

Identification of coordinately dysregulated subnetworks in complex phenotypes

Salim A. Chowdhury¹ and Mehmet Koyutürk^{1,2}

Case Western Reserve University

(1)Electrical Engineering & Computer Science

(2)Center for Proteomics & Bioinformatics

Dynamics of Molecular Interaction Networks

Pacific Symposium on Biocomputing

January 8, 2010

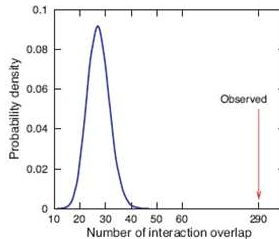
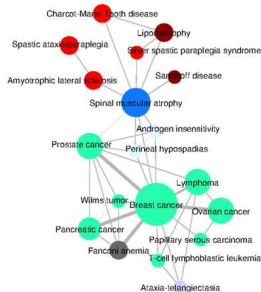
Complex phenotypes

- Many diseases/phenotypes are based on complex interactions between multiple genetic and environmental factors.
 - Heart disease, high blood pressure, Alzheimer's disease, diabetes, cancer, obesity, etc.
- Traditional approaches: single genes.
- Characterization of multiple markers and the dynamics of their interactions is important.



Networks and complex phenotypes

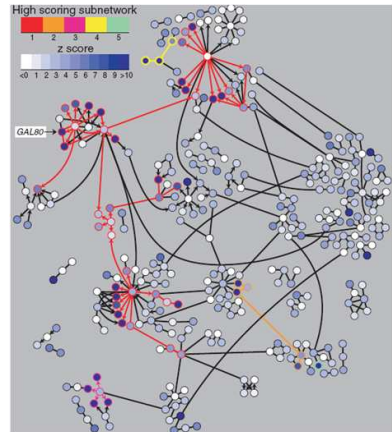
- Protein-protein interactions (PPIs) highlight functional relationships among proteins.
 - Products of genes implicated in similar diseases are highly connected in PPI networks.



Goh *et al.*, *PNAS*, 2007

Dysregulated subnetworks

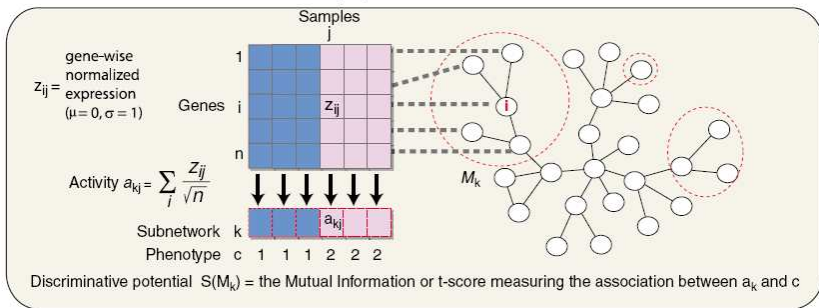
- Network data + Molecular expression data \Rightarrow Network dynamics of phenotype.
- Earlier studies: Connected subgraphs of the PPI network that are rich in differentially expressed genes.
 - Subnetwork score: $\sum_{g_i \in S} z_i / \sqrt{|S|}$.
 - Differential expression is assessed individually for each gene!



Ideker et al., *Bioinformatics*, 2002

Coordinate dysregulation

- By assessing dysregulation collectively for a subnetwork, coordination of genes can be captured at a sample-specific level.













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

- Algorithmic challenge: combinatorial objective function, greedy algorithms!

Cover-based formulation

- **Key idea:** For paired samples, assess the differential expression of each gene for each sample.
 - A gene **positively covers**/ **negatively covers** a sample if it is **up-regulated**/**down-regulated** in the phenotype sample.
 - Differential expression for a single sample can be assessed by properly quantizing gene expression levels.

	s_1	s_2	s_3	s_4	s_5	
Phenotype:						$\mathcal{P} = \{s_4\}$
Control:						$\mathcal{N} = \{s_2, s_5\}$

- **Objective:** Identify subnetworks composed of genes that complement each other in covering all samples.

