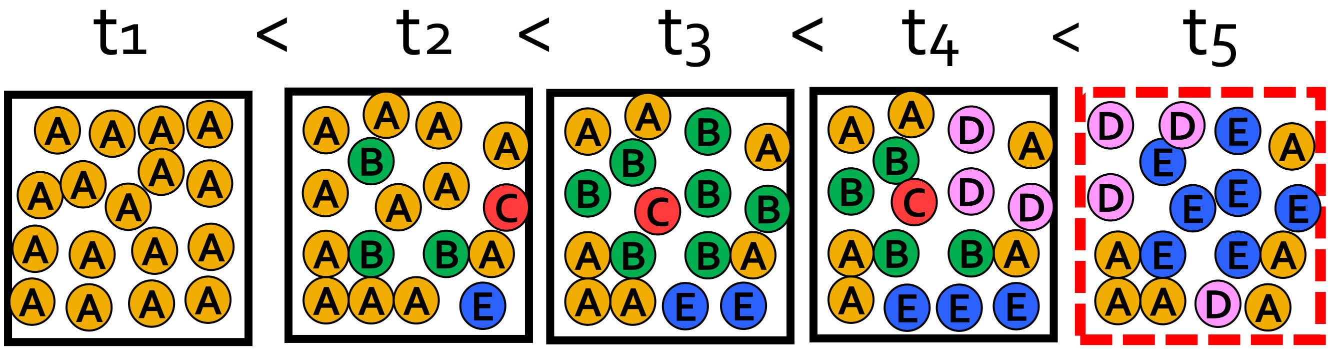
Charalampos (Babis) E. Tsourakakis
charalampos.tsourakakis@aalto.fi

Modeling Intratumor Gene Copy Number Heterogeneity using Fluorescence in Situ Hybridization data

WABI 2013, France

Tumor heterogeneity



C (4)

B (3)

A (2)

D (1)

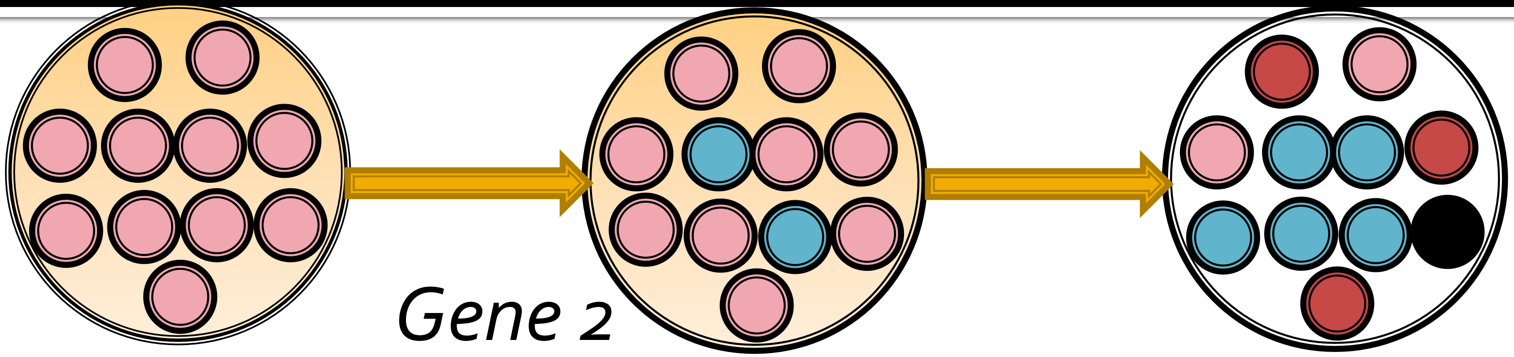
E (0)

