Next Generation Bioinformatics Tools

Fall 2012

Day 3 – Systems Biology and Biological Networks

Hesham H. Ali
UNO Bioinformatics Research Group
College of Information Science and Technology

UNIVERSITY OF NEBRASKA AT OMAHA

UNO BIOINFORMATICS

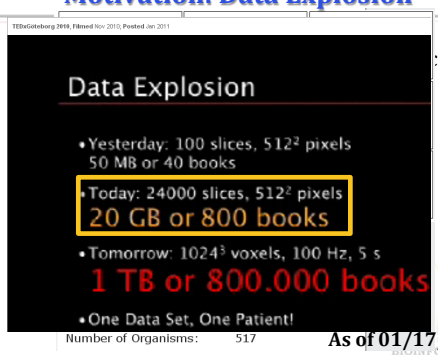
Overview

- Motivation
- Types of Biological Networks
  - Applications of Biological Networks
- Structural Network Concepts
  - Global Measures
  - Local Measures
  - Lethality and Enrichment
    - Assessment of Central Nodes
  - Annotation of Clusters
    - Traditional Node Enrichment
    - Edge Enrichment
  - Structural Summary
- Network Integration
  - Network Alignment
  - Data-based Integration
  - Knowledge-based Integration
- Case-studies

UNO BIOINFORMATICS

Motivation: Data Explosion

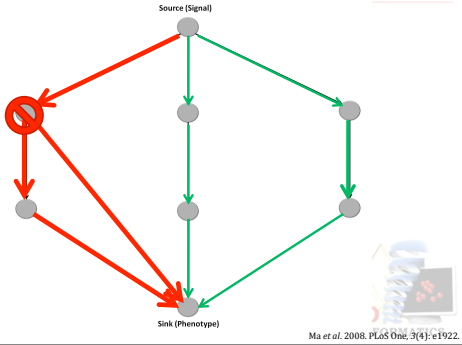
- 's'



Number of Organisms: 517

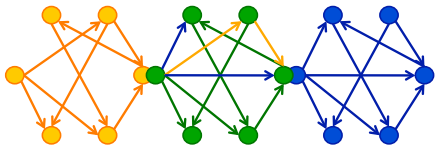
As of 01/17/2011

Motivation: Compensation



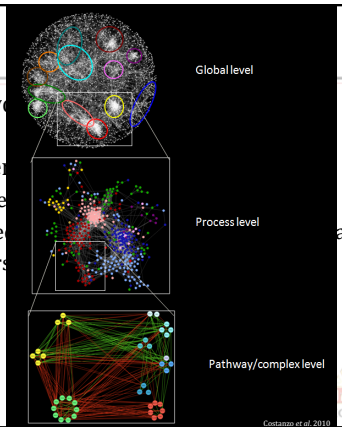
Ma *et al.* 2008. PLoS One. 3(4): e1922.

Motivation: Network Alignment



Normal **Diseased**

- Netw
- Syste
  - Inte
  - Use
  - Pers



Costanzo *et al.* 2010

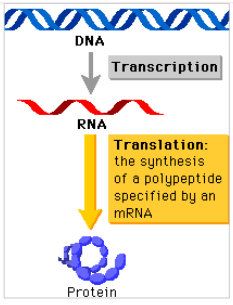
Systems Biology Approach

- Realistic and Innovative:
  - Networks model relationships, not just elements
  - Discover groups of relationships between genes and gene products
- Discovery Aspect
  - Examine changes in systems
    - Normal vs. diseased
    - Young vs. old
    - Stage I v. State II v. Stage III v. Stave IV

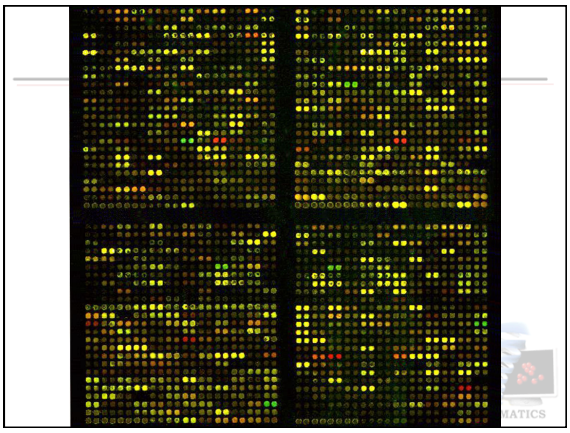
Motivation

- How does the network allow us to achieve these ICD goals?
  - Layers of information
  - Integration of different types of knowledge
  - High performance computing
    - Key to analysis of large, complex sets of data with multiple layers

Central Dogma

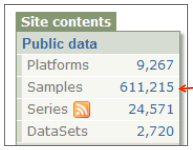


- DNA → RNA → Protein
  - Occurs millions of times within a cell
  - Results in different cell/ tissue types
  - *Measurable relationships*
- High-throughput analyses: measuring thousands of relationships in one experiment



Why Networks?

- Explosion of biological data



Each sample can have over 40,000 genes

- Average microarray experiment: 1200 pages of data*
- How can we extract information from data?

* http://www.ncbi.nlm.nih.gov/geo/info/GEO2.html

Systems Biology

- Holist view of the system
  - Ability to zoom in/out to view critical system components
- Past: Reductionist biology
  - Find a gene/protein of interest
  - Examine under different conditions
- Systems biology: examine an entire system at different conditions

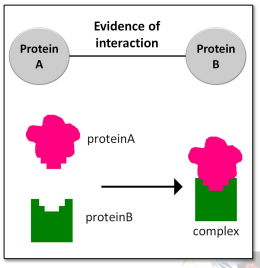
Types of Biological Networks

Biological Networks

- A biological network represents elements and their interactions
- Nodes → elements
- Edges → interactions
- Can represent multiple types of elements and interactions

Types of Biological Networks

- **Protein-protein interaction network**
- Synthetic lethality
- Metabolome
- Signal transduction
- Correlation/co-expression network



Protein-Protein Interaction Networks

- Built directly from Y2H, Co-IP, TAP
  - Physical detection of interactions
- Databases house PPI data
  - Pathway commons (warehouse)
  - BioGrid
  - HumanCyc
  - ……

Pathway Commons Quick Stats:

Number of Pathways:	1,623
Number of Interactions:	585,237
Number of Physical Entities:	105,949
Number of Organisms:	564

BIOINFORMATICS

PPI data

- Experimental PPI datasets:
  - Data sets are available for yeast as it has been studied extensively
  - Ito full data and Ito core data
  - Utez data
  - Gavin complexes
  - Ho Complexes

UNO BIOINFORMATICS

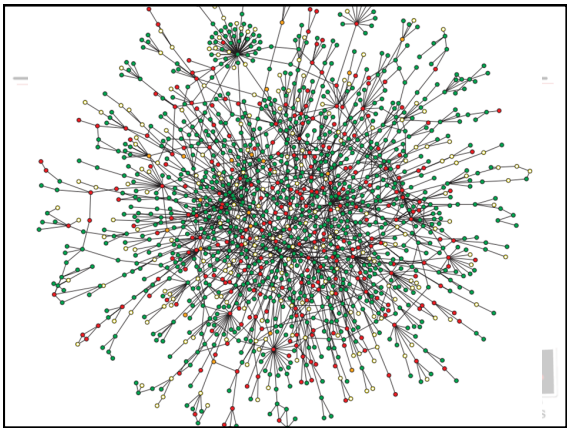
Public PPI databases

- MIPS
- DIP
- BIND
- BioGRID
- MINT
- IntAct
- HPRD

UNO BIOINFORMATICS

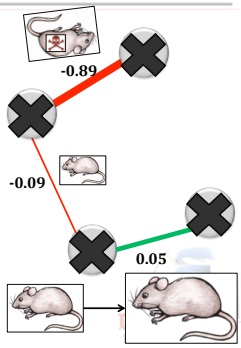
Protein-Protein Interaction Networks

- "Hub" proteins in biological networks began from the study of PPI's
- Study done by Jeong (2001)
  - 1870 proteins (nodes)
  - 2240 interactions (edges)
- Forms a scale-free network (*incomplete*)

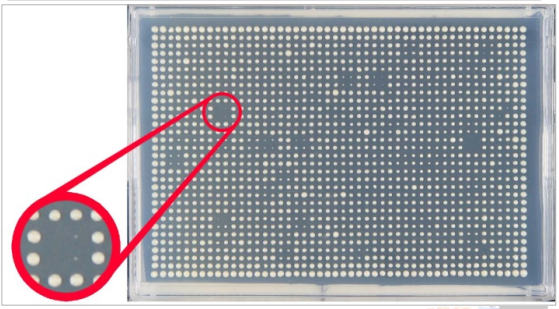


Types of Biological Networks

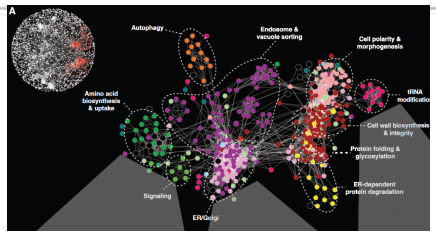
- Protein-protein interaction network
- **Synthetic lethality**
- Metabolome
- Signal transduction
- Correlation/co-expression network



Genetic Interaction Networks



Genetic Interaction Networks



- Typically used in *S. cerevisiae*
  - Entire library of query genes available
- Methods being developed for *E. coli, S. pombe*
- Identification of relationships that are ***beneficial and detrimental***

Genetic Interaction Networks Applications

BRCA1 + PARP1 → synthetic lethal interactors

Tumor cells only

Normal

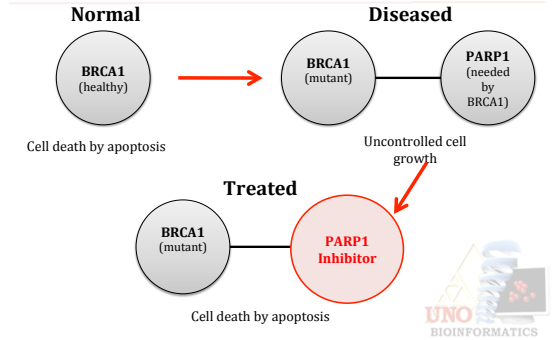
BRCA1 (healthy)

Diseased

BRCA1 — **PARP1**

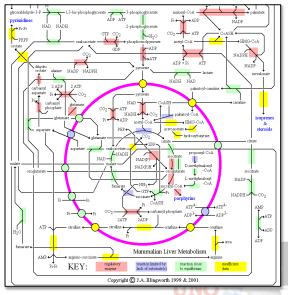
PARP1 only expressed in tumor cells

Genetic Interaction Networks Applications



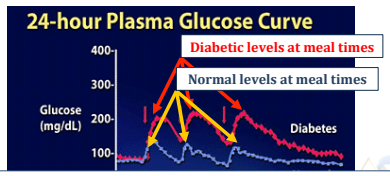
Types of Biological Networks

- Protein-protein interaction network
- Synthetic lethality
- **Metabolome**
- Signal transduction
- Correlation/co-expression network



Metabolome

Goal: Provide a model that can handle a *very dynamic system* – i.e., metabolism



This measures just ONE metabolite in the human body at a time over the period of one day - - over 2,500 known metabolites can be found in the body

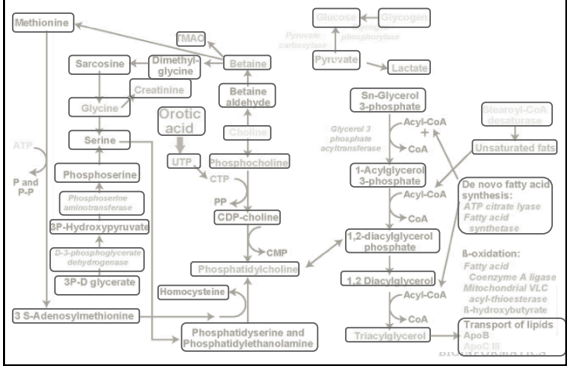
Metabolome

- *Dynamic network*- constantly changing
  - Which metabolites are present?
- Node: Compound/chemical/metabolite
- Edge: Reaction metabolizing node
  - Edges are directed
- OR
- Node: Reaction (Process)
- Edges: Metabolites required for that process

•Can we measure the change in metabolites?
•Can the model of metabolism reveal the mechanisms behind disease?

UNO BIOINFORMATICS

Metabolome

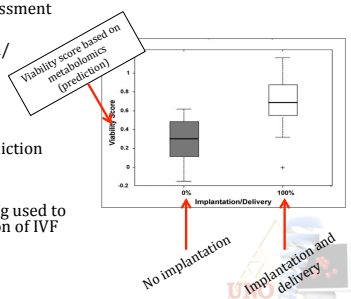


Summary of metabolomic databases

Database name	URL or web address	Comments
Human metabolome database	http://www.hmdb.ca	Largest and most complete of its kind. Specific to humans only
BioMagResBank (BMRB – metabolimics)	http://www.bmrb.wisc.edu/metabolomics/	Emphasis on NMR data, no biological or biochemical data Specific to plants (Arabadopsis)
BiGG (database of biochemical, genetic and genomic metabolic network reconstructions)	http://bigg.ucsd.edu/home.pl	Database of human, yeast and bacterial metabolites, pathways and reactions as well as SBML reconstructions for metabolic modeling
Fiehn metabolome database	http://fiehnlab.ucdavis.edu/compounds/	Tabular list of IDId metabolites with images, synonyms and KEGG links
Golm metabolome database	http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/gmd.html	Emphasis on MS or GC–MS data only No biological data Few data fields Specific to plants
METLIN metabolite database	http://metlin.scripps.edu/	Human specific Mixes drugs, drug metabolites together Name, structure, ID only
NIST spectral database	http://webbook.nist.gov/chemistry/	Spectral database only (NMR, MS, IR) No biological data, little chemical data Not limited to metabolites
Spectral database for organic compounds (SDBS)	http://www.aist.go.jp/RIODB/SDBS/cgi-bin/direct_frame_top.cgi?lang=eng	Spectral database only (NMR, MS, IR) No biological data, little chemical data Not limited to metabolites

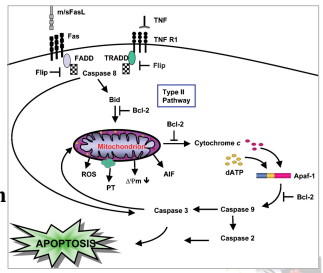
Metabolome Applications

- Pharmacological assessment
- Toxicology prediction/ assessment
- Nutrigenomics
- Drug interaction prediction
- Others (IVF)
  - Metabolic profiling used to predict implantation of IVF embryos



Types of Biological Networks

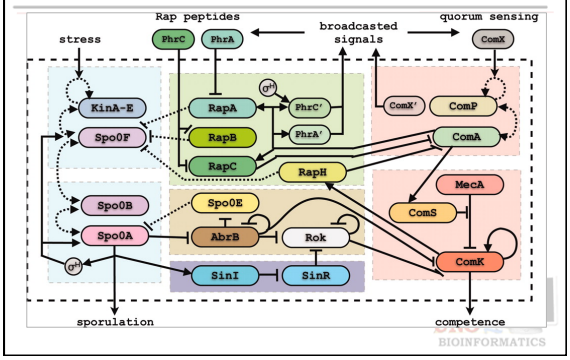
- Protein-protein interaction network
- Synthetic lethality
- Metabolome
- **Signal transduction**
- Correlation/co-expression network

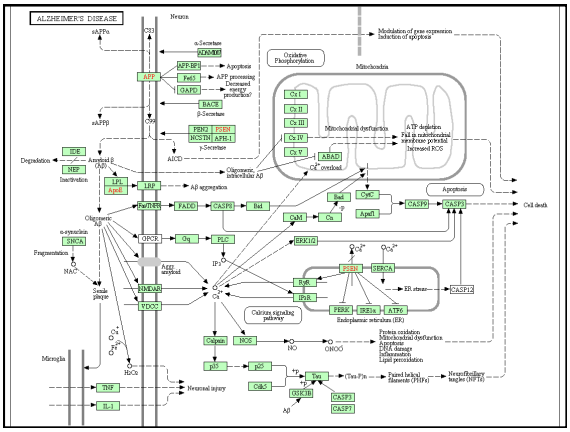


Signal Transduction Networks

- Map of cellular communication
  - Nodes: Proteins, RNAs
  - Edges: To-/from- relationships
    - Edges are *directed*
    - Represent communication, stimulation, or repression
- KEGG Pathway or developed by individual groups

