Mining Genetic Determinants of Human Disease with Graph-based Learning

TaeHyun Hwang and Rui Kuang

Department of Computer Science and Engineering, University of Minnesota Twin Cities

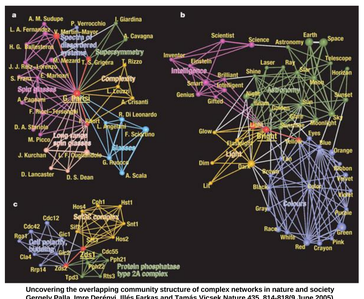
E-mail: thwang@cs.umn.edu, website: http://compbio.cs.umn.edu/

Understanding causative relations between genes and disease phenotypes has been a long standing challenge in genomics research. The recent 'omics projects offer the opportunity for a genome-wide study of the associations with the large scale human protein-protein interaction network, phenotype similarity network and their association network. Integrating the heterogenous genomic and phenotypic data and exploring the modularity among the genomic and phenotypic objects becomes the central computational problem. The key challenge is two-fold: how to integrate heterogenous data and how to utilize the modular structures. We propose to integrate the data as a large heterogeneous network and introduce an algorithm to explore the modules in the heterogenous network. The proposed algorithm is a mathematically principled and yet intuitive label propagation method. Extensive empirical experiments support that the algorithm can improve the accuracy of disease phenotype-gene association prediction over currently available methods. The algorithm is also a general method that is applicable to similar network integration problems in other research areas.

BACKGROUND

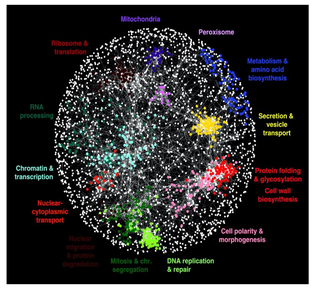
Motivation

- Many real world networks contain clusters, communities, cohesive groups or modules (such as functionally related genes, groups of people, and) among the objects in the network
- Several graph-based learning algorithms have been developed to utilize network structures to improve performance in different learning tasks on networks such as classification and ranking
- However, there is no general framework to explore network structures in the heterogeneous network containing different types of vertices and edges



Uncovering the overlapping community structure of complex networks in nature and society Gergely Palla, Imre Derényi, Illés Farkas and Tamás Vicsek Nature 435, 814-818(9 June 2005)

The community structure in three different networks



Michael Costanzo et al., The Genetic Landscape of a Cell, Science 22 January 2010: Vol. 327. no. 5964, pp. 425 - 431

Genetic interactions of cell in yeast

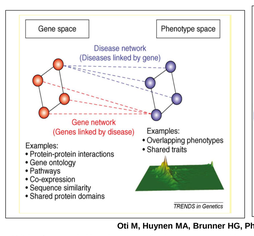
Challenge

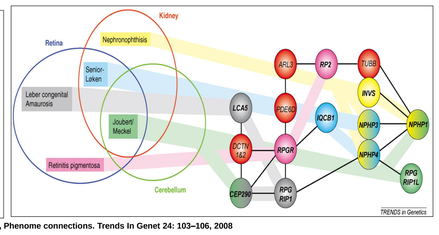
Network Data Integration perspective:

- **Heterogeneity**
  : The heterogeneous networks may contain several subnetworks consisted of different types of vertices and edges
  : Some vertices are connected to many others, some to very few, and interaction strengths and dynamics may vary widely
- **Bias**
  : The heterogeneous networks may have subnetworks containing unbalanced sizes, different noisy levels and different edge-weight scales
- **Scalability**
  : The heterogeneous networks may have more than tens of thousands vertices and edges

Biological Network Data Integration perspective:

- **Network View**
  : Available genomic data needs to be integrated with “network” view of systems biology
- **Exploring modular structure of complex biological systems**
  : Need to develop an efficient technique for exploring communities on a large scale
- **Interpretability**
  : Results from the technique should be interpretable and biologically meaningful.





