Integrative 'omics approaches in systems biology of complex phenotypes

Mehmet Koyutürk*

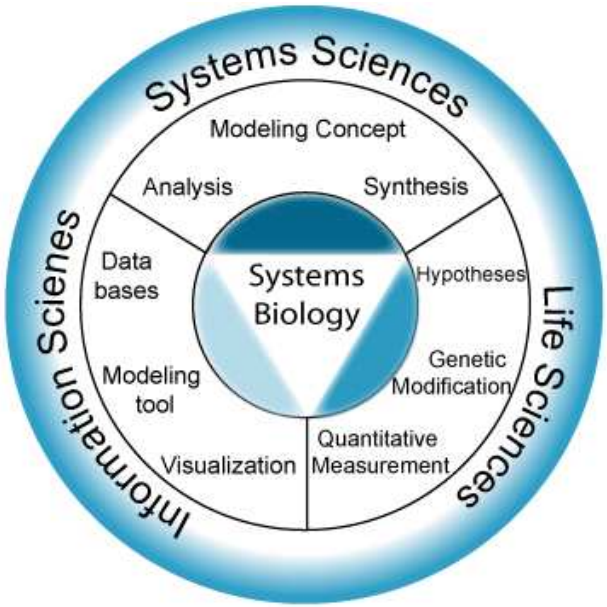
Case Western Reserve University
(1)Electrical Engineering & Computer Science
(2)Center for Proteomics & Bioinformatics

Department of Computer Engineering, Bahçeşehir University
April 7, 2010

*Joint work with Sinan Erten, Rod K. Nibbe, Salim A. Chowdhury, and Mark R. Chance.

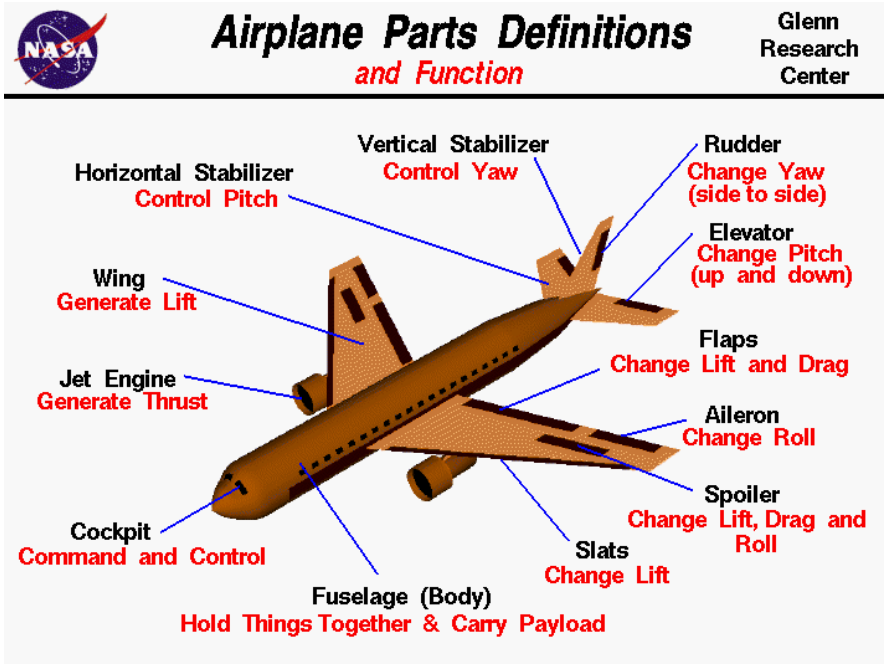
Systems biology

- Life is an emergent property.
  - "To understand biology at the system level, we must examine the structure and dynamics of cellular and organismal function, rather than the characteristics of isolated parts of a cell or organism." (Kitano, *Science*, 2002).



- Systems biology complements molecular biology.

Organization and dynamics of complex systems



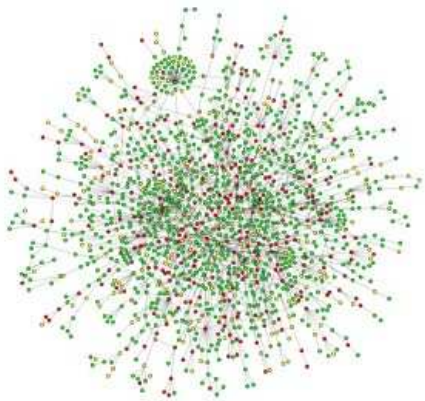
- Understanding how an airplane (cell) works:
  - Listing parts (genes, proteins).
  - Understanding how parts are connected (interactions).
  - Characterizing the electrical and mechanical dynamics (cellular dynamics).

Complex diseases

- Many diseases are based on a set of complex interactions between multiple genetic and environmental factors.
  - Heart disease, high blood pressure, Alzheimers disease, diabetes, cancer and obesity, etc.
- Characterization of multiple markers and their interactions is important for effective diagnosis, prognosis, modeling, and intervention.

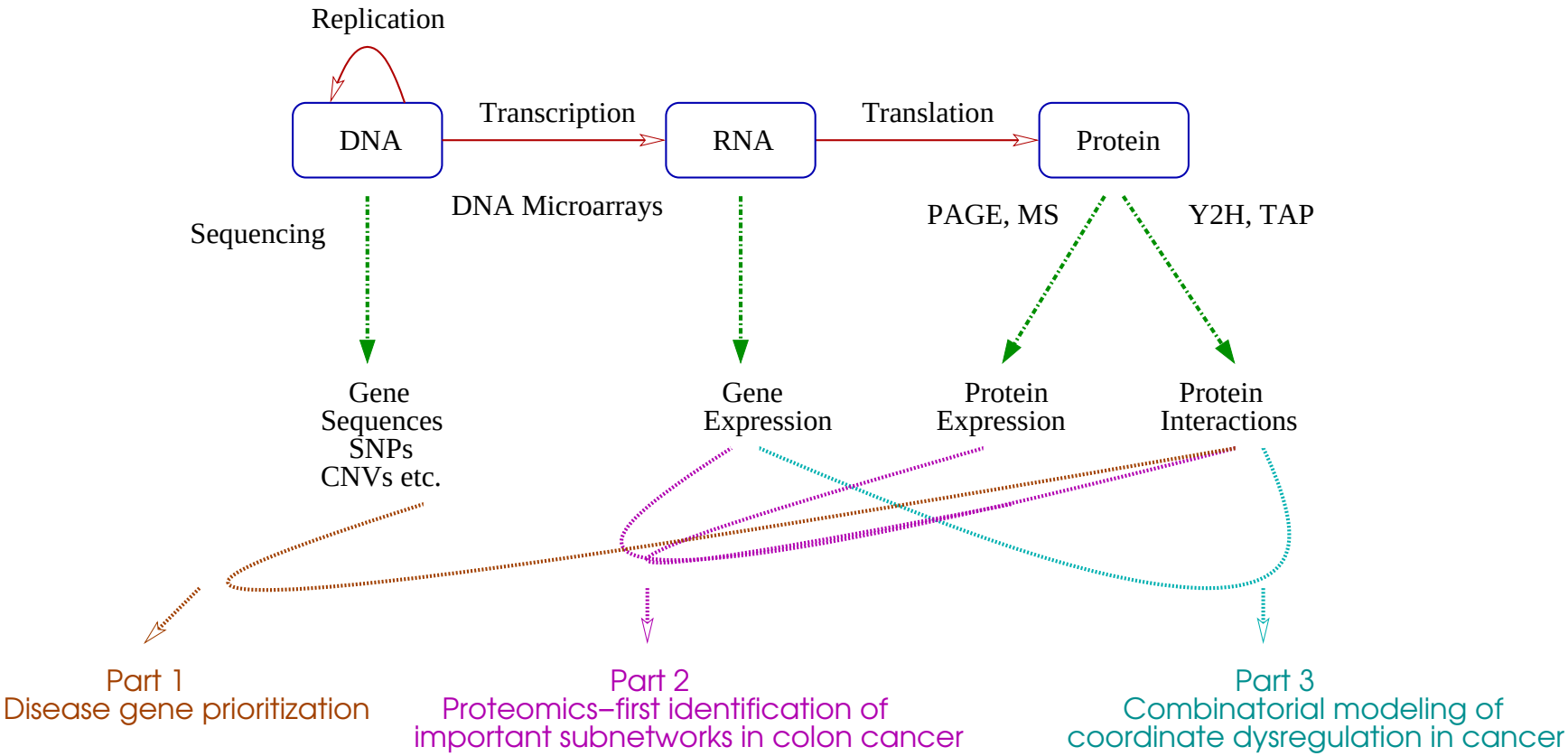
Protein-protein interaction (PPI) networks

- Physically interacting proteins can be identified via high-throughput screening.
- Nodes represent proteins.
- Edges represent interactions.
  - Binding, regulation, modification, transport, complex membership...



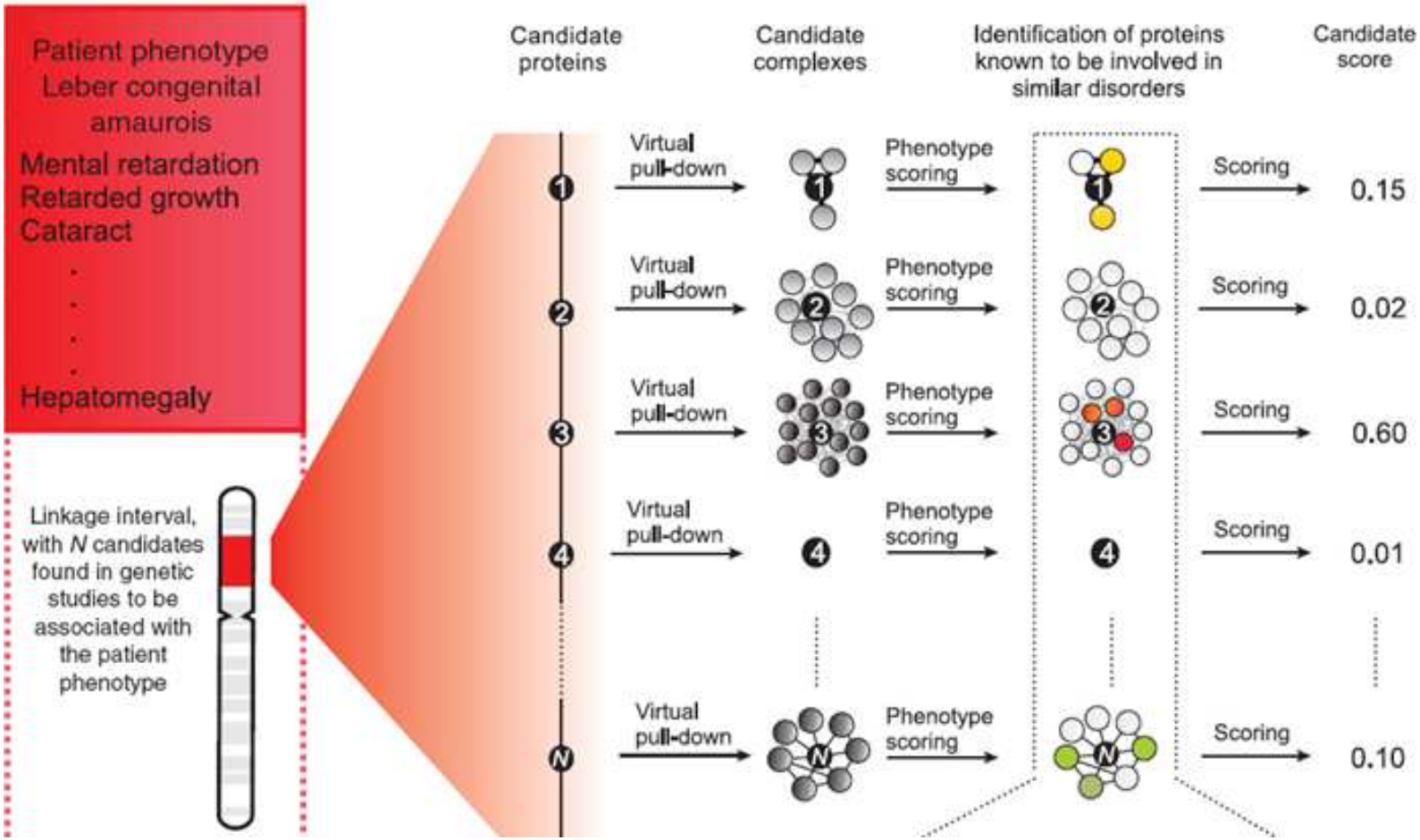
S.cerevisiae (Baker's yeast)
Protein Interaction (PPI) Network

