CS 6104: Projects, Topics, and Schedule

T. M. Murali

September 9, 2004

Choice of Topics and Weeks

Sep 9	Topics and schedule
Sep 16	Diagnostic genes, stem cells: Andrew, Eric, John
Sep 23	Cancer classification: Deept, Nilanjan
Sep 30	Outcome prediction: Greg, Jonathan
Oct 7	Comparative systems biology: Chaitanya, Kiran, Rob, Shenghua
Oct 14	Chemical genomics and pharmacogenomics: Corban, Shivaram
Oct 21	Genome variation and disease (invited lecture)
Oct 28	Mid-term project reviews
Nov 4	Functional annotation: Satish, Venkat
Nov 11	Malaria (invited lecture)
Nov 18	RNA interference: Rajat, Shenghua, Srinivas
Dec 2	Gene and literature datasets
Dec 9	Project presentations
Dec 16	Project presentations

Projects

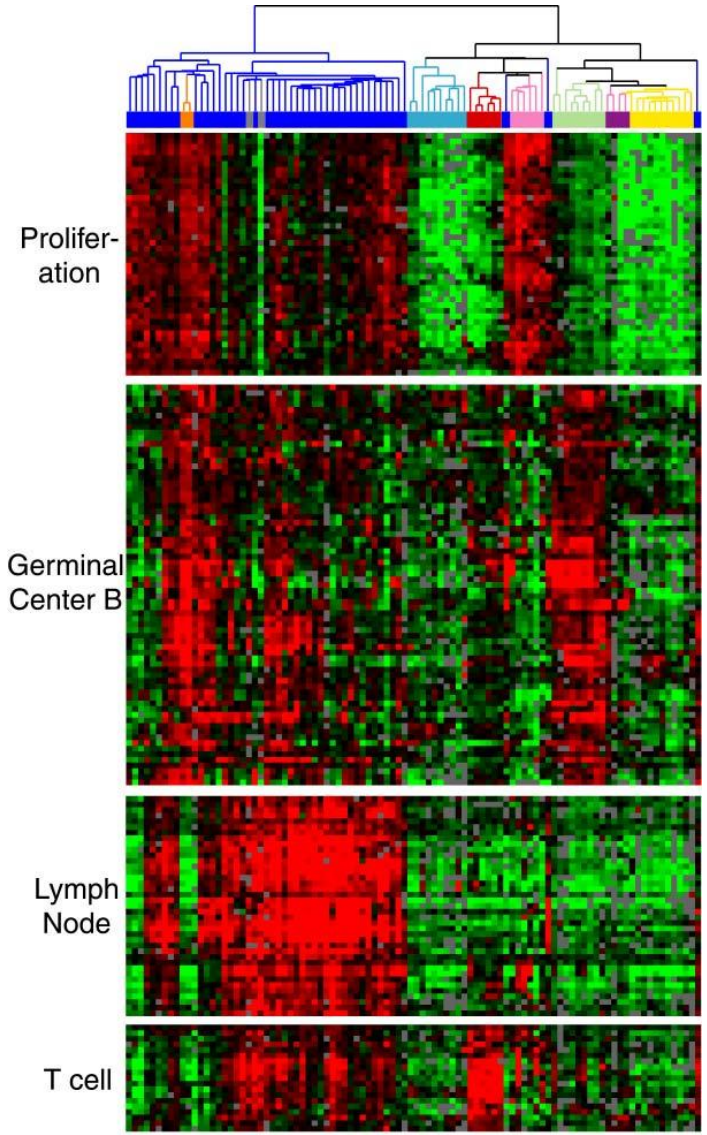
- Gene expression analysis using biclustering
  - Cross-condition gene expression signatures: related to cancer, treatment outcome, stem cells.
  - Bicluster database and web-server.
- Cellular network analysis using ActiveNetworks
  - ActiveNetworks in various cancers.
  - Cross-species systems biology: ActiveNetworks common to and different between various organisms.

Projects Continued

- Whole genome functional annotation
  - Improvements to the GAIN algorithm.
  - Functional annotation web server: Web-server and database for querying and probing functional linkage networks
  - Cross-species functional annotation
  - Human genome: Annotation of the human genome using HPRD and/or (cancer) gene expression data sets
  - Malaria Functional annotation of Malaria genes using gene expression data.
- Association of SNPs with disease
- Prediction of microRNA targets

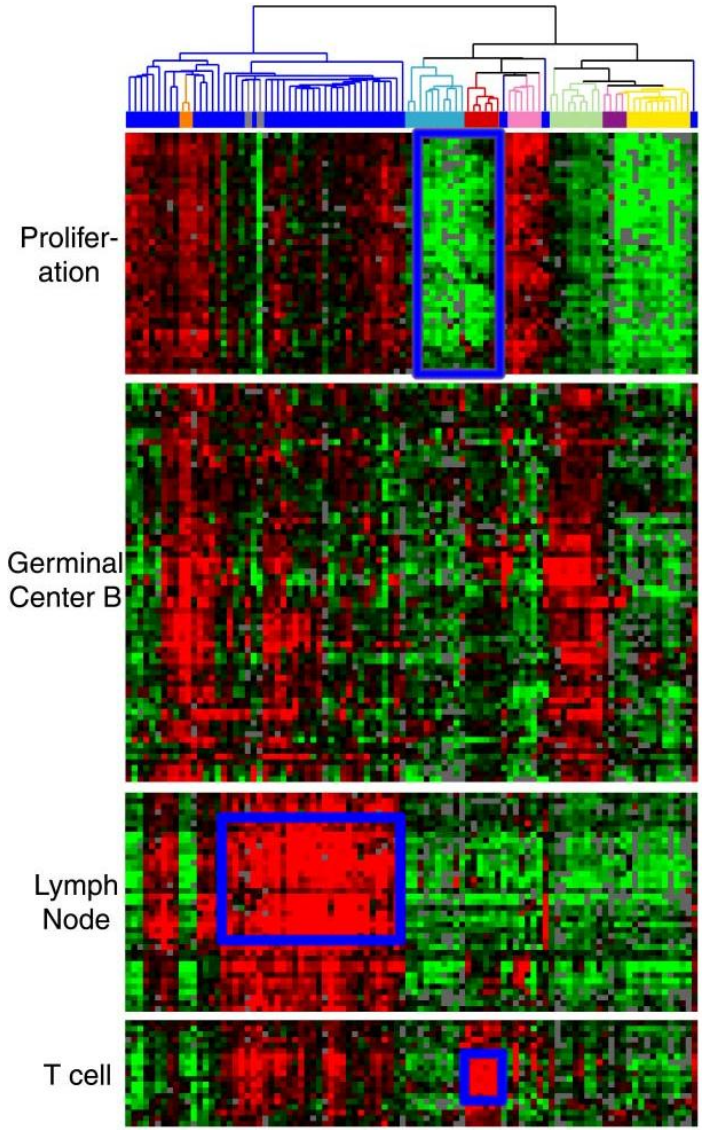
Motivation for Biclusters

- Clustering: Reveals coarse patterns in the data.



Motivation for Biclusters

- Clustering: Reveals coarse patterns in the data.

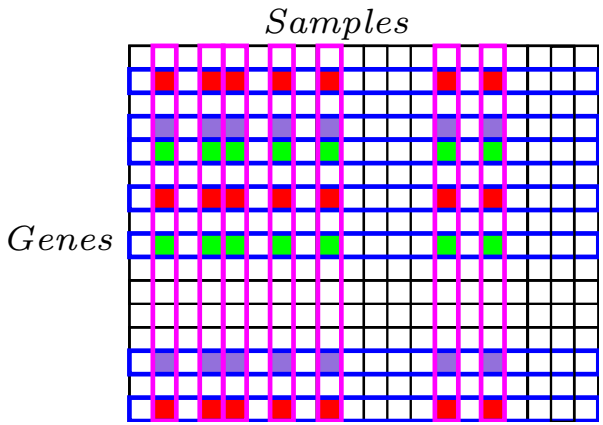


Overall Goal

- Develop a clustering algorithm for detecting condition-specific patterns of gene co-expression.

