

Densely-connected Biclustering for Mining Interaction Networks with Expression Data

Martin Ester

CMPT 884 Spring 2018

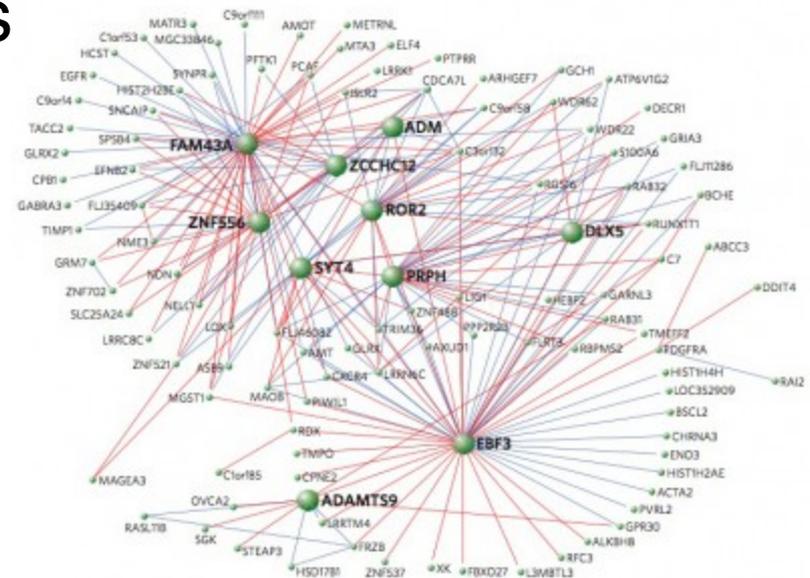
- Introduction
- Related work
- Densely-connected biclustering
[Moser et al, SDM 2009]
- Discovery of functional modules
[Colak et al, PLoS ONE 2010]
- Inference of cancer subnetwork markers
[Dao et al, Bioinformatics 2010]
- Conclusion

Outline

- Biological networks: molecules and their interactions and attributes
- Nodes: proteins, protein/protein complexes, genes, mRNAs, . . .
- Attributes: properties of molecules such as expression values
- Edges: interactions / reactions between corresponding molecules

Introduction

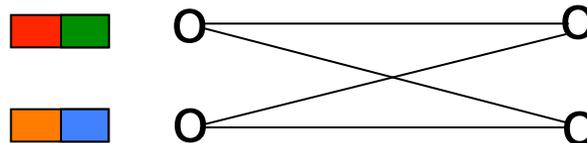
- Protein-protein interaction (PPI) networks interactions of protein molecules, which are important for many biological functions
- Gene regulatory networks / Transcriptional regulatory networks promotion / repression interactions between genes, transcription factors (TFs)



Introduction

- High-throughput data to infer networks
- Attributes such as gene expression (microarray, SAGE)
- Interactions such as regulatory protein-gene interactions (ChIP)
- Interactions are often condition-specific

if some genes are regulated by the same transcription factors, they are expressed at similar levels under the same conditions



Introduction

Advanced biclustering methods

- Order preserving submatrices [Ben-Dor et al 2002]
 - considers relative expression values
 - subset of genes (rows) whose expression values induce an order across a subset of conditions (columns)
- Significant biclusters (Samba) [Tanay et al 2002]
 - subset of genes that jointly respond across a subset of conditions
 - gene responds if its expression level changes significantly

Related Work

Finding dense components

- MCL (Markov Clustering) [Enright et al., 2002] simulates flow in a network by computing successive powers of the adjacency matrix
- Quasi-cliques [Pei et al 2005] subnetworks of nodes that all have more than a specified percentage of all possible out-edges within the subnetwork

Related Work

Motivation for integrated methods

- High-throughput data of either type is noisy and/or incomplete
e.g. high-throughput PPI data can contain up to 50% false positives
- Single data types provide only partial information on the underlying biological system
e.g. interaction network is static picture of the interactome and cannot yield insights into the dynamic behavior of cellular systems

Related Work

Integrated methods

- Co-clustering [Hanisch et al. 2002]
distance function combining expression and network distance, apply standard (distance-based) clustering algorithm
→ does not guarantee connectivity, clusters partitioning
- Matisse [Ulitsky & Shamir 2007]
test the hypothesis of a group of genes being co-expressed, find connected subnetworks which pass the test on a high significance level
→ does not guarantee density, clusters partitioning

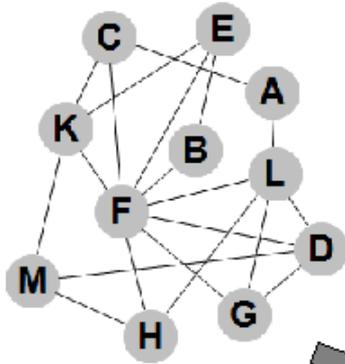
Related Work

- Given a network (graph) with d -dimensional (node) attributes.
 - Want to find subnetworks that
 - are dense,
 - connected,
 - have similar attribute values in a subspace.
- can overlap

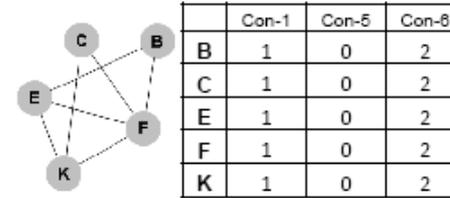
Densely-connected Biclustering

- *Densely-connected bicluster (DCB)*:
subgraph G' satisfying three conditions:
 - (1) subspace *homogeneity*, i.e. attribute values are within a range of at most ω in at least δ dimensions,
 - (2) *density*, i.e. has at least α of all possible edges, and
 - (3) *connectedness*, i.e. each pair of nodes has a connecting path in G' .
- Task
Find all *maximal* DCBs.

