## *Table of Contents*

## *Project Proposal*

### 1. Motivation & Challenges in Discovering Gene Regulatory Relationships

Living organisms are complex systems to understand. Biologists and medical scientists have been pushing the frontiers of genomics to understand the evolutionary mechanisms and study biological problems in organisms ever since the discovery of Watson-Crick's DNA model. Activities within an organism are normally understood at the molecular biology and cell level where proteins form the backbone of the entire machinery. It is through the production of different proteins and binding between them that a cell carries out its particular functionalities, and it is through the collective efforts of the cells that an organism carries out daily functions. Therefore, one step towards the understanding a life form amounts to understanding the interactions between different protein molecules. However, because of the complex structure of protein molecules, it is often easier to analysis the interactions by looking at the genes that encodes protein, and instead analyze how expressions of genes relates to each other, or in a biological term, the regulatory relationships between genes.

In the current state, regulatory relationships are found by doing experiments in the lab. For example, biologist may manually put two proteins together and see if they interact. It often takes one year to find the regulatory function of one gene. With the recent development of DNA Microarray technique to simultaneously measure thousands of gene's expression activity, researches start to apply systematic algorithms such as Bayesian network [3] [1] and regression tree model [7] to automatically discover the regulatory mechanisms from the vast amount of data obtained. These studies have only been carried out in the past 3-4 years, and the regulatory network for organisms, as simple as yeast, is far from fully understood.

Our project "Generative models for Analyzing Molecular Biology Data" is different from these previous researches in that we will not only look at the gene-expression profile, but will consider three different datasets in conjunction. These datasets profile from different perspectives, the possible interactions that exists in *Saccharomyces Cerevisiae* (yeast), namely, how the absence of one gene affects the expression of other genes, how transcriptional factors binds to the DNA sequences, and how any two protein molecules

interacts. We will use machine learning algorithms on these data to identify known regulatory relationships and discover previously unknown relationships between genes & proteins in yeast.

The main challenges in this project are a) the biological data that we have are noisy and have many missing values, b) very little prior knowledge is known about the true interactions between genes and c) the inherent complexity of interactions between genes in living organism allows many possible way of interpreting the data and therefore, inferring the true underlining interactions becomes difficult. Machine learning techniques such as clustering method will be used in early stage of the project to understand some of the patterns in each individual dataset, and a confidence measure will be adopted in the later stage to account for the probabilistic nature of the data.

## 2. Regulatory Network Representation

The ultimate goal of this project is to infer a graph showing the regulatory mechanism of genes in yeast and to develop software for visualizing this graph. Such a graph shall provide a comprehensible view of the complex interactions among the genes and also allow for systematic methods such as graph traversal algorithms to be applied directly for extracting relevant information for both biological and biomedical purposes. To achieve this goal, we will adopt a representation of the regulatory graph based on the ones used in [2][3][4], with some enhancements.

The network will be a directed graph with each node representing individual gene; all other form of molecules such as proteins are mapped to their corresponding genes. An edge in the graph indicates that the expression level of the gene in the source node will have a **direct** effect on the expression of the gene in the target node. Figure 1 shows a mock up of the representation we have in mind. A (+) sign on the edge shows that the effect is activation, while (-) means inhibition. In the case that the type of effect is not inferable from the dataset, we will denote it by a question mark. Furthermore, the width of an edge indicates our confidence of its presence, with the darkest and widest ones indicate the most confidence. We define the confidence of an edge by how well its existence can be justified by the data: the edge can on one extreme be supported strongly by all three datasets while on another extreme weakly by only one dataset.

In the software we will develop, the network will be shown on the screen, and supporting information for each edge will be shown when the user clicks on that edge of the network.