Outline


Replication
DNA
Transcription
RNA
Translation
Protein
Sequencing
DNA Microarrays
PAGE, MS
Y2H, TAP
Gene
Sequences
SNPs
CNVs etc.
Gene
Expression
Protein
Expression
Protein
Interactions
Part 1
Disease gene prioritization
Part 2
Proteomics–first identification of
important subnetworks in colon cancer
Part 3
Combinatorial modeling of
coordinate dysregulation in cancer

Network-based Prioritization of Candidate Disease Genes

PPI networks in disease gene prioritization



Lage *et al.*, *Nature Biotechnology*, 2007.

The problem

- Input:
  - $\mathcal{Q}$: Set of known disease genes (*seeds*).
  - $\sigma(s)$ for $s \in \mathcal{Q}$: Degree of association between s and the disease of interest.
  - $\mathcal{C}$: Set of candidate genes in the disease.
  - $(\mathcal{V}, \mathcal{E})$: Network of PPIs among human proteins (edges can be weighted representing reliability of interactions).
- Output:
  - Ranking of candidate genes in $\mathcal{C}$ based on their likelihood of association with disease.

Driving hypothesis

Products of genes implicated in similar diseases are likely to interact with each other.

Random walk with restarts

- Quantifies the crosstalk between products of known disease genes $\mathcal{Q}$ (seed set) and candidate genes $\mathcal{C}$ (Köhler *et al.*, *Am. J. Hum. Gen.*, 2008; Chen *et al.*, *BMC Bioinf.*, 2009).
  - Accounts for multiplicity of paths and indirect interactions!

- Simulates a random walk on human PPI network, making frequent restarts at known disease genes.

$$\phi_0 = r,\ \phi_{t+1} = (1-c)P\phi_t + cr,\ \phi = \lim_{t\to\infty} \phi_t$$

  - r : Restart vector; $r(s) = \sigma(s)/\sum_{s\in\mathcal{Q}} \sigma(s)$ for $s \in \mathcal{Q}$, 0 otherwise.
  - c: Restart probability (tunable parameter).
  - P: Stochastic network derived from (weighted) adjacency matrix of the PPI network.

Network propagation

- In random walk with restarts, P is the stochastic matrix derived from the adjacency matrix of the network.
  - Only outgoing flow is normalized.

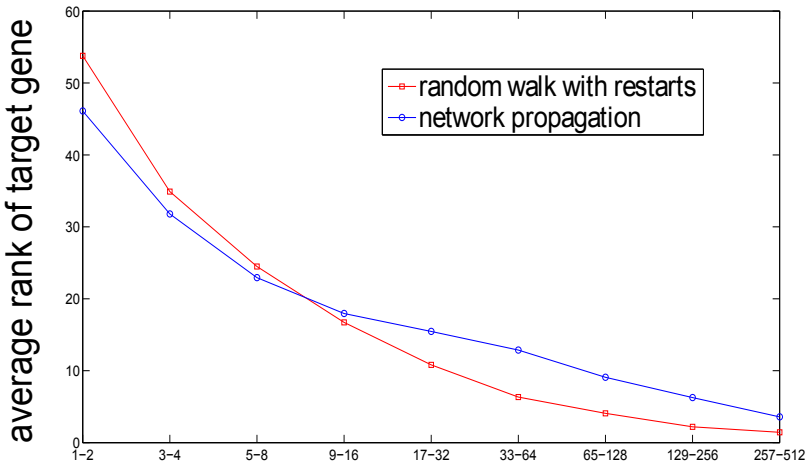
$$P_{\mathrm{RW}}(u, v) = 1/|\mathcal{N}(v)| \text{ for } uv \in E,\ 0 \text{ otherwise.}$$

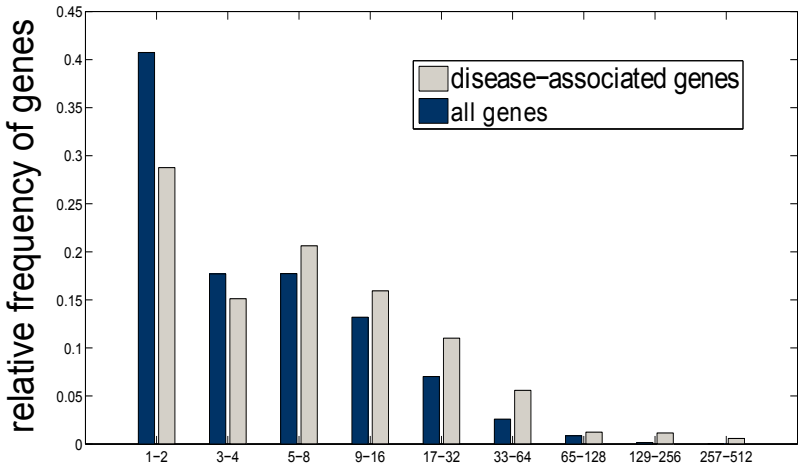
- On the contrary, network propagation models the "disease association information" being pumped from the seed set and propagated across the network (Vanunu *et al.*, *PLoS Comp. Biol.*, 2010).
  - Both incoming and outgoing flows are normalized.

$$P_{\mathrm{NP}}(u, v) = 1/\sqrt{|\mathcal{N}(u)||\mathcal{N}(v)|} \text{ for } uv \in E,\ 0 \text{ otherwise.}$$

$\mathcal{N}(v)$: Set of interacting partners of protein $v \in \mathcal{V}$.

Performance depends on network degree





- Leave-one-out cross-classification experiments using OMIM database demonstrate success of information flow based methods.
- But stratification according to degree clearly shows that these methods are significantly biased by network centrality.

Assessing significance with respect to centrality

- Can we statistically adjust information flow based association scores using reference models that accurately represent the degree distribution of the network?

- Three statistical adjustment schemes:

  - Reference model based on **seed degree**.

  - Reference model based on **candidate degree**.

  - Likelihood-ratio test with respect to **eigenvector centrality**.

Reference model based on seed degree

- Generate random seed sets that represent the degree distribution of original seed set.
  - $\mathcal{S}^{(1)}$, $\mathcal{S}^{(2)}$, ..., $\mathcal{S}^{(n)}$ with sufficiently large n.
- Compute scores $\phi^{(1)}$, $\phi^{(2)}$, ..., $\phi^{(n)}$ w.r.t. random seed sets, estimate population mean and standard deviation.
  - $\mu_{\mathcal{S}} = \sum_{1 \leq i \leq n} \alpha^{(i)} / n$.
  - $\sigma^2_{\mathcal{S}} = \sum_{1 \leq i \leq n} ((\alpha^{(i)} - \mu_{\mathcal{S}})(\alpha^{(i)} - \mu_{\mathcal{S}})^T)/(n-1)$.
- Adjust scores based on these sample statistics:

$$\phi_{\mathrm{SD}}(v) = (\phi(v) - \mu_{\mathcal{S}}(v))/\sigma_{\mathcal{S}}(v).$$

Reference model based on candidate degree

- For each candidate $v \in \mathcal{C}$, generate population $\mathcal{M}(v)$ that contains proteins with degree similar to v.

- Estimate population mean and standard deviation for this degree regime.
  - $\mu(v) = \sum_{u \in \mathcal{M}(v)} \alpha(u)/|\mathcal{M}(v)|$.
  - $\sigma^2(v) = \sum_{u \in \mathcal{M}(v)} (\alpha_{\mathcal{S}}(u) - \mu(v))/(|\mathcal{M}(v)| - 1)$.

- Adjust scores based on these sample statistics:

$$\phi_{\mathrm{CD}}(v) = (\phi(v) - \mu(v))/\sigma(v).$$

Likelihod w.r.t. eigenvector centrality

- The random walk score for $r = 0$ is a measure of network centrality (equivalent to Google page-rank).

- Perform likelihood-ratio test using this score as background:

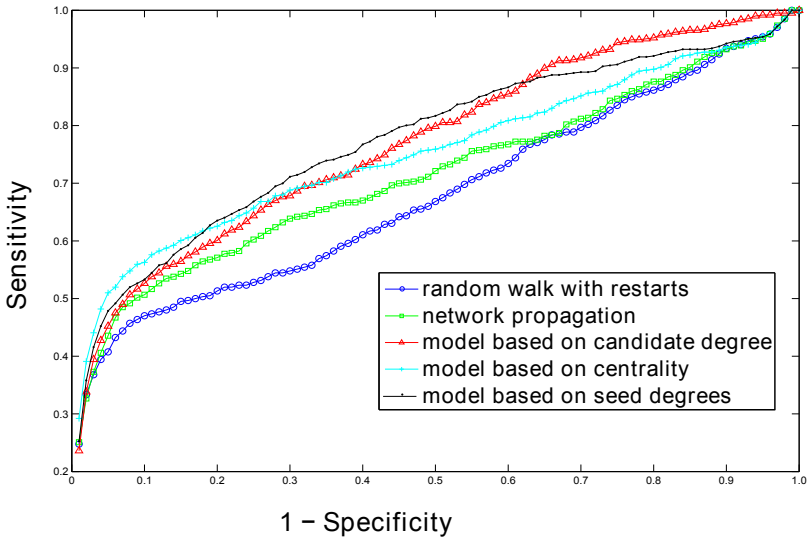
$$\phi_{\mathrm{EC}}(v) = \log \frac{\phi^{(r>0)}(v)}{\phi^{(r=0)}(v)}.$$

Experimental setup

- Human PPI network: NCBI Entrez Gene database.
  - 3528 binary interactions between 8959 proteins.
- Disease-gene associations: Online Mendelian Inheritance in Man (OMIM) database.
  - 206 diseases with at least 3 known associated genes.
  - Number of associations per disease ranges from 3 to 36, mean ≈ 6.
- Leave-one-out cross validation. For each disease:
  - Remove a gene from the seed set (target gene).
  - Generate an artificial linkage interval from its 99 chromosomal neighbors.
  - Rank candidates in this interval, see how target gene is ranked.

