Executive Summary

Recent advances in data mining, bioinformatics and computational biology show promises in revolutionizing global healthcare, drug discovery and biotechnology. One of the key steps in this process is to understand interactions between genes and proteins. Our project "Generative Models for Analyzing Molecular Biology Data" centers on the use of machine learning algorithms to identify known relationship and discover previously unknown relationships between genes & proteins in yeast genome data. The methodologies can be employed to analyze mouse and human genome.

Current attempts to discover novel relationships are carried out manually by biological experiments, which yield very little understanding compared to the complexity of the problem. The goal of this project is to construct a network and develop a visualization tool that will allow molecular biologists and geneticists to systematically and efficiently examine and interpret data, advance their knowledge and discover the above-mentioned relationships between thousands of genes.

Researches in this area are being conducted only recently and no complete understanding of the entire regulatory mechanisms in genes has been gained. However, such an understanding will inevitably have huge impacts in areas such drug discovery, disease diagnosis and disease treatments, and ultimately our society as a whole. Therefore, we can foresee that the outcome of our project will have great market value and social benefits.

Unlike previous projects which are only based on one or two datasets, our project will focus on datasets profiling three different indications of the possible interactions exists in 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. 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.

To overcome these problems, we will adopt data preprocessing techniques such as KNN impute, and select genes with relatively trustable measurements in subsequent analysis. Secondly, we will use unsupervised learning algorithms such as clustering to explore the patterns in the data; such algorithms need very little prior knowledge. Lastly, we will introduce a confidence measure for each interaction we derived from the dataset. By different formulations of such measure, and by adjusting the threshold for an acceptable measure, we will be able to infer the most likely relationships from the existing datasets. Furthermore, to ensure the completeness of the final network, gene-sequencing information will be use to add newly predicted interactions.

All four members of the team– Charudutt Shah, Qian Tang, Jason Lin and Daisy Wang, possess diverse skills ranging from expertise in biology, machine learning algorithms, graphical programming and probabilistic modelling respectively. Our project consists of the following distinct tasks: 1) Pre-processing of Data 2) Clustering analysis and constructing sub-networks 3) combining sub-network with probabilistic measure and 4) Evaluating the model, verifying existing relationships with literature and novel relationships by performing gene knock-out experiments. 5) Developing a visualization tool.

We anticipate the project to be successfully completed by the end of February 2005. Our estimated budget stands at $700, which includes the expenses for Gene Knock-out Experiments, posters and other miscellaneous costs.

In conclusion, the software tool introduced in our project will enable medical professionals and scientists to determine new genetic relationships of biological importance that will potentially have a major impact on national and global health care and society.