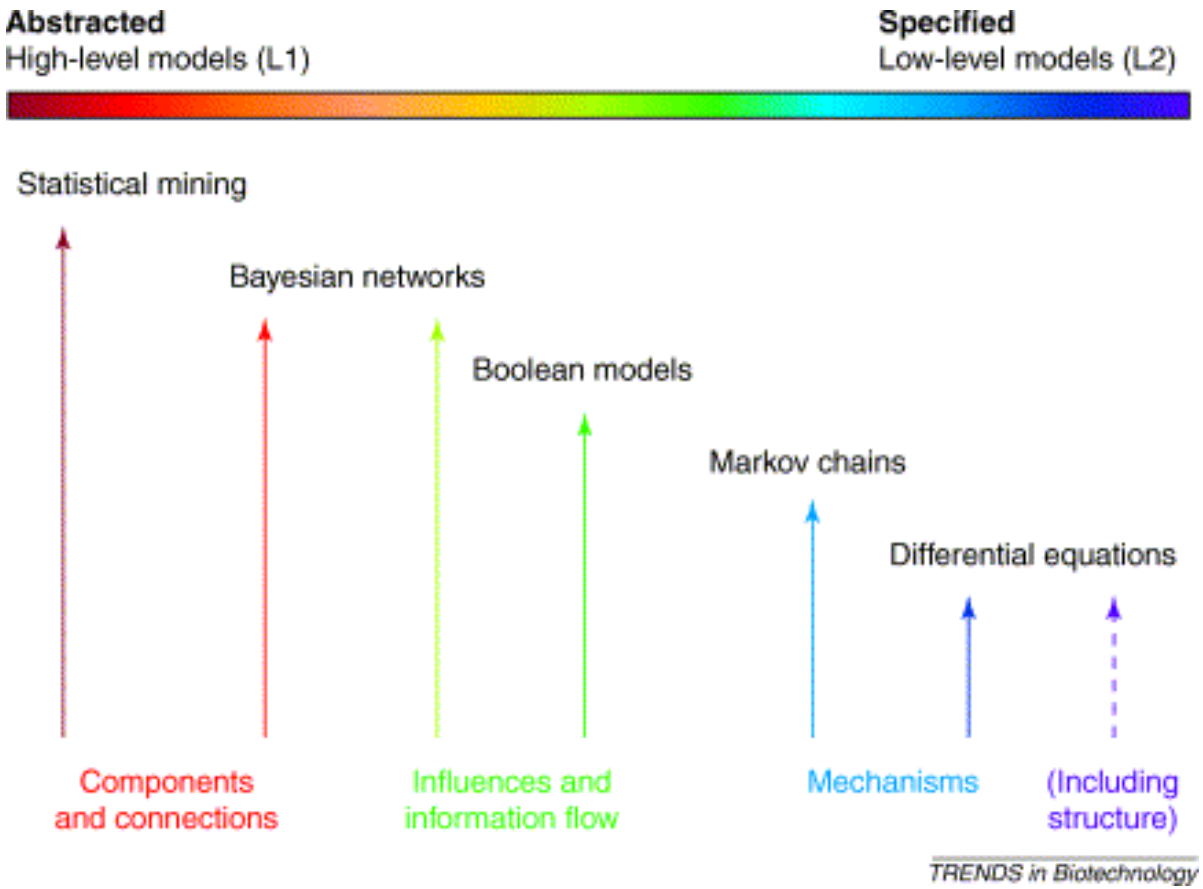
Network Legos: Building Blocks of Cellular Wiring Diagrams

T. M. Murali

Department of Computer Science
Virginia Polytechnic Institute and State University

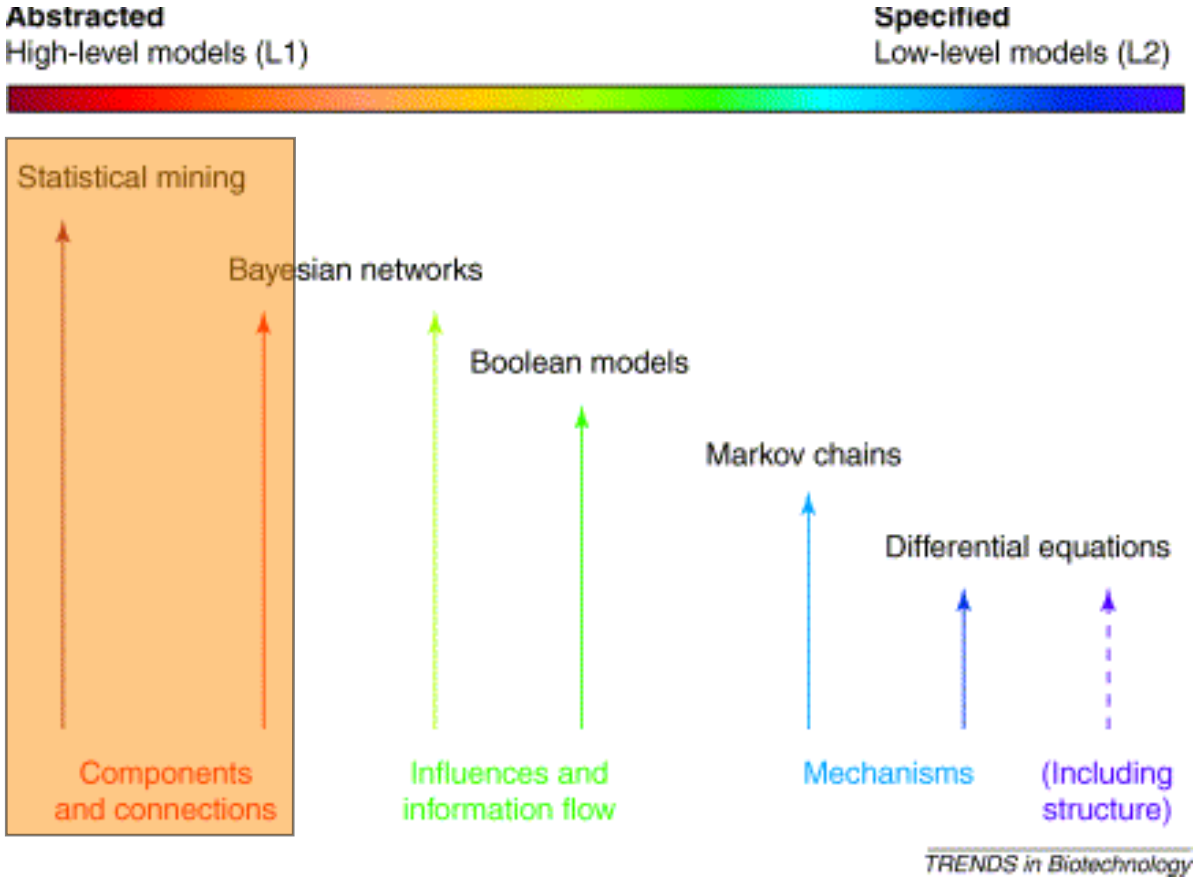
March 7, 2008

Models in Molecular Systems Biology



(From *Building with a scaffold: emerging strategies for high- to low-level cellular modeling*, Ideker and Lauffenburger, Trends in Biotechnology, 2003.)

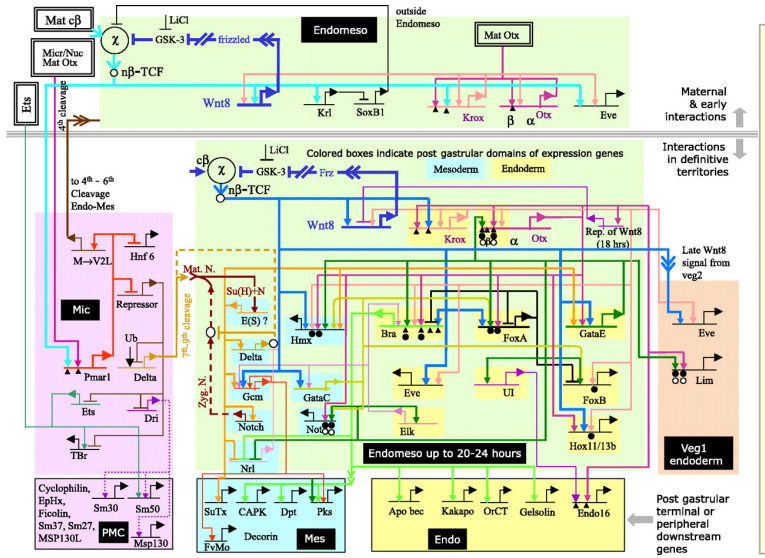
Data-Driven Models in Molecular Systems Biology

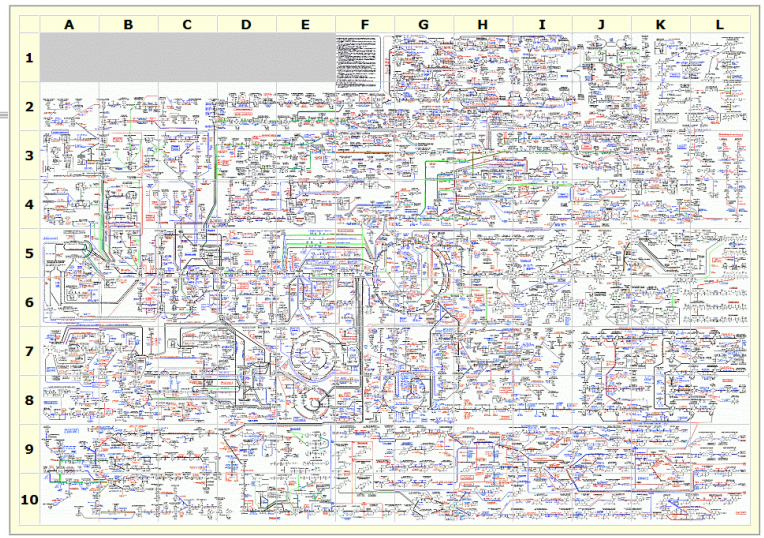


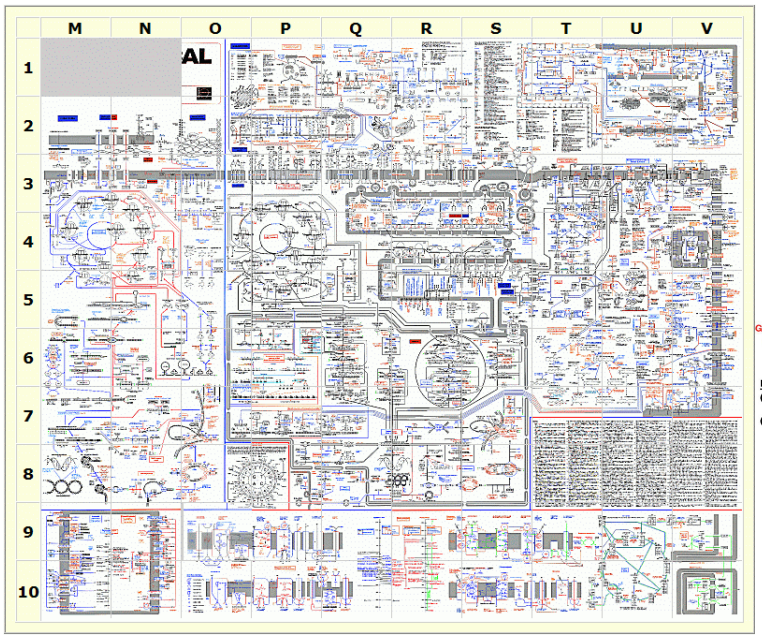
(From *Building with a scaffold: emerging strategies for high- to low-level cellular modeling*, Ideker and Lauffenburger, Trends in Biotechnology, 2003.)

