Simultaneous Identification of Multiple Driver Pathways in Cancer

Mark D. M. Leiserson, et.al
Goal

• To distinguish the functional driver mutations responsible for cancer development from the random passenger mutations that have no consequences for cancer.
Distinguishing driver vs. passenger

- Strategies for identifying the driver mutations ([Ding2014](#))
  - Recurrence and frequency assessment
  - Variant effect prediction
  - Pathway or network analysis
Multi-Dendrix

• Dendrix – De novo Driver Exclusivity
• **Important Assumption:**
  1) High Coverage- most patients have at least one mutation in the set, i.e, set of potential mutated genes of a particular pathway
  2) High Exclusivity- nearly all patients have no more than one mutation in the set
Justification by the author

will have a mutation in some gene in the pathway. Second, a driver mutation in a single gene of the pathway is often assumed to be sufficient to perturb the pathway. Combined with the fact that driver mutations are relatively rare, most patients exhibit only a single driver mutation in a pathway. Thus, we expect that the genes in a pathway exhibit a pattern of mutually exclusive driver mutations, where driver mutations are observed in exactly one gene in the pathway in each patient (Vogelstein and Kinzler 2004; Yeang et al. 2008). There are numerous examples of pairs of

From Vandin, et al, 2012
Dendrix - Method

From Vandin, et al, 2012
Dendrix Method

**Maximum Coverage Exclusive Submatrix Problem:** Given an $m \times n$ mutation matrix $A$ and an integer $k > 0$, find a mutually exclusive $m \times k$ submatrix of $M$ of $k$ columns (genes) of $A$ with the largest number of nonzero rows (patients).

**Coverage Overlap**

Denote the set of patients in which $g$ is mutated

$$\omega(M) = \sum_{g \in M} |\Gamma(g)| - |\Gamma(M)|.$$  

Denote the set of patients in which at least one of the genes in $M$ is mutated

**Weight**

$$W(M) = |\Gamma(M)| - \omega(M) = 2|\Gamma(M)| - \sum_{g \in M} |\Gamma(g)|.$$  

From Vandin, et al, 2012
Dendrix Method

**Maximum Weight Submatrix Problem:** Given an $m \times n$ mutation matrix $A$ and an integer $k > 0$, find the $m \times k$ column submatrix $M$ of $A$ that maximizes $W(M)$.

**Problem**
Computationally Difficult to Solve

Size $k = 6$ of 20,000 genes

$10^23$ subsets

**Solution**
A greedy Algorithm for independent genes
Markov Chain Monte Carlo (MCMC)

From Vandin, et al, 2012
Limitation of Dendrix

- Mutations in different pathways may not be mutually exclusive.
- Mutations in different pathways may exhibit significant patterns of co-occurrence across patients.
- Solution -> Multi-Dendrix Algorithm
Multi-Dendrix Algorithm

• 1) Find sets of genes with high coverage as an integer linear program (ILP)
• 2) Generalize the ILP to simultaneously find multiple driver pathways
• 3) Additional Analysis: Subtype-specific mutations, stability measures, permutation test, compute enrichment states
The Multi-Dendrix Pipeline
Multi-Dendrix Method - the same as the first step of Dendrix

**Maximum Coverage Exclusive Submatrix Problem:** Given an $m \times n$ mutation matrix $A$ and an integer $k > 0$, find a mutually exclusive $m \times k$ submatrix of $M$ of $k$ columns (genes) of $A$ with the largest number of nonzero rows (patients).

**Coverage Overlap**

Denote the set of patients in which $g$ is mutated

$$\omega(M) = \sum_{g \in M} |\Gamma(g)| - |\Gamma(M)|.$$

Denote the set of patients in which at least one of the genes in $M$ is mutated

**Weight**

$$W(M) = |\Gamma(M)| - \omega(M) = 2|\Gamma(M)| - \sum_{g \in M} |\Gamma(g)|.$$
ILP- Integer Linear Programming

- Mathematical optimization or feasibility program where variables are restricted to be integers

The graph on the right shows the following problem.

\[
\begin{align*}
\text{max } y \\
-x + y &\leq 1 \\
3x + 2y &\leq 12 \\
2x + 3y &\leq 12 \\
x, y &\geq 0 \\
x, y &\in \mathbb{Z}
\end{align*}
\]

From Wikipedia
ILP for the Maximum Weight Submatrix Problem

Mutation matrix:

\[ A_{ij} = \begin{cases} 
1 & \text{if gene } j \text{ is mutated in patient } i \\
0 & \text{otherwise.} 
\end{cases} \]  
(1)

For each gene \( j \), a gene set \( M \) is determined by

\[ I_M(j) = \begin{cases} 
1 & \text{if gene } j \text{ is a member of gene set } M, \\
0 & \text{otherwise.} 
\end{cases} \]  
(3)

For each patient \( I \), the coverage is determined by

\[ C_I(M) = \begin{cases} 
1 & \text{if gene set } M \text{ is mutated in patient } i, \\
0 & \text{otherwise.} 
\end{cases} \]  
(4)
Then, $Dendrix_{ILP}(k)$ is defined as follows:

$$\text{maximize } \sum_{i=1}^{m} \left( 2 \cdot C_i(M) - \sum_{j=1}^{n} I_M(j) \cdot A_{ij} \right)$$  \hspace{1cm} (5a)$$

$$k_{\text{min}} \leq \sum_{j=1}^{n} I_M(j) \leq k_{\text{max}}.$$  \hspace{1cm} (5b)

$$\omega(M) = \sum_{g \in M} |\Gamma(g)| - |\Gamma(M)|.$$  \hspace{1cm} (5c)

Denote the set of patients in which $g$ is mutated

Denote the set of patients in which at least one of the genes in $M$ is mutated
Multiple Maximum Weight Submatrices Problem: Given an \( m \times n \) mutation matrix \( A \) and an integer \( t > 0 \), find a collection \( M = \{ M_1, M_2, \ldots, M_t \} \) of \( m \times k \) column submatrices that maximizes

\[
W'(M) = \sum_{\rho=1}^{t} W(M_{\rho}).
\]

\[
\text{maximize } \sum_{\rho=1}^{t} \sum_{i=1}^{m} \left( 2 \cdot C_i(M_{\rho}) - \sum_{j=1}^{n} I_{M_{\rho}}(j) \cdot A_{ij} \right) \quad (7a)
\]

subject to

\[
\sum_{j=1}^{n} I_{M_{\rho}}(j) \cdot A_{ij} \geq C_i(M_{\rho}), \quad (7b)
\]

for \( 1 \leq i \leq m, 1 \leq \rho \leq t, \)

\[
\sum_{\rho=1}^{t} I_{M_{\rho}}(j) \leq 1, 1 \leq j \leq m. \quad (7c)
\]
Multi-Dendrix Results on the GBM Dataset
Multi-Dendrix Results on the BRCA Dataset

Network Diagram Legend
- p53 Signaling
- PI(3)/K/ATK Signaling
- Cell cycle checkpoints
- p38-JNK1
- Other

Subtype annotations legend
- Basal-like
- Luminal A
- Luminal B
- HER2-enriched
- Normal-like

Coverage: 61% (308/507) of samples

Coverage: 56% (283/507) of samples
• Thank you for your attention!