Enumerating Fewer than $2^n$ Partitions for PhySigs

Baqiao Liu
open for the case where \( k = O(n) \). Second, PhySigs exhaustively enumerates all \( 2^n \) partitions of the \( n \) nodes of input tree \( T \). It will be worthwhile to develop efficient heuristics that return solutions with small error. Third, we plan to assess statistical significance of solutions returned by PhySigs using permutation tests or bootstrapping, similarly to Huang et al.[160]
Problem

● Previous methods: either infer identical exposures (to mutational signatures) for all clones, or infer it independently for each clone
● PhySigs: generalize previous approaches
Problem

- Previous methods: either infer identical exposures (to mutational signatures) for all clones, or infer it independently for each clone
- PhySigs: generalize previous approaches
- (The next couple of slides are almost completely copied from Sarah Christensen’s talk)
Clonal Evolution Theory of Cancer
[Nowell, 1976]
Problem

- No tree required
- Large sample size
- Less overfitting
- No heterogeneity

- Heterogeneity
- Small sample size
- May overfit
- Clones required

- Heterogeneity
- Medium sample size
- Less overfitting
- Tree required

Same exposures for every clone

Different exposures for every clone

Clone exposures separated by shifts
Problem 3 (Tree-constrained Exposure (TE)). Given feature matrix $P$, corresponding count matrix $C$, signature matrix $S$, phylogenetic tree $T$ and integer $k \geq 1$, find relative exposure matrix $D$ such that $\|P - SDC\|_F$ is minimum and $D$ is composed of $k$ sets of identical columns, each corresponding to a connected subtree of $T$.

$P$: Mutation Count Matrix  
$S$: Signature Matrix  
$D$: Relative Exposure Matrix  
$C$: Weight Matrix

$E$: Exposure Matrix
Problem

- For a fixed k, PhySigs enumerates all k-partitionings of the input tree T, minimizing some error
- Since using more k-partitionings should make the model fit better, to avoid overfitting, the k with the best Bayesian Information Criterion (BIC) is selected
- This implies that we are enumerating $\sum_{k=1}^{n} \binom{n-1}{k-1} = 2^{n-1}$ partitions, where n is the number of tumor clones

$$\text{BIC}(L(k)) = mn \log(L(k)/(mn)) + kr \log(mn).$$
Goal

- Enumerate fewer partitions
First line of attack: the BIC is (almost) unimodal?

- In simulated dataset: 98.3% BIC unimodal
- In TRACERx dataset (lung cancer): 99.3% BIC unimodal
First line of attack: the BIC is (almost) unimodal?

- Obvious (maybe dangerous) heuristic: use some search algorithm for discrete unimodal distribution, say ternary search
- Asymptotically instead of searching for $\Theta(n)$ choices of $k$, we now search for $\Theta(\lg n)$ choices of $k$
  - With the input sizes this partially does not matter for now
- Somehow I don’t think this was what the thesis had in mind...
Results of Applying Ternary Search

- In simulated dataset: 100% correct BIC recovered (average # total clones=6)
- In TRACERx dataset: 100% correct BIC recovered (average # total clones=6.63)
- I believe that with more total number of clones this can do better
  - Need to generate more data
Second (potential) line of attack: recursive bipartitioning

- First try all bipartitions the tree, select the best bipartitioning (by error)
- Choose the larger (or by some criterion) cluster, recurse and bipartition until the BIC starts to increase with the solution