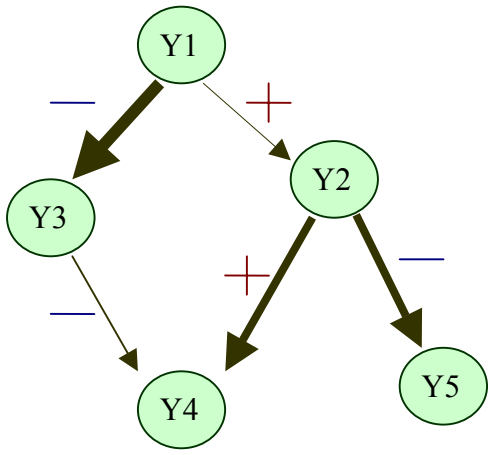


**Figure 1**: Illustration of A regulatory Network. The node Y1, …,Y5 represents the gene, and edges represents the relationship. For example, presence of gene Y1 will inhibit the expression of gene Y3 (an edge with (-) sign), while activate the production of Y2 (an edge with (+) sign). Thickness of the edge indicates the confidence of such a link. For example, we are more confident about Y1's inhibition effects on Y3 than its activation effect on Y2.

It is possible that several alternative paths exist in the network and each of them is a valid path given the data set. In the final visualization of the network, animation will be used to show all the possible configurations of the network by switching periodically from one to another.

## 3. Network Structure and Parameter Inference

The three datasets we will consider in this project are gene knock-out dataset [5], ChIP chip data for 106 regulatory factors [6], and BIND dataset for protein-protein interactions. In this section, we will outline the methods to be used to construct the network described in section 2 from a computational point of view.

### 3.1 Data pre-processing and filtering:

For knock-out dataset and protein-DNA dataset, we will remove the genes that have missing value > 5% out of the 300 knock-out experiments performed. The missing values in the dataset will be substituted using the K-NN (K=7) impute method. The data will also be centered so that each array has zero mean. The same gene removal criterion and missing value imputation will be used for the ChIPchip data. In addition, we will convert the data to a 0-1 matrix by setting all measurements with p-value<1e-4 to 1, indicating that a binding exists between the corresponding regulator and gene exists, while all others to 0, indicating the opposite. The BIND dataset will be converted to an NxN matrix where N is the number of proteins, while each entry $a_{ij}$ = 1 if protein i interacts with protein j.

### 3.2 Data Exploration and Sub-network Construction

By the end of this step, we aim to build three independent regulatory networks, in a form similar to the final network described in section 2, each for one of the datasets. For knock-out data, we will explore the pattern by two hierarchical clustering analysis first on the genes and secondly on the arrays (or experiment), using only measurements that have p-value (indicating the uncertainty of a measurement different from zero) below a certain threshold (e.g. 0.01). We will examine significant clusters and make observations about groups of co-expressed genes and down-regulated/up-regulated relationships between groups. We will summarize our findings in a network in which the node represents a cluster of genes and edges denote the regulatory relationship between groups. Constructing the sub-networks for ChIPchip data and BIND dataset is straightforward by including an edge for every '1' entry in the data matrices. For the ChIPchip data, the direction of the edge always points from the node representing the protein to the node representing the gene. However, we cannot infer the type of the regulatory effect (activation/inhibition) from the ChIPchip data. From the BIND dataset, neither the direction nor the type of the edge can be inferred. The final networks for the two datasets will be absent of the pieces of information mentioned above.

### 3.3 Sub-network Combination

We will combine the three sub-networks obtained from the above step into a single network. To do this, we will map all the protein names to its corresponding gene name and find the genes that are present in any pairs of the three data sets (hereafter referred as linkage gene). We will also identify genes in each dataset that have behaviour similar to the linkage genes (hereafter referred as semi-linkage genes). The sub-networks will be combined by first including in it all the nodes for the linkage genes, and then augmenting it by adding edges selected from the three sub-networks. The edges from the sub-network are considered in the following order: firstly, the edges between linkage genes, secondly, the edges between linkage and semi-linkage genes, and between two semi-linkage genes, and lastly the edges that connect genes unique to one dataset.