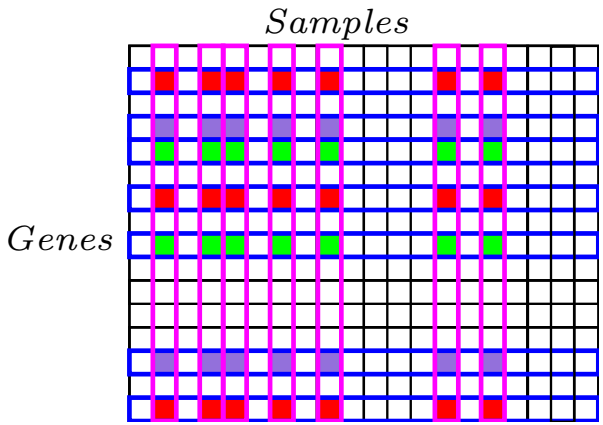
Overall Goal

- Develop a clustering algorithm for detecting condition-specific patterns of gene co-expression.
- A *gene expression signature* is
  - a subset of genes and a subset of samples



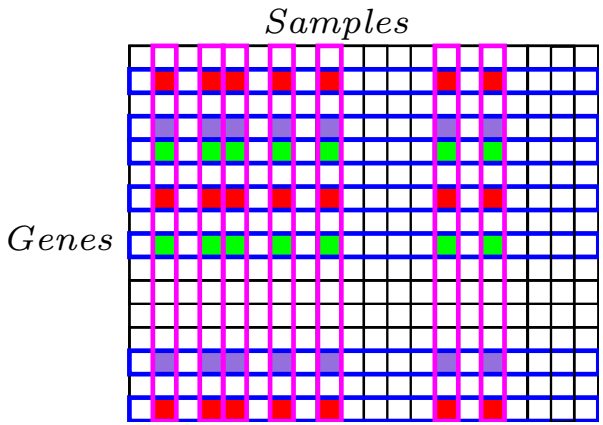
Overall Goal

- Develop a clustering algorithm for detecting condition-specific patterns of gene co-expression.
- A *gene expression signature* is
  - a subset of genes and a subset of samples
  - such that each selected gene is expressed with the same abundance in all the selected samples.



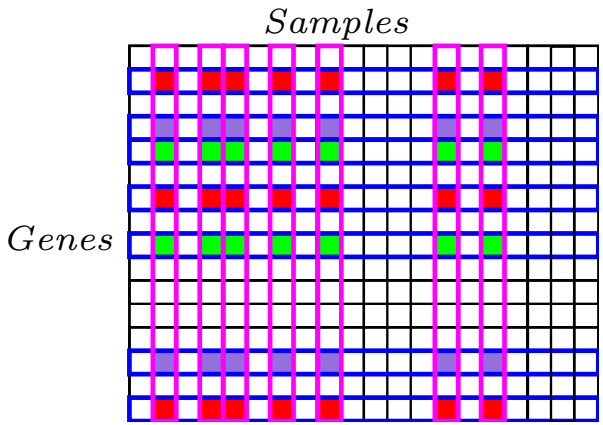
Overall Goal

- Develop a clustering algorithm for detecting condition-specific patterns of gene co-expression.
- A *gene expression signature* is
  - a subset of genes and a subset of samples
  - such that each selected gene is expressed with the same abundance in all the selected samples.
- These signatures combine clustering and dimension reduction/feature selection.



Overall Goal

- Develop a clustering algorithm for detecting condition-specific patterns of gene co-expression.
- A *gene expression signature* is
  - a subset of genes and a subset of samples (a bicluster)
  - such that each selected gene is expressed with the same abundance in all the selected samples.
- These signatures combine clustering and dimension reduction/feature selection.



Advantages

- Biclusters capture activity of genes in combination under specific conditions.
- Biclusters are easy to interpret: each gene is expressed in a particular "state" in all the samples in the Bicluster.
- Genes in an Bicluster may share the same function, be co-regulated, or be active in the same pathway.
- Biclusters may help us distinguish between or charactertise subtly-different classes of samples when no single gene is predictive.

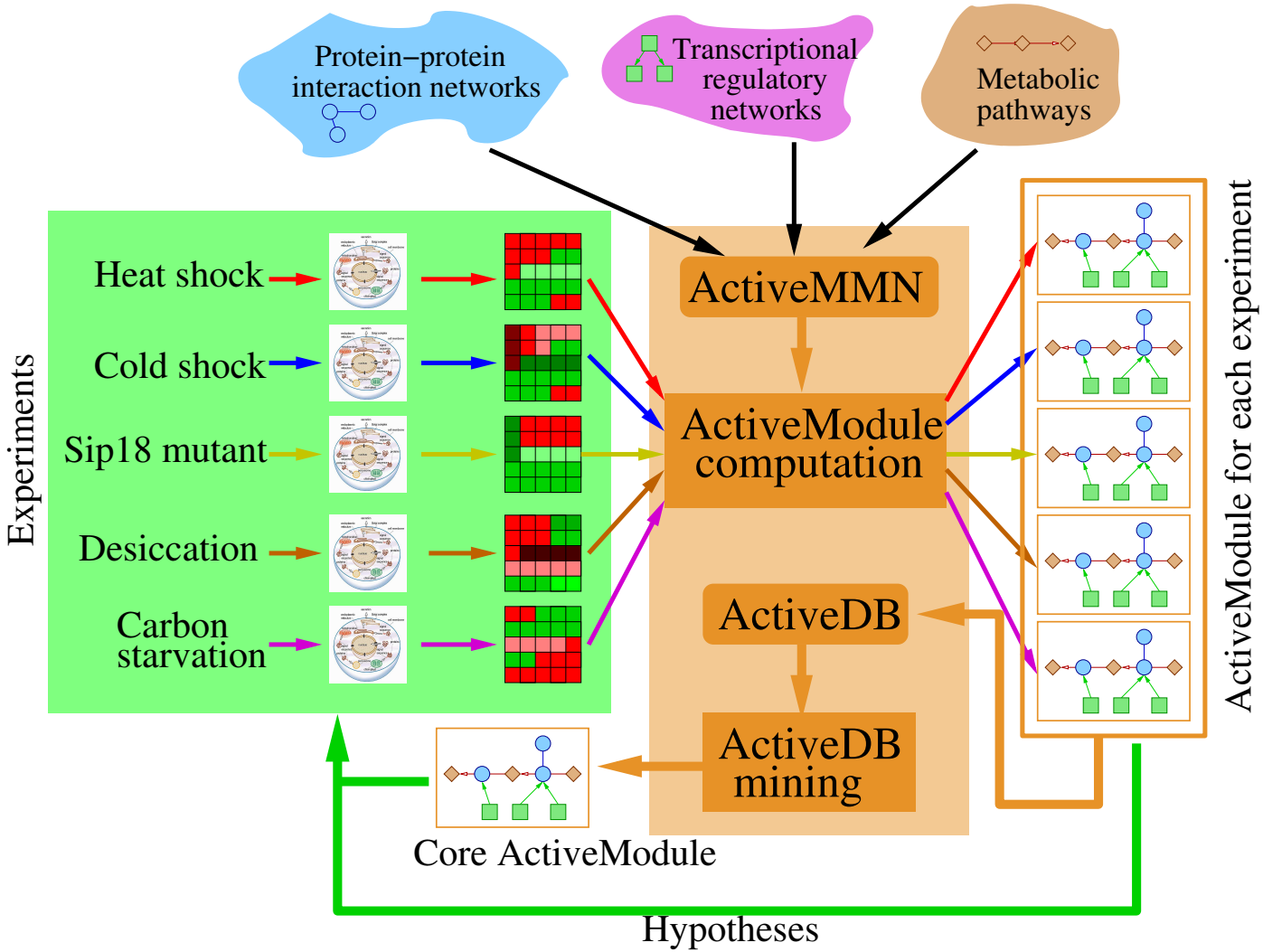
Project: Cross-condition Molecular Signatures

- Input:
  - Microarray data set where each sample belongs to a class.
  - Interact with bicluster database group for input data.
- Output:
  - Find biclusters with samples belonging to multiple classes. The expression patterns may be different from one class to another.
  - Functionally characterise each bicluster.
- Applications: different types of cancer, disease outcomes, stem cells.