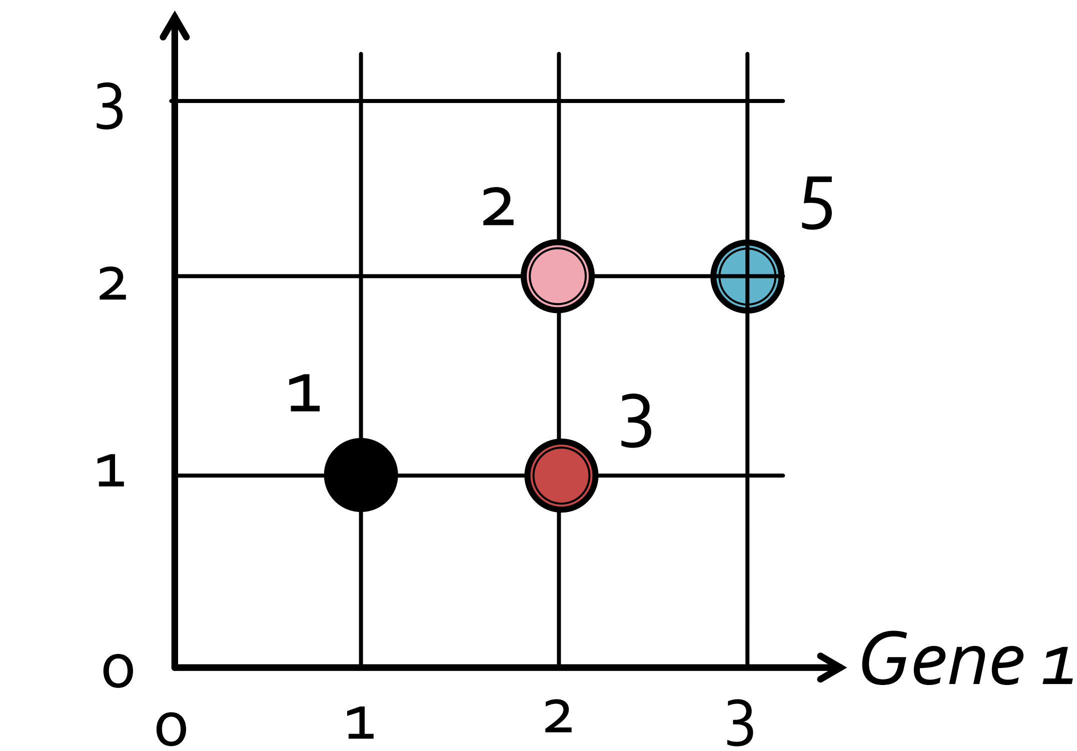
Copy numbers for a single gene

Tumor heterogeneity

- Inverse problem approach
  - High-throughput DNA sequencing data by *Oesper, Mahmoody, Raphael (Genome Biology 2013)*
  - SNP array data by Van Loo et al. (PNAS 2010), Carter et al. (Nature Biotechnology 2012)

Tumor heterogeneity

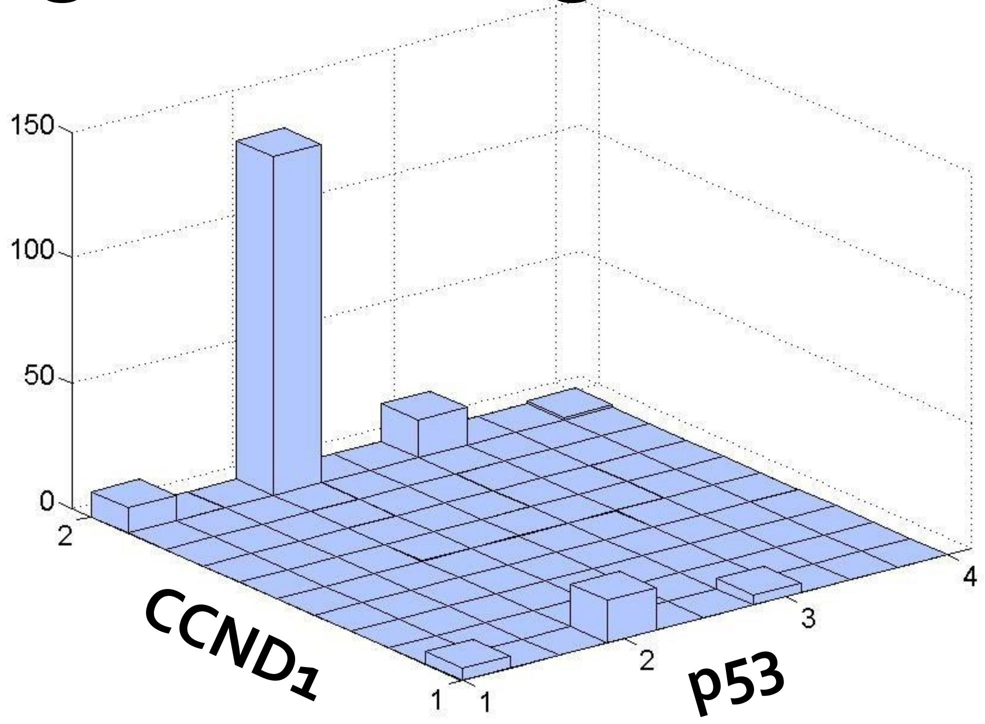




FISH data, direct assessment

FISH data

Multidimensional histogram on the positive integer cone, e.g., for 2 dimensions



FISH data

- Let xij be the number of copies of gene j in the i-th cell, where i=1,..,n(~100) and j=1,..,g(~10).
- The bounding box's size
  $$|[\min_i\ x_{i1}, \max_i x_{i1}] \times .. \times [\min_i\ x_{ig}, \max_i x_{ig}]|$$
  typically grows exponentially in the number of probes for the breast cancer datasets
  - This feature seems to be tumor dependent , i.e., does not hold necessarily for all cancers