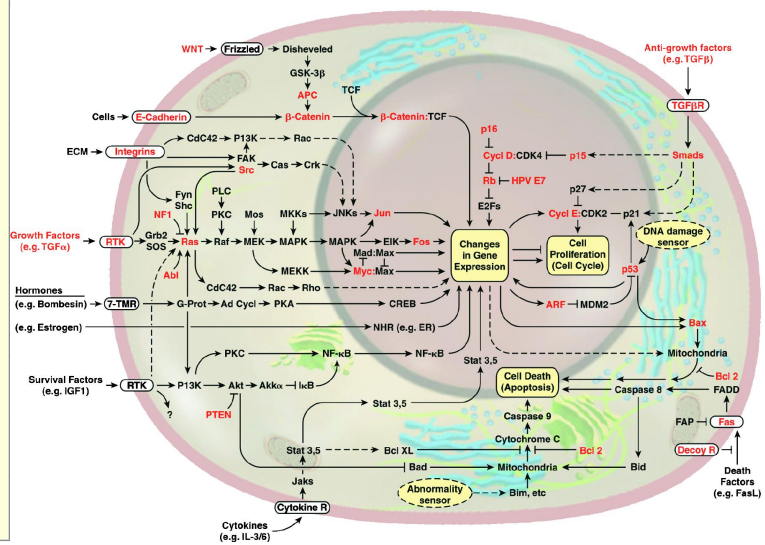
- Emphasise a data-driven approach to molecular systems biology.
- Focus on large-scale properties of molecular wiring diagrams.

Molecular Interactomes

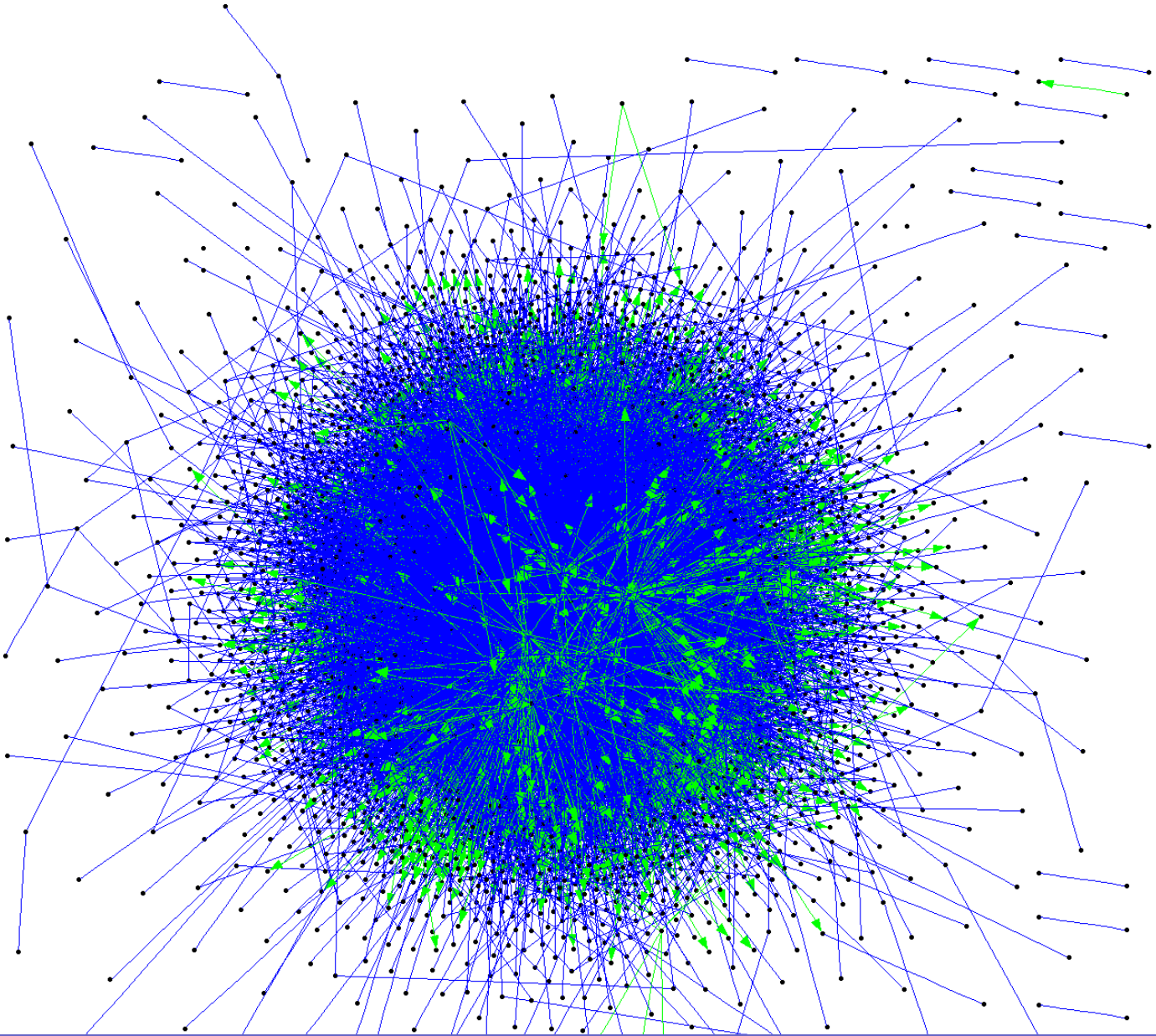








Molecular Ridiculomes



Goals of Data-Driven Molecular Systems Biology

- Identify the building blocks of wiring diagrams.
- Interconnect the building blocks to build high level models of the cell.
- Understand the interaction of the building blocks over time and under different conditions.

Goals of Data-Driven Molecular Systems Biology

- Identify the building blocks of wiring diagrams.
- Interconnect the building blocks to build high level models of the cell.
- Understand the interaction of the building blocks over time and under different conditions.

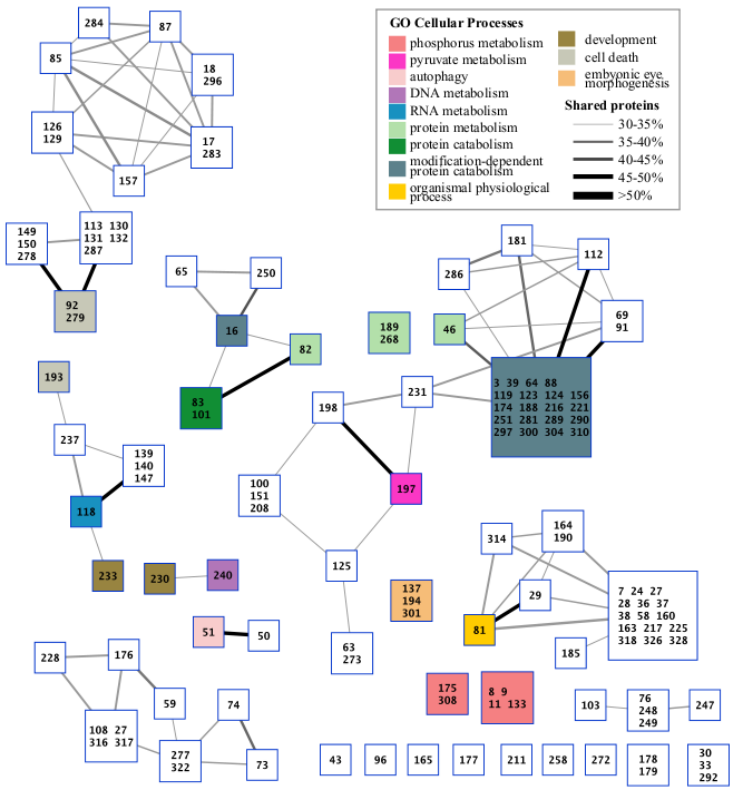
How do we automatically construct these building blocks?

Molecular Interactomes $\longrightarrow$ Modules

- Number of existing techniques decompose interactomes into modules (reviewed by Sharan and Ideker, *Nat. Biotech.*, 2006).
- Map computed modules to known protein complexes, pathways, biological processes, functions, etc.

Molecular Interactomes → Modules

- Number of existing techniques decompose interactomes into modules (reviewed by Sharan and Ideker, *Nat. Biotech.*, 2006).
- Map computed modules to known protein complexes, pathways, biological processes, functions, etc.



(From Sharan et al., *PNAS*, 2005)

Molecular Interactomes $\longrightarrow$ Modules

- Number of existing techniques decompose interactomes into modules (reviewed by Sharan and Ideker, *Nat. Biotech.*, 2006).
- Map computed modules to known protein complexes, pathways, biological processes, functions, etc.

Molecular Interactomes → Modules

- Number of existing techniques decompose interactomes into modules (reviewed by Sharan and Ideker, *Nat. Biotech.*, 2006).
- Map computed modules to known protein complexes, pathways, biological processes, functions, etc.

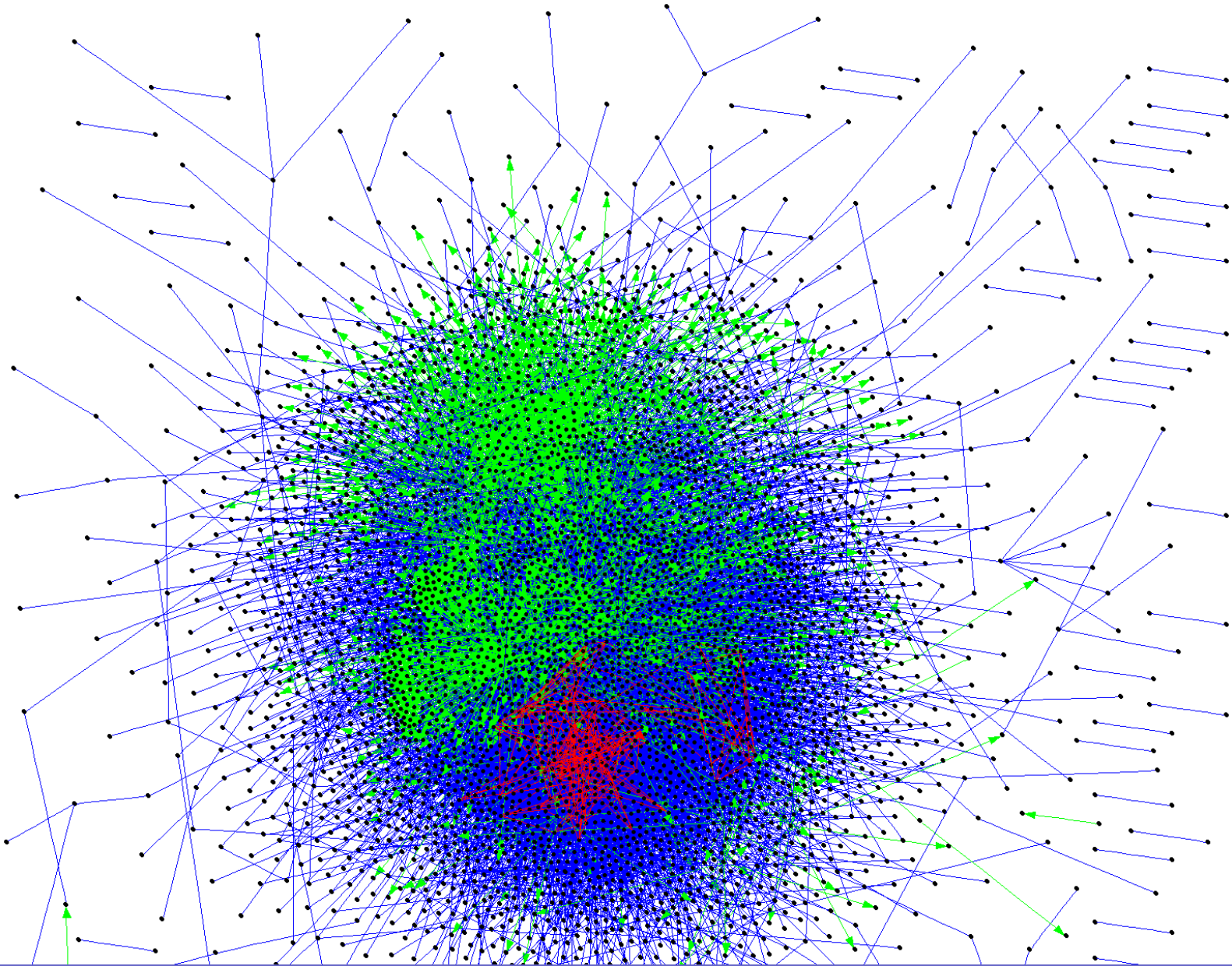
But cell state is dynamic!

