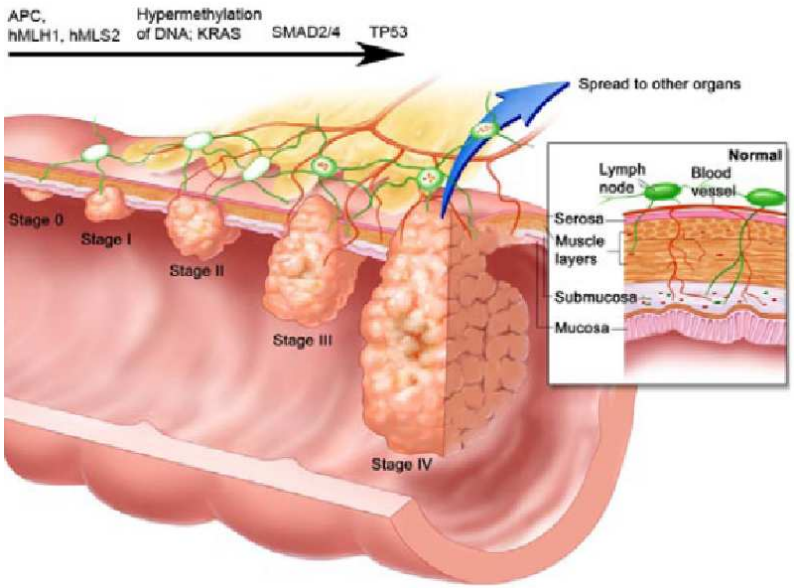
Subnetwork state functions define dysregulated subnetworks in cancer

Salim A. Chowdhury[1], Rod K. Nibbe[2],
Mark R. Chance[2], and **Mehmet Koyutürk**[1,2]

Case Western Reserve University
(1)Electrical Engineering & Computer Science
(2)Center for Proteomics & Bioinformatics

14th Int'l Conf. on Research in Computational Molecular Biology
Lisboa, Portugal; August 12, 2010

Cancer is a complex and progressive disease

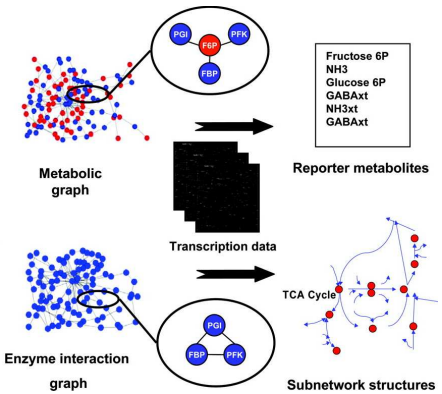


- Complex interactions among multiple genetic and environmental factors.
- Identification of *multiple* markers and their *interactions* $\Rightarrow$ More effective diagnosis, prognosis, modeling, and intervention.

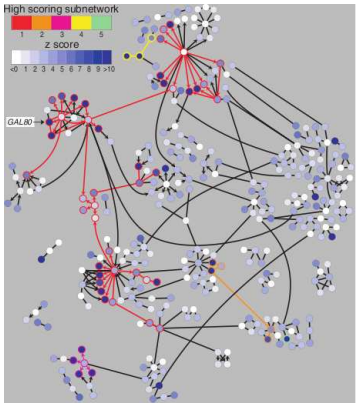
Network-based identification of multiple markers

- Protein-protein interactions (PPIs) highlight functional relationships among proteins.
- Gene expression data hints on transcriptional regulation of proteins in different samples.

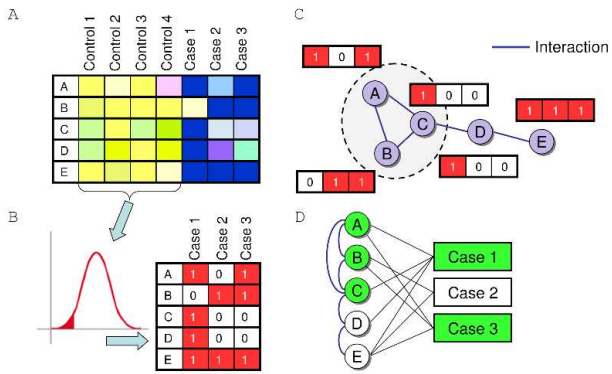
$\Rightarrow$ Identify subnetworks with significant differential expression in pathogenic samples (*dysregulated subnetworks*).