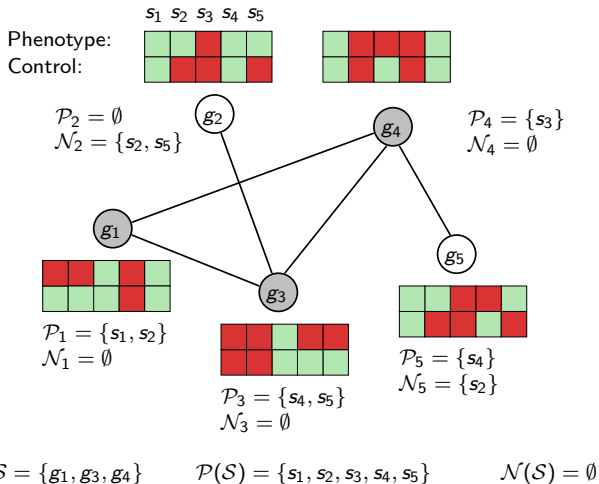
Cover and dysregulation

- How is the cover of a gene related to its dysregulation?
- Information-theoretic formulation of dysregulation.
 - Normalized expression of gene g_i in sample s_j : E_{ij} .
 - Phenotype of sample j : C_j .
 - Dysregulation of gene g_i : $I(E_i; C) = H(C) - H(C|E_i)$.
- Cover of a gene.
 - Binarized expression of gene g_i in sample s_j : \hat{E}_{ij} .
 - Positive cover of gene g_i : $\mathcal{P}_i = \{s_j : \hat{E}_{ij}(Ph) = \uparrow, \hat{E}_{ij}(Co) = \downarrow\}$.

Theorem

For any two genes g_i and g_j , if $||\mathcal{P}_i| - |\mathcal{N}_i|| > ||\mathcal{P}_j| - |\mathcal{N}_j||$, then $I(\hat{E}_i; C) > I(\hat{E}_j; C)$.

Cover of a subnetwork



Cover and coordinate dysregulation

- How is the cover of a subnetwork related to the coordinate dysregulation of the genes in the subnetwork?
- Coordinate dysregulation.
 - Subnetwork activity of \mathcal{S} : $E(\mathcal{S}) = \sum_{g_i \in \mathcal{S}} E_i / \sqrt{|\mathcal{S}|}$.
 - Coordinate dysregulation of \mathcal{S} : $I(E(\mathcal{S}); C) = H(C) - H(C|E(\mathcal{S}))$.
- Cover of a subnetwork.
 - Positive cover of \mathcal{S} : $\mathcal{P}(\mathcal{S}) = \bigcup_{g_i \in \mathcal{S}} \mathcal{P}(g_i)$.
 - Negative cover of \mathcal{S} : $\mathcal{N}(\mathcal{S}) = \bigcup_{g_i \in \mathcal{S}} \mathcal{N}(g_i)$.
- Conjecture: $I(E(\mathcal{S}); C)$ can be maximized by maximizing $|\mathcal{P}(\mathcal{S}) \setminus \mathcal{N}(\mathcal{S})|$.

Problem definition

Minimal covering subnetwork associated with a gene

The minimal covering subnetwork associated with gene g_i is defined as a subnetwork S_i satisfying the following conditions:

1. $g_i \in S_i$.
2. $\forall g_j \in S_i, \exists g_k \in S_i$ such that $\delta(g_j, g_k) \leq \ell$, where δ denotes network distance and ℓ is an adjustable parameter.
3. $\mathcal{P}(S_i) = \mathcal{U}$ or $\mathcal{N}(S_i) = \mathcal{U}$, where \mathcal{U} denotes the set of all samples.
4. If $\mathcal{P}(S_i) = \mathcal{U}$ ($\mathcal{N}(S_i) = \mathcal{U}$), then $|\mathcal{N}(S_i)|$ ($|\mathcal{P}(S_i)|$) is minimum over all subnetworks that satisfy the above three conditions.
5. $\forall g_j \in S_i$, subnetwork $S_i \setminus \{g_j\}$ does not satisfy the above conditions.

NETCOVER

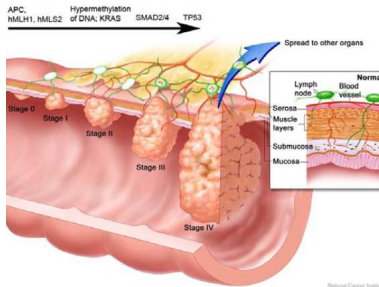
- Identifies a minimal covering subnetwork associated with each gene in the network.
 - Implements an adaptation of Chvátal's (*Math Op Res*, 1979) algorithm for the set-cover problem.

