HotNet

Background

- Determine significantly mutated subnetworks in a large gene interaction network

- Problems with current methods
  - Frequency doesn’t always predict significance
  - Naïve subnetwork analysis
    - Enumeration prohibits subnetworks of reasonable size
    - Large number of hypotheses makes statistically significant difficult
    - Hub genes make for small gene diameters

HotNet Overview

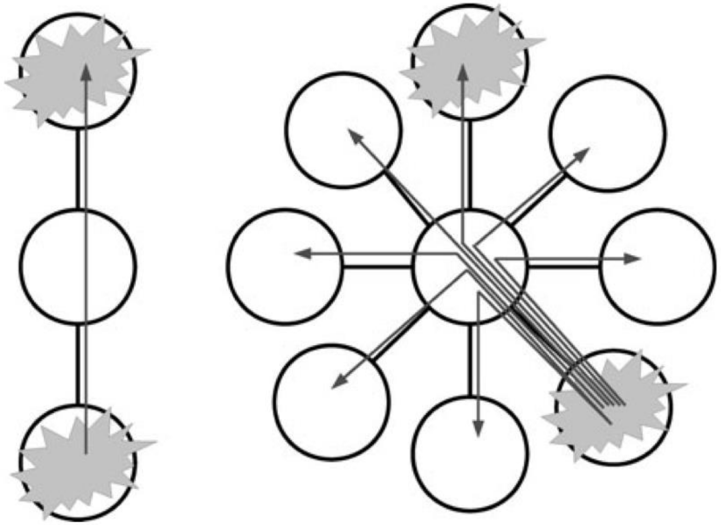
1. Formulate an influence measure between pairs of genes in the network

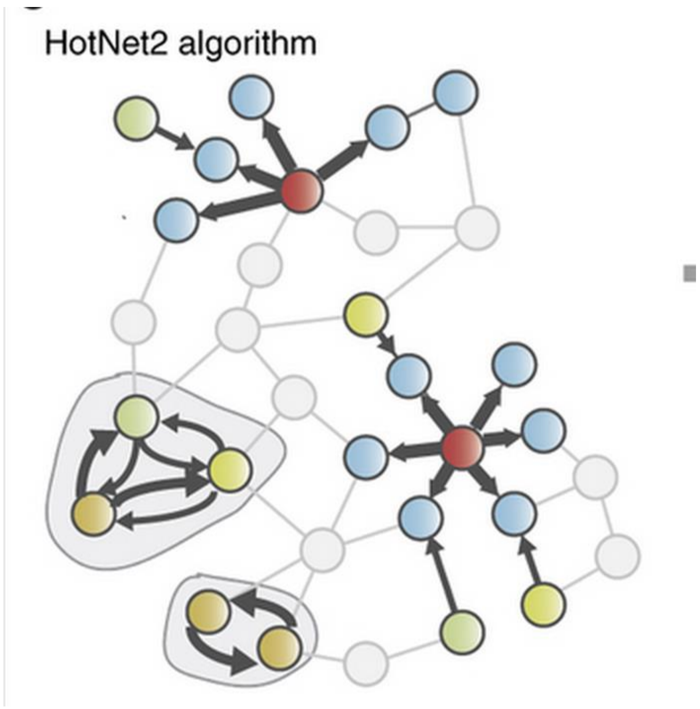
2. Identify subnetworks with *Combinatorial Model* or *Enhanced Influence Model*

3. Two-stage multiple hypothesis test to mitigate testing of large number of hypotheses

Influence Graph

- Identify subnetworks that are significant with respect to a set of mutated genes





