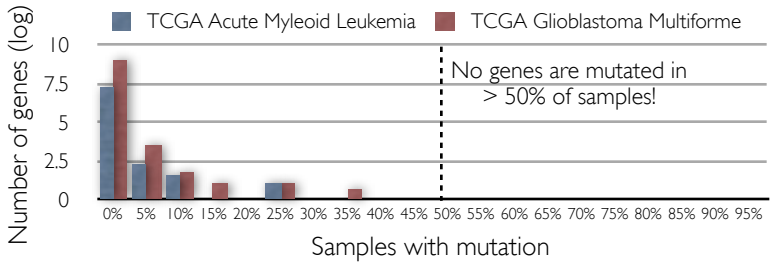
Methods for Identifying Driver Pathways in Cancer

Max Leiserson[1], Hsin-Ta Wu[1], Dima Blokh[2], Fabio Vandin[1], Roded Sharan[2], Benjamin Raphael[1]
[1]Dept. of Computer Science and Center for Computational Molecular Biology, Brown University
[2]School of Computer Science, Tel-Aviv University
mdml@cs.brown.edu

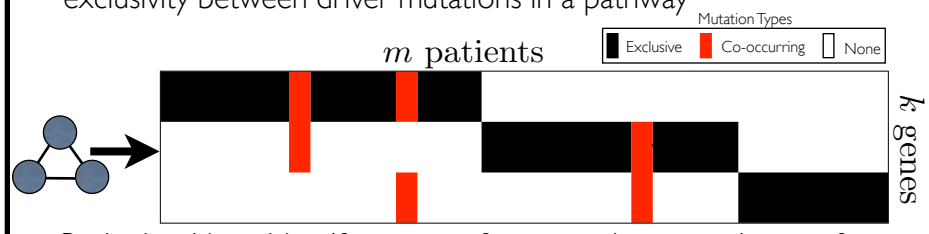
I. Motivation

- **Key challenge**: distinguish the somatic mutations that *drive* cancer development from random *passenger* mutations, and identify driver pathways, i.e. meaningful combinations of driver mutations
- Different combinations of driver mutations are observed in patients with the same cancer type because driver mutations target cellular signaling and regulatory pathways
- Measurement of somatic mutations in large numbers of cancer genomes provides opportunity to identify **novel** driver pathways



II. Contributions

- Two algorithms, **Dendrix++** and **Multi-Dendrix**, for identifying driver pathways
- Both algorithms extend Dendrix (De novo driver exclusivity) [Vandin *et al.* 2012]
- Both algorithms use the combinatorial constraint of mutual exclusivity between driver mutations in a pathway



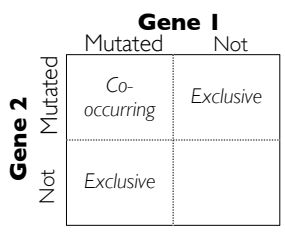
- Both algorithms identify groups of recurrently mutated genes from genome-scale data in hundreds of cancer patients
- Both algorithms rigorously assess the statistical significance of discovered groups

III. Algorithms and Methods

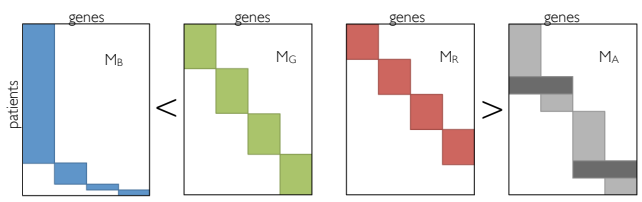
Dendrix++

Goal: Identify all combinations of genes that show surprising exclusivity among their mutations, regardless of coverage (number of mutations).

Main Idea: quantifies exclusivity of a set of genes probabilistically (with contingency tables).



