# Table of Contents

Dedication i  
Acknowledgements ii  
Abstract iv  
Contents viii  
List of Tables xii  
List of Figures xiii  
Publications based on this work xv  

## CHAPTER 1  INTRODUCTION

1.1 Motivation 1  
1.2 The Main Objectives of this Research 3  
1.3 Introduction to Bioinformatics 4  
1.4 Bioinformatics – The Genomic Revolution 9  
1.5 Biological Databases 10  
1.6 Thesis Organization 11  

## CHAPTER 2  LITERATURE REVIEW

2.1 Text Mining 13  
2.1.1 Text Mining from Databases 17  
2.1.1.1 On the Horizon 21  
2.1.1.2 The Human Factor 21  
2.1.2 Working Principle of Text Mining 22  
2.1.3 Text Mining Framework 25  
2.1.4 Knowledge Discovery 27  
2.1.5 Text Encoding 29  
2.1.6 Text Preprocessing 30  
2.1.7 Data Mining Methods for Text 30  
2.2 Biomedical Literature 31
2.2.1 Electronic Journals 33
2.3 National Centre for Biotechnology Information Medline 33
   2.3.1 MeSH 35
2.4 Pubmed Literature Database 37
   2.4.1 PubMed 37
      2.4.1.1 PubMed Features 38
      2.4.1.2 PubMed is easy to use 39
2.5 National Centre for Biotechnology Information Entrez Global Search 42
   2.5.1 Understanding Entrez 43
      2.5.1.1 The Entrez Engine : EG Query, E Search and E Summary 43
      2.5.1.2 Entrez Databases : Einfo, Efetch and ELink 44
      2.5.1.3 Building Customized Data Pipelines using the Entrez Programming utilities in BioPython 45
2.6 Mining Scientific Data from PubMed Database 47
2.7 About Python 48
   2.7.1 Organization of Python 51
   2.7.2 Essence of Python 51
   2.7.3 Python Cannot be Used 52
2.8 Introduction to BioPython 52
   2.8.1 BioPython Features 53

CHAPTER 3 STATEMENT OF THE PROBLEM 55
3.1 Problem Definition 55
CHAPTER 4 METHODOLOGY

4.1 EUTILS

4.1.1 The Seven Entrez Eutils

4.1.2 Parser Design

4.1.2.1 Scanners

4.1.2.2 Consumers

4.1.3 Events

4.1.4 Parsing PubMed with Element Tree

4.2 Additional Packages/Modules

4.3 Experimental SetUp

4.3.1 Software Configuration

4.3.1.1 Objectives

4.3.1.2 Installation of Python 2.5.4

4.3.1.3 Installation of BioPython

4.3.1.4 Installation of Numpy

4.3.2 Guidelines for Software Installation

4.4 Text Mining from NCBI PubMed Databases

4.5 Python Algorithm to Mine Textual Information from PubMed Databases

4.6 A Retrieval Performance Study of PubMed, Citeseer, ACM and IEEE Databases

CHAPTER 5 RESULTS AND DISCUSSION

5.1 Python Program to Mine Data from PubMed databases

5.1.1 Running Program in the Python IDLE

5.1.2 Providing Search Patterns, onset and edge year volume record count for salvage
5.1.3 Total records available in PubMed Databases each year based on Search Term

5.1.4 Transcends Fetched Planted in terms of search patterns

5.2 Mining the Scientific Data from PubMed Databases

5.2.1 Boolean Search

5.2.2 Geographical Contributions

5.3 PubMed, Citeseer, ACM and IEEE Databases Performance Evaluation

CHAPTER 5 CONCLUSIONS

6.1 Present Work

6.2 Future Enhancements

BIBLIOGRAPHY

APPENDIX: Authors Publications