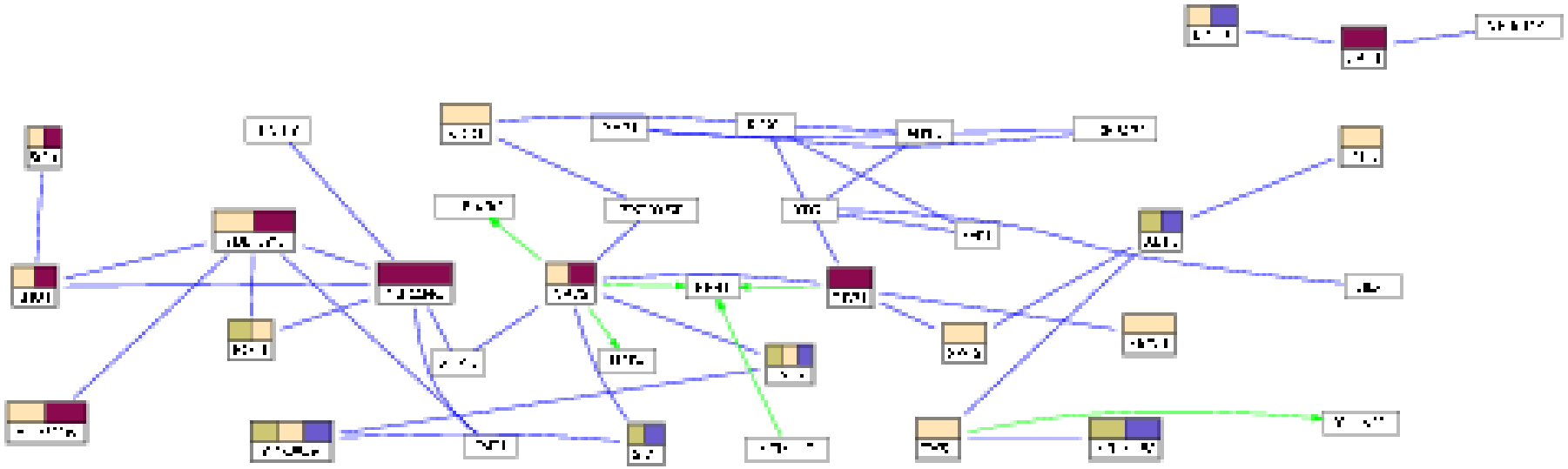
Project: Bicluster Database and Web Interface

- Input:
  - Comprehensive cancer-oriented (or disease-oriented) gene expression data base (only human, or include other species).
  - Detailed annotations for the data sets and for the genes.
  - Biclusters from the previous project.
- Output: Database that stores the gene expression data sets, annotations, and biclusters and allows complex queries on the biclusters and visualisations of the result.
- The group must interact with biologists to determine what types of queries they will find interesting.
- Collaboration with Prof. Simon Kasif of Boston University.

Introduction to ActiveNetworks



Introduction to ActiveNetworks



Project: ActiveNetworks in Cancer

- Input: Cancer gene expression data sets.
- Construct interaction networks for each cancer.

- Output: Cancer-specific and cross-cancer ActiveNetworks with proper functional characterisation.

Project: ActiveNetworks in Cancer

- Input: Cancer gene expression data sets.
- Construct interaction networks for each cancer.
- What is the source of edges?

- Output: Cancer-specific and cross-cancer ActiveNetworks with proper functional characterisation.

Project: ActiveNetworks in Cancer

- Input: Cancer gene expression data sets.
- Construct interaction networks for each cancer.
- What is the source of edges?
  - HPRD.
  - Use cover tree data structure to induce edges.
  - Use biclusters to induce edges.
- Output: Cancer-specific and cross-cancer ActiveNetworks with proper functional characterisation.

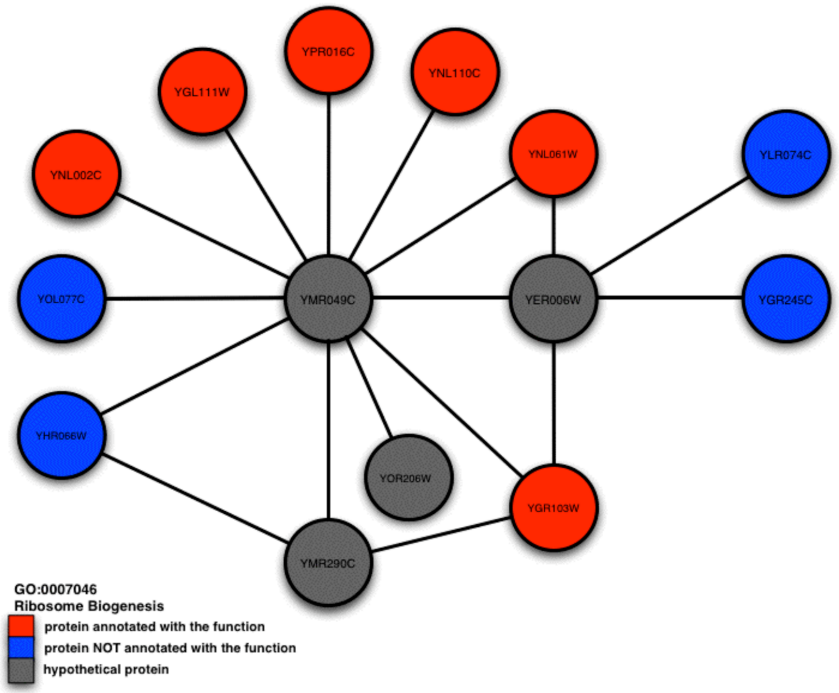
Project: Cross-species Systems Biology

- Input: Gene expression data sets and interaction data sets in various organisms.
- Apply ActiveNetworks system to this data.
- Output: Organism- or condition-specific and cross-organism or cross-condition ActiveNetworks.

Introduction to Functional Annotation

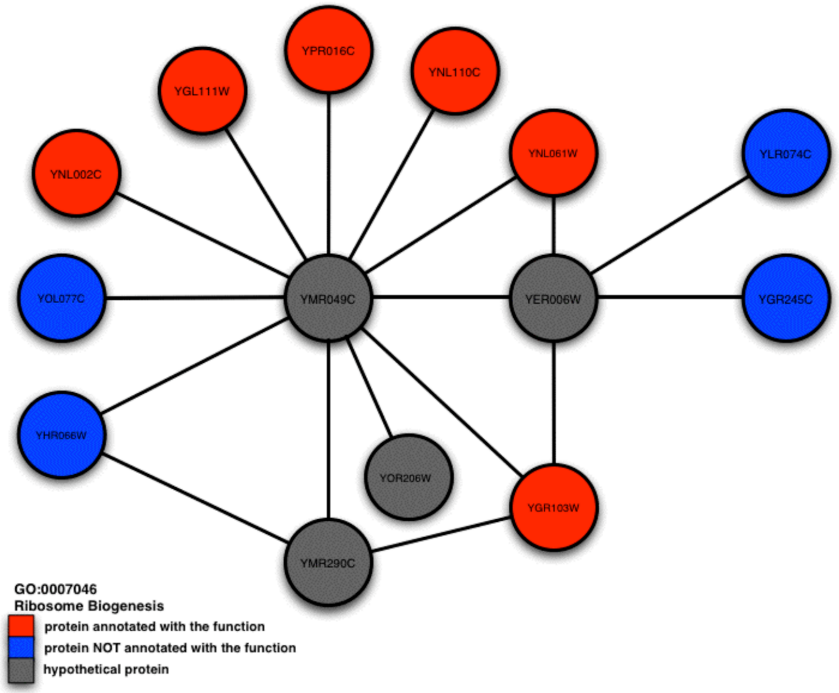
- Sequence similarity most commonly used to annotate genes in a newly-sequenced genome.
- More than 30% of the genes are not annotated.
- Use functional links between genes to construct a functional linkage graph (FLN).

Why is Functional Annotation Difficult?



- Neighbourhood structure is ambiguous.
- 20–30% of hypothetical proteins have only hypothetical neighbours (in GRID data set).
- Source data is very noisy.

