THESIS

University of Western Sydney
Australia

presented on the 13th of October 2010
by
Mazin AOUF

Analysis of High Dimensionality Yeast Gene Expression Data using Data Mining

Supervised by

Principal Supervisor
Dr Liwan Liyanage
School of Computing and Mathematics

Co-Supervisors
Dr Mark Temple
School of Biomedical and Health Sciences

A/Prof Steve Hansen
School of Computing and Mathematics

© M Aouf 2010
THESIS

A dissertation presented to the
University of Western Sydney
School of Computing and Mathematics

fulfillment of the requirements for the degree of

Master of Science (Honours)
in Bio-Informatics

October 2010

By

Mazin AOUF

Analysis of High Dimensionality Yeast Gene Expression Data using Data Mining

© M Aouf 2010
Dedicated to

My Wife Laura

and

My Parents
Statement of Authentication

The work presented in this thesis is, to the best of my knowledge and belief, original except as acknowledged in the text. I hereby declare that I have not submitted this material, either in full or in part, for a degree at this or any other institution.

Mazin Aouf
University of Western Sydney
October 2010
Copyright

I authorize the University of Western Sydney to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose scholarly research.

Mazin Aouf
University of Western Sydney
October 2010
Abstract

Data Mining is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories.

The aim of this thesis was to enhance the effectiveness of the integration of hydrogen peroxide response data related to yeast gene expression data to obtain a protein response process model and to label a set of important genes related to this approach.

From biological studies, the Yeast Proteome Database (YPD) is a model/repository for the organization and presentation of genome-wide functional data. Accordingly, a yeast gene expression which is a unicellular DNA is selected which contains 6500 genes and the database combined with a number of related dataset to create a general dataset. DNA-binding transcriptional regulators interpret the genome’s regulatory code by binding to specific sequences to induce or repress gene expression. The gene products including RNA and protein are responsible for the development and functioning of all living membranes by 2 steps process, transcription and translation. Various transcription factors control gene transcription by binding to the promoter regions. Translation is the production of proteins from mRNA produced in transcription. In this study, out of the 169 transcription factors known to access yeast, we are considering those thought to be involved in the response of Hydrogen Peroxide (H_2O_2). They are 22 transcription factors. Each one is partitioned to 3 parts: TF with No H2O2, TF with Low H_2O_2 and TF with High H2O2.

Data were collected from multiple yeast datasets: “Harbison data” which holds the 22 Transcription Factors features, “Environ dataset” which includes the peroxide times features that help to create the Microarray Data Output (mRNA phase), and “Microarray dataset” which contains the Protein Response to H2O2 feature, to build a general dataset with 110 variables and 6103 observations.

Data Processing phase is carried out by using Enterprise Miner and the process consists of data integration, cleaning, variable transformation and then constructing data for modeling. Decision Tree Model is used to identify possible clusters within the data. This analysis prepared by three ways: gene to mRNA, gene to protein through mRNA.
and gene to protein without mRNA. Same analysis was done with the 5 transcription factors of Alpha treatment and demonstrates that there is no correlation between it and Protein Response phase on contrast as the H2O2 treatment.

Various studies have attempted to make genetic regulatory networks based on datasets derived from the whole-genome methodologies. In addition, several computational methods based on microarray data are currently used to study genome-wide transcriptional regulation. The previous research, prepared by Causton, H et al., in yeast gene expression data explains that a network describes interactions between diverse heterogeneous data leading to protein induction or repression in response to H2O2 treatment. The result for this study that applies multiple stresses to yeast cells had shown that the partition of data is still noisy and the work needs to evaluate their biological possibility.

The purpose of our study is to demonstrate that a huge numbers of yeast genes should be involved in various response biological changes and identify the global set of genes induced and repressed by binding DNA sequence, with initially the good processing of the data source. This process concludes a several important genes in each stage of the 4 ways discussed above.

**Research Methodology**

The research methodology is organized as follows:

- Understanding the application domain for gene expression data.
- Review of Data Mining Techniques.
- Review of existing clustering strategies for high dimensional dataset analysis.
- Understanding the yeast data.
- Data Preparation: collecting, integrating and collating the gene expression data on a database.
- Data analysis via Decision Tree methods using SAS software and data mining methods using Enterprise Miner Software.
- Compare H2O2 treatment result with Alpha treatment.
- Comparing the result with Causton Result.
- Conclusion and future direction.
Thesis Related Publications
(Refereed Conference papers)


Acknowledgments

For this thesis, I have the honor to be supervised by the Dr. Liwan Liyanage, Dr. Mark Temple and Associate Professor Steve Hansen. Dr. Liyanage, Dr. Temple and Ass/Prof Hansen have shown an incredible commitment to reach this research.

I must acknowledge my gratefulness to my primary supervisor, Dr. Liyanage, who’s given me a lot of time and information during my research. Dr. Liyanage is in my opinion the leader of data mining and has a reservoir of information in data mining. As part of her current interesting area for research, she always shows progress in this technology (e.g. conference papers, books) that is partially included in my thesis. I appreciate her critical provisions that aim to eliminate weakness points in my research outcomes (e.g. more analysis, explanation, originality). I would also like to thank her assistance in writing reports and conference papers.

I would like to express my gratitude to my secondary supervisor, Ass/Prof Hansen, who I have fortunately known prior to this degree by approximately three years. I appreciate his vast knowledge and skills in many areas (e.g. computing, Web technology, planning) that enhanced my research outcomes, and his dedicated efforts to improve my research skills. I can’t also ignore his assistance in writing reports and conference papers. I would also like to appreciate his efforts to develop this research through research collaboration schemes.

Also, I would like to thank my other secondary supervisor, Dr. Mark Temple from School of Biomedical and Health Science who has helped me from his deep background in Biology to collect the dataset related to gene expression data. I am also appreciative in his pains in analyzing the data.

Appreciation also goes out to Dr. Vojislav (Vic) Ilic from School of Engineering for the help in starting this research in its early initiative.

I must also acknowledge Professor Simeon Simoff, Professo...
Professor Yan Zhang, Professor Wei Xing Zheng, and other involved academic and office staff for their positive contribution to succeed my research by providing resources (e.g. related books) and approving two travels to attend two international conferences.

I am grateful to my University for the support it provided me through my thesis preparation. Thanks also go out to the School of Computing and Information Technology for all their computer and technical assistance throughout my program.

Very special thanks go out to Mrs. Marianne Oswald “School Finance and Travel Officer” who helped in my travels during my degree for thesis related publications.

I would like thank Mr. Les Krawczyk his support by installing the SAS and Answer Tree software used in this thesis.

I am also thankful to my wife Laura which is never stopped supporting me. I know that now she is proud of me and I hope that in my coming life, I shall know how to prove to be worthy of her.

To my parents, thanks for their support and for joy that they made me live.

A special thanks to my brother Rashad. He has stood behind me until the finish of the thesis and has supported me all the time.

I wish to express my sincere thanks to those who spent valuable time proofreading the manuscript.
# TABLE OF CONTENTS

1. **INTRODUCTION** .......................................................................................................................... 18
   1.1 Research Problem .......................................................................................................................... 18
   1.2 Background of the Study ............................................................................................................. 2
   1.3 Thesis Objectives .......................................................................................................................... 4
   1.4 Research Questions ....................................................................................................................... 5
   1.5 Future Direction ............................................................................................................................ 6
   1.6 Rationale ....................................................................................................................................... 7
   1.7 Current Status: Overview ............................................................................................................. 7
   1.8 Methodology ............................................................................................................................... 7
   1.9 Organization of the Thesis ........................................................................................................... 9

2. **RESEARCH FOUNDATIONS** ........................................................................................................ 12
   2.1 Introduction ................................................................................................................................. 12
   2.2 Data Mining and Knowledge Discovery ...................................................................................... 13
   2.3 Data Mining Process .................................................................................................................... 15
      2.3.1 Problem identification and Definition .................................................................................... 16
      2.3.2 Obtaining and Processing Data .............................................................................................. 16
      2.3.3 Data Mining Extracting and Knowledge ............................................................................... 16
      2.3.4 Result Interpretation and Evaluation ..................................................................................... 17
      2.3.5 Using Discovered Knowledge ............................................................................................... 17
   2.4 Data Mining Tasks and Functionalities ....................................................................................... 18
   2.5 Literature Review of Gene Expression Data Analysis .................................................................. 21
   2.6 Literature Review of Data Mining Techniques ............................................................................ 22
      2.6.1 Subspace Clustering Technique ......................................................................................... 22
      2.6.2 Decision Tree Technique .................................................................................................... 22
      2.6.3 Neural Networks Techniques .............................................................................................. 23
   2.7 Subspace Clustering Methods ..................................................................................................... 23
   2.8 Critical Review ............................................................................................................................ 26
      2.8.1 K-means Clustering ............................................................................................................... 27
      2.8.2 Kohenen Networks .............................................................................................................. 27
      2.8.3 Growing Cell Structures Networks ..................................................................................... 28
      2.8.4 Fuzzy Kohenen Neural Networks ...................................................................................... 28
      2.8.5 Methods’ Limitation ............................................................................................................ 29
   2.9 Recent Studies ............................................................................................................................. 29
      2.9.1 Gram Method ........................................................................................................................ 29
      2.9.2 Ihmels Method ...................................................................................................................... 30
      2.9.3 Sudarsanam Method ............................................................................................................ 30
2.10 ONGOING RESEARCH .......................................................................................................................... 30
2.11 CONCLUSION ............................................................................................................................................. 31

3. DATA PREPROCESSING ................................................................................................................................. 32

3.1 OVERVIEW OF YEAST GENE EXPRESSION DATA ANALYSIS ........................................................................ 32
3.2 RELATION BETWEEN GENE AND PROTEIN: BACKGROUND ......................................................................... 33
3.3 DATA COLLECTION AND DESCRIPTION ....................................................................................................... 35
3.4 VARIABLES DESCRIPTION ............................................................................................................................. 37
  3.4.1 The Relation between Attributes: ........................................................................................................... 38
  3.4.2 Summary of definition of the 110 variables: ............................................................................................... 40
3.5 DATA PREPROCESSING ................................................................................................................................... 46
  3.5.1 Data Integration ........................................................................................................................................ 47
  3.5.2 Data Cleaning ............................................................................................................................................ 48
3.6 CONCLUSION ................................................................................................................................................... 49

4. DATA MODELING .............................................................................................................................................. 50

4.1 RESEARCH DESIGN ....................................................................................................................................... 50
  4.1.1 Research Purpose ..................................................................................................................................... 51
  4.1.2 Research Approach ................................................................................................................................. 52
  4.1.3 Research Strategy .................................................................................................................................... 53
4.2 RESEARCH PROCESS ..................................................................................................................................... 53
4.3 SAS ENTERPRISE MINER ................................................................................................................................... 56
4.4 VARIABLE CONSTRUCTION .......................................................................................................................... 56
  4.4.1 Target variable Construction ..................................................................................................................... 58
  4.4.2 Predictor Variables/Variables Construction ............................................................................................. 58
    • From Gene to mRNA .................................................................................................................................... 58
    • From gene to Protein through mRNA ........................................................................................................ 59
    • From gene to Protein without mRNA ........................................................................................................ 59
4.5 GENE SELECTION ........................................................................................................................................... 60
  4.5.1 Predictor Variables/Variables Construction ............................................................................................. 60
4.6 DATA SAMPLING FOR TRAINING, VALIDATION AND TEST SET ................................................................... 61
4.7 DECISION TREE CLASSIFICATION ............................................................................................................... 62
4.8 BUILDING DECISION TREE MODELS USING TRAINING AND VALIDATION DATA SETS .................................. 62
  4.8.3 From Gene to Protein without mRNA ..................................................................................................... 68
4.9 MODELS EVALUATION ................................................................................................................................... 71
4.10 CONCLUSION ................................................................................................................................................ 72

5. RESULTS AND ANALYSIS ....................................................................................................................................... 73

5.1 RESULT OF VARIABLE CONSTRUCTION .....................................................................................................
5.2 RESULT OF GENE SELECTION ......................................................................................................................
5.3 RESULT OF DECISION TREE CLASSIFICATION MODEL ................................................. 78
   5.3.1 From Gene to mRNA ........................................................................................ 78
   5.3.2 Gene to Protein through mRNA...................................................................... 80
   5.3.3 From Gene to Protein without mRNA: ............................................................ 83
5.4 PREDICTION RESULT .............................................................................................. 85
5.5 IMPORTANT GENES AND TFs.............................................................................. 87
5.6 COMPARE RESULT.................................................................................................. 89
   5.6.1 Causton Study..................................................................................................... 89
   5.6.2 Alpha Treatment ............................................................................................... 90

6. CONCLUSION AND FURTHER RESEARCH .............................................................. 94
   6.1 CONCLUSION ........................................................................................................ 94
   6.2 RESEARCH LIMITATIONS .................................................................................. 96
   6.3 MANAGEMENT IMPLICATIONS .......................................................................... 97
   6.4 SUGGESTIONS FOR FURTHER RESEARCH ...................................................... 98

7. APPENDIX A: YEAST DATASET ANALYSIS USING HYDROGEN PEROXIDE
   TREATMENT ................................................................................................................ 100
   7.1 DATASET............................................................................................................... 100
   In this section, we select the main cases that are used to prepare dataset for
   analysis, including:........................................................................................................... 100
   7.2 IMPORTANT RESPONSES GENES....................................................................... 104
   7.3 ALL IMPORTANT GENES AND TFs..................................................................... 105

8. APPENDIX B: YEAST DATASET ANALYSIS USING ALPHA TREATMENT ...... 108
   8.1 DATASET............................................................................................................... 108
   8.2 IMPORTANT RESPONSES GENES....................................................................... 111

9. REFERENCES.................................................................................................................. 113
FIGURES

FIGURE 1.1: THE TWO-STEP PROCESS FROM GENE TO PROTEIN, TRANSCRIPTION AND TRANSLATION, BY WHICH THE INFORMATION IN GENES FLOWS INTO PROTEINS: DNA → RNA → PROTEIN [4]........... 3

FIGURE 1.2: GRAPHICAL PRESENTATION OF COMPARISON METHODOLOGY ........................................... 9

FIGURE 1.3: THESIS STRUCTURE ............................................................................................................. 11

FIGURE 2.1: THE KDD PROCESS [29] ..................................................................................................... 16

FIGURE 2.2: TASKS IN KDD PROCESS [29] ........................................................................................ 18

FIGURE 2.3: DATA MINING FUNCTIONALITIES ..................................................................................... 19

FIGURE 2.4: MATRIX OF GENE EXPRESSION DATA IN MICROARRAY .................................................. 22

FIGURE 2.5: ELEVEN OBJECTS IN ONE UNIT BIN (BY PARSONS, L.[34]) ............................................ 24

FIGURE 2.6: SIX OBJECTS IN ONE UNIT BIN (BY PARSONS, L. [34]) .................................................. 24

FIGURE 2.7: FOUR OBJECTS IN ONE UNIT BIN (BY PARSONS, L. [34]) ............................................. 24

FIGURE 3.1: THE CENTRAL DOGMA OF MOLECULAR BIOLOGY DESCRIBES THE TWO-STEP PROCESS, TRANSCRIPTION AND TRANSLATION, BY WHICH THE INFORMATION IN GENES FLOWS INTO PROTEINS: DNA → RNA → PROTEIN. [60] ........................................................................... 33

FIGURE 3.2: THERE ARE FOUR TYPES OF RNA, EACH ENCODED BY ITS OWN TYPE OF GENE. THE GENOMIC DNA CONTAINS ALL THE INFORMATION FOR THE STRUCTURE AND FUNCTION OF AN ORGANISM. IN ANY CELL, ONLY SOME OF THE GENES ARE EXPRESSED, THAT IS, TRANSCRIBED INTO RNA. [61] ......................................................................................................................................... 34

FIGURE 3.3: THE SEQUENCE OF A EUKARYOTIC PROTEIN-CODING GENE IS TYPICALLY NOT COLLINAR WITH THE TRANSLATED mRNA; THAT IS, THE TRANSCRIPT OF THE GENE IS A MOLECULE THAT MUST BE PROCESSED TO REMOVE EXTRA SEQUENCES (INTRONS) BEFORE IT IS TRANSLATED INTO THE POLYPEPTIDE. [62] ........................................................................................................... 34

FIGURE 3.4: THE "LIFE CYCLE" OF AN mRNA IN A EUKARYOTIC CELL. RNA IS TRANSCRIBED IN THE NUCLEUS; ONCE COMPLETELY PROCESSED, IT IS TRANSPORTED TO THE CYTOPLASM AND TRANSLATED BY THE RIBOSOME. AT THE END OF ITS LIFE, THE mRNA IS DEGRADED. [64] ........... 35

FIGURE 4.1: RESEARCH DESIGN OF THE STUDY ....................................................................................... 51

FIGURE 4.2: GENERAL OVERVIEW OF RESPONSE MODEL ..................................................................... 54

FIGURE 4.3: RESEARCH PROCESS .......................................................................................................... 55

FIGURE 4.4: RESEARCH PROCESS – DETAILED PROCEDURE .................................................................. 55

FIGURE 4.5: TREE FOR MICROARRAY DATA OUTPUT AS TARGET VARIABLE IN mRNA PHASE .......... 64

FIGURE 4.6: AVERAGE SQUARE ERROR FORM GENE TO mRNA ............................................................ 65

FIGURE 4.7: TREE FOR PROTEIN RESPONSE TO H₂O₂ AS TARGET VARIABLE IN PROTEIN INDUCED BY H₂O₂ PHASE WHEN mRNA INDUCE ......................................................................................... 66

FIGURE 4.8: AVERAGE SQUARE ERROR FROM GENE TO PROTEIN THROUGH mRNA WHEN PROTEIN INDUCED BY H₂O₂ ......................................................................................................................... 67
FIGURE 4.9: TREE FOR PROTEIN RESPONSE TO H₂O₂ AS TARGET VARIABLE IN PROTEIN REPRESSED BY H₂O₂ PHASE WHEN mRNA REPRESS ................................................................. 68
FIGURE 4.10: AVERAGE SQUARE ERROR FROM GENE TO PROTEIN WITHOUT mRNA WHEN PROTEIN REPRESSED BY H₂O₂ .................................................................................. 68
FIGURE 4.11: TREE FOR PROTEIN RESPONSE TO H₂O₂ AS TARGET VARIABLE IN PROTEIN INDUCED BY H₂O₂ PHASE WHEN WITHOUT mRNA ............................................................ 69
FIGURE 4.12: TREE FOR PROTEIN RESPONSE TO H₂O₂ AS TARGET VARIABLE IN PROTEIN REPRESSED BY H₂O₂ PHASE WHEN WITHOUT mRNA ............................................................ 70
FIGURE 4.13: AVERAGE SQUARE ERROR FROM GENE TO PROTEIN WITHOUT mRNA ............... 70
FIGURE 5.1: PREPARATION DATA FOR ANALYSIS ........................................................................ 78
FIGURE 5.2: TREE FOR MICROARRAY DATA OUTPUT AS TARGET VARIABLE IN mRNA PHASE ...... 79
FIGURE 5.3: AVERAGE SQUARE ERROR FROM GENE TO mRNA ................................................... 80
FIGURE 5.4: TREE FOR PROTEIN RESPONSE TO H₂O₂ AS TARGET VARIABLE IN PROTEIN INDUCED BY H₂O₂ PHASE WHEN mRNA INDUCE ................................................................. 81
FIGURE 5.6: AVERAGE SQUARE ERROR FROM GENE TO PROTEIN THROUGH mRNA ............... 82
FIGURE 5.7: TREE FOR PROTEIN RESPONSE TO H₂O₂ AS TARGET VARIABLE IN PROTEIN INDUCED BY H₂O₂ PHASE WHEN mRNA HAS NO CHANGE VALUE ................................. 83
FIGURE 5.8: AVERAGE SQUARE ERROR FROM GENE TO PROTEIN OF PROTEIN INDUCED BY H₂O₂ WHEN mRNA HAS NO CHANGE VALUE ................................................................. 84
FIGURE 5.9: TREE FOR PROTEIN RESPONSE TO H₂O₂ AS TARGET VARIABLE IN PROTEIN REPRESSED BY H₂O₂ PHASE WHEN mRNA HAS NO DATA ................................................................ 84
FIGURE 5.10: AVERAGE SQUARE ERROR OF PROTEIN REPRESSED BY H₂O₂ WHEN mRNA HAS NO DATA ...................................................................................................................... 85
FIGURE 5.11: WHOLE MODEL - 1756 GENES AFTER FILTERING DATASET FROM GENE TO PROTEIN THROUGH mRNA ................................................................................................................... 86
FIGURE 5.12: TREE FOR MICROARRAY DATA OUTPUT AS TARGET VARIABLE IN mRNA INDUCE PHASE ................................................................................................................................. 91
FIGURE 5.13: TREE FOR MICROARRAY DATA OUTPUT AS TARGET VARIABLE IN mRNA REPRESS PHASE ............................................................................................................................... 92
FIGURE 5.14: TREE FOR PROTEIN RESPONSE TO H₂O₂ AS TARGET VARIABLE IN PROTEIN INDUCED BY H₂O₂ PHASE WITH MICROARRAY DATA OUTPUT ......................................................... 92
FIGURE 7.1: 1756 GENES AFTER FILTERING DATASET AND THE WAY FROM GENE TO PROTEIN THROUGH mRNA BY ADDING H₂O₂ TREATMENT ........................................................................... 102
FIGURE 7.2: THE FEATURES NAMES USED FOR ANALYSIS ................................................................ 103
FIGURE 7.3: THE FEATURES NAMES USED FOR ANALYSIS (CONT.) .................................................. 103
FIGURE 8.1: 378 GENES AFTER FILTERING DATASET AND THE WAY FROM GENE TO PROTEIN THROUGH mRNA BY ADDING ALPHA TREATMENT ................................................................. 110
FIGURE 8.2: THE FEATURES NAMES USED FOR ANALYSIS ............................................................... 111
TABLES

Table 3.1: All cases for 22 transcription factors outcomes .................................................. 42
Table 3.2: All cases for 2 experiments (A and B) at zero time ........................................... 44
Table 3.3: All cases for two experiments at 40 and 60 minutes time .................................. 45
Table 3.4: All cases for the microarray data output .......................................................... 45
Table 4.1: Filter method ................................................................................................... 60
Table 4.2: Error for all models ......................................................................................... 71
Table 5.1: First step of variable construction .................................................................. 73
Table 5.2: Second step of variable construction .............................................................. 75
Table 5.3: Filter method of gene selection ...................................................................... 77
Table 5.4: The three important genes for 50% of mRNA induce by RPN4 outcome and HSF1
      outcome transcription factors ................................................................................... 79
Table 5.6: The important gene of protein induced by H2O2 when mRNA induce ........... 81
Table 5.7: The important gene of protein repressed by H2O2 when mRNA repress ....... 82
Table 5.8: The important gene of protein induced by H2O2 when mRNA has no change
      value ....................................................................................................................... 83
Table 5.9: The important gene of protein repressed by H2O2 when mRNA has no data .... 85
Table 5.10: The distribution output of the protein response to H2O2 with the outcome of
      microarray data output .......................................................................................... 86
Table 5.11: The twenty important genes with all their directions ................................... 87
Table 5.12: The thirteen important TFs with all their directions ..................................... 88
Table 5.13: The two important genes when mRNA induce ............................................. 91
Table 5.14: The three important genes when mRNA repress ........................................ 92
Table 5.15: The two important genes from gene to protein ............................................ 93
Table 6.1: Summary result ............................................................................................... 96
Table 7.1: 6103 genes before filtering dataset .................................................................. 101
Table 7.2: The distribution of genes between protein and mRNA phases ..................... 102
Table 7.3: Eleven genes when mRNA induce and protein induced by H2O2 with three
      important genes ...................................................................................................... 104
Table 7.4: Eight genes when mRNA repress and protein induced by H2O2 with no
      important genes ...................................................................................................... 104
Table 7.5: Five genes when mRNA induce and protein repressed by H2O2 with two
      important genes ....................................................................................................... 104
Table 7.6: Thirteen genes when mRNA repress and protein repressed and repressed by
      H2O2 with four important genes ............................................................................. 105
Table 7.7: Four genes when mRNA no change and protein induced and repressed by H2O2
      with one important gene ......................................................................................... 105
TABLE 7.8: SEVEN IMPORTANT GENES WHEN mRNA NO DATA AND PROTEIN INDUCED AND RERESSED BY H₂O₂.................................................................................................................. 105
TABLE 7.9: ALL IMPORTANT GENES IN ALL THE CASES WITH DESCRIPTIONS .................. 106
TABLE 7.10: ALL IMPORTANT TFs IN ALL THE CASES WITH DESCRIPTIONS ...................... 107
TABLE 8.1: 6103 GENES BEFORE FILTERING DATASET ................................................................ 109
TABLE 8.2: THE DISTRIBUTION OF GENES BETWEEN PROTEIN AND mRNA PHASES ............... 110
TABLE 8.3: THE TWO GENES WHEN mRNA REPRESS AND NO DATA VALUES WITH PROTEIN REPRESSED BY H₂O₂ AND TEC1 TRANSCRIPTION FACTOR UNBOUND ALPHA OUTCOME ...................... 112
TABLE 8.4: THE TWO IMPORTANT GENES WHEN mRNA INDUCE AND NO RESPONSE OR UNKNOWN PROTEIN BY H₂O₂........................................................................................................ 112
TABLE 8.5: THE THREE IMPORTANT GENES WHEN mRNA REPRESS AND NO RESPONSE OR UNKNOWN PROTEIN BY H₂O₂........................................................................................................ 112
Chapter One

1. Introduction

In this chapter a prologue of the thesis is introduced and provides the reader with an insight to the research area. The chapter begins with a background concerning the study area, followed by the thesis objective, which guides the reader to the research questions and related topics. Also a brief presentation of the overview of current status is provided. In addition, the methodology is outlined in this chapter. The study’s importance and the structure of the thesis are also highlighted.