FISH data

- Breast and cervical cancer data publicly available from NIH

ftp://ftp.ncbi.nlm.nih.gov/pub/FISHtrees/data

Motivation

- Understanding tumor heterogeneity is a key step towards:
  - find first mutation events, hence identify new drugs and diagnostics
  - predict response to selective pressure, hence develop strategies to avoid drug resistance
  - identify tumors likely to progress, hence avoid over- and under-treatment.

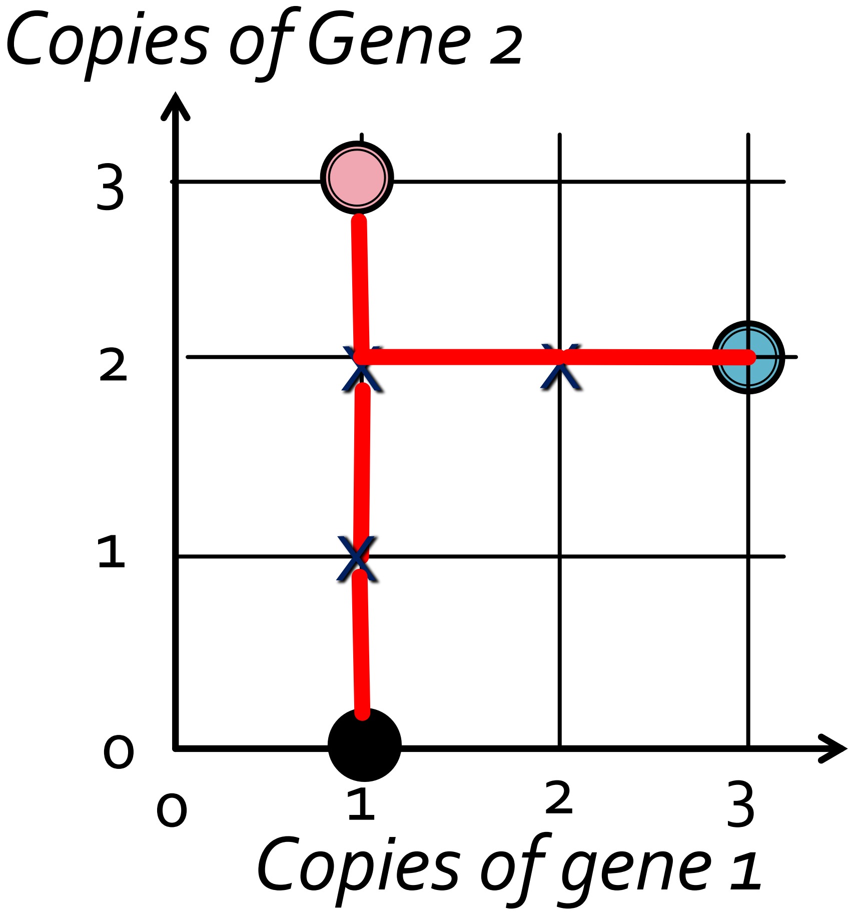
Related work

- Pennington, Smith, Shackney and Schwartz (J. of Bioinf. and Comp. B. 2007)
  - Two probes
  - Random walk where coordinate *i* is picked independently and with probabilities pi0,pi-1,pi1 is modified by {0,-1,+1} respectively.
  - Efficient heuristic to maximize a likelihood function over all possible trees and parameters.

Related work

- Chowdhury, Shackney, Heselmeyer-Haddad, Ried, Schäffer, Schwartz (Best paper in ISMB'13). Among other contributions:
  - Methods which are able to handle large number of cells and probes.
  - Exponential-time exact algorithm and an efficient heuristic for optimizing their objective
  - New test statistics, tumor classification
  - Extensive experimental evaluation

Related work



- Chowdhury et al.:
  - Problem: Find tree (and possibly Steiner nodes) to minimize cost of connecting all input (terminal) vertices

Contributions I

- Probabilistic approach
  - We summarize the empirical distribution based on a model that captures complex dependencies among probes without over-fitting.
  - Allows us to assign weights on the edges of the positive integer di-grid which capture how likely a transition is.
  - *And now, how do we derive a tumor phylogeny?..*