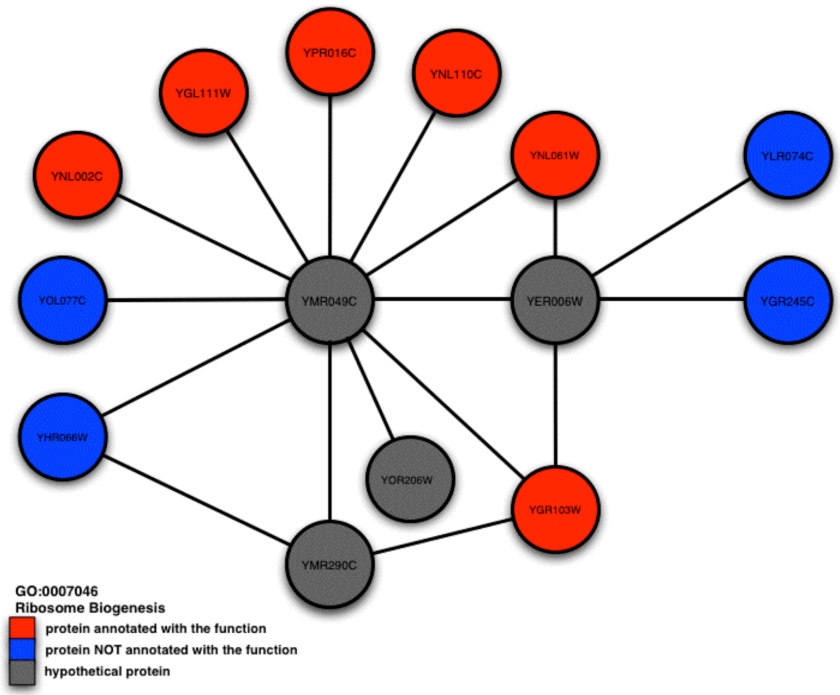
Goals



- Propagate evidence across functional-linkage graph.
- Intergate information from different sources to improve robustness.

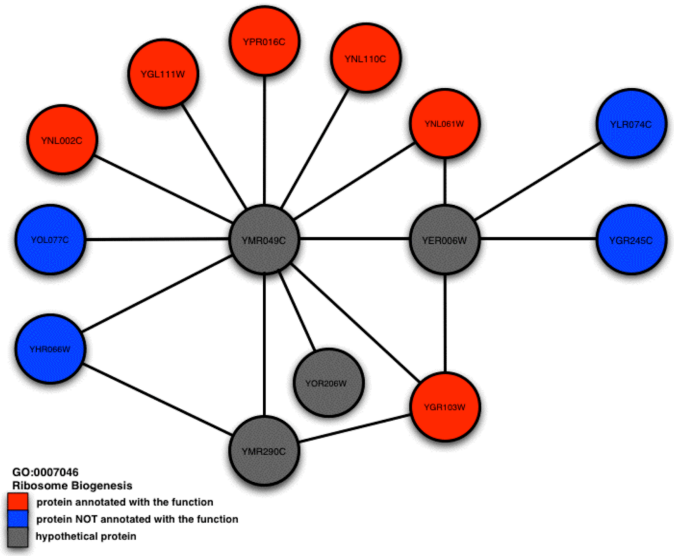
Hopfield Networks

- Functional linkage graph $\longrightarrow$ discrete Hopfield network.
  - Protein $\equiv$ node, interaction $\equiv$ edge.
  - Build a separate Hopfield network for each function.



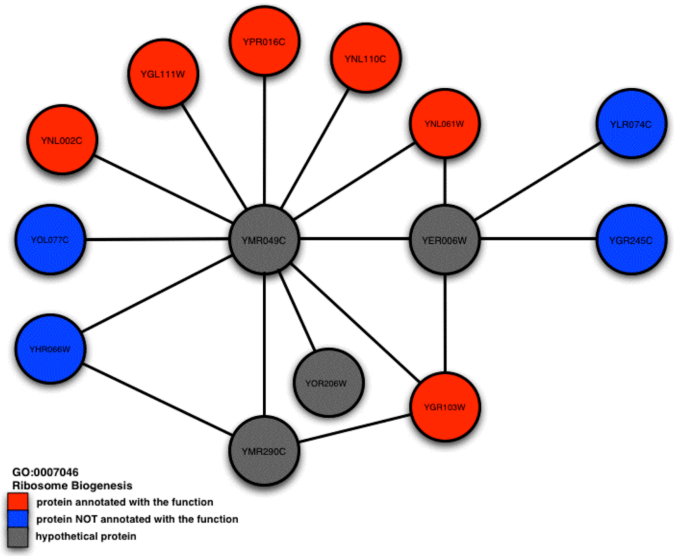
- Given a function f, each node i has an associated state s_i:
  - $s_i = 1$: protein i is annotated with f.
  - $s_i = 0$: protein i is hypothetical.
  - $s_i = -1$: protein i is annotated with another function f'.
- An edge between nodes i and j has a weight w_{ij}.

Goal: Maximally-Consistent Assignments



- An edge is *consistent* if it is incident on nodes with the same state.
- *Maximally-consistent assignment*: number of consistent edges is maximised.

Goal: Maximally-Consistent Assignments

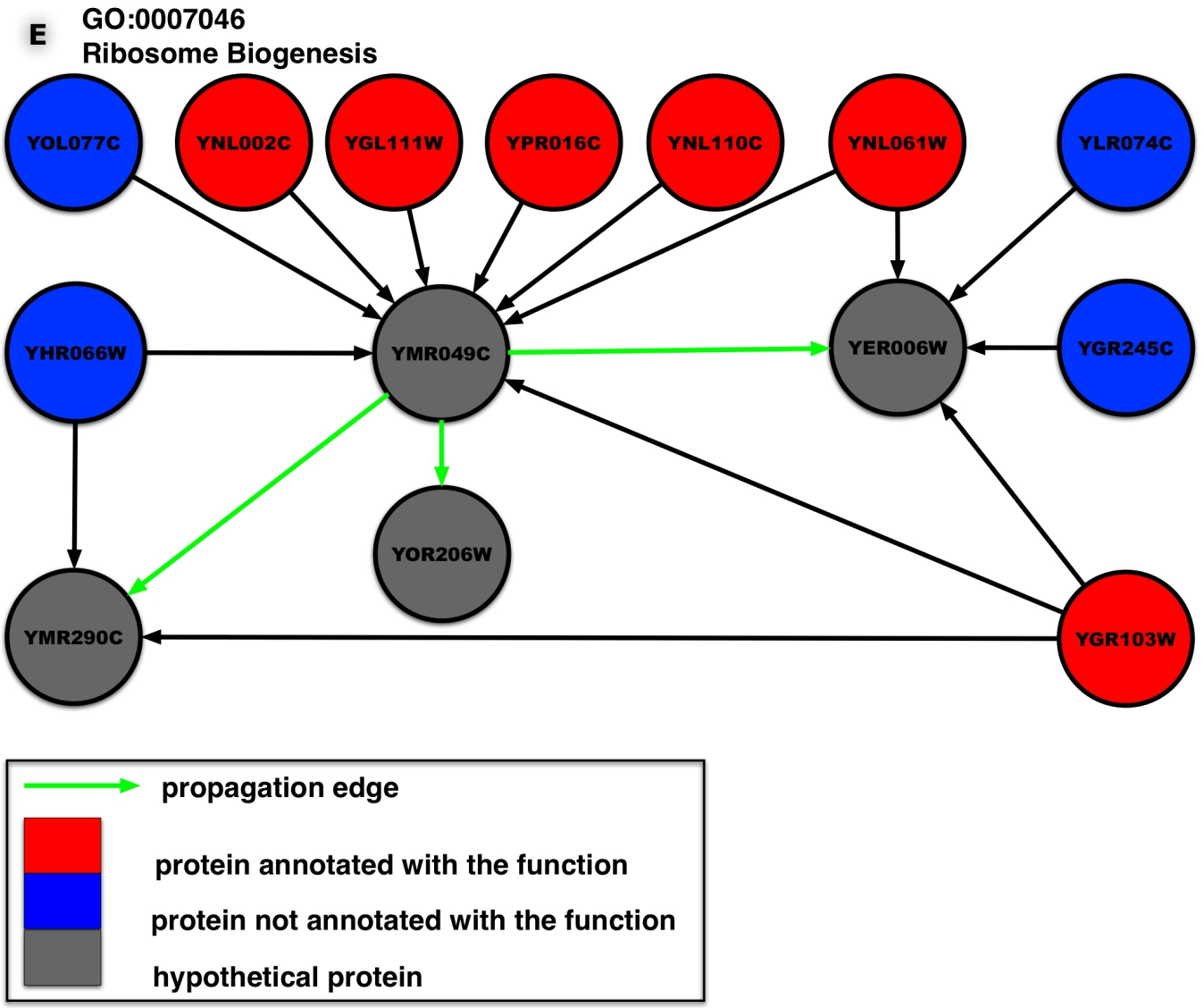


- An edge is *consistent* if it is incident on nodes with the same state.
- *Maximally-consistent assignment*: number of consistent edges is maximised.

