A PROBABILISTIC MODEL FOR RECONSTRUCTING INTRA-TUMOR PHYLOGENIES

NIKO BEERENWINKEL AND FLORIAN MARKOWETZ

1. Introduction

We would like to reconstruct the evolutionary history of individual tumors. Each tumor consists of multiple subclonal genetic variants that have evolved after its last clonal expansion. We assume that the mixed tumor cell population has been bulk sequenced, such that the resulting reads provide a statistical sample of the underlying population. Ignoring structural variation and copy number alterations, we aim at inferring a phylogenetic tree model that represents the evolutionary history of the tumor clones from the observed short read data focusing only on single-nucleotide variations (SNVs). This task is different from classical phylogenetic approaches in at least three ways: (i) The observed data are short reads covering only a small fraction of the cancer genotypes that define the subclones. (ii) Due to the bulk sequencing approach of mixed samples, it is unknown which read originated from which cancer genotype. (iii) Read counts allow for inference of relative frequencies of mutations, which are informative about the order in which mutations have occurred.

We present an intra-tumor oncogenetic tree (ITOT) model for inferring tumor evolutionary histories from NGS data obtained from bulk seqeuncing of mixed tumor samples.

2. Methods

An SNV x is defined by its genomic position $\mathrm{pos}(x)$ and its variant nucleotide $\mathrm{nt}(x) \in \mathcal{A} := \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. Let $\mathcal{S}$ denote the set of all SNVs in a given tumor. NGS of the tumor produces short reads and we consider only those reads that overlap with at least one segregating site $j \in \{\mathrm{pos}(x) \mid x \in \mathcal{S}\}$. We assume that each read contains either one or two SNVs (relative to the most frequent nucleotide at the respective position). Reads without SNVs are not informative and those with more than two SNVs are so rare that we can ignore them. Then the data X consists of N_1 reads carrying one SNV, $x_1^{(i)} \in \mathcal{S}$, $i = 1, \dots, N_1$, and N_2 reads with pairs of SNVs, $x_2^{(i)} \in \mathcal{S}^2$, $i = 1, \dots, N_2$.

An intra-tumor oncogenetic tree (ITOT) is a labeled binary rooted tree $T = (V, E, r, g, f, s)$ (Figure **??**). The edges $e \in E$ are labeled with sets of SNVs, $s : E \to 2^{\mathcal{S}}$, that have occurred on this branch such that they define a partition of the set of all SNVs, i.e., $s(e_1) \cap s(e_2) = \emptyset$ for all edges $e_1 \neq e_2$ and $\cup_{e \in E} s(e) = \mathcal{S}$. The vertices $v \in V$ represent subclones and are labeled (uniquely) with genotypes. A genotype is a set of co-ocurring SNVs and each vertex is labeled by the set of all SNVs that have given rise to it. Formally, let $(e_1, \dots, e_M)$ be the unique path (ordered set of edges) connecting the root vertex r with vertex $v \in V$. Then the genotype label $g : V \to 2^{\mathcal{S}}$ of vertex v is $g(v) = \cup_{m=1}^{M} s(e_m)$. It follows that $g(v) = g(w_1) \cap g(w_2)$ for all branching points v into daughter clones w_1 and w_2. The root of the tree is the genotype after the last clonal expansion of the tumor, the leaves of the tree are the subclones at the time of observation (diagnosis), and the interior vertices are extinct common ancestor clones. We denote by $L \subset V$ the set of leaves (contemporary clones). In addition, vertices are labeled with their relative frequencies

Date: June 28, 2013.

```
  ----------- A   0.7
 |
-+ r
 |     ------ BC  0.2
  ----+ B
       ------ BD  0.1
```

FIGURE 1. ITOT for three clones. Edges are labeled from top to bottom by SNV sets A, C, B, D. Vertices are labeled from top to bottom by genotypes A (leaf), r (root), BC (leaf), B (internal), BD (leaf), and by clone frequencies 0.7, 1.0, 0.2, 0.3, 0.1.

within the tumor, $f : V \to (0, 1]$, subject to the following constraints: (i) $f(r) = 1$, (ii) $\sum_{v \in L} f(v) = 1$, and (iii) $f(v) = f(w_1) + f(w_2)$ for all branching points v into daughter clones w_1 and w_2. Note that if (ii) holds for a function $f : L \to (0, 1]$, then f can always be extended in a unique way to V such that (i) and (iii) also hold.