Nielsen & Patil, *PNAS*, 2005



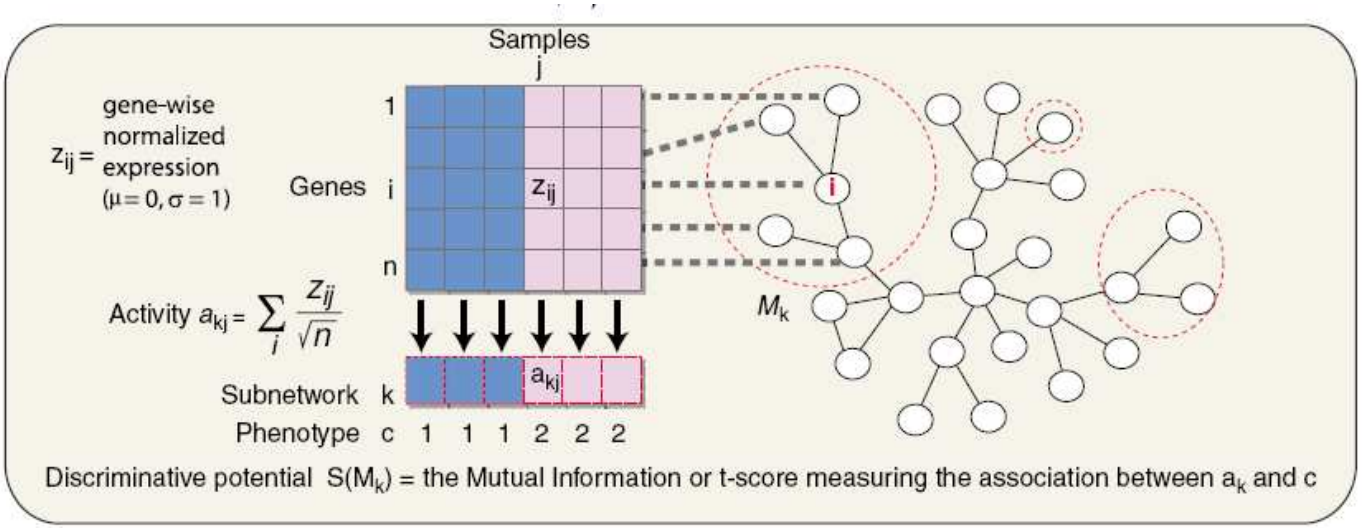
Ideker *et al.*, *ISMB*, 2002



Ulitsky *et al.*, *RECOMB*, 2008

Additive coordinate dysregulation

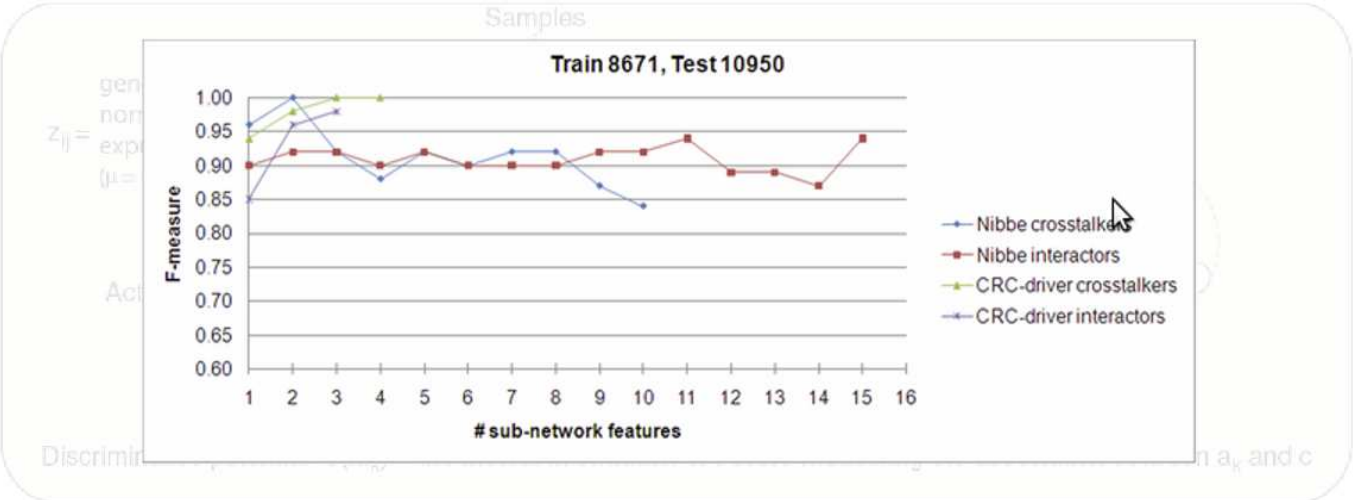
- Subnetwork activity: Aggregate expression of the genes coding for the proteins in the subnetwork.
- Dysregulated subnetworks: Those with differential aggregate expression in pathogenic samples.
  - Captures coordinate dysregulation at a sample-specific resolution.



Chuang *et al.*, *Nature Mol. Sys. Biol.*, 2007

Additive coordinate dysregulation

- Subnetwork activity: Aggregate expression of the genes coding for the proteins in the subnetwork.
- Dysregulated subnetworks: Those with differential aggregate expression in pathogenic samples.
  - Captures coordinate dysregulation at a sample-specific resolution.
  - Enables use of subnetworks as markers for classification.



Nibbe *et al.*, *PLoS Comp. Biol.*, 2010

Finding coordinately dysregulated subnetworks

- Limitations of existing methods:
  - *Additive* formulation coordinate dysregulation.
    - How about interacting proteins that are regulated in different directions?
  - *Greedy* algorithms.
    - But the objective function is combinatorial in nature.

Our approach

- Combinatorial formulation of coordinate dysregulation.
- Exhaustive, but efficient search algorithms.

Formulating coordinate dysregulation

- $\mathcal{S} = \{g_1, g_2, ..., g_m\}$: A subnetwork of the human PPI network.
- $E_i(j)$: Expression of gene g_i in the jth sample.
- $C(j)$: Phenotype of jth sample (*e.g.*, metastatic *vs.* primary).