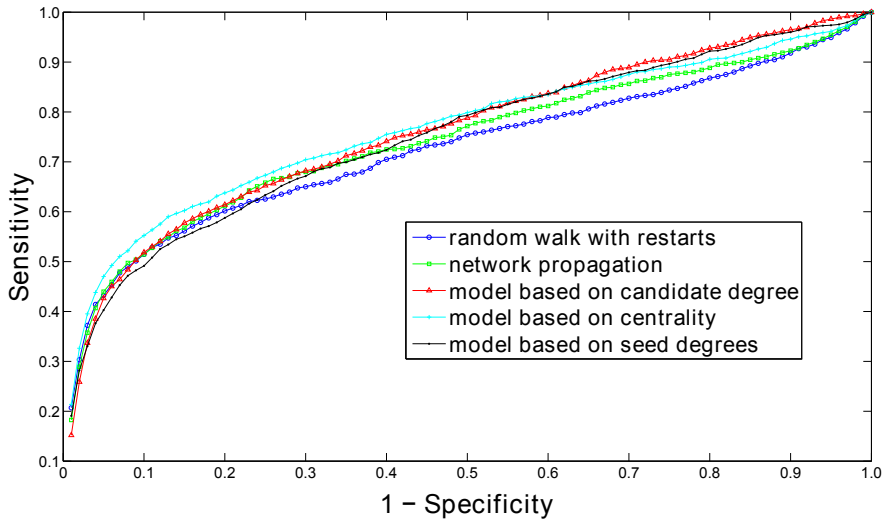
Effect of statistical adjustment

Degree ≤ 5:



(a)

All genes:



(b)

- Statistical adjustment greatly improves performance for loosely connected genes.
- However, the overall improvement is marginal.

Uniform prioritization

- Can we combine raw and statistically adjusted scores to compute a unique rank for each gene?
  - Based on candidate degree (local):

$$R^{(\mathrm{C})}_{\mathrm{UNI}}(v) = \begin{cases} R_{\mathrm{RAW}}(v) & \text{if } |\mathcal{N}(v)| > \lambda \\ R_{\mathrm{ADJ}}(v) & \text{otherwise} \end{cases}$$

  - Optimistic prioritization (local):

$$R^{(\mathrm{O})}_{\mathrm{UNI}}(v) = \begin{cases} R_{\mathrm{RAW}}(v) & \text{if } R_{\mathrm{RAW}}(v) < R_{\mathrm{ADJ}}(v) \\ R_{\mathrm{ADJ}}(v) & \text{otherwise} \end{cases}$$

  - Based on seed degree (global):

$$\overline{d}(\mathcal{S}) = (\sum_{u \in \mathcal{S}} |\mathcal{N}(u)|)/|\mathcal{S}|.$$

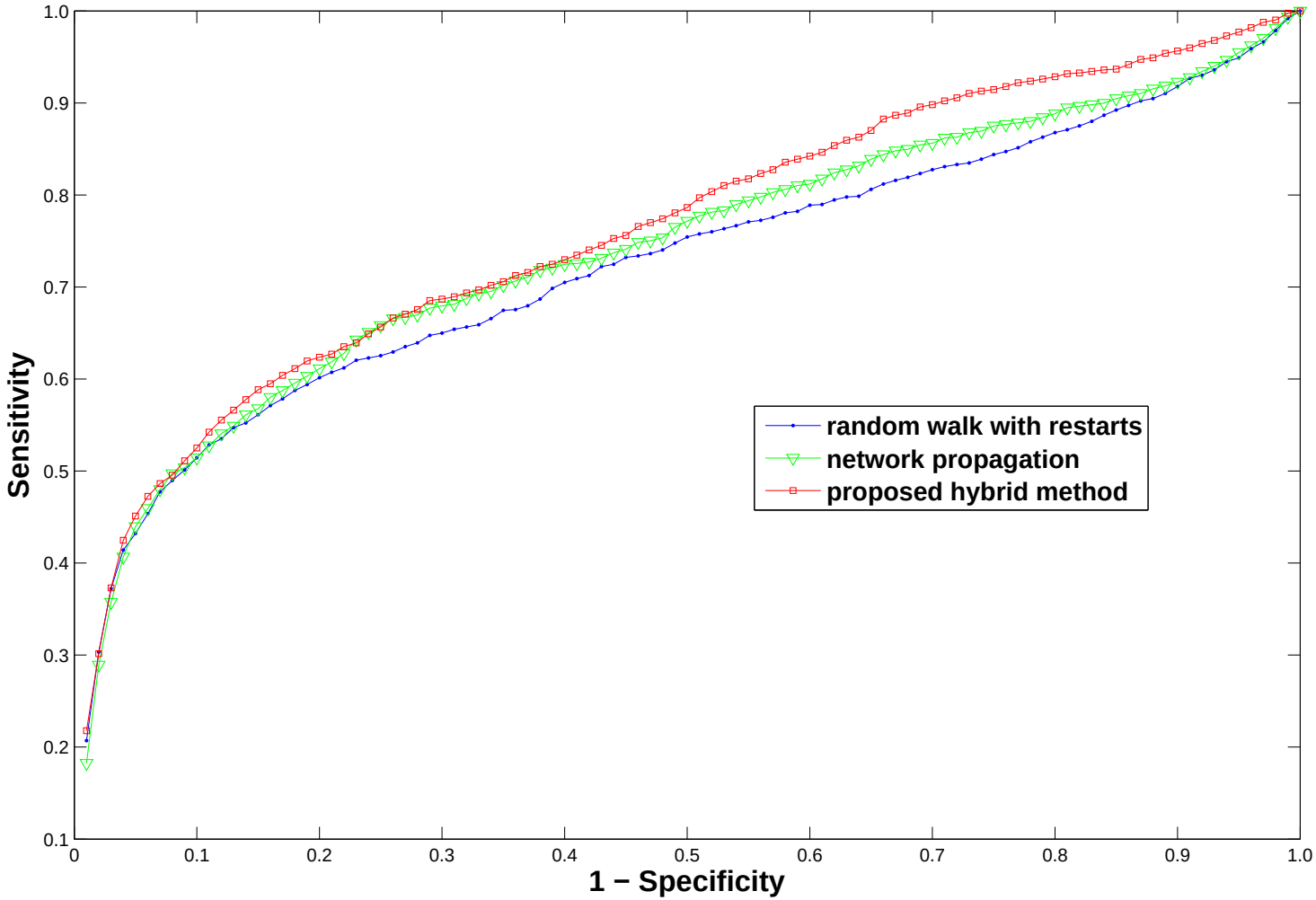
$$R^{(\mathrm{S})}_{\mathrm{UNI}}(v) = \begin{cases} R_{\mathrm{RAW}}(v) & \text{if } \overline{d}(\mathcal{S}) > \lambda \\ R_{\mathrm{ADJ}}(v) & \text{otherwise} \end{cases}$$

Performance of uniform prioritization schemes

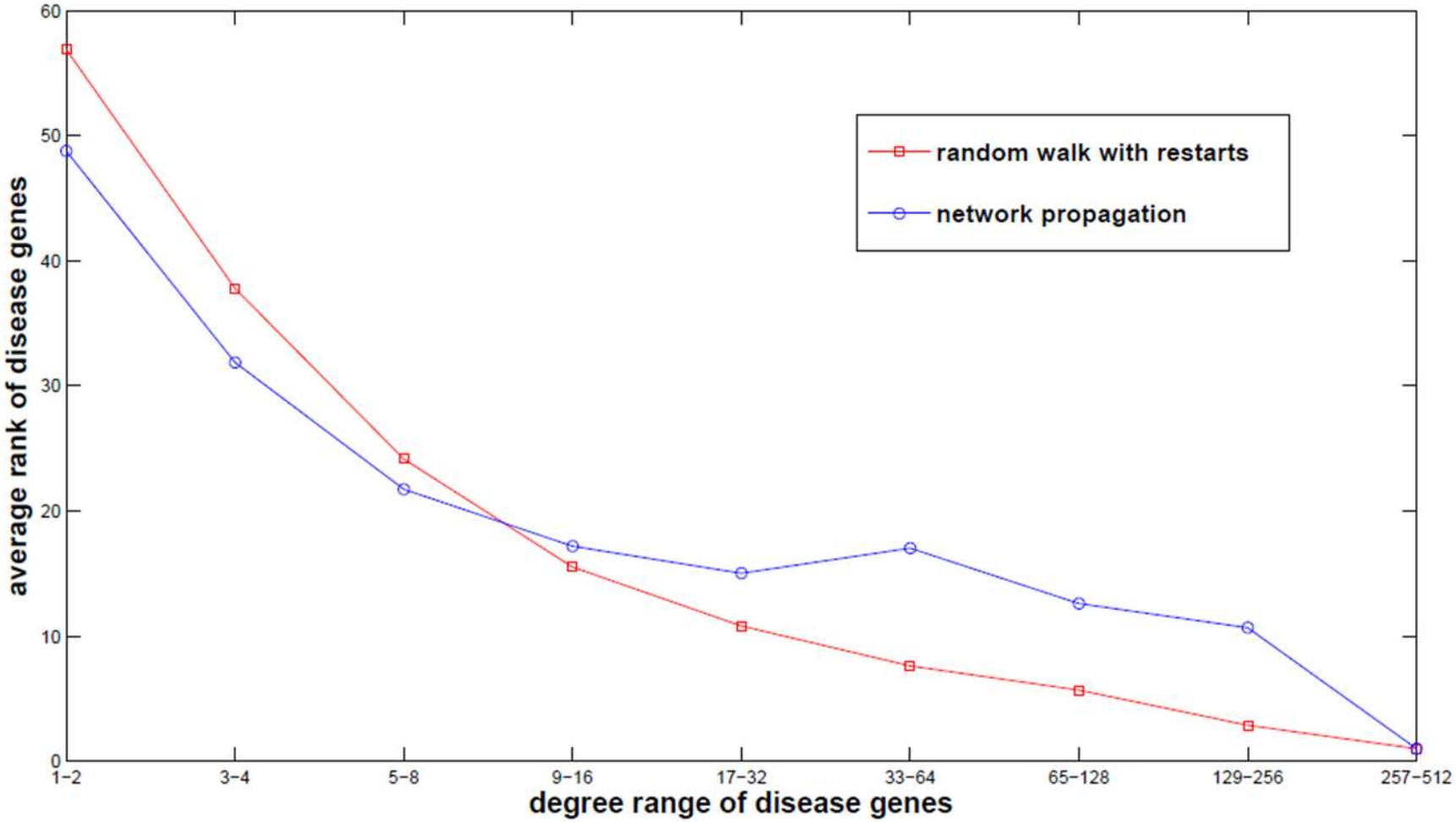
	Candidate deg.			Seed deg.			Centrality		
	$R_{\text{UNI}}^{(\text{C})}$	$R_{\text{UNI}}^{(\text{O})}$	$R_{\text{UNI}}^{(\text{S})}$	$R_{\text{UNI}}^{(\text{C})}$	$R_{\text{UNI}}^{(\text{O})}$	$R_{\text{UNI}}^{(\text{S})}$	$R_{\text{UNI}}^{(\text{C})}$	$R_{\text{UNI}}^{(\text{O})}$	$R_{\text{UNI}}^{(\text{S})}$
Avg. Rank	**23.22**	24.33	23.30	25.01	25.29	25.42	24.95	24.92	24.02
AUROC	0.76	0.76	**0.77**	0.75	0.75	0.76	0.75	0.75	0.76
Top 1%	**21.7**	19.4	14.7	18.4	18.5	19.3	20.0	20.5	21.3
Top 5%	45.1	44.4	42.1	45.5	44.1	41.2	46.3	45.7	**47.0**

- No clear winner, but models based on candidate degree perform consistently well together.

