Role of Centrality in Network Based Prioritization of Disease Genes

Sinan Erten[1] and **Mehmet Koyutürk**[1,2]

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

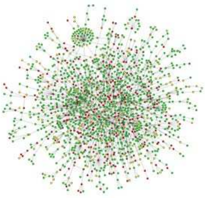
8th European Conference on Evolutionary Computation, Machine Learning, and Data Mining in Bioinformatics
April 8, 2010

Complex diseases

- Many diseases are based on a set of complex interactions between multiple genetic and environmental factors.
  - □ Heart disease, high blood pressure, Alzheimer's disease, diabetes, cancer, obesity, etc.
- Genome-wide association studies (GWAS) hint on where disease-associated genes might be located on the genome (linkage interval), but such intervals might contain up to 300 genes.

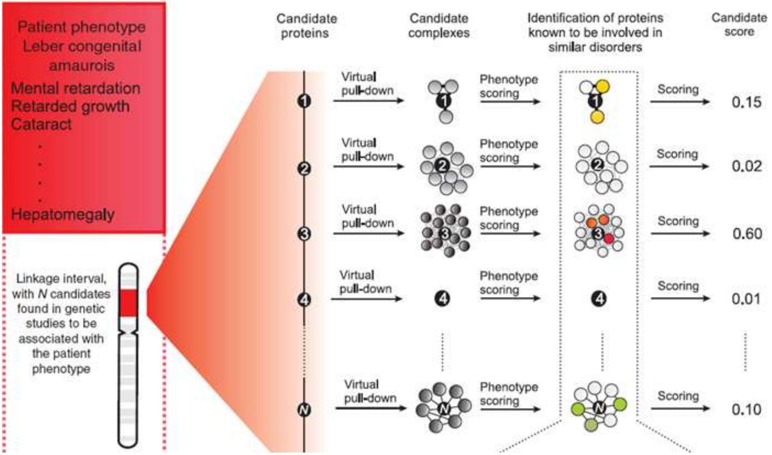
Protein-protein interaction (PPI) networks

- Physically interacting proteins can be identified via high-throughput screening.
- Nodes represent proteins.
- Edges represent interactions.
  - □ Binding, regulation, modification, transport, complex membership...
- Many public databases of PPIs (*e.g.*,HPRD, DIP, BIOGRID).



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

PPI networks in disease gene prioritization



Patient phenotype
Leber congenital amaurois
Mental retardation
Retarded growth
Cataract
.
.
.
.
Hepatomegaly
Linkage interval, with *N* candidates found in genetic studies to be associated with the patient phenotype
Candidate proteins
Candidate complexes
Identification of proteins known to be involved in similar disorders
Candidate score
1
2
3
4
N
Virtual pull-down
Phenotype scoring
Scoring
0.15
0.02
0.60
0.01
0.10

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.

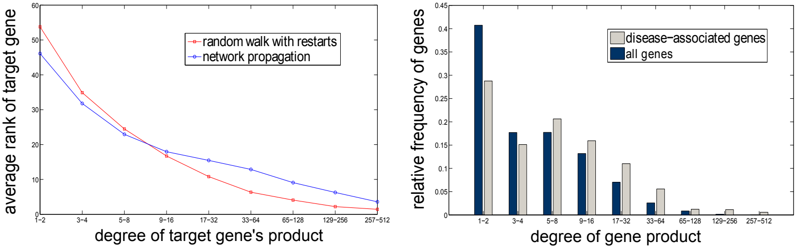
$$P_{\mathrm{RW}}(u, v) = 1/|\mathcal{N}(v)| \text{ for } uv \in E,\ 0 \text{ otherwise.}$$

- 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



(a)

(b)

- 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 $c = 0$ is a measure of network centrality (equivalent to Google page-rank).

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

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

Experimental Setup

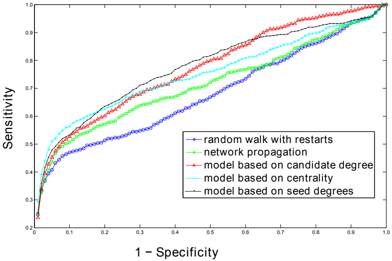
- Human PPI network: NCBI Entrez Gene database.
  - 33528 binary interactions between 8959 proteins.

- 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 $\approx$ 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.