Computational goal: Assign state of -1 or $+1$ to nodes with initial state 0 to achieve maximal consistency by minimising

$$E = -\frac{1}{2}\sum_i \sum_j w_{ij} s_i s_j$$

Propagation Diagrams



Project: Improvements to the GAIN Algorithm

- Incorporate correlations between *functions*: predict that a gene has function f when the gene's interactors have function g.
- Minimise
$$\sum_f \sum_g \sum_i \sum_j w_{ijfg} s_i^f s_j^g .$$
- Can factor w_{ijfg} into product $u_{ij} v_{fg}$.
- How do we estimate the weights v_{fg}?
- Develop new algorithms or modify current algorithm to optimise this function.

Project: Functional Annotation of the Human Genome

- Construct FLN using HPRD and/or cancer expression data sets.
- Use cover tree data structure to implement fast similarity queries on gene expression profiles.
- Carefully assess which groups of functions can be accurately predicted for the human genome.

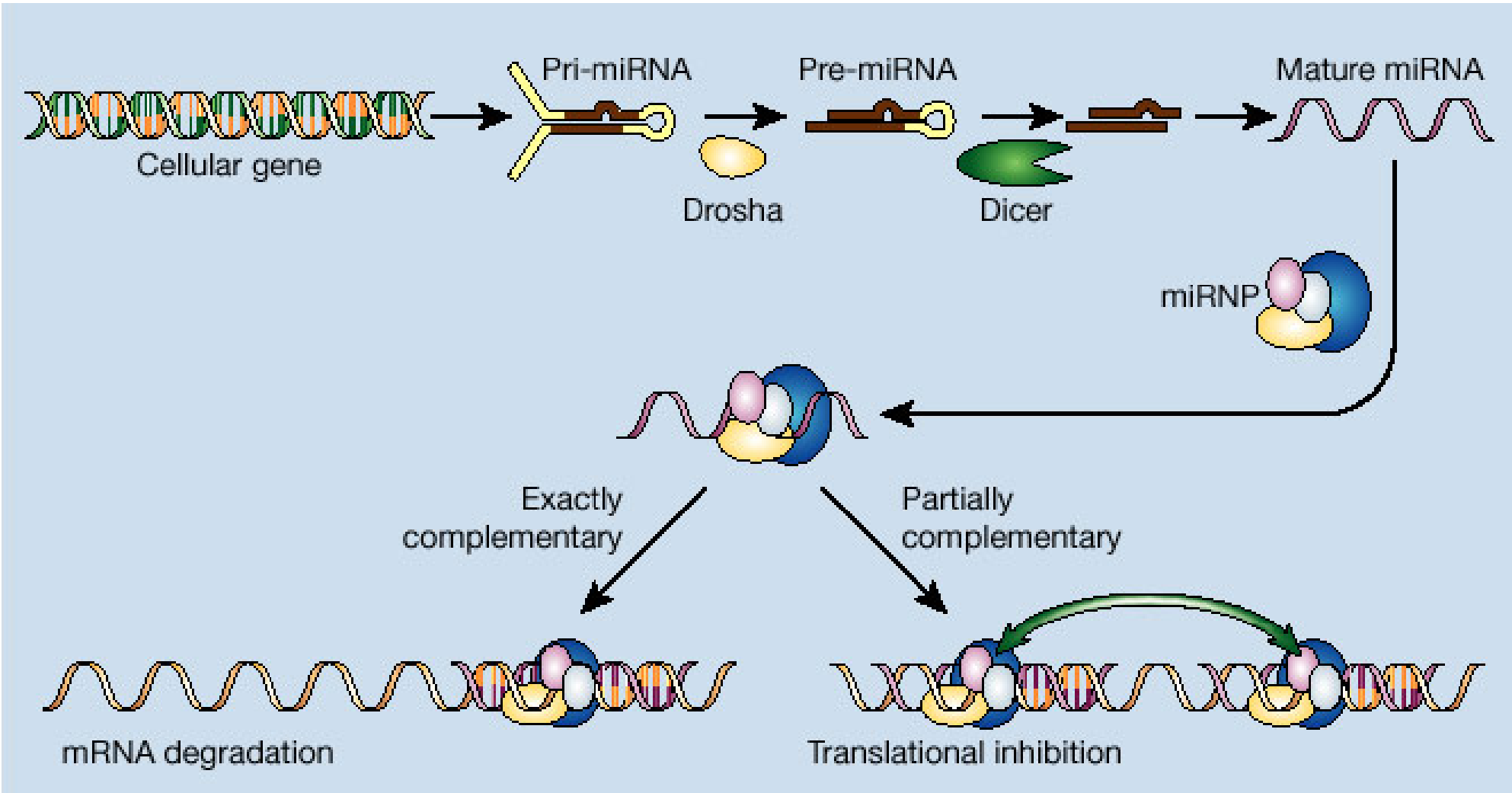
Project: Cross-Species Functional Annotation

- Use metagene and gene expression data sets of Stuart et al. to construct FLN.
- Alternatively, use the microarray data sets collected by Ihmels et al.
- Use GAIN to functionally annotate genes across multiple species.

Project: Whole-Genome Functional Annotation of Malaria

- Build FLN using a combination of gene expression data and proteomics data.
- Use GAIN on this FLN to annotate malaria genome.
- Potential collaboration with Dr. Dharmendar Rathore of VBI.

Project: Prediction of microRNA Targets



Project: Prediction of microRNA Targets

- Novina and Sharp: "...miRNAs regulate separate genes—perhaps hundreds or more per miRNA. Furthermore, the degree of translational inhibition by miRNAs is thought to depend on how many of these molecules are bound to the target mRNA. Typically, such an mRNA contains many binding sites at one end (the 3'-untranslated region), and several different miRNAs can target the same 3' region."
- For each miRNA and for each gene, find potential binding sites in the 3'-UTR of that gene.
- Use biclustering algorithms to find sets of miRNAs that target the same group of genes.

Project: Prediction of microRNA Targets

- Novina and Sharp: "... miRNAs regulate separate genes—perhaps hundreds or more per miRNA. Furthermore, the degree of translational inhibition by miRNAs is thought to depend on how many of these molecules are bound to the target mRNA. Typically, such an mRNA contains many binding sites at one end (the 3'-untranslated region), and several different miRNAs can target the same 3' region."
- For each miRNA and for each gene, find potential binding sites in the 3'-UTR of that gene.
- Use biclustering algorithms to find sets of miRNAs that target the same group of genes.

Project: Genome Variation and Disease

- Only a small fraction of SNPs are in coding regions.
- Correlate presence of SNPs (in promotor motifs) with changes in gene expression and suggest how these changes may cause disease.
- Collaboration with Prof. Liqing Zhang in the Department of Computer Science.

Ground Rules for Projects

- Weekly 1 hour meetings with each group on Thursdays.
- Maintain web pages describing your project (will decide location).
- Project descriptions (motivation, background, related and previous research, approach, data, any preliminary results) due on October 14.
- Project reviews on October 28 in class.
- Final project presentations on December 9 and 16 in class.