- Active molecular interactions change with time, external signals, and perturbations.
- Decompositions of static and *universal* interactomes may miss many important aspects of cellular activity.
- We must integrate interactomes with dynamic measurements of cell state to compute the cell's response to different conditions.

Molecular Interactomes $\longrightarrow$ Active Networks

- Gene expression data provide dynamic snapshots of cellular activity.
- *Active network*: Molecular interactions activated by the cell in response to a stimulus.
- Methods to integrate interactomes with transcriptional measurements to compute active networks:
  - Ideker et al., *Bioinformatics* 2002, *Mol. Sys. Bio* 2007.
  - Bar-Joseph et al., *Nat. Biotech.*, 2003.
  - Luscombe et al., *Nature* 2004.
  - Ulitsky and Shamir, *BMC Sys Bio* 2007.
  - This work: Dense subgraphs (Charikar, *Proc. APPROX* 2000).

Molecular Interactomes → Active Networks



Molecular Interactomes $\longrightarrow$ Active Networks

- Gene expression data provide dynamic snapshots of cellular activity.
- *Active network*: Molecular interactions activated by the cell in response to a stimulus.
- Methods to integrate interactomes with transcriptional measurements to compute active networks:
  - Ideker et al., *Bioinformatics* 2002, *Mol. Sys. Bio* 2007.
  - Bar-Joseph et al., *Nat. Biotech.*, 2003.
  - Luscombe et al., *Nature* 2004.
  - Ulitsky and Shamir, *BMC Sys Bio* 2007.
  - This work: Dense subgraphs (Charikar, *Proc. APPROX* 2000).

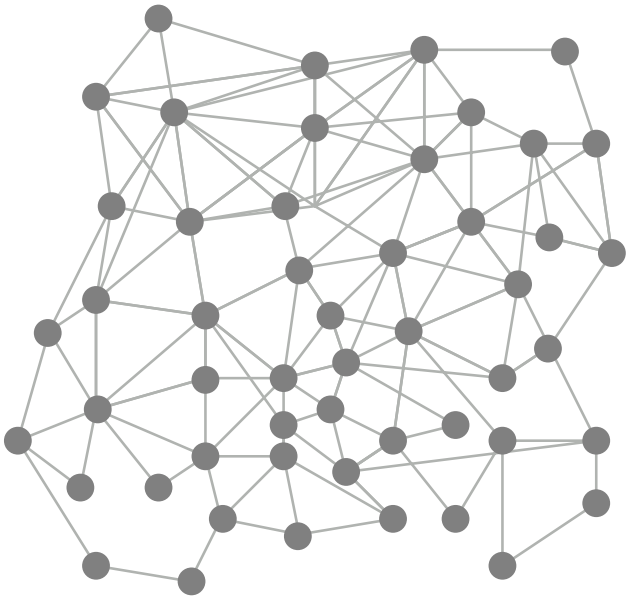
Molecular Interactomes → Active Networks

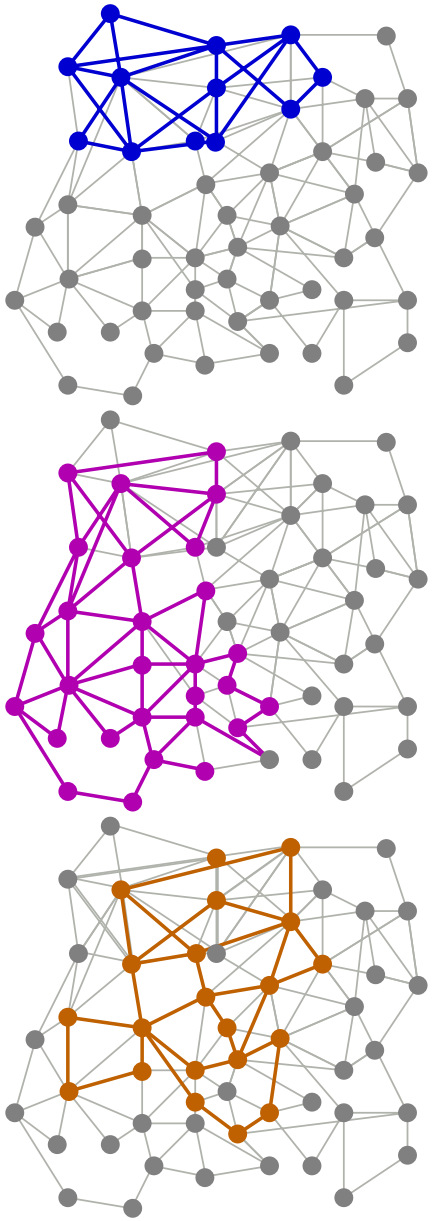
- Gene expression data provide dynamic snapshots of cellular activity.
- *Active network*: Molecular interactions activated by the cell in response to a stimulus.
- Methods to integrate interactomes with transcriptional measurements to compute active networks:
  - Ideker et al., *Bioinformatics* 2002, *Mol. Sys. Bio* 2007.
  - Bar-Joseph et al., *Nat. Biotech.*, 2003.
  - Luscombe et al., *Nature* 2004.
  - Ulitsky and Shamir, *BMC Sys Bio* 2007.
  - This work: Dense subgraphs (Charikar, *Proc. APPROX* 2000).
- These methods usually compute active networks one condition at a time or simultaneously across multiple conditions.

Goals of the Network Lego Approach

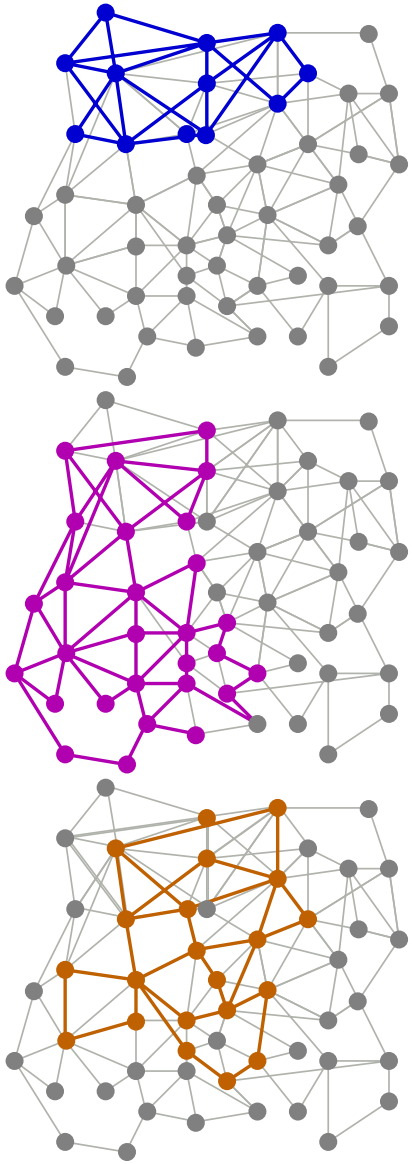
- Combine active network computation with module detection to compute *network legos*: context-sensitive building blocks of wiring diagrams.
- Potential applications:
  1. Identify pathways uniquely activated in one or more conditions.
  2. Compare and contrast responses of different cell types to the same stress.
  3. Develop a formalism for expressing any active network as a combination of network legos.

Step 1: Molecular Interactome to Active Networks

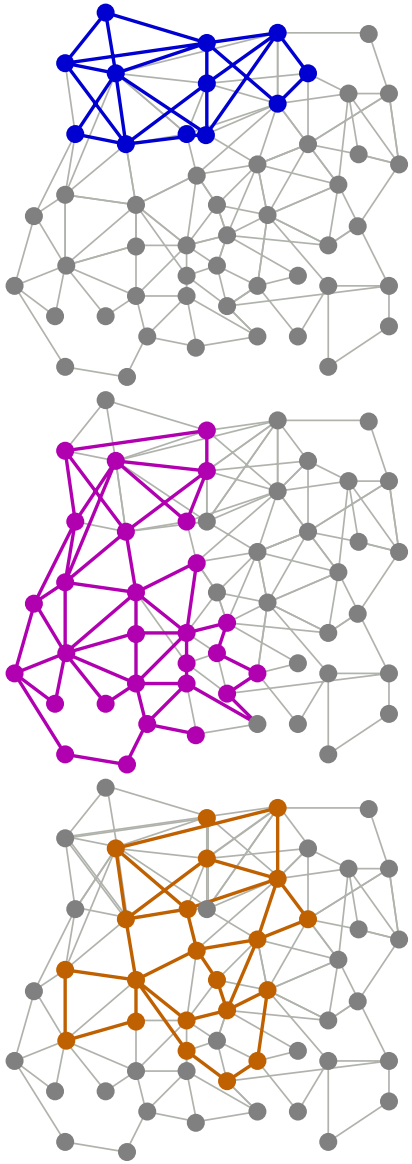


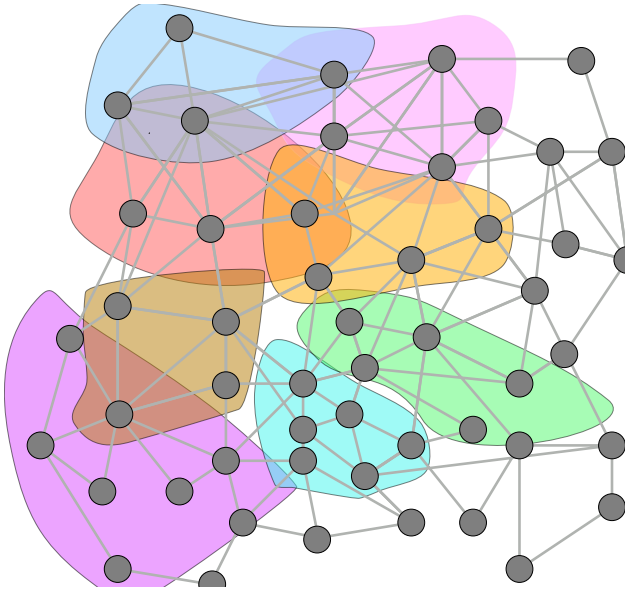


Step 2: Active Networks to Network Legos

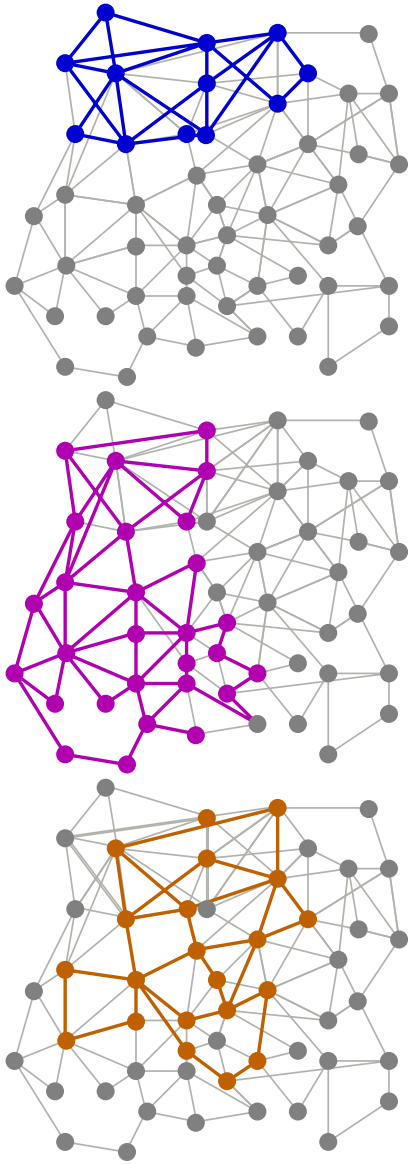


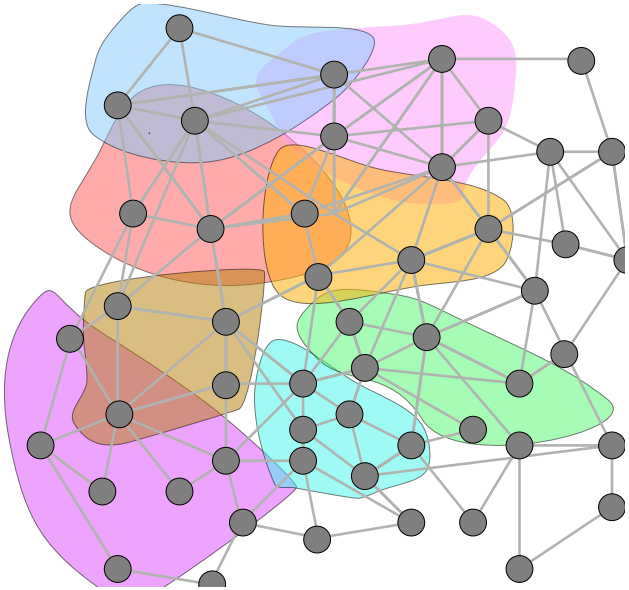
Step 2: Active Networks to Network Legos