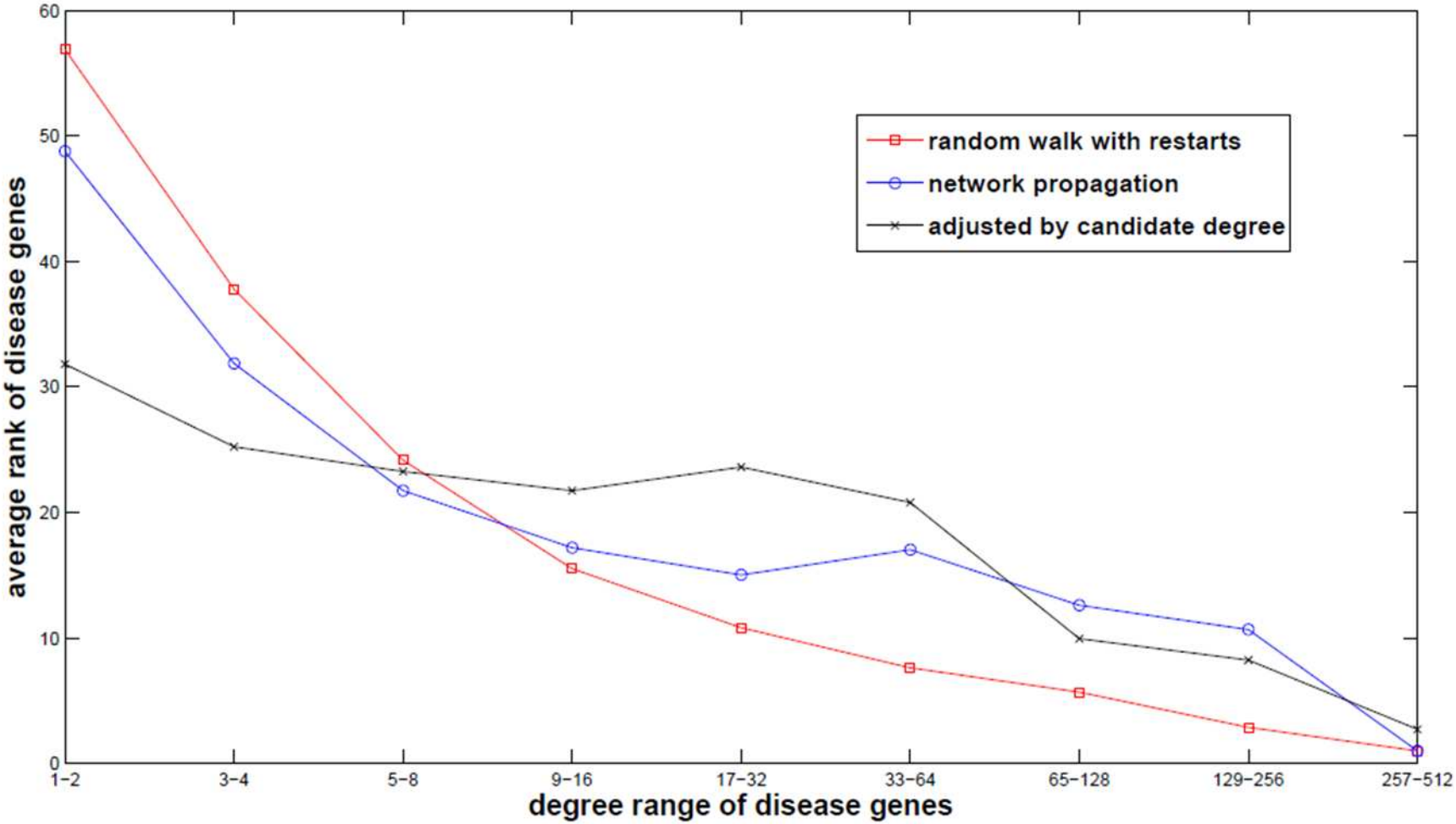
Overall performance



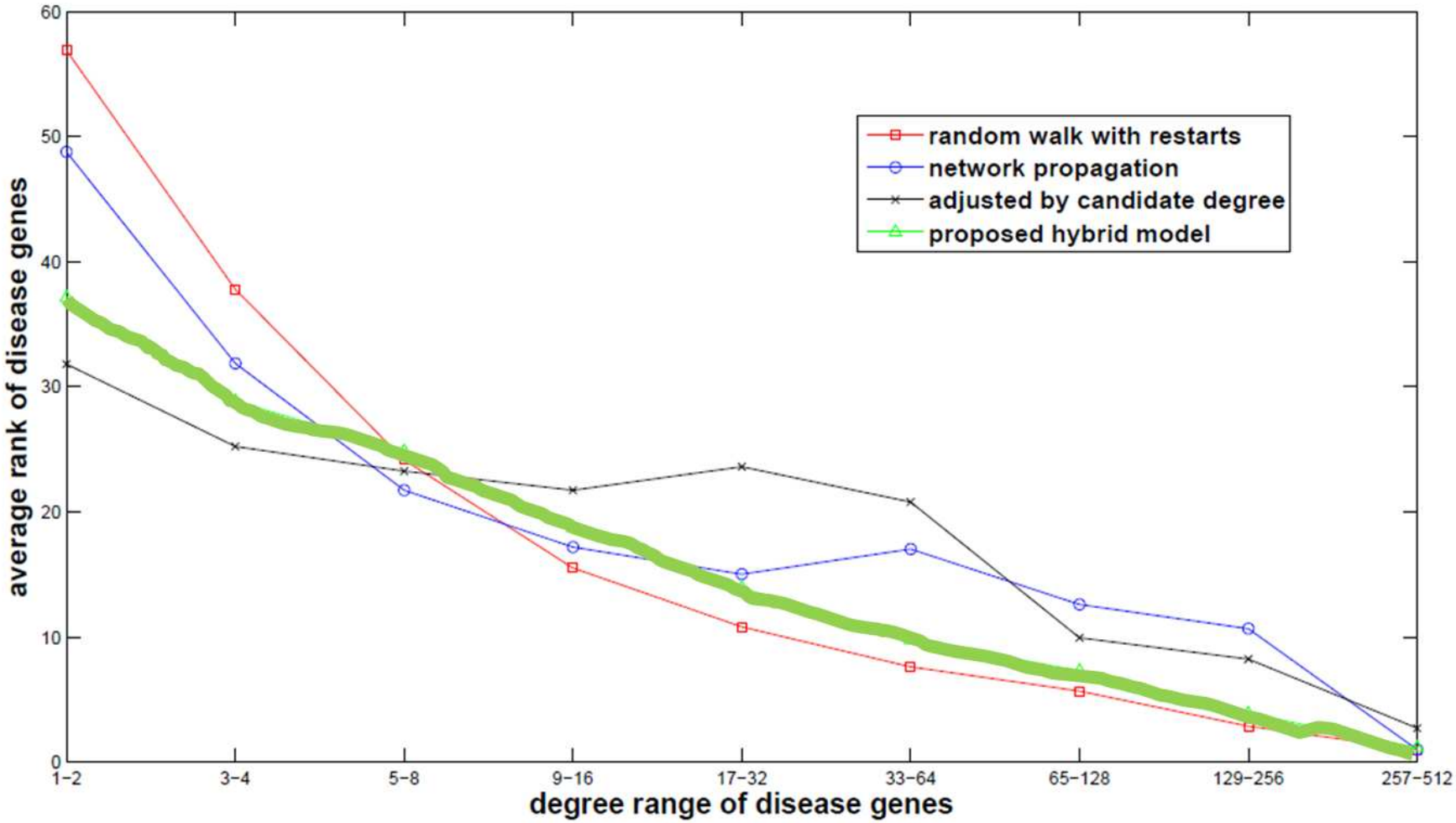
Effect of network degree



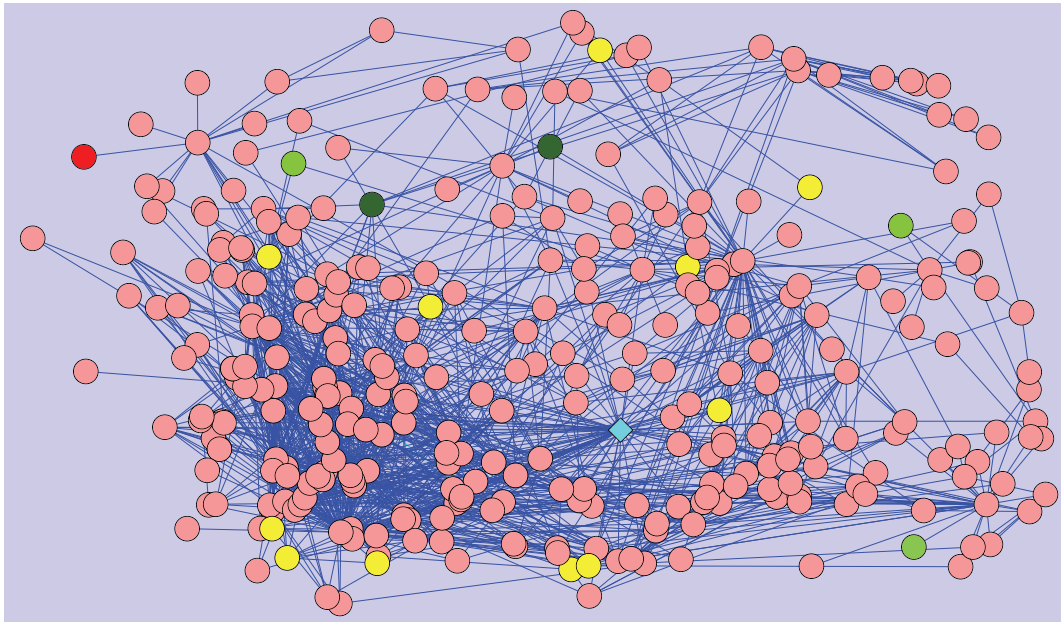
Effect of network degree



Effect of network degree



Case example

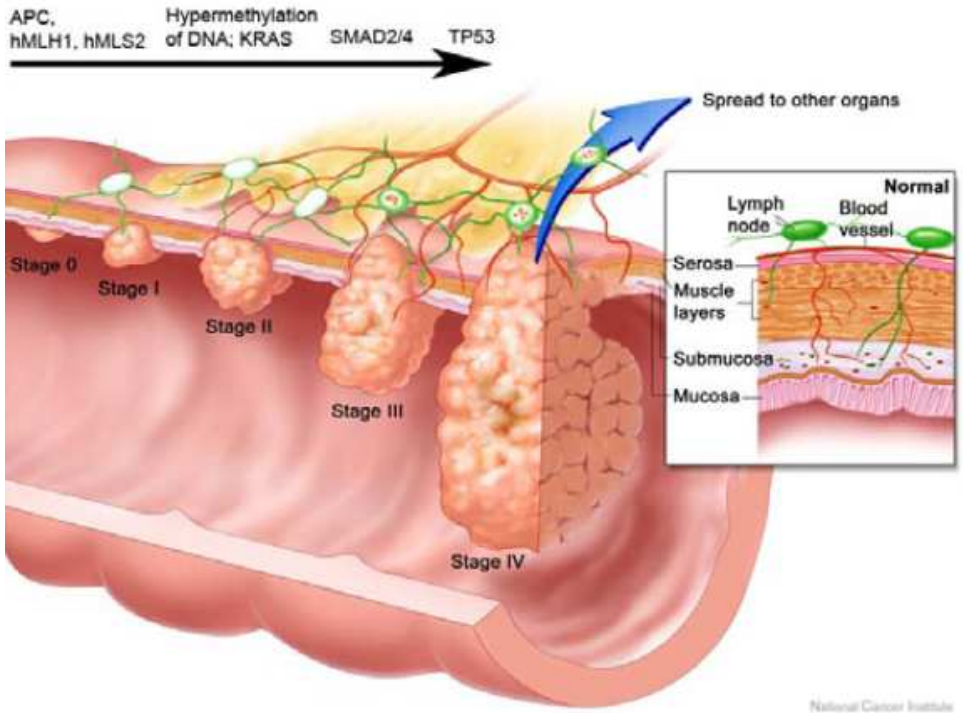


- Microphtalmia disease
  - Three associated genes: *SIX6*, *CHX10*, *BCOR*
  - Target gene: *BCOR* (red circle), Other candidate genes: Yellow circles
  - Level of assoication with Microphtalmia: Shade of green
  - *AKT1*: Diamond, ranked 1st by both competing methods
  - *BCOR* ranked 1st by our approach, 16th by both competing methods

