ActiveNetworks
Cross-Condition Analysis
of Functional Genomic Data

T. M. Murali

April 18, 2006
Motivation: Manual Systems Biology

- Biologists want to study a favourite stress, e.g., oxidative stress or desiccation tolerance.
- Measure gene expression, apply clustering algorithms, and find genes whose expression level change in response to the stress.
Biologists want to study a favourite stress, e.g., oxidative stress or desiccation tolerance.

Measure gene expression, apply clustering algorithms, and find genes whose expression level change in response to the stress.

Motivation: Manual Systems Biology

Redescription R5
Heat Shock, 30 min \( \leq -1 \) \( \rightarrow \) T\(_2\) vs. T\(_1\) \( \geq -5 \) AND NOT T\(_2\) vs. T\(_1\) \( \geq -1 \)

Redescription R5 Gene List
ARO4, ASN1, CLN2, GAS3, HEM13, HIS1, IMD4, PHO3, RPL-7A, 7B, 13A, 17B, 27B, 40B, RPS-0B, 9B, 10A, 16B, 22B, 26B, SAH1, SAM1, SUN4, TEF4, TPO2, URA7, UTR2, YHB1, YBR238C, YER156C, YFR055W, YOR309C

Can we automate this process?

T. M. Murali April 18, 2006 ActiveNetworks
Redescription R5
Heat Shock, 30 min \( \leq -1 \) \( \rightarrow \) \( T_2 \) vs. \( T_1 \) \( \geq -5 \) AND NOT \( T_2 \) vs. \( T_1 \) \( \geq -1 \)

Redescription R5 Gene List
ARO4, ASN1, CLN2, GAS3, HEM13, HIS1, IMD4, PHO3, RPL-7A, 7B, 13A, 17B, 27B, 40B, RPS-0B, 9B, 10A, 16B, 22B, 26B, SAH1, SAM1, SUN4, TEF4, TPO2, URA7, UTR2, YHB1, YBR238C, YER156C, YFR055W, YOR309C

Can we automate this process?
Requirements for Automation

- Wiring diagram of the cell: protein-protein interactions, metabolic pathways, transcriptional regulatory networks, ...
- Measurement of molecular profiles (gene expression, protein expression, metabolite levels) under different conditions or cell states.
- Algorithms for combining these types of information.
High-throughput Biology Provides Wiring Diagram

- Large amounts of information on different types of cellular interactions are now available.
- Protein-protein interactions: genome-scale yeast 2-hybrid experiments, in-vivo pulldowns of protein complexes.
- Transcriptional regulatory networks: ChIP-on-chip experiments yield protein-DNA binding data.
- Metabolic networks: databases culled from the literature (KEGG).
- Techniques that extract interactions automatically from abstracts.
S. cerevisiae Wiring Diagram

- **Physical network**
  - 15,429 *protein-protein interactions* from the Database of Interacting Proteins (DIP).
  - 5869 *protein-DNA interactions* (Lee et al., Science, 2002).
  - 6,306 *metabolic interactions* (proteins operate on at least common metabolite) based on KEGG.

- **Genetic network**
  - 4,125 *synthetically lethal/sick interactions* (Tong et al., Science, 2004).
  - 687 synthetically lethal interactions (MIPS).

- **Overall network** has 32,416 (27,604 physical and 4,812 genetic) interactions between 5601 proteins (Kelley and Ideker, Nature Biotech., 2005).
Challenges in Utilising the Wiring Diagram

- Networks are large; they contain tens of thousands of interactions.
- High-throughput experiments contain many errors.
- Networks are incomplete; experiments are expensive and have biases.
Challenges in Utilising the Wiring Diagram

- Networks are large; they contain tens of thousands of interactions.
- High-throughput experiments contain many errors.
- Networks are incomplete; experiments are expensive and have biases.
- A biologist wants to explore and analyse system of interest.
- How do we zoom into the appropriate parts of the wiring diagram?
**ActiveNetworks**

**ActiveNetwork**: *network of interactions activated in response to a stress or in a particular condition.*

1. Overlay molecular profile for a particular stress on wiring diagram to obtain **ActiveNetwork** for that stress.
2. Combine computed **ActiveNetworks** for each stress to find
   2.1 **ActiveNetwork** common to multiple stresses.
   2.2 **ActiveNetwork** unique to a particular stress or group of stresses.
**ActiveNetworks**

**ActiveNetwork:** network of interactions activated in response to a stress or in a particular condition.

1. Overlay molecular profile for a particular stress on wiring diagram to obtain ActiveNetwork for that stress.

2. Combine computed ActiveNetworks for each stress to find
   2.1 ActiveNetwork common to multiple stresses.
   2.2 ActiveNetwork unique to a particular stress or group of stresses.
The ActiveNetworks pipeline

- Protein–protein interaction networks
- Transcriptional regulatory networks
- Metabolic pathways

Experiments:
- Heat shock
- Cold shock
- Carbon starvation
- Desiccation

ActiveNetworks for each condition:
- Universal network
- ActiveNetwork computation
- ActiveNetwork mining

Condition–specific ActiveNetworks

Cross–condition ActiveNetworks

Hypotheses
Weight of an interaction is the Pearson correlation between the expression profiles of the interacting genes.