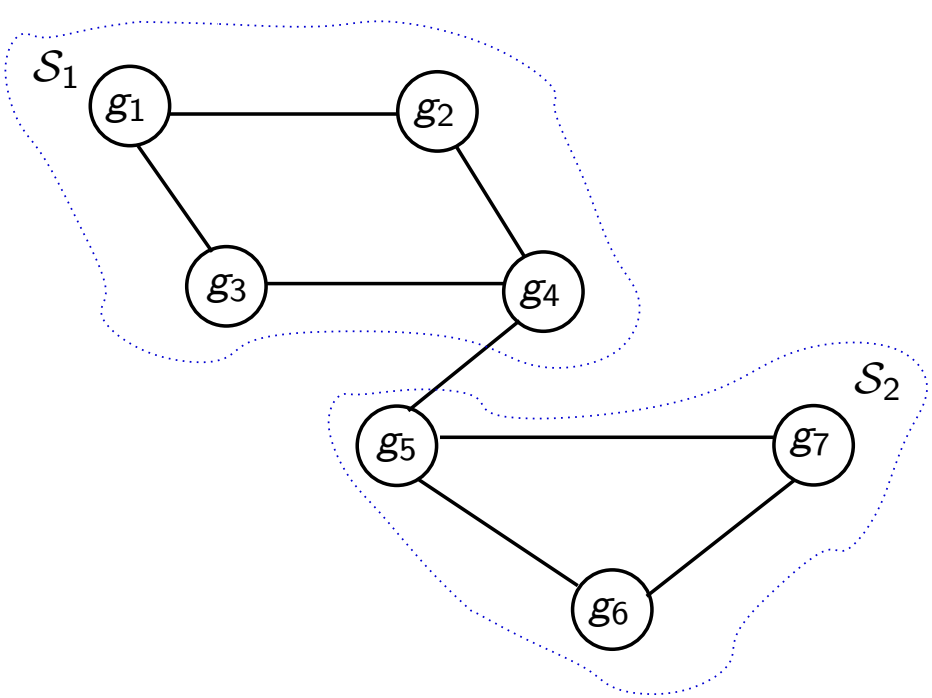
Additive coordinate dysregulation

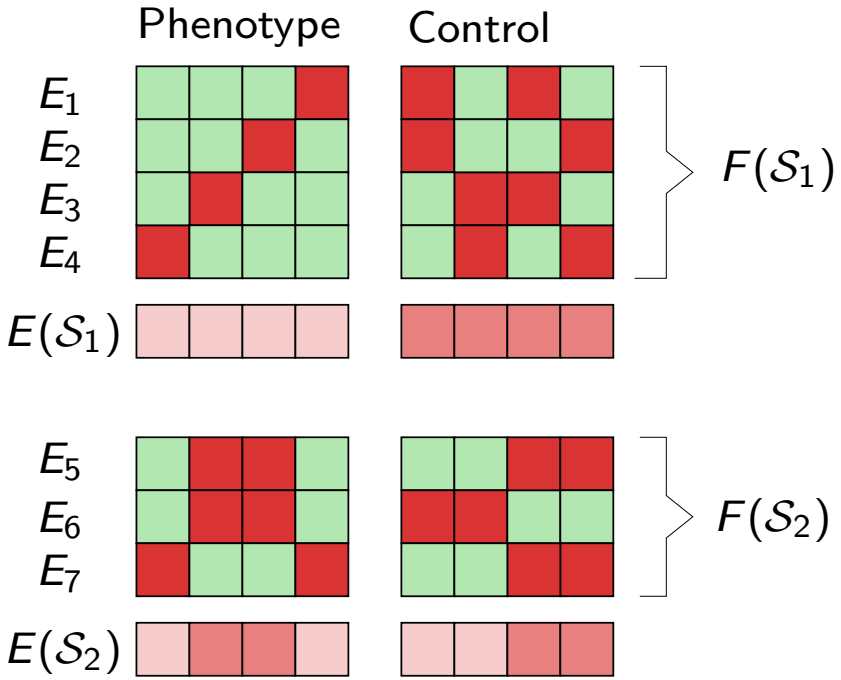
- Subnetwork activity: $E_{\mathcal{S}} = \sum_{i=1}^{m} E_i / \sqrt{m}$
- Additive coordinate dysregulation: $I(E_{\mathcal{S}}; C) = H(C) - H(C|E_{\mathcal{S}})$

Combinatorial coordinate dysregulation

- Subnetwork state: $F_{\mathcal{S}} = \{\hat{E}_1, \hat{E}_2, ..., \hat{E}_m\} \in \{\mathtt{H}, \mathtt{L}\}^m$
- Combinatorial coordinate dysregulation: $I(F_{\mathcal{S}}); C) = H(C) - H(C|F_{\mathcal{S}})$

Combinatorial *vs.* additive coordinate dysregulation

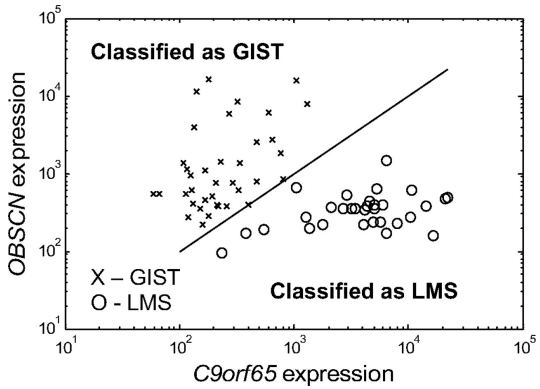




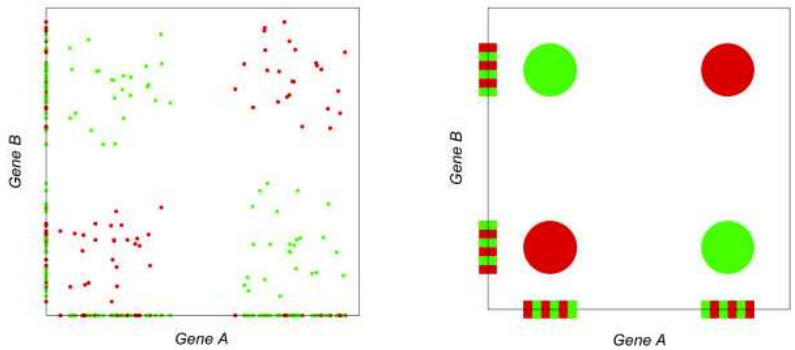
- Additive formulation can capture the dysregulation of $\mathcal{S}_1$, but not that of $\mathcal{S}_2$.
- Combinatorial formulation captures both.