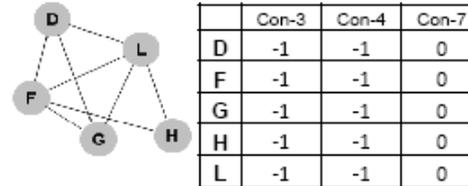
Densely-connected Biclustering



density = 7/10



$\alpha = 0.7$
 $\delta = 3$
 $\omega = 0.0$



density = 8/10

	Con-1	Con-2	Con-3	Con-4	Con-5	Con-6	Con-7
A	2	-2	1	0	2	1	2
B	1	1	0	-2	0	2	1
C	1	0	0	1	0	2	1
D	0	0	-1	-1	2	1	0
E	1	1	2	0	0	2	2
F	1	1	-1	-1	0	2	0
G	0	2	-1	-1	2	0	0
H	1	1	-1	-1	1	1	0
K	1	0	2	0	0	2	1
L	2	2	-1	-1	-1	-1	0
M	2	0	1	1	1	-2	-1

Densely-connected Biclustering

- DCB problem is NP-hard
its decision version can be reduced from the Max-Clique problem.
- Search space of all potential DCBs (subgraphs) is a lattice:
 - a subnetwork is a *child* of another one if it can be obtained by adding exactly one node and the corresponding edges to the parent network
 - *bottom* of lattice: the complete network
 - top* of lattice: pairs of nodes

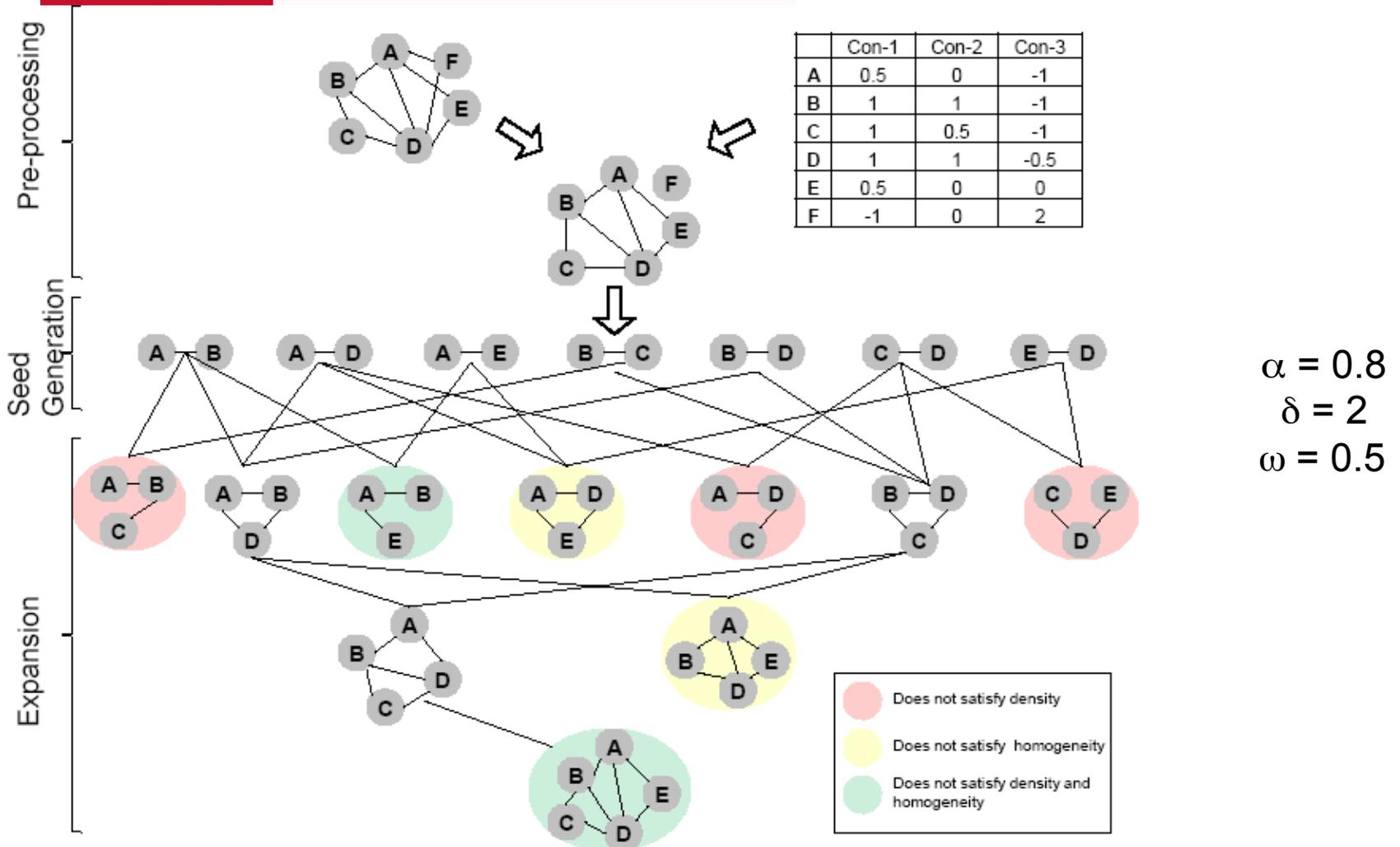
Densely-connected Biclustering

- A constraint is *loose anti-monotone*:
if for each network G of size n that satisfies the constraint,
there is *at least one* induced subnetwork G' of G of size $n - 1$ satisfying the constraint.
- For $\alpha \geq 0.5$, the DCB constraints are loose anti-monotone.
- level-wise (breadth-first) search of the lattice structure
in a top-down manner
→ construct only connected subgraphs
→ prune all candidates that do not satisfy the constraints
of density and homogeneity