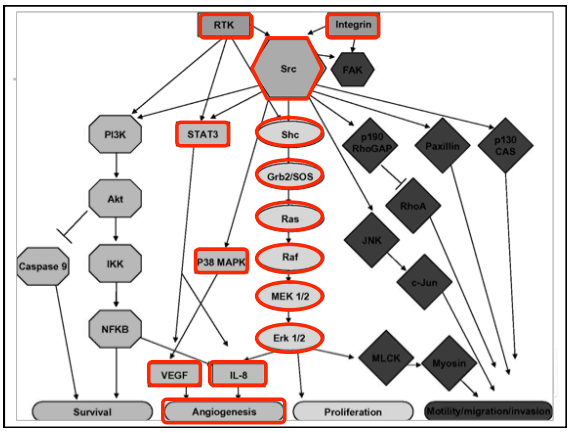
Signal Transduction Networks





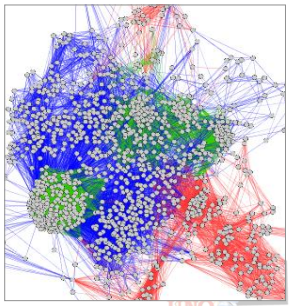
Signal Transduction Networks: Applications

- Identification of new pathways
  - Leads to identification of new target genes
- Identification of compensatory pathways
- General schema of cellular communication



Types of Biological Networks

- Protein-protein interaction network
- Synthetic lethality
- Metabolome
- Signal transduction
- **Correlation/co-expression network**



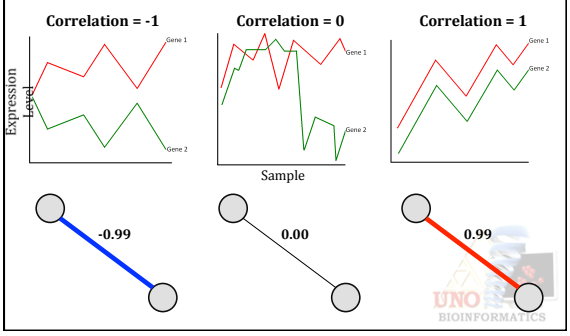
Correlation Networks

Gene ID	Sample1	Sample2	Sample3	Sample4	
A_52_P616356	5.9813	6.0079	5.9525	7.2753	6.2
A_52_P580582	7.7845	7.7512	8.0943	8.3608	8.1
A_52_P403405	5.9301	6.5153	6.0526	7.1707	6.1
A_52_P819156	6.5748	6.8645	6.981	7.4937	6.9
A_51_P331831	7.1732	7.8754	7.7632	8.1875	7.6
A_51_P430630	6.0661	6.4009	5.9525	7.1208	6.6
A_52_P502357	5.936	6.3206	5.9525	7.1819	6.2
A_52_P299964	6.3452	6.8025	6.6457	7.3445	6.7
A_51_P356389	6.5088	7.0545	7.2346	7.631	7.6
A_52_P684402	10.0915	10.7124	10.2245	10.377	10.

- 10,000-45,000+ probes
- UNO Blackforest cluster
- HCC Firefly

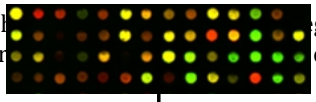
0 **Correlation = 1**

Correlation Network



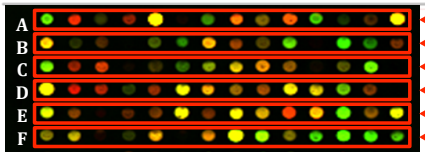
Correlation Networks

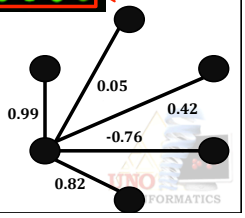
A graph[illegible] degree of corr[illegible] entity



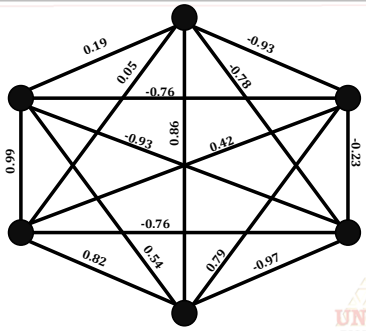
	Sample 1	Sample 2	Sample 3
Gene 1	10.5	11.0	12.1
Gene 2	3.2	3.3	2.9
Gene 3	1.4	1.5	0.9
Gene 4	7.8	7.1	8.2

Correlation Networks



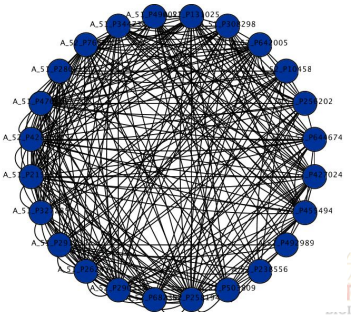


Correlation Networks



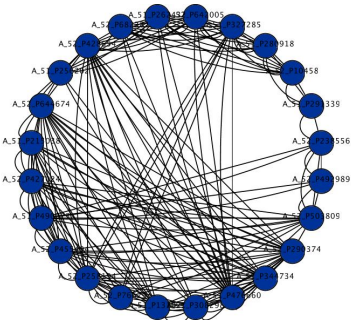
Correlation Networks

24 node sample
Threshold: 0.00-1.00



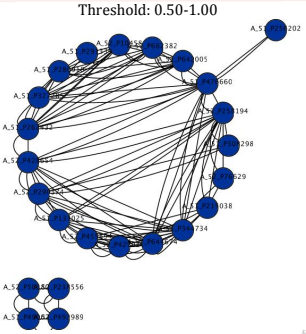
Correlation Networks

24 node sample
Threshold: 0.30-1.00



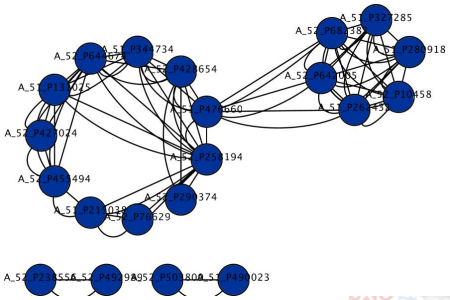
Correlation Networks

24 node sample
Threshold: 0.50-1.00



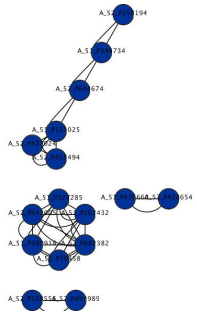
Correlation Networks

24 node sample
Threshold: 0.60-1.00



Correlation Networks

24 node sample
Threshold: 0.80-1.00



Correlation Network Applications

- "Versus" analysis
  - Normal vs. disease
  - Times/environments
- Model for high-throughput data
  - Especially useful in microarrays
- Identification of groups of causative genes
  - Ability to rank based on graph structure
  - Identify sets of co-regulated, co-expressed genes

Network Concepts

- **Biological networks have structural properties**
  - Can differ from one network to another
- **Specific structures/characteristics have biological meaning**
  - Degree can indicate essentiality
  - Cluster density can indicate relevance
- **Networks do not have to be static**
  - Most interesting discoveries coming from temporal or state-change network alignment & comparison

Structures & their Functions

Network structures correspond to key cellular structures

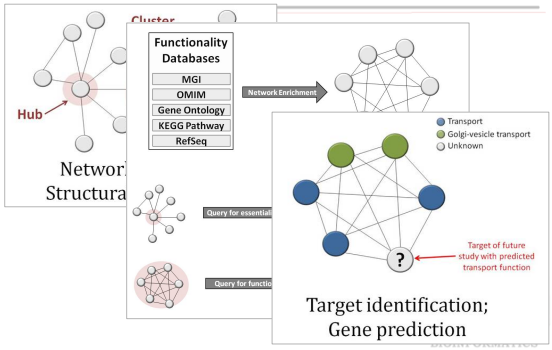
Hypothesis

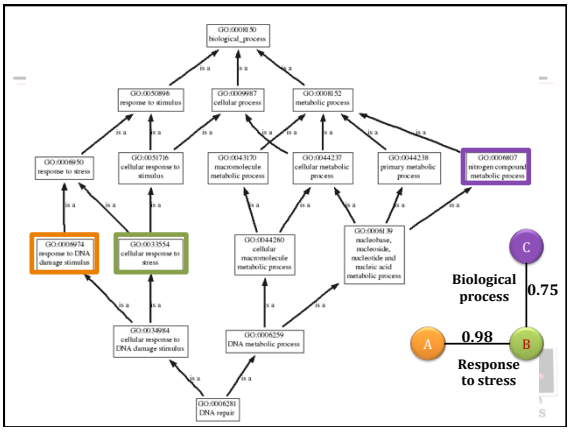
Correlation networks are an excellent tool for mining relationship rich knowledge from high-throughput data

Using systems biology approach, CN can help identify:
- *Critical Genes* that are essential for survival
- *Subsets of genes* that are responsible for biological functions

Measures of centrality to identify key elements:
Proves existence of structure/fucntion relationship in correlation networks

Integrated Data Model

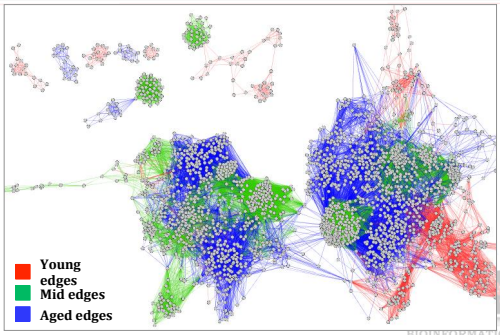


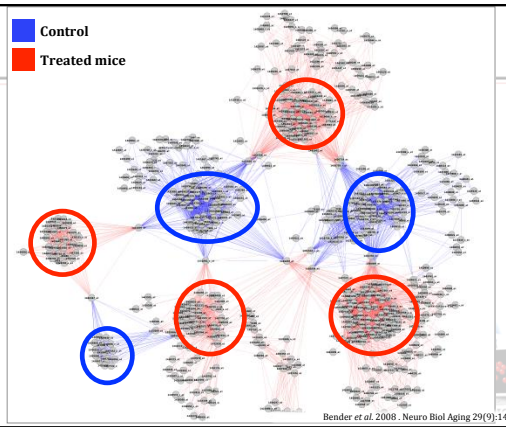


Objectives

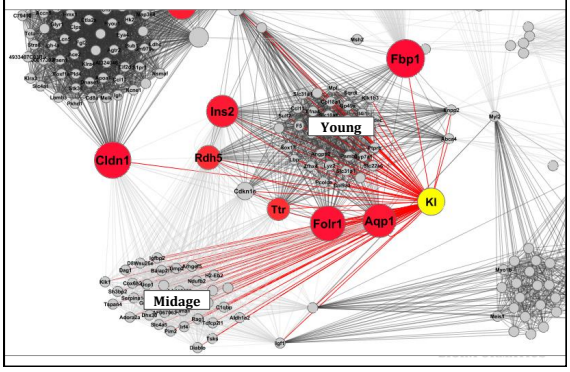
- Confirm structure/ function relationships in integrated biological networks
- Uncover genetic drivers of aging and disease using application of graph theory

Case Study: Aging

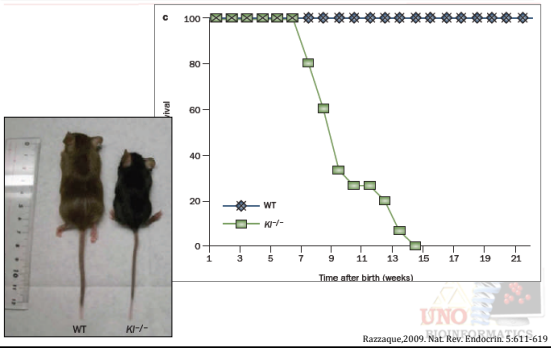




Critical Gene Identification



Validation



Razzaque,2009. Nat. Rev. Endocrin. 5:611-619

Structural Network Concepts

Basic Concepts

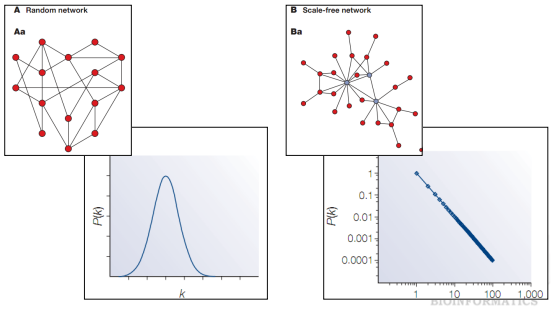
- **Biological networks have structural properties**
  - Can differ from one network to another
- **Specific structures/characteristics have biological meaning**
  - Degree can indicate essentiality
  - Cluster density can indicate relevance
- **Networks do not have to be static**
  - Most interesting discoveries coming from temporal or state-change network alignment & comparison

Key Characteristics of Networks

- **Small-world property**: all nodes can be reached quickly from any node via a few hops to its immediate neighbors
  - Average shortest-path length between any two nodes in a network is relatively small
- **Scale-free distribution**: have few high degree hub nodes, while most nodes will have only a few connections
  - Highly connected hub nodes are infrequently connected to each other in contrast to social networks where well-connected people tend to have direct connections to each other.