Research Problem

The analysis of gene expression is not a new concept. Over the last decade, it has become possible to analyze genome wide gene expression using microarray technology. Many experiments have been done to study gene expression data under a variety of conditions. This study examines gene expression data using Data Mining Decision Tree models.

Data mining is the process of extracting patterns from large values of data. It is becoming an increasingly important tool to transform data into information and is used in wide range applications such as profiling practices, scientific discovery, surveillance and fraud detection.

Our research is applied to yeast gene expression data. It explores the path to go from gene to protein through mRNA. The way from gene to mRNA is by using Transcription Factors. When the mRNA is on, the path from mRNA to protein is by translation to check the response of protein.
Various studies have attempted to make genetic regulatory networks based on datasets derived from the whole-genome methodologies. In addition, several computational methods based on microarray data are currently being used to study genome-wide transcriptional regulation [1]. Research by Causton [2] introduced a central method in analyzing yeast gene expression data which gives that a network describing interactions between diverse genes data leading to protein induction or repression in response to H$_2$O$_2$ treatment. The results of this study which applies multiple stresses to yeast cells had shown that the data is noisy and farther work to evaluate the biological response is needed (see section 5.7 for more details).

The aim of this study is to improve the accuracy of this result of predicting protein response to hydrogen peroxide (H$_2$O$_2$). In addition, the purpose of our study is to demonstrate that a large numbers of yeast genes should be used to select the important genes that have protein responses to hydrogen peroxide and to highlight the importance of initial data cleaning and preprocessing of the data source. This will leads to identification of biological changes and the global set of genes induced or repressed by binding DNA sequence. Data preprocessing was achieved by using SAS Enterprise Miner software filtering techniques and classification was done using Decision Tree Models.

### 1.1 Background of the Study

Expressed genes are genes that are transcribed into messenger RNA (mRNA) and then translated into protein. Gene expression is a highly specific process in which a gene is switched on at a certain time and "speaks out" [3].

The figure below shows the central dogma of molecular biology from gene to protein phases.
Figure 1.1: the two-step process from gene to protein, transcription and translation, by which the information in genes flows into proteins: DNA → RNA → protein [4].

From biological studies, the Yeast Proteome Database (YPD) is a model/repository for the organization and presentation of comprehensive protein information [5]. Based on the scientific literature for the yeast Saccharomyces cerevisiae, YPD contains more than 50000 annotations lines derived from the review of 8500 research publications. The information concerning each of the approximately 6500 yeast proteins are structured around a convenient one-page format, the Yeast Protein Report, with additional information provided as pop-up windows.

Data Mining Techniques:

Decision Tree, Cluster Analysis and subspace clustering, as applied in Data Mining, seeks to discover groups, or clusters within different subspaces within a dataset. Even though technology has aided increased data collection, this has resulted in larger and more complex and noisy datasets. Traditional data mining methods take all the dimensions of an input dataset. However, high dimensional data usually includes irrelevant dimensions [6] which can mask existing models particularly in noisy data. This is particularly the case in applications such
expression data. There have been a number of approaches [7-13], to address the problems arising from this high dimensionality, such as Decision Tree.

However researchers have identified various limitations of these methods particularly in the areas of information integration systems, text-mining and bioinformatics [6]. So it is required to reduce the dimensionality of the given dataset in order to improve the efficiency and accuracy [14]. Moreover the running time of an algorithm certainly has to be minimized to achieve the desired results. This is being done (in this study) by using the same data set and applying the same decision tree technique with and without H$_2$O$_2$ treatment in the transcription phase from gene to mRNA. In comparison to the results of other studies, the results of decision tree were shown to be easy to understand.

In the bio-informatics domain, existing methodologies for gene expression data analysis are an ongoing topic focusing on disease diagnosis [15]. Addressing this intrinsic problem, various approaches have been developed. This research contributes to this body of knowledge by introducing data mining methodologies to discover new knowledge in bio-informatics domain.

### 1.2 Thesis Objectives

This study aims to:

**Primary objective:**

- “Investigation of methodologies to enhance the effectiveness of the integration of hydrogen peroxide response data related to yeast gene expression data to obtain a protein response process model and to label a set of important genes related to this approach.”

**Secondary objective:**

- Create a framework to collect, integrate and collate yeast gene expression data, gathering from number of different data sources, on a single database and analyze using data mining methods to study dynamics of protein response process model.
1.3 Research Questions

Following are the research questions covering steps taken in the context of data mining methodologies to analyze gene expression data with high dimensionality.

1. **What are the difficulties facing the collection of data?**

There are two kinds of data: primary and secondary data. In this study, secondary data are used. Three types of yeast gene expression data were collected from different datasets: Harbison dataset [16-19], Environ Dataset [20] and Microarray dataset [2] (See chapter three). The main difficult is finding data sources with relevant data representing the characteristic of the DNA sequence process with the common systematic name of gene (ORF) which is essential for data integration.

2. **What is the role of integration of yeast gene expression data with hydrogen peroxide ($\text{H}_2\text{O}_2$)?**

In order to develop yeast gene expression data, the methods used are to implement Hydrogen Peroxide ($\text{H}_2\text{O}_2$) in the analysis and see the reaction for Microarray Data in mRNA and protein phases, if they are induced or repressed and how can control them (See chapter three).

3. **What are the challenges facing high dimensionality data in gene expression?**

Due to the high dimensionality of genetic data set, subspace clustering has been introduced where it seeks to find clusters in different subspaces within a dataset. Traditional clustering is performed by feature selection and removing irrelevant and redundant dimensions by analyzing the entire dataset (See chapter three). In this study, they are 110 dimensions illustrating high dimensionality problem.

4. **How can filter method help for analysis?**

In order to obtain reliable outcomes, it is essential to use a cleaned data set. For that reason, a filter method was applied to delete all the unwanted data (See chapter four).

5. **How to analyze the yeast gene expression data process?**
Chapter 1. Introduction

After filtering data set, the yeast gene expression data was analyzed in SAS Enterprise Miner software by using the partition data sets and then building decision tree models to study the response of protein to identify the genes responding to treatment (See chapter four and chapter five).

6. How can Alpha treatment be used to improve the result?

In this study, hydrogen peroxide (H2O2) was added as a treatment in mRNA phase to test the response of protein (See chapter five). To validate this result, another experiment was conducted by addition of another treatment (Alpha treatment) instead of H2O2 and compared the results. Comparison is done by using correlation coefficient between gene and protein phases of the two treatments. In addition, this result was evaluated with a previous similar approach - Causton approach (See chapter six).

1.4 Future Direction

Additional datasets can be added to include more genes in the given dataset to create a general framework to analyze yeast gene expression data. This framework can further extend to analyze gene expression data at PhD level.
1.5 Rationale

The insufficient control of huge amount data within gene expression makes the study for this case hard. For this reason, the data mining analysis and the use for algorithm make the learning easier.

This study makes use of one of the important models in data mining, which aims to improve the prediction based on the Decision Tree Model, to examine widespread of protein in the genes. This model is briefly explained in Chapter Four.

1.6 Current Status: Overview

Data mining study is one of the major tools that can assist in many disciplines. From the last studies, classification of gene expression data has emerged as an active area of research recently. Consequently, various techniques are proposed by researchers to enhance the outcome of their genetic data analysis. In [2002], Bellaachia, A. et al. developed a clustering algorithm for the analysis of gene expression data. In this example, the results have shown a better performance than hierarchical algorithm [21]. Following, Jiang, D. et al. [22] have proposed another method based on indexing patterns to facilitate clustering methods to explore coherent patterns in gene expression data [22]. The results have indicated a good agreement with other studies. However, both studies were still noisy and additional work is still required in order to obtain an effective method for mining gene expression with accurate classification. This type of methods is highly recommended in the field of bio-informatics to help in diagnosis and treatment of protein response to $\text{H}_2\text{O}_2$.

From the above examples, it is observed that classification of gene expression data requires further enhancement due to the complexity of its molecular data structure.

1.7 Methodology

Laboratory experiment:
As part of this study, a gene directs the synthesis of a protein by a two-step process. First, the instructions in the gene in the DNA are copied into a messenger RNA (mRNA) molecule. The sequence of nucleotides in the gene determines the sequence of nucleotides in the mRNA. This step is called transcription. Second, the instructions in the messenger RNA are used by ribosomes to insert the correct amino acid residue in the correct sequence to form the protein coded for by that gene. The sequence of nucleotides in the mRNA determines the sequence of amino acid residues in the protein. This step is called translation.

The research methodology is organized as follows:

- Understanding the application domain for gene expression data.
- Review of Data Mining Techniques.
- Review of existing clustering strategies for high dimensional dataset analysis.
- Understanding the yeast data.
- Data Preparation: collecting, integrating and collating the gene expression data on a database.
- Data analysis via Decision Tree methods using SAS software and data mining methods using Enterprise Miner Software.
- Compare H₂O₂ treatment result with Alpha treatment.
- Comparing the result obtained in this study with Causton Results.
- Conclusion and future directions.
Chapter Three details the dataset preparation (collect, integrate and collate datasets) Chapter Four the analysis process (clean, explore and classification data).

1.8 Organization of the Thesis

This thesis consists of six chapters. Figure 1.3 shows a graphical representation of the thesis structure. In the first chapter, the background of the selected research area is presented followed by research objectives that end with the research questions covering the entire process.

The literature and critical reviews in Chapter Two will give the reader an overall review on theories relevant to research area and a
related to the topic. This chapter provides an overview of gene expression data analysis and data mining techniques include subspace clustering techniques. In addition, it contains recent studies of data mining methodologies and gene expression data.

In Chapter Three, an overview of yeast gene expression data analysis is described. This chapter consists of detail description on data collection, integration and cleaning. The collected database consists of 110 attributes and brief descriptions of these are given in this chapter.

In Chapter Four, first the design and process of this research is explained. Further, the decision tree model building process procedure is explained in detail step by step for each objective listed. Finally, model evaluation is discussed.

Chapter Five give result for each listed objective. Also, a comparison of our result is mentioned with Causton study and Alpha treatment (the same process as H2O2 treatment was done for Alpha treatment for comparison).

Chapter Six is just to summarize and present the highlights of the study. In addition, limitations of this study and suggestions for further research are provided in this chapter.
Figure 1.3: Thesis structure
Chapter Two

2. Research Foundations

This chapter is based on the previous introduction and the problem presented in chapter one. The aim of this chapter is to provide the reader with a literature review concerning the research area. It will provide an introduction, background of gene expression data and data mining techniques including subspace clustering techniques applied for genetic data classification. Also a brief explanation of and a critical review of some clusters and the recent studies to make genetic regulatory networks based on datasets is derived from the whole-genome methodologies. In the end, the ongoing research related to this area and the conclusion.

2.1 Introduction

Classification of gene expression data has explored in the recent years. This can aid in the development of efficient methodology in the field of bio-informatics for used in tumours diagnosis and treatment. Data mining is an effective technique being used in this field. One of the difficulties facing this technology is the inappropriate classification methods that examine complex structure of gene expression data. In this chapter, we give a brief introduction of gene expression data with experiment and we have made a critical review of major techniques being applied in the field of gene expression data with help of data mining. It can be seen that researchers have developed various techniques for gene data classification. In addition, they may differ from one to another whereas results are still showing the need for enhancement in this field. Some of these techniques are addressed in this chapter in term of advantages and disadvantages. Accordingly, the deoxyribonucleic acid (DNA) is considered as the smarts of the tumour-derived factors. Anal:
gene expression may give rise for diagnosis enhancement of affected tissues in their early stages.

For that reason, an ongoing research is addressing the problem of subspace clustering methodologies suitable for high dimensional datasets and verification of the new methodologies using appropriate datasets, particularly suitable for the analysis of gene expression data. In this context, researchers have identified various limitations of these methods particularly in the areas of information integration systems, text-mining and bio-informatics. This chapter aims to provide an overview of the published literature with a particular focus on the current status of subspaces clustering for knowledge discovery. This is considered to be an essential step in attempt to overcome the limitations and provide effective statistical model in sense of genetic knowledge discovery.

2.2 Data Mining and Knowledge Discovery

The digital revolution has made digitized information easy to capture, process, store, distribute, and transmit [23]. With significant progress in computing and related technologies and their ever-expanding usage in different walks of life, huge amount of data of diverse characteristics continue to be collected and stored in databases. The rate at which such data are stored is growing phenomenally [24]. We can draw an analogy between the popular Moore’s law [25] and the way data are increasing with the growth of information in this world of data processing applications. The advancement of data processing and the emergence of newer applications were possible, partially because of the growth of the semiconductor and subsequently the computer industry. According, to Moore’s law, the number of transistors in a single microchip is doubled every 18 months, and the growth of the semiconductor industry has so far followed the prediction. We can correlate this with a similar observation from the data and information domain. If the amount of information in the world doubles every twenty months, the size and number of databases probably increases at a similar pace. Discovery of knowledge from this huge volume of data is a challenge indeed. Data mining is an attempt to make sense of the information explosion embedded in this huge volume of data.
There exist several domains where large volumes of data are stored in centralized or distributed database including gene data clustering. Raw data are rarely of direct benefit. Its true value is predicated on (i) the ability to extract information useful for decision support or exploration and (ii) understanding the phenomenon governing the data source. In most domains, analyst process data manually and, with the help of statistical techniques, provide summaries and generate reports. However, such an approach rapidly breaks down as the size of data grows and the number of dimensions increases. Nowadays databases contain number of data on the order of $10^9$ or above and dimension on the order of $10^3$ are becoming increasingly common. When the scale of data manipulation and exploration goes beyond human capacities, people need the aid of computing technologies for automating the process.

All these have prompted the need for intelligent data analysis methodologies, which could discover useful knowledge from data. Data mining is nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [23]. There are other definitions:

- Data Mining is the process of exploration and analysis of large quantities of data in order to discover meaningful patterns and rules from large amounts of data [27]

- Data Mining is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories. Simply stated, data mining refers to extracting or “mining” knowledge from large amounts data [28]

Actually the major reason that data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge [28]. The data mining process is sometimes referred to as knowledge discovery or KDD (knowledge discovery in databases). The term “KDD” refer to the overall process of discovering useful knowledge from data. There is a difference in understanding the terms “knowledge discovery” and “data mining” between people from different
Knowledge discovery in databases is the process of identifying valid, novel, potentially useful, and ultimately understandable patterns/models in data. Data mining is a step in the knowledge discovery process consisting of particular data mining algorithms that, under some acceptable computational efficiency limitations, finds patterns or models in data [29].

Shapiro in 1989 invented the KDD term. He said that the term “Knowledge Discovery in Databases” (KDD) became popular in the AI and Machine Learning Community. However, the database researchers were on better speaking terms with the business folks and the press, and the term “data mining” became much more popular in the business press. Data mining term is older than KDD term which was invented in Statistics Data Analysis Association [30]. Data mining is actually a misnomer term. Mining of gold from rocks or sand is referred to as gold mining rather than rock or sand mining. Thus, data mining should have been more appropriately named “Knowledge mining from data,” which is unfortunately somewhat long. “Knowledge mining,” a shorter term may not reflect the emphasis on mining from large amounts of data. Nevertheless, mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material. Thus, such a misnomer that carries both “data” and “mining” became a popular choice [28]. Although Knowledge Discovery term is more appropriate but is assumed that data mining is equal to Knowledge Discovery.

2.3 Data Mining Process

Generally KDD is an iterative and interactive process involving several steps. This KDD process was chosen (Figure 2.1) according to UNESCO definition because of its simplicity and comprehensiveness.
2.3.1 Problem identification and Definition

The first step is to understand the application domain and to formulate the problem. This step is clearly a prerequisite for extracting useful knowledge and for choosing appropriate data mining methods in the third step according to the application target and the nature of data.

2.3.2 Obtaining and Processing Data

The second step is to collect and preprocess the data. Today’s real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size (often several gigabytes), and their likely origin in multiple, heterogeneous sources. Low quality data will lead to low quality mining results. Data preprocessing is an essential step for knowledge discovery and data mining. Data preprocessing include the data integration, removal of noise or outliers, the treatment of missing data, data transformation, and reduction of data, etc. This step usually takes the most time needed for the whole KDD process. Detail description of data preprocessing methods and techniques are discussed in Chapter Three.

2.3.3 Data Mining Extracting and Knowledge

The third step is data mining that extracts patterns and/or models hidden in data. This is an essential process where intelligent methods are appli
data patterns. In this step, we have to select first data mining tasks and then data mining method. The major classes of data mining methods are predictive modeling such as classification and regression; segmentation (clustering) and association rules which are explained in detail in section 2.4.

2.3.4 Result Interpretation and Evaluation

The fourth step is to interpret (post-process) discovered knowledge, especially the interpretation in terms of description and prediction which is the two primary goals of discovery systems in practice. Experiments show that discovered patterns or models from data are not always of interest or direct use, and the KDD process is necessarily iterative with the judgment of discovered knowledge. One standard way to evaluate induced rules is to divide the data into two sets, training on the first set and testing on the second. One can repeat this process a number of times with different splits, and then average the results to estimate the rules performance.

2.3.5 Using Discovered Knowledge

Putting the results into practical use is certainly the ultimate goal of knowledge discovery. The information achieved by data mining can be used later to explain current or historical phenomenon, predict the future, and help decision-makers make policy from the existed facts [29].

Tasks in KDD process in more details are shown in Figure 2.2
2.4 Data Mining Tasks and Functionalities

The CRISP-DM methodology distinguishes between the reference model and the user guide and is a new blueprint for data mining models [31].

Several data mining problem types or analysis tasks are typically encountered during a data mining project. Depending on the desired outcome, several data analysis techniques with different goals may be applied successively to achieve a desired result. In general, data mining tasks can be classified into two categories: descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions [28].

Based on the different mining tasks, we can categorize data mining functionalities (methods) as classification, clustering, regression, association rules, sequence discovery, prediction and so on [32]. Data mining functionali
the kind of patterns to be found in data mining tasks [28]. Data mining functionalities are shown in Figure 2.3.

**Figure 2.3: Data mining functionalities**

According to Berry and Linoff [33] basic data mining functionalities are: classification, estimation, prediction, affinity grouping or associating rules, clustering, and description and visualization. The first three are all examples of directed data mining, where the goal is to find the value of particular target variable. Affinity grouping and clustering are undirected tasks where the goal is to uncover structure in data without respect to a particular target variable. Profiling is a descriptive task that may be either directed or undirected.

- **Classification (supervised learning):** Classification maps data into predefined group or classes. Because the classes are determined before examining the data, classification is often considered as supervised learning. Classification algorithms require that the classes be defined based on data attribute values. They often describe these classes by looking at the characteristics of data which are already known to belong to the classes. Classification techniques are: Decision Tree: CART, C4.5, Bayesian Classification: Consists of two types, Naïve Bayesian Classification and Bayesian Belief Networks, Neural Networks, Support Vector Machines, Associative Classification, Lazy Learners (Learning from Your Neighbors):
k-Nearest Neighbor Classifiers, Case-Based Reasoning. Other classification methods: Genetic Algorithms, Rough Set Approach and Fuzzy Set Approach [33].

- **Estimation:** Estimation deals with continuously valued outcomes. Given some input data, estimation is used to assign a value for some unknown continuous variable [33].

- **Prediction:** Based on past and current data, many real-world data mining applications can be considered as predicting future data states. Prediction is viewed as a type of classification. The difference is that prediction is predicting a future state rather than a current state. Actually the difference is on the emphasis, since in predictive tasks the records are classified according to some predicted future behavior or estimated future value. With prediction, the only way to check the accuracy of the classification or the estimation is to apply the model and then evaluate if its performance was the preferred. Prediction applications include gene expression morphology [33].