Densely-connected Biclustering

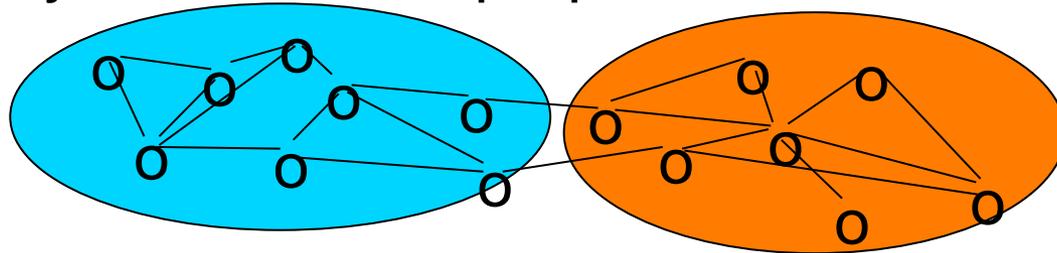
- Pre-processing
remove edges between nodes that cannot belong to same DCB
 - Seed generation
DCBs of size 2
 - Expansion
level-wise search of the lattice structure
- Algorithm is efficient
as long as the network is not too dense

Densely-connected Biclustering



Densely-connected Biclustering

- Modularity is fundamental design principle in natural and man-made systems
- Modules are formed according to common physical, regulatory, or functional properties



- Many interactions within modules,
 few interactions between different modules
→ Easy debugging and replacement of modules

Discovery of Functional Modules

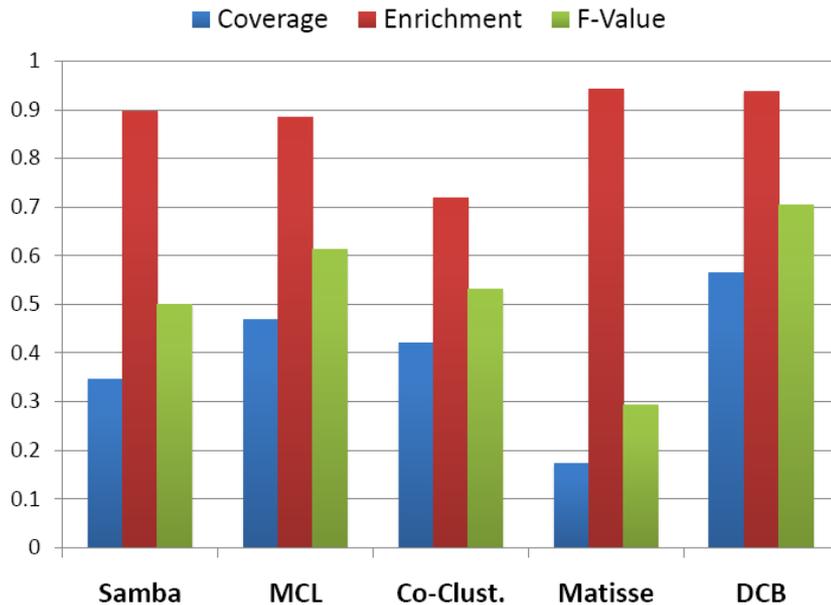
- Functional modules in PPI network
 - static molecular complex (e.g. ribosome)
 - or a dynamic signaling pathway
 - e.g. MAP kinase cascade
- Functional modules in gene regulatory networks
 - regulatory modules, in which every gene is controlled by the same TFs under the same environmental conditions
- Here: discovery of modules in PPI networks using DCB