Proteomics-Driven Identification of Important Subnetworks in Human Colorectal Cancer

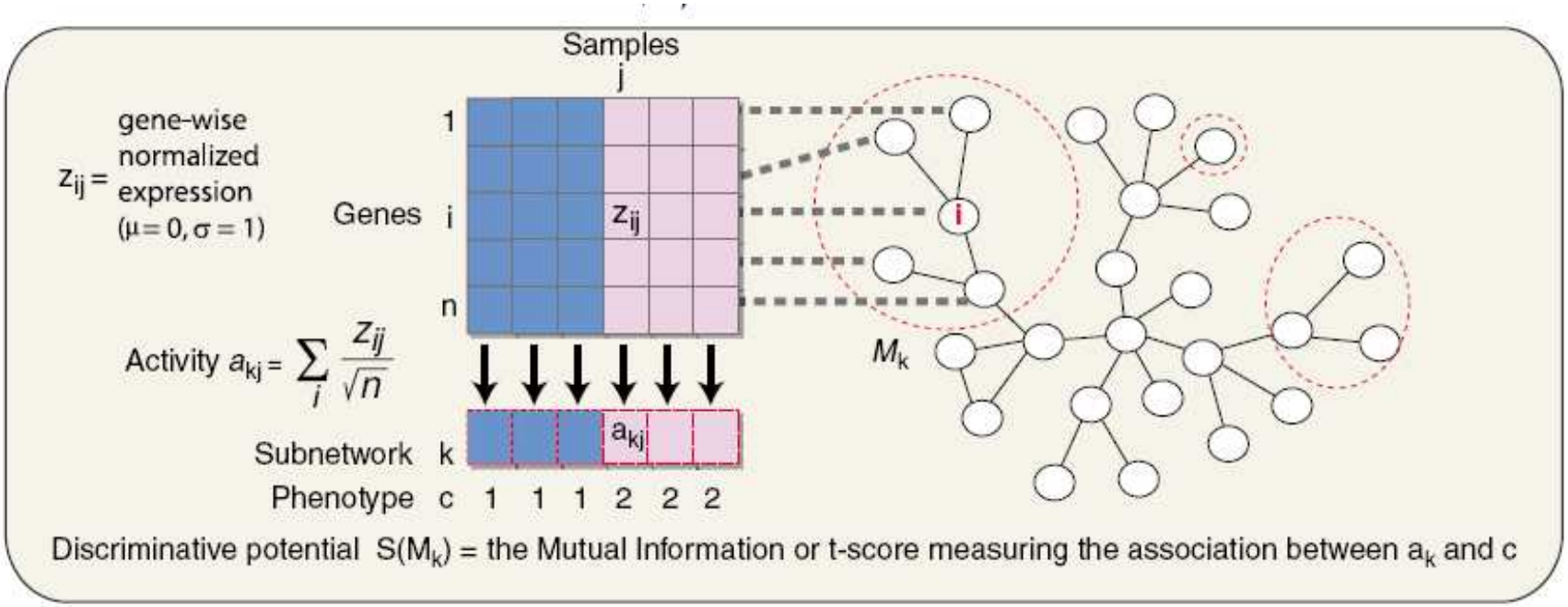
Human colorectal cancer (CRC)

- Second leading cause of cancer deaths in the United States.
- One out of every 19 individuals will be diagnosed with CRC in their lifetime.
- CRC is a complex, progressive disease.
  - Identification of **multiple markers** is important for effective prognosis and intervention.



Network-based identification of multiple markers

- Protein-protein interactions (PPIs) highlight functional relationships among proteins.
- We can identify **subnetworks** that are **coordinately** dysregulated in tumorigenic (or metastatic) samples.



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

Searching for coordinately dysregulated subnetworks

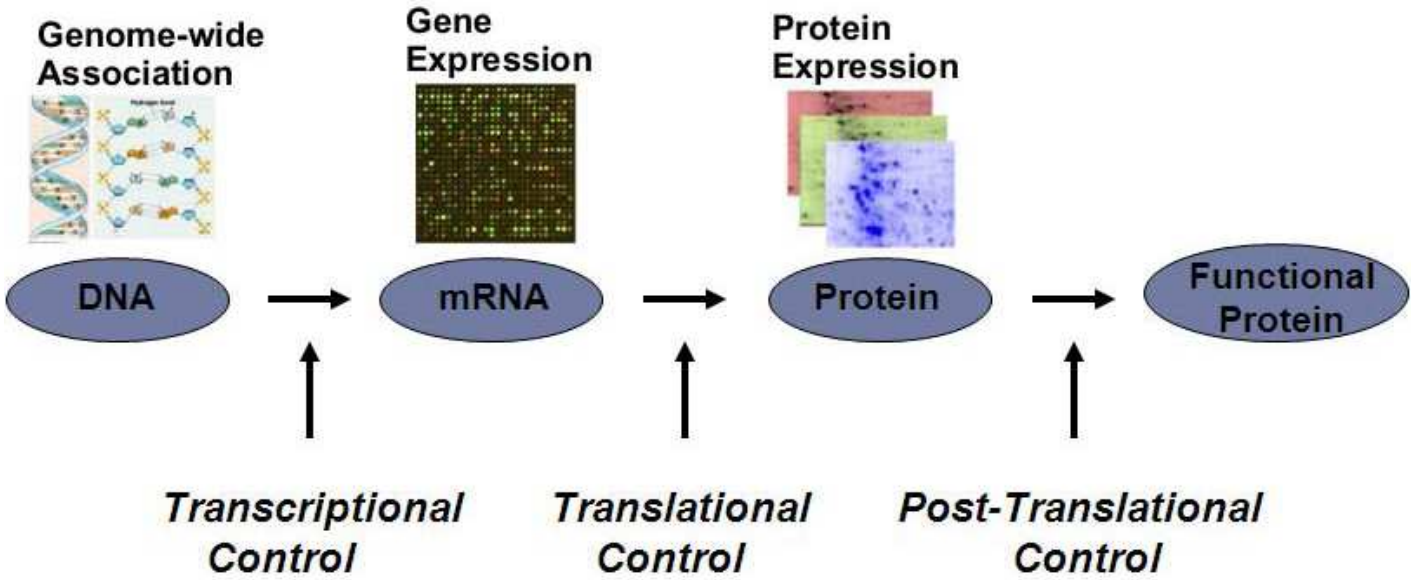
- Existing approaches use **mRNA** expression data and **greedy** algorithms based on **additive** formulation of coordinate dysregulation.

Our approach

Utilize other *omic* datasets to gain additional biological insights and improve upon greed using computational insights

1. **Proteomics-driven** identification of subnetwork markers (Part 2)
2. NetCover: **Combinatorial** algorithms for identification of subnetwork markers (Part 3)

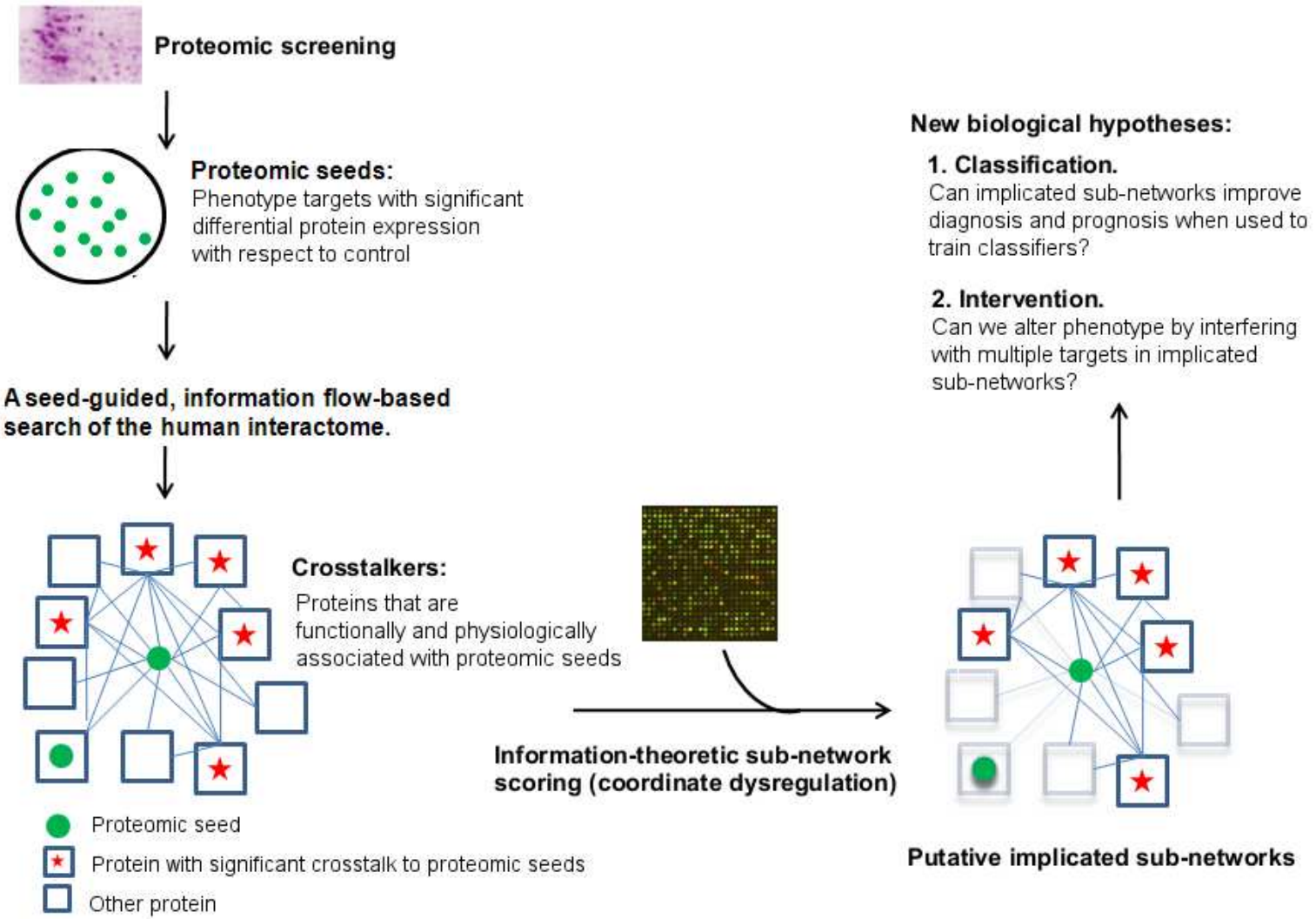
Utilizing protein expression data



Protein *vs.* mRNA (gene) expression

- **Transcriptomic data:** genome-wide monitoring of mRNA expression.
- **Proteomic data:** more reliable information at the functional level.