Scale-Free Networks

Degree (*k*): Number of edges connected to a node



Scale-Free Networks

- **Follows a power-law degree distribution**
  - High-degree nodes, known as hubs
  - Contains clusters:
    - Groups of highly connected nodes
- **Has "small world" characteristic**
  - Average path length ~ $\log_2 \log_2 N$
    - N = number of nodes
    - $\text{Log}_2 \log_2$ (6 billion) ~ 5.0215

Albert, R., 2005. Journal of Cell Science 118(21):4947-4957

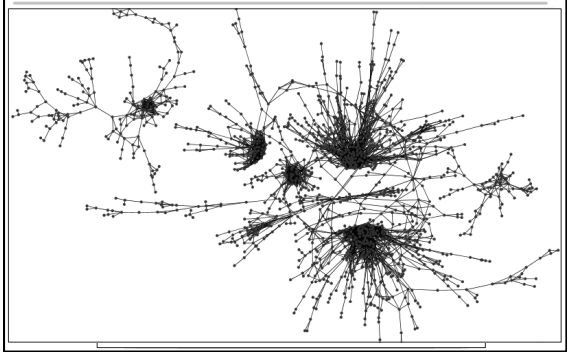
Modular Networks

- Module: is a region with *dense internal connections* and *sparse external interconnections* to other regions.
  - Modular nodes: high degree nodes at core of a module
  - Peripheral nodes : low degree, linked to modular nodes or to other peripheral nodes in same module
  - Interconnecting nodes: nodes connecting modules
    - Edge connecting two modules : BRIDGE
  - Existence of modular structure: presence of high average clustering coefficients

Modularity Analysis

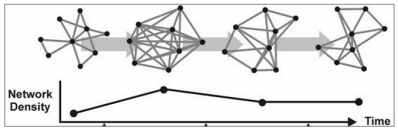
- Functional modules
  - Proteins that participate in a particular cellular process while binding to each other at various times and places.
  - Detection of these groupings → modularity analysis

Network Filtering

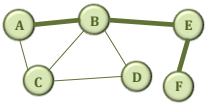


Global Network Structures

- **Density:** Edges in the network/total edges
  For network with *n* nodes, total number of edges: *n(n-1)/2*



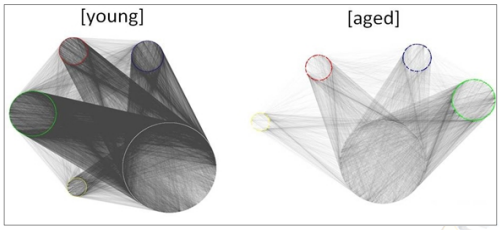
- **Radius:** Longest shortest path in the network



Longest shortest path:
A→ B→E→F

Diameter: 3

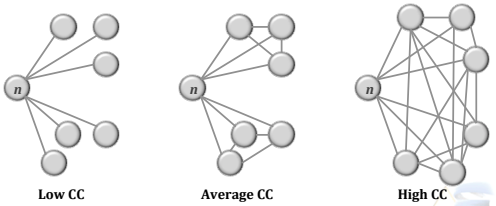
Global Network Structures



Network density changes over time for specific cellular functions

Global Network Structures

- **Transitivity/ clustering coefficient:** For any node *n*, transitivity is how connected *n*'s neighbors are



Low CC Average CC High CC

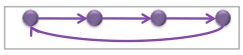
- Transitivity of the entire network is the *average* of clustering coefficient for all nodes in the network

Local Network Structures

Local Network Structures

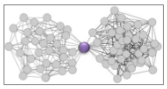
- **Cliques**
  Protein complexes, regulatory modules

- **Pathways**
  Signaling cascades

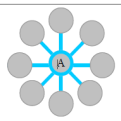


- **Hubs**
  Regulators, TFs, active proteins

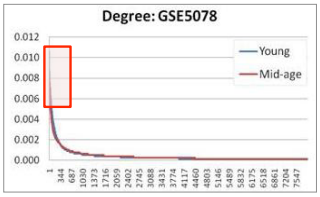
- **Articulation points**



Local Network Structures: Centrality



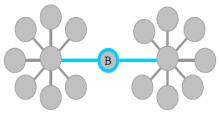
High Degree Nodes —
Node with more neighbors



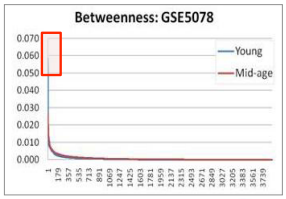
Basic Centralities

- **Degree centrality of a node** = degree of vertex
  - Indicator of importance of node
- **Distance-based centralities**
  - Importance of components based on distance between vertices in a graph
  - Shortest path is used to measure the topological properties of a graph component
- **Eccentricity:** of a vertex, *v*, is the greatest distance between *v* and any other vertex in a graph.
  - Eccentricity represent distance of a vertex from center of graph.
  - Vertices at center have zero eccentricity
- **Closeness:** reciprocal of total distance from a vertex v to all other vertices in a graph.
  - Mean-shortest path length from vertex v to all other vertices in a graph
  - Time needed for information to spread from a particular vertex to others in a network
  - Not applicable for disconnected networks

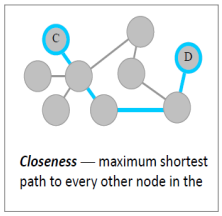
Local Network Structures: Centrality



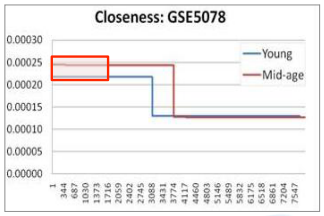
Betweenness—How often a node exists on the shortest path between nodes



Local Network Structures: Centrality

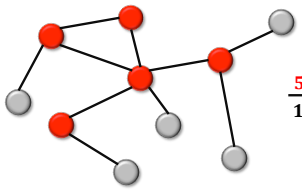


Closeness — maximum shortest path to every other node in the



Centrality Assessment: Lethality

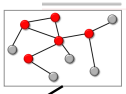
$$Lethality_T = \frac{\#\ lethal\ genes\ in\ T}{\#\ tested\ genes\ in\ T} \quad (1)$$

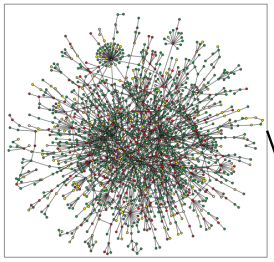


$$\frac{\text{5 lethal nodes}}{\text{10 nodes total}} = \text{50\% lethality}$$

Centrality Assessment: Enrichment

$$Enrichment = \frac{\#\ lethal\ genes\ in\ T\ /\ \#\ tested\ genes\ in\ T}{\#\ lethal\ genes\ in\ B\ /\ \#\ tested\ genes\ in\ B}$$





$$\frac{\text{5 lethal nodes}}{\text{10 nodes total}} = \text{50\% lethality}$$

Target

$$\frac{\text{200 lethal nodes}}{\text{1000 nodes total}} = \text{20\% lethality}$$

Background

Centrality Assessment

$$Enrichment = \frac{\#\ lethal\ genes\ in\ T\ /\ \#\ tested\ genes\ in\ T}{\#\ lethal\ genes\ in\ B\ /\ \#\ tested\ genes\ in\ B} \quad (2)$$

5 lethal nodes = **50% lethality**
10 nodes total
Target

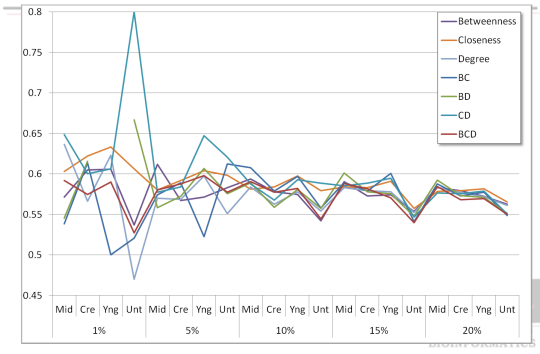
200 lethal nodes = **20% lethality**
1000 nodes total
Background

= 2.5 Enrichment

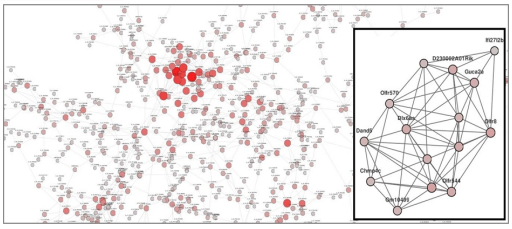
Centrality Assessment

- Enrichment < 1
  - Background enriched in lethal nodes (poor)
- Enrichment = 0
  - Background and target have equal rates of lethal nodes (decent)
- Enrichment > 1
  - Target enriched in lethal nodes (poor)

Lethality of Essential Nodes



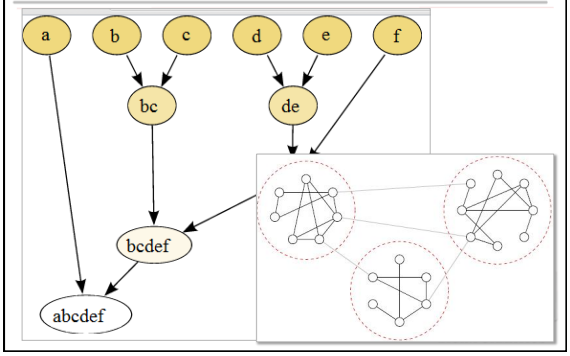
Local Network Structures: Clusters



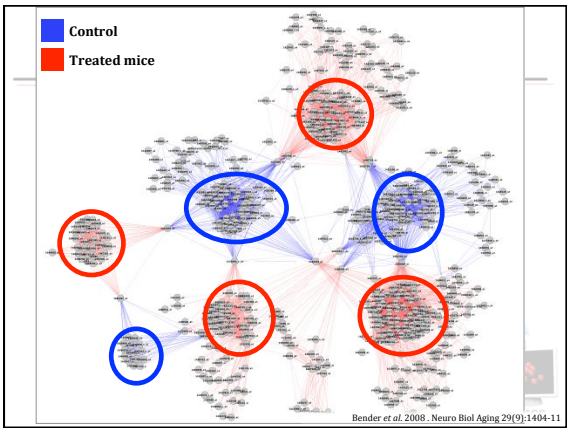
Over-represented GO functions in shown cluster:

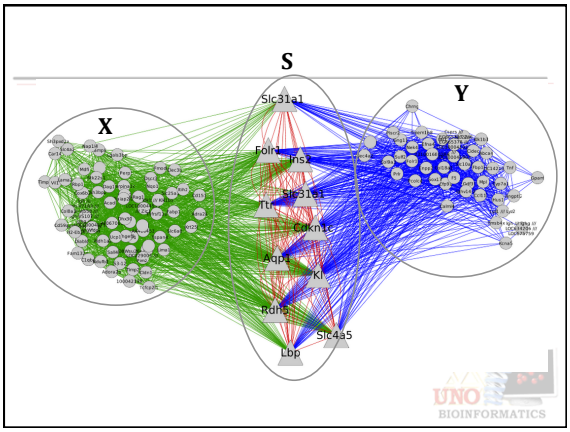
- cell surface receptor linked signaling pathway,
- sensory perception & cognition
- signal transmission & signaling process

Clustering



Gateway Nodes





Node Gatewayness

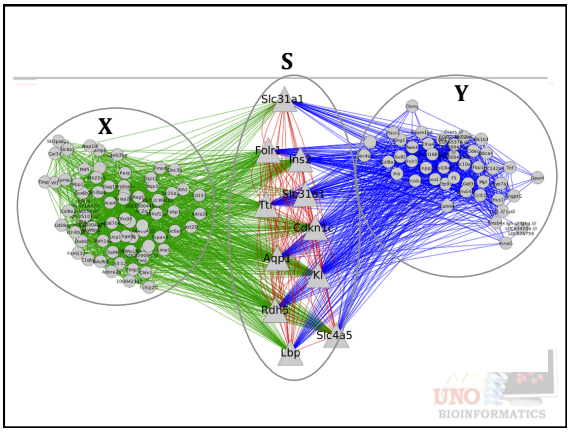
- Let undirected graphs *G1 = (V, E1)* and *G2 = (V, E2)* such that graphs *G1* and *G2* share same node set *V* with different edge sets *E1* and *E2*.
- For each graph we identify clusters (dense subgraphs) such that:
  - Cluster *X* represents some dense subgraph in *G1*
  - Cluster *Y* represents some dense sub-graph in *G2*
- Compute *G'* such that *G'* = (*V,(E1 ∪ E2)*)

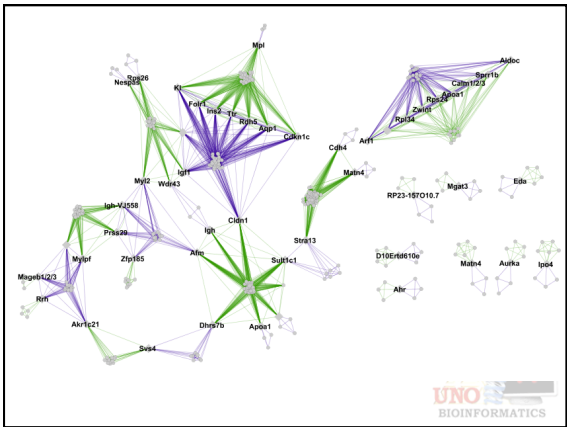
Node Gatewayness

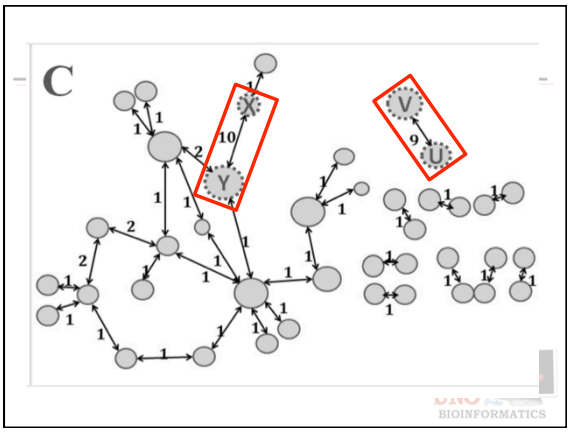
- Define subset of nodes $S = V(X) \cap V(Y)$
- For any node s in S, $E_{(s)}$ is the set of edges connecting s to any node in the set X from graph $G1$ and the set of edges connecting s to any node in the set Y from graph $G2$.
- Using these definitions we define gatewayness as the following:

$$gatewayness_s = \frac{E_{(s)}}{(E1(X) + (E2(Y))}$$

- E(s) = Total edges connecting s to X and Y
- E1(X)|E2(y) = Total edges connecting S to X and Y

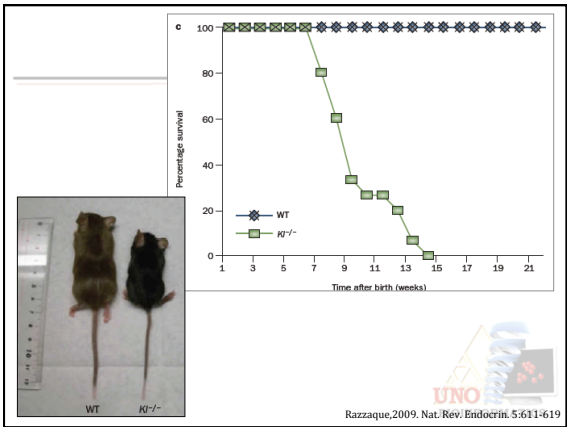




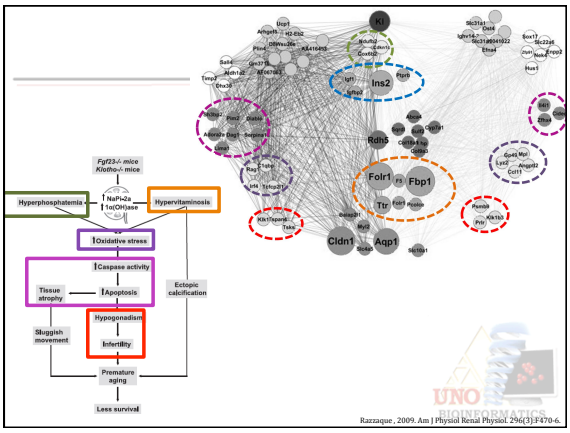


	ID	GENE	DEGREE X	DEGREE Y	DEGREE TOTAL	GATEWAY SCORE	PERCENT	RANK	TARGETED KNOCKOUT LITERATURE
X,Y Gateway Nodes	100956_at	Kl	67	43	110	0.1484	14.8448%	1	Kl-/- mice are growth retarded with shortened lifespan, mouse model of aging [Kurosu 1997]
	95350_at	Ttr	45	37	82	0.1107	11.0661%	5	RNAi in *C. elegans* increases lifespan by 14+% [Hansen 2005]
U,V Gateway Nodes	93293_at	Calm1|Calm2|Calm3	9	6	15	0.0806	8.0600%	9	Loses calcium binding ability with age [Tarcsa 2000]
	101016_at	Arf1	12	4	16	0.0860	8.6000%	8	Expression results in age-correlated change in proliferation [Kim 2006]
	160546_at	Aldoc	12	7	19	0.1022	10.2200%	5	Associated with age-dependent cellular decline and apoptosis [MGI]
	97696_r_at	Rps24	16	9	25	0.1344	13.4400%	2	Identified as up-regulated in late stages of cognitive aging [Kadish 2009]
	160455_s_at	Zwint	17	9	26	0.1398	13.9800%	1	Negatively regulates cell proliferation [Endo 2011]
	96307_s_at	Rpl34	16	9	25	0.1344	13.4400%	2	-
	Cluster Total		121	65	186				