- **Affinity grouping or associating rules:** Association rules alternatively referred to as affinity analysis. An association rule is a model that identifies specific types of data associations. They are usually used in the retail sales community to identify items which are often purchased together. The task of affinity grouping is to determine which things go together (e.g. what usually goes together at gene expression data) [33].

- **Clustering (unsupervised learning):** Clustering is the task of segmenting a diverse group into a member of more similar subgroups or clusters. What distinguishes clustering from classification is that clustering does not rely on predefined classes, examples, or target concepts. Clustering analysis data objects without consulting a known class label. In general, the class labels are not present in the training data simply because they are not known to begin with. Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity. That is, clusters of
that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Each cluster that is formed can be viewed as a class of objects, from which rules can be derived. Clustering is often done as prelude to some other form of data mining or modeling. Clustering techniques are: Partitioning: K-means, and K-medians, Hierarchical, Density based, and Model based [33].

- **Descriptive and visualization:** Sometimes the purpose of data mining is simply to describe what is going on in a complex database, in a way that increase our understanding of gene expression data such as in disease diagnosis. A good enough description of a behavior will often suggest an explanation for it as well, or at least where to start looking for it [33].

### 2.5 Literature Review of Gene Expression Data Analysis

In order to compare messenger RNA (mRNA) abundance, a typical experiment is being applied in two different samples (e.g., treatment versus control), known by microarray experiment. The extraction of RNA from the treatment and control cells make use of two different fluorescent approaches (i) red dye for the RNA from the treatment population and (ii) a green dye for that from the control population. Both extracts are washed over the microarray. Following, the array is excited by a laser. If the RNA from the treatment population is in abundance, the spot will be red, if the RNA from the control population is in abundance, it will be green. If both populations are equal then, the spot will be yellow. Thus, the relative expression levels of the genes in the populations can be estimated.

In microarray, the gene expression data set can be noted by the following matrix $M \{ w_{ij} | 1 \leq i \leq n, 1 \leq j \leq m \}$, where the rows ( $G \{ g_1, ..., g_n \}$ ) from the expression patterns of genes, the columns ( $S \{ s_1, ..., s_m \}$ ) from the expression profiles of samples, and $w_{ij}$ is the measured expression level of gene $i$ in sample $j$ [34]. Thus, $M$ is defined as:
The figure 2.4 above has the $w_{ij}$ values which are usually the mean or median values of the ratios of red / green fluorescent pixels. Our experiments showed that the microarray data is noisy and that several normalization methods can be applied.

### 2.6 Literature Review of Data Mining Techniques

#### 2.6.1 Subspace Clustering Technique

Due to the high dimensionality of genetic data set, subspace clustering has been introduced where it seeks to find clusters in different subspaces within a dataset. Traditional clustering is performed by feature selection and removing irrelevant and redundant dimensions by analyzing the entire dataset. However, the application of this method is inappropriate and thus will hide existing clusters particularly when data is noisy. As given in Parsons and Liu, subspace clustering algorithms localize the search for relevant dimensions allowing them to find clusters that exist in multiple, possibly overlapping subspaces [35]. From their research, two major branches of subspace clustering based are examined. Top-down algorithms where full set of dimensions are used to find the initial cluster and evaluate the subspaces of each cluster, iteratively improving the results. Bottom-up approaches where dense regions in low dimensional spaces are selected initially and combine them to form clusters. The following section addresses other subspace methods that are currently in use.

#### 2.6.2 Decision Tree Technique

A decision tree is a predictive model that, as its name implies, can be viewed as a tree. Specifically each branch of the tree is a classification question and the leaves of the tree are partitions of the dataset with their classification.
Often times this technique was originally developed for statisticians to automate the process of determining which fields in the database were actually useful or correlated with the particular problem that it was trying to understand. Partially because of this history, decision tree algorithms tend to automate the entire process of hypothesis generation and then validation much more completely and in a much more integrated way than any other data mining techniques.

Because decision trees were originally developed to understand simple predictive model, so they can’t be used in a wide variety of gene expression data for prediction. Concerning this problem, other methods will take place to analyze critical features with data mining.

2.6.3 Neural Networks Techniques

Neural networks are very powerful predictive modeling techniques but some of the power comes at the expense of ease of use and ease of deployment. As we will see in this section, neural networks, create very complex models that are almost always impossible to fully understand even by experts. In case of gene expression data expression, this can be much more complicated.

Because of the complexity of these techniques much effort has been expended in trying to increase the clarity with which the model can be understood by the end user. These efforts are still in their infancy but are of tremendous importance since most data mining techniques including neural networks are an active area of research to analyze problems in genetic data classification.

2.7 Subspace Clustering Methods

Recently, subspace clustering methods have been a rich area of investigation for genetic data classification. Therefore, an outline of advanced methods of subspace clustering is given in this section. In addition, a critical review has taken a place to present weakness of these methods.

In their survey, Parsons, L. et al [35], have identified two difficulties that challenge clustering methods. In high dimensional data, many
often irrelevant. These irrelevant dimensions can confuse clustering algorithms by hiding clusters in noisy data. Second reason that many clustering algorithms struggle with high dimensional data is the curse of dimensionality. As the number of dimensions in a dataset increases, distance measures become increasingly meaningless. Additional dimensions spread out the points until, in very high dimensions. As resultant they are almost equidistant from each other. In Figure 1, 2, and 3, Parsons observed a progressive decrease of the portion of the points when dimensions are increased. It can be seen the distribution of 20 points between 1 and 3 dimensions.

Figure 2.5: Eleven objects in one unit bin (by Parsons, L.[35])

Figure 2.6: Six objects in one unit bin (by Parsons, L. [35])

Figure 2.7: Four objects in one unit bin (by Parsons, L. [35])

This thought has been although proved by Jiang, D. et al [36], in their review, studies shown that in high dimensional data set, distance among features become meaningless. One of the commonly used methods to measure the distance between two data objects is the Euclidean distance. This is defined as:
where \( o_{id} \) is the value of the \( d \)th feature of the \( i \)th data object and \( p \) is the number of the features. Object is defined as a numerical vector \( \hat{O}_i = \{o_{ij} \mid 1 \leq j \leq p\} \)

However, in data gene expression, the individual magnitudes of each feature are not meaningful. Therefore, the overall shapes of the gene expression patterns are of greater interest. To measure the similarity the shapes of two expressions patterns, Pearson’s correlation coefficient is used. This is defined as follows:

\[
Pearson(O_i, O_j) = \frac{\sum_{d=1}^{p} (o_{id} - \mu_i) (o_{jd} - \mu_j)}{\sqrt{\sum_{d=1}^{p} (o_{id} - \mu_i)^2} \sqrt{\sum_{d=1}^{p} (o_{jd} - \mu_j)^2}}
\]

Where \( \mu_i \) and \( \mu_j \) are the means of \( \hat{O}_i \) and \( \hat{O}_j \), respectively. With Pearson’s correlation coefficient, each object is viewed as a random variable with observations and the similarity between two objects is measured by calculating the linear relationship between the distributions of the two corresponding random variables.

This coefficient has proven effective as a similarity measure for gene expression data [36-39]. However, empirical study has shown that it is not robust with respect to outliers [40], thus potentially yielding false positives which assign a high similarity score to a pair of dissimilar patterns. If two patterns have a common peak or valley at a single feature, the correlation will be dominated by this feature, although the patterns at the remaining features may be completely dissimilar.

Toward addressing these challenges, method called geometric hashing for comparison of 3d geometric structures was proposed by Tan and Tung, in which he...
argues that finding clusters of molecules satisfying a required property is more important than finding precise structural difference between pair of molecules [41].

In 2005, a new framework have been proposed by Kriegel, H. P. et al, which is based on an efficient filter refinement architecture that scales at most quadratic with respect to the data dimensionality and the dimensionality of the subspace clusters. Major advantage of this framework is that it can be applied to any clustering notions including notions that are based on a local density threshold. A broad experimental evaluation on synthetic and real-world data empirically shows that this method achieves a significant gain of runtime and quality in comparison to state-of-the-art subspace clustering algorithms [42]. This framework also shows that the gene expression datasets is not highly addressed. From here is the obvious need for meaningful biological subspace clusters.

Gan, G. & Wu, J. [43] has also designed an iterative algorithm called SUBCAD for clustering high dimensional categorical data sets, based on the minimization of an objective function for clustering. They deduced some cluster memberships changing rules using the objective function. The subspace associated with each cluster is also achieved making use an objective function. They proved various properties of this objective function that are essential for us to design a fast algorithm to find the subspace associated with each cluster. Finally, they carried out some experiments to show the effectiveness of their proposed method and the algorithm. A major problem that is challenging this method is that it required the number of clusters as input.

Finally, subspace clustering is still an open area for more contribution from researchers. We will be exploring all the listed clustering methods to select the most appropriate method and refine the existing methodologies to improve prediction and detection of affected tissues.

2.8 Critical Review

Generally classification technique aims to discover and describe classed, categories, or groups of objects contained within data. Some of the classical
techniques, such as the k-means algorithm and Kohonen nets, assign one class to each object [44].

However this can be a simplification, because in some case, objects can be partially assigned to two or more classes. Therefore, this problem is addressed by fuzzy clustering algorithms by allowing objects to be gradual members of two or more groups or classes. In the following subsections we briefly describe a classical clustering method (k-means), two neural methods (Kohonen and growing cell structures networks), and a fuzzy-neural hybrid method (fuzzy Kohonen network). In parallel, we throw the light on the weakness of these methods.

### 2.8.1 K-means Clustering

K-means is one of the common studied clustering methods. Following is a sample given by Anderberg [45]. He initiates a simple algorithm with certain number of clusters (the parameter k). Then, they implement the following: (1) the cluster centroids (k points) are chosen randomly; (2) the case are assigned to the clusters depending on their distance to the centroids; (3) next new centroids of the clusters are calculated by averaging the positions of each point in the cluster along each dimension moving the position of each centroid. This process aims to stop changing in boundaries of clusters. Therefore it is repeated from step (2). The performance of k-means clustering is highly dependent on the initial seed centroids. Therefore the result of this method is often suboptimal.

### 2.8.2 Kohonen Networks

This method is also known as self-organizing feature maps [46]. It consists of input and output layer networks that comprised of sensory receptors and of processing units respectively. From the Manuals of the data Engine, it is given that “Each processing unit is connected with a set of sensory receptors. Accordingly, the lateral interaction between the processing units follows a centre-excitatory and surrounding-inhibitory scheme. The latter represents a useful tool to implement the local cooperation and global competition between processing units. To optimize the potential complexity of this computation strategy, a logic defined as "winner takes all" is usually applied. Based on this, the winner for a particular
unit with highest activation. Synaptic weights of the winning unit and those within a defined neighborhood are then updated. Arising from this, there is an issue with Kohonen networks is that the number of classes that a network can determine is directly related to the number of neurons or units in the network. Consequently, the formation of new classes is impossible as the number of the classes is pre-defined and fixed. This issue is addressed by the so-called growing cell structures neural networks. Other issues of the Kohonen net approach include its dependence on the sequence in which the learning data are presented and its lack of robust convergence criteria.” All these issues are addressed by fuzzy Kohonen networks.

2.8.3 Growing Cell Structures Networks

Growing cell structures (GOS) neural networks are a variation of the Kohonen networks. Fitzke [47] did in his study Growing Cell Structures: A Self-Organizing Network for Unsupervised and Supervised learning. Major advantages of the GCS method over Kohonen nets can be presented as follows: (1) no-predefinition necessary as it is self-adaptive topology which is highly independent of the user; (2) in contrast to Kohonen nets, a GCS net requires only a small number of constant parameters and there is no need to define time-dependent or decay schedule parameters; and (3) construct an incremental and dynamic learning system which is derived from its ability to interrupt a learning process or to continue a previously interrupted.

2.8.4 Fuzzy Kohonen Neural Networks

As given by Kohonen in his study of Self-organized Formation of Topologically Correct Feature Maps. A Fuzzy Kohonen network is a model that combines the fuzzy sets and Kohonen networks. It consists of two parts: Kohonen network and a fuzzy c-means algorithm. The use of both techniques aims to overcome some of the shortcomings of each individual technique such as the Kohonen learning parameters discussed before [48].

2.8.5 Methods’ Limitation

The major limitation or problem with methods 2.8.1, 2.8.2, 2.8.3 and 2.8.4 is the irrelevance of forming new classes describing a diverse set d
gene expression data clustering, SOM method has been recently used in some studies [49, 50]. However, it is not effective in term of covering most of genes within a data set. Therefore, it is used by Naoki, Y. et al [51] in conjunction with K-means clustering in order to get obvious clustering boundaries for gene expression data. Consequently, the results have shown the possibility to determine the functions of unknown genes. Consequently this research will make use of the current clustering techniques, including above, and extend them for more advanced method. It also aims to examine a subspace clustering method to examine huge amount of data depicted in gene expression of affected bio-tissue. This examination will provide a useful benchmark to produce a more effective algorithm.

2.9 Recent Studies

Various studies have attempted to make genetic regulatory networks based on datasets derived from the whole-genome methodologies. In addition, several computational methods based on microarray data are currently used to study genome-wide transcriptional regulation. Different levels of success have been achieved by exploiting exclusively one of these databases [52-54].

2.9.1 Gram Method

The method GRAM [55] combines two databases and had some advantages over the previous ones. The GRAM algorithm determines gene modules that are a set of co expressed genes to which the same set of transcription factors binds. Roughly, the algorithm GRAM scans all subsets of factors (factorsets) and analyzes the expression profiles of genes to which each factorset binds. If the expression profiles of these genes are significantly similar, they would be controlled by the factorset. However, scanning all factorsets in the subset space is an extremely demanding task. GRAM uses some heuristic rules to tackle this problem, but when the number of factors, genes and interactions between them are large, dealing with all the possible subsets become an infeasible task.
2.9.2 Ihmels Method

Ihmels et al [52] offered an alternative method for the global analysis of genome-wide expression data. Their experiment allocates genes potentially overlapping transcription modules. They used a method to clarify regulatory properties of cellular pathways and to characterize regulatory elements. The algorithm used for this experiment, was applied for all the expression data which they selected and classified a complete set of transcriptional modules. As a result, they suggested that the practical predictions for many genes recognize relations between modules and a general view on the transcriptional network.

2.9.3 Sudarsanam Method

Here, Sudarsanam et al [53] describe an approach using microarray data to discover novel functional motif combinations in the promoters of Saccharomyces cerevisiae. They classified a combination of design expression during the cell cycle and studied regulatory cross-talk among this processes to offer general view of transcription networks. The results specified the power of combinatorial analyses of microarray data and a small number of transcription factors compared with the current experiment of clustering expression data and then applying motif-finding algorithms [56-58].

2.10 Ongoing Research

As explained in the previous sections, traditional clustering methods take all the dimensions of an input dataset. As noted, high dimensional data usually includes irrelevant dimensions [35] which can mask existing clusters particularly in noisy data. This is particularly the case in applications such as occurring in gene expression data. There have been a number of approaches [7-13], to address the problems arising from this high dimensionality, such as subspace clustering, decision tree and self organizing maps. Primarily, we conduct a survey on methods on dimensionality reduction in general and then study how this applies to gene expression sub domain.
However researchers have identified various limitations of these methods particularly in the areas of information integration systems, text-mining and bio-informatics [35].

In the bio-informatics domain, existing methodologies for gene expression analysis are an ongoing topic focusing on disease diagnosis [59]. Addressing this intrinsic problem, various approaches have been developed. This research will contribute to this body of knowledge in analyzing gene expression using a data mining technique (decision tree).

By addressing the research problem, the following set of methodology was developed:

- Data Preparation: collecting, integrating and collating the yeast gene expression data from multiple databases on a database.
- Data analysis via Decision Tree methods using SAS software and data mining methods using Enterprise Miner Software.
- Compare H$_2$O$_2$ treatment result with Alpha treatment. Integrate the TF binding in response to Alpha factor and relate this to H$_2$O$_2$ without response. This was done as a negative control.
- Comparing the result with the Causton Results.

2.11 Conclusion

From our general review, we observed among numerous studies that various data mining techniques are currently in use in order to develop an appropriate methodology for effective gene expression data classification. However, further investigation is required in terms of the gene data optimization and elimination of all the redundant data. The following chapters describe the methods applied for yeast gene expression data classification using decision tree technique.
Chapter Three

3. Data Preprocessing

In this chapter the technical approaches used in order to reach the objective and answer the research question are described. It starts by an overview of yeast gene expression data. It is followed by a background of the relation between gene and protein and the list of variables used to build the dataset. All the methods and techniques are mentioned in this chapter: data collection and data preprocessing.

3.1 Overview of Yeast Gene Expression Data Analysis

In the bio-informatics domain, existing methodologies for gene expression analysis are an ongoing topic focusing on disease diagnosis. Using gene expression data measured by microarrays, estimating gene networks has received considerable attention in the field of bioinformatics. Yeast is a unicellular fungi of the genus Saccharomyces studies reproducing by budding and from ascospores and capable of fermenting carbohydrates. The Yeast Proteome Database (YPD) is a model for the organization and presentation of genome-wide functional data.

Due to the inappropriate classification of the datasets, we extract the proposed data into databases to discover maximum groups of factors where each group (or factorset) binds to a set of genes (geneset) and regulates their expression. In other words, we want to find all transcriptional regulatory modules having a form factorset→geneset, where geneset is a set of genes that are bound by factorset and similar in their expression profiles.

In addition, our method shows how gene is changed in answer to Hydrogen peroxide (H$_2$O$_2$) changes in extracellular situation. We will demonstrate that a huge
numbers of genes should be involved in various response biological changes and identify the global set of genes induced and repressed by this condition.

3.2 Relation between Gene and Protein: Background

From biological studies, it is known that gene products including RNA and protein are responsible for the development and functioning of all living cells. Subsequently in eukaryotes, various transcription factors control gene expression by binding to the promoter regions. With combinatorial control, it is not necessary to have a single, simply definable function as commander of a particular battery of genes or specified of a particular cell type, for a given transcription factor. Rather, transcription factors can be likened to the words of a language: they are used with different meanings in a variety of contexts and rarely alone; it is the well-chosen combination that conveys the information that specifies a gene regulatory event [60].

Example: Some concept from gene to protein

![Figure 3.1: The central dogma of molecular biology describes the two-step process, transcription and translation, by which the information in genes flows into proteins: DNA → RNA → protein. [61]](image)
Figure 3.2: There are four types of RNA, each encoded by its own type of gene. The genomic DNA contains all the information for the structure and function of an organism. In any cell, only some of the genes are expressed, that is, transcribed into RNA. [62]

Figure 3.3: The sequence of a eukaryotic protein-coding gene is typically not collinear with the translated mRNA; that is, the transcript of the gene is a molecule that must be processed to remove extra sequences (introns) before it is translated into the polypeptide. [63]

Each cell is the product of specific gene expression programs that involve direct or indirect interactions between DNAs and transcription factors, which are in turn the products of gene expression in previous time course. Such genetic regulatory networks and mechanisms inside them have long been investigated. Traditionally these studies required labor-intensive and gene-specific work.
complete genome sequences of a number of organisms and the development of several high-throughput genomic technologies, such studies have shifted to a new level, whole-genomic scale [64].

In this work, we address a methodology for data collection and processing of yeast gene expression microarray data for knowledge discovery. As part of this methodology, it is proposed to examine a yeast dataset obtained from a collaborative research study between departments of science and computing at University of Western Sydney.

Figure 3.4: The "life cycle" of an mRNA in a eukaryotic cell. RNA is transcribed in the nucleus; once completely processed, it is transported to the cytoplasm and translated by the ribosome. At the end of its life, the mRNA is degraded. [65]

3.3 Data Collection and Description

It is necessary to know how the observed data were collected. According to Zikmund [66], there are two kinds of data normally used in research: primary and secondary data. Primary data is the data that a person gathers on his/her own with a specific purpose in mind while secondary data is data that already has been gathered by other researchers like online research centre’s or websites.
analysis saves time that would otherwise be spent collecting data and, particularly in the case of quantitative data, provides larger and higher-quality databases than would be feasible for any individual researcher to collect on their own [67]. In addition to that, analysts considers secondary data essential, since it is impossible to conduct a new survey that can sufficiently capture past change and/or developments.

As mentioned above, for data mining purpose, secondary data are used. For conducting this study, three types of yeast gene expression data were collected from different sources:

1. **Harbison Dataset and Description:**

   Harbison [16-19] dataset research identified 169 Transcription Factors binding sites before treated harshly by a treatment/experiment in this case H$_2$O$_2$ (22 TFs) and has the systematic name of gene (ORF) variable and the Transcription Factors (TFs) with H$_2$O$_2$ variables.

   As Harbison et. al believed in their research [20] that DNA-binding transcriptional regulators interpret the genome’s regulatory code by binding to specific sequences to induce or repress gene expression. Comparative genomics has recently been used to identify potential cis-regulatory sequences within the yeast genome on the basis of phylogenetic conservation, but this information alone does not reveal if or when transcriptional regulators occupy these binding sites. They have constructed an initial map of yeast’s transcriptional regulatory code by identifying the sequence elements that are bound by regulators under various conditions and that are conserved among Saccharomyces species. The organization of regulatory elements in promoters and the environment-dependent use of these elements by regulators are discussed. They find that environment-specific use of regulatory elements predicts mechanistic models for the function of a large population of yeast’s transcriptional regulators.

2. **Environ Dataset and Description:**

   Causton et al. dataset research contains ORF the peroxide times and the peroxide call for each gene [2] in the mRNA step. These variables help to create new variables related to these variables such as Average at Zero Time for two
before adding $\text{H}_2\text{O}_2$ and Average between two different times (40’ and 60’) to see the reaction of $\text{H}_2\text{O}_2$ in the experiment. These two average variables are helped to introduce the microarray data output variable which demonstrate if the microarray gene data goes up (induce) or down (repress).

They used genome-wide expression analysis to explore how gene expression in Saccharomyces cerevisiae is remodeled in response to various changes in extracellular environment, including changes in temperature, oxidation, nutrients, pH, and osmolarity. The results demonstrate that more than half of the genome is involved in various responses to environmental change and identify the global set of genes induced and repressed by each condition.

3. Microarray Dataset and Description:

Saccharomyces Genome Database (SGD) [68] in a new technology discovered and has a lot of data and their genes descriptions. In this case, we select the gene data contains ORF with gene name, description, the three organizing principles of GO: cellular component, biological process and molecular function and the descriptions of the proteins: Deletant Sensitivity to $\text{H}_2\text{O}_2$, Protein Response to $\text{H}_2\text{O}_2$, Localization Summary and Protein Abundance.