Proposed method

- Let X_j = #copies of gene j
  - integer valued random variable
  - Let I_j be the domain of X_j
- We model the joint probability distribution

$X = (X_1, .., X_g)$ as

$$\Pr(x) = \frac{1}{Z} \prod_{A \subseteq [g]} e^{\varphi_A(x)}$$

$x = (x_1, .., x_g)$

Potential function

Proposed method

- with the following properties of hierarchical log-linear model
  - log-linearity: the logarithm of each potential depends linearly on the parameters, e.g., for $\mathrm{g} = 2, \mathrm{I}_1 = \mathrm{I}_2 = \{0,1\}$ then,

$$\begin{aligned}\log \mathbf{Pr}\left[x\right] = {} & w_0 + w_{(1)0}\mathbb{1}\{x_1 = 0\} + w_{(1)1}\mathbb{1}\{x_1 = 1\} + w_{(2)0}\mathbb{1}\{x_2 = 0\} \\ & + w_{(2)1}\mathbb{1}\{x_2 = 1\} + w_{(12)00}\mathbb{1}\{x_1 = 0, x_2 = 0\} + w_{(12)01}\mathbb{1}\{x_1 = 0, x_2 = 1\} \\ & + w_{(12)10}\mathbb{1}\{x_1 = 1, x_2 = 0\} + w_{(12)11}\mathbb{1}\{x_1 = 1, x_2 = 1\},\end{aligned}$$

Proposed method

- Hierarchical:
  - $\mathrm{A} \subseteq B, w_A = 0 \rightarrow w_B = 0$
    - For instance $\mathrm{w}_{\{1,2,3\}}$ can be non-zero only if $\mathrm{w}_{\{1,2\}}, \mathrm{w}_{\{1,3\}}, \mathrm{w}_{\{2,3\}}$ are non-zero.
  - Allows significant computational savings compared to the general form
  - Biologically meaningful: if a set A of genes does not interact, any superset of A maintains this property.

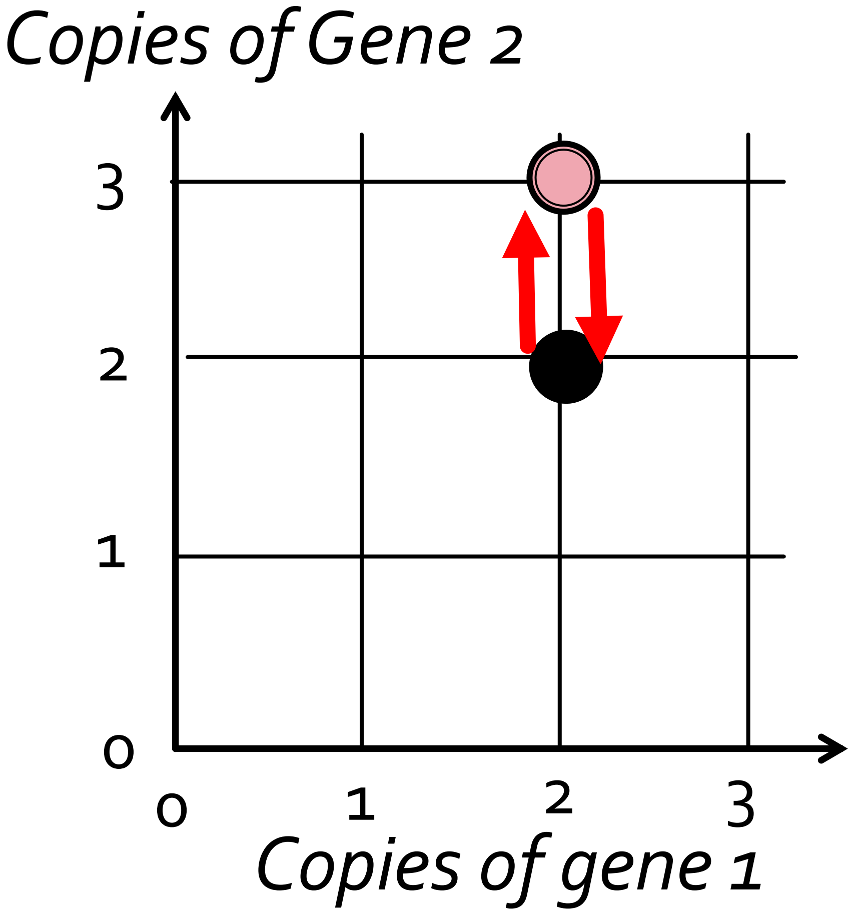
Proposed method

- A lot of related work and off-the-shelf software for learning the parameters
  - Based on Zhao, Rocha and Yu who provide a general framework that allows us to respect the ‘hierarchical’ property ..
  - … Schmidt and Murphy provide efficient optimization algorithms for learning a hierarchical log-linear model