Diffusion

$f_v^s(t)$ Amount of fluid @ node *V* at time *T*

$\mathbf{f}^s(t) = [f_1^s(t), \ldots, f_n^s(t)]^T$ Amount of fluid at all nodes

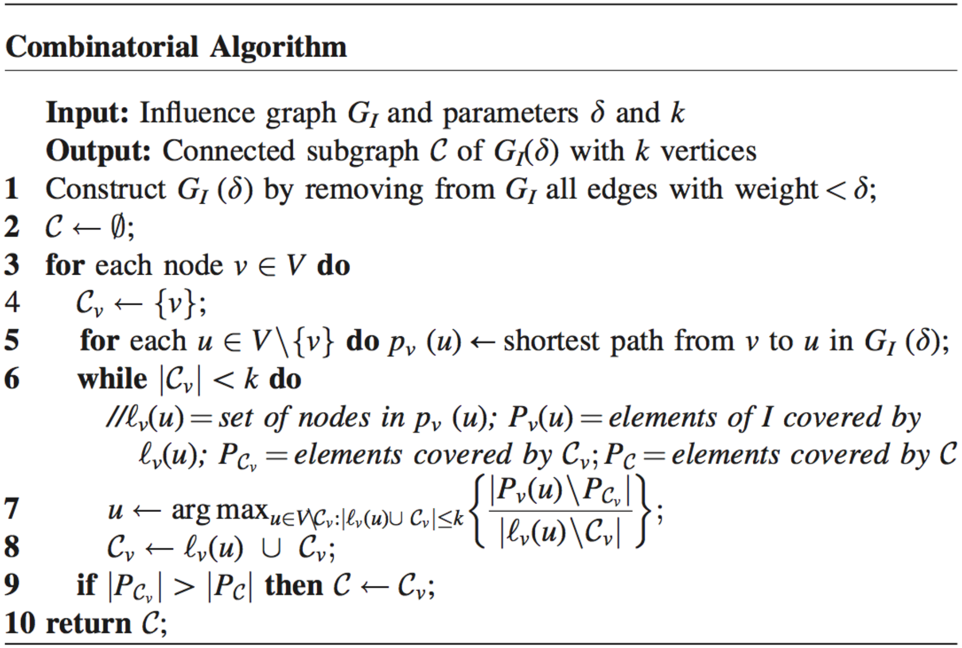
$L_\gamma = L + \gamma I.$ L is the laplacian matrix of the graph

$\frac{d\mathbf{f}^s(t)}{dt} = -L_\gamma \mathbf{f}^s(t) + \mathbf{b}^s u(t),$ Dynamics of the continuous process

- Interpret f_i as the influence of gene g_s on g_i

Combinatorial Model

- Takes in influence measure between genes to discover significant subnetworks

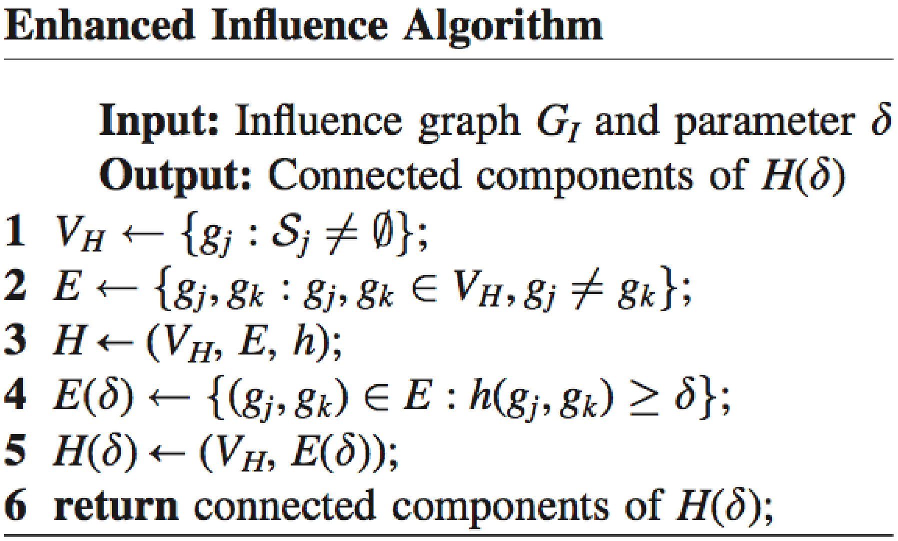


Combinatorial Algorithm

```
Input: Influence graph $G_I$ and parameters $\delta$ and $k$
Output: Connected subgraph $\mathcal{C}$ of $G_I(\delta)$ with $k$ vertices
Construct $G_I(\delta)$ by removing from $G_I$ all edges with weight $< \delta$;
$\mathcal{C} \leftarrow \emptyset$;
for each node $v \in V$ do
    $\mathcal{C}_v \leftarrow \{v\}$;
    for each $u \in V \setminus \{v\}$ do $p_v(u) \leftarrow$ shortest path from $v$ to $u$ in $G_I(\delta)$;
    while $|\mathcal{C}_v| < k$ do
        //$\ell_v(u) =$ set of nodes in $p_v(u)$; $P_v(u) =$ elements of $I$ covered by
          $\ell_v(u)$; $P_{\mathcal{C}_v} =$ elements covered by $\mathcal{C}_v$; $P_{\mathcal{C}} =$ elements covered by $\mathcal{C}$
        $u \leftarrow \arg\max_{u \in V \setminus \mathcal{C}_v : |\ell_v(u) \cup \mathcal{C}_v| \leq k} \left\{ \frac{|P_v(u) \setminus P_{\mathcal{C}_v}|}{|\ell_v(u) \setminus \mathcal{C}_v|} \right\}$;
        $\mathcal{C}_v \leftarrow \ell_v(u) \cup \mathcal{C}_v$;
    if $|P_{\mathcal{C}_v}| > |P_{\mathcal{C}}|$ then $\mathcal{C} \leftarrow \mathcal{C}_v$;
return $\mathcal{C}$;
```