Finding combinatorially dysregulated subnetworks

- Identification of combinatorially dysregulated subnetworks is computationally intractable.
  - *Synergistic dysregulation* is also defined combinatorially, but in a more conservative manner (Anastassiou, *Mol. Sys. Biol.*, 2007).
  - Current applications of synergytic dysregulation are limited to pairs of genes.



Price *et al.*, *PNAS*, 2007



Watkinson *et al.*, *BMC Sys. Biol.*, 2008

State functions

- Decompose the objective function:

$$I(F_{\mathcal{S}}; C) = \sum_{f_{\mathcal{S}} \in \{\mathtt{H},\mathtt{L}\}^m} J(f_{\mathcal{S}}; C)$$

  where

$$J(f_{\mathcal{S}}; C) = p(f_{\mathcal{S}}) \sum_{c \in \{0,1\}} p(c|f_{\mathcal{S}}) \log(p(c|f_{\mathcal{S}})/p(c)).$$

  - $F_{\mathcal{S}}$: Random variable that represents the expression state of subnetwork $\mathcal{S}$.
  - $f_{\mathcal{S}}$: A specific expression state of $\mathcal{S}$ (termed state function).

- High $J(f_{\mathcal{S}}; C) \Rightarrow$ *State function* $f_{\mathcal{S}}$ is *informative* of phenotype.

Algorithmic insight

- $J(.)$ can be bounded for larger state functions using statistics on smaller state functions.
  - Based on a similar result on association rule mining (Smyth & Goodman, *IEEE TKDE*, 1992).

Theorem

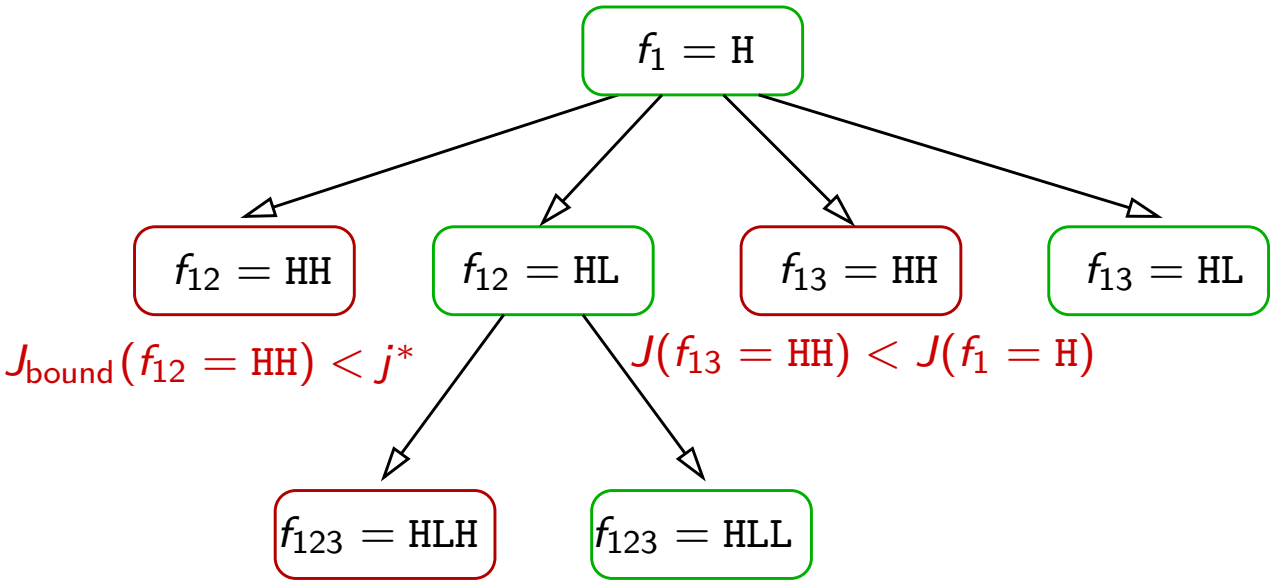
For any superstate $f_{\mathcal{R}}$ of state function $f_{\mathcal{S}}$ (where $\mathcal{S} \subseteq \mathcal{R}$), the following bound holds:

$$J(f_{\mathcal{R}}; C) \leq p(f_{\mathcal{S}}) \max_{c \in \{0,1\}} \left\{ p(c|f_{\mathcal{S}}) \log \frac{1}{p(c)} \right\}.$$

$\Rightarrow$ We can search **exhaustively** for state functions that **indicate** phenotype.

