IMO: Inferring Mutation Order from Cross-sectional Data

July 2, 2018

Abstract

1 Introduction

TODO: Make a table for all terms getting used in all paper, i.e., make the notation uniform. TODO: Define that the genetic aberration can be CNA, Mutation, etc.

2 Structure-based Categories

2.1 Tree Models

2.1.1 Inferring Tree Models for Oncogenesis from Comparative Genome Hybridization Data [2]

Authors of [2] try to generalize the Volgenstien linear notion of tumor progression [3]. They consider a graph whose vertices are genetic events[1] Instead of a linear (path/chain) growth of mutations, they assume that there is an underlying directed tree (branching) that describes the ordering of the events.

Model: Consider the branching[2] $T = (\mathcal{V}, \mathcal{E}, r, \alpha)$ where r is the root representing normal state of the cell and α is a probability function of each edge. They also introduce a variant where the edges also have attached time random variable and there is a total tumor evolution time where control the end of the random process.

The set of mutations $\mathcal{S}$ occur if there is a path from the root to every element of $\mathcal{S}$. Therefore, all the edges on the path should happen and all the edges neighbor to the path should not happen (otherwise we include something that is not in $\mathcal{S}$).

$$\mathbb{P}(\mathcal{S}) = \prod_{e \in \mathcal{S}} \alpha(e) \prod_{e \in \mathcal{N}(\mathcal{S}) \setminus \mathcal{S}} 1 - \alpha(e) \tag{1}$$

[1] They work with chromosome copy number alterations of renal cancer.

[2] Branching is a directed rooted tree where all edges have orientations away from the root.

where $\mathcal{N}$ is the set of all edges neighbor to the corresponding path. If there is no path to even a single member of $\mathcal{S}$, then $\mathbb{P}(\mathcal{S}) = 0$.

Another way to look at the OncoTree model is through the graphical models lens. Consider the tree graphical model with binary random variables whose structure is represented by the branching T. First note that the probability of set $\mathcal{N}(\mathcal{S})$ can be find through marginalization. In other words, we do not care about the value of other nodes:

$$\mathbb{P}(\mathcal{N}(\mathcal{S})) = \sum_{X_i \notin \mathcal{N}(\mathcal{S})} \prod_{X_i \in \mathcal{V}} \mathbb{P}(X_i | \text{pa}(X_i)) \tag{2}$$

When we are interested in set $\mathcal{S}$ happening, it translates to a specific assignment of binary values to $\mathcal{N}(\mathcal{S})$, i.e., all of the random variables in the set $\mathcal{S}$ are one and all of the neighbors of those nodes in the tree (minus themselves) are zero, i.e., $\mathcal{N}(\mathcal{S}) \setminus \mathcal{S}$. Therefore we can write the probability of $\mathcal{S}$ happening as:

$$\begin{aligned}
\mathbb{P}(\mathcal{S}) &= \mathbb{P}(\forall X_i \in \mathcal{S} : X_i = 1, \forall X_i \in \mathcal{N}(\mathcal{S}) \setminus \mathcal{S} = 0) \\
&= \prod_{X_i \in \mathcal{S}} \mathbb{P}(X_i = 1 | \text{pa}(X_i) = 1) \prod_{X_i \in \mathcal{N}(\mathcal{S}) \setminus \mathcal{S}} \mathbb{P}(X_i = 1 | \text{pa}(X_i) = 1) \\
&= \prod_{X_i \in \mathcal{S}} \mathbb{P}(X_i = 1 | \text{pa}(X_i) = 1) \prod_{X_i \in \mathcal{N}(\mathcal{S}) \setminus \mathcal{S}} (1 - \mathbb{P}(X_i = 1 | \text{pa}(X_i) = 1))
\end{aligned}$$

It is easy to see that the edge probabilities are the conditional probabilities, i.e., $\alpha(e_{ij}) = \mathbb{P}(X_j = 1 | X_i = 1)$.

Algorithm: Note that we primarily want to learn the structure of the tree not the conditional probabilities $\alpha(e)$s. To simplify the notation, let $p_i = \mathbb{P}(X_i = 1)$, $p_{ij} = \mathbb{P}(X_i = 1, X_j = 1)$, and $p_{j|i} = \mathbb{P}(X_j = 1 | X_i = 1)$. We define the weight of edges between mutation i and j as: $w_{ij} = \log \frac{p_i}{p_i + p_j} \frac{p_{j|i}}{p_j}$.

The weights captures the intuition behind "i being the parent of j": If i is more frequent and when it happens it increases the probability of j happening. Authors show that the maximum branching or the graph constructed with edge weights w_{ij} is exactly the tree T. To compute the weights from the data we replace the probabilities with the frequencies: $w_{ij} = \log \frac{f_{ij}}{f_j (f_i + f_j)}$.

