Structured High Dimensional Data Sharing Model

May 9, 2018

Abstract

1 Introduction

2 Related Work

2.1 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?

2.2 Inferring Tree Models for Oncogenesis from Comparative Genome Hybridization Data - 1999

Authors try to generalize the Volgenstien linear notion of tumor progress. They consider a graph whose vertices are mutations (in the original paper based on the state of the art data availability, authors work with chromosome copy number alterations). Instead of a linear (path) growth of mutations, they assume that there is an underlying directed tree (branching) that describes the ordering or the events.

Note that this is far back in 1999 and PGM or SCM are not popular yet. So, digesting the notion or probabilistic tree of the authors needs a little bit of effort. The tree is

$T = (\mathcal{V}, \mathcal{E}, r, \alpha)$ where r is the root representing no/zero mutation state of the tumor 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 $\mathcal{S}$.

$$\mathbb{P}(\mathcal{S}) = \prod_{e \in \text{path}(r \to \mathcal{S})} \alpha(e) \prod_{e \in \mathcal{N}(\text{path}(r \to \mathcal{S}))} 1 - \alpha(e) \tag{1}$$

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 $\mathbb{P}(\mathcal{S}) = 0$.

Authors define the weight of edges between mutation i and j as: $w_{ij} = \log \frac{p_i}{p_i + p_j} \frac{p_{ij}}{p_i p_j}$. Note that if the probability of j happening given i is high and also i is more frequent than j then it makes sense to assume that i is the cause/parent of j and also w_{ij} becomes bigger than w_{ji}. Therefore, we are after a maximum directed tree in the complete weighted graph. This can be computed $O(n^2)$ using generic maximum spanning tree algorithm using the Edmond's algorithm. Note that $w_{ij} < 0$ and therefore the maximum branching problem here can be converted to the minimum branching problem which is exactly what Edmond's algorithm do. Since the Edmond's method takes in the full graph, here they use a heuristic (I didn't look at it, and I think it was not mentioned in the paper) to remove some edges and then start the Edmond's method.

So here is the pseudo-code of the algorithm:

- Compute the weights as $w_{ij} \leftarrow \log \frac{p_i}{p_i + p_j} \frac{p_{ij}}{p_i p_j}$ where ps are counted from data.
- Negate the weights $w_{ij} \leftarrow -w_{ij}$.
- Do some *smart* pre-processing and remove some weights (We can skip this for now).
- Run the Edmonds' algorithm for optimum branching (look it up in Wikipedia and there should be Python code on the Web) and find the minimum branching.

Side note: note that p_i is the summation over all possible mutations where i occured, i.e., there are valid (reachable from r) nodes in the $\mathcal{S}$ and also $i \in \mathcal{S}$.

Authors prove that, the found maximum spanning tree is the correct one when number of samples goes to infinity, 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.

So, basically their result is the following:

For an untimed tree, with enough number of samples with high probability we find the ground truth tree. For timed tree, as long as it is unskewed we can recover it.

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

2.3 Estimating an oncogenetic tree when false negatives and positives are present - 2002

Here authors present another algorithm for finding the maximum spanning tree which has the same complexity but the proof is resistant to any monotone transform of the above suggested w_{ij}. Also they show if the false positive and negative rates are somehow bounded with enough number of samples (to estimate the p_is, if we know them the recovery is exact) recovery is possible.

The algorithm is simple:

- Start with an empty set $\mathcal{S}_0$
- $i = \text{argmin}_{k \in \mathcal{V} \setminus \mathcal{S}_t} p_k$
- $\mathcal{S}_{t+1} = \mathcal{S}_t \cup \{\text{argmax}_{j \in \mathcal{V} \setminus \mathcal{S}_t} w_{ji}\}$

2.4 Conjunctive Bayesian Network - 2006