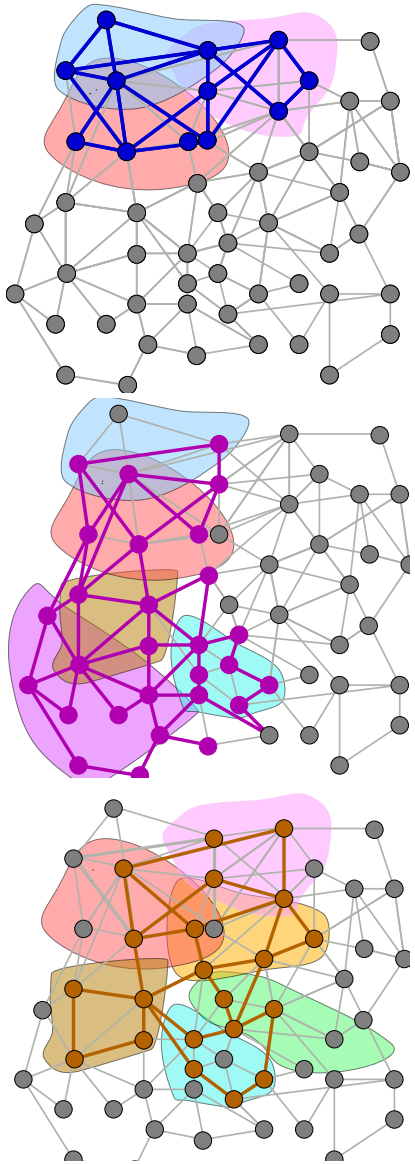




Step 2: Active Networks to Network Legos







Caveats

- Interactomes are incomplete and noisy.
- Gene expression measurements miss many aspects of cellular state.
- We will consider only presence or absence of an interaction in an active network.
- Network legos are only a mental model of how the cell may operate.

Network Blocks

- Suppose we have gene expression datasets for a number of conditions.
- Compute the active network for each condition.
  - Consider each active network to be a set of interactions.
  - Any set operation on these active networks will yield another network of interactions.

Network Blocks

- Suppose we have gene expression datasets for a number of conditions.
- Compute the active network for each condition.
  - Consider each active network to be a set of interactions.
  - Any set operation on these active networks will yield another network of interactions.
- Let $\mathcal{A}$ be the set of all active networks.
- A *network block* is a triple $(G, \mathcal{I}, \mathcal{E})$ where
  - $\mathcal{I} \subseteq \mathcal{A}$, $\mathcal{I}$ is non-empty.
  - $\mathcal{E} \subseteq \mathcal{A}$, disjoint from $\mathcal{I}$.
  - $\mathcal{I}$ and $\mathcal{E}$ are inclusion-maximal
  - G is a network where each interaction
    - is present in every active network in $\mathcal{I}$.
    - is absent in every active network in $\mathcal{E}$.

$$G = \left(\bigcap_{P \in \mathcal{I}} P \right) \bigcap \left(\bigcap_{N \in \mathcal{E}} !N \right)$$

ALL, AML, and MLL

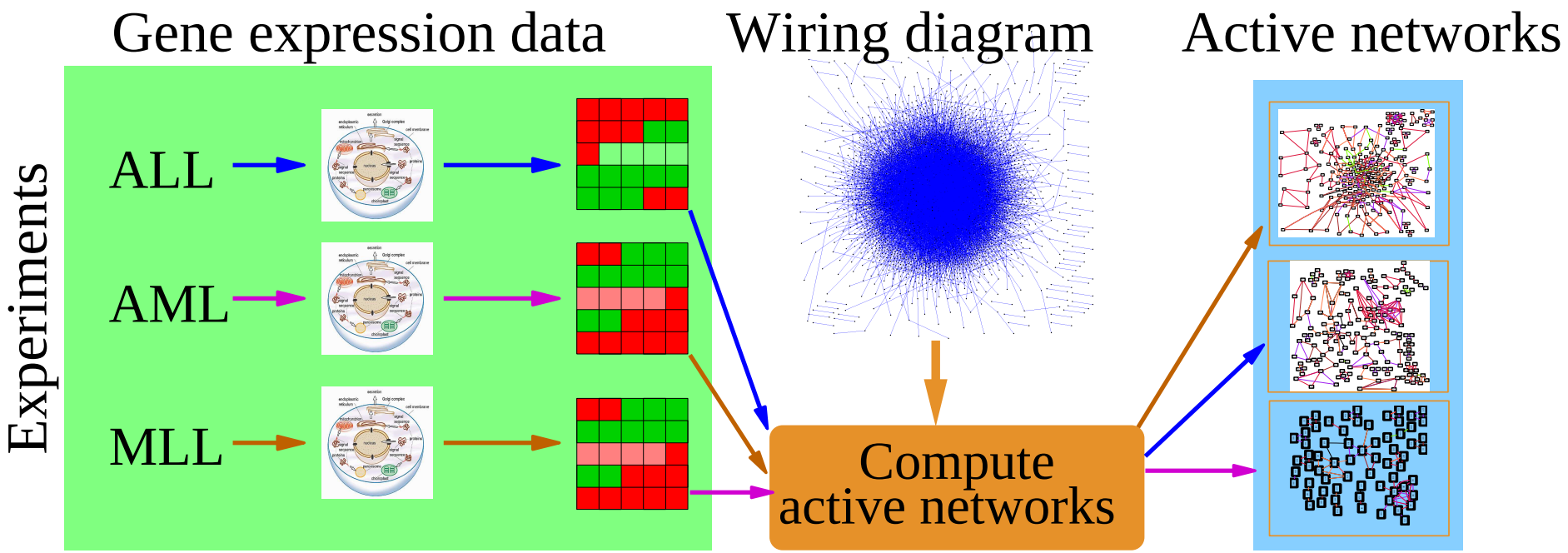
- Acute Lymphoblastic Leukaemia (ALL) and Acute Myeloid Leukaemia (AML) are two types of leukaemia.
- Armstrong et al., Nature Genetics 2003 argued that translocations in the Mixed Lineage Leukaemia (MLL) gene identify a disease distinct from ALL and AML.

ALL, AML, and MLL

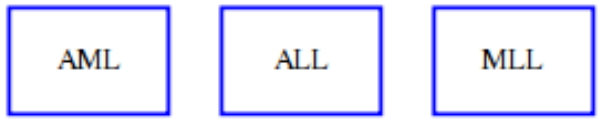
- Acute Lymphoblastic Leukaemia (ALL) and Acute Myeloid Leukaemia (AML) are two types of leukaemia.
- Armstrong et al., Nature Genetics 2003 argued that translocations in the Mixed Lineage Leukaemia (MLL) gene identify a disease distinct from ALL and AML.

Can we compare active networks to identify subsets of interactions differentially activated in each leukaemia?