Adding edges from BIND dataset and ChIPchip data is straightforward (simple additions). For each edge (edges both between clusters and within cluster links) in the graph from knock-out data, we will examine the two end nodes against the existing network, and choose a path that the nodes will likely be on; "likely" in the sense that product of the confidence measure of edges in this path is the highest. The confidence measure for each edge will be calculated using the formula

$$Conf(e) = w1*c1+ w2*c2+ w3*c3$$

where w1, w2,w3 represents the weight assigned to each sub-network, and c1,c2,c3 represents how well such edge is supported in the sub-network. An edge will be included in the final network if it has a confidence measure above a certain threshold. We will try different possible ways of assigning the value for c's and w's. One approach is to assign c1 for knock-out data as proportional to the $\log_{10}$(ratio) measurements, while c2 and c3 for BIND and ChIP-chip data are binary numbers reflecting the existence or non-existence of an edge. Several combinations of the network weight "w" are also possible, including making them all the same (i.e.,w1=w2=w3), or favouring the BIND and ChIP-chip dataset( i.e, w1<w2=w3). The ultimate goal of these different assignments is to find a network configuration that can best describe the regulatory relationship in the data. If time permits, we will also consider sequencing information to add more useful links to the final network.

**3.4 Network Robustness and Correctness Evaluation**

The combined network will be evaluated by looking at the number of false-positive edges and false-negative edges. The former can be found by randomly reordering the data samples and seeing if the network still shows some links, while the latter can be found by comparing the network with annotated data (e.g. the TRANSFAC® database) of known relationships. In addition, Professor Hughes has agreed to perform knock-out experiments for us, which is an additional way of verifying the correctness of the network.

## 4. Work Plan

Data pre-processing will be done before any analysis is carried out on the datasets. The construction of three sub-networks will be done in parallel by three different group members and the evaluation program to evaluate the correctness of the sub-networks will be developed by the fourth member in the meantime. Both the sub-networks and the final regulatory networks will be represented in MATLAB as an array of edges, each edge is a tuple consisting of source node name, destination node name, edge type and confidence measure.

Preliminary visualization will be done in parallel with the network combination step. The visualization tool will be developed using MATLAB GUI which provides a graphical interface to all the functions from data pre-processing to the sub-networks combination. The visualization tool will also contain input areas for user to adjust some parameters used in the computation such as threshold used for filter the data, weights assigned to each sub-networks when combining the networks. In addition, as we inserting edges to the final network, we will also populate a table, arranged by edge index, which records the evidences that support a particular edge. Some examples of such evidences are "protein 1 encoded by gene A binds to the promoter region of gene B", and "the knock-out of gene A, cause gene B to be down-regulated".

In case of failure in combining the sub-networks, we will focus our attention on the common genes among all the three datasets, and try to construct a network only for those common genes. Since the number of common genes will be significantly less than the entire datasets, we can design algorithms much simpler than that in the original case. Another alternative is to adopt the existing method as presented in Yeang et. al.[10] or the Bayesian approach as presented in [3] to construct the network. In any cases, we are still be able to successfully obtain three sub-networks, since the method we based on for constructing them are well researched and well practiced.

The testing of the sub-networks will be carried out as soon as they are constructed. The evaluation will be focused on computational correctness by testing it on a small set of data of some known relationships. For example, we will choose the genes from knockout data that are known to encode regulators, and the

genes that are known to be bind by those regulators. In this case, we know the links and their direction among these genes, and we therefore can verify manually that the sub-network constructed include these links correctly. The evaluation of the final network will be done by testing for false-positive and false-negative edges as described in section 3.4. In addition, we will also adjust parameters used when constructing the network, and for each resulting network configuration, plot an ROC curve with false-positive links/total number of links on x-axis and false-negative links/total number of links on y axis. The objective is to find the best set of parameters with which the final network constructed is able to find out a majority (e.g. 60-70%) of the known links among the same set of genes, while the ratio of false-negative edge to all the edges is relatively small (e.g. below 10-15%).