Proteomics-driven approach to subnetwork discovery



Crosstalk to proteomic seeds

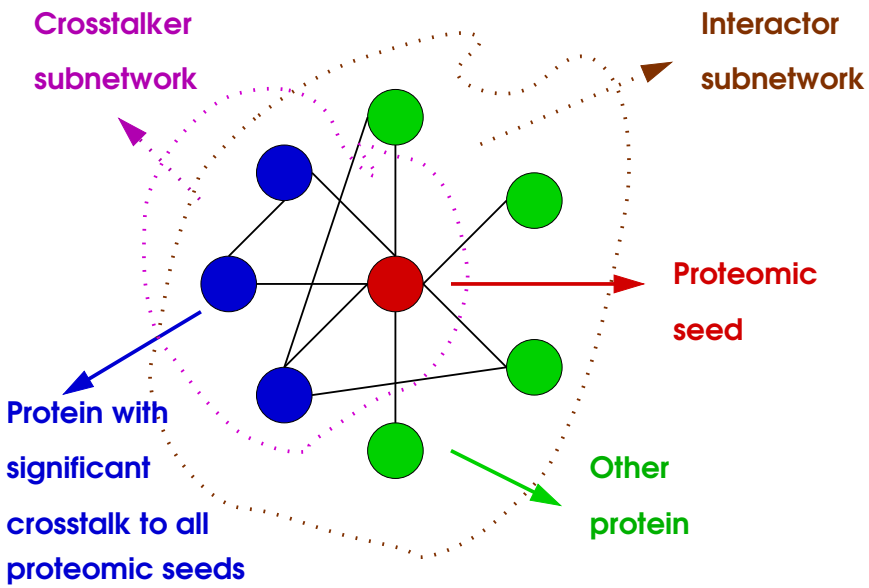
- Quantify the crosstalk between the set of proteomic seeds $\mathcal{Q}$ and each protein in human PPI network.
- Random walk with restarts: Simulate a random walk that makes frequent restarts at proteomic seeds!

$$\phi_0 = r,\ \phi_{t+1} = (1-c)P\phi_t + cr,\ \phi = \lim_{t\to\infty} \phi_t$$

  - r : Restart vector; $r(s) = 1/|\mathcal{Q}|$ for $s \in \mathcal{Q}$, 0 otherwise
  - c: Restart probability

- *Significant* $\phi \Rightarrow$ functional association with proteomic seeds $\Rightarrow$ involved in the progression of CRC?

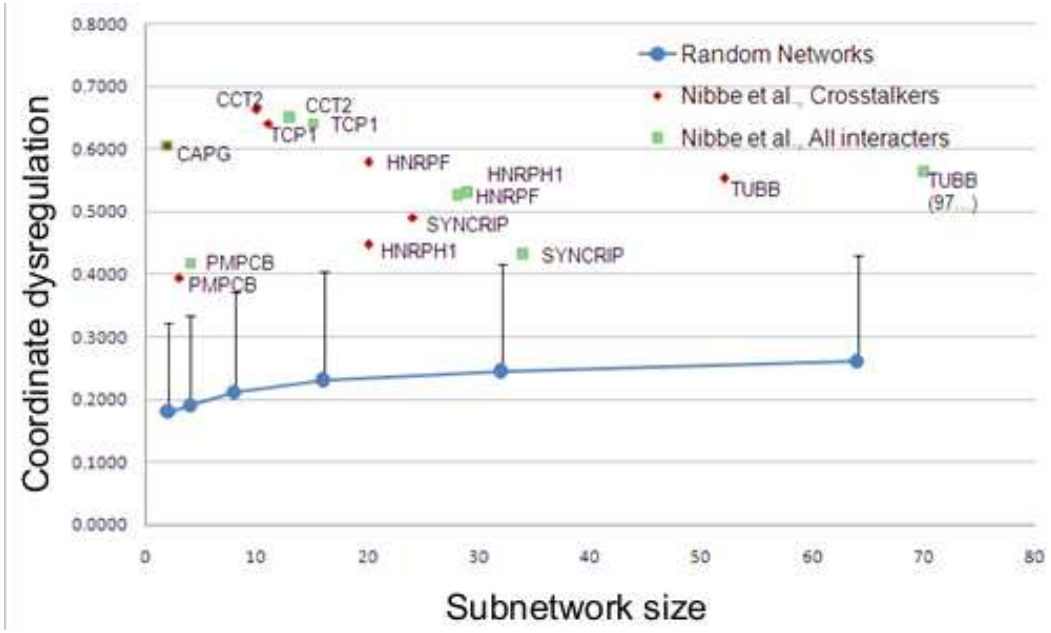
Crosstalk to seeds and coordinate dysregulation



Hypothesis

Proteins with significant crosstalk to proteomic seeds are likely to exhibit significant *coordinate* mRNA-level dysregulation in CRC.

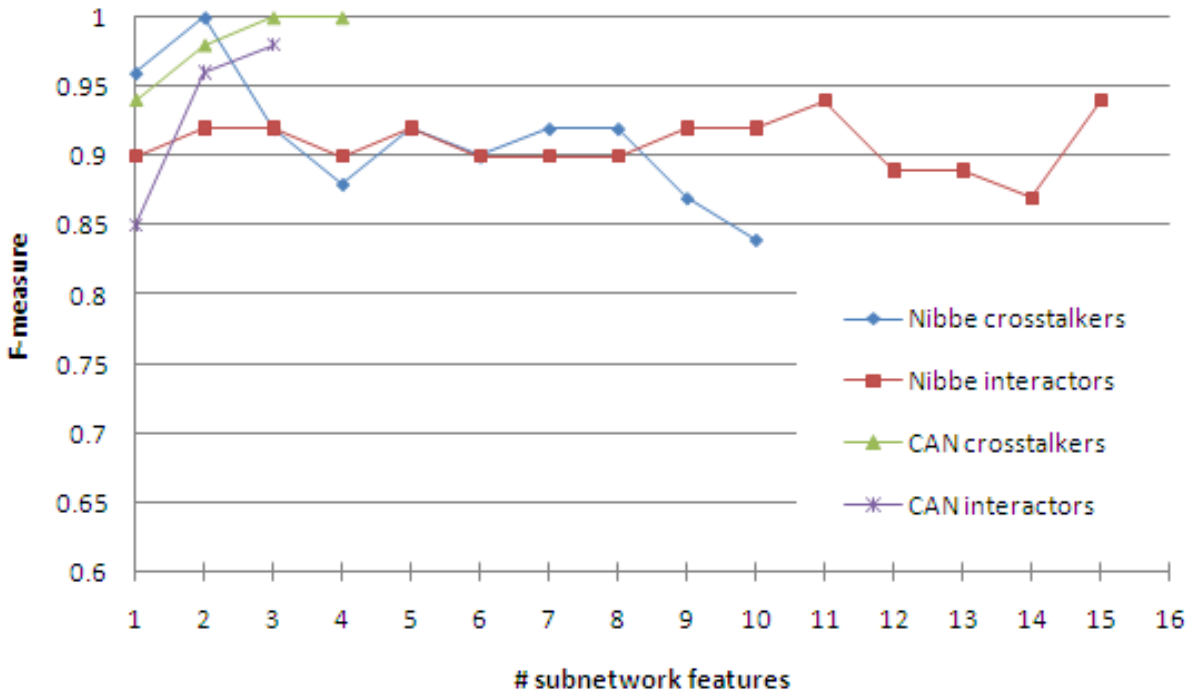
Crosstalkers *vs.* interactors



- **Proteomic seeds:** 67 proteins with significant ($p < 0.05$) differential protein expression in paired samples from 12 patients with late-stage CRC (Nibbe *et al.*, *Mol Cell Prot*, 2009).
- **Gene expression data:** *GSE8671*, 32 prospectively collected adenomas paired with those of normal mucosa (Sabates-Beliver *et al.*, *Mol Cancer Res*, 2007).

Classification performance

- Subnetworks identified on *GSE8671* are used to train classifiers to classify samples in *GSE10950* (Yu *et al.*, *Cancer Cell*, 2008).
- The "subnetwork activity" (aggregate expression profile) of each subnetwork is used as a feature.

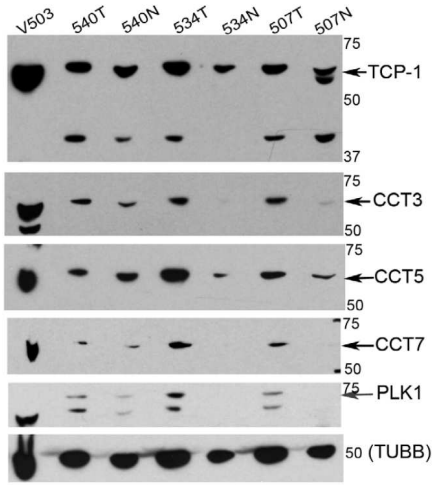


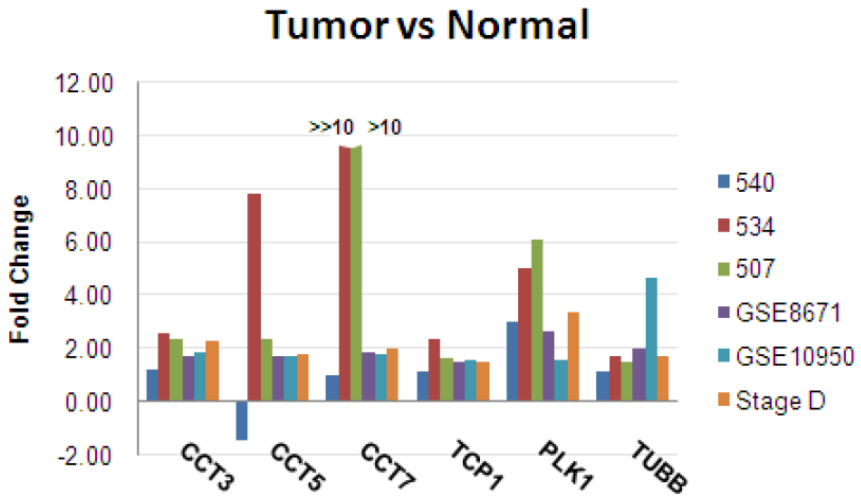
Experimental validation

- Subunits CCT1, CCT3, and CCT7 of the CCT (Chaperonine containing TCP1) complex exhibit significant crosstalk to proteomic seeds and optimize classification performance.
- But they are not reported to be implicated in CRC.

Prediction

These proteins will exhibit significant post-translational dysregulation in CRC.