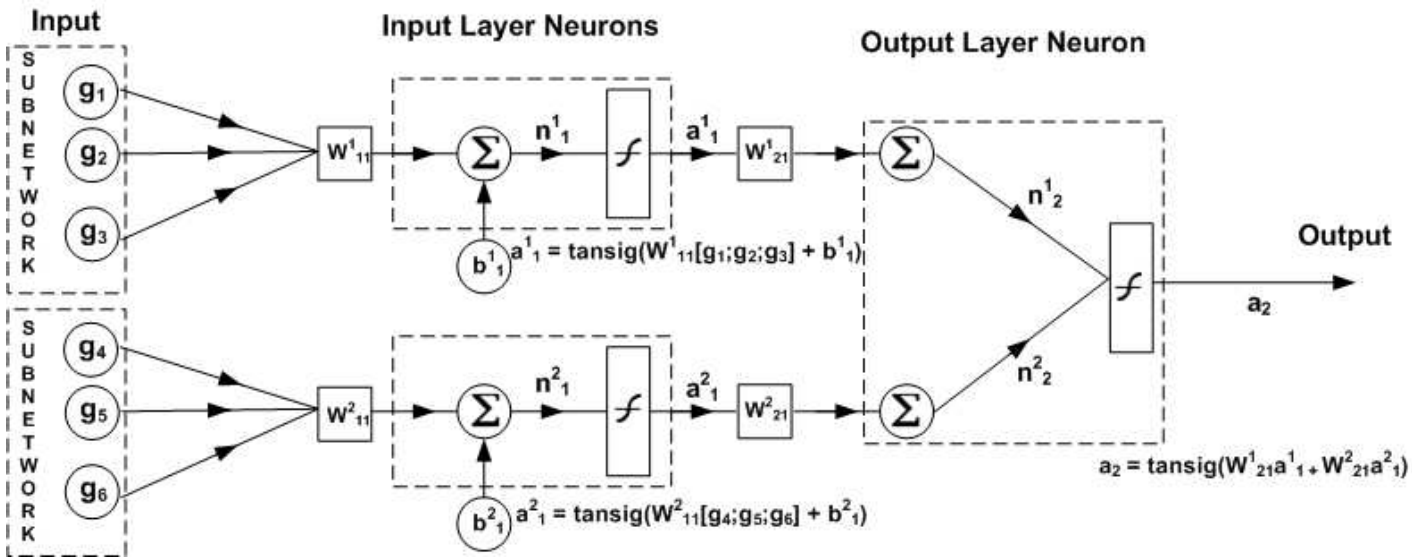
Crane

- Algorithm for the identification of **C**ombinatorially Dys-**R**egul**a**ted Sub-**Ne**tworks.
  - $j*$: Threshold on J-value.
  - b: Breadth of search.
  - d: Depth of search.

