

**Universiti Teknologi MARA**

**Mining the DNA Sequence for Race Classification**

**Siti Nurihan Hj.Ariffin**

Thesis submitted in fulfillment of the requirements for  
**Bachelor of Science (Hons) Information System Engineering**  
**Faculty of Information Technology And**  
**Quantitative Science**

May 2008

## ACKNOWLEDGEMENT

In the name of Allah, The Most Gracious, The Most Merciful and Him alone is worthy of all praise. Thank you the Almighty God for giving me the strength and ability towards completing my final year project.

First and foremost, I would like to thank to my supervisor Pn. Fauziah Redzuan, for her patient, advice and criticism for betterment to this research. This project wasn't made possible without her full support.

A very special thanks to Pn. Jamaliah bt. Taslim and PM Rashidah bt. Rawi, who has never hesitate in helping me upon the completion of my research .With their guidance and ideas, they help me lots in doing my research.

I also would like to thank to all my friends for their dedication to this research and their never ending passion. And also to my entire respondent that give their cooperation towards my research.

A millions thanks to my parents, especially to my mother Jamaliah@Samsiah bt. Mat who are really understand, encouraging and giving full of her support to me to do my research and my study.

Last but not least, thanks to all those who are assisting me directly or indirectly until I complete my research. Thank you very much for all your support.

Thank you.

## **ABSTRACT**

Bioinformatics is a new field that arises in Malaysia. It offers many benefits especially in the genes and genomic research. By using the computers technology, it helps the researchers to understand the large amount of data in the bioinformatics. By implementing data mining techniques, the researchers are able to analyze the data from different perspective and summarizing the data into useful information. This research focusing on identifying the data mining methods/ techniques for the race classification and predict the races based on their DNA sequence using the appropriate technique. The comparison between classification techniques was done in order to get the best technique for classification. By implementing the existing tool of neural network, the race classification was achieved. From this research, it helps in identifying the techniques for classification.

# TABLE OF CONTENTS

<b>TITLE</b>	<b>PAGE</b>
<b>DECLARATION</b>	<b>i</b>
<b>APPROVAL</b>	<b>ii</b>
<b>ACKNOWLEDGEMENT</b>	<b>iii</b>
<b>ABSTRACT</b>	<b>iv</b>
<b>TABLE OF CONTENT</b>	<b>v</b>
<b>LIST OF TABLES</b>	<b>viii</b>
<b>LIST OF FIGURES</b>	<b>ix</b>
<b>LIST OF ABBREVIATIONS</b>	<b>xi</b>
 <b>CHAPTER ONE: INTRODUCTION</b>	
1.0 Introduction	1
1.1 Research Background	1
1.2 Problem Statement	3
1.3 Objective of the Research	3
1.4 Scope of the Research	4
1.5 Significant of the Research	4
 <b>CHAPTER TWO: LITERATURE REVIEW</b>	
2.0 Bioinformatics	6
2.1 Deoxyribonucleic Acid (DNA)	9
2.2 DNA Sequence	11
2.3 Knowledge Discovery in Database	13
2.4 Data Mining	16
2.4.1 Classification	18

2.4.2	Prediction	19
2.4.3	Visualization	20
2.4.4	Affinity Grouping	20
2.4.5	Clustering	21
2.4.6	Estimation	22
2.5	Technique in Data Mining	23
2.5.1	Artificial Neural Network	23
2.5.2	Market Basket Analysis (MBA)	31
2.5.3	Cluster Analysis/ Target Analysis	33
2.5.4	Decision Tree	34
2.5.5	Genetic Algorithms	37
2.6	Tool for Sequencing	38
2.6.1	FASTA	38
2.6.2	BLAST	39

### **CHAPTER THREE: RESEARCH APPROACH AND METHODOLOGY**

3.0	Introduction	41
3.1	Problem Assessment and Resech Study	43
3.2	Knowledge Acquisition	43
3.2.1	Primary Data	43
3.2.2	Secondary Data	44
3.3	Data Analysis	44
3.4	Knowledge Representation	46

### **CHAPTER FOUR: ANALYSIS AND FINDINGS**

4.0	Introduction	47
4.1	Classification Method	48
4.1.1	Neural Network	48
4.1.2	Nearest-Neighbor Classifier	49