Combinatorial Modeling of Coordinate Dysregulation in Cancer

Searching for coordinately dysregulated subnetworks

Coordinate dysregulation

- Subnetwork: $\mathcal{S} = \{g_1, g_2, ..., g_m\}$
- Subnetwork activity: $E_{\mathcal{S}} = \sum_{i=1}^{m} E_i / \sqrt{m}$
- Coordinate dysregulation: $I(E_{\mathcal{S}}; C) = H(C) - H(C|E_{\mathcal{S}})$

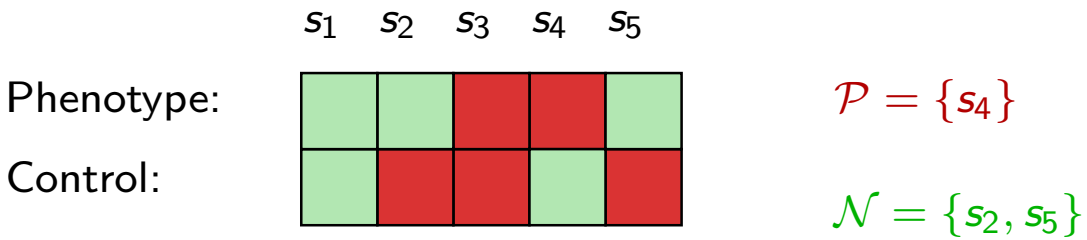
Computational problem

Given a PPI network and gene expression dataset, find subnetworks with maximal $E_{\mathcal{S}}$.

- Algorithms that aim to *greedily* maximize $E_{\mathcal{S}}$ may not suit well to the combinatorial nature of this problem.

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.



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

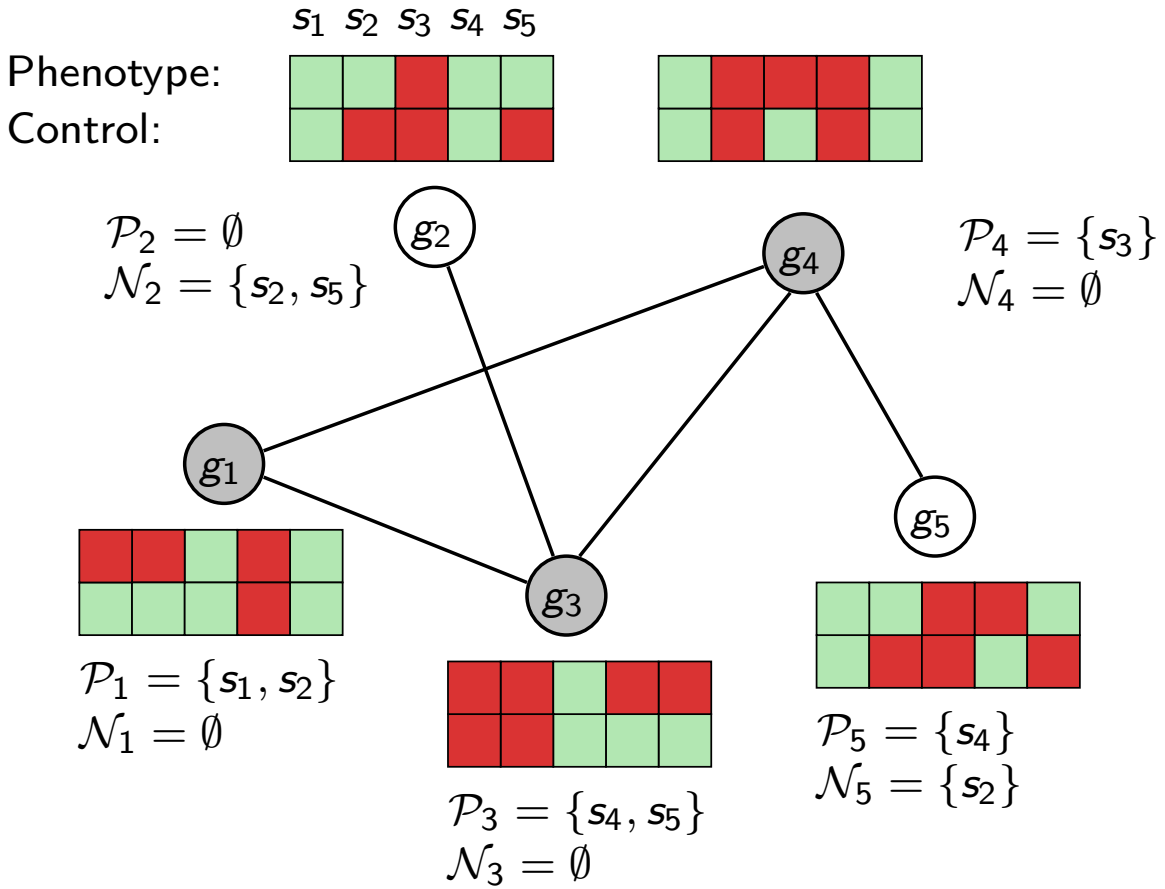
Cover and dysregulation

- 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



$\mathcal{S} = \{g_1, g_3, g_4\}$ $\mathcal{P}(\mathcal{S}) = \{s_1, s_2, s_3, s_4, s_5\}$ $\mathcal{N}(\mathcal{S}) = \emptyset$

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.

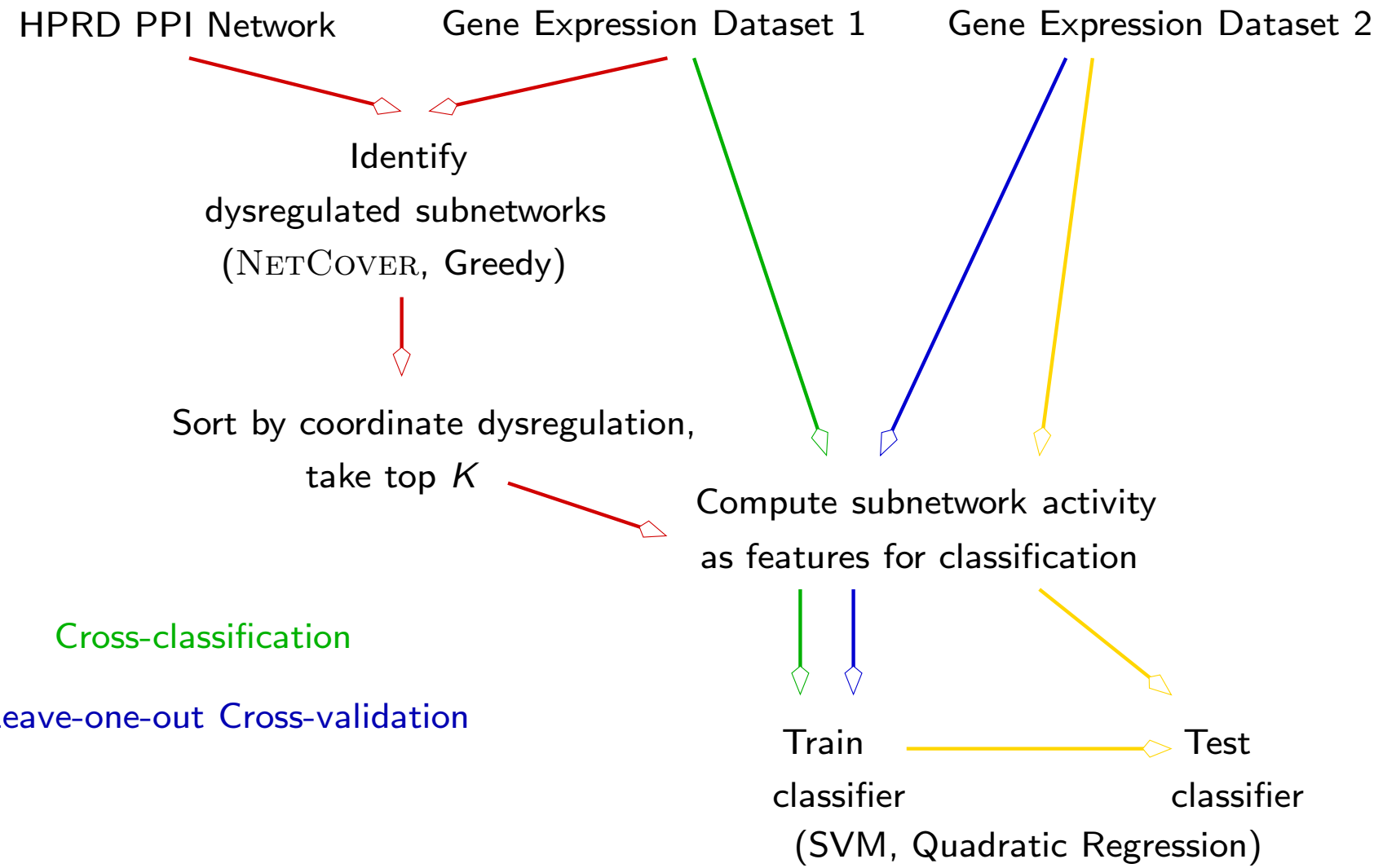
NetCover

- 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).

