The proposed project is to construct a regulatory network that identifies the protein-protein and gene-protein interactions in the yeast genome. The nodes in the network represent either proteins or genes that are important in yeast's life cycles, and a directed edge between two nodes indicates that the production of the gene/protein in the source node either enhances or inhibits the production of the gene/protein in the target node. In the broader context of molecular biology and bioengineering, constructing such regulatory network is one method to understand the complex interactions between protein molecules, through which an organism responds to environmental and biological situations.

The approach used in this project can be extended to understand the protein-protein and protein-gene regulations in human genome. Such extension will have huge impacts to the society, particularly in areas such as drug discovery, disease diagnosis, disease treatments and other areas in biotechnology and biomedical engineering. For example, the cause of cancer can be viewed as uncontrolled transcription of some "bad" genes. If we know the proteins that regulate the transcription of those "bad" genes, we can design drug that carries out similar functions of the regulating proteins and have a cure for cancer.

In the current state, regulatory relationships are found using experiments. For example, biologist will manually put two proteins together and see if they interact. As a result, discovering the regulatory function of one gene often take a whole year. With the recent development of DNA microarray technique which simultaneously measures thousands of gene's expression activity, researchers start to apply systematic algorithms such as Bayesian network and regression models to discover such regulatory relationships. However, these studies have only been carried out for the past 3-4 years, and the regulatory network for organisms, as simple as yeast, is far from fully understood.

In our project, we will consider three yeast datasets that contain information on 106 yeast transcription factors (i.e. proteins) and 273 gene knock-off experiments. Each dataset describes one of following relationships: protein-DNA interactions, pair-wise protein interactions, and gene-gene relationship. The reason for choosing yeast is that the yeast genome is relatively small in size (~6000 genes) and relatively complete. We will construct a single network for the three datasets using the following approach:

1. Problem set-up: we normalize all numerical measurements in three datasets to log10 based measurements, ranging from -1 to 1. We do this by assigning the log of the maximum measurement in each dataset to 1, the log of the minimum to -1, and ranging other measurements accordingly. Next, we organize the data into

three matrices, with rows representing the genes, and column represent the features (intensity measures at each the spot of the array). Missing values are filled in using K-Nearest Neighbour method.

2. Initial Analysis: we perform a k-mean clustering algorithm on each of the dataset to find groups of genes/proteins that behave similarly, suggesting possible regulatory relationships within groups. The groups are represented by cliques in the network. Next, we calculate the correlation coefficients between the groups and use these measures, together with a preset threshold, to find possible links between groups.
3. Networks Combination and Reduction: we delete unnecessary edges in the networks and combine the three different networks into one. In general, we analyze each cluster, and figure out the most likely regulator in that cluster, and remove other edges. We may introduce probabilities to measure the confidence in removing the edges. Any two nodes from the three networks are merged if we find that the gene represented by one node codes for the protein represented by the other nodes.
4. Advanced techniques and enhancements: we consider some advanced techniques such as Bayesian network, or other probabilistic graphical model to enhance our network as reviewed in the paper by Freidman et.al.

Further researches on probabilistic graph models, the BIND database and results on known gene/protein interactions in yeast are needed to complete the project.

On completion of the project, we expect to achieve the following objectives:

- A correct regulatory network that reveals significant protein-protein, gene-protein interactions in yeast. That is, both the type of interaction(enhancement/inhibition) and the links between proteins and genes are correct
- The resulting network can be reasonably explained and interpreted in biological context.
- A colorful and informative visualization for the network, showing the network it self, and when the user click on a link in the network, the tool pops up a window to provide evidences to justify the link.

We will test the correctness of the network by the following measures:

- Motif: we examine the co-regulated genes as suggested in the network, and evaluate how many of them have a common sequence in their upstream region (the genes are expected to have such common sequence if they are co-regulated). A more complicated task is to look at the motif of all genes available, and calculate how many pairs of genes have sequence in common but no links shown in the network.
- Annotated data: we collect published data on known regulatory relationships in yeast, and verify these relationships with our network. We expect that our network can discover a majority of those relationships.

- Knock-off Experiments: the Banting and Best Department of Medical Research in U of T can perform knock-off experiments to verify the regulatory relationship between genes. If time permits, we will also verify some significant links in the network experimentally.

The following table summarizes work divisions between the group members. In order to complete the project, we need computing facilities to perform computational intensive algorithms such as data clustering, pre-processing, and graphical model building. We believe that the ECF lab server provides enough computing power in our case. The cost for knock-off experiments is $100/expr. Professor Hughes from the Banting and Best department has agreed to do 10 experiments for us for free. We estimate that we need 10-15 experiments, which amount to $500. Each group member can contribute $50, and for the rest of the money, we are requesting support from the Design Center.

Main Tasks	***Detailed Subtasks***	***J.Y. Lin***	***C. Shah***	***D. Wang***	***Q.Y. Tang***
Preprocessing	Data collection		A		R
	Data normalization/missing values	R			
	Data format conversion	R			A
	Document work/results	R			
Modeling	clustering of dataset			A	R
	Correlation calculation/network building	A	R(dataset2)	R(dataset3)	R(dataset1)
	Develop test and verify the network	R			
	Document work/results		R(dataset2)	R(dataset3)	R(dataset1)
	Network reduction	A	R(dataset2)	R(dataset3)	R(dataset1)
	collecting bio info for merging 3 networks		R		
	Implement networks merging algorithm				R
	Develop test and verify the network	A			R
	Document work/results				R
	Network enhancement	A	A	R	A
	Develop test and verification	A		R	
	Document work and results			R	
Evaluation	Motif analysis and verification		R		
	Annotated data		R		
	Document work/results		R		
Visualization	A software to show the Network	R			
	Software Testing	R			
	Document work	R			
Prepare poster, oral presentation and final report		R	A	A	A