Enhanced Influence Model

- Enhance the influence measure between genes by the number of mutations observed in each gene



Enhanced Influence Algorithm

```
    **Input:** Influence graph $G_I$ and parameter $\delta$
    **Output:** Connected components of $H(\delta)$
$V_H \leftarrow \{g_j : \mathcal{S}_j \neq \emptyset\}$;
$E \leftarrow \{g_j, g_k : g_j, g_k \in V_H, g_j \neq g_k\}$;
$H \leftarrow (V_H, E, h)$;
$E(\delta) \leftarrow \{(g_j, g_k) \in E : h(g_j, g_k) \geq \delta\}$;
$H(\delta) \leftarrow (V_H, E(\delta))$;
**return** connected components of $H(\delta)$;
```

Statistics

- Calibrates with H_o^{sample} and H_0^{gene}
  - – Sample: mutations placed at random nodes
  - – Gene: Move genes around…..?
- Compute significance of **number** of subnetworks
- Bound FDR

Experimental Data

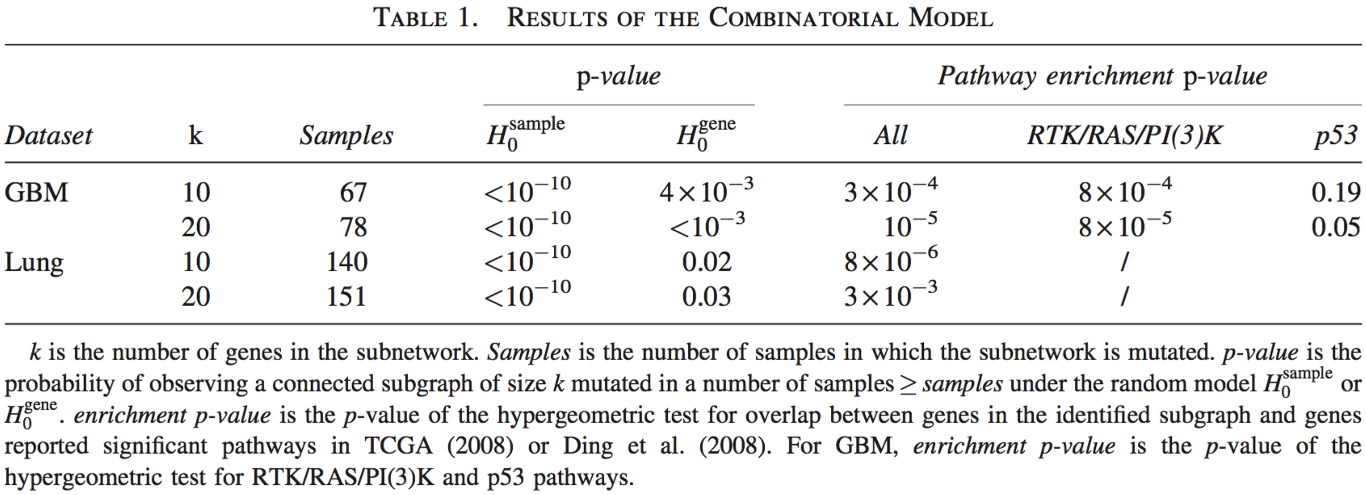
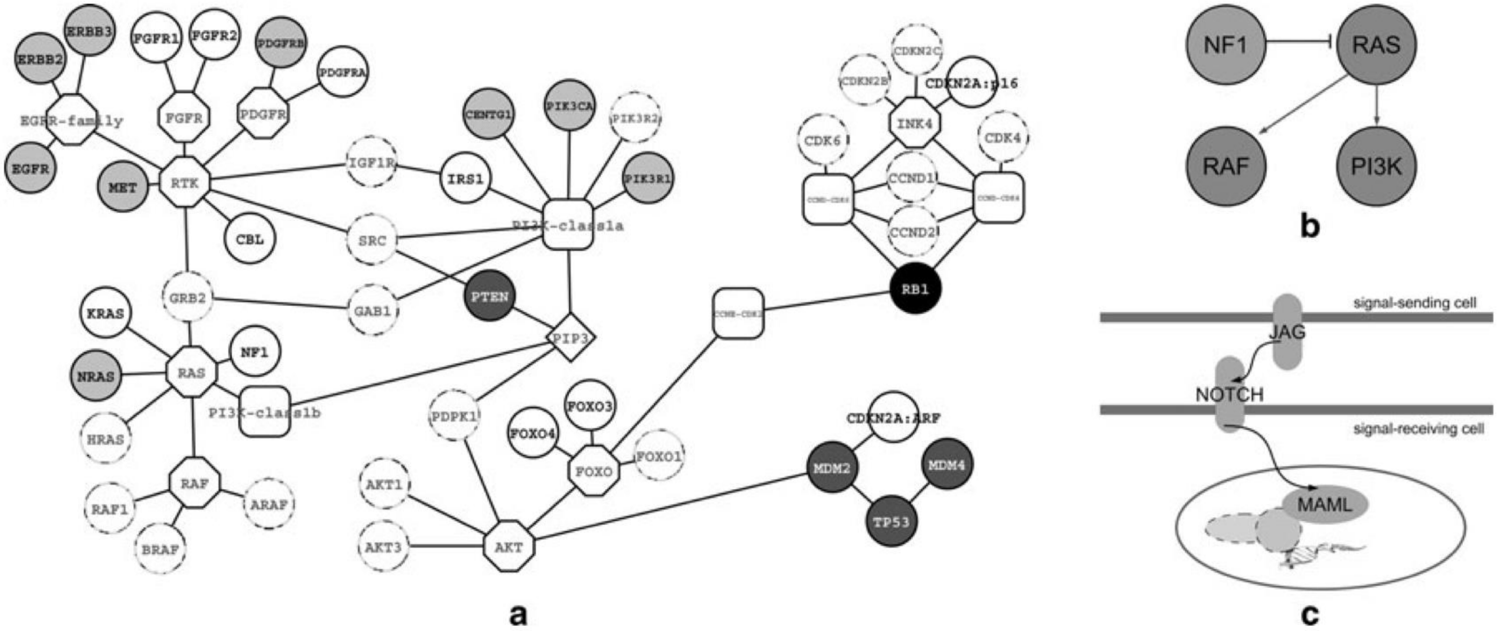