Discovery of Functional Modules

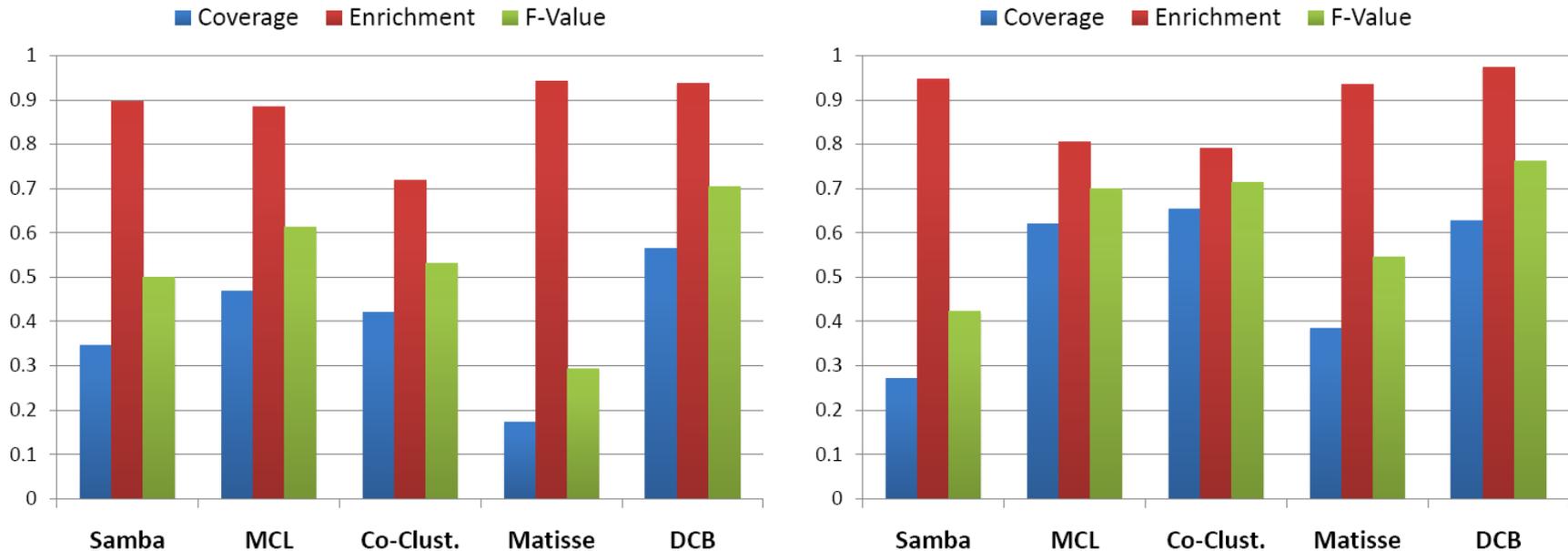
- Yeast dataset
 - PPI network from BioGRID database
 - gene expression data from yeast compendium dataset
- Human dataset
 - PPI network from BioGRID database
 - human tissue expression dataset
- GO-based evaluation
 - enrichment*: percentage of modules that are enriched with at least one GO term
 - coverage*: percentage of GO-terms that are enriched in at least one module
 - F-value*: harmonic mean of enrichment and coverage

Discovery of Functional Modules

Yeast



Human

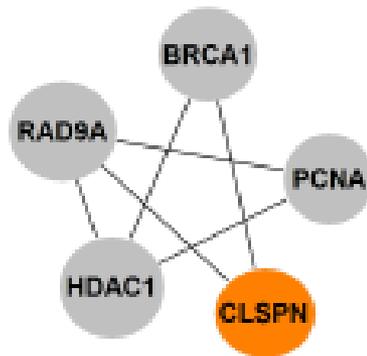


→ DCB consistently outperforms the comparison partners

Discovery of Functional Modules

Annotation Prediction

- Using annotations of neighboring genes in an enriched module → based on the guilt-by association principle
- Examine genes that do not have detailed GO term annotations but have some related work in the literature



Annotations predicted for **Claspin**:

DNA replication checkpoint
regulation of DNA replication initiation
regulation of DNA repair
DNA damage checkpoint

Discovery of Functional Modules

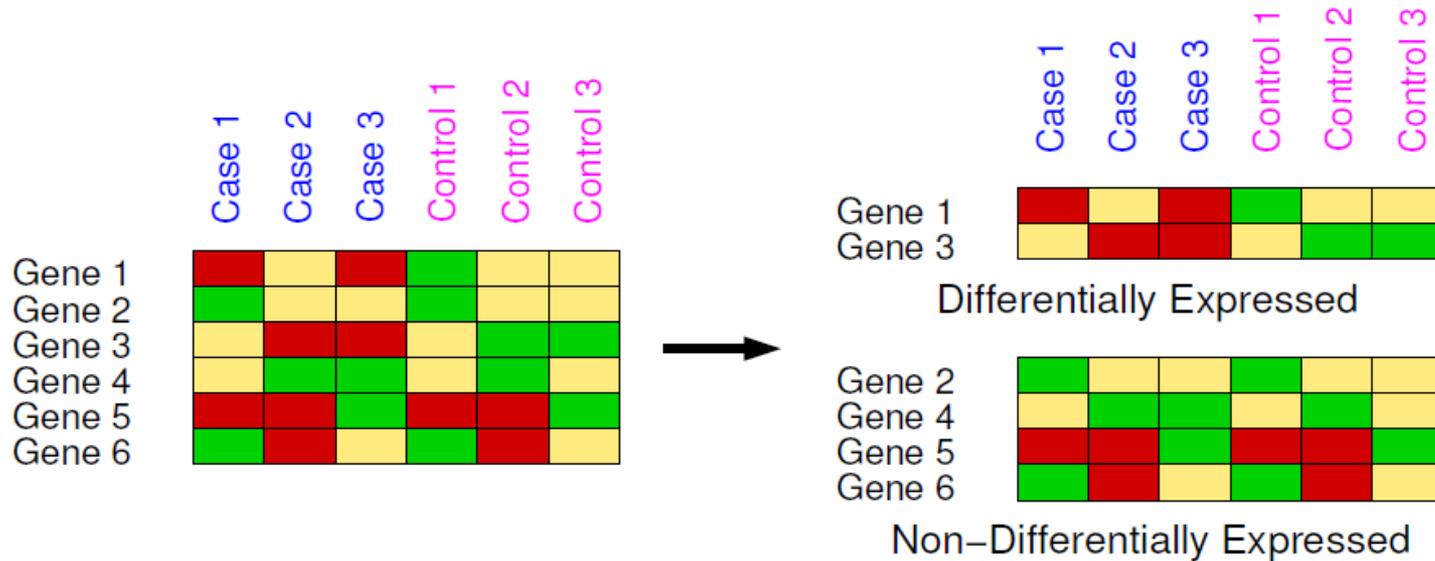
Biomarker Discovery

Biomarker: an indicator of a biological state

- Applications:
diagnosis of cancer, prognosis of metastasis
- Variables: gene expression, mutations
- **Single gene markers:** one gene per marker
- **Multigenic markers:** a set of gene(s) per marker