In addition, new attributes related to these datasets calculated by ourselves as twenty two TF Outcomes, Difference between the Average of Peroxide and Microarray Data Output. After uploading all these datasets, we merged them on a general database using Visual Basic Editor Functions.

3.4 Variables Description

Variable construction is an important part of supervised classification. It is a process that discovers missing information about the relationships between variables and augments the space of variables by creating additional variables [69]. In attribute construction, new attributes are constructed from the given attributes and added in order to help improve the accuracy and understanding of structure in high-dimensional data. By combining attributes, attribute construction can discover
missing information about the relationships between data attributes that can be useful for knowledge discovery.

The data of the yeast genome presents profiles up to 6500 genes of mRNA are known Transcription Factors (169 transcription factors). From this data, we select the twenty two Transcriptions Factors related to the absence or present of Hydrogen Peroxide ($H_2O_2$) in the experiments, where each transaction factor has a unique gene identifier (ORF) and contains a set of factors that binds to its promoter with the confidence less than 0.001. The numbers of remaining transactions in the database is 6500 which is equal to the number of yeast genes and each gene can make a single protein. This process is regulated by several factors.

### 3.4.1 The Relation between Attributes:

To go from Gene to Protein, several steps should follow them:

1. Regions before the gene bind proteins known as transcription factors that regulate gene activity. It has some attributes for explanation the gene itself (ORF, Gene Name, GO Biological Process, GO Molecular Function, GO Cellular Component and Description). The dataset contains 6500 Observations and 110 variables.

2. The first step from gene is to mRNA through Transcription Factors (TFs). 169 different TFs describe this bidding way. Our case chooses *twenty two TFs* binds by Hydro peroxide ($H_2O_2$) experiments.

3. Each TF has three values: without $H_2O_2$, low $H_2O_2$ and high $H_2O_2$. The genome-wide location data, the bidding value for the TF, identified 11,000 unique interactions between regulators and promoter regions at high confidence ($P < 0.001$: Strength Binding). It is not necessary for each TF to have a value for each gene. If the gene doesn’t have a value so it can bind by unknown TF or not bind by any TF.
4. In this case, when TFs are “On”, the gene can bind with different TFs. The outcome for each TF can be: TF Always Bound, TF Bound H$_2$O$_2$, Not Response to H$_2$O$_2$ and Not Represent (explained above).

At the mRNA phase, to check if the microarray data is up or down. The two different experiments of Peroxide (Peroxide0A and Peroxide0B) were applied at the zero time. It is the zero time point of the time course experiment and has value of mRNA expression before it was affected by H$_2$O$_2$. We are looking for the difference between any time point and the initial zero time point to see if there is a change. Each peroxide value has a call. The call can be 1 or 0. 1 means the data point was adequately measured in the experiment and 0 means there is no value for the data point due to a technical issue. The outcome of Zero Time is on table 3.2.

In addition, other experiments were done in different times point. We select two times of the value of peroxide after 40 minutes time and value of peroxide after 60 minutes time. The average of these two values was prepared by the same way as Zero Time with the using of call values (table 3.3).

The difference between values of Zero Time and Average Response 40 and 60 indicated in the variable Difference Average 0 and Average 40 and 60.

The Microarray Data Output variable of Zero Time and Average Response 40 and 60 has four outcomes (table 3.4).

5. The translation is the technique to go from mRNA to Protein to check the response of protein. At this point, 4 attributes were used in the analysis:

- Deletant Sensitivity to H$_2$O$_2$ identifies if the gene is deleted from the cell. Therefore the gene is needed for survival.
- Protein Response to H$_2$O$_2$ states if the protein induced or repressed to H$_2$O$_2$.
- Localization Summary indicates the location of protein in cell.
- Protein Abundance specifies the number of proteins in cell.
3.4.2 Summary of definition of the 110 variables:

ORF, Gene Name, Category or GO molecular process and Description have a reference from “Microarray gene list” which is yeast array comprises 6500 genes of DNA fragments.

In this study, three main objectives are considered:

1. From gene to mRNA (more details in the two sections below 4.4.1 and 4.4.2)

2. From gene to Protein through mRNA (more details in the two sections below 4.4.1 and 4.4.2)

3. From gene to Protein without mRNA (more details in the two sections below 4.4.1 and 4.4.2)

- **ORF**: Open Reading Frame is the systematic name of a gene which is the name generated by the systematic sequencing project [70, 71].

- **Gene name**: A Gene Name for example COX2 or CDC28 is a name conferred on a gene by a researcher. Names may be conferred on the basis of genetic as well as biochemical or molecular characterization of the gene. Thus, Gene Names may be conferred on any type of gene or variable that can be characterized genetically. Most genes having Gene Names are ORFs, but tRNAs and other non-protein coding RNAs have also received Gene Names. In addition, there are named genes in SGD that have not yet been mapped to a physical location on the chromosome [70].

The three organizing principles of GO: are cellular component, biological process and molecular function. A gene product might be associated with or located in one or more cellular components; it is active in one or more biological processes, during which it performs one or more molecular functions. For example, the gene product cytochrome c can be described by the molecular function term oxidoreductase activity, the biological process terms oxidative phosphorylation and induction of cell death, and the cellular component terms mitochondrial matrix and mitochondrial inner membrane [72].
The Gene Ontology project is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases.

- **GO biological process:** A biological process is series of events accomplished by one or more ordered assemblies of molecular functions. Examples of broad biological process terms are cellular physiological process or signal transduction. Examples of more specific terms are pyrimidine metabolic process or alpha-glucoside transport. It can be difficult to distinguish between a biological process and a molecular function, but the general rule is that a process must have more than one distinct steps. A biological process is not equivalent to a pathway; at present, GO does not try to represent the dynamics or dependencies that would be required to fully describe a pathway.

- **GO molecular function:** Molecular function describes activities, such as catalytic or binding activities, that occur at the molecular level. GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where or when, or in what context, the action takes place. Molecular functions generally correspond to activities that can be performed by individual gene products, but some activities are performed by assembled complexes of gene products. Examples of broad functional terms are catalytic activity, transporter activity, or binding; examples of narrower functional terms are adenylate cyclase activity or Toll receptor binding.

- **GO cellular component:** A cellular component is just that, a component of a cell, but with the proviso that it is part of some larger object; this may be an anatomical structure (e.g. rough endoplasmic reticulum or nucleus) or a gene product group (e.g. ribosome, proteasome or a protein dimer).

- **Description:** the gene description is a chromosomal location which recorded for each gene [73].
* From *Harbison dataset* [16-19], to analyze yeast data which use the transcription (TF) factor in absence and presence (low and high) of Hydrogen Peroxide ($H_2O_2$). Select 22 different TFs:

- 22 TFs without $H_2O_2$: TF in Absence of $H_2O_2$.
- 22 TFs with $H_2O_2$ Lo: TF in Low Presence of $H_2O_2$.
- 22 TFs with $H_2O_2$ Hi: TF in High Presence of $H_2O_2$.
- 22 TFs Outcome: is our definition to compare the 3 majors of TF.

<table>
<thead>
<tr>
<th>No$H_2O_2$</th>
<th>$H_2O_2$Lo</th>
<th>$H_2O_2$Hi</th>
<th>TFOutcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>TF Always Bound</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>TF Always Bound</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>TF Always Bound</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>TF Unbound $H_2O_2$</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>TF Bound $H_2O_2$</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>TF Bound $H_2O_2$</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>TF Bound $H_2O_2$</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Null</td>
</tr>
</tbody>
</table>

The two values on the table 1 and 0 mean:
- 1: it is a value (The value is P<= 0.001: Strength Binding)
- 0: no value (Bind with unknown TF or no TF)

  TF Always Bound: Bound in Absence and Presence of H\textsubscript{2}O\textsubscript{2}.

  TF Unbound H\textsubscript{2}O\textsubscript{2}: Bound only in Absence of H\textsubscript{2}O\textsubscript{2}.

  Altered Enabled: Bound only in Presence of H\textsubscript{2}O\textsubscript{2}.

  Null: Not Bound in Absence or Presence of H\textsubscript{2}O\textsubscript{2}.

Other variables were extracted from Environ dataset [2]:

- Peroxide 0 min (A) and (B): 2 experiments for biological duplicate at the zero time point of the time course experiment, i.e. the value of mRNA expression before it was affected by H\textsubscript{2}O\textsubscript{2}. We are looking for the difference between any time point and the initial zero time point to see if there is a change, more complex analysis could also be considered such as the trend across all time points.

- Call: 1 = P, 0 = A: P for Present (good) and A for Absent (Bad). 1 means the data point was adequately measured in the experiment, 0 means there is no value for the data point due to a technical issue. For each time we have this value, the values for the same gene at Zero time average and the Average Response 40 and 60 compare together (below).

- Zero Time (Average 0 min): at the 0 minute, the average of the value of the 2 experiments used to compare it with Average Response 40 and 60 variable.

- Peroxide 40 and 60 min: two values after 40 and 60 minutes of adding H\textsubscript{2}O\textsubscript{2}. It’s the zero time point of the time course experiment, i.e. the value of mRNA expression after it was affected by H\textsubscript{2}O\textsubscript{2}.

- Average Response 40 and 60: is the average for the use of H\textsubscript{2}O\textsubscript{2} after 40 and 60 minutes. Compare this value with the value of Zero Time.
• **Difference Average Zero Time and Average 40 and 60**: the difference between Average Response 40 and 60 time point and the initial zero time points.

• **Microarray data Output**: the difference between Average Response 40 and 60 time point and the initial zero time points to see if there is a change, more complex analysis could also be considered such as the trend across all time points.

  a- If the value of call at Peroxide 0’ (A) and Peroxide 0’ (B) are 0: ignore this gene, in the other way do the average for the value of Peroxid0A and Peroxide0B.

<table>
<thead>
<tr>
<th>Call at Peroxide 0’ (A)</th>
<th>Call at Peroxide 0’ (B)</th>
<th>Zero Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>Ignore</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>Mid of Peroxide 0 (A) and Peroxide 0 (B)</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>Mid of Peroxide 0 (A) and Peroxide 0 (B)</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>Mid of Peroxide 0 (A) and Peroxide 0 (B)</td>
</tr>
</tbody>
</table>

b- Now, for Average Response 40 and 60 variable, it is the mid of Peroxide 40’ and Peroxide 60’.
Table 3.3: All cases for two experiments at 40 and 60 minutes time

<table>
<thead>
<tr>
<th>Call at Peroxide 40°</th>
<th>Call at Peroxide 60°</th>
<th>Average Response 40 and 60</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>Ignore</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>Mid of Peroxide 40 and Peroxide 60</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>Mid of Peroxide 40 and Peroxide 60</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>Mid of Peroxide 40 and Peroxide 60</td>
</tr>
</tbody>
</table>

c- Give a value for Microarray Data Output:

Table 3.4: All cases for the Microarray Data Output

<table>
<thead>
<tr>
<th>Zero Time</th>
<th>Average Response 40 and 60</th>
<th>Microarray Data Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Data</td>
<td>No Data</td>
<td>No Data</td>
</tr>
<tr>
<td>No Data</td>
<td>Value</td>
<td>No Data</td>
</tr>
<tr>
<td>Value</td>
<td>No Data</td>
<td>No Data</td>
</tr>
<tr>
<td>Value</td>
<td>Value</td>
<td>mRNA Induce/ mRNA Repress/ mRNA No change</td>
</tr>
</tbody>
</table>

- A comparison of the Zero Time and Average Response 40 and 60 has four outcomes:
  - mRNA Induce: if the value at zero time is at least smaller three times than the value at Average Response 40 and 60.
  - mRNA Repress: if the value at zero time is greater at least three times than the value at Average Response 40 and 60.
- **mRNA No change:** if the value at zero time is not *greater three times* the value at Average Response 40 and 60 and not is *smaller three times* the value at Average Response 40 and 60.

- **No Data:** When one at least of Average Zero Time or Average Response 40 and 60 are ignored.

  In addition, extra variables were added from Saccharomyces Genome Database [68]:

  - **Deletant sensitivity to H\textsubscript{2}O\textsubscript{2}:** If the gene is deleted from the cell, the cell is viable/happy except when treated harshly by a treatment/experiment - in this case H\textsubscript{2}O\textsubscript{2}. Therefore the gene is needed for survival. 1 means it was very sensitive to H\textsubscript{2}O\textsubscript{2} compared to the normal (no deletion) cell, 7 means is a little bit more sensitive to H\textsubscript{2}O\textsubscript{2} compared to the normal cell.

  - **Protein response to H\textsubscript{2}O\textsubscript{2}:** states if the protein induced or repressed to H\textsubscript{2}O\textsubscript{2} and if the protein has No response or unknown.

  - **Localization summary:** it is a predict sub cellular of protein and indicates the location of protein in cell.

  - **Protein abundance:** it is related to mRNA expression in the yeast gene through many different cellular processes. And it is a ration. It indicates how many proteins in cell.

### 3.5 Data Preprocessing

In the whole data mining process, data preparation is a significant process. Some books [28, 74, 75] state that if data mining is considered as a process then a data preparation is at the heart of this process. However, nowadays databases are highly susceptible to noisy, missing, inconsistent data and due to their typically huge size, and their likely origin from multiple, heterogeneous sources. The database may contain fields that are redundant, missing values, outliers, data in a form not suitable for data mining models and values not consistent with policy or common sense [74]. Low quality data will lead to low quality mining results. So data preprocessing is an essential step for knowledge discovery and data mining and improve the efficiency and ease of the data mining process, this becomes an import
consulting firms have approved that data preparation costs 50% to 80% resource of the whole data mining process [28]. From this view, it really needs to pay attention to data preparation.

There are a number of data preprocessing techniques: data cleaning, data integration and data reduction. Data cleaning can be applied to remove noise, supply missing values and correct inconsistencies in the data. Data integration merges data from multiple sources into a coherent data store, such as data warehouse. Data reduction can reduce the data size by aggregating, eliminating redundant variables, or clustering, for instance. These techniques may work together. Data preprocessing step usually takes the most time needed for the whole KDD process, Pyle [75] estimates that data preparation alone accounts for 60% of all the time and effort expanded in the entire data mining process. In the following sections data cleaning and data integration are explained in more details.

3.5.1 Data Integration

It is likely that your data analysis task will involve data integration, which combines data from multiple sources into a coherent data store, as in data warehousing. These sources may include multiple databases, data cubes, or flat files. As explained above the datasets were collected from different databases. All the genes had a unique ID (ORF), which was the same in all the datasets. In this step all the datasets were merged using Macro Microsoft Visual Basic into one table in database. In this study database was created in Microsoft Office Excel and datasets were integrated into one table in Microsoft Office Excel.

Several issues should be considered during data integration, such as schema integration, correlation analysis for detecting redundancy, and detection and resolution of data value conflicts. Some attributes representing a given concept may have different names, causing inconsistencies and redundancies in different publications. For example, gene name may be inconsistent across many publications. Careful integration of the data can help improve the accuracy and speed of the mining process [28].
3.5.2 Data Cleaning

After data was integrated into one table, data had to be cleaned to meet the requirements of the data mining. Real world data tend to be incomplete (lacking attribute values or certain attributes of interest, or containing only aggregate data), noisy (containing errors, or outlier values that deviate from the expected), and inconsistent (conflict attributes). Data cleaning routines attempt to fill in missing values (a record has no value for particular field), smooth out noise while identifying outliers, and correct inconsistencies in the data. Actually data cleaning preprocessing consists of basic operations such as the removal of noise or outliers if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, accounting for time sequence information and known changes [28].

According to Han [28], there are several strategies to deal with missing values such as: ignore the tuple, fill in the missing value manually, and to smooth out the data to remove the noise, use a global constant to fill in the missing value, use the attribute mean to fill in the missing value and use the most probable value to fill in the missing value. Noise is a random error or variance in a measured variable. The solution smooth out the data and remove the noise using different techniques. Such techniques include binning, regression and clustering [28].

For our dataset, a little data cleaning was required. Only some variables had missing values. All records missing values and also all the data which was only an overhead and was not helpful to us in any way was deleted from table.

Data cleaning was done in SAS Enterprise Miner program as follow:

- Delete all records from table have missing value.

As discussed above, there are two Peroxide variables obtained from Environ Dataset. Some of these variables don’t have value. Thus in this step, all the genes has no values was deleted from table.

In this way, after cleaning, dataset was reduced from 6500 genes (records) to 6103 genes (records).
3.6 Conclusion

Following this discussion, it can be seen that complex structures of datasets including genetic data, at molecular level, need an appropriate method suitably to classify the dataset. In this chapter, we presented an overview for yeast gene expression. As well, we gave a brief definition of the variables and the relations between them during the variables description. In addition, we mentioned the way of preprocessing data by collecting datasets, their integration and cleaning. In the coming chapters, we will explain how the analysis is done using SAS Enterprise Miner software and how to build decision tree models using training and validation data.
Chapter Four

4. Data Modeling

In this chapter methodological approaches were used to achieve the objective and answer the research question. It starts with the research design and then presents the research process used in this thesis. The entire decision tree model building process is explained in detail step by step for each objective listed.

4.1 Research Design

The research design is a framework for conducting marketing research [76]. Consequently, it’s a basic plan that guides the collection and analysis phase of the research. It specifies the type of information to be collected, the sources of the data, and the data collection procedure. A good research design will ensure that the information collected will be consistent with the objectives of the study and the procedures regarding data collection are accurate and efficient. The research design of this study is show in Figure 4.1. Detail descriptions are explained below:
4.1.1 Research Purpose

Research can be classified in terms of its purposes. Accordingly, Saunders [77] mentioned that they are most often classified as Examining and Descriptive. The different types of research are explained below:

- Examining research aims to develop initial feeling and provide direction for any further research needed.
- The primary purpose of exploratory research is to shed light on the nature of a situation and identify and specific objectives or data needs to be addressed through additional research.
- Examining is particularly useful when researchers lack a clear idea of the problems they will meet during the study.
• Descriptive research could be in direct connection to examining research, since research might have started off by wanting to gain insight to a problem, and after having stated it their research becomes descriptive [77].

According to data mining definition, the term data mining describes the process of discovering knowledge from databases stored in data warehouses. The purpose of data mining is to identify valid, novel, useful and ultimately understandable patterns in data. Data mining is a useful tool, an approach that combines examining and discovery with confirmatory analysis. Since the focus of this study is data mining, thus the purpose of this research is exploratory.

4.1.2 Research Approach

There are two main research approaches to choose from when conducting research in social science: qualitative and quantitative methods [78]. There is one significant difference between these two approaches. In the quantitative approach results are based on numbers and statistics that are presented in figures, whereas in qualitative results lies on describing an event with the use of words.

Quantitative result seeks to quantify data typically apply some form of statistical analysis. Quantitative research approach transforms the information to numbers that later gets analyzed statistically. It is also characterized by being study few of variables on a large number of entities [79].

Qualitative research is an unstructured, primarily designed based on small samples which are intended to provide some insight and understanding. Qualitative research approach aims at reaching better understanding of the phenomenon being studied, they also tend to be relative flexible using this approach the researcher tries to separate the specific or add and strive to create a complete understanding of the situation [78]. Characteristics of qualitative study are that they are based largely on the researcher’s own description, emotions and reaction. It is also in closeness of respondents or to the source that data is being collected from [79].

Examining data relies more heavily on qualitative techniques but in some cases, there exist sufficient data to allow data mining or examining of
individual measurements to take place. The concept of data mining allows decision makers to be supported through examining quantitative research [80]. This focus of this study is data mining, so the research approach of this research is to be considered a quantitative research.

4.1.3 Research Strategy

Research strategy is a general plan of how to go about answering the research question set [81]. A research strategy is a particular way that the researcher wants to collect data. There are a number of approaches for a researcher to make when conducting empirical data collection. Depending on the character of research question, the researcher can choose between an experiment, a survey, history, secondary data analysis and case study [78].

As mentioned in Chapter Three, there are two kinds of data normally used in researchers: primary and secondary data. Primary data are created by a researcher for the specific purpose of addressing the problem at hand. Secondary data are data that have already been collected for purposes other than the problem at hand. Secondary data can be classified as internal and external. The internal data are those generated within organization for which the research is being conducted and the external data are those generated by sources outside the organization [80]. Internal Secondary data were gathered from Harbison dataset [16-19], Environ dataset [2] and Microarray dataset [68]. See the data collection section for more details. Within secondary data examination, a researcher should start first with an organization’s own data archive [82].

In conclusion, the purpose and approach of this study is an examination quantitative research and the research strategy is the analysis of secondary data.

4.2 Research Process

The objective of this research is to develop a response model in a direct biological setting. The goal is to integrate data from experiments concerning TF binding, mRNA expression and protein induction (in each case these data were obtained in the presence and absence of H$_2$O$_2$) to identify set of genes tha
response. In figure 4.2 a general overview of the proposed model for predicting a response to the H$_2$O$_2$ and if it is yes or not.

![Diagram of response model](image)

**Figure 4.2: General Overview of Response Model**

Response model is a classification model. The task is to classify which genes will respond to H$_2$O$_2$ based on information collected about the genes and divides the customers into two classes’ respondents and non-respondents. As can be seen in Figure 4.2 the input variables of response model are the H$_2$O$_2$ Transcription factors, which show the transform from gene to mRNA and translation from mRNA to protein. Genes with these variables are induced to the model and then model classify them as respondents or non-respondents. Protein response model procedure consists of several steps and different data mining techniques and algorithms have been used for implementing each step of modeling.

Several studies were carried out by using steps such as; data preprocessing, variable construction, gene selection, classification and model evaluation. As can be seen in some related work on protein response model, very few articles deal with all steps. Most of them focus on two or three steps of modeling and for the rest used the result of previous works such as [83-85]. The purpose of this study is to focus on all these steps. In this thesis, the research process presented in the following has been followed. The following process was constructed based on previous methodology (work) on protein response model. Due to the nature of this study different steps of modeling were collected from previous work and with some changes and modifications integrated them into a unique process. The schematic diagram of the entire procedure is shown in figure 4.3. The overall procedure consists of multiple steps such as preprocessing, variable construction, gene selection, classification and evaluation.
Each step of above procedure consists of many different courses of actions. In figure 4.4 the overall procedure in detail is shown.

As can be seen in research process (Figure 4.4) response model procedure consists of several steps. In order to implement the overall research process, an extensive exploration was required. The exploring related to each step was prepared in Microsoft Excel using Macro Visual Basic and SAS Enterprise Miner.
In Chapter Three, we mentioned the data collection and data preparation. In the following sections, the detailed of each step and the methods associated to each step are explained.