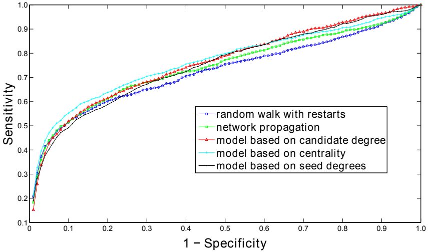
Effect of statistical adjustment

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):

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

  - □ Optimistic prioritization (local):

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

  - □ Based on seed degree (global):

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

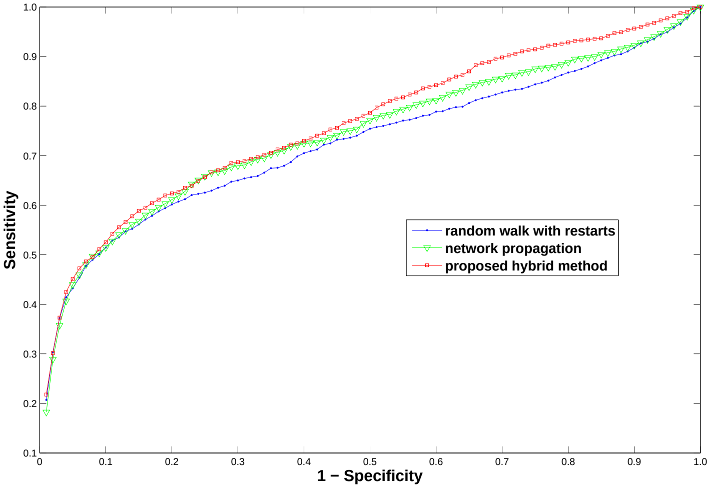
$$R^{(\mathsf{S})}_{\mathsf{UNI}}(v) = \begin{cases} R_{\mathsf{RAW}}(v) & \text{if } \overline{d}(\mathcal{S}) \; > \lambda \\ R_{\mathsf{ADJ}}(v) & \text{otherwise} \end{cases}$$

Performance of uniform prioritization schemes

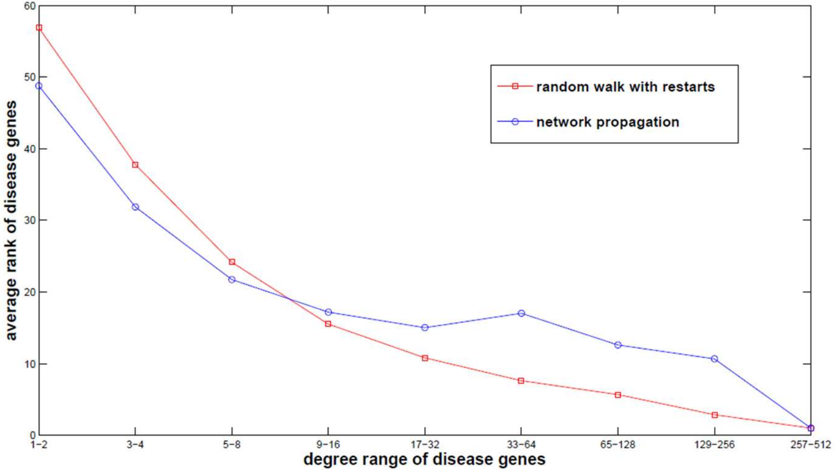
	Candidate deg.			Seed deg.			Centrality		
	$R^{(C)}_{\mathsf{UNI}}$	$R^{(O)}_{\mathsf{UNI}}$	$R^{(S)}_{\mathsf{UNI}}$	$R^{(C)}_{\mathsf{UNI}}$	$R^{(O)}_{\mathsf{UNI}}$	$R^{(S)}_{\mathsf{UNI}}$	$R^{(C)}_{\mathsf{UNI}}$	$R^{(O)}_{\mathsf{UNI}}$	$R^{(S)}_{\mathsf{UNI}}$
Avg. Rank	**23.22**	24.33	23.30	25.01	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.