Weight $\equiv$ “activity” level of the interaction.
Weight of an interaction is the Pearson correlation between the expression profiles of the interacting genes.

Weight \equiv \text{“activity” level of the interaction.}

Discard interactions based on a threshold.

- Unsatisfactory since we test each interaction individually.
Weight of an interaction is the Pearson correlation between the expression profiles of the interacting genes.

Weight $\equiv$ “activity” level of the interaction.

Discard interactions based on a threshold.

* Unsatisfactory since we test each interaction individually.

We find the most highly active subnetwork.
Defining Highly-Active Subnetworks

How do we measure the activity/weight of a subnetwork?
Defining Highly-Active Subnetworks

- How do we measure the activity/weight of a subnetwork?
- Sum or average of edge weights?
Defining Highly-Active Subnetworks

▶ How do we measure the activity/weight of a subnetwork?
▶ Sum or average of edge weights?
▶ The \textit{density} of a network with $n$ nodes is the total weight of the edges divided by $n$.
▶ Problem: Compute the subnetwork with highest density.
Computing Most Dense Subnetwork

- $O(n^3)$ time network flow-based approach gives optimal result (Gallo, Grigoriadis, Tarjan, SIAM J. Comp, 1989).
- Can also be solved by linear programming.
Computing Most Dense Subnetwork

- **Greedy algorithm:**
  - Weight of a node $\equiv$ total weight of incident edges.
  - Repeatedly delete nodes with the smallest weight.
  - Keep track of density of remaining network.
  - Return the most dense subnetwork.

![Graph showing weight of edges between nodes with weights 0.1, 0.3, 0.4, 0.5, 0.7, 0.8, 0.9.](image)
Computing Most Dense Subnetwork

- **Greedy algorithm:**
  - Weight of a node $\equiv$ total weight of incident edges.
  - Repeatedly delete nodes with the smallest weight.
  - Keep track of density of remaining network.
  - Return the most dense subnetwork.

Computed subnetwork is at least half as dense as the most dense subnetwork (Charikar, Proc. APPROX, 2000).
Computing Most Dense Subnetwork

Greedy algorithm:
- Weight of a node $\equiv$ total weight of incident edges.
- Repeatedly delete nodes with the smallest weight.
- Keep track of density of remaining network.
- Return the most dense subnetwork.

Computed subnetwork is at least half as dense as the most dense subnetwork (Charikar, Proc. APPROX, 2000).
Computing Multiple Dense Subnetworks

Repeat
1. Apply greedy algorithm to compute most dense subnetwork.
2. Remove edges of computed subnetwork from the network.

Until remaining network has density less than the original network.

Output is a sequence of decreasingly dense subnetworks that can share nodes but not edges.
Advantages of Dense Subnetworks

- Uses no parameters.
- Avoid inclusion of interactions that appear active due to noise.
- Relatively weakly correlated interactions can reinforce each other.
Example of an ActiveNetwork
Example of an ActiveNetwork
Further Analysis of an ActiveNetwork

- Visualise the network (Graphviz package) and the gene expression profiles.
- Measure functional enrichment.
  - Use hypergeometric distribution to calculate the significance of functions enriched in an ActiveNetwork.
  - Use Bonferroni correction to adjust for testing multiple hypotheses.
AA Starvation: ActiveNetworks

T. M. Murali

April 18, 2006
ActiveNetworks for Multiple Stresses

**ActiveNetwork**: network of interactions activated in response to a stress or in a particular condition.

1. Overlay molecular profile for a particular stress on universal network to obtain ActiveNetwork for that stress.
2. Combine computed ActiveNetworks for each stress to find
   2.1 ActiveNetwork common to multiple stresses.
   2.2 ActiveNetwork unique to a particular stress or group of stresses.
ActiveNetworks for Multiple Stresses

**ActiveNetwork:** network of interactions activated in response to a stress or in a particular condition.

1. Overlay molecular profile for a particular stress on universal network to obtain ActiveNetwork for that stress.
2. Combine computed ActiveNetworks for each stress to find
   2.1 ActiveNetwork common to multiple stresses.
   2.2 ActiveNetwork unique to a particular stress or group of stresses.
Comparative ActiveNetwork Analysis

Richard and Malcolm want to compare desiccation with other conditions to find similarities and differences. Can we automate this process?
Richard and Malcolm want to compare desiccation with other ActiveNetworks to find similarities and differences.

Can we automate this process?
A “conserved” **ACTIVENETWORK** is a set of conditions and a set of interactions, such that each interaction appears in the **ACTIVENETWORK** for each condition.
Computing Conserved ActiveNetworks

Construct a 0-1 interaction-by-condition matrix.
### Computing Conserved ActiveNetworks

Construct a 0-1 interaction-by-condition matrix.

A "conserved" **ActiveNetwork** is a set of conditions and a set of interactions, such that each interaction appears in the ActiveNetwork for each condition.