### 4.3 SAS Enterprise Miner

All the analysis was carried out with the SAS Enterprise Miner. SAS Analytics provides an integrated environment for predictive and descriptive modeling, data mining, text analytics, forecasting, optimization, simulation, experimental design and more [86]. From predictive analytics to model deployment and process optimization, SAS provides a range of techniques and processes for the collection, classification, analysis and interpretation of data to reveal patterns, anomalies, key variables and relationships, leading ultimately to new insights and better answers faster.

In this study, SAS Data Mining was applied which is an iterative process of creating predictive and descriptive models, by uncovering previously unknown trends and patterns in vast amounts of data from across the enterprise, in order to support decision making [87]. Text mining applies the same analysis techniques to text-based documents. The knowledge gleaned from data and text mining can be used to fuel strategic decision making.

### 4.4 Variable Construction

We briefly described in Chapter Three the extracting of 110 attributes from three different datasets and merged them together to create a framework of yeast gene dataset. Here is a summary explanation of the variables. The dataset contains after cleaning 6103 observations. There are two steps to go from gene to proteins through two ways

- Gene to mRNA during several Transcription factors (without \( \text{H}_2\text{O}_2 \), with low \( \text{H}_2\text{O}_2 \) and with High \( \text{H}_2\text{O}_2 \)) applied for each gene. It is not necessary for each TF to have a value for each gene. If the gene doesn’t have a value so it can bind by unknown TF or not bind by any TF.
On this phase to check if the microarray data is up or down. The two different experiments of Peroxide (Peroxide0A and Peroxide0B) were applied at the zero time. It is the zero time point of the time course experiment and has value of mRNA expression before it was affected by \( \text{H}_2\text{O}_2 \). We are looking for the difference between any time point and the initial zero time point to see if there is a change. Each peroxide value has a call. The call can be 1 or 0. 1 means the data point was adequately measured in the experiment and 0 means there is no value for the data point due to a technical issue. In addition, other experiments were done in different times point. We select two times of the value of peroxide after 40 minutes time and value of peroxide after 60 minutes time. The average of these 2 values was prepared by the same way as Zero Time with the using of call values. The difference between values of Zero Time and Average Response 40 and 60 indicated in the variable ‘Difference Average 0 and Average 40 and 60’.

- The translation is the technique to go from mRNA to Protein to check the response of protein. At this point, four attributes were used in the analysis Deletant Sensitivity to \( \text{H}_2\text{O}_2 \), Protein Response to \( \text{H}_2\text{O}_2 \), Localization Summary and Protein Abundance.

In this study, three main processes/objectives and their associated constructed databases are considered:

1. Process of translation from gene to mRNA (more details in the two sections below 4.4.1 and 4.4.2)

2. Process of translation from gene to Protein through mRNA (more details in the two sections below 4.4.1 and 4.4.2)

3. Process of translation from gene to Protein without mRNA (more details in the two sections below 4.4.1 and 4.4.2)

The decision tree models have been built to analyze the objectives. In following sections described the construction of dependent variable (target variable) and independent variables (predictor variables) of these models.
4.4.1 Target variable Construction

Response model is a classification (supervised) model. The task is to classify which genes will respond to protein based on collected about genes (see Figure 4.2). The first step in the predictive modeling process is to determine the objective variable to be modeled from gene to mRNA and the second and third steps are from gene to protein through mRNA. When building models for targeting protein response to H$_2$O$_2$, the objective variable is usually based on the gene’s response data. So in this thesis the target variable protein response to H$_2$O$_2$ was used.

- The first step in creating the target (dependent) or objective variable from gene to mRNA is to select the Microarray Data Output in consideration. All the genes should be categorized as mRNA Induce, mRNA Repress, mRNA No Change and No Data.

- All these genes, for the second and third steps from gene to Protein, were used in creating Protein Response to H$_2$O$_2$ as target (dependent) or objective variable. These inputs must be classified as Protein Induce by H$_2$O$_2$, Protein Repressed by H$_2$O$_2$ and No Response or Unknown.

4.4.2 Predictor Variables/Variables Construction

The predictor variables are composed in three parts:

- From Gene to mRNA

In this research, transaction factors (TFs) were used as independent variables for the first step from gene to mRNA. Since then the literature has provided so many uses of these variables categories and they are an important set of predictors for protein response model. In this study, all TFs variables were constructed based on knowledge based approach. Using domain knowledge to construct new variables is often recommended because one can thus quickly rule out many impossible combinations and determine an effective measure to evaluate new compound variables. Some information of data is obtained from different datasets. This allowed us to derive and construct all the necessary TFs variables for a total sample size of 6103 observations. According to literature various TFs variable
protein response model. TFs variables for this study were defined as follow: twenty two TFs and each one of these twenty two are composed to four variables: TF without $H_2O_2$, TF with Low $H_2O_2$, TF with High $H_2O_2$ and TF Outcome. The first three variables indicate if the gene effects by these inputs. TF Outcome variable compare the first three variables and check if TF Always Bound, TF Bound $H_2O_2$, TF Unbound $H_2O_2$ or Null.

- **From gene to Protein through mRNA**

The same as above, Microarray data Output and the twenty two TF Outcomes were used as independent variables for the second step from mRNA to Protein. In this study, Microarray Data Output variable was constructed based on knowledge based approach. Some information of data is obtained from different datasets. This allowed us to derive and construct all the necessary Microarray Data Output variables for a total sample size of 6103 observations. According to literature various Microarray Data Output variable was constructed for protein response model. It is composed to four outputs: mRNA Induce, mRNA Repress, mRNA No Change and No Data.

- **From gene to Protein without mRNA**

The same as from gene to protein through mRNA, but when Microarray data Output had No Data value.

In addition, for the three objectives above, to study the effects of individual independent variables (predictors), also it is important to study their interactions. An interaction effect is a change in the simple main effect of one variable over levels of the second [88]. After construction individual independent variables, the second step is to construct interactions variables. First interactions between the predictor variables should be identified with an appropriate method or model which searches for two way interactions between candidate predictors, and then after identifying which variables interacted, the pool of candidate predictors is augmented by adding products between the interacting variables. In this study, the TFs are all related to each other and the Microarray Data Output is also related. In the next section, the gene selection step and the techniques were used for this step is
4.5 Gene selection

Gene selection is a critical step in protein response model. Variable subset selection can be formulated as an optimization problem which involves searching the space of possible variables to identify a subset that is optimum or near-optimal with respect to performance measures such as accuracy [89]. Various ways exist to perform variable subset selection. Gene selection can then either be performed as a preprocessing step, independent of the induction algorithm, or explicitly make use of it. The former approach is termed filter, the latter wrapper [90]. Filter methods operate independently of the target learning algorithm. Unwanted inputs are filtered out of the data before induction commences. Wrapper methods make use of the actual target learning algorithm to evaluate the usefulness of inputs.

In this study, because the numbers of constructed variables (individual and interaction variables) were large, only the filter method was used as a preprocessing step.

4.5.1 Predictor Variables/Variables Construction

To go from gene to protein through mRNA, the transcription factors are the predictor variables and the protein response is the target variable.

As we discussed on Chapter Three, the twenty two transcription factors outcomes variables (22 TFs Outcome) have four values: TF Always Bound, TF Bound with H₂O₂, TF Unbound H₂O₂ and Null. In addition, the protein response to H₂O₂ variable has three values: Protein Induced by H₂O₂, Protein Repressed by H₂O₂ and No Response or Unknown.

<table>
<thead>
<tr>
<th>22 TFs Outcome</th>
<th>Protein Response to H₂O₂</th>
<th>6103 Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Nulls</td>
<td>No Response or Unknown</td>
<td>4347 Genes (Deleted)</td>
</tr>
<tr>
<td>All Nulls</td>
<td>Protein Induced or Repressed</td>
<td>56 Genes (Used)</td>
</tr>
<tr>
<td>Bound, Always Bound and Unbound</td>
<td>Protein Induced or Repressed</td>
<td>27 Genes (Used)</td>
</tr>
<tr>
<td>Bound, Always Bound and Unbound</td>
<td>No Response or Unknown</td>
<td>1673 Genes (Used)</td>
</tr>
</tbody>
</table>
In this study, we are focus on the gene that has at least one value (not Null value) from the twenty two transcriptions factors or the value on the protein response to \( \text{H}_2\text{O}_2 \) is Protein Induced or Repressed to \( \text{H}_2\text{O}_2 \). The dataset reduced, after using this filtering method, from 6103 genes to 1756 genes as it is shown on the table above.

After filtering data, data should be partitioned for the use of Decision Tree techniques.

### 4.6 Data Sampling for Training, Validation and Test Set

Data mining is looking for patterns and relationships in (sometimes large volumes of) data. Many methods, such as recursive partitioning and neural nets, are extremely sensitive to the sample of data being mined. Data miners separate their data into three different subsets to ensure model accuracy. They are training data, testing data and validation data. There is no set rule as to the proportion of data in each data set.

The training data is used to build the model. When the pattern was modeled and not the noise, run the model on the testing data set. The results, especially if the data sets are large, should be similar. If the results are different, the model created over-fit the data (i.e., modeled the noise). There will be some back and forth in this process. Once both the training and testing results are similar, a good model was found that it would expect to be repeatable. This model is put to test with the validation data set. All three models should have a similar fit; otherwise, the underlying pattern has not uncovered in the data.

Prefer Data Partition to compose the dataset into Training, Validation, and Test dataset Percentages.

Actually, researchers build a classifier using the training and validation sets and evaluate it using the test set. They randomly split data into training, validation and test sets. Training is typically done on a large proportion of the total data available, whereas testing is done on small percentage of the data that has been held out exclusively for this purpose (usually 2/3 for train and validate, 1/3 for test).
In this study, the dataset was partitioned into just training, validation and test sets for performance evaluation. We tried many partition and finally we decided to take 40% and 30% were randomly assigned respectively for training and validation sets and analysis in decision tree model, while the other 30% was to test set for.

### 4.7 Decision Tree Classification

The Decision Tree classification fits decision tree models to the data. The implementation includes variables found in a variety of popular decision tree algorithms. The nodes support both automatic and interactive training. When the Decision Tree runs nodes in automatic mode, it automatically ranks the input variables based on the strength of their contribution to the tree. This ranking may be used to select variables for use in subsequent modeling. It may override any automatic step with the option to define a splitting rule and prune explicit tools or sub trees. Interactive training (used in this study) enables the exploration and evaluation of a large set of trees. This study used the classification and regression tree (CART) algorithm with the default parameters. The fine details of after splitting the trees, as well as the validation and missing data management, constitute both the strengths and weaknesses of these types of approaches.

### 4.8 Building Decision Tree Models using Training and Validation Data sets

To build decision tree models for the three objectives discussed above: gene to mRNA, gene to protein with and without mRNA, we used the partition of training and validation set for the 1756 genes. We tried many decision tree models for each objective to get the best model.

#### 4.8.1 From Gene to mRNA:

This case has the Microarray Data Output as target variable and the twenty two transcription factors outcomes as independent variables. Many analyses had been done to process this case, and the important model is shown below.
When the protein has “No Response or Unknown” value, Microarray Data Output (target variable) has mRNA Induce value and the best model is: RPN4 Outcome and HSF1 Outcome (two independent TF outcomes variables). On the other hand, when mRNA Repress, the best model is: RPN4 Outcome and ROX1 Outcome (two independent TF outcomes variables) as showed in Figure 4.5.

Each tree node in Figure 4.5 demonstrates the target concentrations for training and validation sets. The first box in the tree window presented the distribution of Target: Microarray Data Output. It has 702 genes for training data, 527 genes for validation data and composes to mRNA Repress, mRNA No Change, No Data and mRNA Induce values. After selecting the Split Node from the menu, it appears a dialog box that shows the relative value, -log (p) or logworth, of partitioning the training data using the indicated input. It can choose to split the data on the selected input. The inputs variables are all the TFs Outcome.

Repeat this process split the tree until arrived to a particular result and it is partitioned to five leaves with three splitting rules: RPN4 Outcome, HSF1 Outcome and ROX1 Outcome with eight predicted values. Below is the Tree that explains the method from Gene to mRNA through transcription factors.
Using the first input from the list “RPN4 Outcome” for the first leaf, the training data is portioned into three subsets. The first subset corresponding to cases with a “Null” RPN4 Outcome don’t mean anything and we stopped to split for this subset. The second subset with RPN4 Bound to H$_2$O$_2$ has 40 genes, with the mRNA Induce value grown from 17 % to 37%. So continue the same process on the second subset by using HSF1 Outcome and it had “TF Bound H$_2$O$_2$” value. The mRNA Induce value grew from 37% to 60%.

- mRNA Repress Output

The same way as above, from the first node in the tree, split by RPN4 Outcome. The third subset (RPN4 Unbound H$_2$O$_2$) had 33% for mRNA repress. So split this subset with ROX1 Outcome to two subsets and had shown (Figure 4.5) 67 % of mRNA Repress in the second subset when ROX1 Bound H$_2$O$_2$ and 17 % of the second subset when ROX1 had a Null value.
In addition to check if this result was analyzed in high-quality way, it should be checked the average square error for the tree. The Average Square Error method (Gini Index) (Figure 4.6) is the most widely used measure of impurity. It selects the tree that has the smallest average square error.

![Figure 4.6: Average Square Error form Gene to mRNA](image)

As we can see in the Figure 4.6, the average square error kept down in all the five leaves for training and validation sets. This means that this model is good.

### 4.8.2 From Gene to Protein through mRNA

The same process as 4.8.1 section but had different target and independents variables. This case has the Protein Response to H$_2$O$_2$ as target variable with Microarray Data Output and the twenty two transcription factors outcomes as independent variables. Many analyses had been done to process this case, and the important model is shown below.

When Protein Response to H$_2$O$_2$ (target variable) has Protein induced by H$_2$O$_2$, the best model is: mRNA Induce and YAP7 Outcome (Microarray Data Output and one independent TF outcome variables) as showed in Figure 4.7. And when Protein repressed to H$_2$O$_2$, the best model is: mRNA Repress and HSF1 Outcome (Microarray Data Output and one independent TF outcome variables) as showed in Figure 4.9.
The first box in the tree window (Figure 4.7) presented shows the distribution of Target: Protein Response to \( \text{H}_2\text{O}_2 \). It has 702 genes for training data, 527 genes for validation data and composes to No Response or Unknown, Protein Induced by \( \text{H}_2\text{O}_2 \) and Protein Repressed by \( \text{H}_2\text{O}_2 \) values. After selecting the Split Node from the menu, it appears a dialog box that shows the relative value, -log (p) or logworth, of portioning the training data using the indicated input. It can choose to split the data on the selected input. The inputs are the Microarray Data Output and all the TFs Outcome.

Repeat this process split the tree and it is partitioned to five leaves, three splitting rules: Microarray Data Output, HSF1 Outcome and YAP7 Outcome with nine predicted values.

- Protein induced by \( \text{H}_2\text{O}_2 \)

Using the Microarray Data Output from the list for the first leaf, the training data is portioned into three subsets. The second and third subsets, corresponding respectively to No Change and No Data values, don’t mean anything and we stopped to split for this subset. The fourth subset, with Protein induced by \( \text{H}_2\text{O}_2 \) value, was grown from 3 % to 4%. So continue the same process on the second subset by using HSF1 Outcome and it had “TF Bound \( \text{H}_2\text{O}_2 \)” value. The Protein induced by \( \text{H}_2\text{O}_2 \) value grew to 6%.

Figure 4.7: Tree for Protein Response to \( \text{H}_2\text{O}_2 \) as Target variable in Protein induced by \( \text{H}_2\text{O}_2 \) phase when mRNA Induc
Also, the Average Square Error method selects the tree that has the smallest average square error. As shown in Figure 4.8, the average square error is kept down on the training and validation sets, which signifies a good model.

Figure 4.8: Average Square Error from Gene to Protein through mRNA when Protein induced by H$_2$O$_2$

- Protein repressed by H$_2$O$_2$

As we said above, the tree is portioned to five leaves with three splitting rules and nine predicted values. The way for protein repressed by H$_2$O$_2$: the first subset of Microarray Data Output (mRNA repress) when Protein repressed by H$_2$O$_2$ has 5% and HSF1 Outcome (TF Bound H$_2$O$_2$) when Protein repressed by H$_2$O$_2$ grew to 25% as mentioned in the Figure 4.9.
Figure 4.9: Tree for Protein Response to H$_2$O$_2$ as Target variable in Protein repressed by H$_2$O$_2$ phase when mRNA Repress

Also, the Average Square Error method selects the tree that has the smallest average square error. As shown in Figure 4.10, the average square error is kept down on the training and validation sets. That means that this model has a good quality.

Figure 4.10: Average Square Error from Gene to Protein without mRNA when Protein repressed by H$_2$O$_2$

4.8.3 From Gene to Protein without mRNA

Repeat the same process for section 4.8.2 but without mRNA. That means directly from Gene to Protein.

- Protein induced by H$_2$O$_2$
The best way for Protein induced by H$_2$O$_2$ without mRNA is: SKN7 Outcome (TF Bound H$_2$O$_2$) and FKH2 Outcome (TF Always Bound) with 13% of Protein induced by H$_2$O$_2$ as shown in the figure below.

Figure 4.11: Tree for Protein Response to H$_2$O$_2$ as Target variable in Protein induced by H$_2$O$_2$ phase when without mRNA

- Protein repressed by H$_2$O$_2$

The best way for Protein repressed by H$_2$O$_2$ without mRNA is: NRG1 Outcome (TF Always Bound) and YAP6 Outcome (TF Always Bound) with 20% of Protein induced by H$_2$O$_2$ as shown in the figure below.
Figure 4.12: Tree for Protein Response to H$_2$O$_2$ as Target variable in Protein Repressed by H$_2$O$_2$ phase when without mRNA

Protein Induced by H$_2$O$_2$ and Protein repressed by H$_2$O$_2$ have the same Average Square Error method. As shown in Figure 4.13, the average square error is kept down on the training and validation sets. That means that this model has a good quality.

Figure 4.13: Average Square Error from Gene to Protein without mRNA
4.9 Models Evaluation

Once the data models have been created and tested, their performances are analyzed. In this study the performances of the models were measured by using the errors for each objective.

To get the best models after the partition of data set into training, validation and test sets, it should check the error for each model and these three sets should have accuracies close to each other. We used these models because the errors as showed on the Table 4.2 are small. In general, the error should be less than 50% to be a good model.

<table>
<thead>
<tr>
<th>Objectives</th>
<th>Training</th>
<th>Validation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene to mRNA</td>
<td>0.181293</td>
<td>0.184176</td>
<td>0.182042</td>
</tr>
<tr>
<td>Gene to Protein through mRNA</td>
<td>0.032224</td>
<td>0.031109</td>
<td>0.030494</td>
</tr>
<tr>
<td>Gene to Protein without mRNA</td>
<td>0.031154</td>
<td>0.031287</td>
<td>0.030624</td>
</tr>
</tbody>
</table>

- In testing many simple random experiments from gene to mRNA, we selected the best model which had about 18% errors of training, validation and test data set which is a good result (table above).
- Also from gene to protein through mRNA, we tried many simple random experiments and we concentrated in one model mentioned above which had a very small error for training, validation and test sets (table above).
From gene to protein without mRNA, a lot of random experiments tested and they have errors for the combination training, validation and test sets about 3%. The model chosen for this case had the better result with the smallest error (table above).

### 4.10 Conclusion

It can be seen that complex structures of datasets including genetic data, supports the need for an appropriate method suitably to classify the dataset. In this chapter, we presented the data models methodologies to produce results (See chapter five) that are more solid and full than those from previous studies to identify and interpret the transcriptional role of regulator combinations on sets of genes.
Chapter Five

5. Results and Analysis

The process analysis was a brief introduction of SAS software used for modeling, followed by results for each listed objective. Also, the important genes and TFs are declared. Finally a comparison of our result is mentioned with the Causton study and Alpha treatment.

5.1 Result of Variable Construction

After data preprocessing (collecting, integration and cleaning), target variable (dependent variable) and all necessary variables (independent variables) for modeling, as explained in section 4.4, were constructed from three different datasets. Based on domain knowledge and literature, twenty two predictor variables and target variable were constructed from gathered information for the first step from gene to mRNA. In addition, for the second step from gene to Protein through mRNA, twenty three predictor variables and one target variable were created from the collected information. These variables are respectively listed in table 5.1 and 5.2. A detailed description of variable construction was presented in Chapter Four (see section 4.4).