This project involves data mining, algorithm design, MATLAB programming, graphics design and biological analysis. The four members of the team, Charudutt, Qian, Daisy and Jian Ye, and have combined abilities to engage work in the aforementioned areas. Charudutt has previous experiences in DNA microarray experiments and are very familiar with different databases for yeast genome. He will be mainly responsible for biological interpretation and evaluation part of the project. Daisy has solid background in database systems, and has excellent mathematical and problem-solving skills. Qian is experienced in MATLAB programming, and familiar with several different machine learning algorithms such as clustering. Daisy and Qian will work together for design algorithms to construct the network. Jian Ye is good at graphics design, and will be mainly responsible for developing visualization tool for the network. A complete breakdown of the works is shown in the table below, where R means responsible and A means assisting. A Gantt chart is provided on page no. 10 outlining the estimated timelines for each tasks and milestones.

| | ***Tasks/Milestones*** | ***J.Y. Lin*** | ***C. Shah*** | ***D. Wang*** | ***O.Y. Tang*** |
|---|---|---|---|---|---|
| 1 | Researches in respective areas: gene regulatory network and clustering algorithm | | | R | R |
| 2 | Collecting datasets and understand the meaning of the numerical measurements in each dataset. | A | R | | A |

| | *Tasks/Milestones* | *J.Y. Lin* | *C. Shah* | *D. Wang* | *Q.Y. Tang* |
|---|---|---|---|---|---|
| 3 | Data preprocessing, including data import, conversion to matlab matrix, and missing value imputation. | R | | | A |
| 4 | Clustering Analysis on knock-out dataset | A | | | R |
| 5 | Constructing networks for BIND dataset and ChIPchip dataset | | A | R | |
| 6 | Explore the regulatory relationships in the clusters of the knock-out dataset, and construct a network for the dataset. | | | A | R |
| 7 | Three sub network testing and verification | | R | A | |
| 8 | Visualization tool development | R | A | | |
| 9 | Analysis of the connection between three sub-networks in terms of common nodes and edges | | | R | A |
| 10 | Network combination with different combination of network weights and confidence measurement assignments | | A | | R |
| 11 | Integrating the visualization GUI with the computational modules | R | | A | |
| 12 | Collecting and processing annotated data for evaluating the network | A | R | | |
| 13 | Design programs for testing the networks robustness and correctness | A | | R | |
| 14 | Collecting and summarize testing data | R | | | A |
| 15 | Gene-sequencing information analysis (motif), and improve the network based on this information | | R | | A |
| 16 | Report and poster preparation | A | A | A | R |

## 5. Financial Plan

The bulk of the costs for our project are performing biological experiments such as gene-knock out experiments. We intend to build software, develop algorithms and create the graphs using open-source code and non-proprietary software. License for MATLAB and computational resources for the project have been secured through Prof. Frey's Laboratory and IPL group at University of Toronto. The following two tables show a rough budget for our project and the possible sources of obtaining those supports.

Table 1: Expected cost for completing the project

| **Projected Expenses** | **Unit Price** | **Quantity** | **Total** |
|---|---|---|---|
| Gene Knock-Out Experiments (Complimentary -Courtesy Prof. Hughes from Banting and Best Institute) | ` | 10 | $0 |
| Gene Knock-out Experiments | $100 | 5 | $500 |
| Poster Board (Presentation) | $15 | 2 | $30 |
| Pamphlets (Demo) | $0.06 | 250 | $15 |
| Miscellaneous | - | | $100 |
| Supplies | - | | $55 |
| Other | $0 | 0 | $0 |
| ***Subtotals*** | | | $700 |
| ***Total Expenses*** | | | **$700** |

Table 2: estimated source of monetary supports

| **Sources of Income** | | | |
|---|---|---|---|
| Personal Contribution | | | 0 |
| - Daisy Wang | | | $50 |
| - Charudutt Shah | | | $50 |
| - Qian Ying | | | $50 |
| - Jason Lin | | | $50 |
| Support from Design Centre | | | $300 |
| Other sources of funding (Research Groups/ Departments) | | | $200 |
| Other | | | - |
| ***Total Income*** | | | $700 |