TABLE 1. RESULTS OF THE COMBINATORIAL MODEL

			p-*value*		*Pathway enrichment* p-*value*		
Dataset	k	*Samples*	H_0^{sample}	H_0^{gene}	*All*	*RTK/RAS/PI(3)K*	*p53*
GBM	10	67	$<10^{-10}$	4×10^{-3}	3×10^{-4}	8×10^{-4}	0.19
	20	78	$<10^{-10}$	$<10^{-3}$	10^{-5}	8×10^{-5}	0.05
Lung	10	140	$<10^{-10}$	0.02	8×10^{-6}	/	
	20	151	$<10^{-10}$	0.03	3×10^{-3}	/	

k is the number of genes in the subnetwork. *Samples* is the number of samples in which the subnetwork is mutated. *p-value* is the probability of observing a connected subgraph of size *k* mutated in a number of samples ≥ *samples* under the random model H_0^{sample} or H_0^{gene}. *enrichment p-value* is the *p*-value of the hypergeometric test for overlap between genes in the identified subgraph and genes reported significant pathways in TCGA (2008) or Ding et al. (2008). For GBM, *enrichment p-value* is the *p*-value of the hypergeometric test for RTK/RAS/PI(3)K and p53 pathways.


ERBB3
ERBB2
FGFR1
FGFR2
PDGFRB
PDGFRA
EGFR-family
FGFR
PDGFR
EGFR
MET
RTK
CENTG1
PIK3CA
PIK3R2
PIK3R1
IGF1R
IRS1
CBL
SRC
PI3K-class1a
PTEN
GRB2
GAB1
PIP3
KRAS
NF1
NRAS
RAS
PI3K-class1b
HRAS
PDPK1
RAF
RAF1
ARAF
BRAF
AKT1
AKT3
AKT
FOXO3
FOXO4
FOXO
FOXO1
CDKN2C
CDKN2B
CDKN2A:p16
INK4
CDK6
CDK4
CCND-CDK6
CCND1
CCND-CDK4
CCND2
CCNE-CDK2
RB1
CDKN2A:ARF
MDM2
MDM4
TP53
a
NF1
RAS
RAF
PI3K
b
signal-sending cell
JAG
NOTCH
signal-receiving cell
MAML
c