Comparing ALL, AML, and MLL

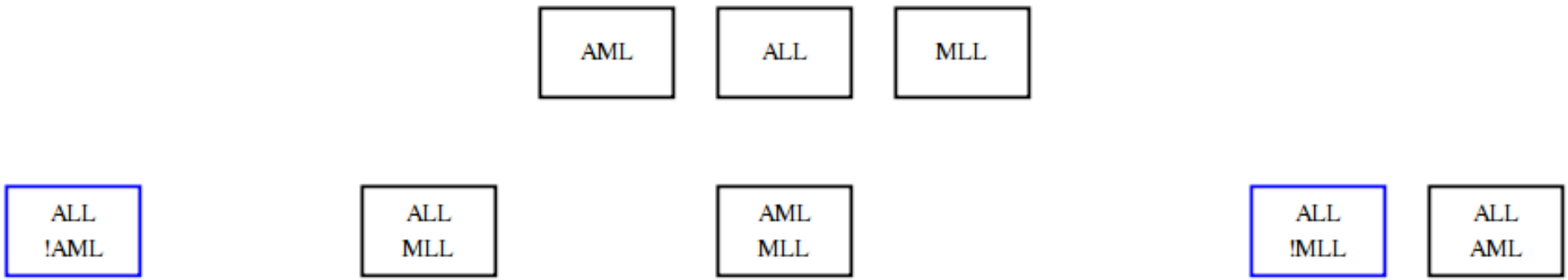


Computing 17 Comparisons

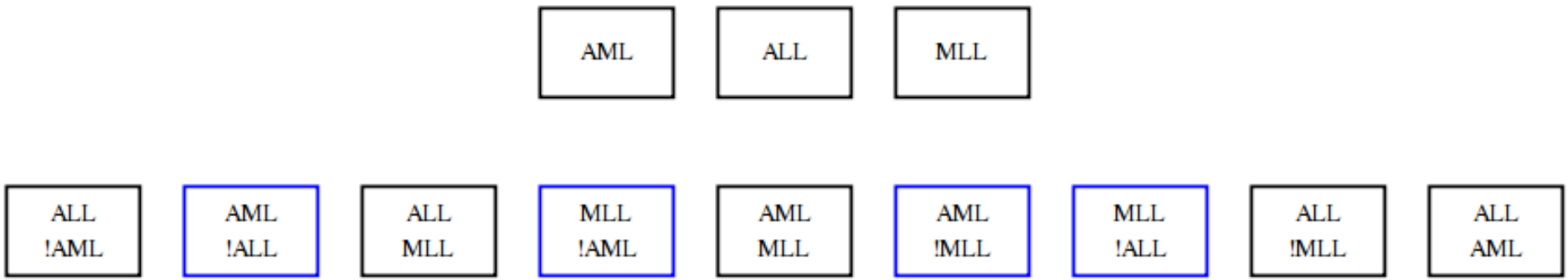


Computing 17 Comparisons

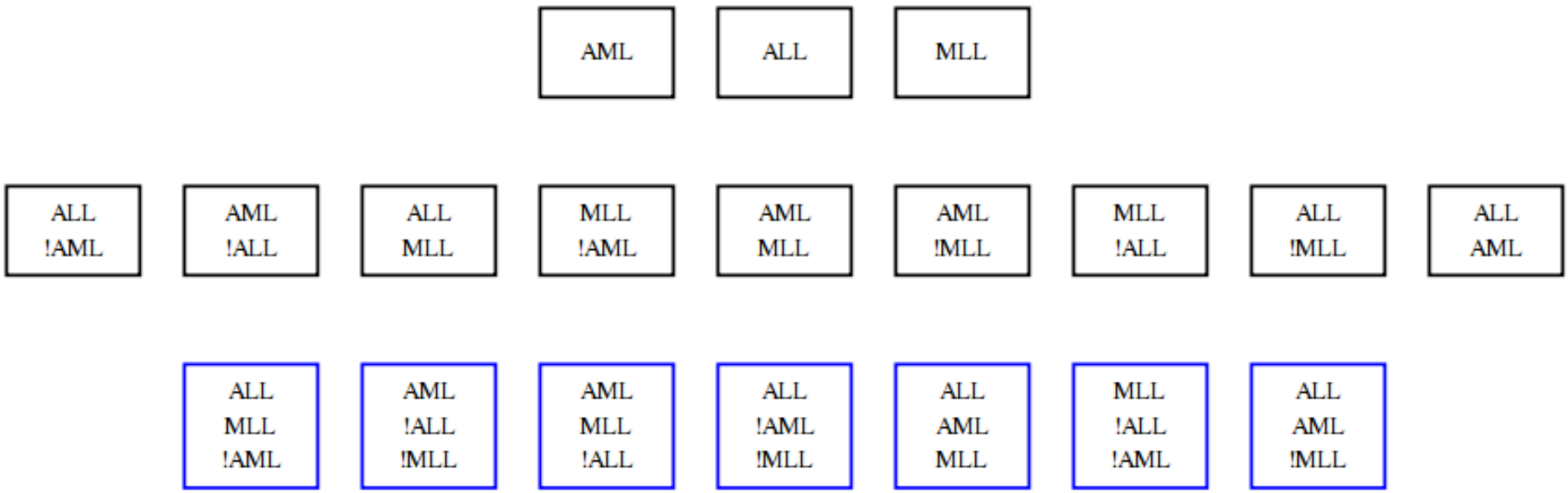
Computing 17 Comparisons



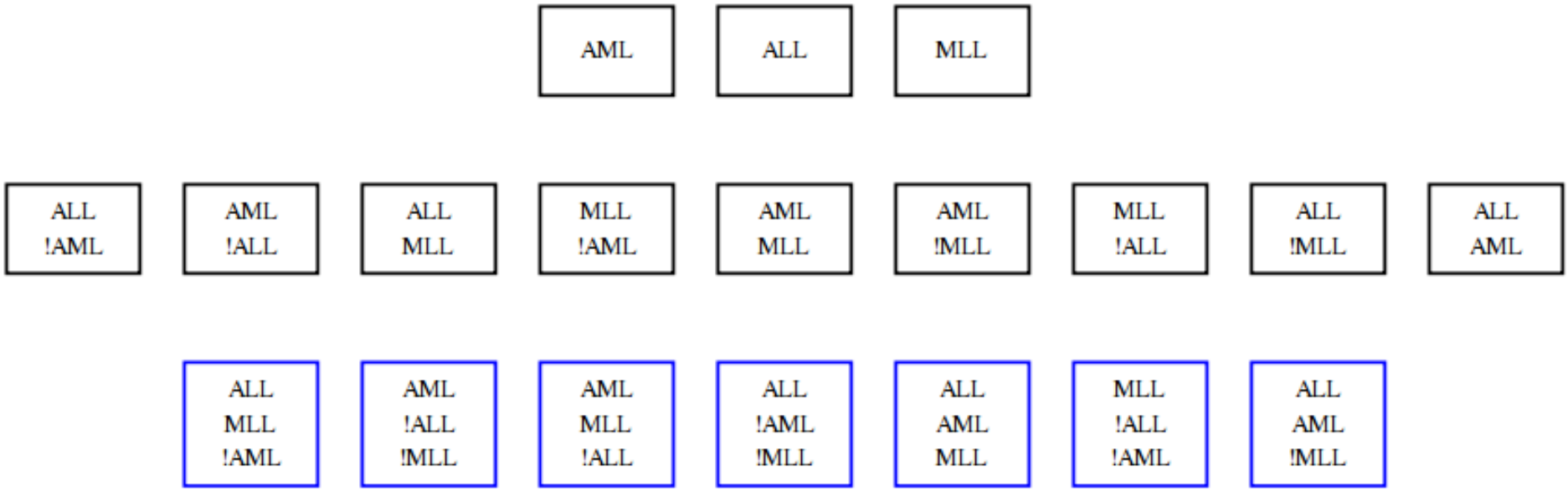
Computing 17 Comparisons



Computing 17 Comparisons

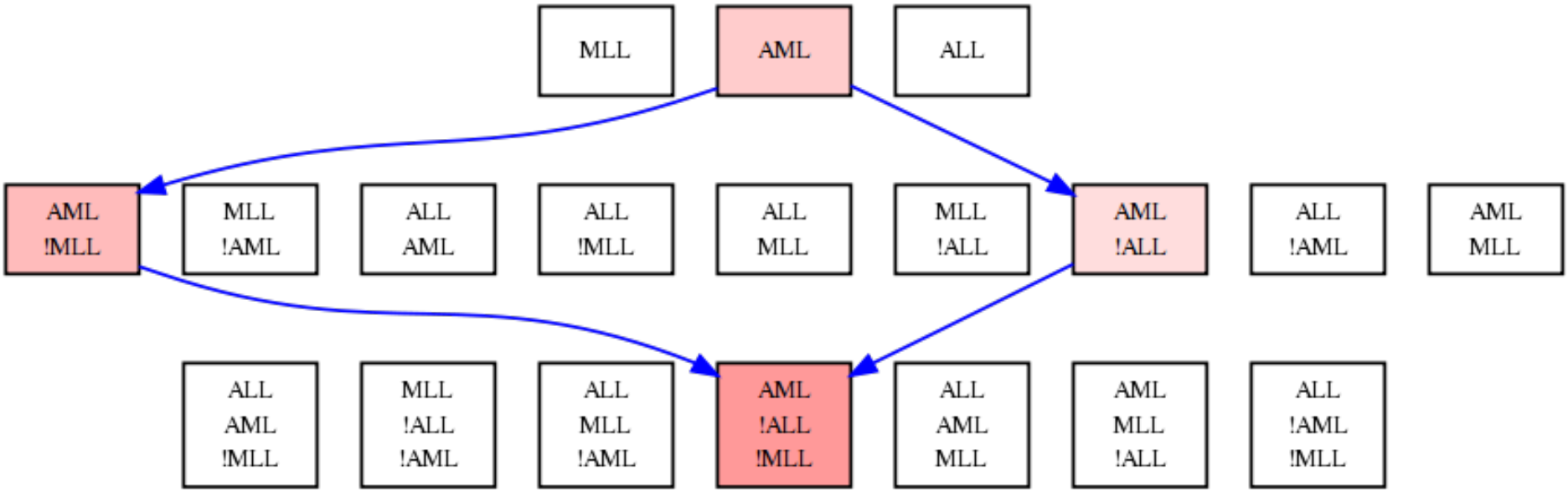


Computing 17 Comparisons



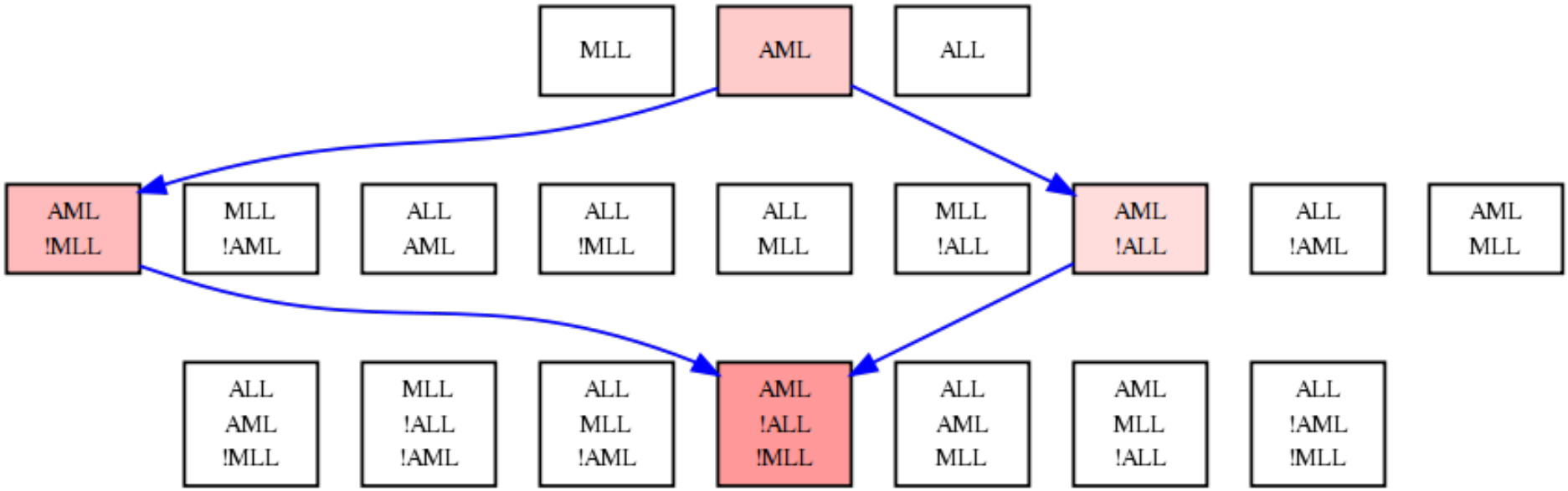
- Each node represents
  1. a boolean conjunction of (possibly negated) conditions and
  2. a network of interactions
- Can compute enrichment of known processes or pathways (Gene Ontology, Netpath, REACTOME, etc.) in each network.

Differential Activation of the Kit Receptor Pathway in AML



- AML: p-value 2×10^{-4}
- AML $\cap$!ALL: p-value 1×10^{-3}
- AML $\cap$!MLL: p-value 6.7×10^{-5}
- AML $\cap$!ALL $\cap$!MLL: p-value 3.5×10^{-7}

Differential Activation of the Kit Receptor Pathway in AML



- AML: p-value 2×10^{-4}
- AML $\cap$!ALL: p-value 1×10^{-3}
- AML $\cap$!MLL: p-value 6.7×10^{-5}
- AML $\cap$!ALL $\cap$!MLL: p-value 3.5×10^{-7}
- c-KIT receptor is activated in almost all subtypes of AML but not in ALL (Reuss-Borst et al., *Leukemia*, 1994, Bene et al., *Blood*, 1998, Schwartz et al., *Leuk Lymphoma.*, 1999).

Challenges in Comparing Arbitrary Numbers of Active Networks

- How can we efficiently compute all combinations?
- How do we identify which combinations are the network legos?
- How do we demonstrate that the network legos we have found are building blocks?

Challenges in Comparing Arbitrary Numbers of Active Networks

- How can we efficiently compute all combinations?
  - Construct binary matrix of interactions vs. active networks and use closed itemset mining algorithms.
- How do we identify which combinations are the network legos?
- How do we demonstrate that the network legos we have found are building blocks?

Challenges in Comparing Arbitrary Numbers of Active Networks

- How can we efficiently compute all combinations?
  - Construct binary matrix of interactions vs. active networks and use closed itemset mining algorithms.
- How do we identify which combinations are the network legos?
  - Compute statistical significance of each combination and exploit DAG structure.
- How do we demonstrate that the network legos we have found are building blocks?

Challenges in Comparing Arbitrary Numbers of Active Networks

- How can we efficiently compute all combinations?
  - Construct binary matrix of interactions vs. active networks and use closed itemset mining algorithms.
- How do we identify which combinations are the network legos?
  - Compute statistical significance of each combination and exploit DAG structure.
- How do we demonstrate that the network legos we have found are building blocks?
  - Define and measure stability and recoverability.