Algorithm NETCOVER

1. Initialize $S_i \leftarrow \{g_i\}$, $\mathcal{T} \leftarrow \mathcal{U} \setminus \mathcal{P}_i$, $\mathcal{Q} \leftarrow \{g_j \in \mathcal{V} : \delta(g_i, g_j) \leq \ell\}$.
2. For all $g_j \in \mathcal{Q}$, compute $\mathcal{P}'_j \leftarrow \mathcal{P}_j \cap \mathcal{T}$
3. Find the genes in \mathcal{Q} with maximum $|\mathcal{P}'_j|$ and let g_k be the gene among these genes with minimum $|\mathcal{N}_j|$.
4. $S_i \leftarrow S_i \cup \{g_k\}$.
5. $\mathcal{T} \leftarrow \mathcal{T} \setminus \mathcal{P}'_k$.
6. $\mathcal{Q} \leftarrow \mathcal{Q} \cup \{g_j \in \mathcal{V} : \delta(g_k, g_j) \leq \ell\} \setminus \{g_k\}$.
7. If $\mathcal{T} = \emptyset$ or $\mathcal{Q} = \emptyset$, return S_i ; otherwise, go to step (2).

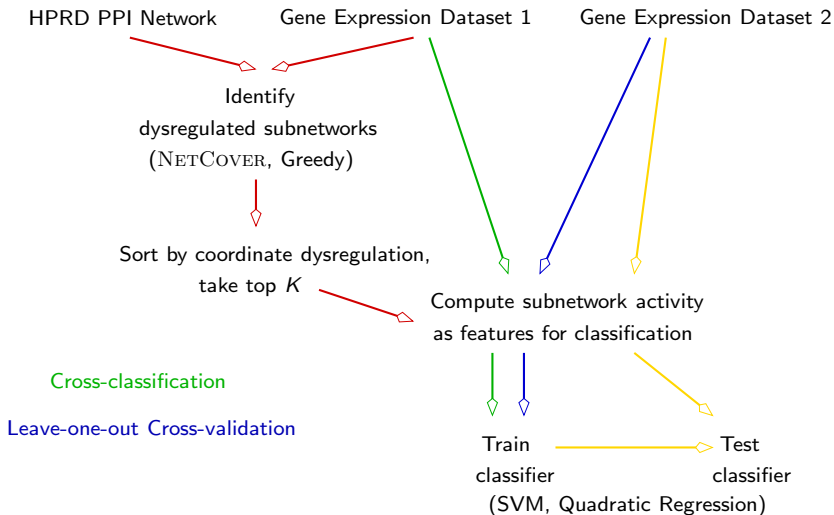
Human colorectal cancer (CRC)

- Second leading cause of cancer deaths in the United States (American Cancer Society, 2009).
- CRC is a complex, progressive disease.
 - Identification of multiple markers is important for effective **diagnosis**, **prognosis**, modeling, and intervention.



National Cancer Institute

Classification framework



Experimental Setup

■ Classification tasks

- Diagnosis: Discriminating tumor samples from normal.
- Prognosis: Discriminating metastatic samples from primary tumor.

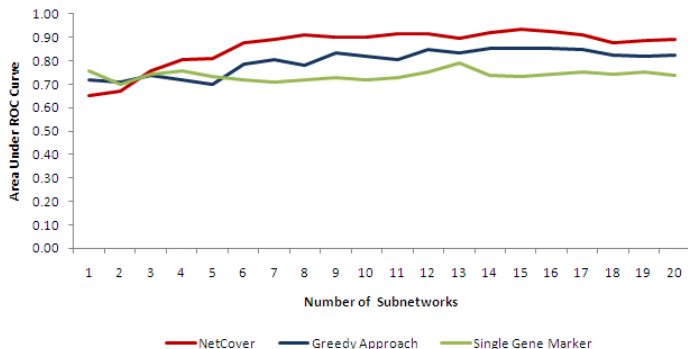
■ Datasets

- GSE8671: 32 adenoma samples paired with normal mucosa.
- GSE10950: 24 normal and tumor pairs.
- GSE6988: 27 liver metastasis, 20 primary colorectal tumors, 25 normal mucosa.