Inference of Cancer Subnetwork Markers

Single gene markers

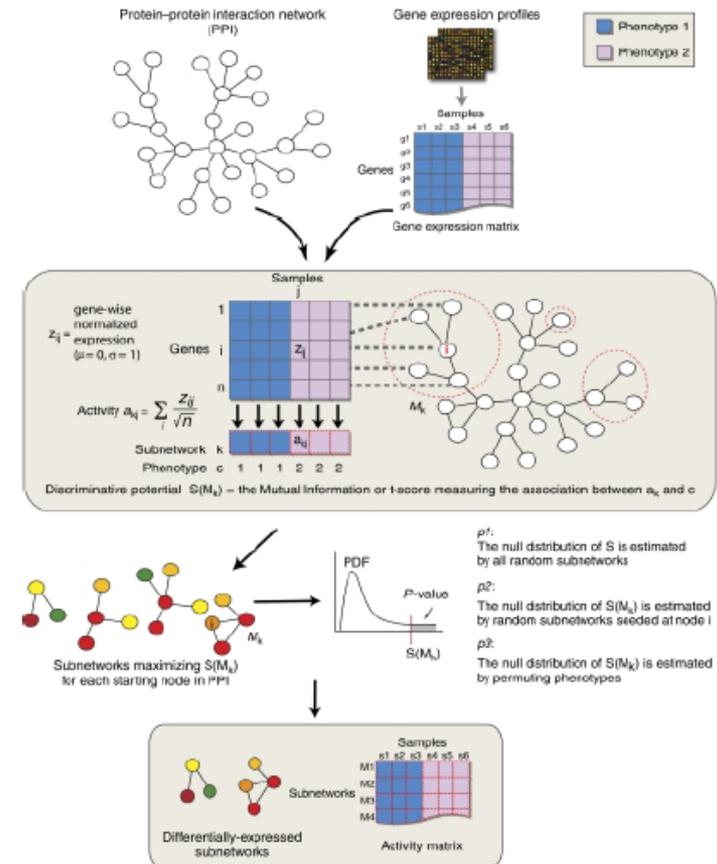


Inference of Cancer Subnetwork Markers

Multigenic markers

[Chuang et al., Mol.Sys.Biol. (2007)]:

- Markers are connected subgraphs with aggregate expression profiles which correlate the most with the labels of the samples
- Greedy heuristics for searching for optimal subnetwork markers
- Predicting progression of breast cancer



Inference of Cancer Subnetwork Markers

Multigenic markers

[Chowdhury et al., PSB 2010]:

- Each marker is a small connected subnetwork S such that genes in S **cover** all disease samples (gene g **covers** sample s if g is differentially expressed in S)
- Greedy heuristics for searching for the smallest subnetwork markers
- Predicting progression of colon cancer

Inference of Cancer Subnetwork Markers

Motivations

- Subnetwork markers from previous works are not functionally enriched
- [Goh et al., PNAS (2007)]:
 - The protein products of genes related to the same disease tend to interact with one another.
 - Genes related to a disease have coherent functions with respect to the Gene Ontology hierarchy.

Inference of Cancer Subnetwork Markers

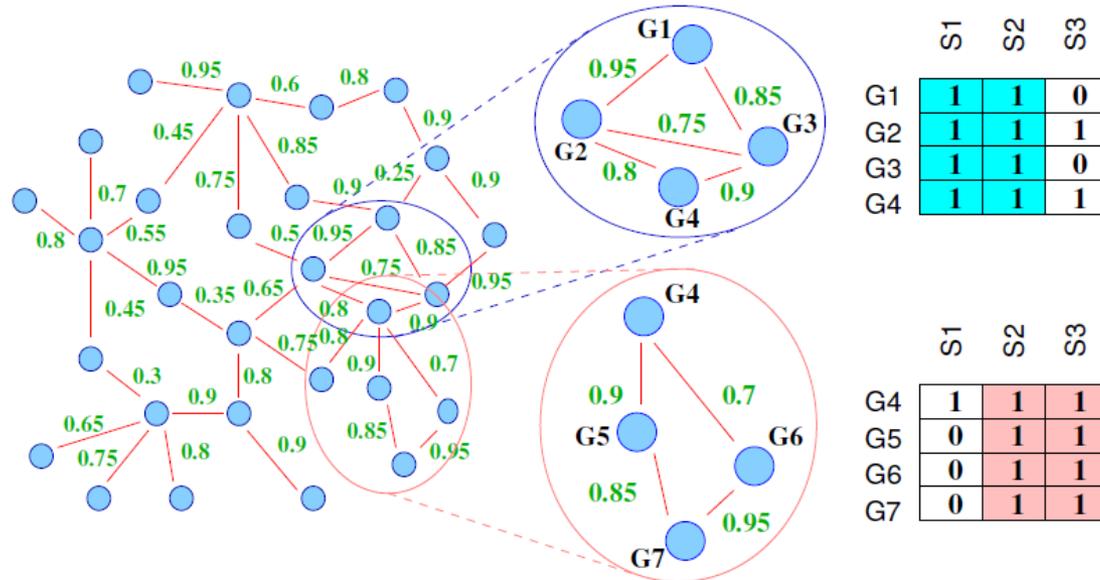
Our Approach

Our subnetwork markers:

- Subnetwork of genes that have more interactions among themselves than expected (**densely connected subnetworks**).
- Genes are differentially expressed in a fraction of all the samples (**partially differential expression**).