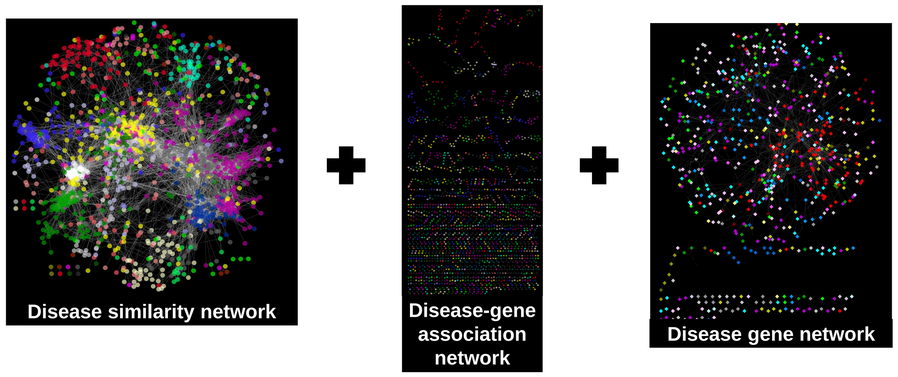
Oti M, Huynen MA, Brunner HG, Phenome connections. Trends In Genet 24: 103–106, 2008

“Network” view of human disease and gene association.
Each node indicates disease and gene. Phenotypically similar disease can be grouped together in the disease network, and genes have similar functions can be grouped together in gene networks. Disease genes for phenotypically similar diseases may reside in similar biological modules and subnetworks in the protein interaction networks.

METHODOLOGY

Problem Formulation

- **Given**: Multiple network data. Each network has its own cluster structures containing different types of vertices and edges.

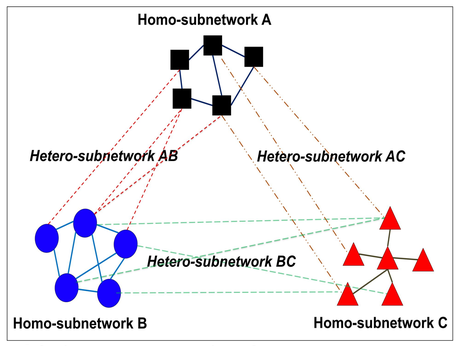


- **Objective**:
  a) Predict label / activation value to the unlabeled data for classification or ranking
  b) Find community structure in the heterogeneous network

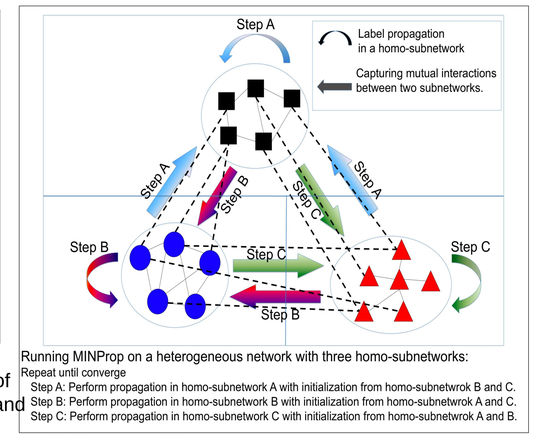
Approach

Mutual Interaction based Network Propagation (MINProp)

:To handle label propagation on a complex heterogeneous network, MINProp sequentially performs network propagations on each individual homo and hetero-subnetwork



• A heterogeneous network
This heterogeneous network contains three types of vertices, and accordingly three homo-subnetwork and three heterogeneous network



Running MINProp on a heterogeneous network with three homo-subnetworks:
Repeat until converge
Step A: Perform propagation in homo-subnetwork A with initialization from homo-subnetwork B and C.
Step B: Perform propagation in homo-subnetwork B with initialization from homo-subnetwork A and C.
Step C: Perform propagation in homo-subnetwork C with initialization from homo-subnetwork A and B.

• MINProp workflow
Illustration of the MINProp algorithm. Label propagation initialized by the interactions with the other homo-subnetworks is sequentially performed on each individual homo-subnetwork

Regularization framework

$$\Omega(f) = \sum_{i=1}^{k} \left(f_i^T \Delta f_i + \mu_i \left\| f_i - y_i \right\|^2 \right) + \frac{1}{2} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \mu_{ij} \left[f_i^T \; f_j^T \right] \sum^{(i,j)} \begin{bmatrix} f_i \\ f_j \end{bmatrix}$$