BIOINFORMATICS

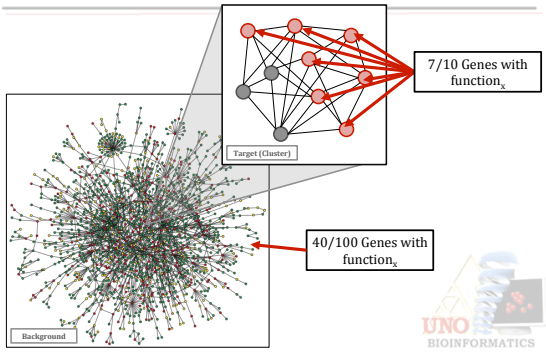


Razzaque,2009. Nat. Rev. Endocrin. 5:611-619

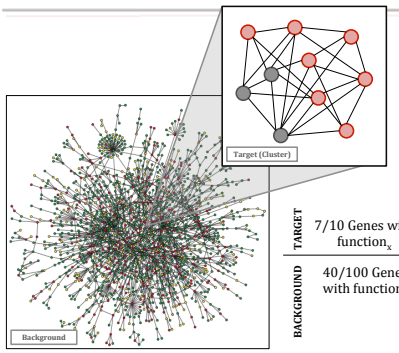


Cluster Annotation

Cluster Annotation: Traditional Method



Cluster Annotation: Traditional Method

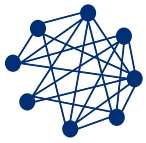


$$\frac{\text{TARGET: } 7/10 \text{ Genes with function}_x}{\text{BACKGROUND: } 40/100 \text{ Genes with function}_x} = 1.75$$

Enrichment of function$_x$ in the cluster

Cluster Annotation

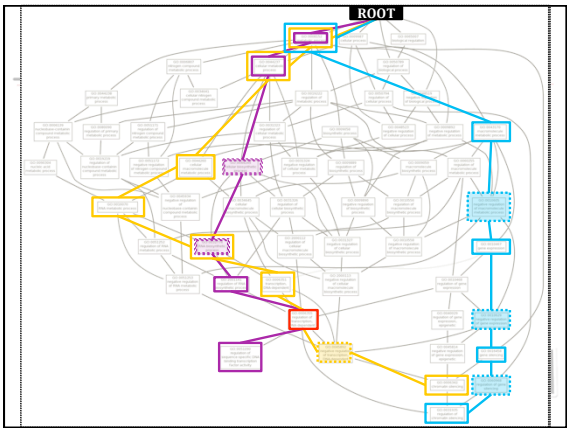
- Functional annotation of clusters
  - Can it be automated?
- Traditional methods
  - Based only on gene groups
  - New concept: Enrichment based on *relationships?*

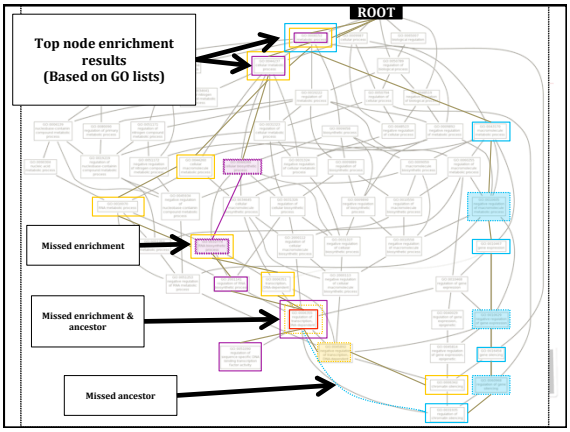


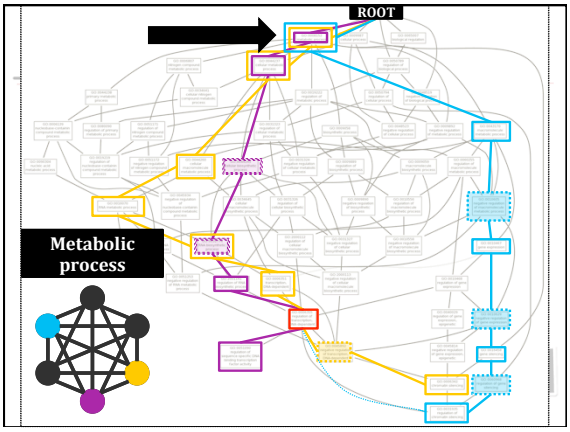
Same # of nodes
Different # of edges

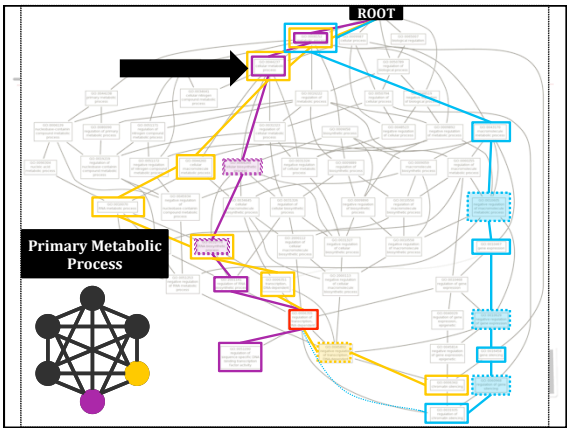
Traditonal GO analysis:
Clusters are same based on genes, not looking at edges

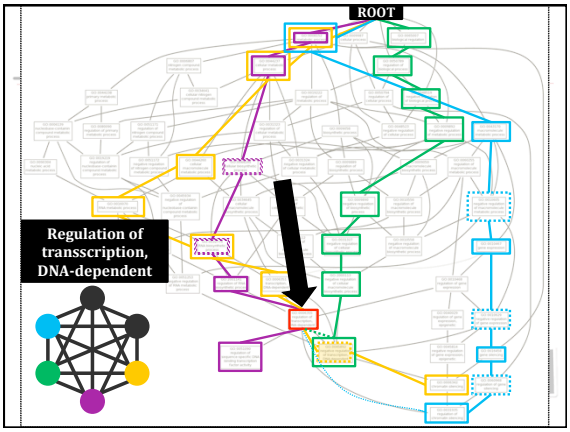
Edge Enrichment Method


ROOT


ROOT
Top node enrichment results (Based on GO lists)
Missed enrichment
Missed enrichment & ancestor
Missed ancestor


ROOT
Metabolic process





Results: Density vs. Score

Many clustering algorithms rank based on density and size

Our method indicates that

density does not necessarily imply functional relevance.

mid 10	0.5000	2.3333	cellular protein metabolic cellular macromolecule biosynthetic process	4 3	Both clusters have 10 nodes!

Cluster Annotation

						Our Method			Traditional Method		
Cluster	NC	EC	TPEs	Density	Avg. Depth	Annotation	Anno. Depth	%	Annotation	Anno. Depth	%
mid 1	41	645	820	0.7866	3.0512	multicellular organismal process	1	7.13%	metabolic process	1	8.14%
mid 2	28	184	378	0.4868	2.7554	transport	3	11.54%	metabolic process	1	5.24%
mid 3	41	239	820	0.2915	2.5565	signal transduction	3	9.43%	cell communication	2	5.01%
mid 4	41	208	820	0.2537	2.8750	transport	3	8.51%	primary metabolic process	2	5.45%
mid 5	10	43	45	0.9556	2.6047	regulation of cellular process	3	25.93%	primary metabolic process	2	5.26%
mid 6	11	47	55	0.8545	2.8511	signal transduction response to chemical stimulus regulation of cell proliferation	3 2 4	9.09% 9.09% 9.09%	cell communication	2	8.54%
mid 7	65	225	2080	0.1082	2.9733	metabolic process	1	6.09%	metabolic process	1	6.65%
mid 8	10	30	45	0.6667	3.3000	positive regulation of transcription from RNA polymerase II promoter primary metabolic process	8 2	15.38% 15.38%	primary metabolic process metabolic process	2 1	8.97% 8.97%
mid 9	58	171	1653	0.1034	2.6199	regulation of cellular process	3	10.48%	primary metabolic process	2	5.35%
mid 10	12	33	66	0.5000	2.3333	regulation of cellular process cellular protein metabolic cellular macromolecule biosynthetic process	3 4 3	12.50% 12.50% 12.50%	primary metabolic process cell communication signal transduction metabolic process	2 2 3 1	7.55% 7.55% 7.55% 7.55%

Structural Summary

	Centrality			
Network Type	Degree	Betweenness	Closeness	Eccentricity
PPI	Essential gene	?	?	?
Metabolome	Abundant metabolite	?	?	?
Correlation	Essential genes/genetic co-regulators			
Genetic Interaction	Essential gene?	?	?	?
Signal Transduction	Receptor/ essential	?	?	?

Structural Summary

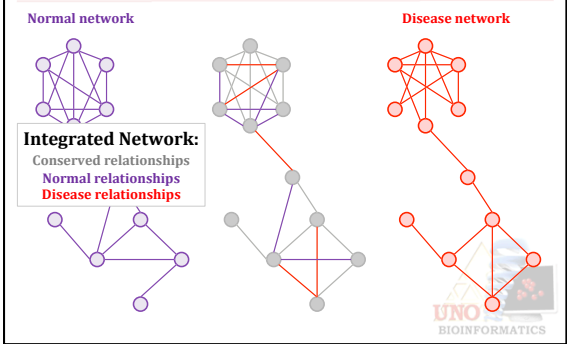
	Clusters			Pathways	Integration	
Network Type	Motifs	Semi-cliques	Cliques		Articulation Points	Overlapping Clusters
PPI	?	Protein complex	Protein complex	N/A	?	N/A
Metabolome	Feedback loop	?	?	Metabolic process	?	
Correlation	Co-reg. set	Regulatory cohort	Regulatory cohort	N/A	Gene drivers	Compensatory clusters
Genetic Interaction	Co-lethal set			N/A	?	?
Signal Transduction	Feedback loop	Exist?	Exist?	Feedback loop/ processes	?	?

Network Integration

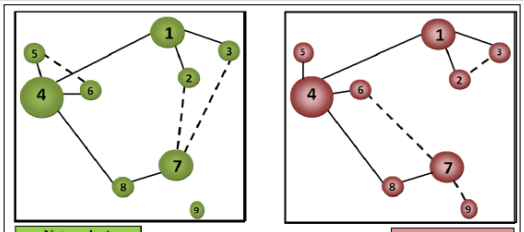
Network Integration

- **Network Alignment**
  - Homozygous (PPI aligned with PPI)
  - Heterozygous (Phenome aligned with transcriptome)
- **Network Combination**
  - Union, Intersection, Difference
- **Data Integration**
  - Knowledge-driven
  - Data-driven

Network Alignment



Network Alignment: Homozygous

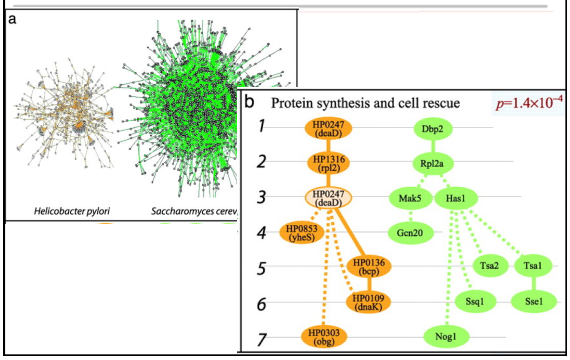


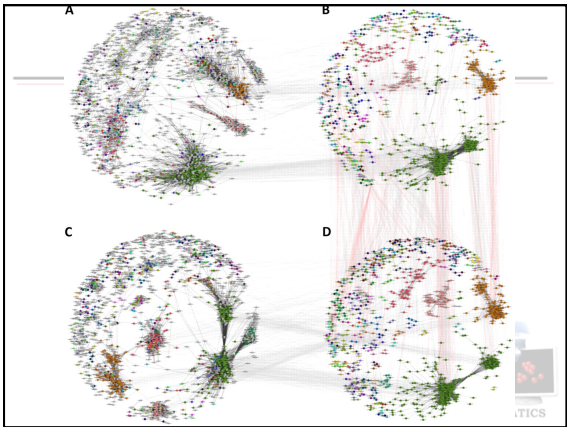
Network A	
gene	degree
4	4
1,7	3
2, 3, 5, 6, 8	2
9	0

Top genes based on comparing the degrees	5, 9
Top genes based on the proposed algorithm	7, 2, 3, 6

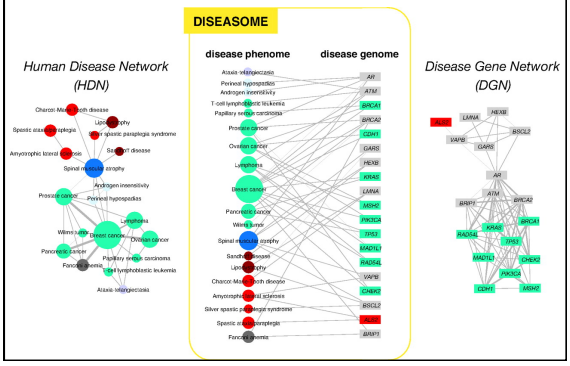
Network B	
gene	degree
4	4
1, 7	3
2, 3, 6, 8	2
5,9	1