Network Blocks

- Let $\mathcal{A}$ be the set of all active networks.
- A *network block* is a triple $(G, \mathcal{I}, \mathcal{E})$ where
  - $\mathcal{I} \subseteq \mathcal{A}$, $\mathcal{I}$ is non-empty.
  - $\mathcal{E} \subseteq \mathcal{A}$, disjoint from $\mathcal{I}$.
  - $\mathcal{I}$ and $\mathcal{E}$ are inclusion-maximal such that

$$G = \left(\bigcap_{P \in \mathcal{I}} P \right) \bigcap \left(\bigcap_{N \in \mathcal{E}} !N \right)$$

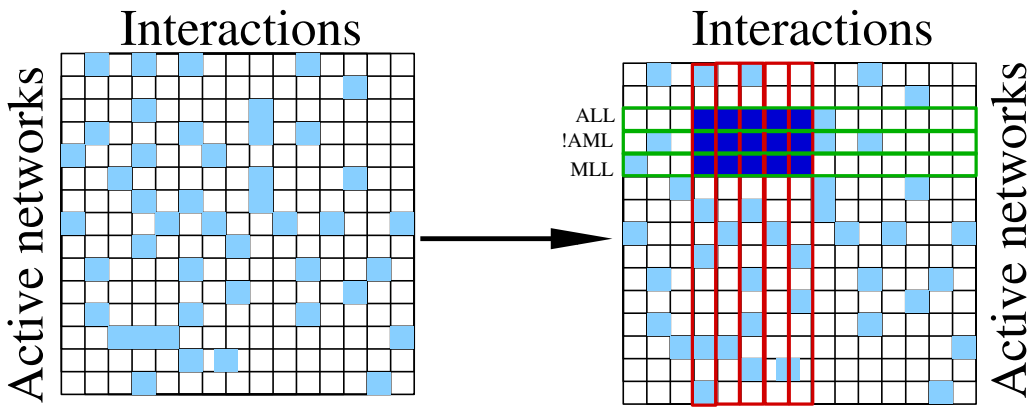
Network Blocks

- Let $\mathcal{A}$ be the set of all active networks.
- A *network block* is a triple $(G, \mathcal{I}, \mathcal{E})$ where
  - $\mathcal{I} \subseteq \mathcal{A}$, $\mathcal{I}$ is non-empty.
  - $\mathcal{E} \subseteq \mathcal{A}$, disjoint from $\mathcal{I}$.
  - $\mathcal{I}$ and $\mathcal{E}$ are inclusion-maximal such that

$$G = \left(\bigcap_{P \in \mathcal{I}} P \right) \bigcap \left(\bigcap_{N \in \mathcal{E}} !N \right)$$

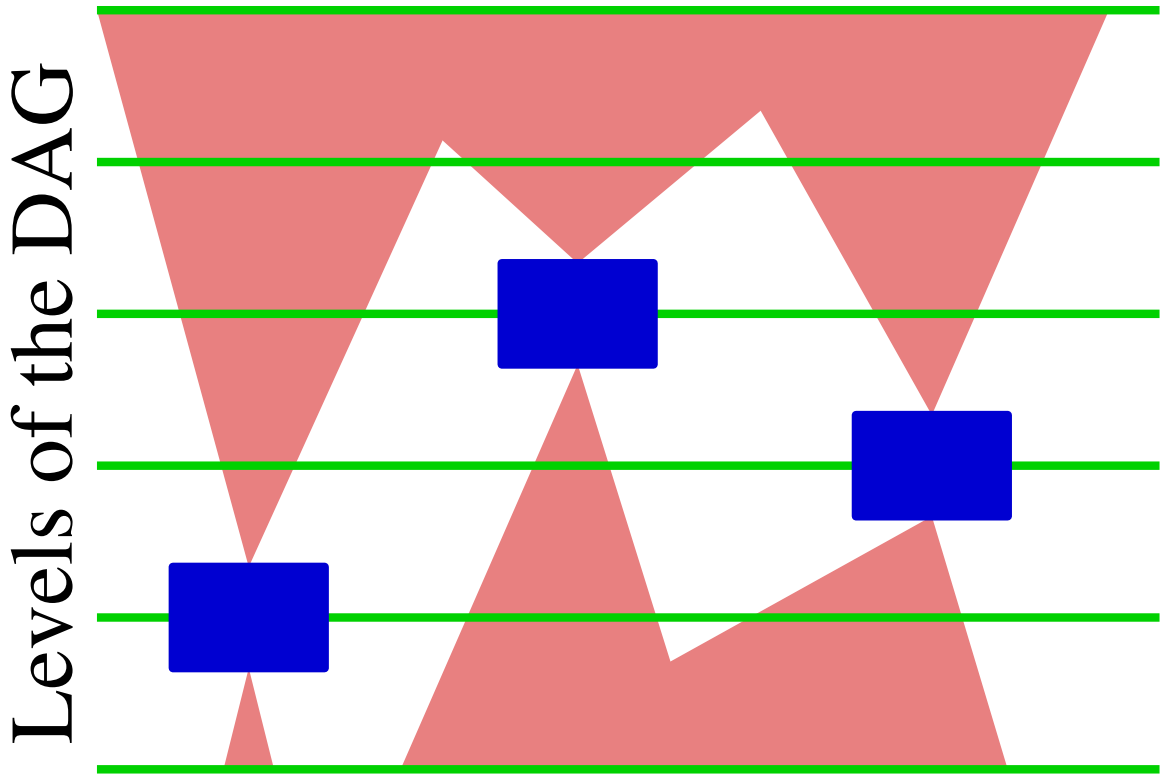
- Partial order exists between network blocks, e.g.,
  - $\mathrm{ALL} < \mathrm{ALL} \cap \mathrm{AML}$.
  - $\mathrm{ALL} \cap \mathrm{MLL} < \mathrm{ALL} \cap \mathrm{MLL} \cap !\mathrm{AML}$.

Efficiently Compute Network Blocks



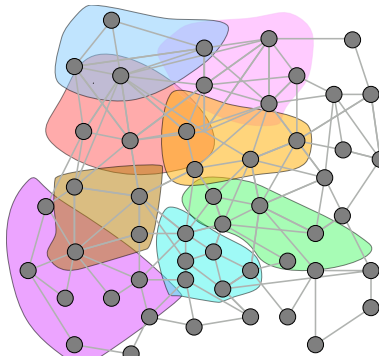
- Construct a binary matrix M whose columns are interactions.
- Represent each active network and its complement in M's rows.
- A *bicluster* is a subset of rows and subset of columns such that M only has 1s in this submatrix.
  - Rows of bicluster $\equiv$ formula.
  - Columns of bicluster $\equiv$ network.
- Compute all closed biclusters in M.
- Connect biclusters in the DAG induced by the partial order.

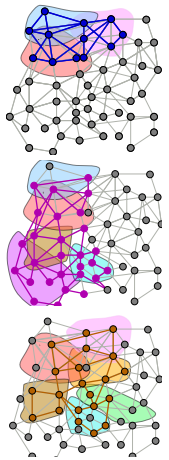
Identify Network Blocks that are Network Legos



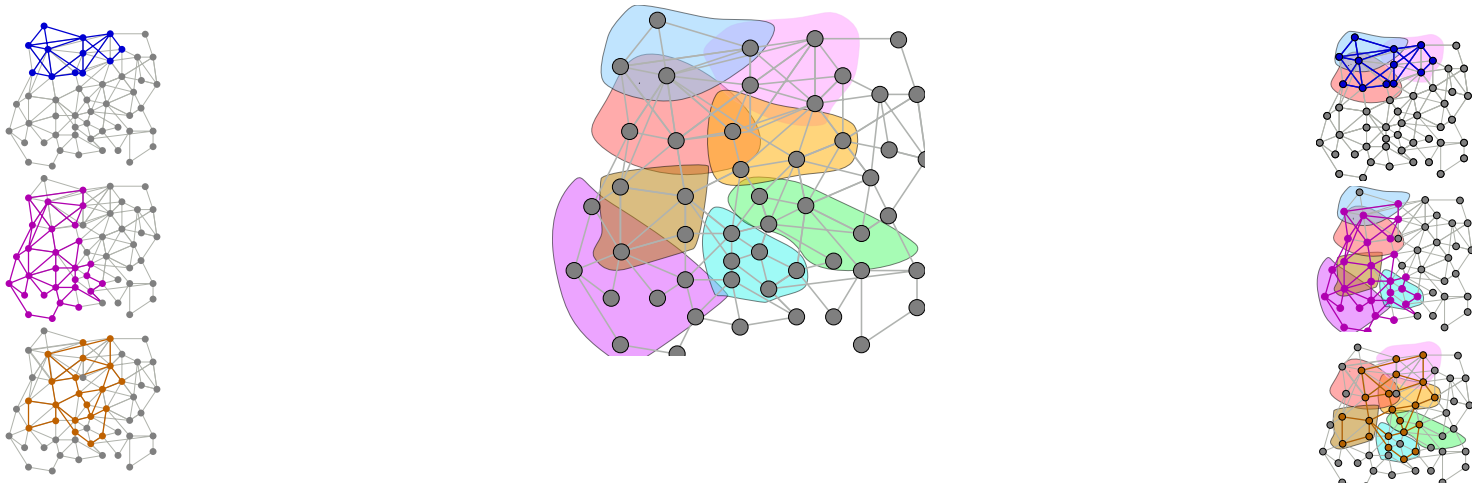
- Assess the statistical significance of each bicluster by simulation.
- B is a *network lego* if it is more significant than any of its ancestors or descendants in the DAG.

Show that Network Legos are Building Blocks



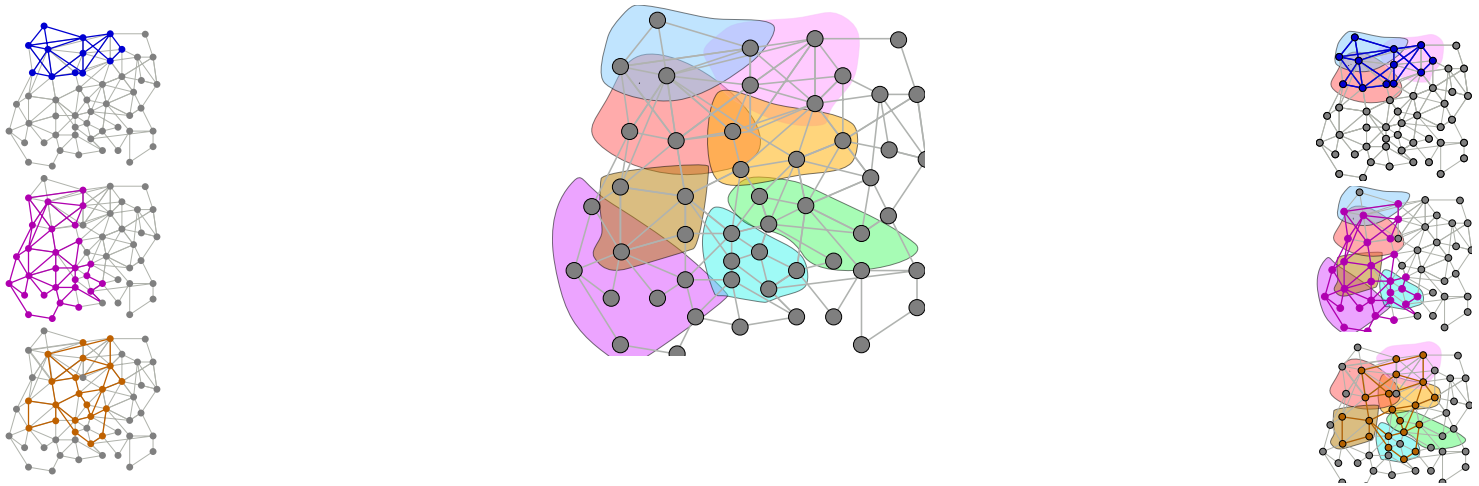


Show that Network Legos are Building Blocks



- Stability
  - Remove each active network and recompute network legos.
  - For each original network lego, compute the fraction of leave one out datasets for which the network lego occurs with at least t% fidelity.

