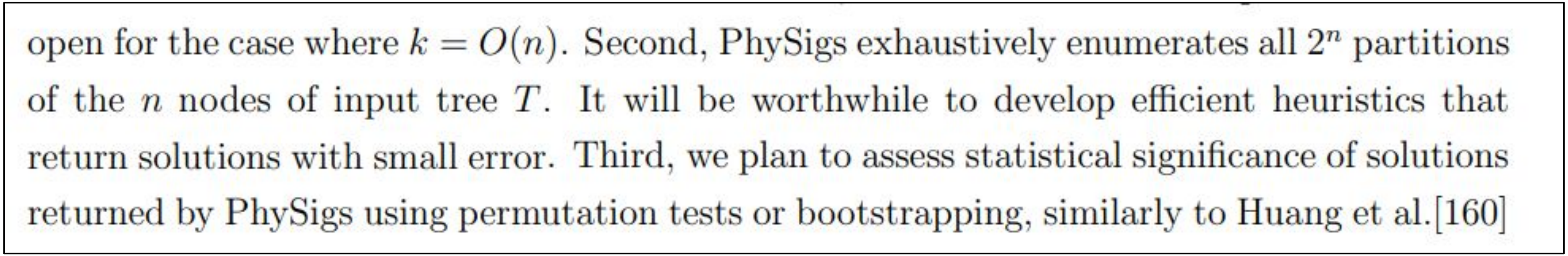
Enumerating Fewer than 2^n Partitions for PhySigs

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

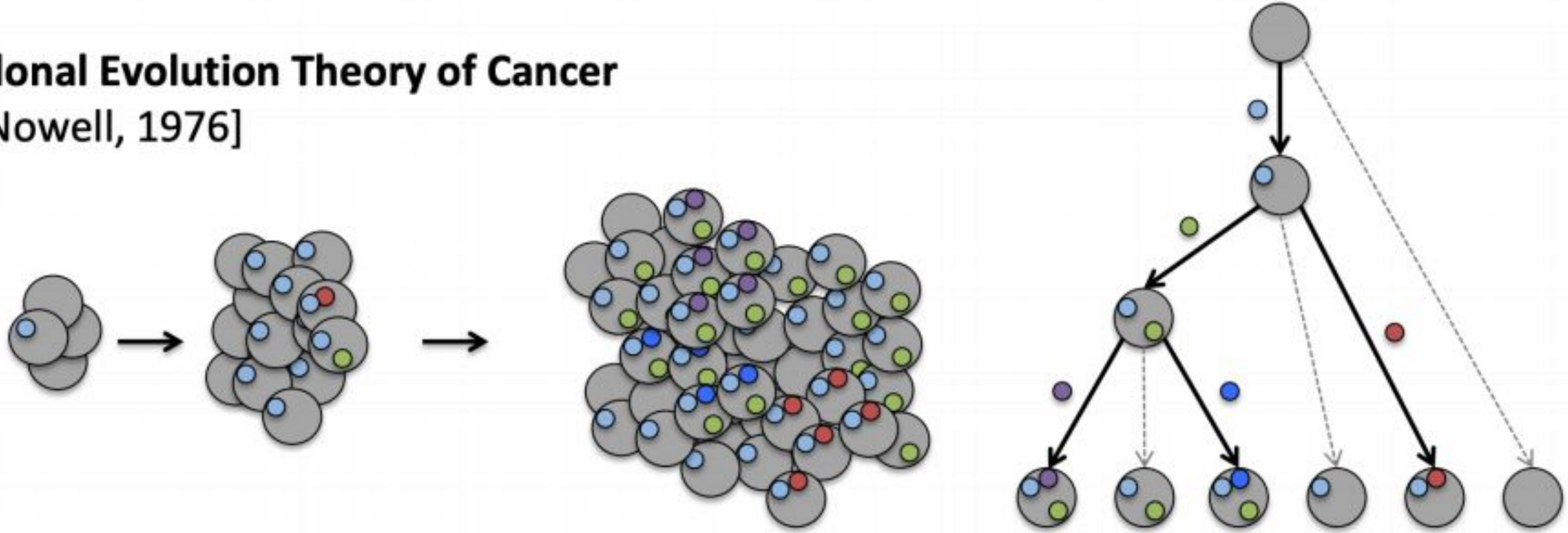
- Previous methods: either infer identical exposures (to mutational signatures) for all clones, or infer it independently for each clone
- PhySigs: generalize previous approaches