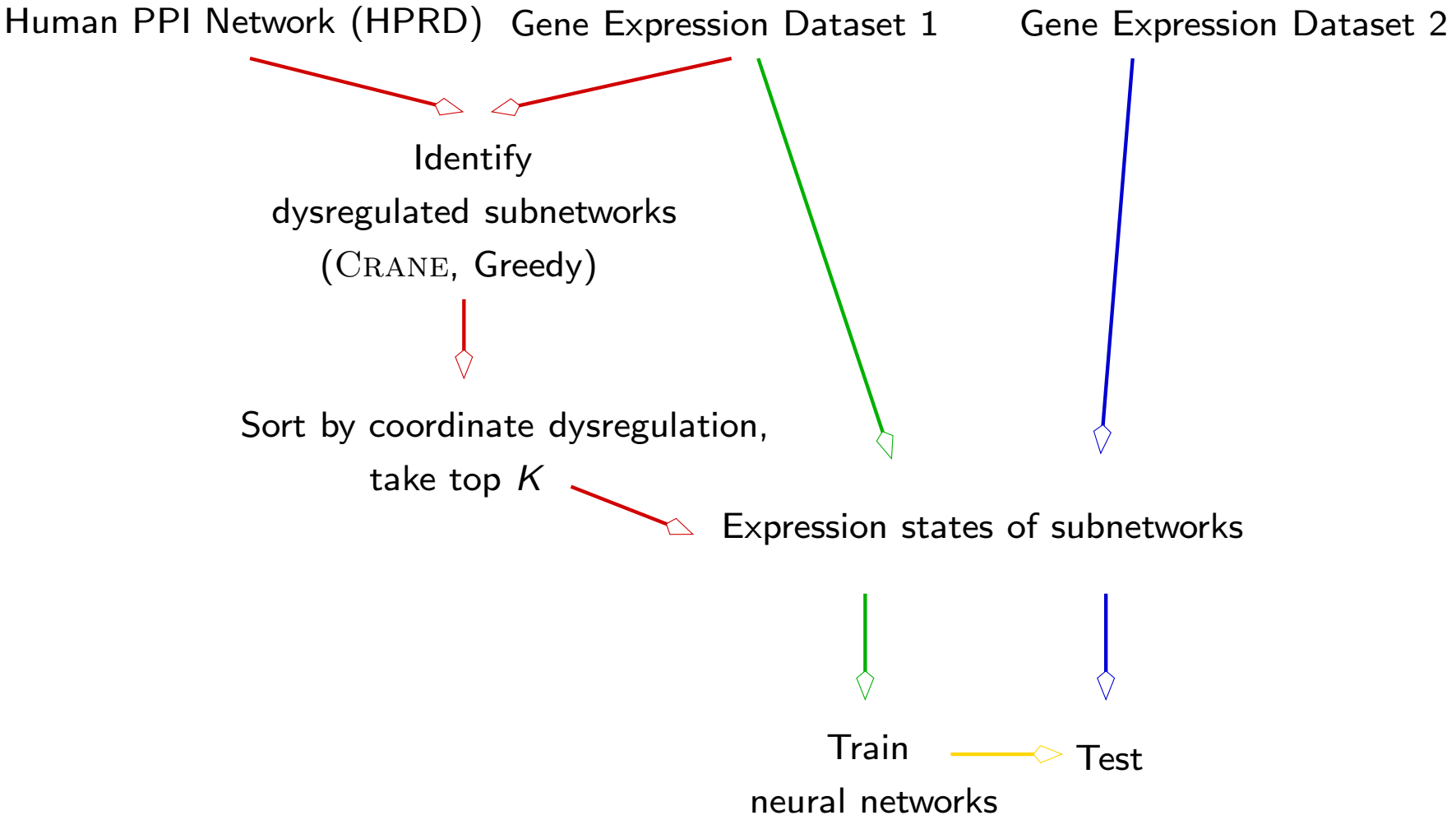


Using informative state functions for classification

- Not straighforward to represent the combinatorial relationship among multiple genes using traditional classifiers (e.g., SVMs).
- We build neural networks in which each subnetwork is represented by an input layer neuron.



Experimental Setup



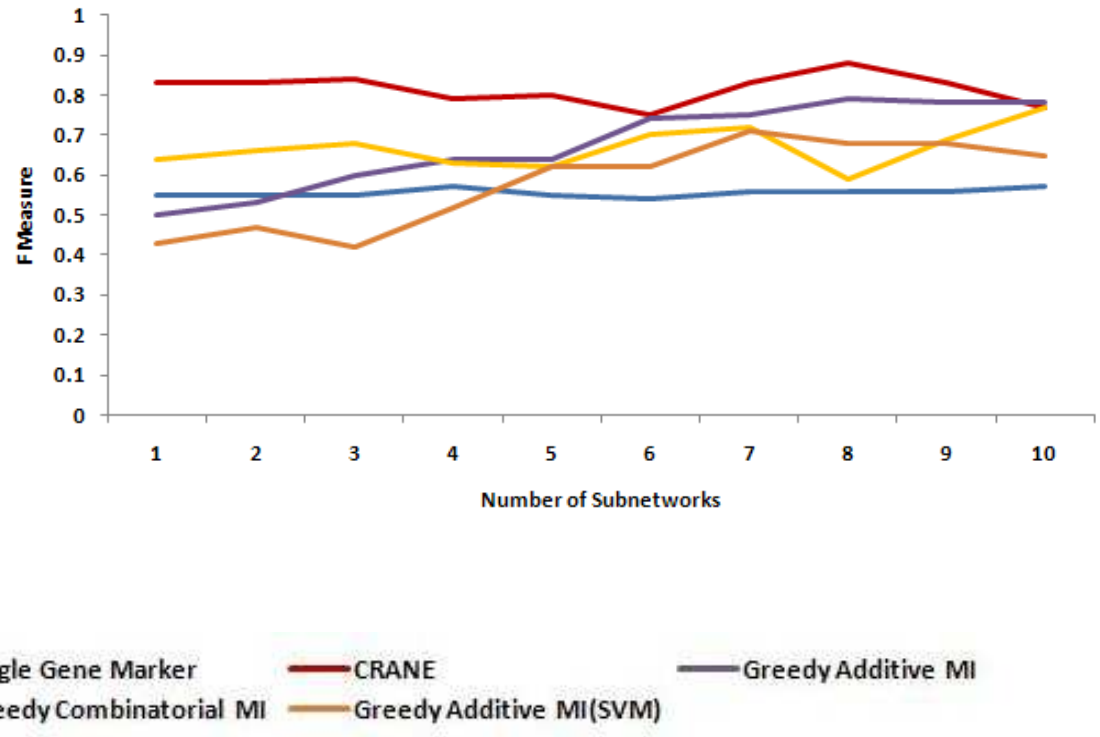
Predicting colon cancer metastasis

- Datasets:
  - GSE6988: 27 *vs.* 20 tumor samples w/ *vs.* w/o liver metastasis (Ki *et al.*, *Int J Cancer*, 2007).
  - GSE3964: 30 *vs.* 18 tumor samples w/ *vs.* w/o liver metastasis (Graudens *et al.*, *Genome Biol*, 2006).

- Algorithms:
  - CRANE.
  - Greedy algorithm with combinatorial dysregulation.
  - Greedy algorithm with additive dysregulation (NN+SVM).
  - Single gene markers (no network information).