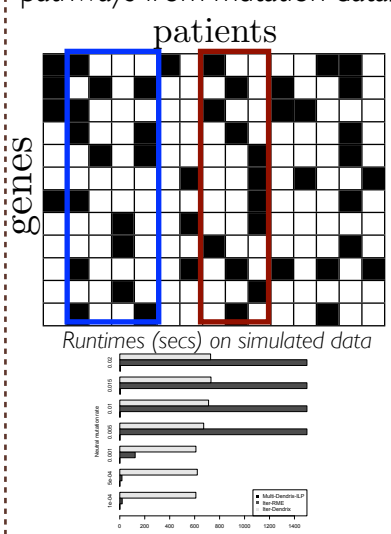
- The scoring function is less influenced by overall frequency than the Dendrix scoring function.
- Dendrix++ can find combinations with lower frequency, rare mutations.
- Dendrix++ distinguishes between combinations with more surprising mutation patterns.



http://compbio.cs.brown.edu/projects/dendrix++

Multi-Dendrix

Goal: Find t cancer pathways from mutation data.



Main idea: it is well known that mutations in several pathways are generally required for cancer. By simultaneously identifying *sets* of driver pathways, **Multi-Dendrix** finds more subtle combinations of pathways.

- **W'(M)** = weight of sets **M** of genes that quantifies coverage and exclusivity within each set.
- Finding **M*** that maximizes W'(**M**) is NP-hard.
- ILP rapidly identifies **M*** on genome-scale data.
- Statistical significance computed via a permutation test.

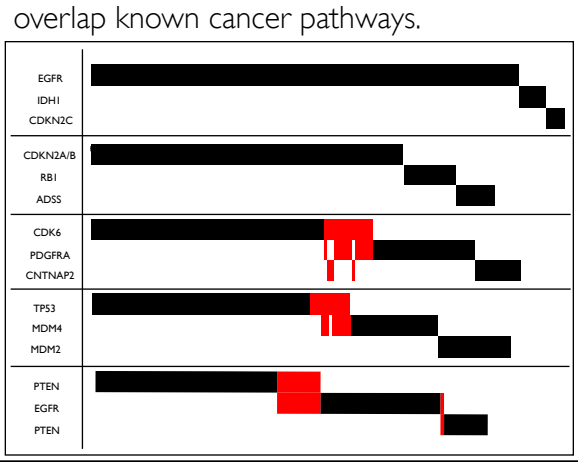
http://compbio.cs.brown.edu/projects/multi-dendrix

IV. Results

TCGA Glioblastoma Multiforme

Data: Whole-genome and array copy number data from 224 GBM patients.

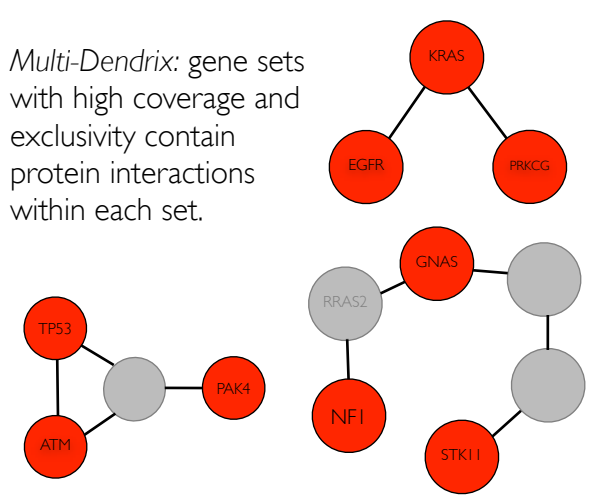
Multi-Dendrix: high coverage & exclusive sets overlap known cancer pathways.



Lung adenocarinoma [Ding *et al.*]

Data: Targeted gene sequencing of 190 genes in 163 Lung Adenocarcinoma patients.

Multi-Dendrix: gene sets with high coverage and exclusivity contain protein interactions within each set.



TCGA Acute Myeloid Leukemia [unpublished]

Data: Somatic mutations (whole-exome sequencing) and fusion genes (RNA-seq) from 200 AML patients.

Dendrix++: high coverage and exclusive sets overlap gene categories implicated in AML development.