The ITOT model consists of an ITOT $T = (V, E, r, g, f, s)$ for generating genotypes G and a probabilistic model for generating reads, i.e., single SNVs X_1 and SNV pairs X_2, from genotypes. Let $L = \{g_1, \dots, g_K\}$ denote the genotypes of the contemporary clones, i.e., the leaves of the tree. Each leaf genotype G is generated with probability

$$\Pr(G = g_k) = f_k \tag{1}$$

where $f_k \in (0, 1]$ are parameters such that $\sum_{k=1}^{K} f_k = 1$. The parameters f_k define the clone frequencies of the ITOT via $f(g_k) = f_k$ and its unique recursive extension to V. The conditional probabilities of observing single and pair SNVs are, respectively,

$$\Pr(X_1 = s \mid G = g) \;=\; \begin{cases} \varepsilon & \text{if } |\{s\} \cap g| = 0 \\ 1 - \varepsilon & \text{if } |\{s\} \cap g| = 1 \end{cases} \tag{2}$$

$$\Pr(X_2 = (s, t) \mid G = g) \;=\; \begin{cases} \varepsilon^2 & \text{if } |\{s, t\} \cap g| = 0 \\ 2\varepsilon(1 - \varepsilon) & \text{if } |\{s, t\} \cap g| = 1 \\ (1 - \varepsilon)^2 & \text{if } |\{s, t\} \cap g| = 2 \end{cases} \tag{3}$$

where ε is the error probability accounting for misassignment of SNVs to their clones. The ITOT model is the mixture model

$$\Pr(X_m \mid T) \;=\; \sum_{k=1}^{K} \Pr(X_m, G = g_k) = \sum_{k=1}^{K} \Pr(X_m \mid G = g_k) \Pr(G = g_k), \quad m = 1, 2$$

To learn the structure and parameters of the ITOT model for fixed K from observed SNV data, we devise a structural EM algorithm. In the E step, we compute the responsibility of each leaf genotype for each observed read. In the M step, we re-estimate the ITOT.

2.1. **E step.** The responsibility of leaf genotype g_k for observation $x_m^{(i)}$, $m = 1, 2$, $i = 1, \dots, N_m$, is

$$\gamma_{mik} \;:=\; \Pr(G = g_k \mid X_m = x_m^{(i)}) \tag{4}$$

$$\;=\; \frac{\Pr(X_m = x_x^{(i)} \mid G = g_k) \Pr(G = g_k)}{\sum_{k=1}^{K} \Pr(X_m = x_x^{(i)} \mid G = g_k) \Pr(G = g_k)} \tag{5}$$

2.2. **M step.** The ML step consists of maximization over the discrete structures tree topologies and SNV partitions, and over the continuous parameters f (clone frequencies) and ε (error probability). For given leaf genotypes $g_1, \ldots, g_K$, the MLEs of f and ε are, respectively,

$$\hat{f}_k = \frac{\sum_{m=1,2} \sum_{i=1}^{N_m} \gamma_{mik}}{N_1 + N_2}, \quad k = 1, \ldots, K \tag{6}$$

$$\hat{\varepsilon} = \frac{\sum_{m=1,2} \sum_{i=1}^{N_m} \gamma_{mik}(m - |x_m^{(i)} \cap g_k|)}{N_1 + 2N_2} \tag{7}$$

Estimation of the leaf genotypes involves reconstruction of the ITOT from the responsibilities γ_{mik}. Exact ML estimation seems difficult, but we can employ heuristics, such as distance-based phylogenetic methods. The distance between clone k and clone l may be defined as

$$d_{kl} = \sum_{m=1,2} \sum_{i=1}^{N_m} (1 - 2 \min\{\gamma_{mik}, \gamma_{mil}\}) \tag{8}$$

The resulting tree (V, E) can be labeled recursively from the leaves to the root by setting $g(g_k) = \cup_{m,i}\{x_m^{(i)} \mid \gamma_{mik} \geq \eta\}$ for all leaves g_k, and $g(v) = g(w_1) \cap g(w_2)$ for all branchings v into daughters w_1, w_2. Here, the paramter η controls the multiple, hard assignment of SNVs to genotypes. $\eta \approx \varepsilon$ for single and $\eta \approx \varepsilon^2$ for pairs of SNVs appears a reasonable choice, or else a preset fixed cutoff.

Alternatively, the triplet search may be employed here ...

From $g : V \to 2^{\mathcal{S}}$, edge labels $s : E \to 2^{\mathcal{S}}$ can be inferred. Clone frequencies f can then be estimated for all leaves as described above and extended to V as described earlier.

3. Results

4. Discussion