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target: Microarray Data Output</td>
<td>Difference between Average Response 40 and 60 time point and the initial zero time points</td>
</tr>
</tbody>
</table>

22 Predictor variables:

CIN5 Outcome | Outcome of CIN5 transcription factor in presence or absence of H$_2$O$_2$
<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>FKH2 Outcome</td>
<td>Outcome of FKH2 transcription factor in presence or absence of H(_2)O(_2)</td>
</tr>
<tr>
<td>HSF1 Outcome</td>
<td>Outcome of HSF1 transcription factor in presence or absence of H(_2)O(_2)</td>
</tr>
<tr>
<td>MBP1 Outcome</td>
<td>Outcome of MBP1 transcription factor in presence or absence of H(_2)O(_2)</td>
</tr>
<tr>
<td>MOT3 Outcome</td>
<td>Outcome of MOT3 transcription factor in presence or absence of H(_2)O(_2)</td>
</tr>
<tr>
<td>MSN2 Outcome</td>
<td>Outcome of MSN2 transcription factor in presence or absence of H(_2)O(_2)</td>
</tr>
<tr>
<td>MSN4 Outcome</td>
<td>Outcome of MSN4 transcription factor in presence or absence of H(_2)O(_2)</td>
</tr>
<tr>
<td>NRG1 Outcome</td>
<td>Outcome of NRG1 transcription factor in presence or absence of H(_2)O(_2)</td>
</tr>
<tr>
<td>PHO2 Outcome</td>
<td>Outcome of PHO2 transcription factor in presence or absence of H(_2)O(_2)</td>
</tr>
<tr>
<td>RCS1 Outcome</td>
<td>Outcome of RCS1 transcription factor in presence or absence of H(_2)O(_2)</td>
</tr>
<tr>
<td>RIM101 Outcome</td>
<td>Outcome of RIM101 transcription factor in presence or absence of H(_2)O(_2)</td>
</tr>
<tr>
<td>ROX1 Outcome</td>
<td>Outcome of ROX1 transcription factor in presence or absence of H(_2)O(_2)</td>
</tr>
<tr>
<td>RPH1 Outcome</td>
<td>Outcome of RPH1 transcription factor in presence or absence of H(_2)O(_2)</td>
</tr>
<tr>
<td>RPN4 Outcome</td>
<td>Outcome of RPN4 transcription factor in presence or absence of H(_2)O(_2)</td>
</tr>
<tr>
<td>RTG3 Outcome</td>
<td>Outcome of RTG3 transcription factor in presence or absence of H(_2)O(_2)</td>
</tr>
<tr>
<td>SFP1 Outcome</td>
<td>Outcome of SFP1 transcription factor in presence or absence of H(_2)O(_2)</td>
</tr>
<tr>
<td>SKN7 Outcome</td>
<td>Outcome of SKN7 transcription factor in presence or absence of H(_2)O(_2)</td>
</tr>
<tr>
<td>YAP6 Outcome</td>
<td>Outcome of YAP6 transcription factor in presence or absence of H(_2)O(_2)</td>
</tr>
<tr>
<td>Name</td>
<td>Description</td>
</tr>
<tr>
<td>-----------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>YAP7 Outcome</td>
<td>Outcome of YAP7 transcription factor in presence or absence of H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>YJL206C Outcome</td>
<td>Outcome of YJL206C transcription factor in presence or absence of H\textsubscript{2}O\textsubscript{2}</td>
</tr>
</tbody>
</table>

Table 5.2: Second step of variable Construction

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target: Protein Response to H\textsubscript{2}O\textsubscript{2}</td>
<td>States if the protein induced or repressed to H\textsubscript{2}O\textsubscript{2} and if the protein has No response or unknown.</td>
</tr>
</tbody>
</table>

23 Predictor Variables:

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microarray Data Output</td>
<td>Difference between Average Response 40 and 60 time point and the initial zero time points</td>
</tr>
<tr>
<td>AFT2 Outcome</td>
<td>Outcome of AFT2 transcription factor in presence or absence of H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>CIN5 Outcome</td>
<td>Outcome of CIN5 transcription factor in presence or absence of H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>FKH2 Outcome</td>
<td>Outcome of FKH2 transcription factor in presence or absence of H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>HSF1 Outcome</td>
<td>Outcome of HSF1 transcription factor in presence or absence of H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>MBP1 Outcome</td>
<td>Outcome of MBP1 transcription factor in presence or absence of H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>MOT3 Outcome</td>
<td>Outcome of MOT3 transcription factor in presence or absence of H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>MSN2 Outcome</td>
<td>Outcome of MSN2 transcription factor in presence or absence of H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>MSN4 Outcome</td>
<td>Outcome of MSN4 transcription factor in presence or absence of H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>NRG1 Outcome</td>
<td>Outcome of NRG1 transcription factor in presence or absence of H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>RCS1 Outcome</td>
<td>Outcome of RCS1 transcription factor in presence or absence of H\textsubscript{2}C</td>
</tr>
<tr>
<td>Name</td>
<td>Description</td>
</tr>
<tr>
<td>-----------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>RIM101 Outcome</td>
<td>Outcome of RIM101 transcription factor in presence or absence of H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>ROX1 Outcome</td>
<td>Outcome of ROX1 transcription factor in presence or absence of H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>RPH1 Outcome</td>
<td>Outcome of RPH1 transcription factor in presence or absence of H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>RPN4 Outcome</td>
<td>Outcome of RPN4 transcription factor in presence or absence of H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>RTG3 Outcome</td>
<td>Outcome of RTG3 transcription factor in presence or absence of H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>SFP1 Outcome</td>
<td>Outcome of SFP1 transcription factor in presence or absence of H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>SKN7 Outcome</td>
<td>Outcome of SKN7 transcription factor in presence or absence of H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>YAP1 Outcome</td>
<td>Outcome of YAP1 transcription factor in presence or absence of H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>YAP6 Outcome</td>
<td>Outcome of YAP6 transcription factor in presence or absence of H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>YAP7 Outcome</td>
<td>Outcome of YAP7 transcription factor in presence or absence of H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>YJL206C Outcome</td>
<td>Outcome of YJL206C transcription factor in presence or absence of H\textsubscript{2}O\textsubscript{2}</td>
</tr>
</tbody>
</table>

After construction all the necessary variables and their interaction, it is important to select the best subset of variables for modeling. As explained in Chapter Four (section 4.4), for this study gene selection was performed in using filter method. The following sections describe show the result for this step. A brief table related to this data is given in the Appendix A

### 5.2 Result of Gene selection

Decision tree was selected as classifier for this study. It is an effective classification method, but it does not directly obtain the variable importance. Also when the numbers of variable is large, the performance of Deci
significantly. Actually, variable subset selection of the predictor input variables is the main issue that affects the performance of this model. As the numbers of constructed variables (predictor variables and their interaction) for this study were large, so the best subset variables must be selecting for Decision tree modeling.

In this study, gene selection was performed by using a filter method which is usually chosen as a preprocessing step due to its computational efficiency. As explained in section 4.5, the gene selection was done by deleting all record in the twenty two Transaction Factors that have “Null” values for each TF Outcome and have at the same time “Not Represent or Unknown” value on the Protein Response to \( H_2O_2 \) variable as shown in the table below.

<table>
<thead>
<tr>
<th>22 TFs Outcome</th>
<th>Protein Response to ( H_2O_2 )</th>
<th>6103 Genes</th>
<th>mRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Nulls</td>
<td>No Response or Unknown</td>
<td>4347 Genes (Delete)</td>
<td>Delete all</td>
</tr>
<tr>
<td>蛋白 Induced (35) or Repressed (21)</td>
<td>56 Genes (Used)</td>
<td>- 9 mRNA Induce</td>
<td>- 15 mRNA Repress</td>
</tr>
<tr>
<td>蛋白 Induced (15) or Repressed (12)</td>
<td>27 Genes (Used)</td>
<td>- 14 mRNA No Change</td>
<td>- 18 No Data</td>
</tr>
<tr>
<td>蛋白 Induced (15) or Repressed (12)</td>
<td>27 Genes (Used)</td>
<td>- 7 mRNA Induce</td>
<td>- 6 mRNA Repress</td>
</tr>
<tr>
<td>蛋白 Induced (15) or Repressed (12)</td>
<td>27 Genes (Used)</td>
<td>- 6 mRNA No Change</td>
<td>- 8 No Data</td>
</tr>
<tr>
<td>蛋白 Induced (15) or Repressed (12)</td>
<td>27 Genes (Used)</td>
<td>- 289 mRNA Induce</td>
<td>- 332 mRNA Repress</td>
</tr>
<tr>
<td>蛋白 Induced (15) or Repressed (12)</td>
<td>27 Genes (Used)</td>
<td>- 561 mRNA No Change</td>
<td>- 491 No Data</td>
</tr>
<tr>
<td>蛋白 Induced (15) or Repressed (12)</td>
<td>27 Genes (Used)</td>
<td>- 289 mRNA Induce</td>
<td>- 332 mRNA Repress</td>
</tr>
<tr>
<td>蛋白 Induced (15) or Repressed (12)</td>
<td>27 Genes (Used)</td>
<td>- 561 mRNA No Change</td>
<td>- 491 No Data</td>
</tr>
</tbody>
</table>

The dataset reduced, after using this filtering method, from 6103 genes to 1756 genes as it is shown in the table 5.3 above.
5.3 Result of Decision Tree Classification Model

After completing the data preparation, variable construction and selecting the best subset of variables for modeling the response, it is the time to select the classification model. As explained in Chapter Two, various classification methods have been used for protein response of gene expression data, such as machine learning methods: K-means clustering, neural networks and decision trees. In this study, decision tree was used as a classifier for classification (Figure 5.1).

As we discussed in the section 4.8, there are three objectives to check the response of protein to hydrogen peroxide (H₂O₂), gene to mRNA, gene to protein through mRNA and gene to protein without mRNA. We already got the best models for these three objectives. In sections below, we applied the same models but now for the entire data set with 1756 genes, as shown on the Figures 5.2, 5.4, 5.5, 5.7, 5.9, without any partition of data to don’t lose any gene which make affect to the results. The result of each objective is presented in following sections.

5.3.1 From Gene to mRNA

This case has the Microarray Data Output as target variable and the twenty two transcription factors outcomes as independent variables. As mentioned in section 4.8.1, the best model when Microarray Data Output (target variable) has mRNA Induce value is: RPN4 Outcome (TF Bound H₂O₂) followed by HSF1 Outcome (TF Bound H₂O₂). On the other hand, when mRNA Repress, the best model is: RPN4 Outcome (TF Unbound H₂O₂) and ROX1 Outcome (TF Bound H₂O₂). We applied the same model for the full data set now as showed in Figure 5.2.
Figure 5.2: Tree for Microarray Data Output as Target variable in mRNA phase (The highlighted genes were identified by default by the software).

- **mRNA Induce Output**

  On this model, 50% of genes as shown on the figure above have mRNA induce for 6 genes when RPN4 Bound H$_2$O$_2$ and HSF1 Bound H$_2$O$_2$. See the table below for these genes.

<table>
<thead>
<tr>
<th>ORF</th>
<th>Gene Name</th>
<th>RPN4 Outcome</th>
<th>HSF1 Outcome</th>
<th>Microarray Data Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>YGR142W</td>
<td>BTN2</td>
<td>TF Bound H$_2$O$_2$</td>
<td>TF Bound H$_2$O$_2$</td>
<td>mRNA Induce</td>
</tr>
<tr>
<td>YLL039C</td>
<td>UBI4</td>
<td>TF Bound H$_2$O$_2$</td>
<td>TF Bound H$_2$O$_2$</td>
<td>mRNA Induce</td>
</tr>
<tr>
<td>YOR007C</td>
<td>SGT2</td>
<td>TF Bound H$_2$O$_2$</td>
<td>TF Bound H$_2$O$_2$</td>
<td>mRNA Induce</td>
</tr>
</tbody>
</table>

- **mRNA Repress Output**

  When mRNA Repress and by using the same model on 4.8.1 but for the full data set, it is 67% of genes as shown on Figure 5.2 from 6 genes with RPN4 Unbound H$_2$O$_2$ and ROX1 Bound H$_2$O$_2$. The table below dives a descript...
Table 5.5: The three important genes for 67% of mRNA Repress by RPN4 Outcome and ROX1 Outcome

<table>
<thead>
<tr>
<th>ORF</th>
<th>Gene Name</th>
<th>RPN4 Outcome</th>
<th>ROX1 Outcome</th>
<th>Microarray Data Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>YMR239C</td>
<td>RNT1</td>
<td>TF Unbound H$_2$O$_2$</td>
<td>TF Bound H$_2$O$_2$</td>
<td>mRNA Repress</td>
</tr>
<tr>
<td>YNL337W</td>
<td>---</td>
<td>TF Unbound H$_2$O$_2$</td>
<td>TF Bound H$_2$O$_2$</td>
<td>mRNA Repress</td>
</tr>
<tr>
<td>YOR318C</td>
<td>---</td>
<td>TF Unbound H$_2$O$_2$</td>
<td>TF Bound H$_2$O$_2$</td>
<td>mRNA Repress</td>
</tr>
<tr>
<td>YOR319W</td>
<td>HSH49</td>
<td>TF Unbound H$_2$O$_2$</td>
<td>TF Bound H$_2$O$_2$</td>
<td>mRNA Repress</td>
</tr>
</tbody>
</table>

The Average Square Error (Figure 5.3) method selects the tree that has the smallest average square error.

![Average square error (Gini index)](image_url)

Figure 5.3: Average Square Error from Gene to mRNA

In comparison with Figure 4.6, we can see in the Figure 5.3 that the average square error kept down in all the leaves from 0.733 to 0.728 which are nearly the same for the both cases. This means that this model is good.

5.3.2 Gene to Protein through mRNA

This case has the Protein Response to H$_2$O$_2$ as target variable and Microarray Data Output and the twenty two transcription factors outcomes as independent variables. As mentioned on section 4.8.2, the best model when Protein Response to H$_2$O$_2$ (target variable) has Protein induce by H$_2$O$_2$ value is: Microarray Data Output (mRNA Induce) followed by YAP7 Outcome (TF Bound H$_2$O$_2$). On the other hand, when Protein repress by H$_2$O$_2$, the best model is: Microarray...
Repress) and HSF1 Outcome (TF Bound \( \text{H}_2\text{O}_2 \)). We applied the same model for the full data set now as showed in Figure 5.4 and 5.6. (The highlighted genes were identified by default by the software).

- **Protein induced by \( \text{H}_2\text{O}_2 \)**

![Figure 5.4: Tree for Protein Response to \( \text{H}_2\text{O}_2 \) as Target variable in Protein Induced by \( \text{H}_2\text{O}_2 \) phase when mRNA Induce](image)

In this model, 5% of genes as shown on the figure above have Protein induce by \( \text{H}_2\text{O}_2 \) for 43 genes when Microarray Data Output has mRNA Induce and YAP7 Bound \( \text{H}_2\text{O}_2 \). See the table 5.6 below for these genes.

<table>
<thead>
<tr>
<th>ORF</th>
<th>Gene Name</th>
<th>YAP7 Outcome</th>
<th>Microarray Data Output</th>
<th>Protein Response to ( \text{H}_2\text{O}_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>YPL091W</td>
<td>GLR1</td>
<td>TF Bound ( \text{H}_2\text{O}_2 )</td>
<td>mRNA Induce</td>
<td>Protein induced by ( \text{H}_2\text{O}_2 )</td>
</tr>
<tr>
<td>YPL171C</td>
<td>OYE3</td>
<td>TF Bound ( \text{H}_2\text{O}_2 )</td>
<td>mRNA Induce</td>
<td>Protein induced by ( \text{H}_2\text{O}_2 )</td>
</tr>
</tbody>
</table>

- **Protein repressed by \( \text{H}_2\text{O}_2 \)**

In addition, 13% of genes as shown on the figure below have Protein repressed by \( \text{H}_2\text{O}_2 \) for 16 genes when Microarray Data Output has mRNA Induce and YAP7 Bound \( \text{H}_2\text{O}_2 \). See the table 5.7 for these genes.
Figure 5.5: Tree for Protein Response to H$_2$O$_2$ as Target variable in Protein Repressed by H$_2$O$_2$ phase when mRNA Repress

Table 5.6: The important gene of Protein repressed by H$_2$O$_2$ when mRNA Repress

<table>
<thead>
<tr>
<th>ORF</th>
<th>Gene Name</th>
<th>HSF1 Outcome</th>
<th>Microarray Data Output</th>
<th>Protein Response to H$_2$O$_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>YBR082C</td>
<td>UBC4</td>
<td>TF Bound H$_2$O$_2$</td>
<td>mRNA Repress</td>
<td>Protein Repressed by H$_2$O$_2$</td>
</tr>
<tr>
<td>YGR198W</td>
<td>---</td>
<td>TF Bound H$_2$O$_2$</td>
<td>mRNA Repress</td>
<td>Protein Repressed by H$_2$O$_2$</td>
</tr>
</tbody>
</table>

Figure 5.7: Average Square Error from Gene to Protein Through mRNA

In comparison with Figure 4.8 and 4.10, we can see in the Figure above that the average square error kept down in all the leaves from 0.0911 to 0.0.0908 which is better than the Figures 4.8 and 4.10. This means that th...
5.3.3 From Gene to Protein without mRNA:

This case is as the section 5.3.2, with the same target and independent variables. As mentioned on section 4.8.3, the best model when Protein Response to H₂O₂ (target variable) has Protein induce by H₂O₂ value is: SKN7 Outcome (TF Bound H₂O₂) followed by FKH2 Outcome (TF Always Bound). On the other hand, when Protein repress by H₂O₂, the best model is: Outcome (TF Always Bound) and HSF1 Outcome (TF Always Bound). We applied the same model for the full data set now as showed in Figure 5.7 and 5.9.

- Protein induced by H₂O₂

![Figure 5.8: Tree for Protein Response to H₂O₂ as Target variable in Protein Induced by H₂O₂ phase when mRNA has No Change value](image)

On this model, 10% of genes as shown on the figure above have Protein induce by H₂O₂ for 10 genes when SKN7 Bound H₂O₂ and FKH2 Always Bound. See the table 5.8 below for these genes.

Table 5.9: The important gene of Protein induced by H₂O₂ when mRNA has No Change value

<table>
<thead>
<tr>
<th>ORF</th>
<th>Gene Name</th>
<th>SKN7 Outcome</th>
<th>FKH2 Outcome</th>
<th>Microarray Data Output</th>
<th>Protein Response to H₂O₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>YGL008C</td>
<td>PMA1</td>
<td>TF Bound H₂O₂</td>
<td>TF Always Bound</td>
<td>mRNA No Change</td>
<td>Protein induced by H₂O₂</td>
</tr>
</tbody>
</table>
Figure 5.10: Average Square Error from Gene to Protein of Protein induced by H$_2$O$_2$ when mRNA has No Change value

Also, we can see that the average squared error was down here from 0.0911 to 0.0907 which is better than Figure 4.14 which has a 0.98 average error. This means that this model had a good quality.

- Protein repressed by H$_2$O$_2$

Figure 5.11: Tree for Protein Response to H$_2$O$_2$ as Target variable in Protein Repressed by H$_2$O$_2$ phase when mRNA has No Data

In addition, 5% of genes as shown on the figure below have Protein repressed by H$_2$O$_2$ for 19 genes when NRG1 Outcome Always bound and YAP6 Always bound. See the table 5.9 for these genes.
Table 5.12: The important gene of Protein repressed by H$_2$O$_2$ when mRNA has No Data

<table>
<thead>
<tr>
<th>ORF</th>
<th>Gene Name</th>
<th>NRG1 Outcome</th>
<th>YAP6 Outcome</th>
<th>Microarray Output</th>
<th>Data</th>
<th>Protein Response to H$_2$O$_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>YBL029W</td>
<td></td>
<td>TF Always Bound</td>
<td>TF Always Bound</td>
<td>No Data</td>
<td></td>
<td>Protein repressed by H$_2$O$_2$</td>
</tr>
</tbody>
</table>

Figure 5.13: Average Square Error of Protein repressed by H$_2$O$_2$ when mRNA has No Data

Also, we can see that the average squared error was down here from 0.09115 to 0.09072 which is better than Figure 4.14 which has a 0.98 average error. This means that this model had a good quality.

The low percentages of response, when protein induced or repressed by H$_2$O$_2$ as described on the sections 5.3.3 and 5.3.2, are clear because there are just 83 response proteins from the 1756 records.

5.4 Prediction Result

After selecting the best decision tree models, training and validation set was used for building a model. Classification was performed using the twenty four variables selected. A description of the whole model is described in figure 5.11.
To ensure that the model has not overfit the data, the prediction result was used to evaluate the model accuracy. Protein response to $H_2O_2$ variable has 83 outcomes, which is 4.72% of the whole dataset with 1756 genes, portioned between protein induce to $H_2O_2$, protein repressed to $H_2O_2$ as appeared in the table below and the rest 1673 genes had no response or unknown outcome.

**Table 5.15: The distribution output of the protein response to $H_2O_2$ with the outcome of Microarray Data Output**

<table>
<thead>
<tr>
<th>Protein Induced by $H_2O_2$ (50 Genes)</th>
<th>Protein Repressed by $H_2O_2$ (33 Genes)</th>
<th>No Response or Unknown (1673 Genes: each one has at least 1 TF input (not Null))</th>
</tr>
</thead>
<tbody>
<tr>
<td>- 11 mRNA Induce</td>
<td>- 5 mRNA Induce</td>
<td>- 289 mRNA Induce</td>
</tr>
<tr>
<td>- 8 mRNA Repress</td>
<td>- 13 mRNA Repress</td>
<td>- 332 mRNA Repress</td>
</tr>
<tr>
<td>- 13 mRNA No Change</td>
<td>- 7 mRNA No Change</td>
<td>- 561 mRNA No Change</td>
</tr>
<tr>
<td>- 18 No Data</td>
<td>- 8 No Data</td>
<td>- 491 No Data</td>
</tr>
</tbody>
</table>
As can be seen in Table 5.10, 83 out 1756 correspond to the numbers of genes that the model classified as respondent and 1673 out of 1756 as non respondent. 50 genes out of the 83 were predicted as protein induced by \textit{H}_2\textit{O}_2 and 33 out of the 83 predicted as protein repressed by \textit{H}_2\textit{O}_2. All these genes were composed in the Microarray Data Output between the mRNA induce, mRNA repress, mRNA no change and no data values. All these genes are described on the Appendix A. Also all of them have several TF outcomes.

The accuracy and the performance of decision tree will be the figures for each step from gene to mRNA and from gene to protein. Evaluation of these models and the analysis are described in more detail following.

5.5 Important Genes and TFs

The objective of this research is to find the most important genes that have a protein response to \textit{H}_2\textit{O}_2. From the 83 genes response records, 20 of them bounds by several transcription factors and have Microarray Data Output by its four outputs: mRNA induce, mRNA Repress, mRNA No Change and No Data followed by protein induced or repressed by \textit{H}_2\textit{O}_2.

<table>
<thead>
<tr>
<th>ORF</th>
<th>Gene Name</th>
<th>TFs</th>
<th>mRNA \rightarrow Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>YPL240C</td>
<td>HSP82</td>
<td>2 TFs Bound: HSF1 and PHO2</td>
<td>mRNA Induce \rightarrow Protein Induced By \textit{H}_2\textit{O}_2</td>
</tr>
<tr>
<td>YPL171C</td>
<td>OYE3</td>
<td>2 TFs Bound: MSN4 and YAP7</td>
<td>mRNA Induce \rightarrow Protein Induced By \textit{H}_2\textit{O}_2</td>
</tr>
<tr>
<td>YPL091W</td>
<td>GLR1</td>
<td>1 Bound: YAP7</td>
<td>mRNA Induce \rightarrow Protein Induced By \textit{H}_2\textit{O}_2</td>
</tr>
<tr>
<td>YJL031C</td>
<td>BET4</td>
<td>3 TFs Bound: CIN5, YAP6, YAP7</td>
<td>mRNA Induce \rightarrow Protein Repressed By \textit{H}_2\textit{O}_2</td>
</tr>
<tr>
<td>YPR074C</td>
<td>TKL1</td>
<td>1 Bound: RCS1</td>
<td>mRNA Induce \rightarrow Protein Repressed By \textit{H}_2\textit{O}_2</td>
</tr>
<tr>
<td>BR082C</td>
<td>UBC4</td>
<td>1 Bound: HSF1</td>
<td>mRNA Repress \rightarrow Protein Repressed By \textit{H}_2\textit{O}_2</td>
</tr>
<tr>
<td>YGR156W</td>
<td>-</td>
<td>1 Bound: HSF1</td>
<td>mRNA Repress \rightarrow Protein Repressed By \textit{H}_2\textit{O}_2</td>
</tr>
<tr>
<td>YGL008C</td>
<td>PMA1</td>
<td>1 TF Always Bound: FKH2, 2 TFs Bound: HSF1 and SKN7</td>
<td>mRNA No Change \rightarrow Protein Induced By \textit{H}_2\textit{O}_2</td>
</tr>
</tbody>
</table>
In addition, from the 22 transcription factors 13 of them are important and bound several genes with the protein response of H$_2$O$_2$.