Therefore, we are after a maximum directed tree in a weighted graph. To filter out false positives and reduce the size of the problem, we focus on a subset of mutations, i.e., maximum-weight clique of the graph. Then we use the Edmond's algorithm which runs in $O(n^2)$ and finds the minimum branching of the graph with negated edge weights $-w_{ij}$. Algorithm 1 summarizes the pseudo-code of the method.

Authors prove that, the computed branching is the correct one with high probability for enough number of samples and when there is no false positive or negative, and finally the tree is not *skewed*.

Skewness means that for three nodes i, j, and k where k is the lowest common ancestor of the other, $p_{j|i} > p_{i \cup j|k} = p_{i|k} + p_{j|k} - p_{i,j|k}$. Intuitively, in this case it is hard to tell apart causation of $i \to j$ from $k \to j$. They claim that untimed trees are not skewed which I don't get.

Summary: For an un-timed tree, with enough number of samples with high probability OncoTree finds the ground truth tree. For timed tree, as long as it is un-skewed,

Algorithm 1 OncoTree

input: Frequencies of mutations f_i, f_{ij}, a threshold t, and a clique size s
output: Branching $T = (\mathcal{V}, \mathcal{E}, r, p)$
$w_{ij} \leftarrow \mathbf{1}(f_{ij} > t) \log \frac{f_{ij}}{f_j(f_i+f_j)}$
$G = (\mathcal{V}, \mathcal{W})$, where $\mathcal{V}$ and $\mathcal{W}$ are sets of all mutations and edge weights.
Optional: Focus on maximum-weight clique of size s in G
$w_{ij} \leftarrow -w_{ij}$
Compute the minimum branching of G: $T(\mathcal{V}, \mathcal{E}, r) \leftarrow$ Edmond's(G)
Compute edge probabilities of branching as: $p_{j|i} \leftarrow \frac{f_{ij}}{f_i}$.

they can recover it. For path trees, because of the equivalence (for each timed path tree the corresponding distribution can be represented by an un-timed counterpart), they can recover the ground truth.

Dataset: A set of 117 cases of clear cell renal cell carcinoma collected using CGH [5] as described in [6]. The frequency threshold t was set to 5 and the size of maximum-weighted clique s to 7.

2.1.2 Estimating an oncogenetic tree when false negatives and positives are present [7]

Model: Here the model is exactly similar to the OncoTree model of Section 2.1.1 and authors only present another algorithm for finding the maximum spanning branching which has the same complexity of $O(n^2)$ but the proof is invariant to any monotone transformation of the edge weight $w_{ij} = \log \frac{p_i}{p_i+p_j} \frac{p_{j|i}}{p_j}$ and to some degree to false positive and negative.

In particular, let's ϵ_+ and ϵ_- be the uniform (similar for all patients) false positive and negative probabilities respectively. Then, if $\epsilon_+ + \epsilon_- < 1$ and $\epsilon_+ < (p_{\min})^{1/2}(1 - \epsilon_+ - \epsilon_-)$ for enough number of samples the proposed algorithm recovers the true unskewed branching with high probability.

Algorithm: The algorithm is presented in Algorithm 2. Intuitively, we first find the least frequent mutation without an assigned parent and then set its parent to its neighbor with highest weighted.

Dataset: The CGH [5] data for 25 samples with renal cell carcinoma from [4] study was used. The frequency threshold t was set to 5 which resulted in 11 aberrations.

2.2 Forest Models

2.2.1 Paper Title

Model:
Algorithm:
Dataset:

Algorithm 2 OncoTree with False Positive and Negative

input: Frequencies of mutations f_i, f_{ij}, a threshold t
output: Branching $T = (\mathcal{V}, \mathcal{E}, r, p)$
Compute edge weights: $w_{ij} \leftarrow \mathbf{1}(f_{ij} > t) \log \frac{f_{ij}}{f_j(f_i+f_j)}$
Form $G = (\mathcal{V}, \mathcal{W})$, where $\mathcal{V}$ and $\mathcal{W}$ are sets of all mutations and edge weights.
$\mathcal{S}_0 \leftarrow \emptyset$, $\mathcal{E}_0 \leftarrow \emptyset$
for $v = 1$ to $|\mathcal{V}|$ **do**
    $i \leftarrow \text{argmin}_{k \in \mathcal{V} \setminus \mathcal{S}_t} f_k, \quad j \leftarrow \text{argmax}_{k \in \mathcal{V} \setminus \mathcal{S}_t} w_{ki}$
    $\mathcal{S}_{t+1} \leftarrow \mathcal{S}_t \cup j, \quad \mathcal{E}_{t+1} \leftarrow \mathcal{E}_{t+1} \cup (j, i)$
end for
Set root r to the node without parent.
Form the branching $T = (\mathcal{V}, \mathcal{E}_{|\mathcal{V}|}, r)$
Compute edge probabilities of branching as: $p_{j|i} \leftarrow \frac{f_{ij}}{f_i}$.