- f : predicted label
- y : initial label
- Δ : graph laplacian of homo-subnetwork
- k : number of subnetwork
- Σ : graph laplacian of hetero-subnetwork
- μ_i and μ_{ij} : positive constants
- $\left\| f_i - y_i \right\|^2$: fitting term which keeps the final label values consistent with the initial labels
- $f_i^T \Delta^i f_i$: smoothness constrain on the homo-subnetwork

that enforces a consistent labeling of the strongly connected vertices in $V^{(i)}$

- $\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \mu_{ij} \left[f_i^T \; f_j^T \right] \sum^{(i,j)} \begin{bmatrix} f_i \\ f_j \end{bmatrix}$: smoothness term in the hetero-subnetwork

RESULTS

Task

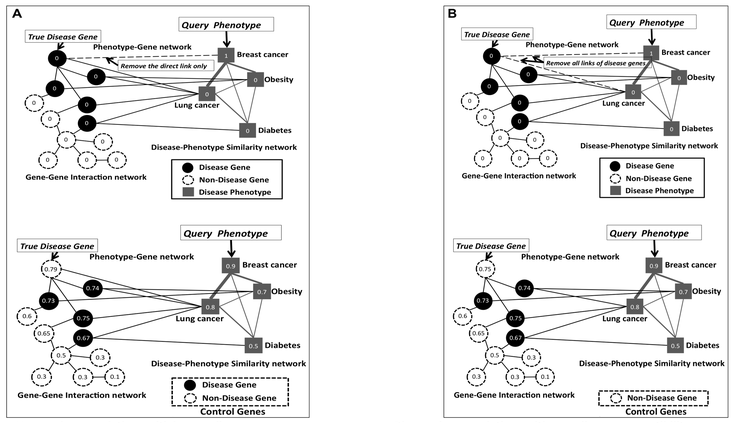
- **Given query disease phenotype, we want to highly rank its disease causative genes. The ranking results can be used to discover novel disease genes.**

Datasets

Three heterogeneous network are used in this study.

- Disease similarity network
  : an undirected graph with 5080 disease vertices. Edges are weighted by the pairwise quantitative measurments of phenotypic overlap in text and clinical synopsis of OMIM
- Human protein-protein interaction network
  :ppi-network contains 34,364 binary-valued undirected interactions between 8919 human proteins (genes)
- Disease-gene association network
  :an undirected bipartite graph-with disease vertices and gene vertices. There are 1126 disease-gene associations in the network

Experiments



Uncovering associations with known disease genes | **Discovering associations with new disease susceptibility genes**

Performance of ranking disease genes in leave-one-out cross validation

Methods	Associations with known disease genes Avg. AUC *(win/draw/loss)*	Associations with new disease genes Avg. AUC *(win/draw/loss)*
MINProp vs. Random Walk	**0.805** vs. 0.797 (738/75/313)	**0.728** vs. 0.648 (1045/2/79)
MINProp vs. CIPHER-DN	**0.863** vs. 0.738 (565/5/288)	**0.821** vs. 0.738 (515/11/332)
MINProp vs. CIPHER-SP	**0.805** vs. 0.734 (678/8/440)	0.728 vs. 0.729 (538/54/534)

Exploring modularity of genes

CC	MINProp vs. RW Avg. AUC	MINProp vs. CIPHER-DN Avg. AUC	MINProp vs. CIPHER-SP Avg. AUC	Hybrid vs. CIPHER-SP Avg. AUC
$[0.1, 1]$	**0.875** vs. 0.776	**0.889** vs. 0.855	**0.875** vs. 0.813	**0.886** vs. 0.813
$[0.01, 0.1)$	**0.902** vs. 0.799	**0.906** vs. 0.799	**0.902** vs. 0.801	**0.911** vs. 0.801
$[0, 0.01)$	**0.653** vs. 0.586	**0.770** vs. 0.688	0.654 vs. **0.693**	0.692 vs. 0.693
Total	**0.728** vs. 0.648	**0.821** vs. 0.738	0.728 vs. 0.729	**0.756** vs. 0.729