Table 5.17: The thirteen important TFs with all their directions

<table>
<thead>
<tr>
<th>TF Name</th>
<th>TF Outcome</th>
<th>ORFs</th>
</tr>
</thead>
<tbody>
<tr>
<td>YAP7 Outcome</td>
<td>5 Bounds</td>
<td>YPL240C, YPL091W, YJR109C, YPR074C</td>
</tr>
<tr>
<td>HSF1 Outcome</td>
<td>4 Bounds</td>
<td>YPL240C, BR082C, YGR156W, YGL008C</td>
</tr>
<tr>
<td>RCS1 Outcome</td>
<td>4 Bounds, 1 Unbound</td>
<td>YPR074C, YPR091W, YBL013W, YGL235</td>
</tr>
</tbody>
</table>

In addition, from the 22 transcription factors 13 of them are important and bound several genes with the protein response of H$_2$O$_2$. 

Table 5.17: The thirteen important TFs with all their directions

<table>
<thead>
<tr>
<th>TF Name</th>
<th>TF Outcome</th>
<th>ORFs</th>
</tr>
</thead>
<tbody>
<tr>
<td>YAP7 Outcome</td>
<td>5 Bounds</td>
<td>YPL240C, YPL091W, YJR109C, YPR074C</td>
</tr>
<tr>
<td>HSF1 Outcome</td>
<td>4 Bounds</td>
<td>YPL240C, BR082C, YGR156W, YGL008C</td>
</tr>
<tr>
<td>RCS1 Outcome</td>
<td>4 Bounds, 1 Unbound</td>
<td>YPR074C, YPR091W, YBL013W, YGL235</td>
</tr>
<tr>
<td>Outcome</td>
<td>Result Description</td>
<td>Bound Proteins</td>
</tr>
<tr>
<td>-----------</td>
<td>---------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>RTG3</td>
<td>2 Bounds</td>
<td>YDL066W, YJR109C</td>
</tr>
<tr>
<td>CIN5</td>
<td>1 Bound, 2 Always Bound</td>
<td>1 Bound: YJL031C, 2 Always Bounds: YBL029W, YDL037C</td>
</tr>
<tr>
<td>FKH2</td>
<td>1 Bound, 2 Always Bound</td>
<td>1 Bound: YHL029C, 2 Always Bounds: YGL008C, YJR109</td>
</tr>
<tr>
<td>SKN7</td>
<td>1 Bound, 1 Always Bound, 2 Unbound</td>
<td>1 Bound: YGL008C, 1 Always Bound: YBL029W</td>
</tr>
<tr>
<td>ROX1</td>
<td>1 Bound, 1 Unbound</td>
<td>1 Bound: YLR465C, 1 Unbound: YBL029W</td>
</tr>
<tr>
<td>MSN4</td>
<td>1 Bound, 2 Unbound</td>
<td>1 Bound: YPL171C, 1 Unbound: YLR465C</td>
</tr>
<tr>
<td>AFT2</td>
<td>1 Bound</td>
<td>YER091C</td>
</tr>
<tr>
<td>PHO2</td>
<td>1 Bound</td>
<td>YPL240C</td>
</tr>
<tr>
<td>NRG1</td>
<td>2 Always Bound</td>
<td>YBL029W, YIL019W</td>
</tr>
</tbody>
</table>

## 5.6 Compare Result

To support our result, it is better to compare it with other related studies. Two of them are followed below.

### 5.6.1 Causton Study

Many researches had been done in this area such as by the Causton study [2]. In this study, Causton used different treatments on the same experiments: heat shock, acid, alkali, hydrogen peroxide, salt and sorbitol. Causton conducted the experiment for two hours after the addition of hydrogen peroxide in the mRNA phase. The response for the expression analysis was focused for some transcription factors. In
contrast, we used all the hydrogen peroxide for analysis and our result displayed the important genes and TFs (see Appendix A for more details).

5.6.2 Alpha Treatment

As mentioned in Chapter Three, there are 169 transcription factors with multiple treatments (H$_2$O$_2$, Alpha, Acid, Heat Shock, and others). In this study, we used the transcription factors that contain hydrogen peroxide (H$_2$O$_2$) treatment to check the protein response of H$_2$O$_2$. To verify that H$_2$O$_2$ treatment is better than the others, we did another experiment using Alpha treatment which have five transcription factors (TFs), same variables and decision tree model too. The same study of H$_2$O$_2$ applied for this case. After cleaning dataset, the genes reduce from 6500 to 6103. Also, the dataset reduced, after using this filtering method, from 6103 genes to 378 genes (see Appendix B for more details). It also has the same target variables and independent for the same three objectives but on here just the five TFs Outcomes related to Alpha treatment. The three objectives are: gene to mRNA and gene to protein with and without mRNA. In addition, we build decision tree models after partition data into training and validation data sets as we did for H$_2$O$_2$ case.

We applied these decision tree models to the full dataset without partition to don’t lose any records as the H$_2$O$_2$ treatment.

For each objective, the best model selected with analysis:

- From Gene to mRNA

  From gene to mRNA, when the protein has “No Response or Unknown” value, Microarray Data Output (target variable) has mRNA Induce or Repress values and the best model is: STE12 Outcome and TEC1 Outcome (two independent TF outcomes variables).

  - The same study for H$_2$O$_2$ treatment is done on this treatment. So we split this objective on the same model of the partition data as shown in the figure below:
Figure 5.18: Tree for Microarray Data Output as Target variable in mRNA Induce phase

Figure 5.12 indicates that the best way to determine “mRNA Induce” is: TEC1Outcome (TF Bound Alpha) -> STE12 Outcome (TF Always Bound) with 40% from five observations. The variables for this case are given in Table 5.13 below.

<table>
<thead>
<tr>
<th>ORF</th>
<th>Gene Name</th>
<th>STE12 Outcome</th>
<th>TEC1 Outcome</th>
<th>Microarray Data Output</th>
<th>Protein Response to H₂O₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>YDR461W</td>
<td>MFA1</td>
<td>TF Always Bound</td>
<td>TF Bound ALPHA</td>
<td>mRNA Induce</td>
<td>No Response or Unknown</td>
</tr>
<tr>
<td>YNR044W</td>
<td>AGA1</td>
<td>TF Always Bound</td>
<td>TF Bound ALPHA</td>
<td>mRNA Induce</td>
<td>No Response or Unknown</td>
</tr>
</tbody>
</table>

- The same work when mRNA Repress:
Figure 5.20: Tree for Microarray Data Output as Target variable in mRNA Repress phase

Figure 5.13 above indicates that the best way to determine “mRNA Repress” is: TEC1 Outcome (TF Always Bound Alpha) -> STE12 Outcome (TF Always Bound) 60% from five observations. The variables for this case are given in table 5.14 below.

Table 5.2122: The three important genes when mRNA Repress

<table>
<thead>
<tr>
<th>ORF</th>
<th>Gene Name</th>
<th>STE12 Outcome</th>
<th>TEC1 Outcome</th>
<th>Microarray Data Output</th>
<th>Protein Response to H$_2$O$_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>YDL127W</td>
<td>PCL</td>
<td>TF Always Bound</td>
<td>TF Always Bound</td>
<td>mRNA Repress</td>
<td>No Response or Unknown</td>
</tr>
<tr>
<td>YNL279W</td>
<td>PRM1</td>
<td>TF Always Bound</td>
<td>TF Always Bound</td>
<td>mRNA Repress</td>
<td>No Response or Unknown</td>
</tr>
<tr>
<td>YNL280C</td>
<td>ERG24</td>
<td>TF Always Bound</td>
<td>TF Always Bound</td>
<td>mRNA Repress</td>
<td>No Response or Unknown</td>
</tr>
</tbody>
</table>

- From Gene to Protein with mRNA

From gene to protein through mRNA, when Protein Response to H$_2$O$_2$ (target variable) has Protein induced by H$_2$O$_2$ doesn’t have any model. And when Protein repressed to H$_2$O$_2$, the best model is: mRNA Repress, SKN7 Outcome and ROX1 Outcome (Microarray Data Output and two independents TF outcomes variables).

The figure below is from gene to protein with mRNA:

Figure 5.2324: Tree for Protein Response to H$_2$O$_2$ as Target variable in Protein Induced by H$_2$O$_2$ phase with Microarray Data Output
In Figure 5.14, the protein response splits the Microarray data output as an independent variable and there are no TF Outcomes that can be split in this case. So we grew the tree manually level by level and we found only two genes as shown in table 5.15.

- From Gene to Protein without mRNA

In addition, from gene to protein without mRNA (when Microarray Data Output has No Data value), there is no model when Protein induces to H2O2. A manual approach was taken for this part and shown in table below that it is one gene when TEC1 Outcome unbound Alpha, mRNA has No Data and protein repressed by H2O2.

<table>
<thead>
<tr>
<th>ORF</th>
<th>Gene Name</th>
<th>TEC1 Outcome</th>
<th>Microarray Data Output</th>
<th>Protein Response to H2O2</th>
</tr>
</thead>
<tbody>
<tr>
<td>YBR082C</td>
<td>UBC4</td>
<td>TF Unbound ALPHA</td>
<td>mRNA Repress</td>
<td>Protein repressed by H2O2</td>
</tr>
<tr>
<td>YLR465C</td>
<td>---</td>
<td>TF Unbound ALPHA</td>
<td>No Data</td>
<td>Protein repressed by H2O2</td>
</tr>
</tbody>
</table>

**Decision:**

It is a relation between Alpha TFs Outcomes and Microarray data output but without any relation between all TFs and Protein. That doesn’t make any sense because we are looking for relation between genes and proteins. When the TFs are ‘On’ the protein response to H2O2 has ‘No Response or Unknown’ value. In contrast, on the protein site, when the gene has a response to the protein, the TFs Outcomes are ‘Null’.

This analysis indicated that there was a poor correlation between Alpha TF binding and a response mRNA expression. The 83 protein response to H2O2 (induced or repressed) didn’t bind any genes because all the five TFs outcomes with these 83 responses had ‘Null’ outcomes except two of them had ‘TF Unbound Alpha’. In contrast with H2O2 treatment, the 83 protein responses by H2O2 are almost covered by TFs Bounds H2O2. So this result means that hydrogen peroxide is a better treatment for this model in comparison with Alpha treatment.
Chapter Six

6. Conclusion and Further Research

In this chapter conclusions are presented with the comparison of our results with studies of others. Also the limitations and management implications of this research will be mentioned. At the end, further research on the area related to this study will be recommended.

6.1 Conclusion

In this thesis, a response model for target selection in yeast gene expression data with data mining techniques was constructed from three different datasets and a core set of genes being highly represented (see Section 5.5). The purpose of this study was identifying whether a protein response to H$_2$O$_2$ could be predicted by transcription factors binding and protein expression. To achieve this aim, we developed a predictive response model with data mining techniques for yeast gene expression data to select the variables that should be targeted in order to obtain a high percentage of positive responses. By using this model, yeast data can obtain a percentage as large as possible of the targeted proteins responds to the gene offer.

Protein response model is usually formulated as a three classification problems. The proteins are divided into three classes, protein induced by H$_2$O$_2$, protein repressed by H$_2$O$_2$ and No response or Unknown. Various classification methods have been used for protein response model of gene expression data, such as machine learning methods: K-means clustering, neural networks and decision trees. In this study, decision tree was used as a classifier for classification.

Protein response model procedure consists of different steps such as: data collection, data preprocessing, variable construction, gene selec
classification and model evaluation. Various data mining techniques have been used for implementing each step. In this study, the research process illustrated in Figure 4.4 was followed. This process was constructed based on previous methodology on protein response model. Due to nature of this study, different steps of modeling were collected from previous work and with some changes and amendments integrated them into a unique process. In order to select a technique for implementing each step of process, different studies related to each step were reviewed and evaluated. After reviewing various techniques and strategies for each step, the best and appropriate ones were selected. In order to implement this research process (building response model), all the techniques run in SAS Enterprise Miner.

The first step of modeling was data collection. As mentioned in section 4.3, the yeast dataset collected from the Harbison dataset, Environ dataset and Microarray dataset for a total sample size of 6500 genes. For data analysis and modeling purpose, all these datasets were combined and integrated into one yeast dataset. As a form of preprocessing, genes which had missing values for each of these were not helpful to us in any way were deleted from dataset. After data cleaning, dataset reduced to 6103 genes.

In variable construction step, target variable and all necessary attributes were constructed. Twenty-two predictor variables and target variable were constructed from gathered information for the first step from gene to mRNA. In addition, for the second step from gene to Protein through mRNA, twenty-three predictor variables and one target variable were created from the collected information. Subsequent to variable construction, gene selection was performed. Gene selection was done by filtering method to reduce the dataset from 6103 genes to 1756 genes. The dataset was partitioned into training and validation data set, took the best model and applied it for the whole data set to don’t lose any data whose affect in the analysis.

Decision tree classification model was used for protein response model. After selecting the best variables for modeling, the twenty-two Transcription Factors Outcomes, Microarray Data Output and Protein Response to H₂O₂, all the analysis was done using SAS Enterprise Miner Data Mining software.
Performance of this model was measured in terms of its percentage. The overall model percentage on test set for the response of protein to H\textsubscript{2}O\textsubscript{2} is shown in the table below:

<table>
<thead>
<tr>
<th>Protein Response to H\textsubscript{2}O\textsubscript{2}</th>
<th>mRNA Output</th>
<th>TF Outcomes</th>
<th>Percentage of Protein Response</th>
<th>Gene predictive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein Induced by H\textsubscript{2}O\textsubscript{2}</td>
<td>mRNA Induce</td>
<td>YAP7 Bound H\textsubscript{2}O\textsubscript{2}</td>
<td>5%</td>
<td>GLR1 and OYE3</td>
</tr>
<tr>
<td>Protein Induced by H\textsubscript{2}O\textsubscript{2}</td>
<td>mRNA No Change</td>
<td>SKN7 Bound H\textsubscript{2}O\textsubscript{2} → FKH2 Always Bound</td>
<td>10%</td>
<td>PMA1</td>
</tr>
<tr>
<td>Protein Repressed by H\textsubscript{2}O\textsubscript{2}</td>
<td>mRNA Repress</td>
<td>HSF1 Bound H\textsubscript{2}O\textsubscript{2}</td>
<td>13%</td>
<td>UBC4 and YGR198W</td>
</tr>
<tr>
<td>Protein Repressed by H\textsubscript{2}O\textsubscript{2}</td>
<td>No Data</td>
<td>NRG1 Always Bound → YAP6 Always Bound</td>
<td>5%</td>
<td>YBL029W</td>
</tr>
<tr>
<td>No Response or Unknown</td>
<td>mRNA Induce</td>
<td>RPN4 Bound H\textsubscript{2}O\textsubscript{2} → HSF1 Bound H\textsubscript{2}O\textsubscript{2}</td>
<td>50%</td>
<td>BTN2, UB14 and SGT2</td>
</tr>
<tr>
<td>No Response or Unknown</td>
<td>mRNA Repress</td>
<td>RPN4 Unbound H\textsubscript{2}O\textsubscript{2} → ROX1 Bound H\textsubscript{2}O\textsubscript{2}</td>
<td>67%</td>
<td>RNT1, HSH49, YNL337W and YOR318C</td>
</tr>
</tbody>
</table>

The low percentages of response, when protein induced or repressed by H\textsubscript{2}O\textsubscript{2}, is clear as we can see on table above because there is just 83 response proteins from 1756 records.

In conclusion, predictive response model help yeast gene expression data to identify a subset of genes who are more likely to respond than others and establish a direct relationship with them. By using this model, the yeast biological dataset not only can significantly reduce the overall percentage but also can prevent to irritate the genes and improve gene relationship.

### 6.2 Research Limitations

This study, like all other previous studies, is not without its limitations. There were some limitations in this research work:
• One of the important limitations of this research was data collection. For data mining purpose secondary data was used (yeast data stored in gene’s databases). Finding a good dataset which store needed information for building a direct response model was very difficult.

• For building a response model, all the variables from gene to protein passed by mRNA are generally used. All of them were collected from the three different datasets mentioned above, except the transcription factors outcomes and the microarray data output variables were excluded. In this study, we determined a way to give these variables values as declared on Chapter Three.

• Powerful and appropriate data mining software and tools are the essential requirement. The only free and available software that we could use for implementing each step of modeling was SAS Enterprise Miner software. Some outcomes were missed after the analyses, like the observations (genes) names that took a lot of time to find an appropriate way to solve this problem.

• Target variable was built based on protein response to $H_2O_2$. One of the major limitations of this work was the yeast gene expression data and another related datasets that use mass datasets as their strategy for offering a new result to genes and they did not have direct method for gene’s response data. So in this thesis, for building the target variable, we had to use gene’s response data to mass yeast gene expression data.

6.3 Management Implications

The findings of this study have important implications for growing numbers of genes, especially yeast gene expression data.

The main objective of this study was to build a protein response model with data mining techniques by integrating hydrogen peroxide yeast’s data to select genes that should be targeted in the next researches.
Our experiment used hydrogen peroxide transcription factors by itself. In contrast to another study like Causton [2], they used multiple treatments together as H$_2$O$_2$, Alpha, Acid, Heat Shock, and others. For this reason, the result shows that by using this model, the yeast dataset can be obtained a percentage as large as possible of the targeted genes responds to the present dataset.

Predictive response model helps the biology researchers to identify a subset of genes who are more likely to respond than others. The main purpose of protein response model is to improve future campaign return on investment and our study shows an increase in protein response to hydrogen peroxide in comparison to other studies.

In summary, researchers should realize the advantages and importance of the data mining in their strategic planning and use the yeast gene data, which is stored in database, to extract knowledge and useful information. Also it can be very useful to build the direct method of response model with data mining techniques to identify a subset of genes. In addition, they are supposed to find a superior method to add multiple treatments together (H$_2$O$_2$, Alpha, Acid, Heat Shock, and others) to don’t lose any targeted genes responds by each treatment by itself.

### 6.4 Suggestions for Further Research

With the development of data mining techniques and databases, some area which are not covered in this study are interesting and need to be explored. Future research could add extension to this study.

The following work ought to be carried out in the future:

- This research mainly focuses on building a predictive response model with decision tree classifier. It was not our aim to compare the performance of different classifiers (like clustering model) when applying to protein response model. Further research is suggested that to compare the performances of different classification models or to create a new algorithm when apply to response model.
• In data cleaning and filtering steps, many records were deleted regarding to the different reasons. It is important to deal with noisy data too. So further research is needed to remove noisy data with clustering techniques.

• Future project can also attempt to identify and develop a general framework how to collect, integrate and collate the gene expression data for high dimensional datasets on one database and analysis. This general framework illustrates using two or more application areas like yeast gene expression data and cancer gene expression data.

• The interactions between proteins are important for the majority of biological functions [91]. This process plays a fundamental role in many biological processes and in many diseases (e.g. yeast, cancers). Proteins might interact for a long time to form part of a protein complex, a protein may be carrying another protein (e.g. cytoplasm to nucleus), or a protein may interact briefly with another protein just to modify it. This modification of proteins can itself change protein to protein interactions. So for future work, we are suggested to study protein to protein Interaction (PPI) when two or more proteins bind together.
7. Appendix A: Yeast Dataset Analysis using Hydrogen Peroxide Treatment

Appendix A contains the analysis developed of this thesis to conduct protein response to \( H_2O_2 \). The concluding consists of examining the behavior of the 1756 genes by adding \( H_2O_2 \) treatment and all the cases that indicates our result using decision tree model.

7.1 Dataset

In this section, we select the main cases that are used to prepare dataset for analysis, including:

- Dataset before filtering (6103 genes)
- Dataset after filtering (1756 genes)
- Distribution of the genes in the protein response to \( H_2O_2 \) phase
- Target and independents features used for analysis
Table 7.1: 6103 genes before filtering dataset

<table>
<thead>
<tr>
<th>22 TFs Outcome</th>
<th>Protein Response to H$_2$O$_2$</th>
<th>6103 Genes</th>
<th>mRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Nulls</td>
<td>No Response or Unknown</td>
<td>4347 Genes (Delete)</td>
<td>Delete all</td>
</tr>
<tr>
<td>All Nulls</td>
<td>Protein Induced (35) or Repressed (21)</td>
<td>56 Genes (Used)</td>
<td>- 9 mRNA Induce</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- 15 mRNA Repress</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- 14 mRNA No Change</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- 18 No Data</td>
</tr>
<tr>
<td>Bound, Bound and Unbound</td>
<td>Protein Induced (15) or Repressed (12)</td>
<td>27 Genes (Used)</td>
<td>- 7 mRNA Induce</td>
</tr>
<tr>
<td>Bound, Bound and Unbound</td>
<td>No Response or Unknown</td>
<td>1673 Genes (Used)</td>
<td>- 289 mRNA Induce</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- 332 mRNA Repress</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- 561 mRNA No Change</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- 491 No Data</td>
</tr>
</tbody>
</table>
Figure 7.1: 1756 genes after filtering dataset and the way from gene to protein through mRNA by adding H$_2$O$_2$ treatment

Table 7.2: The distribution of genes between protein and mRNA phases

<table>
<thead>
<tr>
<th>Protein Induced by H$_2$O$_2$ (50 Genes)</th>
<th>Protein Repressed by H$_2$O$_2$ (33 Genes)</th>
<th>No Response or Unknown (1673 Genes: each one has at least 1 TF input (not Null))</th>
</tr>
</thead>
<tbody>
<tr>
<td>- 11 mRNA Induce</td>
<td>- 5 mRNA Induce</td>
<td>- 289 mRNA Induce</td>
</tr>
<tr>
<td>- 8 mRNA Repress</td>
<td>- 13 mRNA Repress</td>
<td>- 332 mRNA Repress</td>
</tr>
<tr>
<td>- 13 mRNA No Change</td>
<td>- 7 mRNA No Change</td>
<td>- 561 mRNA No Change</td>
</tr>
<tr>
<td>- 18 No Data</td>
<td>- 8 No Data</td>
<td>- 491 No Data</td>
</tr>
</tbody>
</table>
This is a part of our dataset. As shown in the figure above, it has the 26 features used for analysis from the 110 features: ORF, Gene Name, the 22 TFs Outcome, Microarray Data Output and Protein Response to H$_2$O$_2$. 
7.2 Important Responses Genes

The tables below show the genes in all the approaches declared in Chapter Five. In addition the features highlighted means that these attributes are the important ones. Also, they are all explained in important TFs and genes tables in section A.3.