Hardware Support for Projects

- You can use `cuthbert.cs.vt.edu`, `whipple.cs.vt.edu`, and `sundaram.cs.vt.edu`.
  - `cuthbert` and `whipple` are Dells with a 2.8GHz Pentium IV processor, 1GB of RAM, and a 80GB hard drive running Fedora Core 2.
  - `sundaram` runs Mac OS X 10.2.5, has two 1.8GHz PowerPC processors, 3GB of RAM, and a 160GB hard drive.
- Obtain accounts on `bioinformatics.cs.vt.edu` from Douglas Slotta (dslotta@vt.edu) in Torgerson 2160.

Software Support for Projects

- Molecular signatures
  - *xMotif* algorithm implemented in C++ for finding biclusters in gene expression data.
  - Implementation of the *apriori* algorithm for finding itemsets.
- Functional annotation
  - *GAIN* algorithm in C++.
  - *Cover tree* data structure implemented in Java.
  - Perl classes for manipulating functional predictions.
- ActiveNetworks
  - Various elements of the *ActiveNetworks* pipeline.
  - *Cover tree* data structure implemented in Java.
- C++, Java, and Perl classes for manipulating graphs.
- Perl class (also a rudimentary C++ class) for manipulating functional annotations.
- *spring* C++ programme, a high-level interface to *graphviz*.

Projects

- Cross-condition gene expression signatures related to cancer, treatment outcome, stem cells.
- Bicluster database and web-server
- ActiveNetworks in cancers.
- Cross-species systems biology.
- Improvements to the GAIN algorithm.
- Functional annotation web server
- Cross-species functional annotation
- Annotation of the human genome using HPRD
- Functional annotation of Malaria genome
- Prediction of microRNA targets
- Association of SNPs with disease.
- Completion of bound protein-ligand complexes.

Projects

- Cross-condition gene expression signatures related to cancer, treatment outcome, stem cells.
- Bicluster database and web-server
- ActiveNetworks in cancers.
- Cross-species systems biology.
- Improvements to the GAIN algorithm.
- Functional annotation web server
- Cross-species functional annotation
- Annotation of the human genome using HPRD
- Functional annotation of Malaria genome
- Prediction of microRNA targets
- Association of SNPs with disease.
- Completion of bound protein-ligand complexes.

Project Meetings

- Cross-condition gene expression signatures: Greg, Jonathan, and Rajat; Satish, Srinivas, and Venkat.
- Bicluster database and web-server: Greg, Jonathan, and Rajat; Satish, Srinivas, and Venkat.
- Functional annotation web server: Corban and Shivaram
- Cross-species functional annotation: Chaitanya, Kiran, Rob, and Shenghua
- ActiveNetworks in cancers: Deept and Nilanjan.

9-11am	Greg, Jonathan, and Rajat; Satish, Srinivas, and Venkat
11am-12pm	Chaitanya, Kiran, Rob, and Shenghua
1-2pm	
2-3pm	
3-4pm	

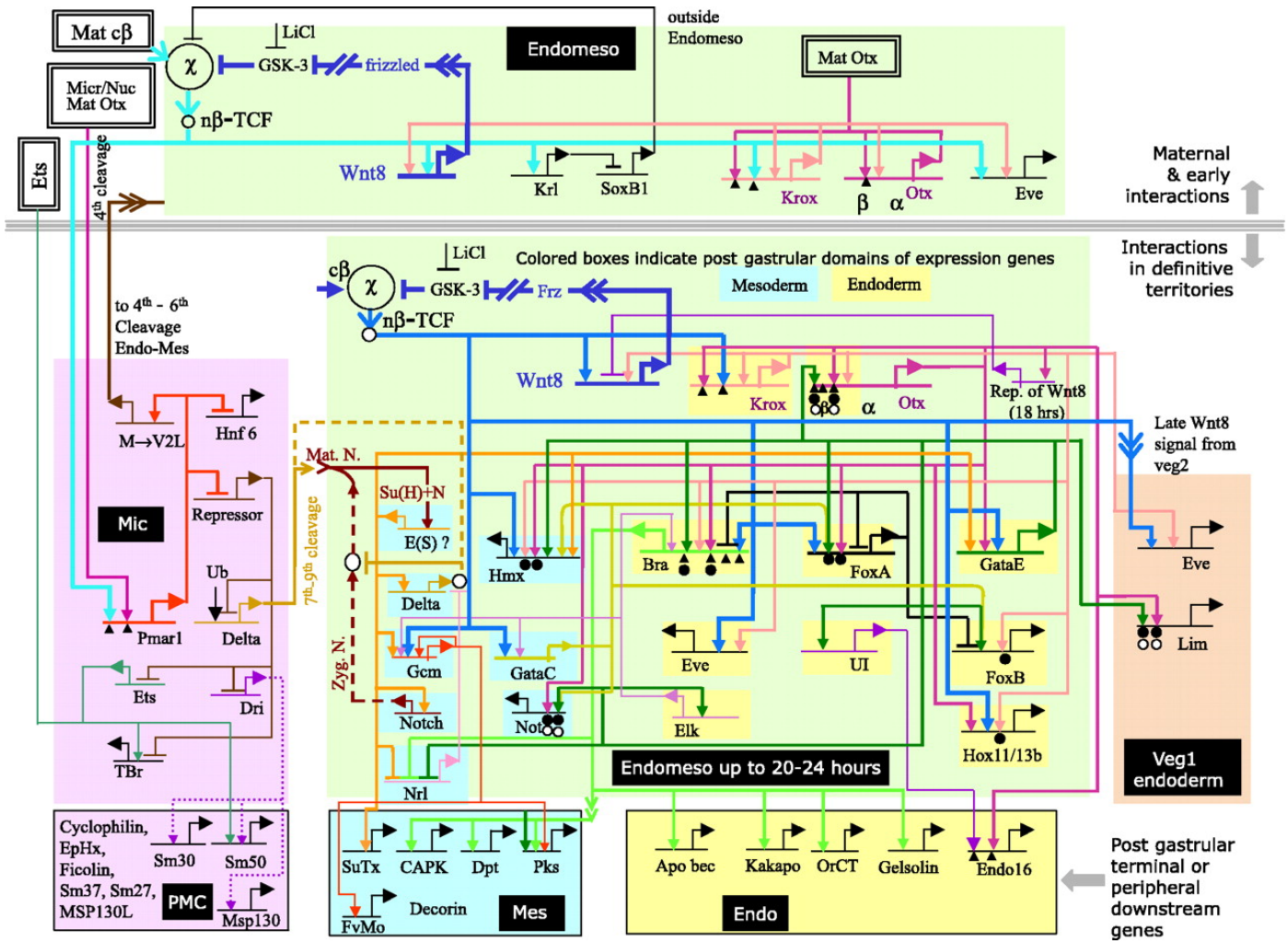
Computational Systems Biology (Fall 2003)

- Fundamental computational ideas and techniques used in systems biology.
- Biotechnological breakthroughs that make systems biology possible.
- Studied research that improves our basic understanding of biology.