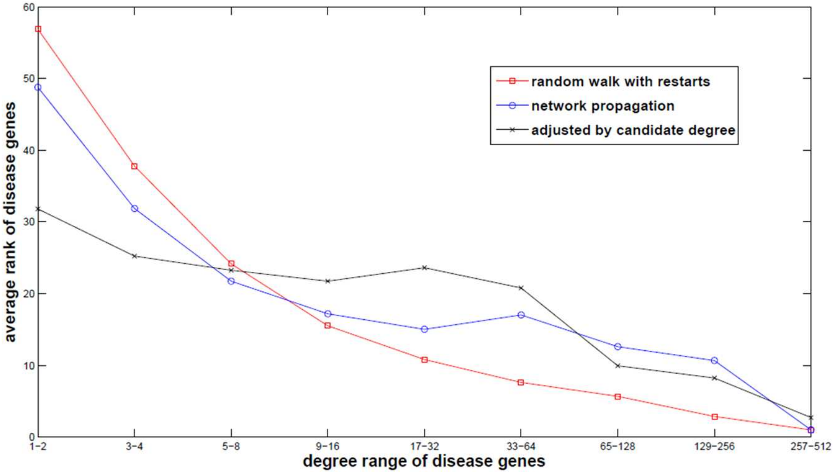
Overall performance



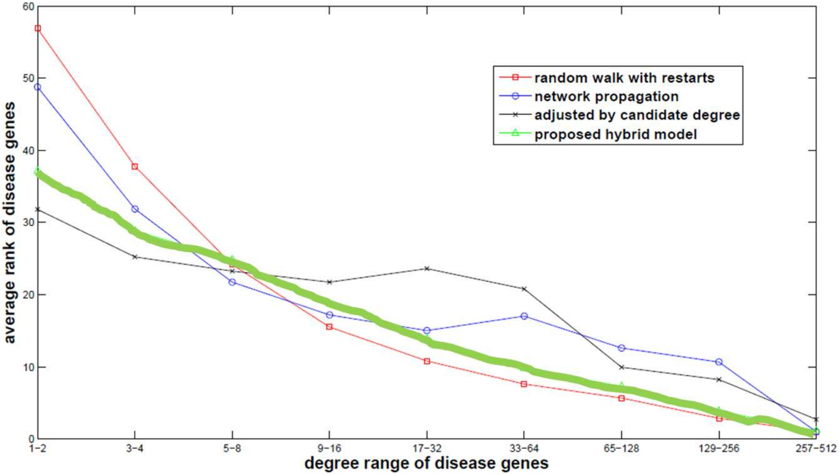
Effect of network degree



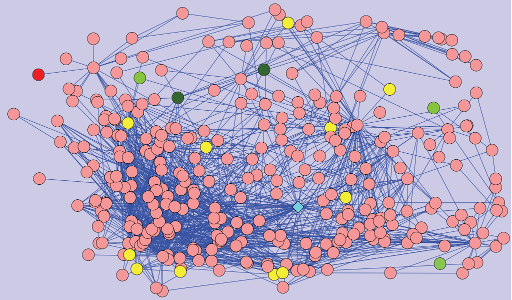
Effect of network degree



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

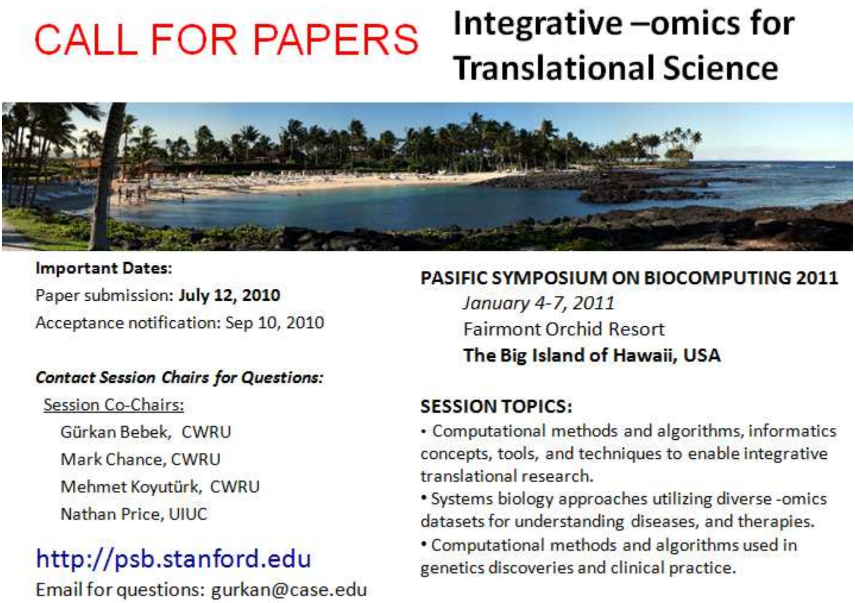
Acknowledgments

Sinan Erten

- Undergraduate students
  - □ Ted Roman (Computer Science/Math), Alex Galante (Biology)
- Graduate students
  - □ Vishal Patel (Genetics)
- Collaborators
  - □ Mark Chance, Rob Ewing, Sudipto Saha, Gurkan Bebek, Rod Nibbe (Center for Proteomics & Bioinformatics)

NSF CAREER Award
CCF-0953195

Case Western Reserve University est. 1826



CALL FOR PAPERS

Integrative –omics for Translational Science

Important Dates:

Paper submission: **July 12, 2010**

Acceptance notification: Sep 10, 2010

Contact Session Chairs for Questions:

Session Co-Chairs:

Gürkan Bebek, CWRU

Mark Chance, CWRU

Mehmet Koyutürk, CWRU

Nathan Price, UIUC

http://psb.stanford.edu

Email for questions: gurkan@case.edu

PASIFIC SYMPOSIUM ON BIOCOMPUTING 2011

January 4-7, 2011

Fairmont Orchid Resort

The Big Island of Hawaii, USA

SESSION TOPICS:

- Computational methods and algorithms, informatics concepts, tools, and techniques to enable integrative translational research.
- Systems biology approaches utilizing diverse -omics datasets for understanding diseases, and therapies.
- Computational methods and algorithms used in genetics discoveries and clinical practice.