Table 7.3: Eleven genes when mRNA Induce and Protein induced by H\textsubscript{2}O\textsubscript{2} with three important genes

<table>
<thead>
<tr>
<th>ORF</th>
<th>Gene Name</th>
<th>Microarray Data Output</th>
<th>Protein Response to H\textsubscript{2}O\textsubscript{2}</th>
</tr>
</thead>
<tbody>
<tr>
<td>YBR026C</td>
<td>MRF1</td>
<td>mRNA Induce</td>
<td>Protein induced by H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>YDL218W</td>
<td>---</td>
<td>mRNA Induce</td>
<td>Protein induced by H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>YDR378C</td>
<td>LSM6</td>
<td>mRNA Induce</td>
<td>Protein induced by H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>YGL026C</td>
<td>TRP5</td>
<td>mRNA Induce</td>
<td>Protein induced by H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>YGR124W</td>
<td>ASN2</td>
<td>mRNA Induce</td>
<td>Protein induced by H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>YKL207W</td>
<td>---</td>
<td>mRNA Induce</td>
<td>Protein induced by H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>YLR395C</td>
<td>COX8</td>
<td>mRNA Induce</td>
<td>Protein induced by H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>YPL091W</td>
<td>GLR1</td>
<td>mRNA Induce</td>
<td>Protein induced by H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>YPL154C</td>
<td>PEP4</td>
<td>mRNA Induce</td>
<td>Protein induced by H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>YPL171C</td>
<td>OYE3</td>
<td>mRNA Induce</td>
<td>Protein induced by H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>YPL240C</td>
<td>HSP82</td>
<td>mRNA Induce</td>
<td>Protein induced by H\textsubscript{2}O\textsubscript{2}</td>
</tr>
</tbody>
</table>

Table 7.4: Eights genes when mRNA Repress and Protein induced by H\textsubscript{2}O\textsubscript{2} with no important genes

<table>
<thead>
<tr>
<th>ORF</th>
<th>Gene Name</th>
<th>Microarray Data Output</th>
<th>Protein Response to H\textsubscript{2}O\textsubscript{2}</th>
</tr>
</thead>
<tbody>
<tr>
<td>YCR008W</td>
<td>SAT4</td>
<td>mRNA Repress</td>
<td>Protein induced by H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>YDR022C</td>
<td>CIS1</td>
<td>mRNA Repress</td>
<td>Protein induced by H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>YGR271W</td>
<td>---</td>
<td>mRNA Repress</td>
<td>Protein induced by H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>YGR274C</td>
<td>TAF145</td>
<td>mRNA Repress</td>
<td>Protein induced by H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>YJR116W</td>
<td>---</td>
<td>mRNA Repress</td>
<td>Protein induced by H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>YOL063C</td>
<td>---</td>
<td>mRNA Repress</td>
<td>Protein induced by H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>YOR102W</td>
<td>---</td>
<td>mRNA Repress</td>
<td>Protein induced by H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>YPL003W</td>
<td>ULA1</td>
<td>mRNA Repress</td>
<td>Protein induced by H\textsubscript{2}O\textsubscript{2}</td>
</tr>
</tbody>
</table>

Table 7.4: Five genes when mRNA Induce and Protein repressed by H\textsubscript{2}O\textsubscript{2} with two important genes

<table>
<thead>
<tr>
<th>ORF</th>
<th>Gene Name</th>
<th>Microarray Data Output</th>
<th>Protein Response to H\textsubscript{2}O\textsubscript{2}</th>
</tr>
</thead>
<tbody>
<tr>
<td>YHR087W</td>
<td>---</td>
<td>mRNA Induce</td>
<td>Protein repressed by H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>YJL031C</td>
<td>BET4</td>
<td>mRNA Induce</td>
<td>Protein repressed by H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>YOR265W</td>
<td>RBL2</td>
<td>mRNA Induce</td>
<td>Protein repressed by H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>YPR069C</td>
<td>SPE3</td>
<td>mRNA Induce</td>
<td>Protein repressed by H\textsubscript{2}O\textsubscript{2}</td>
</tr>
<tr>
<td>YPR074C</td>
<td>TKL1</td>
<td>mRNA Induce</td>
<td>Protein repressed by H\textsubscript{2}O\textsubscript{2}</td>
</tr>
</tbody>
</table>
Table 7.5: Thirteen genes when mRNA Repress and Protein repressed and Repressed by H$_2$O$_2$ with four important genes

<table>
<thead>
<tr>
<th>ORF</th>
<th>Gene Name</th>
<th>Microarray Data Output</th>
<th>Protein Response to H$_2$O$_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>YBL029W</td>
<td>---</td>
<td>mRNA Repress</td>
<td>Protein repressed by H$_2$O$_2$</td>
</tr>
<tr>
<td>YBR082C</td>
<td>UBC4</td>
<td>mRNA Repress</td>
<td>Protein repressed by H$_2$O$_2$</td>
</tr>
<tr>
<td>YDR518W</td>
<td>EUG1</td>
<td>mRNA Repress</td>
<td>Protein repressed by H$_2$O$_2$</td>
</tr>
<tr>
<td>YGR156W</td>
<td>---</td>
<td>mRNA Repress</td>
<td>Protein repressed by H$_2$O$_2$</td>
</tr>
<tr>
<td>YGR198W</td>
<td>---</td>
<td>mRNA Repress</td>
<td>Protein repressed by H$_2$O$_2$</td>
</tr>
<tr>
<td>YHL049C</td>
<td>---</td>
<td>mRNA Repress</td>
<td>Protein repressed by H$_2$O$_2$</td>
</tr>
<tr>
<td>YIL019W</td>
<td>---</td>
<td>mRNA Repress</td>
<td>Protein repressed by H$_2$O$_2$</td>
</tr>
<tr>
<td>YIL153W</td>
<td>RRD1</td>
<td>mRNA Repress</td>
<td>Protein repressed by H$_2$O$_2$</td>
</tr>
<tr>
<td>YJL116C</td>
<td>NCA3</td>
<td>mRNA Repress</td>
<td>Protein repressed by H$_2$O$_2$</td>
</tr>
<tr>
<td>YKL011C</td>
<td>CCE1</td>
<td>mRNA Repress</td>
<td>Protein repressed by H$_2$O$_2$</td>
</tr>
<tr>
<td>YMR204C</td>
<td>---</td>
<td>mRNA Repress</td>
<td>Protein repressed by H$_2$O$_2$</td>
</tr>
<tr>
<td>YOR174W</td>
<td>MED4</td>
<td>mRNA Repress</td>
<td>Protein repressed by H$_2$O$_2$</td>
</tr>
<tr>
<td>YOR209C</td>
<td>NPT1</td>
<td>mRNA Repress</td>
<td>Protein repressed by H$_2$O$_2$</td>
</tr>
</tbody>
</table>

Table 7.5: Four genes when mRNA No Change and Protein Induced and Repressed by H$_2$O$_2$ with one important gene

<table>
<thead>
<tr>
<th>ORF</th>
<th>Gene Name</th>
<th>Microarray Data Output</th>
<th>Protein Response to H$_2$O$_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>YGL008C</td>
<td>PMA1</td>
<td>mRNA No Change</td>
<td>Protein induced by H$_2$O$_2$</td>
</tr>
<tr>
<td>YGR012W</td>
<td></td>
<td>mRNA No Change</td>
<td>Protein induced by H$_2$O$_2$</td>
</tr>
<tr>
<td>YDL066W</td>
<td>IDP1</td>
<td>mRNA No Change</td>
<td>Protein repressed by H$_2$O$_2$</td>
</tr>
<tr>
<td>YHL029C</td>
<td></td>
<td>mRNA No Change</td>
<td>Protein repressed by H$_2$O$_2$</td>
</tr>
</tbody>
</table>

Table 7.6: Seven important genes when mRNA No Data and Protein Induced and Repressed by H$_2$O$_2$

<table>
<thead>
<tr>
<th>ORF</th>
<th>Gene Name</th>
<th>Microarray Data Output</th>
<th>Protein Response to H$_2$O$_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>YBL013W</td>
<td>FMT1</td>
<td>No Data</td>
<td>Protein induced by H$_2$O$_2$</td>
</tr>
<tr>
<td>YDL037C</td>
<td></td>
<td>No Data</td>
<td>Protein induced by H$_2$O$_2$</td>
</tr>
<tr>
<td>YDR135C</td>
<td>YCF1</td>
<td>No Data</td>
<td>Protein induced by H$_2$O$_2$</td>
</tr>
<tr>
<td>YER091C</td>
<td>MET6</td>
<td>No Data</td>
<td>Protein induced by H$_2$O$_2$</td>
</tr>
<tr>
<td>YGL235W</td>
<td></td>
<td>No Data</td>
<td>Protein induced by H$_2$O$_2$</td>
</tr>
<tr>
<td>YJR109C</td>
<td>CPA2</td>
<td>No Data</td>
<td>Protein repressed by H$_2$O$_2$</td>
</tr>
<tr>
<td>YLR465C</td>
<td></td>
<td>No Data</td>
<td>Protein repressed by H$_2$O$_2$</td>
</tr>
</tbody>
</table>

7.3 All Important Genes and TFs
Table 7.7: All important genes in all the cases with descriptions

<table>
<thead>
<tr>
<th>ORF</th>
<th>Gene Name</th>
<th>TFs</th>
<th>mRNA → Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>YPL240C</td>
<td>HSP82</td>
<td>2 TFs Bound: HSF1 and PHO2</td>
<td>mRNA Induce → Protein Induced By H$_2$O$_2$</td>
</tr>
<tr>
<td>YPL171C</td>
<td>OYE3</td>
<td>2 TFs Bound: MSN4 and YAP7</td>
<td>mRNA Induce → Protein Induced By H$_2$O$_2$</td>
</tr>
<tr>
<td>YPL091W</td>
<td>GLR1</td>
<td>1 Bound: YAP7</td>
<td>mRNA Induce → Protein Induced By H$_2$O$_2$</td>
</tr>
<tr>
<td>YJL031C</td>
<td>BET4</td>
<td>3 TFs Bound: CIN5, YAP6, YAP7</td>
<td>mRNA Induce → Protein Repressed By H$_2$O$_2$</td>
</tr>
<tr>
<td>YPR074C</td>
<td>TKL1</td>
<td>1 Bound: RCS1</td>
<td>mRNA Induce → Protein Repressed By H$_2$O$_2$</td>
</tr>
<tr>
<td>YBL029W</td>
<td>-</td>
<td>4 TFs Always Bound: YAP6, SKN7, NRG1, and CIN5. 1 TF Bound: RIM101. 1 Unbound: ROX1</td>
<td>mRNA Repress → Protein Repressed By H$_2$O$_2$</td>
</tr>
<tr>
<td>BR082C</td>
<td>UBC4</td>
<td>1 Bound: HSF1</td>
<td>mRNA Repress → Protein Repressed By H$_2$O$_2$</td>
</tr>
<tr>
<td>YGR156W</td>
<td>-</td>
<td>1 Bound: HSF1</td>
<td>mRNA Repress → Protein Repressed By H$_2$O$_2$</td>
</tr>
<tr>
<td>YGL008C</td>
<td>PMA1</td>
<td>1 TF Always Bound: FKH2 2 TFs Bound: HSF1 and SKN7</td>
<td>mRNA No Change → Protein Induced By H$_2$O$_2$</td>
</tr>
<tr>
<td>YGR012W</td>
<td>-</td>
<td>1 Bound: RCS1</td>
<td>mRNA No Change → Protein Induced By H$_2$O$_2$</td>
</tr>
<tr>
<td>YDL066W</td>
<td>IDP1</td>
<td>1 Bound: RTG3</td>
<td>mRNA No Change → Protein Repressed By H$_2$O$_2$</td>
</tr>
<tr>
<td>YHL029C</td>
<td>-</td>
<td>1 Bound: FKH2</td>
<td>mRNA No Change → Protein Repressed By H$_2$O$_2$</td>
</tr>
<tr>
<td>YDL037C</td>
<td>-</td>
<td>2 TFs Always Bound: CIN5 and YAP6</td>
<td>No Data → Protein Induced By H$_2$O$_2$</td>
</tr>
<tr>
<td>YBL013W</td>
<td>FMT1</td>
<td>1 Bound: RCS1</td>
<td>mRNA No Data → Protein Induced By H$_2$O$_2$</td>
</tr>
<tr>
<td>YDR135C</td>
<td>YCF1</td>
<td>1 Bound: YAP7</td>
<td>mRNA No Data → Protein Induced By H$_2$O$_2$</td>
</tr>
<tr>
<td>YER091C</td>
<td>MET6</td>
<td>1 Bound: AFT2</td>
<td>mRNA No Data → Protein Induced By H$_2$O$_2$</td>
</tr>
<tr>
<td>YJR109C</td>
<td>CPA2</td>
<td>1 TF Always Bound: FKH2 2 TFs Bound: RTG3 and YAP7 1 Unbound: YAP1</td>
<td>No Data → Protein Repressed By H$_2$O$_2$</td>
</tr>
<tr>
<td>YLR465C</td>
<td>-</td>
<td>2 TFs Bound: ROX1 and YAP6 1 Unbound: MSN4</td>
<td>No Data → Protein Repressed By H$_2$O$_2$</td>
</tr>
<tr>
<td>YGL235W</td>
<td>-</td>
<td>1 Bound: RCS1</td>
<td>mRNA No Data → Protein Induced By H$_2$O$_2$</td>
</tr>
<tr>
<td>YIL019W</td>
<td>-</td>
<td>1 Always Bound: NRG1</td>
<td>mRNA Repress → Protein Repressed By H$_2$O$_2$</td>
</tr>
</tbody>
</table>
### Table 7.8: All important TFs in all the cases with descriptions

<table>
<thead>
<tr>
<th>TF Name</th>
<th>TF Outcome</th>
<th>ORFs</th>
</tr>
</thead>
<tbody>
<tr>
<td>YAP7 Outcome</td>
<td>5 Bounds</td>
<td>YPL240C, YPL091W, YJL031C, YDR135C, YJR109C</td>
</tr>
<tr>
<td>HSF1 Outcome</td>
<td>4 Bounds</td>
<td>YPL240C, BR082C, YGR156W, YGL008C</td>
</tr>
<tr>
<td>RCS1 Outcome</td>
<td>4 Bounds, 1 Unbound</td>
<td>4 Bounds: YPR074C, YGR012W, YBL013W, YGL235W</td>
</tr>
<tr>
<td>YAP6 Outcome</td>
<td>2 Bounds, 2 Always Bounds</td>
<td>2 Bounds: YJL031C, YLR465C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 Always Bounds: YBL029W, YDL037C</td>
</tr>
<tr>
<td>RTG3 Outcome</td>
<td>2 Bounds</td>
<td>YDL066W, YJR109C</td>
</tr>
<tr>
<td>CIN5 Outcome</td>
<td>1 Bound, 2 Always Bound</td>
<td>1 Bound: YJL031C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 Always Bounds: YBL029W, YDL037C</td>
</tr>
<tr>
<td>FKH2 Outcome</td>
<td>1 Bound, 2 Always Bound</td>
<td>1 Bound: YHL029C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 Always Bounds: YGL008C, YJR109</td>
</tr>
<tr>
<td>SKN7 Outcome</td>
<td>1 Bound, 1 Always Bound, 2 Unbound</td>
<td>1 Bound: YGL008C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 Always Bound: YBL029W</td>
</tr>
<tr>
<td>ROX1 Outcome</td>
<td>1 Bound, 1 Unbound</td>
<td>1 Bound: YLR465C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 Unbound: YBL029W</td>
</tr>
<tr>
<td>MSN4 Outcome</td>
<td>1 Bound, 2 Unbound</td>
<td>1 Bound: YPL171C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 Unbound: YLR465C</td>
</tr>
<tr>
<td>AFT2 Outcome</td>
<td>1 Bound</td>
<td>YER091C</td>
</tr>
<tr>
<td>PHO2 Outcome</td>
<td>1 Bound</td>
<td>YPL240C</td>
</tr>
<tr>
<td>NRG1 Outcome</td>
<td>2 Always Bound</td>
<td>YBL029W, YIL019W</td>
</tr>
</tbody>
</table>
8. Appendix B: Yeast Dataset Analysis using Alpha Treatment

Appendix B contains the analysis developed in this thesis to conduct protein response to \( H_2O_2 \) by adding Alpha treatment. This case was used to compare our result with \( H_2O_2 \). The concluding consists of examining the behavior of the 378 genes by adding Alpha treatment and all the cases that indicates our result using decision tree model.

8.1 Dataset

In this section, we select the main cases that are used to prepare dataset for analysis, including:

- Dataset before filtering (6103 genes)
- Dataset after filtering (378 genes)
- Distribution of the genes in the protein response to \( H_2O_2 \) phase
- Target and independents features used for analysis
Table 8.1: 6103 genes before filtering dataset

<table>
<thead>
<tr>
<th>5 TFs Outcome</th>
<th>Protein Response to H\textsubscript{2}O\textsubscript{2}</th>
<th>6103 Genes</th>
<th>mRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No Response or Unknown</td>
<td>5725 Genes</td>
<td>Delete</td>
</tr>
<tr>
<td>All Nulls</td>
<td>Protein Induced (50) or Repressed (31)</td>
<td>81 Genes (Used)</td>
<td>- 16 mRNA Induce</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- 20 mRNA Repress</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- 20 mRNA No Change</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- 25 No Data</td>
</tr>
<tr>
<td>2 Unbound</td>
<td>Protein Repressed (2)</td>
<td>2 Genes (Used)</td>
<td>- 1 mRNA Repress</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- 1 No Data</td>
</tr>
<tr>
<td>Bound, Always Bound and Unbound</td>
<td>No Response or Unknown</td>
<td>295 Genes (Used)</td>
<td>- 35 mRNA Induce</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- 72 mRNA Repress</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- 93 mRNA No Change</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- 95 No Data</td>
</tr>
</tbody>
</table>
Figure 8.1: 378 genes after filtering dataset and the way from gene to protein through mRNA by adding alpha treatment

Table 8.2: The distribution of genes between protein and mRNA phases

<table>
<thead>
<tr>
<th>Protein Induced by H\textsubscript{2}O\textsubscript{2} (50 Genes)</th>
<th>Protein Repressed by H\textsubscript{2}O\textsubscript{2} (33 Genes)</th>
<th>No Response or Unknown (295 Genes: each one has at least 1 TF input (not Null))</th>
</tr>
</thead>
<tbody>
<tr>
<td>11 mRNA Induce</td>
<td>5 mRNA Induce</td>
<td>35 mRNA Induce</td>
</tr>
<tr>
<td>8 mRNA Repress</td>
<td>13 mRNA Repress</td>
<td>72 mRNA Repress</td>
</tr>
<tr>
<td>13 mRNA No Change</td>
<td>7 mRNA No Change</td>
<td>93 mRNA No Change</td>
</tr>
<tr>
<td>18 No Data</td>
<td>8 No Data</td>
<td></td>
</tr>
</tbody>
</table>
Figure 8.2: The features names used for analysis

This is a part of our dataset. As shown in the figure above, it has the twenty six features used for analysis from the fifty two features: ORF, Gene Name, the five TFs Outcome, Microarray Data Output and Protein Response to H$_2$O$_2$.

8.2 Important Responses Genes

The tables below show the important genes in all the approaches for protein induce by H$_2$O$_2$, protein repressed by H$_2$O$_2$ and No Response or Unknown in the protein phase through microarray data output and the five transcription factors outcomes.
Table 8.3: The two genes when mRNA Repress and No Data values with Protein repressed by H$_2$O$_2$ and TEC1 transcription factor unbound alpha outcome

<table>
<thead>
<tr>
<th>ORF</th>
<th>Gene Name</th>
<th>TEC1 Outcome</th>
<th>Microarray Data Output</th>
<th>Protein Response to H$_2$O$_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>YBR082C</td>
<td>UBC4</td>
<td>TF Unbound ALPHA</td>
<td>mRNA Repress</td>
<td>Protein repressed by H$_2$O$_2$</td>
</tr>
<tr>
<td>YLR465C</td>
<td>---</td>
<td>TF Unbound ALPHA</td>
<td>No Data</td>
<td>Protein repressed by H$_2$O$_2$</td>
</tr>
</tbody>
</table>

Table 8.4: The two important genes when mRNA Induce and No Response or Unknown Protein by H$_2$O$_2$

<table>
<thead>
<tr>
<th>ORF</th>
<th>Gene Name</th>
<th>STE12 Outcome</th>
<th>TEC1 Outcome</th>
<th>Microarray Data Output</th>
<th>Protein Response to H$_2$O$_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>YDR461W</td>
<td>MFA1</td>
<td>TF Always Bound</td>
<td>TF Bound ALPHA</td>
<td>mRNA Induce</td>
<td>No Response or Unknown</td>
</tr>
<tr>
<td>YNR044W</td>
<td>AGA1</td>
<td>TF Always Bound</td>
<td>TF Bound ALPHA</td>
<td>mRNA Induce</td>
<td>No Response or Unknown</td>
</tr>
</tbody>
</table>

Table 8.5: The three important genes when mRNA Repress and No Response or Unknown Protein by H$_2$O$_2$

<table>
<thead>
<tr>
<th>ORF</th>
<th>Gene Name</th>
<th>STE12 Outcome</th>
<th>TEC1 Outcome</th>
<th>Microarray Data Output</th>
<th>Protein Response to H$_2$O$_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>YDL127W</td>
<td>PCL</td>
<td>TF Always Bound</td>
<td>TF Always Bound</td>
<td>mRNA Repress</td>
<td>No Response or Unknown</td>
</tr>
<tr>
<td>YNL279W</td>
<td>PRM1</td>
<td>TF Always Bound</td>
<td>TF Always Bound</td>
<td>mRNA Repress</td>
<td>No Response or Unknown</td>
</tr>
<tr>
<td>YNL280C</td>
<td>ERG24</td>
<td>TF Always Bound</td>
<td>TF Always Bound</td>
<td>mRNA Repress</td>
<td>No Response or Unknown</td>
</tr>
</tbody>
</table>
9. References


34. Jinn-Yi, T.-S.W., Min-Che, W. and Der-Ming, C. Applying Data Mining Techniques for Cancer Classification from Gene Expression Data. International Conference on Convergence Information Technology, 2007.


38. Tang, C. and Zhang, A. An iterative strategy for pattern discovery in high-dimensional data sets. in Proceeding of 11th International Conference on Information and Knowledge Management (CIKM 02). 2002. McLean, VA.


42. Kriegel, H.-P., Kröger, P. and Zimek, A. A Generic Framework for Efficient Subspace Clustering of High-Dimensional Data. in Proceedings of the Fifth


65. Organizing knowledge for happiness and world peace, http://www.newworldencyclopedia.org/entry/Messenger_RNA.


73. www.hgmd.cf.ac.uk/ac/index.php.