Network Alignment: Homozygous

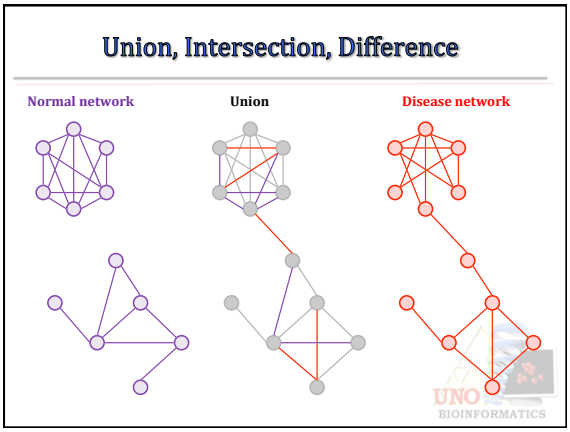




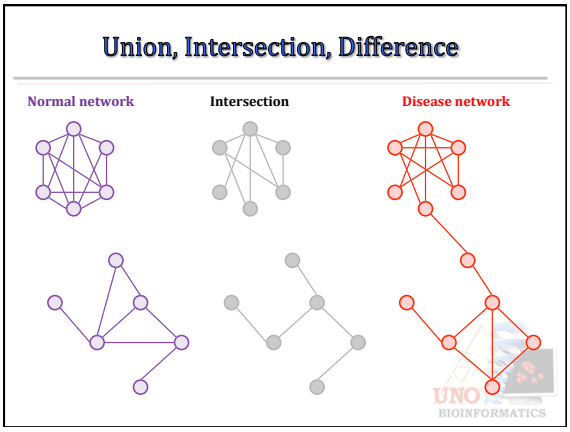
Network Alignment: Heterozygous



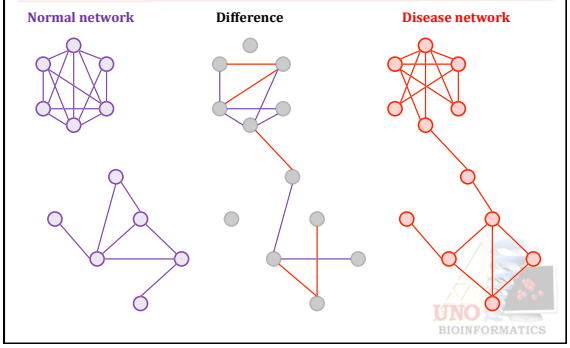
Union, Intersection, Difference



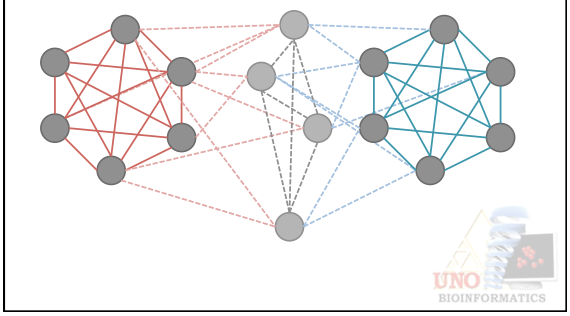
Union, Intersection, Difference



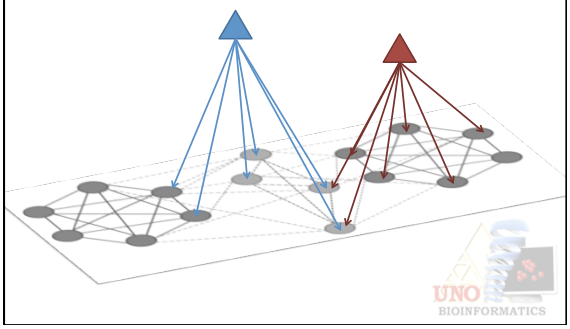
Union, Intersection, Difference



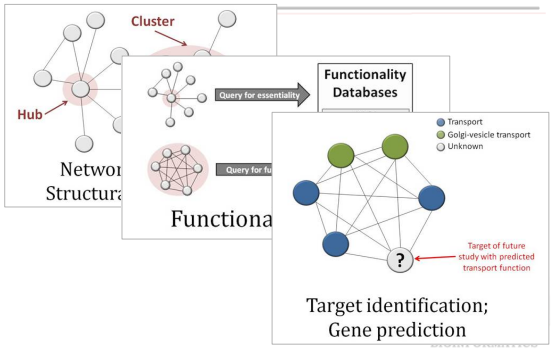
Integration: Data Based



Integration: Data Based

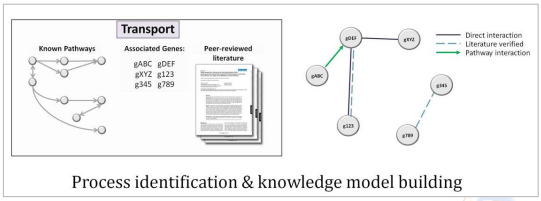


Integration: Data Based Model



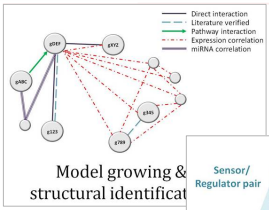
Target identification;
Gene prediction

Integration: Knowledge Based Model

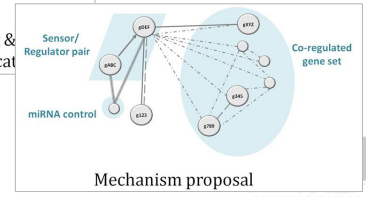


Process identification & knowledge model building

Integration: Knowledge Based Model

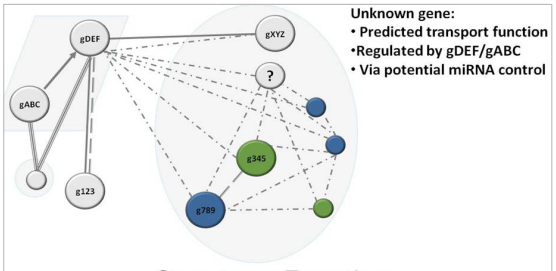


Model growing &
structural identificat



Mechanism proposal

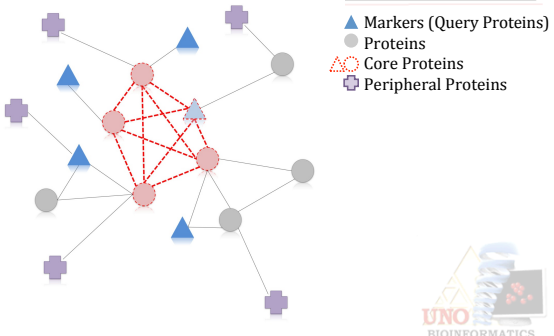
Integration: Knowledge Based Model

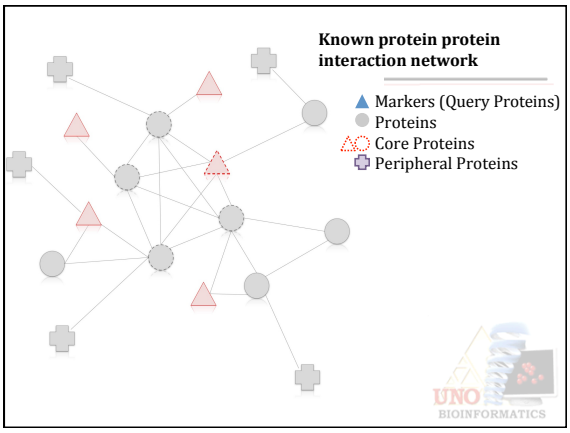


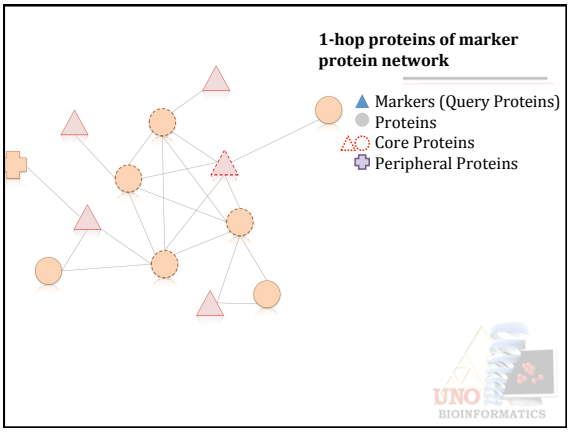
Structure+Function
Integrated Model

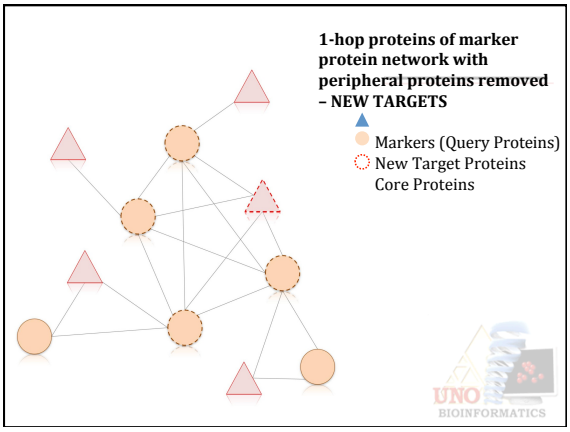
Network Convergence

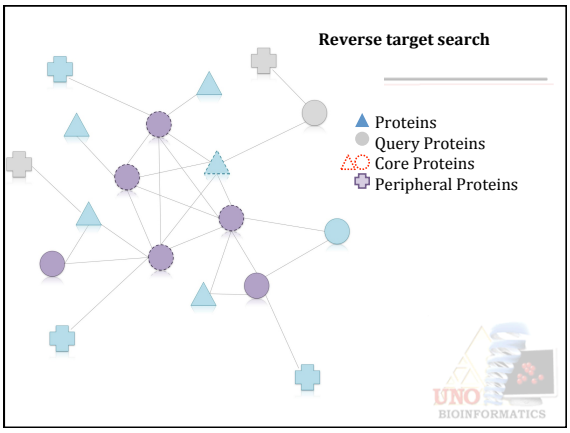
Known protein protein interaction network

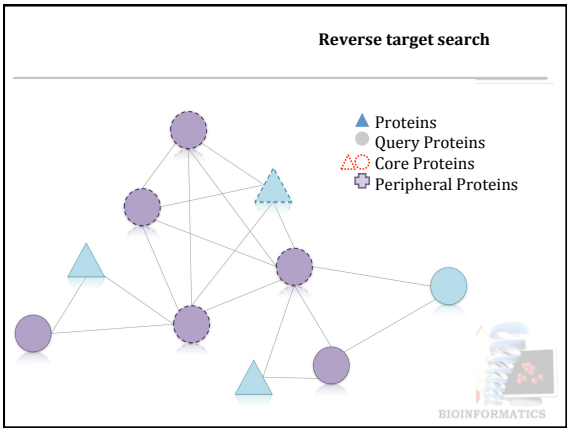


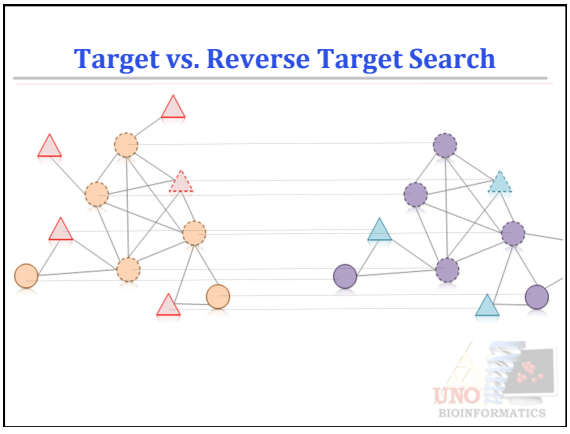

Known protein protein interaction network
Markers (Query Proteins)
Proteins
Core Proteins
Peripheral Proteins
UNO
BIOINFORMATICS


1-hop proteins of marker protein network
Markers (Query Proteins)
Proteins
Core Proteins
Peripheral Proteins
UNO
BIOINFORMATICS

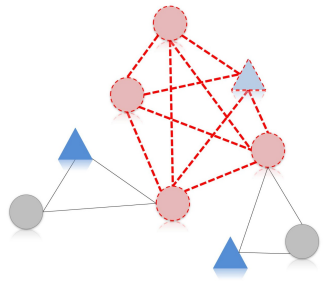

1-hop proteins of marker protein network with peripheral proteins removed – NEW TARGETS
Markers (Query Proteins)
New Target Proteins
Core Proteins
UNO
BIOINFORMATICS


Reverse target search
Proteins
Query Proteins
Core Proteins
Peripheral Proteins
UNO
BIOINFORMATICS


Reverse target search
Proteins
Query Proteins
Core Proteins
Peripheral Proteins
BIOINFORMATICS


Target vs. Reverse Target Search
UNO
BIOINFORMATICS

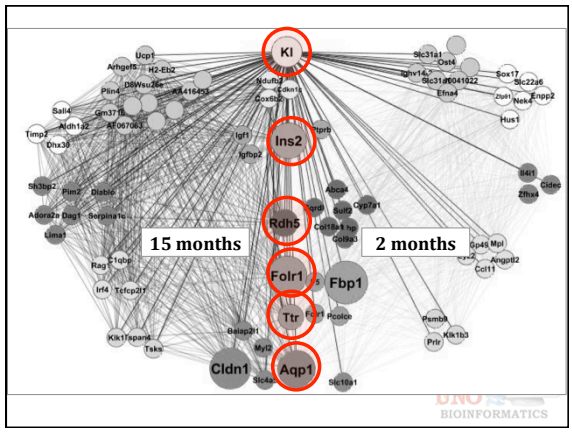
Identification of Core & Mini Cores

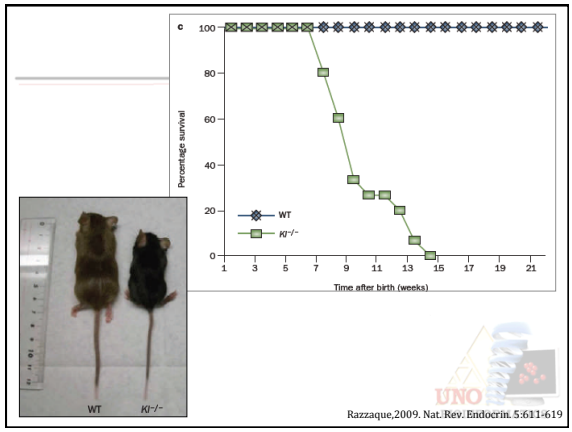


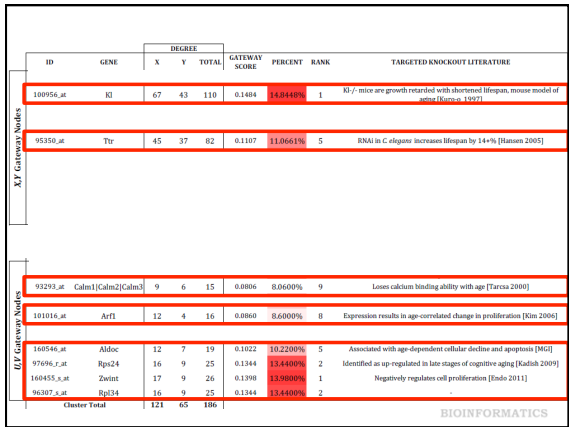
Case Studies using Network Analysis

Case Study I: Aging

- Data: Hippocampal tissue (GSE5078)
  - Young mice (2 months)
  - Aged mice (18 months)
- Method:
  - Identify gateway nodes in 2-state correlation network
- Outcome:
  - Identification of top gateway nodes
  - Notable: Klotho
    - Knockouts of Klotho used as an aging model in mice

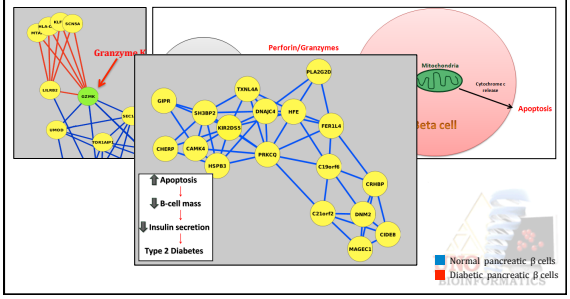

Kl
Ins2
Rdh5
Folr1
Fbp1
Ttr
Cldn1
Aqp1
15 months
2 months
BIOINFORMATICS


c
Percentage survival
Time after birth (weeks)
WT
Kl−/−
WT
Kl−/−
Razzaque,2009. Nat. Rev. Endocrin. 5:611-619


DEGREE
ID
GENE
X
Y
TOTAL
GATEWAY SCORE
PERCENT
RANK
TARGETED KNOCKOUT LITERATURE
X,Y Gateway Nodes
100956_at Kl 67 43 110 0.1484 14.8448% 1 Kl-/- mice are growth retarded with shortened lifespan, mouse model of aging [Kuro-o, 1997]
95350_at Ttr 45 37 82 0.1107 11.0661% 5 RNAi in C. elegans increases lifespan by 14+% [Hansen 2005]
U,Y Gateway Nodes
93293_at Calm1|Calm2|Calm3 9 6 15 0.0806 8.0600% 9 Loses calcium binding ability with age [Tarcsa 2000]
101016_at Arf1 12 4 16 0.0860 8.6000% 8 Expression results in age-correlated change in proliferation [Kim 2006]
160546_at Aldoc 12 7 19 0.1022 10.2200% 5 Associated with age-dependent cellular decline and apoptosis [MGI]
97696_r_at Rps24 16 9 25 0.1344 13.4400% 2 Identified as up-regulated in late stages of cognitive aging [Kadish 2009]
160455_s_at Zwint 17 9 26 0.1398 13.9800% 1 Negatively regulates cell proliferation [Endo 2011]
96307_s_at Rpl34 16 9 25 0.1344 13.4400% 2
Cluster Total 121 65 186
BIOINFORMATICS

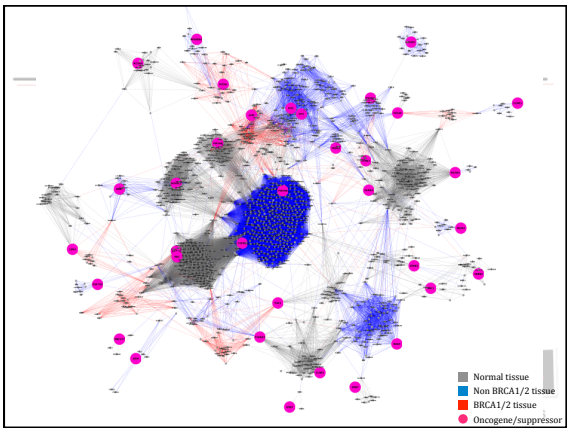
Case Study II: Diabetes

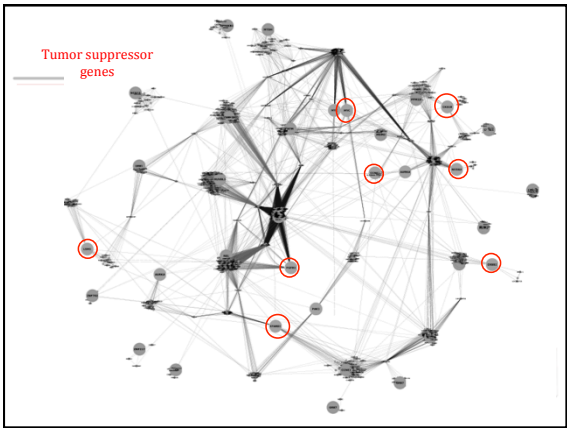
- Data: Pancreatic beta cells (GSE25724)
  - Healthy adults
  - Diabetic adults (Type II, adult onset)
  - Case-matched
- Method:
  - Identify gateway nodes in 2-state correlation network
  - Normal vs. diseased
- Outcome:
  - Identification of top gateway nodes
  - Notable: Granzyme K
    - Facilitates apoptosis in pancreatic beta cells
    - Apoptosis is an upstream step in the development of adult onset Type II diabetes