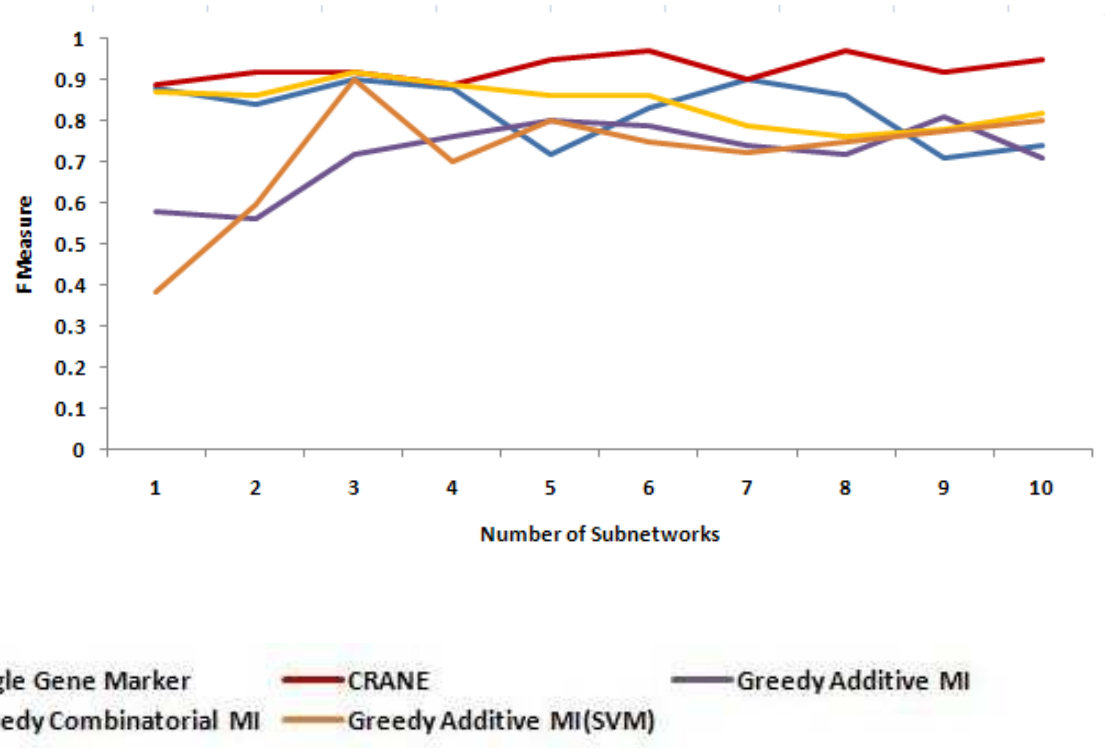
GSE6988 on GSE3964

- Subnetwork discovery & training: GSE6988.
- Testing: GSE3964.



GSE3964 on GSE6988

- Subnetwork discovery & training: GSE3964.
- Testing: GSE6988.

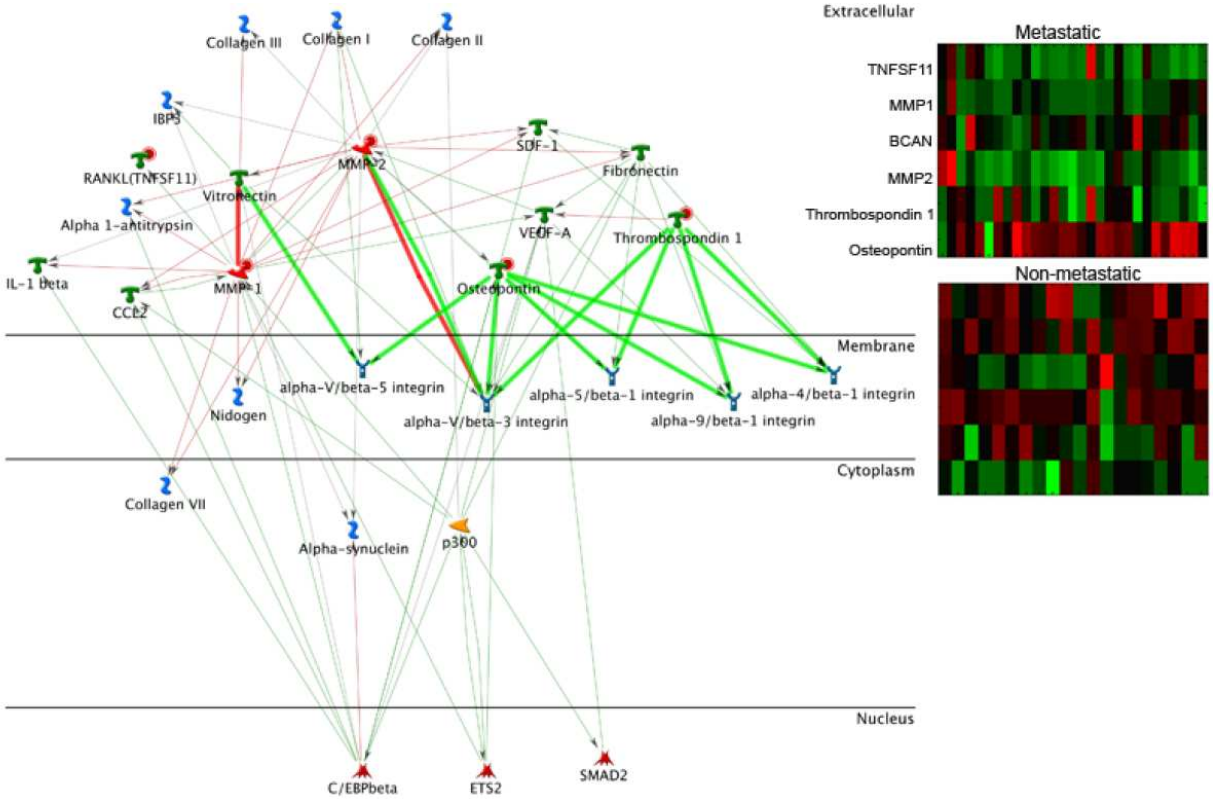


Enrichment analysis

Five subnetworks that are associated with the most informative state functions discovered on GSE6988:

Rank	Proteins	Most Significantly Enriched Process	Enrichment p-value
1	SERPINA3, KLK3, EPOR, GNB2L1, RASA1, RAF1	Inflammation	1×10^{-3}
2	E2F4, CCNE1, GSK3B, HNRPD, SF3B2, RPL13	Cell Movement	1×10^{-3}
3	DMTF1, CCND2. AKAP8, DDX5, FN1, CRP	Cell Migration	1×10^{-4}
4	ANXA11, PLSCR1, EWSR1, PTK2B, ITGB2, HP	Cell Adhesion	1×10^{-4}
5	SKP1A, CCNA2, CDKN1A, GADD45G, EEF1G, RGL2	Inflammation	1×10^{-4}

Generating novel insights



- State function LLLLLH indicates metastasis with $J = 0.33$.
- Overall combinatorial dysregulation: 0.72.
- Overall additive dysregulation: 0.37.