MINProp is capable to discover novel disease genes

MIM#	Phenotype Name	HGNC Symbol	Ranking (%) MINProp	Ranking (%) CIPHER SP	Ranking (%) Random Walk	Status
171300	PHEOCHROMOCYTOMA	VHL	1 (0.000)	1105 (0.138)	1105 (0.138)	new
		GINP	16 (0.002)	1400 (0.157)	1400 (0.157)	known
		KIF1B	228 (0.028)	512 (0.064)	512 (0.064)	new
		RET	2894 (0.362)	1232 (0.154)	3104 (0.388)	new
607174	MENINGIOMA, FAMILIAL	NF2	1 (0.000)	1279 (0.143)	1279 (0.143)	known
		PTEN	5 (0.001)	1307 (0.147)	1307 (0.147)	known
		MN1	5986 (0.748)	1685 (0.211)	5797 (0.724)	new
600880	BUDD-CHIARI SYNDROME	F5	1 (0.000)	398 (0.050)	398 (0.050)	new
		JAK2	51 (0.006)	497 (0.062)	161 (0.020)	new
166710	OSTEOPOROSIS	LRP5	2 (0.000)	1541 (0.173)	1541 (0.173)	known
		CALCR	4 (0.000)	7661 (0.957)	7661 (0.957)	new
		COL1A1	5 (0.001)	8086 (0.907)	8086 (0.907)	known
		VDR	42 (0.005)	1402 (0.157)	1402 (0.157)	known
		COL1A2	3255 (0.407)	7182 (0.897)	3397 (0.424)	new
601626	LEUKEMIA, ACUTE MYELOID	MLF1	3 (0.000)	4323 (0.540)	4323 (0.540)	new
		JAK2	15 (0.002)	354 (0.044)	70 (0.009)	new
		ETV6	23 (0.003)	769 (0.096)	769 (0.096)	new
		GMPS	245 (0.031)	4512 (0.564)	4512 (0.564)	new
		CEBPA	526 (0.066)	2376 (0.297)	708 (0.088)	new
		SHGL1	629 (0.079)	1047 (0.131)	227 (0.028)	new
		FLT3	1860 (0.232)	2107 (0.263)	2463 (0.308)	new
		NPM1	1880 (0.235)	1177 (0.147)	2396 (0.299)	new
		BUNX1	1887 (0.212)	4156 (0.466)	1834 (0.206)	known
		KIT	3377 (0.379)	4024 (0.451)	3227 (0.362)	known
		CHIC2	3881 (0.485)	3807 (0.476)	4376 (0.547)	new
		LPP	4141 (0.517)	4330 (0.541)	3724 (0.465)	new
		NSD1	4388 (0.492)	502 (0.056)	4658 (0.522)	known
		NUP214	4962 (0.620)	5311 (0.664)	4933 (0.616)	new
		ARHGEF12	6202 (0.775)	5714 (0.714)	5841 (0.730)	new
202300	ADRENOCORTICAL CARCINOMA	TP53	1 (0.000)	1249 (0.140)	61 (0.007)	known
300299	NEUTROPENIA	WAS	1 (0.000)	1656 (0.186)	30 (0.003)	known
601367	STROKE, ISCHEMIC	PRKCH	14 (0.002)	449 (0.056)	448 (0.056)	new
		ALOX5AP	154 (0.017)	7892 (0.885)	7892 (0.885)	known
		NOS3	1243 (0.155)	4096 (0.512)	866 (0.108)	new
		F5	2111 (0.237)	4073 (0.457)	2222 (0.249)	known
188470	THYROID CARCINOMA	PTEN	20 (0.002)	710 (0.089)	710 (0.089)	new
		HRAS	60 (0.007)	1182 (0.133)	1182 (0.133)	known
		NRAS	719 (0.081)	2047 (0.230)	2047 (0.230)	new
		MINPP1	2100 (0.262)	4508 (0.563)	2720 (0.340)	new
604131	THALASSEMIAS	HBB	24 (0.003)	2025 (0.227)	2025 (0.227)	known
		HBA2	40 (0.005)	445 (0.056)	91 (0.011)	new
177900	PSORIASIS SUSCEPTIBILITY 1	HLA-C	50 (0.006)	2368 (0.296)	2368 (0.296)	new
		IL23R	397 (0.045)	2297 (0.258)	601 (0.067)	known
273300	TESTICULAR TUMORS	STK11	78 (0.009)	819 (0.092)	633 (0.071)	known
		KIT	88 (0.011)	3437 (0.429)	652 (0.081)	new

Future Works

- Integrate other types of genomic and clinical data into our current model to improve the performance.