■ Algorithms

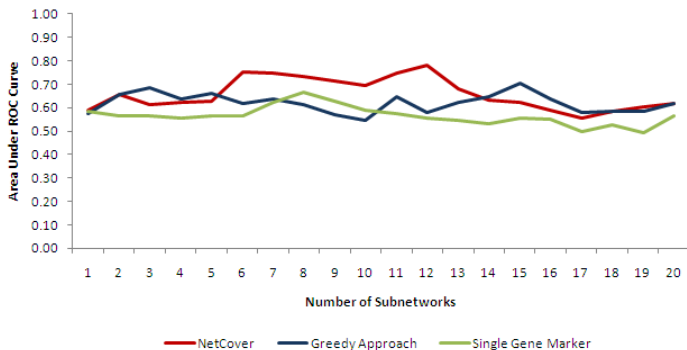
- NETCOVER.
- Greedy algorithm with coordinate dysregulation as the objective function.
- Single gene markers (no network information).

Predicting tumor



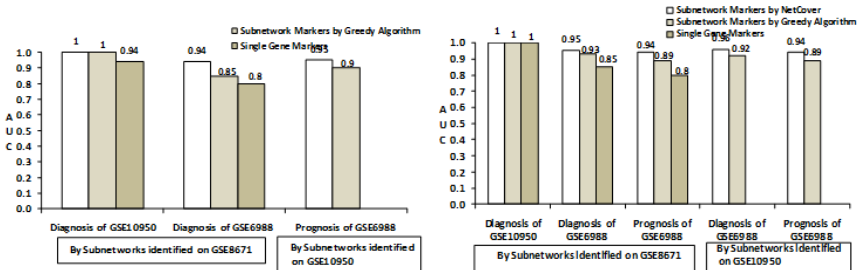
- Subnetwork identification & training: GSE8671.
- Testing: GSE6988.
- Classifier: SVM, Cross-classification.

Predicting metastasis



- Subnetwork identification & training: GSE8671.
- Testing: GSE6988.
- Classifier: Quadratic regression, Leave-one-out Cross-validation.

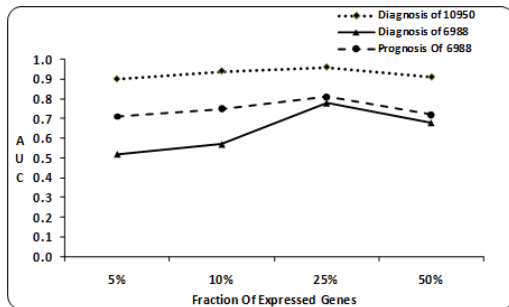
Overall performance



- Classifier: SVM.
- Best performance achieved by each algorithm is reported.

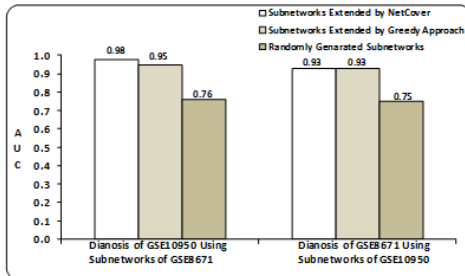
Effect of binarization

- Expression levels are normalized gene-wise ($\mu = 0$, $\sigma = 1$).
- Top α -fraction of expression levels are set to \uparrow , the rest is set to \downarrow .



Integration with genomic targets

- Comparative genomic studies reveal many genes with mutations associated with CRC (Sjöblom *et al.*, *Science*, 2006).
 - Are subnetworks associated with genomic markers more likely to be coordinately dysregulated?



- Genomic targets can be used to seed the search for dysregulated subnetworks!

Conclusions and future work

- Cover-based formulation generates subnetworks with high predictive power and reproducibility.
 - Mechanistic insights: Relevance to within- and between-pathway models?
 - Therapeutic intervention: Effect of interference with the expression of multiple genes?
 - Algorithmic improvements: Application to unpaired samples?
 - Systems-wide perspective: Integration with genomic and proteomic data?

Acknowledgments



Salim A. Chowdhury



Rod K. Nibbe



Mark R. Chance

- Anonymous reviewers and session organizers for their valuable feedback.
- NSF CAREER Award CCF-0953195.