Inference of Cancer Subnetwork Markers

Partially Differential Expression

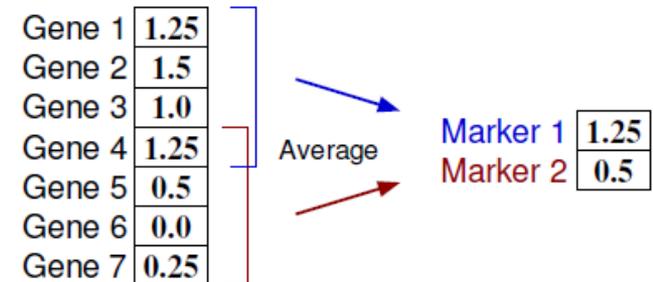
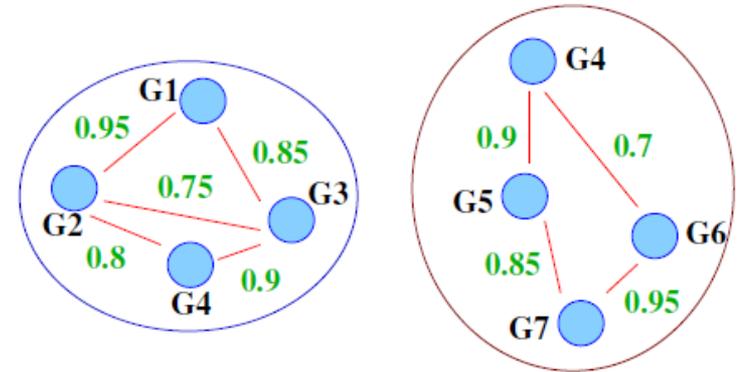


Compute all densely connected subnetworks whose genes are differentially expressed in a subset of patients of size at least k (here: $k = 2$).

Inference of Cancer Subnetwork Markers

Classifier Construction

1. Rank densely-connected biclusters according to density significance
2. Keep only highly-ranked subnetworks (DCBs) with little overlap
3. Feature space dimension = number of markers (DCBs)
4. SVM classification



Gene Expression Profile

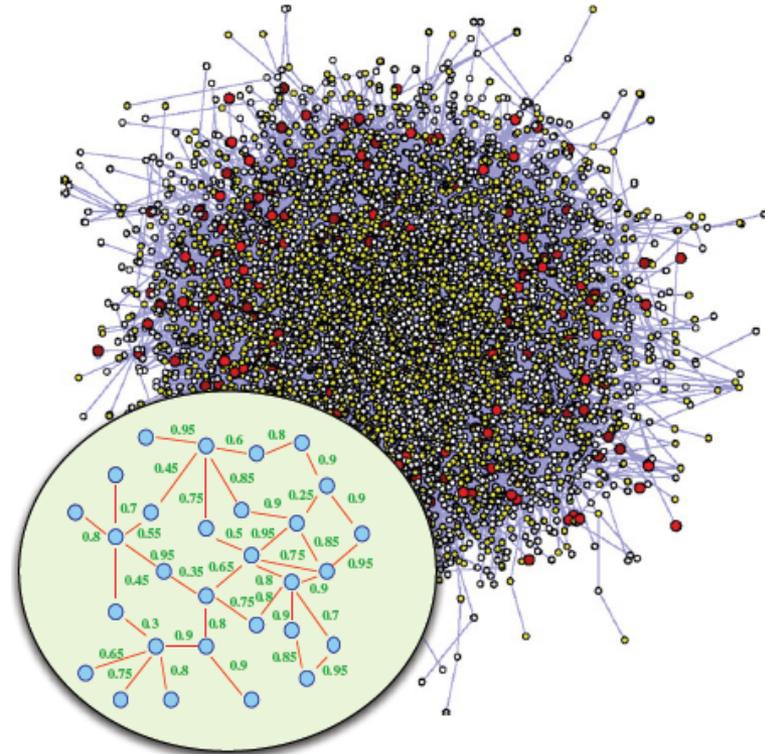
Average Gene Expression Profile

Inference of Cancer Subnetwork Markers

Confidence-scored PPI network

[STRING, von Mering et al., NAR 2009]:

- Edges reflect physical protein-protein interactions
- Confidence scores reflect the probability that the interaction is associated with a cellular phenomenon (and not an experimental artifact)
- Scoring system based on KEGG pathways