Proposed method

- We use the learned hierarchical log-linear model in two ways
  - The non-zero weights provide us insights into dependencies of factors
  - We use them to assign weights on the positive integer di-grid

Proposed method

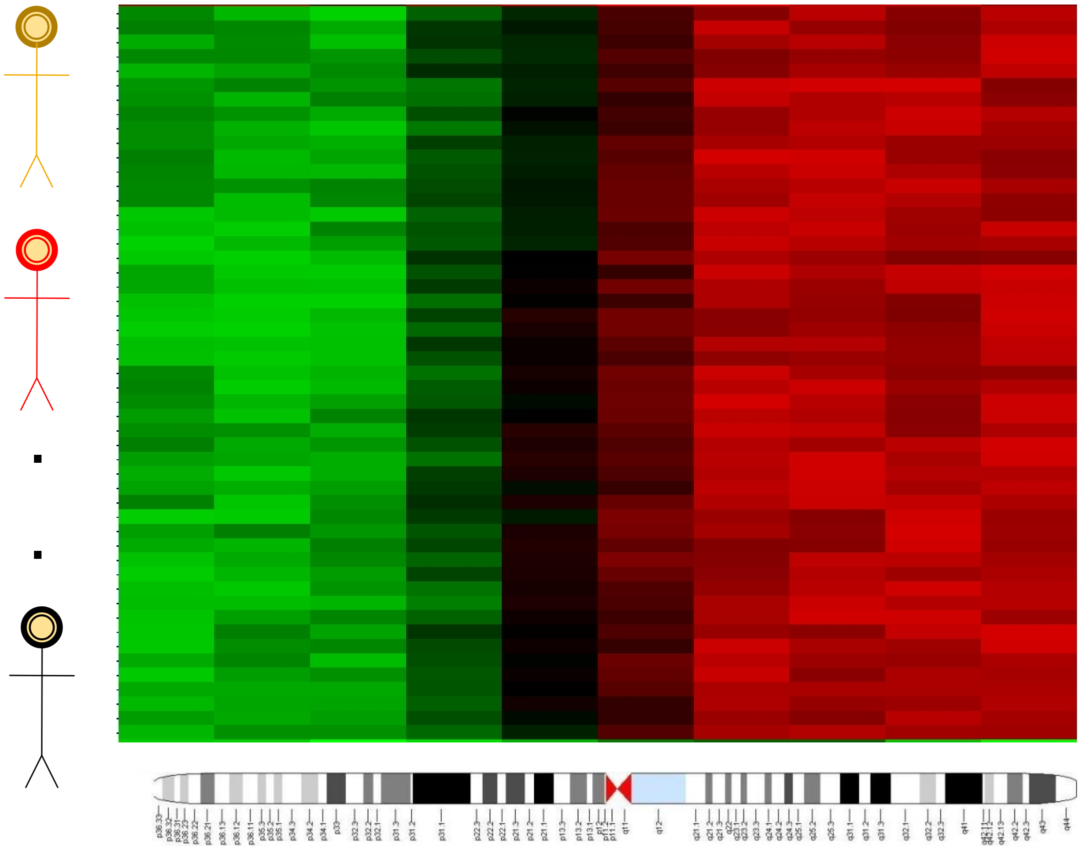


Nicholas Metropolis

Given a probability distribution ***π*** on a state space we can define a Markov Chain whose stationary distribution is ***π***.

Contributions II

- Question: Can we use the *wealth* of inter-tumor phylogenetic methods to understand intra-tumor cancer heterogeneity?



Contributions II

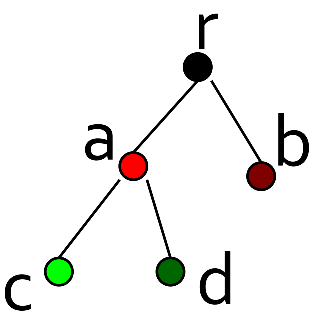
- Motivated by this question:
  - We prove necessary and sufficient conditions for the reconstruction of oncogenetic trees, a popular method for inter-tumor cancer inference
  - We exploit these to preprocess a FISH dataset into an inter-tumor cancer dataset that respects specific biological characteristics of the evolutionary process

Oncogenetic Trees

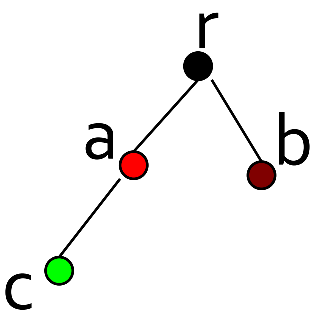
- Desper, Jiang, Kallioniemi, Moch, Papadimitriou, Schäffer
  - T(V,E,r) rooted branching
  - F={A_1,..,A_m} where A_i is the set of vertices of a rooted sub-branching of T.
  - What are the properties that F should have in order to uniquely reconstruct T?
    - Let T be consistent with F if it could give rise to F.