2.3 DAG Models

2.3.1 Conjunctive Bayesian Network [1]

Model: CBN [1] is the first paper that goes beyond tree and forest structure and models the cancer progression with Directed Acyclic Graph (DAG). In its core, CBN is a special form of directed graphical model or Bayesian network [**?**], where genetic events are nodes of the BN and the chance of an event happening if even one of its parent did not happen is zero which justifies the "conjunctive" name.

Mathematically, CBN was introduced [1] as a partial order triplets $(\mathcal{V}, \leq, \Theta)$ where $\mathcal{V}$ is the set of genetic events and $\leq$ is the partial order induced by the DAG and $\Theta = (\theta_1, \ldots, \theta_{|\mathcal{V}|})$ is a tuple of parameters each corresponds to one genetic event. To conform with the graphical model notation, we represent the partial order $\leq$ with a set of directed edges $\mathcal{E}$ where $G = (\mathcal{V}, \mathcal{E})$ is the corresponding Bayesian network. The goal of CBN algorithm is to both learn the structure of the DAG, i.e., the set $\mathcal{E}$ and also all parameters Θ. To handle data points that are inconsistent with the DAG they assume a mixture model where the data is generated by the model with probability λ or is picked uniformly at random from all possible inconsistent data points with probability $(1 - \lambda)$.

Algorithm: The following algorithm 3, has three main steps. First, we remove events that we can not distinguish from each other. Events u and v are indistinguishable if $\sum_{i=1}^{n} \mathbf{u}_i \oplus \mathbf{v}_i == 0$ and no algorithm without prior knowledge can separate them using the given data set. Then, we learn the structure of the graph, where we tolerate $t \times n$ incompatible data points where $t \in [0, 1)$ is the given incompatibility threshold. Note that, $(\mathbf{x}_i \cap \mathbf{1}_{uv}) == \mathbf{1}_u$ on line 13 of the Algorithm 3, flags the inputs in which the child event u happened without the parent event v happening which makes the sample $\mathbf{x}_i$ incompatible with the hypothesis of v being the parent of u, i.e., $(v, u) \in \mathcal{E}$. Finally, after removing the incompatible data we learn the parameters by simple maximum likelihood estimate.

Dataset:

Algorithm 3 CBN: Conjunctive Bayesian Network

input: Input binary data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, a threshold t
output: CBN $= (\mathcal{V}, \hat{\mathcal{E}}, \hat{\Theta})$ and $\hat{\lambda}$
{Merging indistinguishable events}
for $u = 1$ to p **do**
  for $v = u + 1$ to p **do**
    if Events u and v always co-occur **then**
      Lump them together as a new event uv and decrease p by one.
    end if
  end for
end for
{Learning the structure of the DAG}
for $u = 1$ to p **do**
  for $v = u + 1$ to p **do**
    $\mathbf{1}_{uv} = \mathbf{1}_u + \mathbf{1}_v$
    if $\sum_{i=1}^{n} [(\mathbf{x}_i \cap \mathbf{1}_{uv}) == \mathbf{1}_u] \leq t \times n$ **then**
      $\hat{\mathcal{E}} \leftarrow \hat{\mathcal{E}} \cup \{(v, u)\}$
    end if
  end for
end for
{Learning the model's parameters}
Form a new input matrix $\mathbf{X}^{'} \in \mathbb{R}^{n' \times p}$ by removing the incompatible observations $\mathbf{x}_i$ where $(\mathbf{x}_i \cap \mathbf{1}_{uv}) == \mathbf{1}_u$ and $(u, v) \in \hat{\mathcal{E}}$.
for $v = 1$ to $|\mathcal{V}|$ **do**
  $\hat{\theta}_v = \frac{\sum_{i=1}^{n'} (\mathbf{x}_i \cap \mathbf{1}_v)}{\sum_{i=1}^{n'} (\mathbf{x}_i \cap \mathbf{1}_{\text{pa}(v)})}$
end for
$\hat{\lambda} = \frac{n'}{n}$

3 Other Categories

3.1 Causal Models

3.2 Timed Models

3.3 Computational Cancer Biology: An Evolutionary Perspective - 2016:

Introduction:

Only mutation is not a cancer sign! Eyelid epidermal cells harbor lots of mutations that are also present in cancer genes [9]. Intra-tumor heterogeneity is a major problem for targeted therapy, where unknown sub-clone may exist before the treatment [12]. Many mutation can deregulate a same pathway or not so their effect/outcome are not the same.