Show that Network Legos are Building Blocks



- Stability
  - Remove each active network and recompute network legos.
  - For each original network lego, compute the fraction of leave one out datasets for which the network lego occurs with at least t% fidelity.
- Recoverability
  - Compute the union of network legos.
  - Measure the size of the intersection of each active network with union.

Analysis of Human Stress Data

- Human protein-protein interaction network with 9243 proteins and 31000 interactions.
  - PPIs from (Ramani et al., Genome Biology, 2005; Rual et al., Nature, 2005; Stelzl et al., Cell, 2005).
- 13 distinct stresses applied to human cells (Murray et al., Mol. Bio. Cell, 2004).
  - Stress conditions include heat shock, oxidative stress, cell cycle arrest, and crowding.
  - Two cell types: WI38 Fibroblasts and HeLa.
- Murray et al. note that each stress elucidated a unique response.

Human Stress Results

- 13 stresses and their active networks yielded 444201 closed biclusters.
- 143 biclusters are network legos.
- The network legos contained between 165 and 1148 proteins.
- Each network lego has 95% stability.
- The network legos provide better than 86% recoverability for all active networks.
- We recovered 11 active networks at 100%.

#conditions	5	6	7	8	9	10	11	12
#legos	1	6	10	36	34	20	28	8

Human Stress Results without Cell Cycle Arrest Treatment

- The active networks for cell cycle arrest treatments contain interactions that are distinct compared to those in active networks for other treatments.
- Remaining 11 stresses yielded only 15 network legos.
- The network legos provide better than 71% recoverability for all active networks.
- We recovered five active networks at 100%.
- Each formula contained at least 7 active networks.

WI38 Response to Menadione and DTT

- One network lego contained endoplasmic reticulum stress and oxidative stress to fibroblasts in included set.
- Other stresses in network lego appeared in excluded set.
- This network lego is the only one enriched in
  - KEGG "cell cycle" pathway (p-value 3×10^{-30}),
  - REACTOME "G1 to S transition" (p-value 2.3×10^{-24}), and
  - targets of the E2F1 transcription factor (p-value 8×10^{-13}).
- In response to these two stresses, fibroblasts shut down the cell cycle far more aggressively than HeLa cells do.

Software

- All our software is available under the GNU GPL.
- Developed on Linux systems.
- Command-line interface.
- Visit http://bioinformatics.cs.vt.edu/~murali/software

Our Contributions

- Combined representation of biological processes using formulae and network legos.
- A formula relates different cellular states or perturbations by explicitly denoting their participation via intersections and complements.
- Each network lego corresponds to a functional module of coherently interacting genes in the wiring diagram.
- Network legos serve as building blocks of active networks.

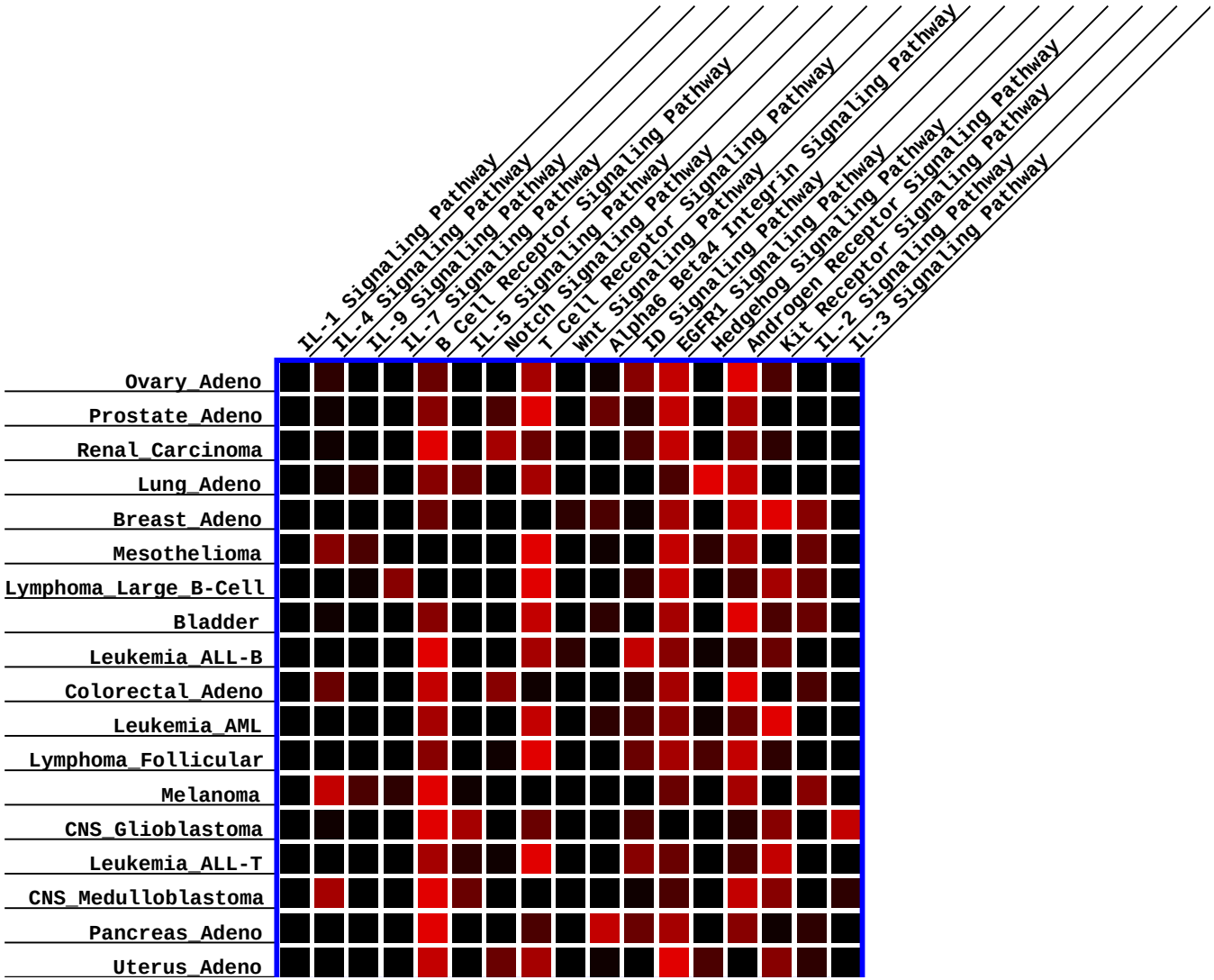
Acknowledgments

- Corban Rivera, Ph.D. student in my group, `cgrivera@vt.edu`.
- Funding from the ASPIRES programme and the Institute for Critical Technologies and Applied Science at Virginia Tech.

Future Work

- Explore network legos in the context of a larger compendium of cellular stresses.
- Develop an algorithm to directly compute network legos without searching the space of all active network combinations.
- Determine rules and grammar for combining network legos into active networks.

Compendium Approach

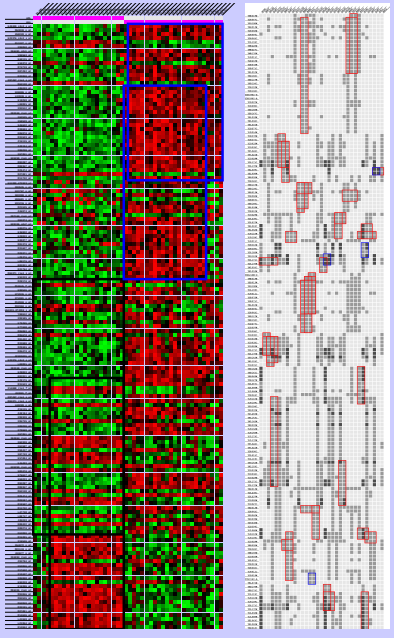


Computational Systems Biology

T. M. Murali, murali@cs.vt.edu, http://bioinformatics.cs.vt.edu/~murali

Biologically-Interpretable Disease Classification

- Detect cancer-specific gene expression signatures.
- Use these signatures to classify distinct cancers.
- An *xMotif* is a subset of conditions and a subset of genes such that each gene is co-expressed in the selected samples.
- Develop a novel nearest-neighbour classifier.
- Classifier is biologically interpretable: xMotifs are enriched in functions relevant to cancers.
- Classifier achieves performance comparable to a support vector machine.



Combinatorial Control of Gene Expression

- Multiple transcription factors often regulate the expression of a group of genes.
- Combinatorial regulation can change dramatically with external stimulations or perturbations.
- We have modified our methods for disease classification to provide insights into combinatorial control.

Research Programme

Rationale

- The genome sequences of 100s of organisms are available.
- High-throughput biological assays provide a dazzling variety of information about the cell.
- Such experiments yield massive quantities of information.
- We can begin considering the cell as an assemblage of interconnected modules of interacting molecules.

Approach

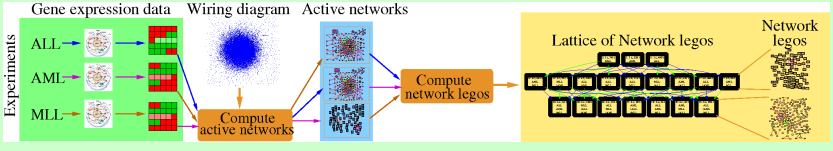
- Develop methods to automatically compute modules of coherently interacting molecules.
- Integrate different types of biological data using principles of graph theory, discrete algorithms, data mining, and machine learning.
- Compare cellular states and responses across different organisms, diseases, stresses, and stimuli.

Applications

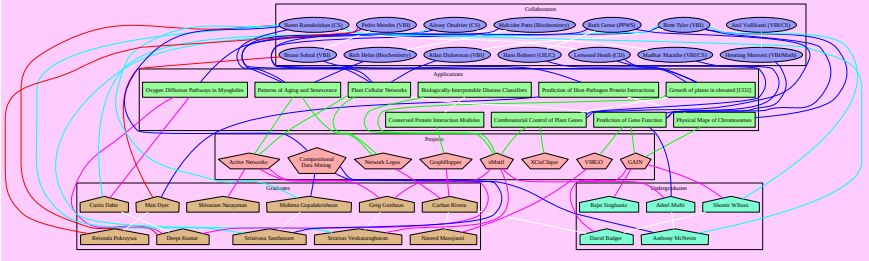
- Find networks activated in the cell in cancer and related diseases.
- Predict protein interactions that enable a pathogen to invade a host.
- Develop biologically-interpretable disease classifiers.
- Detect cryptic components of pathways dis-regulated in human and plant diseases.

Active Networks and Network Legos

- A cell's response to external stimuli or stress is often tuned exquisitely to the stress.
- However, fundamental response networks are common to multiple stresses and are conserved across organisms.
- Our system computes active networks for different stresses and combines them to find network legos.
- An *active network* is a network of interactions activated in the cell in response to a stress.
- A *network lego* is a group of coherently-interacting genes that are completely contained in the active networks of some stresses but are completely disjoint from the active networks for other stresses.
- The network legos we compute are statistically significant, stable, and recoverable.

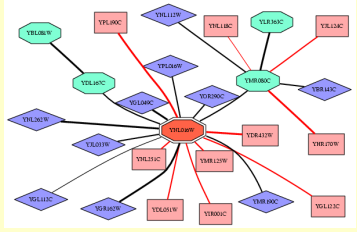


Collaborations



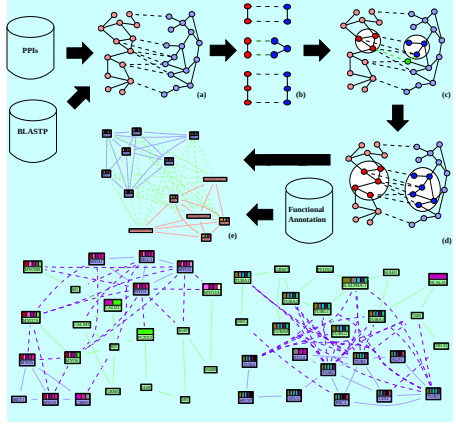
Precise and Robust Prediction of Gene Functions

- Sequenced genomes contains 100,000s of genes.
- Functional roles of 40% of the genes are unknown!
- Another 20% have poorly-known functional roles.
- GAIN predicts gene functions using *Functional Linkage Networks*.
- VIRGO webserver enables biologist to use GAIN to obtain gene function predictions for system of interest.
- GAIN provides informative propagation diagrams.
- GAIN is the only algorithm mathematically guaranteed to make hierarchically consistent predictions.
- *Time validation* of GAIN's predictions for baker's yeast and human reveals that GAIN's predictions are experimentally verified for at least 50% of genes.
- Validated predictions involve several biological processes implicated in cancers and other human diseases.