Example

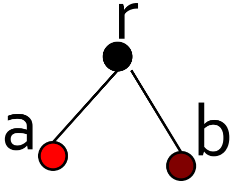
Onco-tree



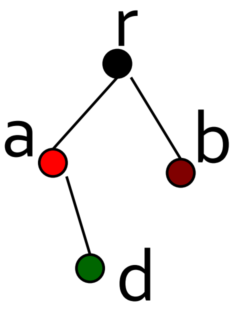
Patient 1, A1 ={ r,a,b,c}



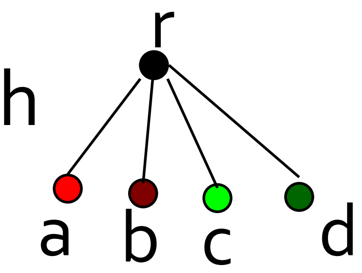
Patient 2, A2 ={ r,a,b}

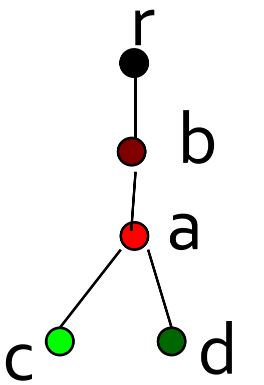


Patient 3, A3 ={ r,a,b,d}



Also, consistent with {A1, A2, A3}

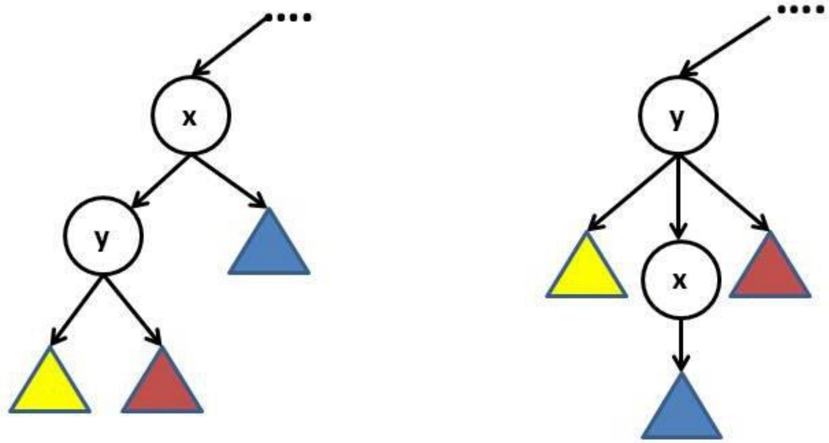




Oncogenetic Trees

- Theorem
  - The necessary and sufficient conditions to reconstruct T from F are the following:
    - x,y such that (x,y) is an edge, there exists a set in the family that contains x but not y.

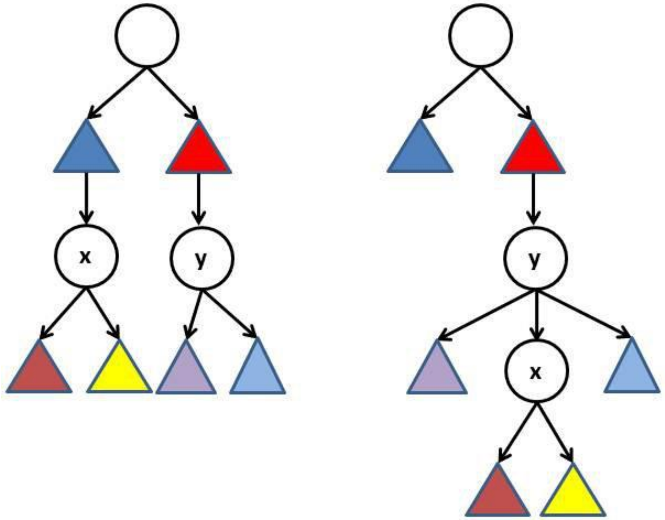
necessity



Oncogenetic Trees

- If x is not a descendant of y and vice versa then there exist two sets A_i, A_j such that
  - x is in A_i but not in A_j
  - y is in A_j but not in A_i