Inference of Cancer Subnetwork Markers

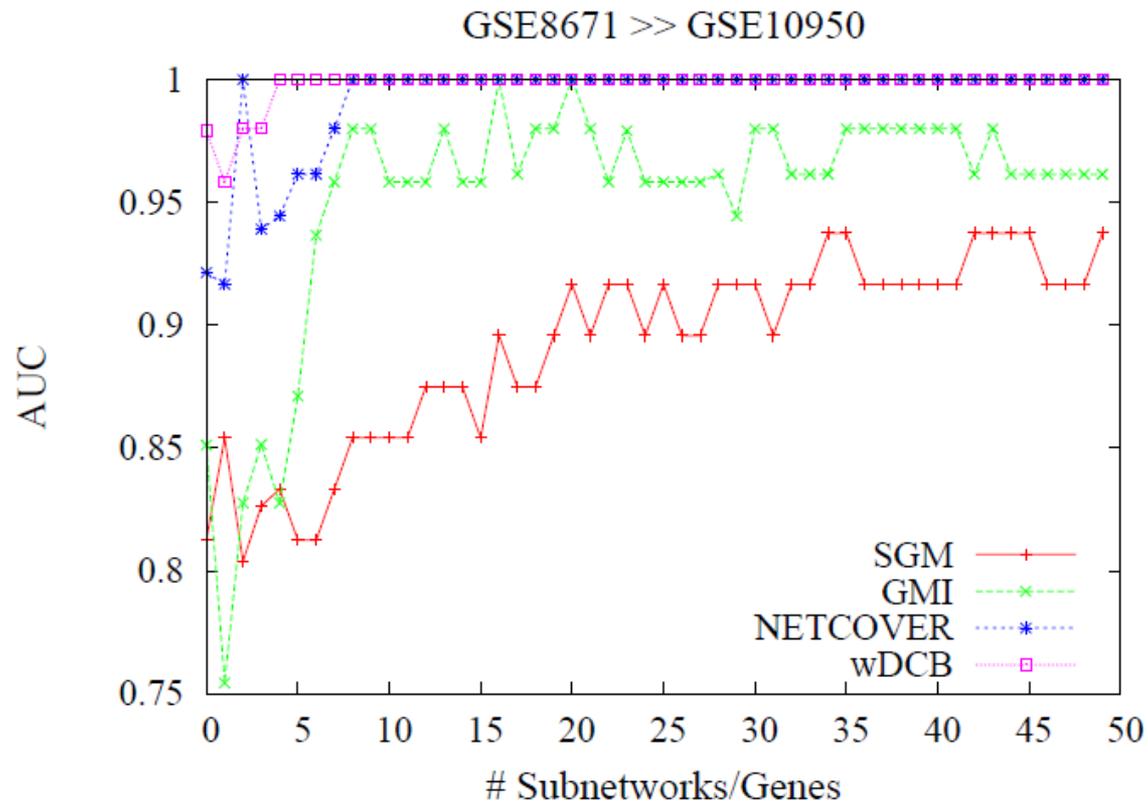
Gene Expression Datasets

Three experiments on colon cancer:

- GSE8671, 32 patients / tissue pairs
- GSE10950, 24 patients / tissue pairs
- GSE6988, 123 samples across several cancer subtypes

Inference of Cancer Subnetwork Markers

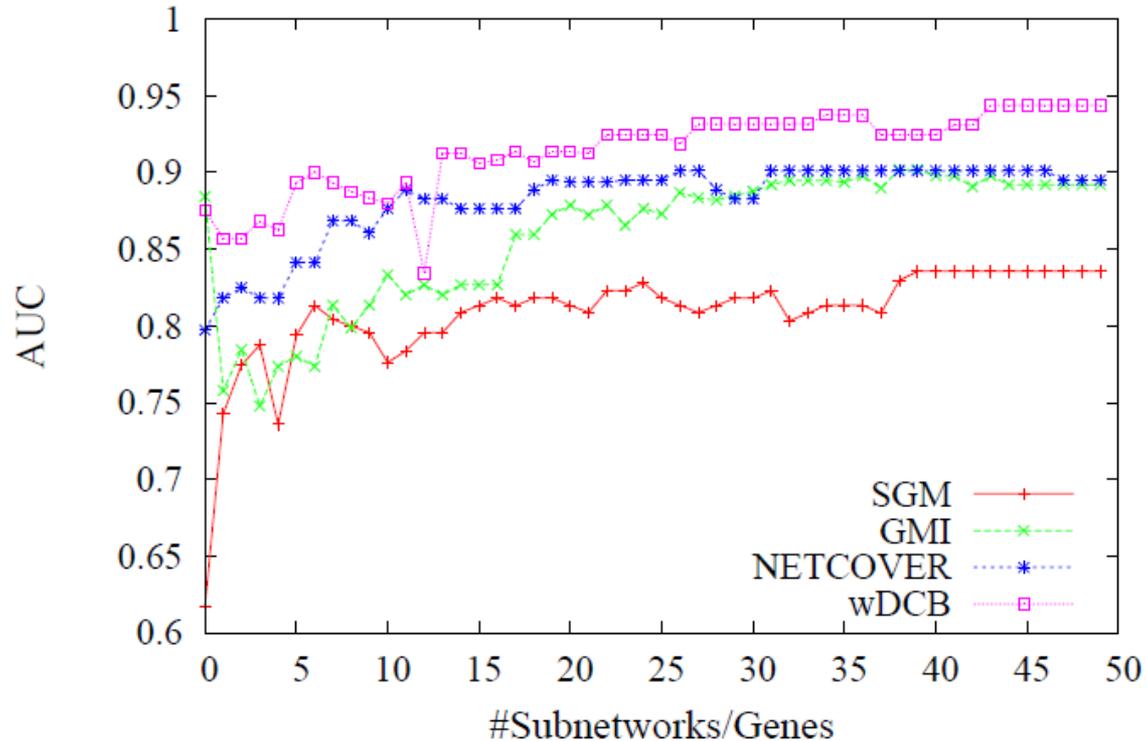
Classification Performance



Inference of Cancer Subnetwork Markers

Classification Performance

GSE8671 >> GSE6988



Inference of Cancer Subnetwork Markers

Subnetwork Marker Statistics

	#	Avg AUC			#	Avg AUC		
		ER-50	6988	10950		ER-50	6988	8671
GMI	806	0.38	0.86	0.95	755	0.34	0.84	0.99
NC	923	0.12	0.87	0.99	N/A	N/A	0.86	N/A
wDCB	282	0.76	0.91	1.00	216	0.74	0.91	1.00
8671 Subnetworks					10950 Subnetworks			

GMI = Greedy Mutual Information (Chuang et al.)