We can compute a conserved ActiveNetwork using techniques for finding itemsets or biclusters.
Construct a 0-1 interaction-by-condition matrix.

A “conserved” ActiveNetwork is a set of conditions.
Construct a 0-1 interaction-by-condition matrix.

A “conserved” ActiveNetwork is a set of conditions and a set of interactions,
Computing Conserved ActiveNetworks

- Construct a 0-1 interaction-by-condition matrix.
- A “conserved” ActiveNetwork is a set of conditions and a set of interactions, such that each interaction appears in the ActiveNetwork for each condition.
Construct a 0-1 interaction-by-condition matrix.

A “conserved” ActiveNetwork is a set of conditions and a set of interactions, such that each interaction appears in the ActiveNetwork for each condition.

We can compute a conserved ActiveNetwork using techniques for finding itemsets or biclusters.
A “large” submatrix of 1’s is a frequent itemset.

Such a submatrix is a special case of a bicluster in gene expression data.
A “large” submatrix of 1’s is a frequent itemset.

Such a submatrix is a special case of a bicluster in gene expression data.

We use the *apriori* algorithm for finding all maximal (closed) itemsets (Agrawal and Srikant 1995) and the *xMotif* algorithm for finding large biclusters (Murali and Kasif, 2003).
Example of a Cross-Condition ActiveNetwork

- Common to “Alternative carbon sources.” “DTT treatment” and “Growth in YPD culture.”
Example of a Cross-Condition ActiveNetwork

Common to "Alternative carbon sources." and "GDT treatment."

Alternative carbon-source
ethanol vs. reference pool car-1
galactose vs. reference pool car-1
glucose vs. reference pool car-1
mannose vs. reference pool car-1
raffinose vs. reference pool car-1
sucrose vs. reference pool car-1
YPD ethanol vs reference pool car-2
YPD fructose vs reference pool car-2
YPD galactose vs reference pool car-2
YPD glucose vs reference pool car-2
YPD mannose vs reference pool car-2
YPD raffinose vs reference pool car-2
YPD sucrose vs reference pool car-2

2.5mM DTT 005 min dtt-1
2.5mM DTT 015 min dtt-1
2.5mM DTT 000 min dtt-1
2.5mM DTT 045 min dtt-1
2.5mM DTT 060 min dtt-1
2.5mM DTT 090 min dtt-1
2.5mM DTT 120 min dtt-1
2.5mM DTT 180 min dtt-1

YPD 2 h ypd-2
YPD 4 h ypd-2
YPD 6 h ypd-2
YPD 8 h ypd-2
YPD 10 h ypd-2
YPD 12 h ypd-2
YPD 1 d ypd-2
YPD 2 d ypd-2
YPD 3 d ypd-2
YPD 5 d ypd-2

YPD stationary phase 2 h ypd-1
YPD stationary phase 4 h ypd-1
YPD stationary phase 6 h ypd-1
YPD stationary phase 8 h ypd-1
YPD stationary phase 12 h ypd-1
YPD stationary phase 1 d ypd-1
YPD stationary phase 2 d ypd-1
YPD stationary phase 3 d ypd-1
YPD stationary phase 5 d ypd-1
YPD stationary phase 7 d ypd-1
YPD stationary phase 13 d ypd-1
YPD stationary phase 22 d ypd-1
YPD stationary phase 26 d ypd-1

T. M. Murali April 18, 2006
ActiveNetworks
ActiveNetworks provide an integrated view of multi-modal universal networks and measurements of molecular profiles.

Compute single stimulus ActiveNetworks using dense subgraphs.

Compare and contrast ActiveNetworks for different stimuli using frequent itemsets.

Automatic extraction of network modules and legos from large scale data.

Promises system-level insights from comparisons between different conditions, disease states, or species.
Closing the Loop

Protein–protein interaction networks
Transcriptional regulatory networks
Metabolic pathways

Heat shock
Cold shock
Carbon starvation
Desiccation

Universal network
ActiveNetwork computation
ActiveNetwork mining

Condition–specific ActiveNetworks
Cross–condition ActiveNetworks

Hypotheses

T. M. Murali April 18, 2006 ActiveNetworks
Project Members

▶ Greg Grothaus
▶ Deept Kumar
▶ Maulik Shukla
▶ Graham Jack
▶ Corban Rivera

▶ Richard Helm
▶ Malcolm Potts
▶ Naren Ramakrishnan
Future Research: Modelling and Algorithmic Improvements

- Integrate other types of data: metabolic measurements, protein expression.
- Explicitly incorporate expression level of a gene.
Future Research: Applications

- **ActiveNetworks** in cancer: integrate gene expression data and protein interaction networks.
- Compare oxidative stress networks across kingdom boundaries (yeast, *Arabidopsis thaliana*, malaria parasite, *P. sojae*).
- Cross-stress networks in *Arabidopsis thaliana*.
- Redox signalling in various plant species.
Related Research

- Discovering regulatory and signalling circuits in molecular interaction networks, Ideker et al. ISMB 2002.
- Physical network models and multi-source data integration, Yeang and Jakkola, RECOMB 2003.
- Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data, Tanay et al., PNAS, March 2004.