necessity



Oncogenetic trees

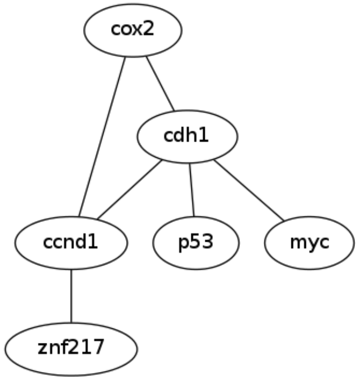
- It turns out that the necessary conditions are sufficient (constructive proof)
- Allows us to force an oncogenetic tree to capture certain aspects of intratumor heterogeneity dynamics

Contributions III

- We evaluate our method on real FISH data where we show findings consistent with cancer literature
    - Here, we show results for a breast cancer dataset

Experimental results

- No ground truth, but
  - concurrent loss of *cdh1* function and *p53* inactivation play a key role in breast cancer evolution
  - subsequent changes in *ccnd1, myc, znf217* according to our tree are consistent with oncogenetic literature



Conclusions

- There exists a lot of interest in understanding intra-tumor heterogeneity
  - Releasement of FISH data that assess it directly can promote this understanding

- Concerning our work:
  - Better algorithms for fitting the model
  - Allow higher-order interactions but use additional penalty (e.g., AIC)

Conclusions

- … concerning our work
  - Other choices of inter-tumor methods
  - Tumor classification applications
  - Consensus FISH trees
  - Allow more mechanisms in copy number changes
- Understand better the connection between our work and Chowdhury et al.