## 6. Technical Risks and Risk Mitigation

There are several risks that we will be confronted with as we move forward trying to complete our project. One of the risks is that the datasets we obtained contain many missing gene expression measurements and the corresponding p-values, due to the limitations of Microarray technology. To mitigate this risk, we will resort to mathematical approaches such as KNN-impute for missing value and gene selection methods based on maximizing the variations in gene expression level across the array. Such methods are followed in the industry and have proven to successfully address the problem. The second risk is the variations, and possibly conflicts, in data due to different collection methods and different experimental techniques followed by various labs. We will address this problem by introducing a weight to each of the dataset, reflecting our confidence of the information obtained from each dataset [9]. A larger weight means more likelihood of the correctness of the measurement while a smaller weight means the opposite. In this way, we can bias toward the information contained in one dataset over others. A closely related risk is that inherent noises in biological data make the inference of correct relationships difficult. It is possible that many relationships that we will discover are effects of noise and in fact they do not exist. To combat this problem, we will adopt a probabilistic measure (detailed description are included in section 3.3) for each relationship, and set a threshold on this measure to select the most likely relationships. The algorithm we propose may not work correctly. We will mitigate this risk by using small subset of the database to evaluate our algorithms and to modify our algorithm accordingly before applying to the entire datasets. In the worst case however, we will still be able to identify one-to-one causal relationships and form minimal connected sub-networks.

Failure to implement a fool-proof visualization tool either due to the large size of the network or the difficulties in interfacing to the MATLAB programs may also impede our progress. To minimize the risk of such failure, we will develop the software in parallel with algorithm design and implementation and interface to the functions as soon as they are written. As a back up plan, we will design a simplified search tool, instead of a full visualization for viewing the relationships found.

## 7. Market Analysis and Benefits to Canada or Society

The market value and social impact of the project come from two folds. First, the regulatory interactions we will discover between genes in yeast can contribute to the current knowledge and study in scientific community. Secondly, the approaches used in this project can be extended to analyze mouse and human genome; in which case, the regulatory interactions discovered could be used in drug discovery and could result substantial benefits in health care.

The visualization software that we develop will facilitate biologists and biomedical researchers for efficient finding of regulatory mechanisms in genes. The discovery of causal relationships from gene expression data in long term will have huge impacts on the society particularly in areas of drug discovery, disease diagnosis and disease treatments. The identification and discovery of new interactions between genes will provide pharmaceutical firms with means to identify drug targets. In addition, such software has great market potentials. It is estimated that current market value for a full-fledged extension of our software is worth $25millon CAD (IOBIO Informatics http://www.iobion.com), and such demand is very likely to grow rapidly within the next decades.

## *Self-Assessment*

### 1. UNDERSTANDING OF THE TECHNICAL PROBLEM AND ITS APPLICATION CONTEXT

D) The team has completed a detailed analysis of both the context and the scope of the technical problem that demonstrates a full understanding of the problem. The team has clearly used this analysis and the basis for the panning of the project.

Our team has researched the project well in advance and fully understands the complexity of the issue facing biologists and how machine-learning techniques can assist them in identifying relationships between genes and proteins. In addition, we have also done many researches in the current state of this problem, and reviewed the papers on existing methods.

### 2. PROPOSED SOLUTION: ADVANCING THE STATE OF THE ART

C) A significant advancement of more than one aspect of technologies or an application area.

Previous project are often focused on one dataset, however, we will look at three different datasets from different sources. Although similar in regulatory network representation as previous project, our project will differ from others with a regulatory network containing more information such as the evidence, and the confidence for each edge. Also, the animation feature of the visualization tool will be unique to our project.

### 3. TECHNICAL METHODOLOGY

C) An evaluation of a range of methodologies and the presentation of a framework for use of a feasible set of state of the art methodologies applied in an effective manner, which are likely to resolve the technical problem.

Our methodology involves preprocessing of data in order to minimize the effects of noise. Several machine learning algorithms exists however only some would work better over the other. Our approach to choosing such methodologies as k-nearest neighbor clustering or classification algorithms requires use to evaluate several methodologies and apply one that is most effective.

## 4. WORK PLAN (Milestones)

C) A detailed, comprehensive and achievable plan that is cost effective, and includes a WBS and a Gantt Chart outlining the scheduled work.

A detailed work plan is provided in section 4 and a Gantt chart is included on page 10. The nature of the project is such that each one of us requires having some basic biology which we have acquired over the summer. The remainder of tasks is split equally based on capabilities and preferences. We believe our methodology lays a clear path to take and outlines detailed procedures towards achieving the end goals.