Problem

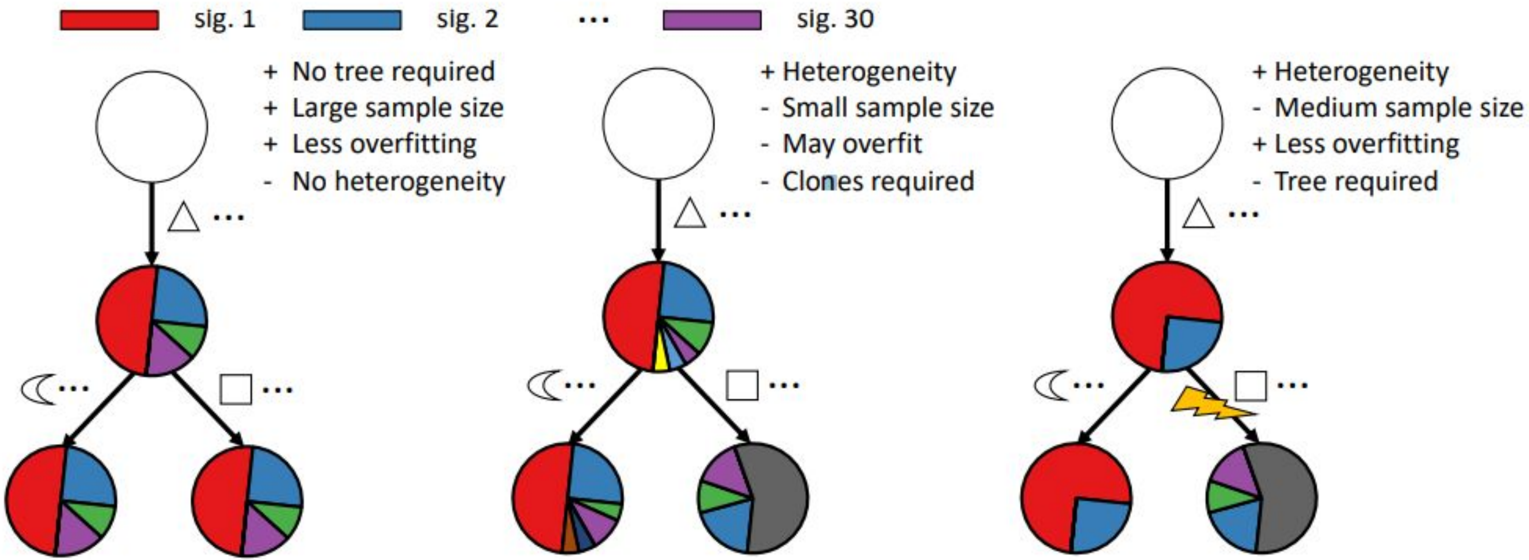
- Previous methods: either infer identical exposures (to mutational signatures) for all clones, or infer it independently for each clone
- PhySigs: generalize previous approaches
- (The next couple of slides are almost completely copied from Sarah Christensen’s talk)

Clonal Evolution Theory of Cancer

[Nowell, 1976]



Problem

- 

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

Same exposures for every clone

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

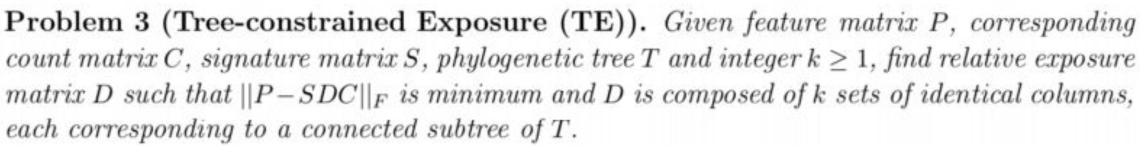
Different exposures for every clone

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

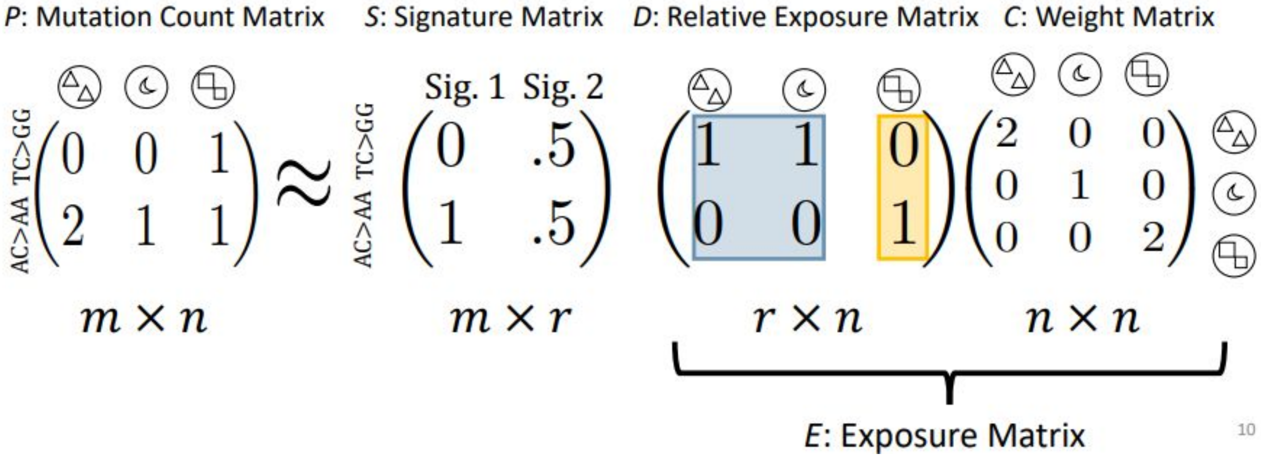
Clone exposures separated by shifts

Problem

-



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

$$\underbrace{\begin{array}{c}\text{TC>GG}\\\text{AC>AA}\end{array}\begin{pmatrix}0&0&1\\2&1&1\end{pmatrix}}_{m\times n} \approx \underbrace{\begin{array}{c}\text{TC>GG}\\\text{AC>AA}\end{array}\begin{pmatrix}0&.5\\1&.5\end{pmatrix}}_{m\times r} \underbrace{\underbrace{\begin{pmatrix}1&1&0\\0&0&1\end{pmatrix}}_{r\times n}\underbrace{\begin{pmatrix}2&0&0\\0&1&0\\0&0&2\end{pmatrix}}_{n\times n}}_{E\text{: Exposure Matrix}}$$