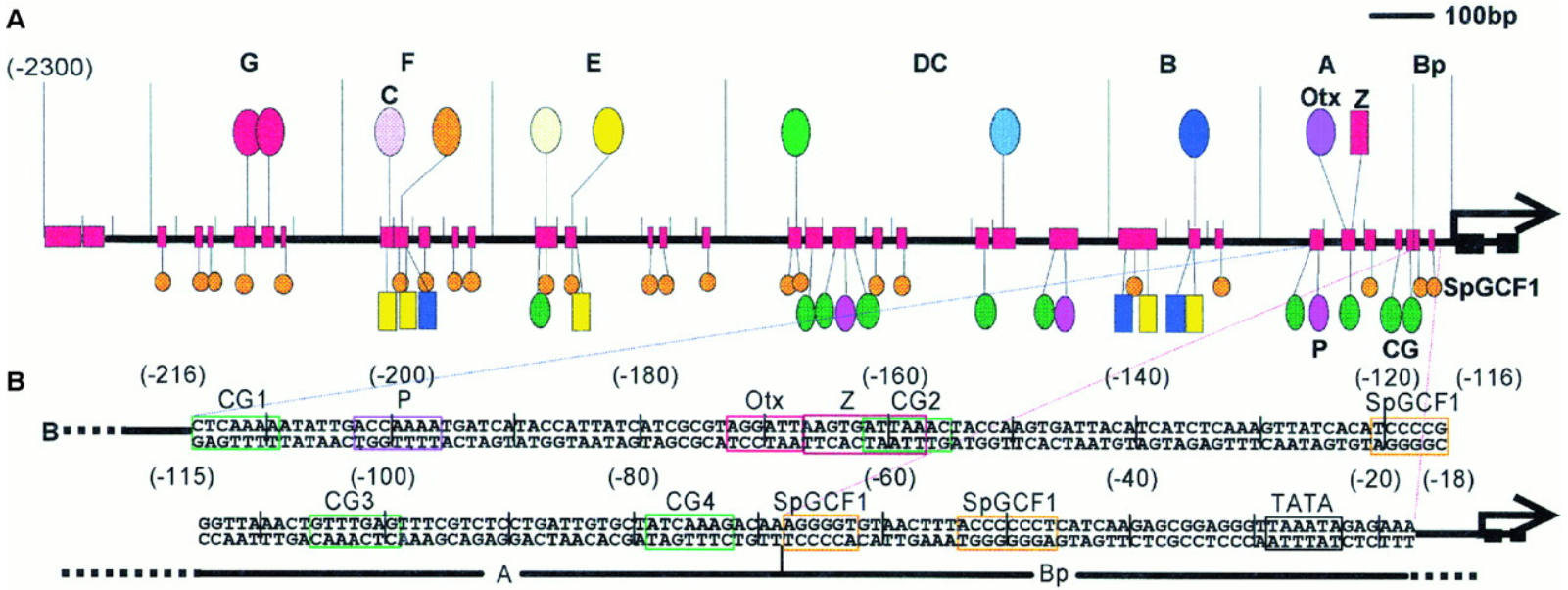
CSB 2003: Topics in Analysis of Gene Expression Data

- Simple DNA microarray clustering
- Biclustering of DNA microarray data

CSB 2003: Transcriptional Regulatory Networks



CSB 2003: Transcriptional Regulatory Networks



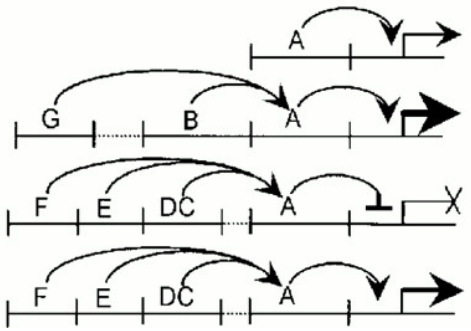
C Module A functions:

Vegetal plate expression in early development:

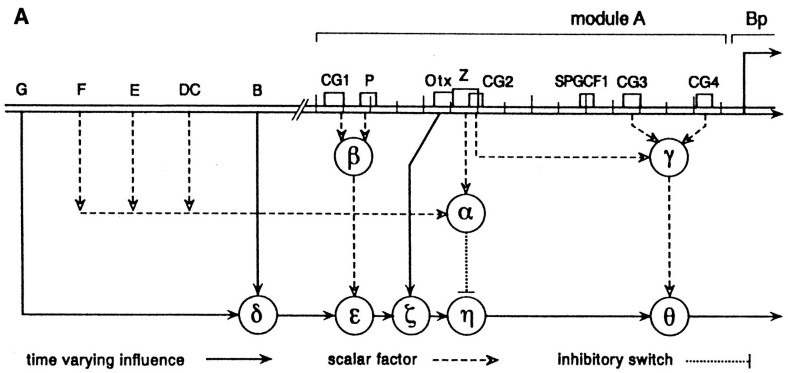
Synergism with modules B and G enhancing endoderm expression in later development:

Repression in ectoderm (modules E and F) and skeletogenic mesenchyme (module DC):

Modules E, F and DC with LiCl treatment:



CSB 2003: Transcriptional Regulatory Networks



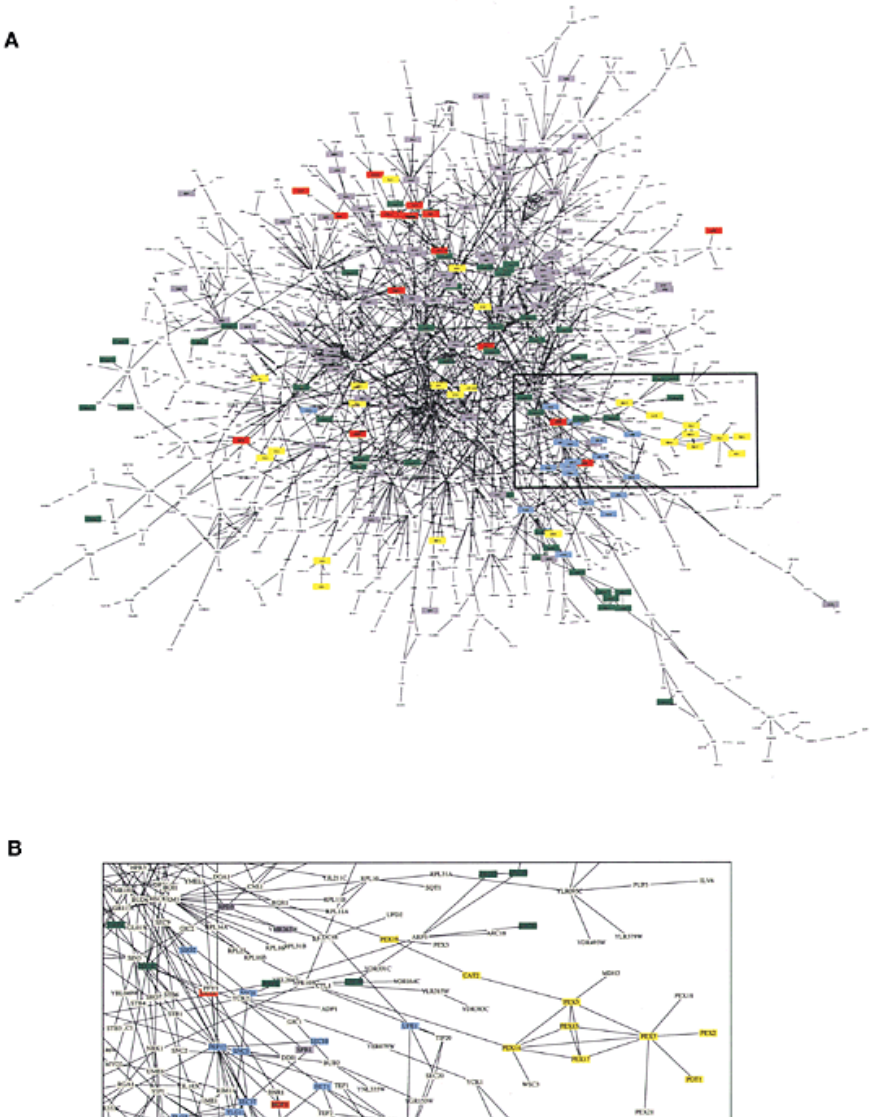
B

if (F = 1 or E = 1 or CD = 1) and (Z = 1)	Repression functions of modules F, E, and DC mediated by Z site
$\alpha = 1$	
else $\alpha = 0$	
if (P = 1 and CG_1 = 1)	Both P and CG_1 needed for synergistic link with module B
$\beta = 2$	
else $\beta = 0$	
if (CG_2 = 1 and CG_3 = 1 and CG_4 = 1)	Final step up of system output
$\gamma = 2$	
else $\gamma = 1$	
$\delta(t) = B(t) + G(t)$	Positive input from modules B and G
$\varepsilon(t) = \beta * \delta(t)$	Synergistic amplification of module B output by CG_1-P subsystem
if ($\varepsilon(t) = 0$)	Switch determining whether Otx site in module A, or upstream modules (i.e., mainly module B), will control level of activity
$\xi(t) = Otx(t)$	
else $\xi(t) = \varepsilon(t)$	
if ($\alpha = 1$)	Repression function inoperative in endoderm but blocks activity elsewhere
$\eta(t) = 0$	
else $\eta(t) = \xi(t)$	
$\Theta(t) = \gamma * \eta(t)$	Final output communicated to BTA

CSB 2003: Topics in Transcriptional Regulatory Networks

- Extracting them from DNA microarray data.
- Finding genes that are regulated together under specific conditions.
- Developmental regulatory networks.
- Modular organisation and network motifs.