Case Study III: Breast Cancer

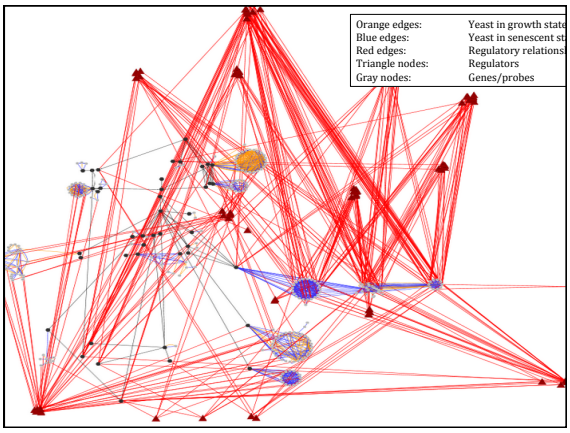
- Data: Breast tissue (GSE17072)
  - BRCA1|2-related mutations (familial breast cancer)
  - Non-BRCA1|2-related mutations (non-familial breast cancer)
  - Control (normal tissue, no breast cancer. Tissue from reduction mammaplasty)
- Method:
  - Identify gateway nodes in 3-state correlation network
  - Normal vs. familial vs. non-familial
  - Identify network distribution/placement of known tumor suppressors and oncogenes
- Outcome:
  - Identification of top gateway nodes
  - Notable: Tumor suppresor genes found to be present more frequently as gateway nodes

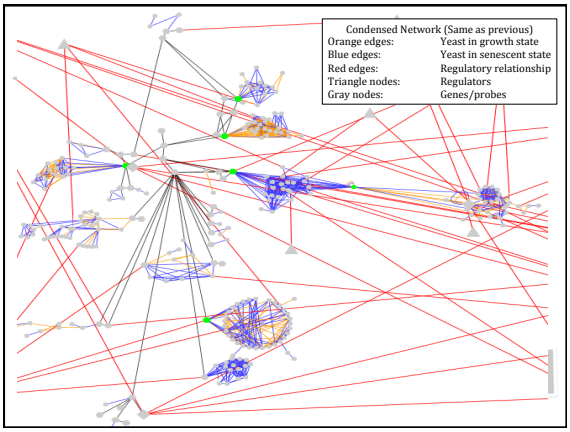


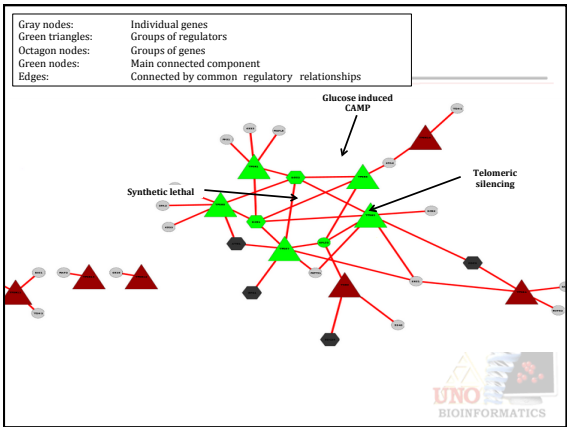


Case Study IV: Yeast Regulation

- Data: *Saccharomyces cerevisiae*
  - Normal growth
  - Senescence
- Method:
  - Identify gateway nodes in 2-state correlation network
  - Integrate regulation data
- Outcome:
  - Identification of top gateway nodes
  - Notable: Identification of major genes involved in regulatory control cohorts

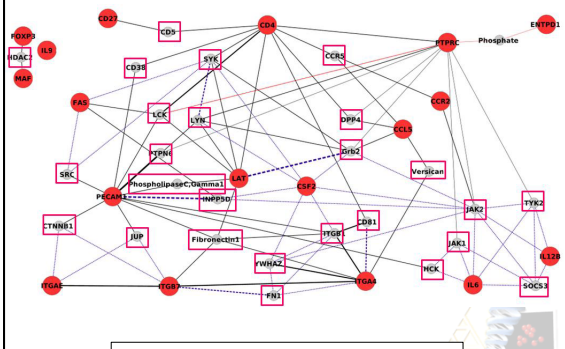

Orange edges: Yeast in growth state
Blue edges: Yeast in senescent st
Red edges: Regulatory relations
Triangle nodes: Regulators
Gray nodes: Genes/probes


Condensed Network (Same as previous)
Orange edges: Yeast in growth state
Blue edges: Yeast in senescent state
Red edges: Regulatory relationship
Triangle nodes: Regulators
Gray nodes: Genes/probes


Gray nodes: Individual genes
Green triangles: Groups of regulators
Octagon nodes: Groups of genes
Green nodes: Main connected component
Edges: Connected by common regulatory relationships
Glucose induced CAMP
Telomeric silencing
Synthetic lethal
UNO BIOINFORMATICS

Case Study V: Parkinson's Disease

- Data: Flow cytometry markers
  - Parkinsons Disease patients
  - Caretakers (non-Parkinsons)
- Method:
  - Create immediate neighbor (1-hop) interactome
  - Identify targets/interactors
  - Identify "key players" based on iterative marker identification
- Outcome:
  - Identification of new marker targets
  - Notable: Identification of 3 major targets based on multiple evidences (from network integration)



Red nodes:	Original markers
Pink boxes:	Marker targets based on connectivity

1-Hop PPI Targets	1-Hop PPI Connected Targets	Pathway Targets	Reverse 1-Hop PPI Targets	Reverse 1-Hop PPI Targets –	Additional Marker Targets
ITGB1	ITGB1	ITGB1	ITGB1	ITGB1	ITGB1
INPP5D	INPP5D		INPP5D	INPP5D	INPP5D
LCK	LCK		LCK	LCK	LCK
PIK3R1	PIK3R1	PIK3R1	PIK3R1	PIK3R1	
SYK	SYK		SYK	SYK	SYK
CD53	CD53		CD53	CD53	
EED	EED		EED	EED	
FYN	FYN		FYN	FYN	
HCK			HCK	HCK	HCK
JUP			JUP	JUP	JUP

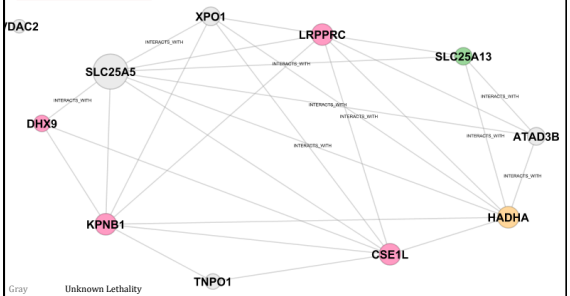
Column 1:	Initial IM network targets
Column 2:	Post-processing IM network targets
Column 3:	Initial Pathway network targets
Column 4:	Reverse - Iterative IM targets – Run 1
Column 5:	Reverse - Iterative IM targets – Run 2
Column 6:	Targets from extraneous data

Case Study VI: Arenavirus

- Data: Affinity assay + Mass spectrometry
  - Extracellular Junin virus Z protein interactors
  - Intracelllular Junin virus Z protein interactors
  - Conserved Junin virus Z protein interactors
- Method:
  - Identify immediate neighbor network of JUNV Z conserved interactors
  - Identify Extracellular/Intracellular markers found in the IM network
- Outcome:
  - 104 of 224 proteins in the Extracellular/Intracellular network identified (evidence for importance)
  - Ongoing study

UNO BIOINFORMATICS

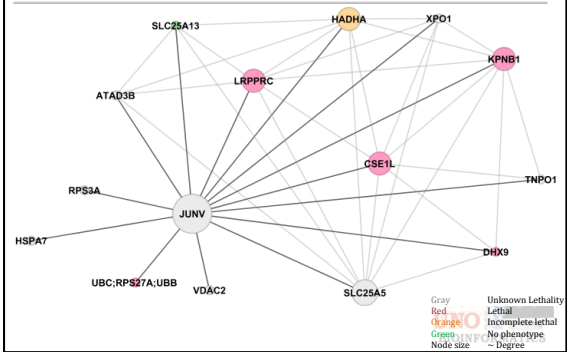
JUNV In Ex Conserved 0-hop Network



Gray	Unknown Lethality
Red	Lethal
Orange	Incomplete lethal
Green	No phenotype
Node size	~ Degree

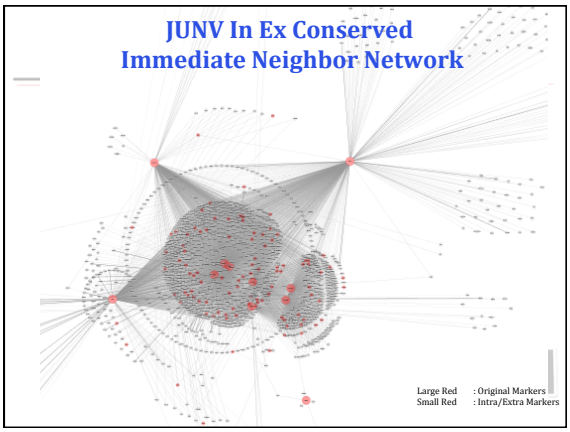
UNO BIOINFORMATICS

JUNV In Ex Conserved 0-hop Merge Network

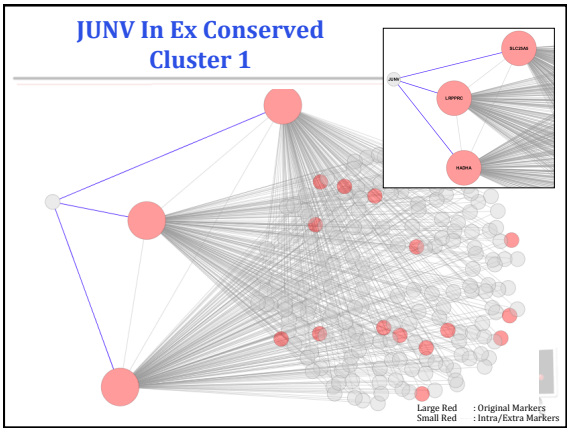


Gray	Unknown Lethality
Red	Lethal
Orange	Incomplete lethal
Green	No phenotype
Node size	~ Degree

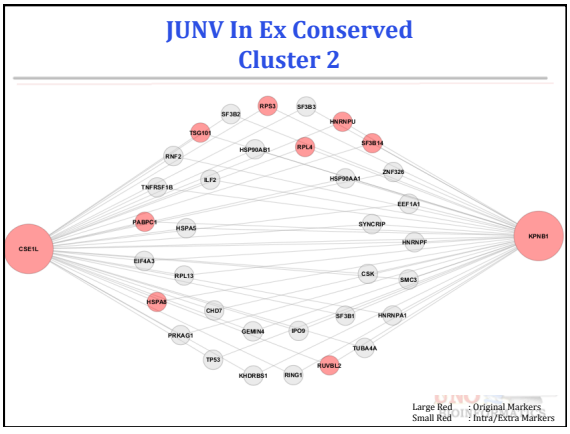
JUNV In Ex Conserved Immediate Neighbor Network



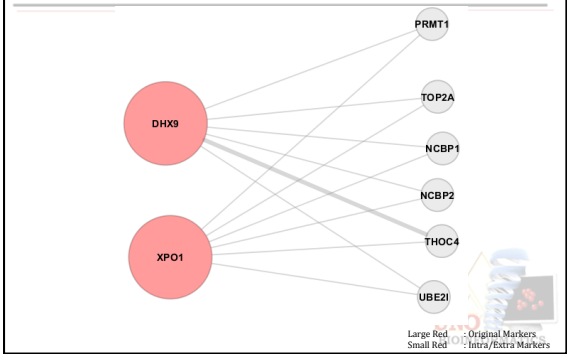
JUNV In Ex Conserved Cluster 1



JUNV In Ex Conserved Cluster 2

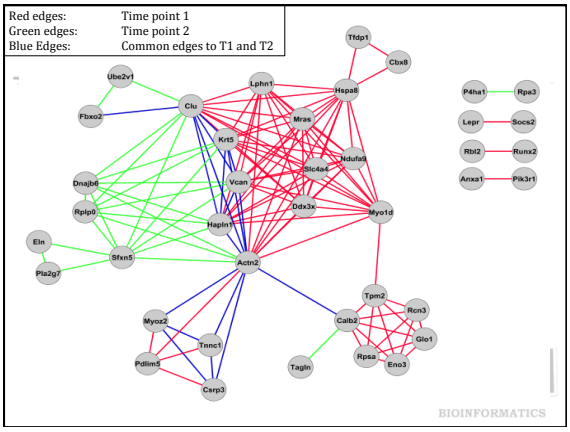


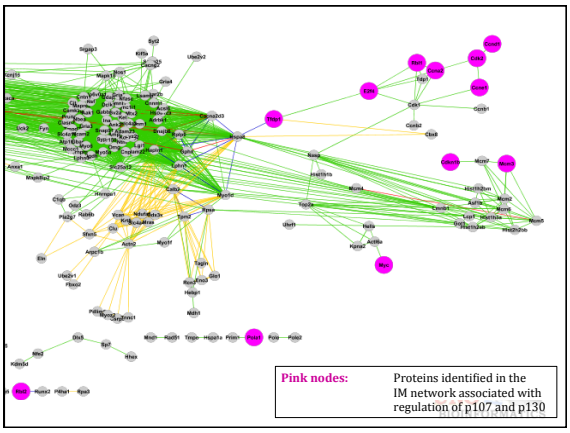
JUNV In Ex Conserved Cluster 3



Case Study VII: Hearing

- Data: Microarray of Hearing Cell (HC) protein knockouts
  - p107 – Deficiency leads to proliferation of HCs
    - Differentially expressed genes when p107 is knocked out
  - p130 – Deficiency leads to lower levels of HCs
    - Differentially expressed genes when p130 is knocked out
- Method:
  - Identify immediate neighbor network of p107 and p130 differentially expressed genes at 3 time points
- Outcome:
  - Identification of proteins known to be involved with cell proliferation and differentiation
  - Ongoing study





Conclusions

- Many types of biological networks
  - Can capture relationships at multiple stages of the central dogma
- Structures in networks can be tied to discrete biological function
  - Depends on noise to signal ratio of network
  - Single nodes or groups of nodes can affect network robustness
- Lethality measures importance of central nodes

Conclusions

- Enrichment measures abundance or lack of function in a set of nodes compared to background
  - Enrichment using nodes vs. edges reveals different functions
- Network Integration is a hot topic in computational biology
  - Proper implementation can lead to important discovery of canditate genes for diease study
  - Achieved through:
    - Network Alignment
    - Data-based models
    - Knowledge-based models
- Multiple case studies to show effectiveness of networks in identifying candidate genes for study from high throughout datasets

Using Cytoscape to analyze networks

- can load any network data in a columnar format
  - eg. for an edge,
    node1 node2
- many plugins available to perform special tasks
  - clustering, extraction of data from public servers,
  - pathway information

Common use case1

- Load a biological network
  - example: protein-protein interaction network
- Highlight the proteins you are interested in
- Find neighbors of your proteins of interest
  - Down sample your network to just these proteins
  - known as one hop network

Common use case1

- Study different graph properties like centrality, betweenness, degree distribution, etc.
- Cluster the network to find groups of proteins which are tightly connected
- Study the biological properties of these clustered proteins

Common use case2

- Load a correlation network of genes with suitable p-value cutoff
- Analyze graph properties like degree distribution, centrality
- Find hubs, driver nodes
- Find clusters, cliques

Common use case2

- discovery
  - overlay biological pathway information to focus on genes of interest and expand on it
- verification
  - find relationship between genes from functional annotation (GO)
  - verify the relationship in the network

Common use case3

- create networks for different conditions/ time points
  - diseased vs. healthy networks
  - young age, middle age and old age networks
- Align graphs to find common and difference between the networks
  - GRAAL

HPC in Analyzing Large Scale Biological Networks

Challenges

(1) Biological networks can be massive in size
Supercomputing access may be limited
Biological network knowledge may be limited.