NC = NetCover (Chowdhury et al.)

wDCB = weighted Density Constrained Biclustering

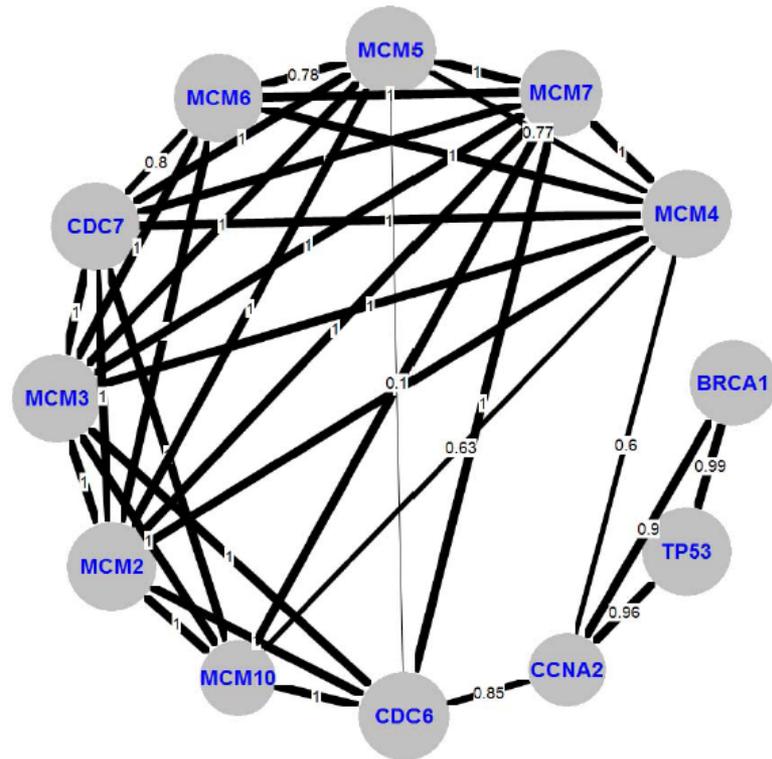
= total number of subnetworks computed

ER-50 = enrichment rate of the top-50 markers

Inference of Cancer Subnetwork Markers

Top Marker from GSE8671

- DNA replication initiation
- DNA metabolic process
- TP53, BRCA1: tumor suppressor genes
- Minichromosome maintenance complex (MCM complex)



Inference of Cancer Subnetwork Markers

Future Work

- Cancer markers that are more robust to different cohorts of the same cancer
- Compare subnetwork signatures of different cancers or subtypes of a particular cancer
- Extend the interaction network with for example ncRNA-protein interactions
- Novel methods to work with real valued continuous phenotype variables

- Ben-Dor et al (2002): “Discovering local structure in gene expression data: the order-preserving submatrix problem”, Proc. RECOMB 2002.
- Chowdhury S, Koyutürk M: “Identification of coordinately dysregulated subnetworks in complex phenotypes”, Pac. Symp. Biocomput. 2010;15:133-144
- Chuang et al: “Network-based classification of breast cancer metastasis”, Mol Syst Biol. 3:140 (2007).
- Colak, R. et al (2010): “Inferring the active modulome: an approach based onDensely Connected Biclustering”, PLoS ONE 5(10): e13348.
- Dao, P. et al (2010): “Inferring cancer subnetwork markers using density-constrained biclustering”, Bioinformatics 26(18).
- Goh, KI et al (2007): "The human disease network", Proc. Natl. Acad. Sci. U.S.A., 104, 21:8685-90.
- Enright, A. J. et al. (2002): “An efficient algorithm for large-scale detection of protein families”, Nucleic Acids Research 30, 1575-1584.
- Hanisch, D. et al. (2002): “Co-clustering of biological networks and gene expression Data”, Bioinformatics, 18, (Suppl. 1): 145-154.

References

- Moser et al (2009): “Mining Cohesive Patterns from Graphs with Feature Vectors”, Proc. SDM 2009.
- Pei, J. et al. (2005): “Mining Cross-Graph Quasi-Cliques in Gene Expression and Protein Interaction Data”, Proc. 21st IEEE Int. Conf. on Data Engineering.
- Tanay A., Sharan R., Shamir R. (2002): “Discovering statistically significant biclusters in gene expression data”, Bioinformatics, 18:Suppl1, S136-S144.
- Ulitsky, I. and Shamir, R. (2007): “Identification of functional modules using network topology and high-throughput data”, BMC Systems Biology, 1:8
- von Mering et al (2003): "STRING: a database of predicted functional associations between proteins“, Nucleic Acids Res 31 (1): 258–261.

References