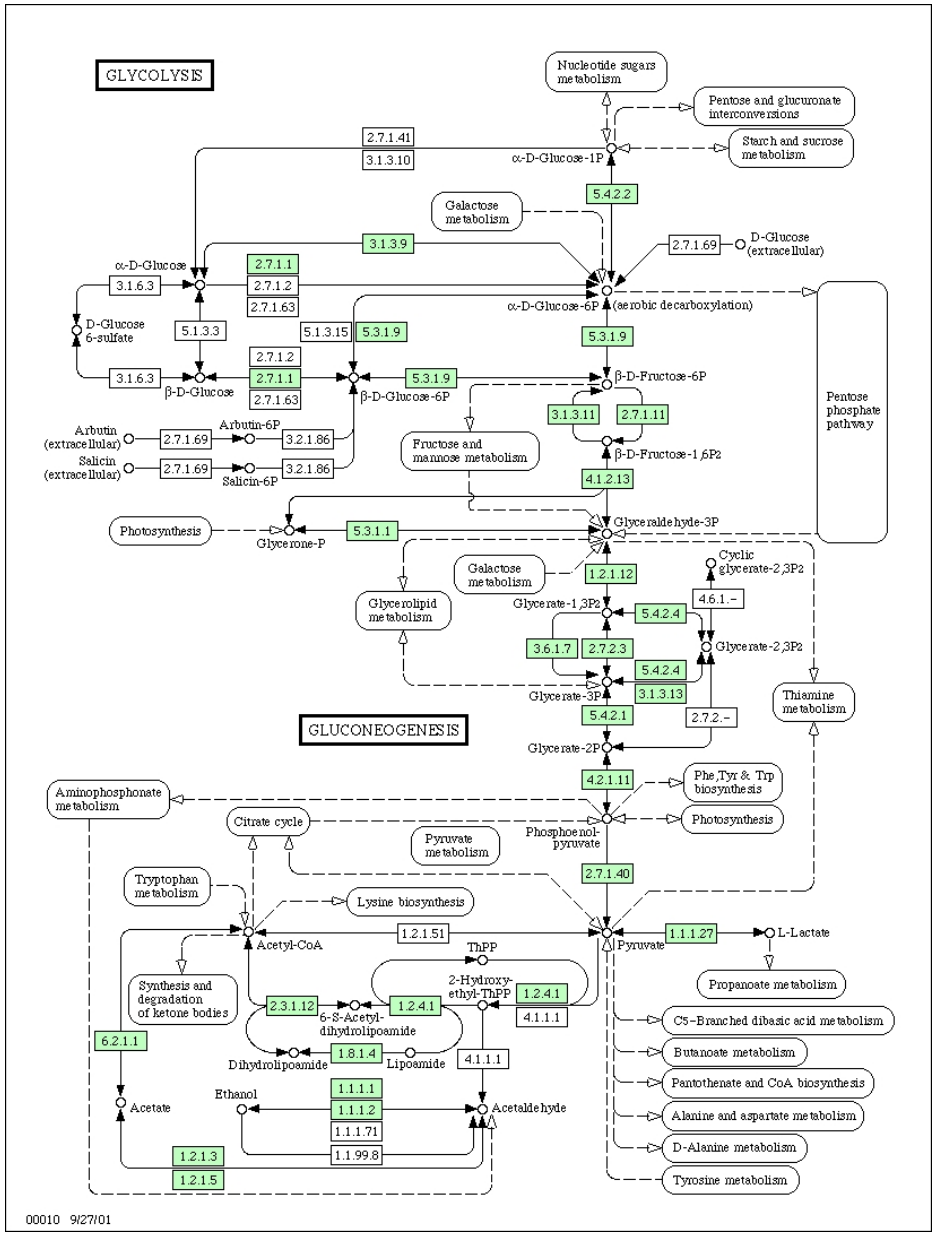
CSB 2003: Protein-Protein Interaction Networks



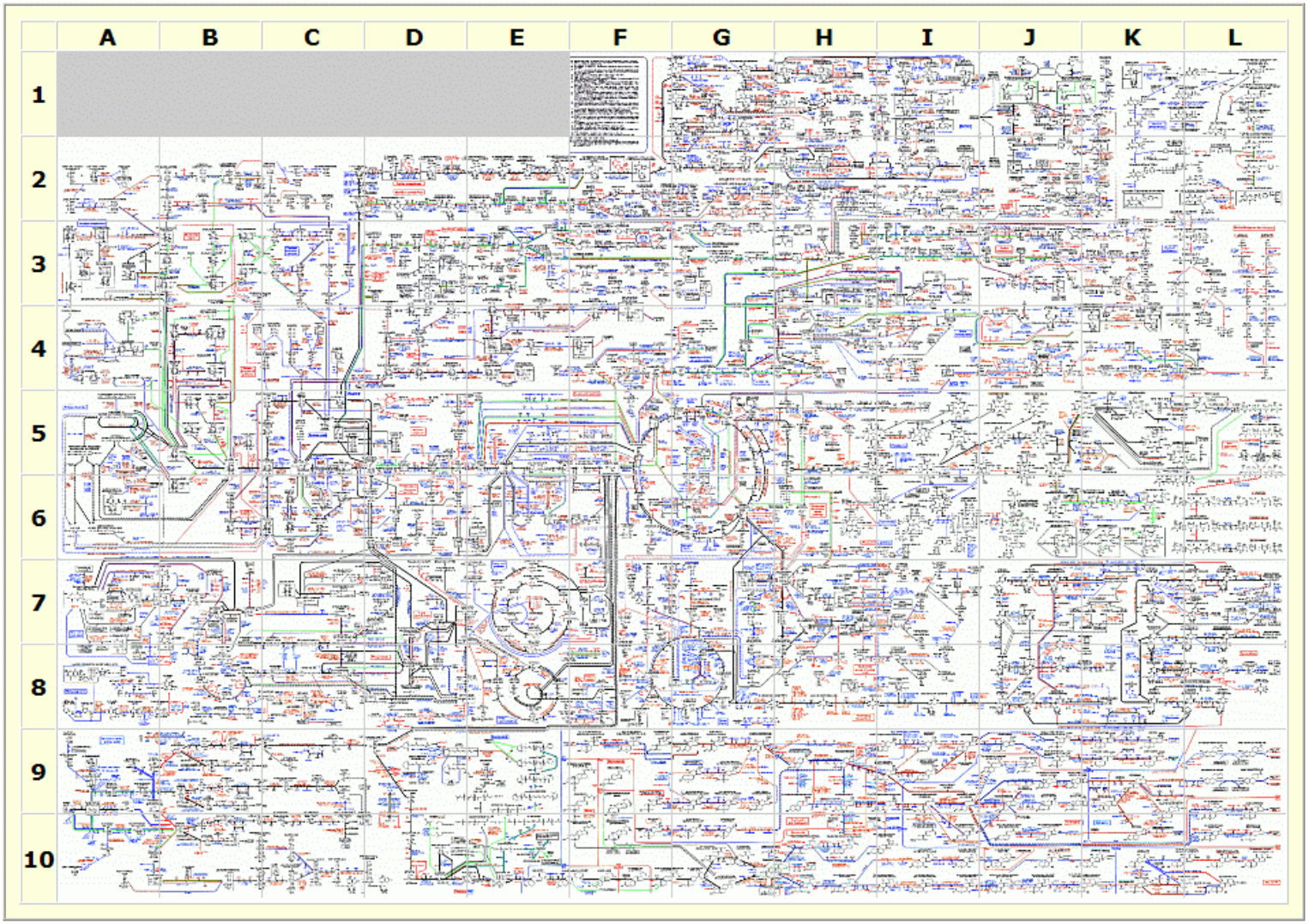
CSB 2003: Topics in PPI networks

- Experimental and computational techniques for determining protein-protein interactions.
- Assessing and improving their reliability.
- Functional annotation using PPI networks (by integrating different sources of evidence).

CSB 2003: Metabolic Networks



CSB 2003: Metabolic Networks



CSB 2003: Topics in Metabolic Networks

- High-level structural properties.
- Modelling and reconstruction.
- Modelling and simulation of cellular networks.