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 # of tumor clones

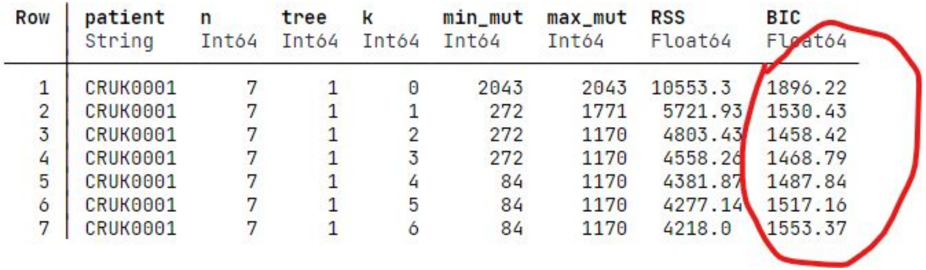
$$\mathrm{BIC}(L(k)) = mn\log(L(k)/(mn)) + kr\log(mn).$$

Goal

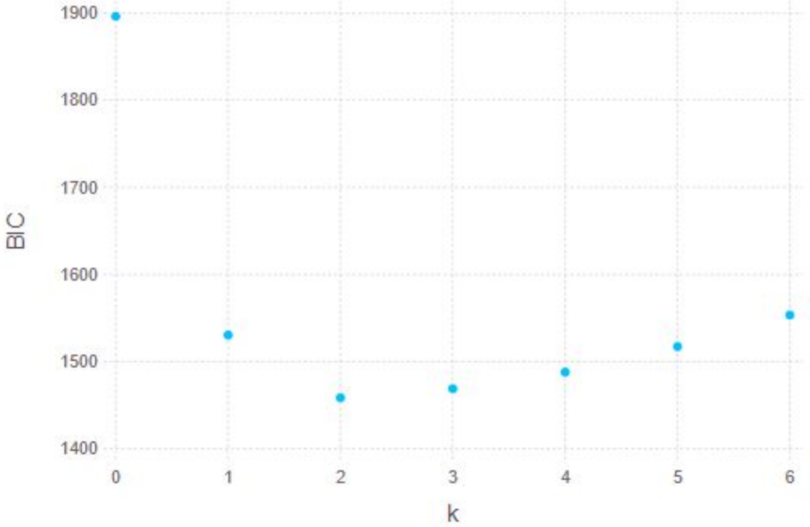
- Enumerate fewer partitions

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



```
Row │ patient   n      tree   k      min_mut  max_mut  RSS      BIC
    │ String    Int64  Int64  Int64  Int64    Int64    Float64  Float64
────┼──────────────────────────────────────────────────────────────────
  1 │ CRUK0001      7      1      0     2043     2043  10553.3  1896.22
  2 │ CRUK0001      7      1      1      272     1771   5721.93 1530.43
  3 │ CRUK0001      7      1      2      272     1170   4803.43 1458.42
  4 │ CRUK0001      7      1      3      272     1170   4558.26 1468.79
  5 │ CRUK0001      7      1      4       84     1170   4381.87 1487.84
  6 │ CRUK0001      7      1      5       84     1170   4277.14 1517.16
  7 │ CRUK0001      7      1      6       84     1170   4218.0  1553.37
```

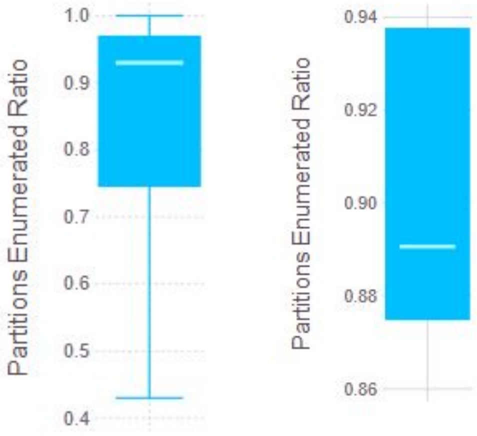


First line of attack: the BIC is (almost) unimodal?

- Obvious (maybe dangerous) heuristic: use some search algorithm for discrete unimodal distribution, say *ternary search*
- Asymptotically instead of searching for ϴ(n) choices of k, we now search for ϴ(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



TRACERx

Simulated

# of partitionings enumerated by PhySigs-TernarySearch / # of partitionings enumerated by PhySigs (Dryrun)

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