(2) Noise within the network is likely
Noise within the network cannot be ignored.

How to address these issues: Network filters
- Reduce network size
- Maintain biological signal
- Improve upon biological signal?

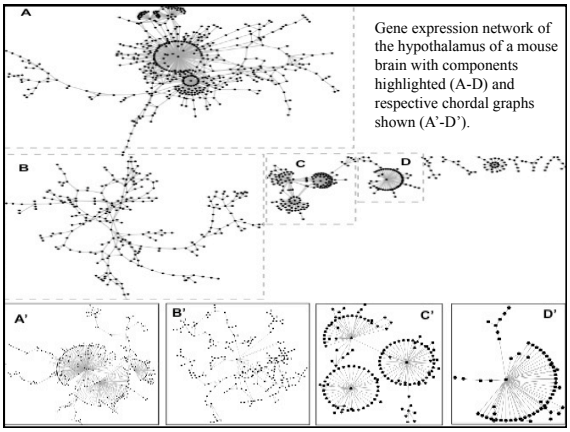
HPC and Biological Networks

- Network creation: 2 weeks on PC
  - 10 hours in parallel, 50 nodes
  - 40,000 nodes = 800 million edges
- Network analysis: Best in parallel
  - Only 3% of entire genome forms complexes
- UNO Sapling Cluster
- Holland Computing Center: Firefly

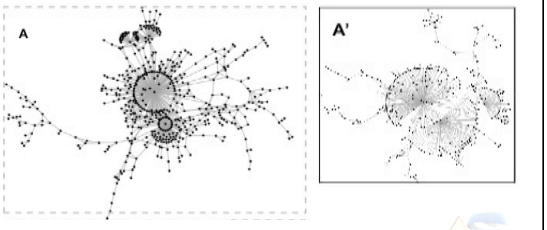
Network Filters

Design a network filter and obtain a sub-network of the original network such that:

- It maintains the important stuff – signal
- Remove unimportant stuff – noise
- Maintain network elements of biological relevance
- Uncover new ones



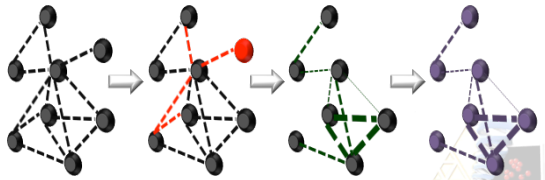
Gene expression network of the hypothalamus of a mouse brain with components highlighted (A-D) and respective chordal graphs shown (A'-D').



A closer view

HPC and Network Analysis

- Network sizes tend to be large
- Signal-to-noise ratio can be high
  - ID biologically relevant relationships?
  - Remove irrelevant nodes/edges?



Original network | ID noisy edges | Weight with literature | Enriched network

Large-Scale Networks and Data Analysis

Many application domain rely on creating networks to model and analyze key relationships among data elements in the domain

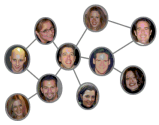
Examples: Biological networks – social networks – inference networks - scheduling networks – Transportation networks

✓ Modeling versus data mining
- Such networks are normally very large
- They are susceptive to significant noise related problems

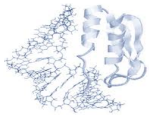
✓ Sampling sub-networks (sub-graphs)
- Reduce network size
- Reduce noise impact
- Questions: does it preserve integrality of the original network

Motivation: Real World Networks

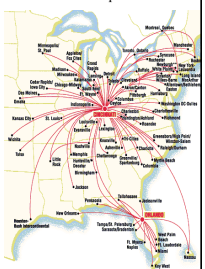
Social Network



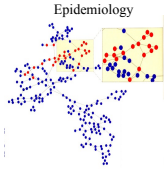
Bioinformatics



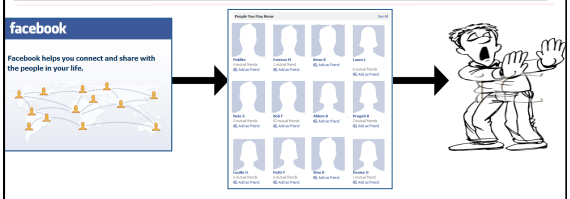
Air Transportation



Epidemiology



Motivation: Signal From Noise

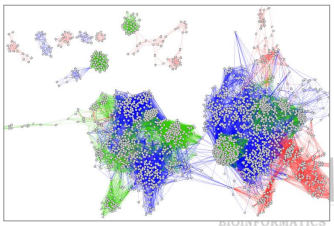


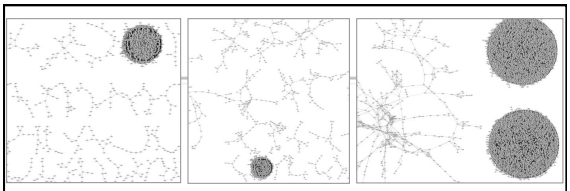
We need a technique that reveals true connections in the network by predicting missing and spurious interactions in a system.

We propose a method that separates wheat from chaff, the signal from the noise

Back to Correlation Networks

- Model for handling high-throughput biological data
- Network contains biologically relevant subgraphs:
  - Hubs
  - Clusters
  - Motifs
  - Bottlenecks





Young	Middle Aged	Aged
28,477 Nodes	29,371 Nodes	29,520 Nodes
77,807 Edges	382,628 Edges	126,354 Edges
1.9% Edge Density	8.8% Edge Density	2.9% Edge Density

- Original Nodes: 41,174
- Original Edges: 847,628,551
- Correlation: Only 1.00
- P-value: p<0.005

Size and Noise

- Network made from average gene expression experiment will have:
  - 40,000 nodes
  - 800 million edges
- Only 3% of genes in entire genome work together to form complexes
- Even with parallel computing resources, unfiltered networks are too noisy for biological discovery

Network Filters

- Chordal graph sampling
  - Keep triangles in expression graphs
  - Remove large cycles, extra edges
  - Keep clusters, identify new clusters
- Spanning tree sampling
  - Keep high degree nodes (maybe?)
  - Remove up to 50% of edges
  - Enhance identification of lethal nodes
- Hybrid chordal-spanning tree method
  - Keep high degree nodes
  - Keep clusters
  - Remove 40-50% of edges
  - Proactively distort/enlarge network structures

Chordal Graph Sampling

Goal: Develop a parallel network sampling technique that *filters noise*, while *preserving the important characteristics of the network.*

✓Maximal Chordal Subgraph

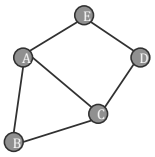
- ➢Spanning subgraph of the network w
- ➢No cycles of length larger than three

✓Properties of Chordal Graph

- ➢Preserves most cliques and highly connected regions of the network
- ➢Most NP hard problems can be solved in polynomial time
- ➢Complexity of finding maximal chordal subgraphs: **O(|E|*max_deg)**

Why chordal graphs?

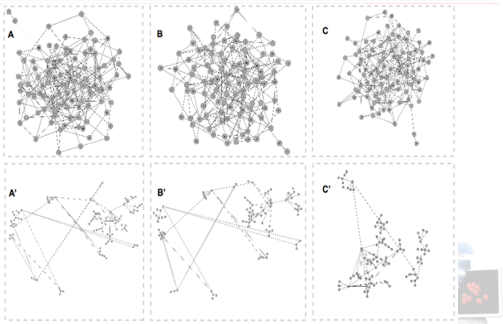
- Chordal graphs are triangulated
  - We want to preserve K_3 subgraphs (triangle)
  - K_3 graphs/motifs are known to represent co-regulated genes
  - Use chordal graphs as a filter for finding co-regulated structures



Subgraph formed by A,B,C is more likely to be biologically relevant.

If gene A and gene B are co-regulated, and if gene A and gene C are co-regulated, then genes B and C will be co-regulated.

Visual Representation

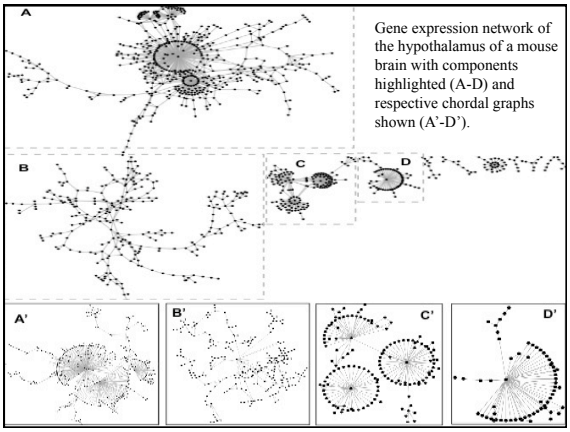


Proposed Approach

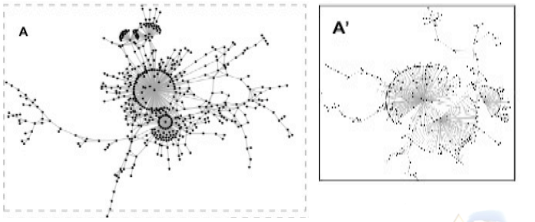
- ✓ Create networks from publicly available data
  - ➢ Aging mice gene expression data –
  - ➢ Young vs. middle-aged mice
- ✓ Test method on networks
- ✓ Assess results by examining biological relevance of network structures
  - ➢ Clusters enriched with function (Gene Ontology)
  - ➢ Do we maintain clusters and function in sampled graphs?
  - ➢ Do we find new functions in sampled graphs?

Hypothesis

- Hypothesis H_0: Given a graph G representing a correlation network, maximal chordal subgraph G_1 will maintain most of the highly dense subgraphs of G while excluding edges representing noise-related relationships in the network.
  - H_{0a} - Key functional properties found in the clusters of unfiltered networks G are maintained in the sampled networks G_1
  - H_{0b} - New clusters with biological function are uncovered. Functional attributes previously lost in noise can now be identified.



Gene expression network of the hypothalamus of a mouse brain with components highlighted (A-D) and respective chordal graphs shown (A'-D').

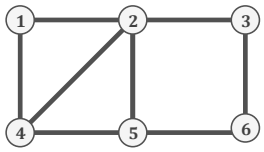


A closer view

Algorithm

- Based on P. M. Dearing *et al.*"Maximal chordal subgraphs", Discrete Applied Mathematics 20(3), 1988.
- Method is based on growing the graph from a starting vertex and adding edges so long as they maintain the chordal characteristics.

1. Select V1
2. Select V2
3. Select V4
4. Select V5
5. Select V3
6. Select V6



Results: Combinatorial Properties

- Young mice, aged 2 months (GSE5078 via NCBI)
  - 5,349 vertices and 7,277 edges

Combinatorial Properties	Original Network	Quasi Chordal Subgraph of young Mice(GSE5078) with				
Number of edges	7277	1,949	4,269	4,029	3,657	3,953
Mean Clust. Coeff	.48	.39	.47	.40	.41	.41
High Degree vertices	146	73(50)	106(72)	96(65)	86(58)	95(65)
Core Numbers	46	26(56)	25(54)	39(84)	36(78)	26(56)

Functional Properties

Cluster Id	Original Network	QCS on 1 Partition	QCS on 2 Partitions	QCS on 4 Partitions	QCS on 8 Partitions	QCS on 16 Partitions
1	protein binding catalytic activity binding	protein binding catalytic activity binding	protein binding catalytic activity binding	protein binding binding enzyme regulator activity	protein binding binding enzyme regulator activity	protein binding catalytic activity binding
2	receptor binding protein binding binding	receptor binding protein binding binding	receptor binding protein binding binding	receptor binding protein binding binding	protein binding binding transmembrane transport. activity	protein binding binding transmembrane transport. activity

- Most of the important functional units are preserved
- Sometimes reveals new clusters in the reduced networks
  - Ex: Enzyme regulator activity in QCS 4, 8
  - Ex: Transmembrane transport activity in QCS 8,16.

Young Mouse – BFS Ordering

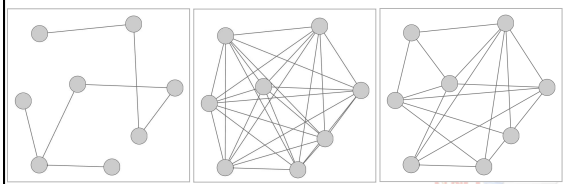
	GO Term	Original	1p	2p	4p	8p	16p	32p
Cluster 1	cellular amino acid, derivative metabolic process		0.15		0.15	0.15	0.15	0.15
	cellular component organization		0.60		0.60	0.60	0.60	0.60
	nitrogen compound metabolic process		0.03		0.03	0.03	0.03	0.03
Cluster2	segment specification				0.24	0.24	0.24	
	cell surface receptor linked signal transduction				2.58	2.58	2.58	
	nervous system development				1.32	1.32	1.32	
Cluster3	nucleo -base/-side/-tide, acid metabolic process	2.06	2.06	2.06	2.06	2.06	2.06	2.06
	cell motion		0.49	0.49	0.49	0.49	0.49	0.49
	system development	1.10	1.10	1.10	1.10	1.10	1.10	1.10
	nervous system development	0.68	0.68	0.68	0.68	0.68	0.68	0.68
	ectoderm development		0.77	0.77	0.77	0.77	0.77	0.77
	nuclear transport		0.05	0.05	0.05	0.05	0.05	0.05
	primary metabolic process		4.53	4.53	4.53	4.53	4.53	4.53
	protein transport		0.84	0.84	0.84	0.84	0.84	0.84
	intracellular protein transport		0.84	0.84	0.84	0.84	0.84	0.84
	mesoderm development	0.84	0.84	0.84	0.84	0.84	0.84	0.84
	intracellular signaling cascade		0.85	0.85	0.85	0.85	0.85	0.85
Cluster4	ion transport			0.21				0.21
	oxygen, reactive oxygen species metabolic process			0.02				0.02
	response to toxin			0.03				0.03
	transport			0.80				0.80
	anion transport			0.04				0.04
Cluster5	vitamin transport	0.04	0.14					
	transport	1.38	4.48					

Identification of New Clusters

Network	Conserved clusters	Newly identified clusters
Young Mouse RCM Ordering	1	6
Young Mouse BFS Ordering	1	3
Middle Aged Mouse BFS Ordering	4	7
Middle Aged Mouse RCM Ordering	4	4

Sampling via Random Walk

- RW: More likely to preserve nodes in clusters; will not preserve edges
- Chordal: Preserves edges and nodes in network

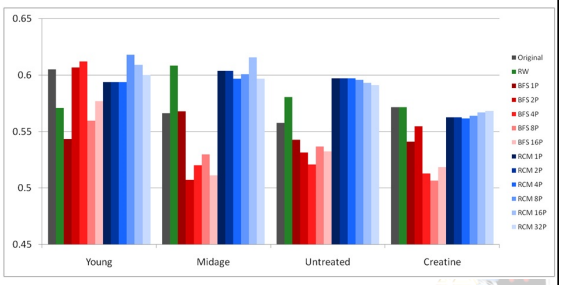


Random Walk Original Chordal

Cluster 1	apoptosis	2.37
	ectoderm development	3.55
	cellular process	16.34
	developmental process	7.55
	immune system process	6.81
	signal transduction	11.13
	cell-cell signaling	3.12
	cell communication	11.53
	nervous system development	3.16
	system development	5.09
	primary metabolic process	20.9
	metabolic process	22
	cellular component organization	3.62
	cell motion	2.25
	cell adhesion	3.3
	intracellular signaling cascade	3.94
	binding	17.57
	receptor binding	3.06
Cluster 2	cell communication	8.26
	signal transduction	7.98
	cell-cell signaling	2.23
	cellular process	11.71
	binding	12.59
	receptor binding	2.19
	protein binding	5.58