## 5. FINANCIAL PLAN

B) Some financial information provided and an indication that project costs have been identified.

There are not many cost components to the project. Hence only brief tables with sources of funding and expected expenses are listed under section 5.

## 6. TECHNICAL RISKS, INCLUDING RISK MITIGATION

C) The proposal has identified the technical-risks likely to be faced by the project. The team has demonstrated a clear understanding of the nature and likelihood of the described risks, within the context of the current state of the art and the proposed technological innovation. The proposal presents a credible plan for the mitigation of the identified risks.

In this project we anticipate problems when combining the three different sources of data. As highlighted in the risks section we are concerned about generating false positives and evaluating the sensitivity of our results. Filling missing information and inconsistencies with known / annotated relationships is a risk that will help us evaluate the accuracy of prediction and generation of models.

## 7. MARKET ANALYSIS, BENEFITS TO CANADA/SOCIETY:

D) Demonstrates an understanding that the project has a significant new and enduring, high-value social benefit to Canada.

The project is a student-based project and clearly the contribution of this project to the society and market was a big motivation in choosing this project. We know that the tool will great assist scientist and people in R&D in field of bioinformatics an drug discovery. We have identified some research groups on campus that could benefit from our tool however we have not explored and concrete market opportunities. The entire field is in its infancy and hence it is unclear at this time of the sales & marketing advantages of our project.

## *References*


1. Chen T, Filkov V, Skiena S. *Identifying gene regulatory networks from experimental data.* Proceedings of the third annual international conference on Computational molecular biology, Lyon, France. 1999 April 11-14 : 4-103.

2. Friedman N, *Inferring cellular networks using probabilistic graphical models.* Science 2004 Feb 6; 303 (5659):799-805.

3. Friedman N, Linial M, Nachman I, Pe'er D. *Using Bayesian networks to analyze expression data.* Journal of Computational Biology 2000;7(3-4):601-20.

4. Hughes TR, Peng WT et al. *A panoramic view of yeast noncoding RNA processing.* Cell. 2003 Jun 27; 113(7):919-33.

5. Hughes TR, Marton MJ et al. *Functional Discovery via a Compendium of Expression Profiles.* Cell. 2000 Jul 7; 102(1): 109-126. (Dataset: http://www.rii.com/publications/2000/cell_hughes.html )

6. Lee TI, Rinaldi NJ et al. *Transcriptional regulatory networks in Saccharomyces cerevisiae.* Science. 2002 Oct 25;298(5594):799-804.
(Dataset available at: http://jura.wi.mit.edu/cgi-bin/young_public/navframe.cgi?s=17&f )

7. Phuong TM, Lee D, Lee KH. *Regression trees for regulatory element identification.* Bioinformatics. 2004 Mar 22; 20(5):750-7.

8. Tobler JB, Molla MN, Nuwaysir EF, Green RD, Shavlik JW. *Evaluating machine learning approaches for aiding probe selection for gene-expression arrays.* Bioinformatics. 2002;18 Suppl 1:S164-71.

9. Weaver DC, Workman CT, Stormo GD. *Modeling regulatory networks with weight matrices.* Pacific Symposium on Biocomputing. 1999:112-23.

10. Yeang CH, Ideker T, Jaakkola T. *Physical network models.* Journal of Computational Biology. 2004;11(2-3):243-62.

## *Glossary/ Terminology*

BIND: Biomolecular Interaction Network Database (http://www.bind.ca)

ChIP : Chromatin Immuno Precipitation

DNA: DeoxyRiboNucleic Acid

KNN: K Nearest Neighbors Algorithm

RNA: RiboNucleic Acid

ROC Curve: Receiver Operating Characteristic curve

P-value: Probability value is probability that is measure of how much evidence we have against the null hypotheses. It is the probability of wrongly rejecting the null hypothesis that the two genes do not interact when in fact they do.

TRANSFAC: A database on eukaryotic transcription factors, their genomic binding sites and DNA-binding profiles available on http://www.gene-regulation.com/