Types of mutation: SNV as the change in a few consecutive base pairs vs. structural

variants (SVs) which ultimately leads to CNVs.

With single snapshot it is hard to do anything, now we have other sources of data: Spatial: multiple biopsies from same tumor, or samples from origin or metastasis sites. Temporal: Before after treatment, Original and relapse.

Amir: still we don't know (and probably won't know) what happened before the diagnosis. So the mutation ordering is always valid.

Functional Interpretation:

Mutations in cancer cells can be broadly divide to drivers and passenger. Drivers are those that cause cancer (fast cell growth) while passengers are non-causal to cancer. Cancer genes are those that have driver mutation or if epigenetically modified cause cancer.

Major assumption about driver mutation: those are most frequent across tumor collection. But still this can't be the only method because mutation freq can have other causes: lower expression genes, and those that replicated during cell cycle (vs. early) can have higher mutation rate without causing any problem.

Also, pathway (i.e., the whole biological network or cause and effect that generate a product or change the cell state) is more important than the single cell. If several genes mutate in a same pathway theoretically the final outcome on the tumor state should be the same. So it seems biologically in efficient to have multiple mutation in a same pathway (why isn't that random?). Therefore mutations occur mutually exclusive in pathways.

Intratumor Phylogeny: We want to reconstruct the evolutionary history of the tumor from a snapshot of it. This reconstruction can use mutations and CNAs in general (there is some info regarding CNA-based reconstruction that I don't understand). Also, we can do it using the single cell data (still not so much publicly accessible data, and the data is biased and noisy.) or bulk sequencing data (which is imperfect and indirect because we have a group of cells probably from different sub-clones and normal cells.)

Perfect phylogeny assumption: Mutations are irreversible and can occur once in a tree. (the first one is reasonable but second one? maybe very low probability)

Spatial genomics and biogeography: The 3D position of the cells in the tumor may have some implication for therapy and also the phylogenetic tree reconstruction. There are two possible ways to get that information in future: barcoding the cells and then extracting and sequencing them or inventing situ single-cell sequencing methods. This is related to the field of biogeography which studies the spatial distribution of species with different genetic makeup. We are looking for models in which cells move because of their neighbor forces or themselves.

Tumor cell population dynamics: I didn't understand this very much, it is mostly about modeling the population change using the branching processes. Important observation: drug resistant is fait accompli in the tumor, meaning that those mutated resisting cells are already in the tumor before the starting of the treatment.

Cancer progression networks: Multi-stage theory of cancer progression: ? Fix the cancer type. There is an evolutionary (random) process that making this tumor. So using different realization of this process we want to infer the process.

At first people only considered the linear progression model but then went for the PGM like: tree, mixters of trees and DAG. This can help to better quantify the stage of

the tumor, predict the waiting time for the mutational accumulation and finally survival prediction.

Ideas: Sparsity based methods for finding the drivers gene? Sparse group lass for detecting genes in the pathway?

References

[1] Niko Beerenwinkel, Nicholas Eriksson, and Bernd Sturmfels. Conjunctive bayesian networks. *Bernoulli*, pages 893–909, 2007.

[2] Richard Desper, Feng Jiang, Olli-P Kallioniemi, Holger Moch, Christos H Papadimitriou, and Alejandro A Schäffer. Inferring tree models for oncogenesis from comparative genome hybridization data. *Journal of computational biology*, 6(1):37–51, 1999.

[3] Eric R Fearon and Bert Vogelstein. A genetic model for colorectal tumorigenesis. *Cell*, 61(5):759–767, 1990.

[4] Feng Jiang, Jan Richter, Peter Schraml, Lukas Bubendorf, Thomas Gasser, Guido Sauter, Michael Jörg Mihatsch, and Holger Moch. Chromosomal imbalances in papillary renal cell carcinoma: genetic differences between histological subtypes. *The American journal of pathology*, 153(5):1467–1473, 1998.

[5] Anne Kallioniemi, Olli P Kallioniemi, Damir Sudar, Denis Rutovitz, Joe W Gray, Fred Waldman, and Dan Pinkel. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, 258(5083):818–821, 1992.

[6] Holger Moch, Joseph C Presti, Guido Sauter, Niels Buchholz, Paul Jordan, Michael J Mihatsch, and Frederic M Waldman. Genetic aberrations detected by comparative genomic hybridization are associated with clinical outcome in renal cell carcinoma. *Cancer research*, 56(1):27–30, 1996.

[7] Aniko Szabo and Kenneth Boucher. Estimating an oncogenetic tree when false negatives and positives are present. *Mathematical biosciences*, 176(2):219–236, 2002.