Acknowledgements

Russell Schwartz

Alejandro Schäffer

Carnegie Mellon University

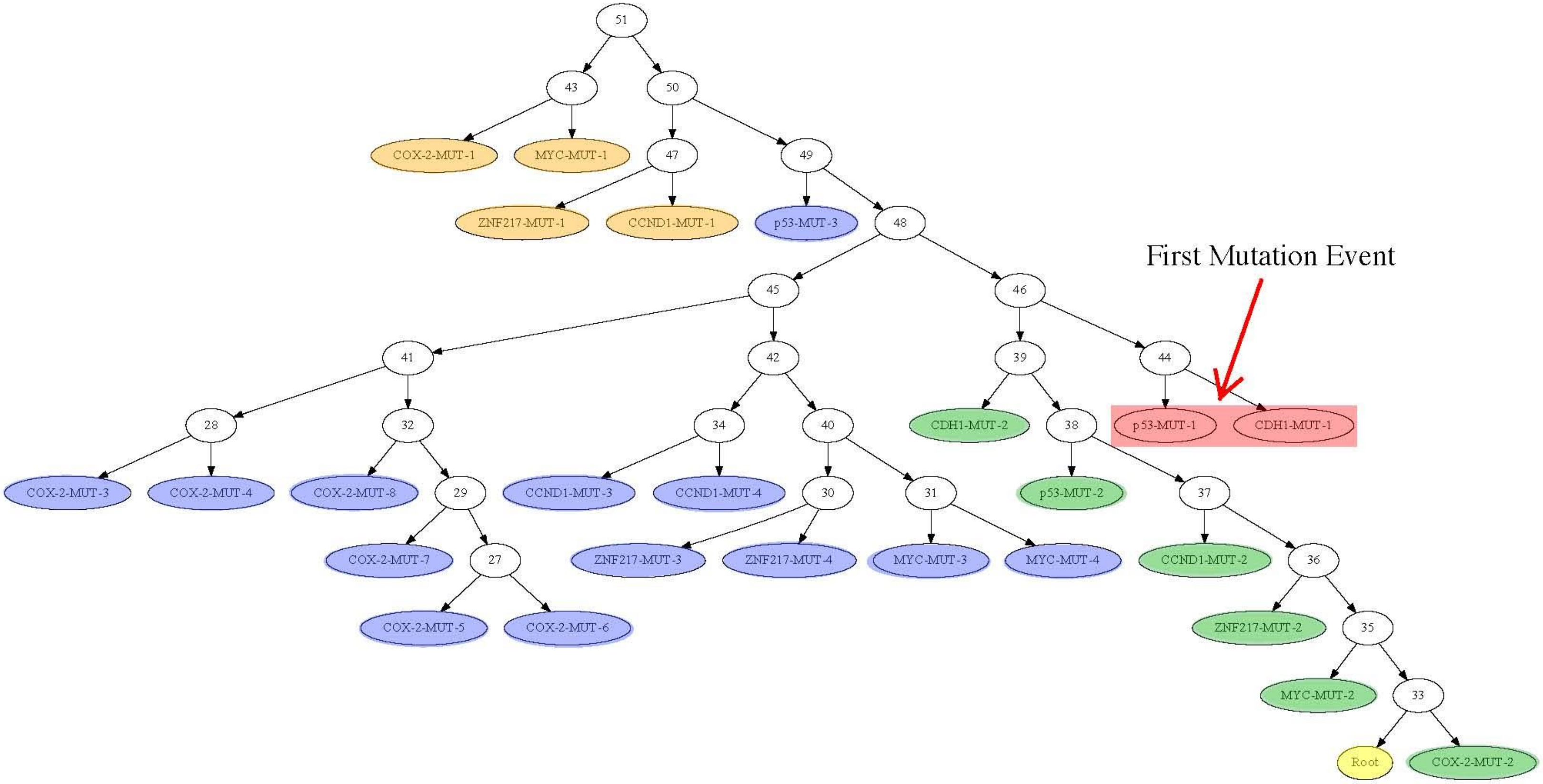
NSF

NSF Grant CCF-1013110

Thanks!

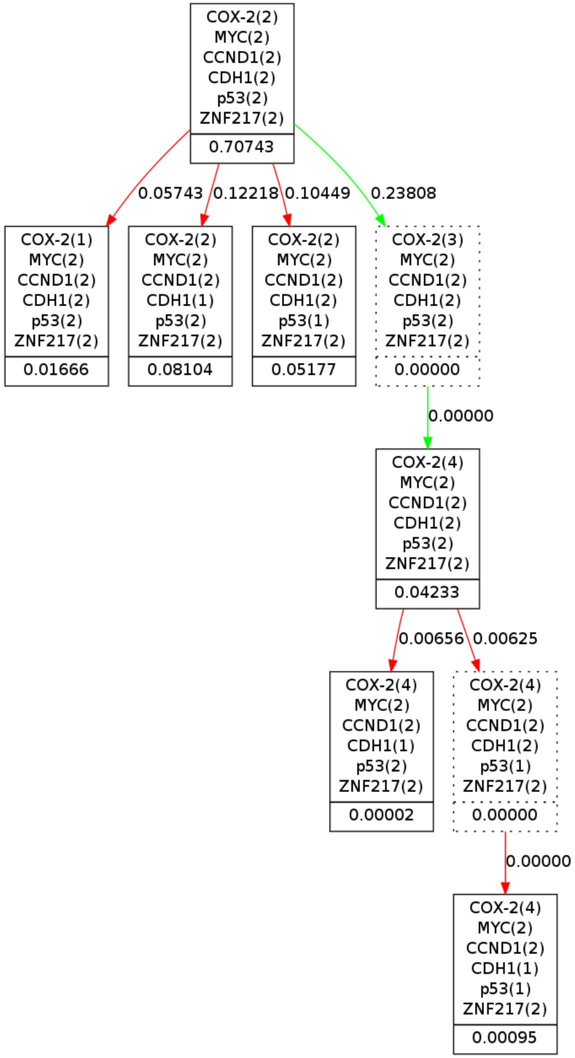
Appendix

Experimental results



First Mutation Event

Experimental results


COX-2(2) MYC(2) CCND1(2) CDH1(2) p53(2) ZNF217(2)
0.70743
0.05743
0.12218
0.10449
0.23808
COX-2(1) MYC(2) CCND1(2) CDH1(2) p53(2) ZNF217(2)
0.01666
COX-2(2) MYC(2) CCND1(2) CDH1(1) p53(2) ZNF217(2)
0.08104
COX-2(2) MYC(2) CCND1(2) CDH1(2) p53(1) ZNF217(2)
0.05177
COX-2(3) MYC(2) CCND1(2) CDH1(2) p53(2) ZNF217(2)
0.00000
0.00000
COX-2(4) MYC(2) CCND1(2) CDH1(2) p53(2) ZNF217(2)
0.04233
0.00656
0.00625
COX-2(4) MYC(2) CCND1(2) CDH1(1) p53(2) ZNF217(2)
0.00002
COX-2(4) MYC(2) CCND1(2) CDH1(2) p53(1) ZNF217(2)
0.00000
0.00000
COX-2(4) MYC(2) CCND1(2) CDH1(1) p53(1) ZNF217(2)
0.00095

Generated with code available at ftp://ftp.ncbi.nlm.nih.gov/pub/FISHtree