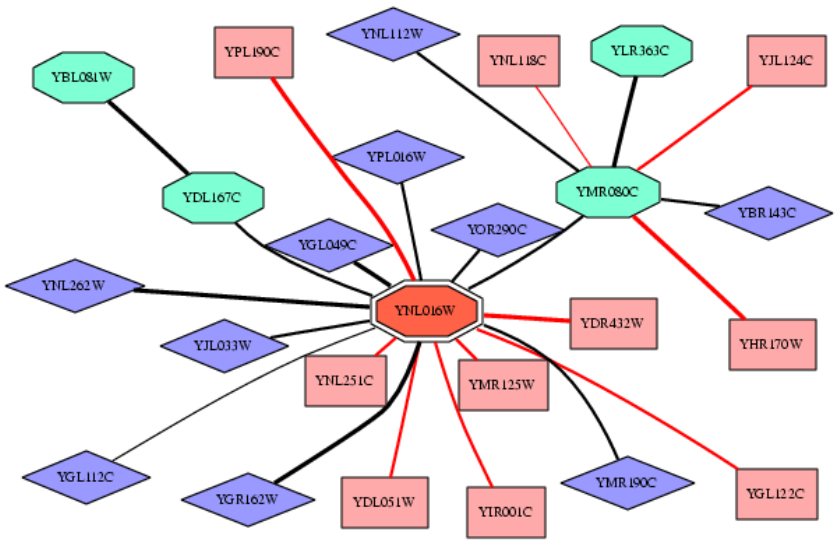
GraphHopper

- Protein-protein interactions (PPIs) for a number of organisms are available.
- A *Conserved Protein Interaction Module* (CPIM) is a group of interacting proteins whose interactions are conserved in more than one species.
- Our *GraphHopper* algorithm detects numerous, biologically-significant, and topologically diverse CPIMs in comparisons between human, fruitfly, and worm interaction networks.



Precise and Robust Prediction of Gene Functions

- Functions of 40% of the genes in sequenced genomes are unknown.
- GAIN predicts gene functions using *Functional Linkage Networks*.
- VIRGO webserver enables biologist to use GAIN to obtain predictions for system of interest.
- *Time validation* for baker's yeast and human reveals that GAIN's predictions are experimentally verified for at least 50% of genes.



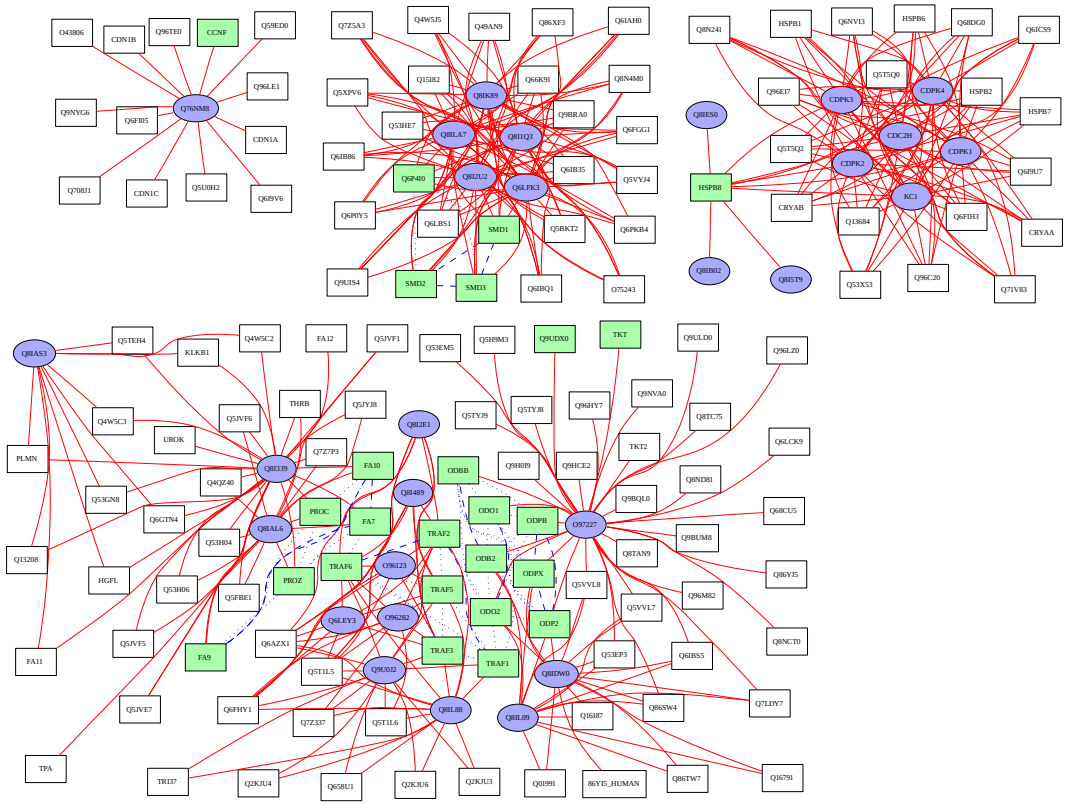
Karaoz, Murali et al., *PNAS* 2004; Massjouni, Rivera, and Murali, *NAR* Web server issue, 2006; Murali, Wu, and Kasif, *Nature Biotechnology*, 2006.

Biologically-Interpretable Disease Classification

- Detect cancer-specific gene expression signatures.
- Use these signatures to classify distinct cancers.
- Novel nearest-neighbour classifier based on biclusters.
- Interpretability: our biclusters are enriched in functions relevant to diseases.
- Classifier achieves performance comparable to a support vector machine.

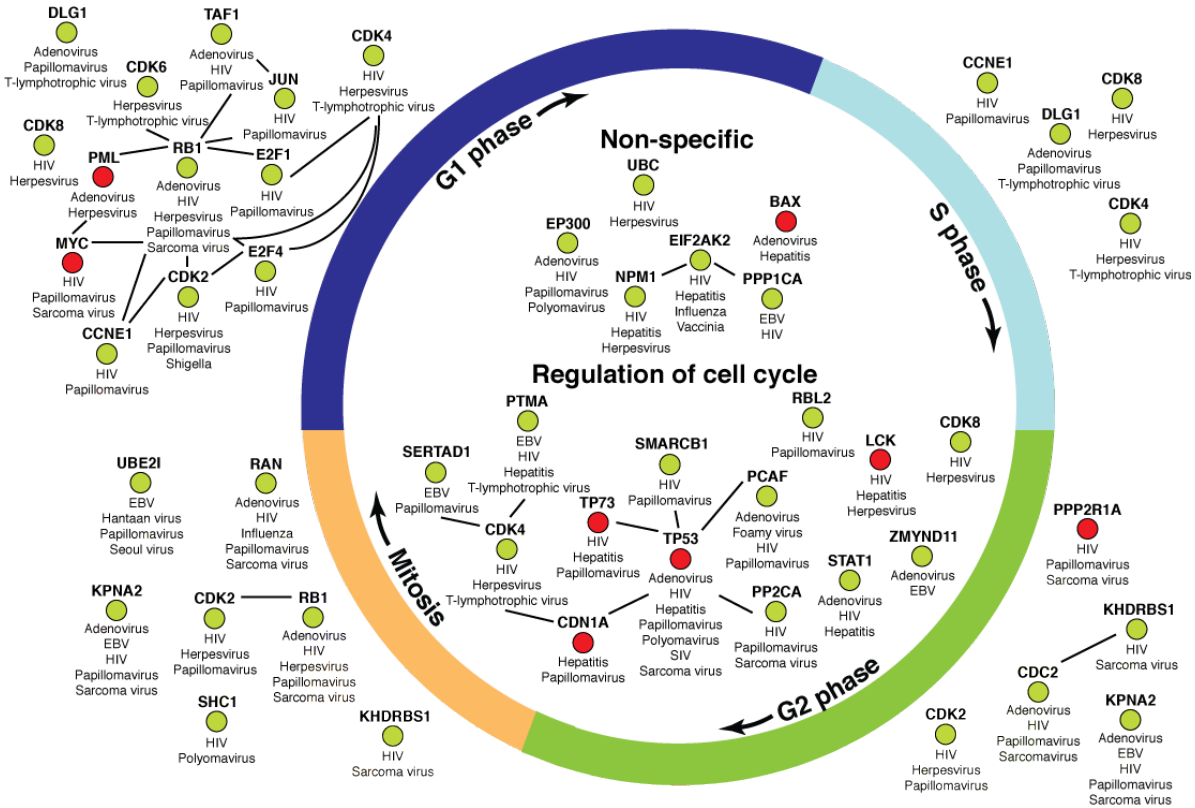
Host-Pathogen Protein Interaction Networks

- Use domain-pair occurrences to predict interactions between host and pathogen proteins Dyer, Murali, and Sobral, *ISMB* 2007.
- Results predict links between *P. falciparum* membrane and dense granule proteins and subtilases and human blood coagulation proteins.



Host-Pathogen Protein Interaction Networks

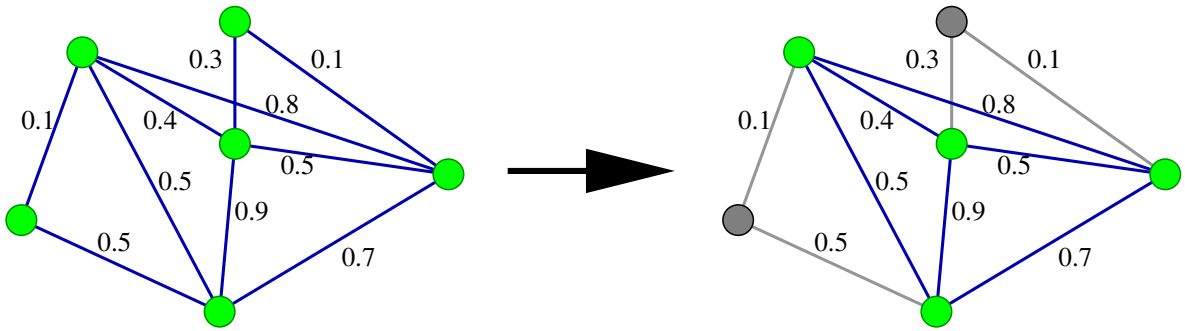
- Study landscape of human proteins interacting with viruses and other and pathogens Dyer, Murali, and Sobral, *PLoS Pathogens*, 2008.
- Pathogens have evolved to interact with hubs and bottlenecks in the human PPI network.
- Many viruses share common pathways to infection.



Algorithmic Ingredients: Active Networks

(i) Assign Pearson's correlation as the interaction weight

(ii) Compute dense subgraphs



- Compute the Pearson's correlation coefficient of the expression profiles of the interacting genes.
- Search for pockets of concerted activity using an algorithm for finding dense subgraphs.

Assessing Statistical Significance of a Bicluster

- Suppose a bicluster B has n included and c excluded active networks.
  1. Pick n active networks and the complements of c active networks repeatedly at random, compute the number of interactions induced by this combination, and build a distribution of the number of interactions.
  2. Set the p-value of B to be the fraction of random biclusters with more interactions than B.