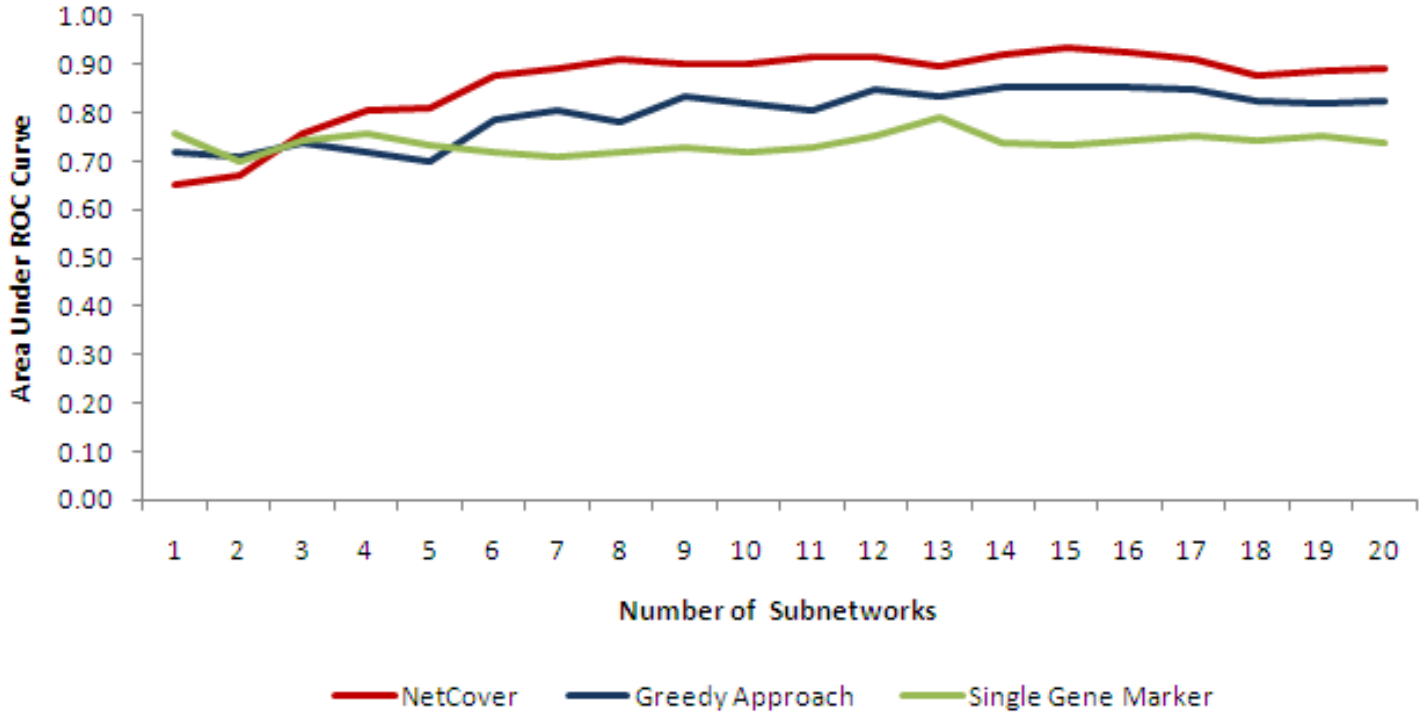
Classification framework



Experimental Setup

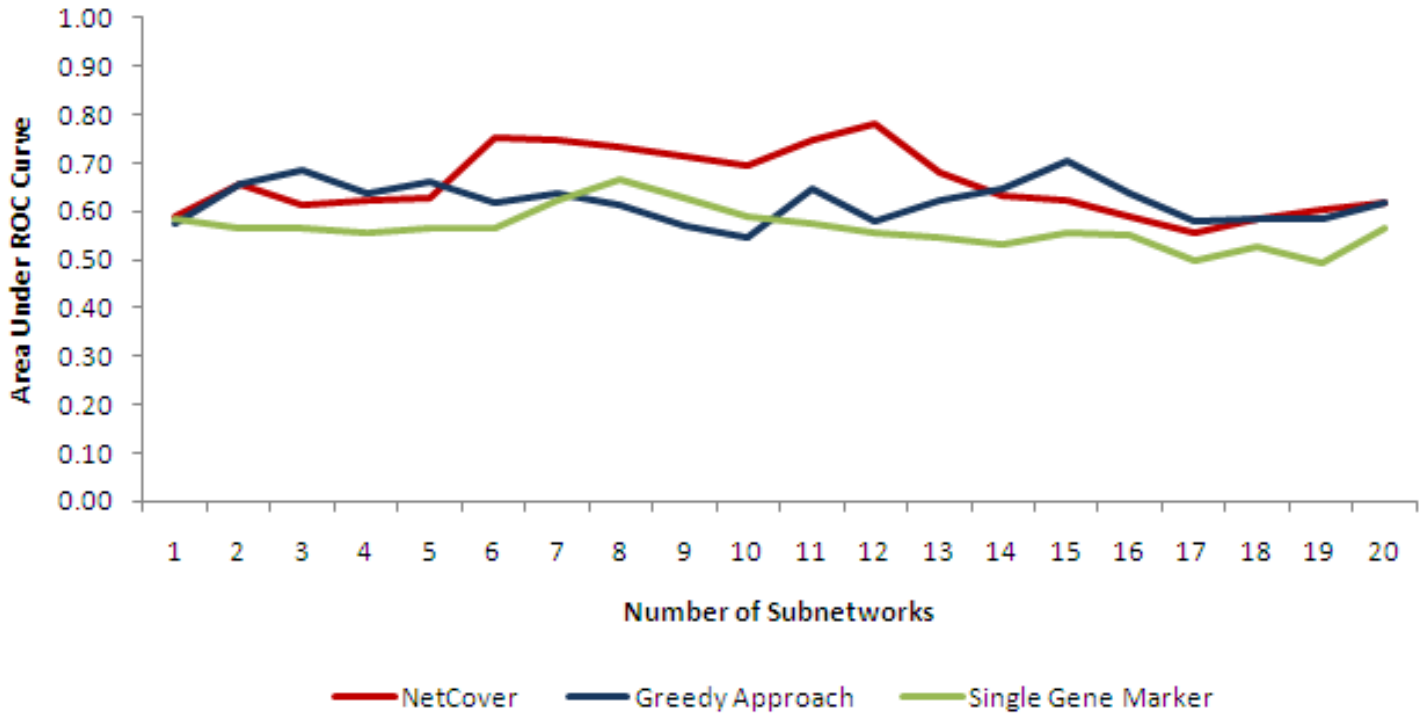
- Classification tasks
  - Diagnosis: Discriminating tumor samples from normal.
  - Prognosis: Discriminating metastatic samples from primary tumor.
- Datasets
  - GSE8671: 32 adenoma samples paired with normal mucosa.
  - GSE10950: 24 normal and tumor pairs.
  - GSE6988: 27 liver metastasis, 20 primary colorectal tumors, 25 normal mucosa.
- Algorithms
  - NetCover.
  - Greedy algorithm with coordinate dysregulation as the objective function.
  - Single gene markers (no network information).

Predicting tumor



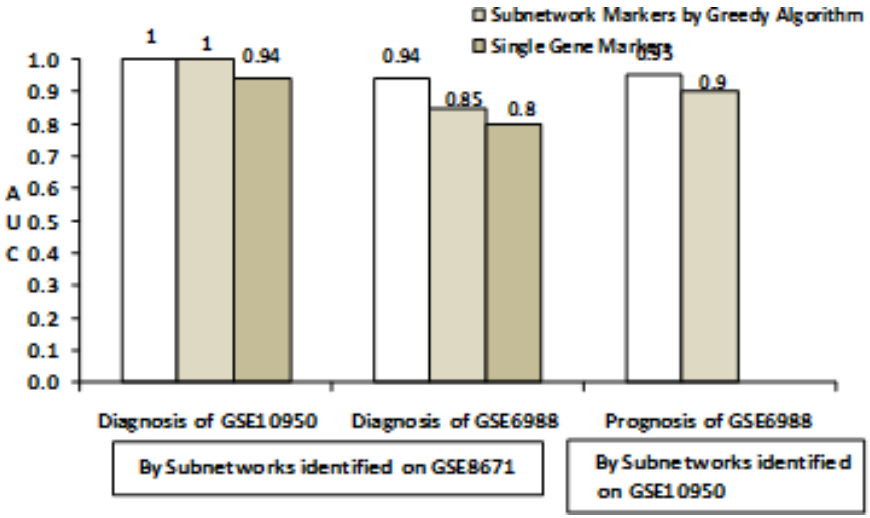
- Subnetwork identification & training: GSE8671.
- Testing: GSE6988.
- Classifier: SVM, Cross-classification.

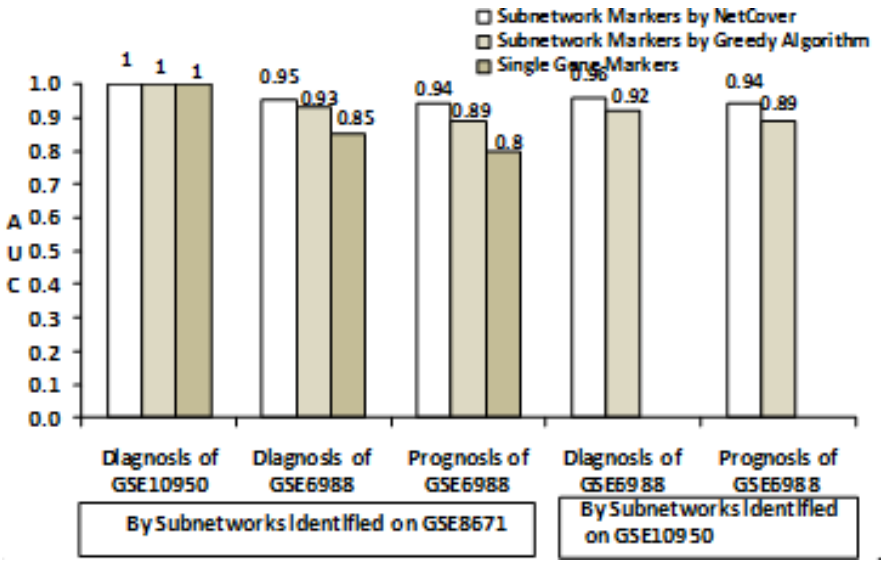
Predicting metastasis



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

Overall performance

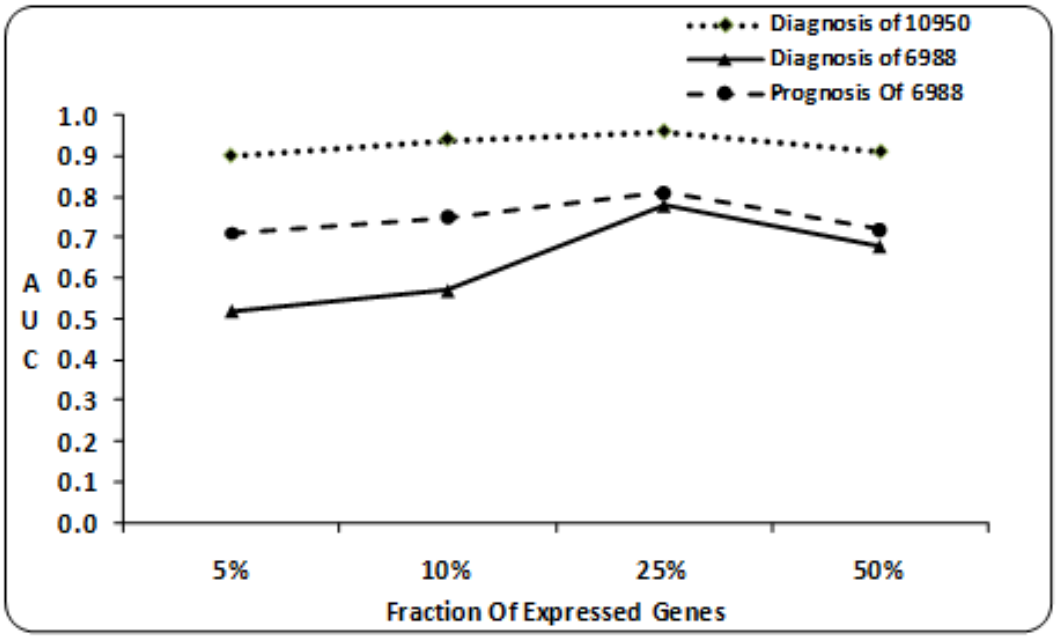




- Classifier: SVM.
- Best performance achieved by each algorithm is reported.

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 $\downarrow$.



Conclusions

1. Statistical significance with respect to degree distribution matters in network-based biological inference.

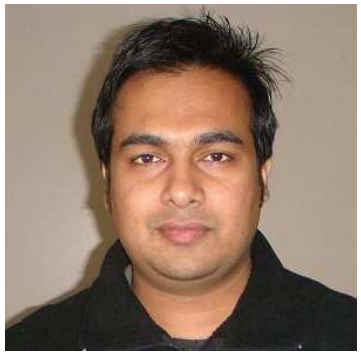
2. Information theoretic formulation of coordinate dysregulation is promising.

3. Genomic and proteomic data can provide shortcuts for important subnetwork identification.

4. Consideration of samples that are discriminated by each gene better captures coordinate dysregulation of multiple genes.

Acknowledgments

Sinan Erten



Salim Chowdhury

Rod Nibbe

Mark Chance

- Gökhan Yavaş, Ted Roman, Xin Li, Jing Li, Meral Özsoyoğlu (EECS); Alex Galante (Biology); Tom LaFramboise (Genetics); Vishal Patel, Gurkan Bebek, Sudipto Saha, Rob Ewing (Proteomics).

- NSF CAREER Award CCF-0953195.