4	Common with BFS
9	Common with RCM
1	Common with Both

Lethality: Original vs. Sampled Networks



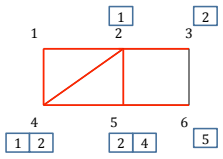
Chordal Graphs

- Chordal graphs: cycle length <=3
- Chordal graphs have polynomial time algorithms for many NP-hard problems
- Are useful in sampling networks to find clusters
- Can be used in detecting impact of change in networks

Maximal Chordal Subgraph

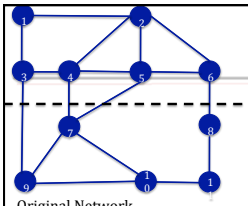
- Obtaining maximum chordal subgraph is NP-hard
- Algorithm for finding maximal chordal subgraph [Dearing et al. 1988]
  - Based on growing a subgraph with chordal edges
  - Depends on the order of the vertices
  - Unweighted Graph
- Complexity O(Ed)

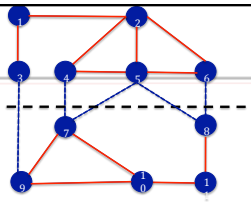
Algorithm

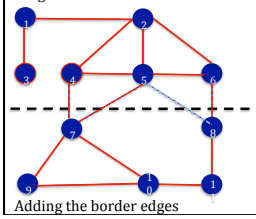


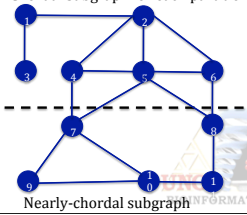
Vertex 1 is selected
Vertex 2 is selected
Vertex 4 is selected
Vertex 5 is selected
Vertex 3 is selected
Vertex 6 is selected

This is a traversal algorithm
Criteria: Seen neighbors of u is a subset of seen neighbors of v


Original Network

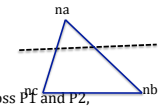

Chordal subgraph for each partition

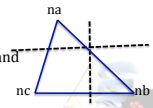

Adding the border edges


Nearly-chordal subgraph

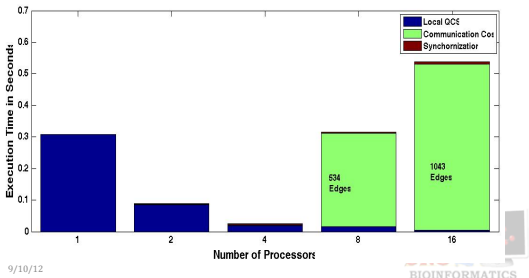
Parallel Chordal Subgraphs (Coarse Grained)

- Input: Original Graph, Output: Maximal Chordal Subgraph (MCS)
- Partition graph across processors
- In each processor Pi (In Parallel)
  - Find local MCS
- Handling Border edges (Communication)
  - Between 2 processors
    - If (na, nb) and (na, nc) are border edges across P1 and P2, and (nb, nc) is a chordal edge in P2
  - Between 3 processors
    - If (na, nb) is a border edge across P1 and P2
    - If (na, nc) are border edges across P1 and P3, and
    - If (nb, nc) is a edge across P2 and p3
- Eliminating Cycles (In Parallel)
  - Run a cycle detection algorithm in each processor
  - Eliminate edge whose vertices have lowest degree



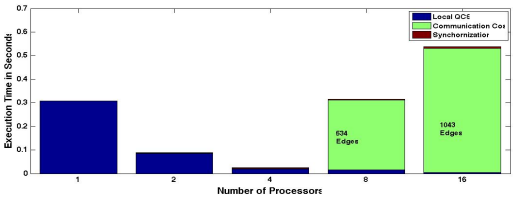


Break Down of Execution Time



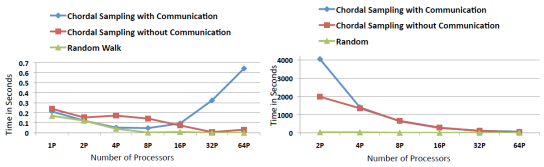
9/10/12

Execution Time

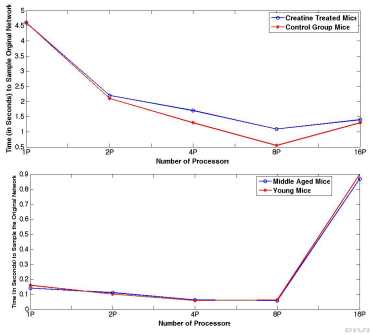


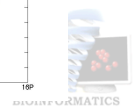
- Parallel algorithm loses efficiency as more processors are added
- Border edges increase with more processors
- We are currently investigating a master-worker approach that might reduce some of the communication costs

Execution Time: Chordal vs. RW

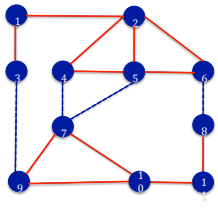


Chordal Graph Speedup



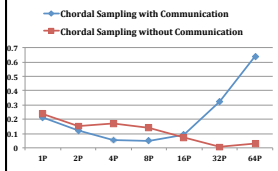


No significant Communication

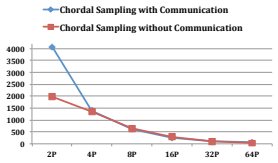


- We can actually remove all communication because the border nodes know the ghosts
- Just check known ghosts for triangles
- Have fewer larger cycles
- Costs duplicate computations in processors

Scalability



Young Mouse 5K vertices and 7K edges



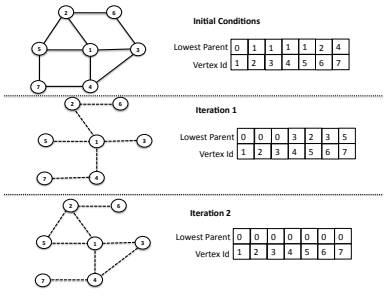
Creatine Treated Mouse 27K vertices and 30K edges

Parallel Chordal Subgraphs (Fine Grained)

- Every vertex associated with an id
- Vertex identifies lowest parent—smallest id number lower than itself
- Find chordal neighbors
- At each iteration check if own chordal neighbors are a subset of the chordal neighbors of lowest parent
  - If yes include LP as chordal neighbor
- Update LP to next highest vertex
- Terminate when no LPs remain

Influenced by Multithreaded Algorithms for Graph Coloring –Catalyurek, Feo, Gebremedhin, Pothen and Halappanavar (preprint)

Algorithm



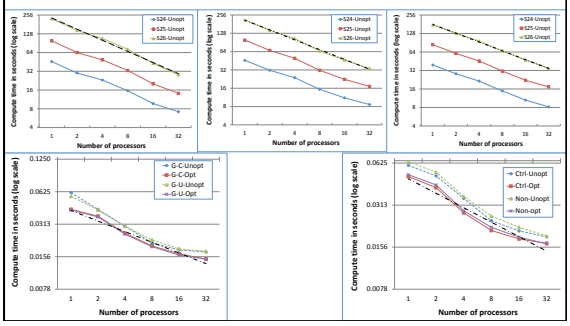
Architecture

- Shared Memory (AMD Opteron)
  - 48 cores, (4 sockets with 12 processors)
  - 256 GB of globally addressable memory
  - 3 level Cache
- Multithreaded (Cray XMT)
  - 128 processors with 128 hardware threads
  - 8 GB per processor
  - No cache—uses hardware hashing for uniform memory access

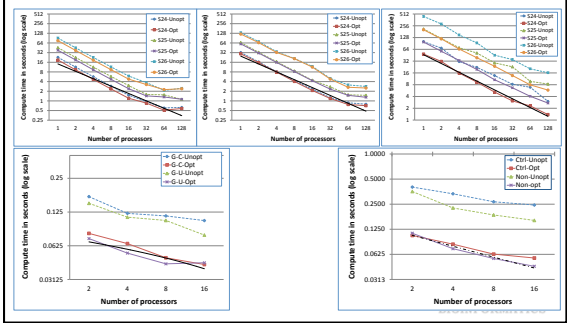
Test Sets

Group	Vertices	Edges	Max Degree	Avg Degree
RMAT-ER (24)	16,777,216	134,217,654	42	16
RMAT-ER (25)	33,554,432	268,435,385	41	16
RMAT-ER(26)	67,108,864	536,870,837	48	16
RMAT-G(24)	16,777,216	134,181,095	1,278	416
RMAT-G (25)	33,554,432	268,385,483	1,489	442
RMAT-G(26)	67,108,864	536,803,101	1,800	469
RMAT-B(24)	16,777,216	133,658,229	38,143	8086
RMAT-B(25)	33,554,432	267,592,474	54,974	9,539
RMAT-B (26)	67,108,864	535,599,280	77,844	11,214
GSE5140 (CRT)	45,023	714,628	690	58
GSE5140(UNT)	45,020	1,311,396	657	32
GSE17072(CTL)	48,803	949,094	365	39
GSE17072(NON)	48,803	1,109,553	463	45

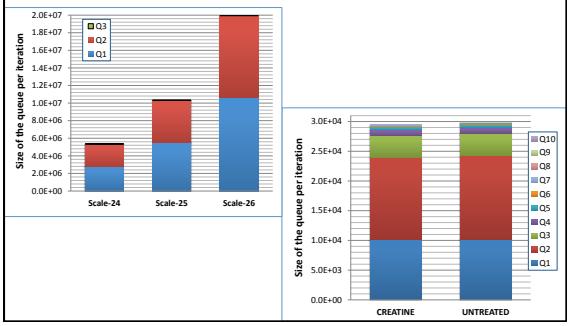
Results on AMD



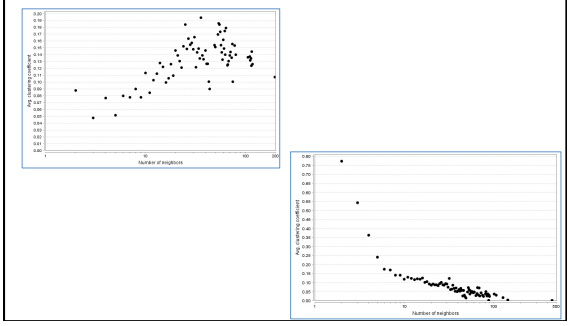
Results on XMT



Number of Iterations XMT



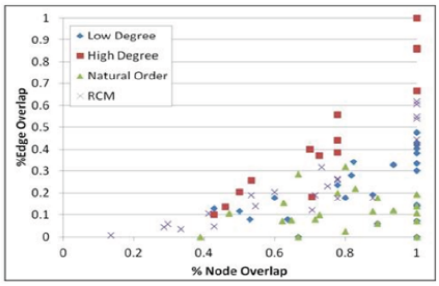
Clustering Coefficient vs Nearest Neighbors



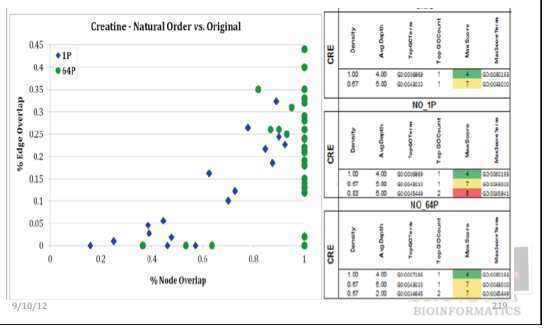
Speedup

Group	XMT (UnOpt)	XMT (Opt)	AMD (UnOpt)
RMAT-ER (24)	32.34	31.70	6.38
RMAT-ER (25)	40.29	29.57	6.96
RMAT-ER(26)	32.25	28.0	7.9
RMAT-G (24)	41.08	41.29	5.33
RMAT-G (25)	44.17	45.97	5.75
RMAT-G(26)	47.35	47.97	6.24
RMAT-B (24)	33.61	35.08	4.80
RMAT-B (25)	21.37	36.16	4.86
RMAT-B(26)	16.70	34.09	5.13
GSE5140 (CRT)	1.40	1.22	3.54
GSE5140(UNT)	1.43	1.14	2.85
GSE17072(CTL)	1.52	1.25	3.41
GSE17072(NON)	2.05	1.65	3.12

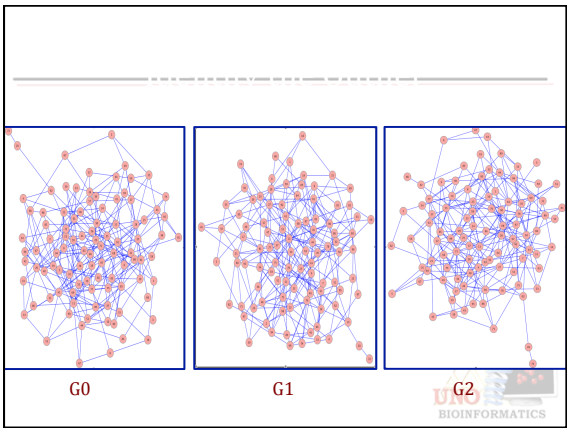
Node and Edge Overlap

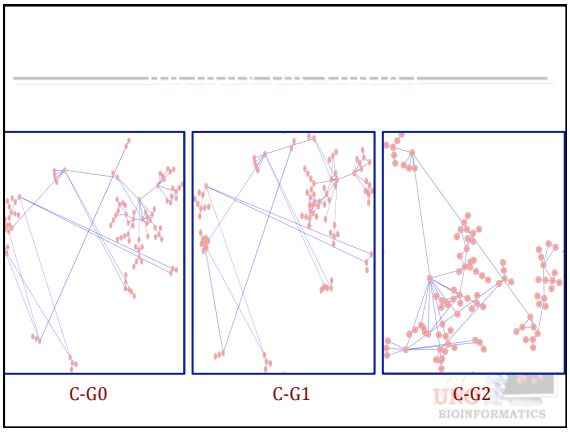


Parallel Results



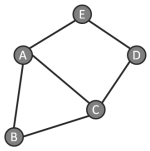
9/10/12





Chordal and other Filters

Chordal-based filters			Tree-based filters		
Filter	*Name*	*Description*	*Filter*	*Name*	*Description*
HD	High Degree	Traversal based on ascending order of vertices	ST	Spanning Tree	Tree determined by Prims Algorithm
LD	Low Degree	Traversal based on descending order of vertices			

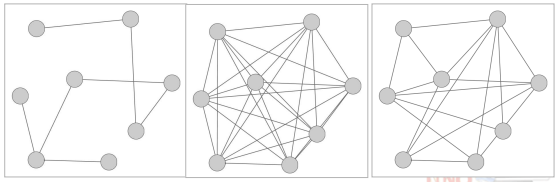


Chordal graphs

- Preserve K_3 subgraphs (triangle)
- Known to represent co-regulated genes
- Chordal graphs as a filter for finding co-regulated structures
- Maintain local dense connections (highly dense subgraphs)

Other Filters

Chordal-based filters			Tree-based filters		
Filter	***Name***	***Description***	***Filter***	***Name***	***Description***
HD	High Degree	Traversal based on ascending order of vertices	ST	Spanning Tree	Tree determined by Prims Algorithm
LD	Low Degree	Traversal based on descending order of vertices			



Spanning Tree Original Chordal

Hypothesis

- *H_0: A tree-based network filter will identify essential high-degree nodes in some network G while reducing the number of edges comparably to a chordal-based filtered network*
  - Propose a spanning tree network filter
    - Reduce the number of edges in the network
    - Nodes with high degree will maintain their high degree n the spanning tree
  - Unweighted edges
    - Edges chosen for the filter will connect to nodes are more likely to connect to central in the original network due to the "friendship paradox" described by [16] for modular networks
  - We use two controls, the original unfiltered network, and multiple chordal graph filters, to verify our findings.