Conclusions

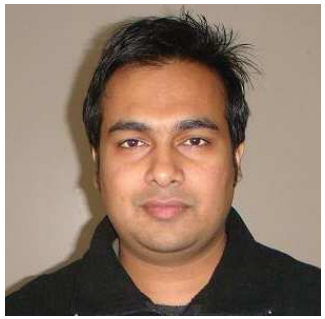
1. Information theoretic formulation of coordinate dysregulation is promising.

2. Consideration of "cellular states" appears to be more effective as compared to "superposition" of information on multiple molecules.

3. Improving upon greedy may improve performance in the search for relevant subnetworks.

4. Combinatorial coordinate dysregulation $\Rightarrow$ novel modeling paradigms for cellular signaling.

Acknowledgments



Salim A. Chowdhury

Rod K. Nibbe

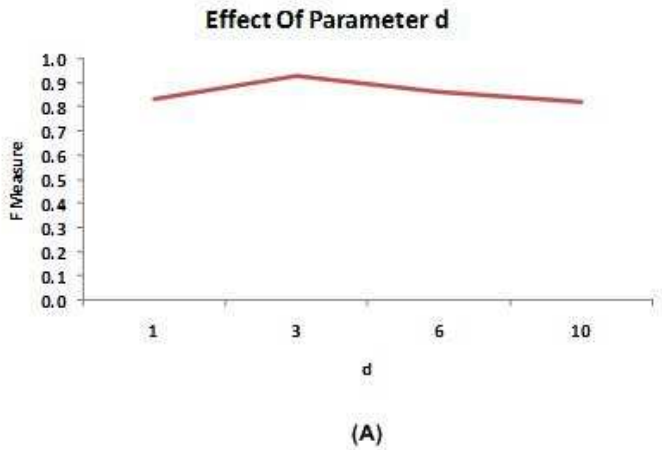
Mark R. Chance

- Sinan Erten (CWRU EECS); Vishal Patel, Gürkan Bebek, Rob Ewing (CWRU Proteomics), Jill Barnholtz-Sloan (Case Comprehensive Cancer Center), Xiaowei Guan (CWRU Biostatistics & Epidemiology).

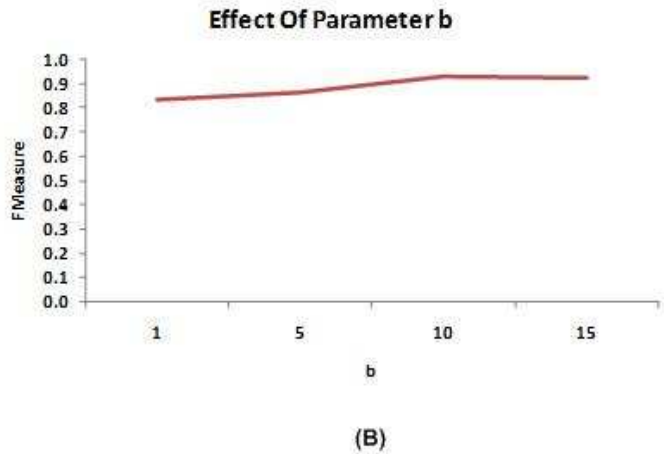
CCF-0953195

UL1-RR024989

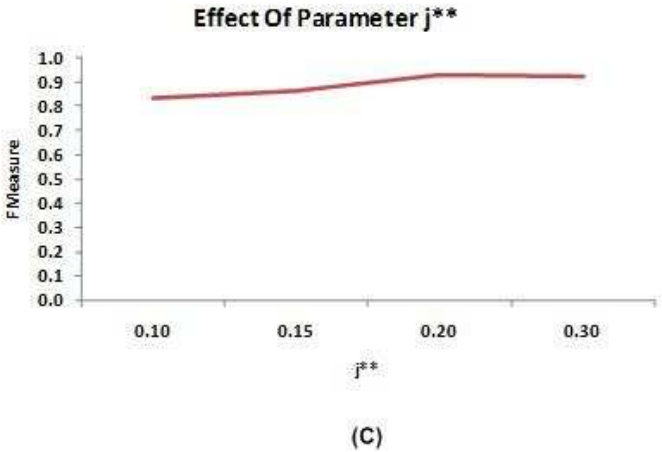
Effect of parameters



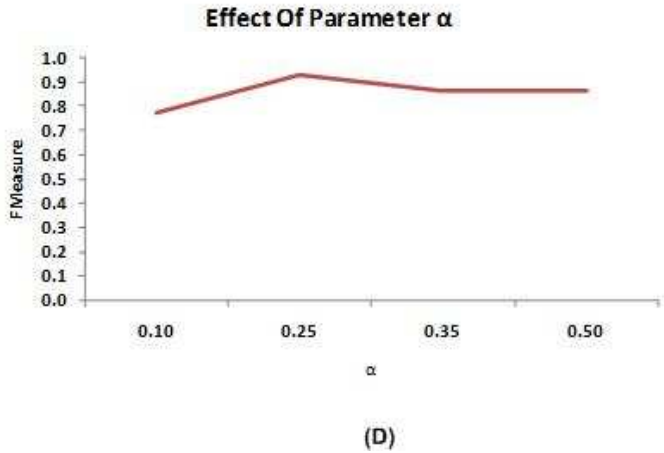
(A)



(B)



(C)



(D)