the leaf set X, and for which the interior vertices have degree at least 3. For the sake of simplicity, we will restrict our discussions to the case where the internal vertices have degree exactly 3; these trees are said to be *trivalent*.

The Random Cluster Model and Motivation. Given a phylogenetic X-Tree $\mathcal{T}$, consider the following process. For each edge e, cut this edge with probability $p(e)$. The resulting forest partitions the vertex set $V(\mathcal{T})$ into non-empty sets according the equivalence relation $u \sim v$ if u and v are in the same connected component. This model is called the random cluster model on the pair $(\mathcal{T}, p)$ and it generates partitions of $V(\mathcal{T})$ and therefore X. We will refer to these partitions as $\bar{\chi}$ and χ respectively in order to reinforce the notion of characters and character completions we studied earlier. Notice that χ generated thusly is a convex character with respect to the tree.

While Markov models have been widely used for modeling the evolution of genetic data (after alignment), there is an increasing interest in genomic characteristics such as gene orderings where these models do not capture the underlying essence. For instance, in these models, when there is a change of state – like gene reshuffling – it is likely that the resulting state remains the same for the rest of evolutionary history. In these settings, the random cluster model is the appropriate limiting case model. This leads to an important question: how many (i.i.d) characters are required to reconstruct the phylogenetic tree?

2 The Main Result

The main result of [1] is stated in Theorem 1.1. The result has two distinct parts and in this discussion, we will focus our attention on the first part. This can be thought of as an "achievability" result. In particular, we will prove the following theorem.

Theorem 1. *Let $0 < a \leq b < 1/2$ and $0 < \epsilon < 1$ be fixed constants. Consider the random cluster model on the pair $(\mathcal{T}, p)$, where $\mathcal{T}$ is a trivalent phylogenetic tree with n leaves and $a \leq p(e) \leq b$ for all the edges e of $\mathcal{T}$. Let k be the number of i.i.d characters generated under this model. Then, with probability exceeding $1 - \epsilon$, the tree can be reconstructed (using a polynomial(n) time algorithm) as long as*

$$k \geq \frac{(1-b)^4}{a(1-2b)^4} \log\left(\frac{n^2}{\epsilon}\right).$$

Before we prove this theorem, we need to introduce some terminology. We will denote by $\mathcal{Q}(\mathcal{T})$ the set of all quartet trees induced by $\mathcal{T}$. Given a collection of quartet trees $\mathcal{Q}$, we say that $\mathcal{Q}$ is displayed by $\mathcal{T}$ if $\mathcal{Q} \subset \mathcal{Q}(\mathcal{T})$.

For any three vertices, a, b, c, we let $\text{med}(a, b, c)$ denote the median of the vertex of the triple, i.e., the unique vertex of $\mathcal{T}$ that is shared by paths connecting a and b, b and c and a and c. We say that $\mathcal{Q}$ is a generous cover of $\mathcal{T}$ if $\mathcal{Q}$ is displayed by $\mathcal{T}$ and for all pairs of internal vertices u and v, there exists a quartet $xx'|yy'$ in $\mathcal{Q}$ such that $u = \text{med}(x, x', v)$ and $v = \text{med}(y, y', u)$.

Given a sequence of characters $\mathcal{C} = (\chi_1, \ldots, \chi_k)$ of characters on X, let

$$\mathcal{Q}(\mathcal{C}) = \left\{xx'|yy' : \exists i \in [k] : \chi_i(x) = \chi_i(x') \neq \chi_i(y) = \chi_i(y')\right\}.$$

Using simple graph theoretic arguments, one can show the following theorem (Theorem 2.4 in [1]).

Theorem 2. *Suppose that $\mathcal{Q}$ is a generous cover of a trivalent phylogenetic tree $\mathcal{T}$. Then, $\mathcal{T}$ is the only phylogenetic $X-$tree that displays $\mathcal{Q}$.*

We will not prove this theorem here but we will note that this implies that if one is given a generous cover $\mathcal{Q}$ of $\mathcal{T}$, then in principle $\mathcal{T}$ can be recovered. Actually, there is an easy (and relatively efficient) procedure to perform this recovery. Find a pair of leaves x, y such that $\mathcal{Q}$ contains no quartet of the form $xx' \mid yy'$. This pair of leaves, therefore has to be a cherry of $\mathcal{T}$. Now, find any quartet that contains x or y in $\mathcal{Q}$ and replace these occurrences with $u \notin X$; this stands for the immediate ancestor of x, y. Notice that this bottom-up procedure can be performed recursively and will yield the tree.

Now, all that remains to be shown is that given enough samples, the i.i.d data (i.e., $\mathcal{Q}(\mathcal{C})$) from the random cluster model gives us a generous cover of $\mathcal{T}$. Towards this end, we define $q(e) = 1 - p(e)$, $p_{\min} = \min_e p(e)$, and $q_{\min} = min_e q(e)$.

Lemma 1. *Let v be an arbitrary internal vertex and a be a neighbor of v, then the probability that there exists a path beginning at v, passing through a and ending at a leaf of $\mathcal{T}$ such that none of the edges are cut is lower bounded by $g(q_{\min}) := \frac{2q_{\min}-1}{q_{\min}}$.*

Proof. Let us denote the quantity we need to bound by $\alpha(v, a)$. We prove this result by induction. Note that if a is a leaf, then $\alpha(v, a) = q(\{v, a\}) \geq q_{\min} \geq g(q_{\min})$. Otherwise, let b and c be the neighbors of a different from v. Then, we have $\alpha(a, v) = q(v, a)(1 - \alpha(a, b))(1 - \alpha(a, c)) \geq q_{\min}(\alpha(a, b) + \alpha(a, c) - \alpha(a, b)\alpha(a, c))$. Applying the induction step concludes the proof. □

The bottom line is that this lower bound is non-vanishing as a function of k. Now, observe that for a pair of internal vertices u and v, the probability that a character χ generated under this model satisfies $\chi(x) = \chi(x') \neq \chi(y) = \chi(y')$ for some $x, x', y, y' \in X$ with $u = \text{med}(x, x', v)$ and $v = \text{med}(y, y', u)$ is lower bounded by $p_{\min} g(q_{\min})^4 =: \beta$. Consequently, the probability that $\mathcal{Q}(\mathcal{C})$ is not a generous cover for $\mathcal{T}$ is upper bounded by

$$\text{\# pairs of internal vertices} \times \mathbb{P}\left\{\text{no character in } \mathcal{C} \text{ can ``cover'' a particular pair } u, v\right\} \leq n^2(1-\beta)^k$$

The quantity on the RHS can be made smaller than ϵ if $k \geq \frac{1}{\beta} \log\left(\frac{n^2}{\epsilon}\right)$. Therefore, with probability exceeding $1 - \epsilon$, $\mathcal{Q}(\mathcal{C})$ is a generous cover of $\mathcal{T}$ and hence can be used to reconstruct $\mathcal{T}$. This concludes the proof of Theorem 1.

References

[1] Mossel, E., Steel, M., *A phase transition for a random cluster model on phylogenetic trees*,Mathematical Biosciences,Volume 187, Issue 2, February 2004, Pages 189203