Significant Pattern Discovery in Gene Location and Phylogeny

Michael C. Riley

Department of Computer Science
University of Wales, Aberystwyth

March

2009

This thesis is submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy of the University of Wales.

Declaration

This thesis has not previously been accepted in substance
for any degree and is not being concurrently submitted in candidature
for any degree.

Signed ..... (candidate)

Date .....

Statement 1

This thesis is the result of my own investigations, except
where otherwise stated.

Other sources are acknowledged by footnotes giving explicit
references. A bibliography is appended.

Signed ..... (candidate)

Date .....

Statement 2

I hereby give consent for my thesis, if accepted, to be made
available for photocopying and for inter-library loan, and for the
title and summary to be made available to outside organisations.

Signed ..... (candidate)

Date .....

Abstract

This thesis documents the investigation into the acquisition of knowledge from biological data using computational methods for the discovery of significantly frequent patterns in gene location and phylogeny.

Beginning with an initial statistical analysis of distribution of gene locations in the flowering plant *Arabidopsis thaliana*, we discover unexplained elements of order. The second area of this research looks into frequent patterns in the single dimensional linear structure of the physical locations of genes on the genome of *Saccharomyces cerevisiae*. This is an area of epigenetics which has, hitherto, attracted little attention. The frequent patterns are patterns of structure represented in Datalog, suitable for analyses using the logic programming methodology Prolog. This is used to find patterns in gene location with respect to various gene attributes such as molecular function and the distance between genes. Here we find significant frequent patterns in neighbouring pairs of genes. We also discover very significant patterns in the molecular function of genes separated by distances of between 5,000 and 20,000 base pairs. However, in complete contrast to the latter result, we find that the distribution of genes of molecular function within a local region of $\pm 20,000$ base pairs is locationally independent.

In the second part of this research we look for significantly frequent patterns of phylogenetic subtrees in a broad database of phylogenetic trees. Here we investigate the use of two types of frequent phylogenetic structures. Firstly, phylogenetic pairs are used to determine relationships between organisms. Secondly, phylogenetic triple structures are used to represent subtrees. Frequent subtree mining is then used to establish phylogenetic relationships with a high confidence between a small set of organisms. This exercise was invaluable to enable these procedures to be extended in future to encompass much larger sets of organisms.

This research has revealed effective methods for the analysis of, and has discovered patterns of order in the locations of genes within genomes. Research into phylogenetic tree generation based on protein structure has discovered the requirements for an effective method to extract elements of phylogenetic information from a phylogenetic database and reconstruct a single consensus tree from that information. In this way it should be possible to produce a species tree of life with high degree of confidence and resolution.

Acknowledgements

I would like to thank Prof. Ross D. King and Dr. Amanda Clare for their supervision and guidance. I would also like to thank Lynne Bowdon for her support and many hours of proof reading, and Jan Struyf and Hendrik Blockeel for technical assistance with the ACE data mining system.

My thanks also go to my friends and colleagues in C57: Dr. David Currie, Janet Mawby, Martin Robbins, Dr. Claire Rocks, Colin Sauze and Alan Woodland for their ideas and inspiration.

I would like to thank Dr. Maria Liakata and Dr. Wayne Aubrey for their feedback on the content of my thesis.

Finally, I would like to thank all the staff in the Computer Science Department of Aberystwyth University for their help and assistance throughout 2004–2008.

Contents

1 Introduction **1**

2 Background **5**
2.1 Genes and Gene Location 6
2.1.1 The Genome 6
2.1.2 Genome Morphology 6
2.1.3 Genome Order 9
2.1.4 Gene Classification 16
2.2 Proteins and Protein Structure 18
2.2.1 Proteins 18
2.2.2 Protein Structure 20
2.2.3 Protein Classification 21
2.3 Comparative Genomics and Proteomics 26
2.3.1 Sequence Alignment 26
2.3.2 The Dynamic Programming Algorithm 26
2.3.3 ClustalW 30
2.3.4 BLAST: Basic Local Alignment Search Tool 31
2.3.5 Profile Based Alignment 33
2.4 Logic Programming 36
2.4.1 Prolog 36
2.5 Frequent Pattern Mining 38
2.5.1 Apriori Algorithm 39
2.5.2 First Order Pattern Mining 43
2.6 Phylogeny 45
2.6.1 Introduction 45
2.6.2 Phylogenetic Trees 46
2.6.3 Broad Sampling and Consensus Trees 47
2.6.4 Phylogenetic Databases 47
2.6.5 New Hampshire/Newick format 50

3 Model Organisms **51**
3.1 The Model Organism *Arabidopsis thaliana* . . . 51
3.1.1 The *Arabidopsis thaliana* Genome . . . 52
3.1.2 *Arabidopsis thaliana* Genome Statistics . . . 53
3.2 The Model Organism *Saccharomyces cerevisiae* . . . 64
3.2.1 The *Saccharomyces cerevisiae* Genome . . . 64
3.2.2 *Saccharomyces cerevisiae* Genome Statistics . . . 66
3.3 Summary . . . 69

4 Statistical Tools and Methods **70**
4.1 Statistical Methods . . . 70
4.1.1 Moments . . . 70
4.1.2 Standard Error . . . 72
4.1.3 Mean Filtering . . . 73
4.2 Probability . . . 74
4.2.1 Factorial . . . 75
4.2.2 Random Selection . . . 76
4.2.3 Poisson Distribution . . . 77
4.2.4 Binomial Coefficient . . . 79
4.2.5 Binomial Distribution . . . 80
4.2.6 Multinomial Distribution . . . 81
4.3 Multiple Hypothesis Testing: Bonferroni Correction . . . 82
4.4 The Greenwood Statistic . . . 83
4.5 Summary . . . 84

5 Gene Location in *A. thaliana* **85**
5.1 Introduction . . . 85
5.1.1 The Locational Distribution of Genes . . . 86
5.1.2 Methodology Overview . . . 88
5.2 Methods . . . 89
5.2.1 Data . . . 89
5.2.2 Removal of Tandem Duplicates . . . 90
5.2.3 Distribution of All Genes . . . 90
5.2.4 The Locational Distribution of Functional Classes of Genes . 94
5.2.5 The Greenwood Statistic . . . 95
5.2.6 Ranking and P-values . . . 96
5.3 Results . . . 96
5.3.1 Distributions of All Genes . . . 96
5.3.2 The Locational Distribution of Functional Classes of Genes . 98
5.3.3 Clustered Distributions . . . 101

5.3.4 Evenly Spaced Distributions . . . 104
5.4 Discussion . . . 105
5.4.1 Tandem Duplicates . . . 105
5.4.2 Evenly Distributed Classes of Genes . . . 106
5.5 Conclusions . . . 107

6 Pattern Mining: Gene Location 108
6.1 Introduction . . . 108
6.1.1 Epigenetics . . . 109
6.1.2 Gene Location . . . 110
6.2 The SPD System . . . 111
6.2.1 Knowledge Base . . . 113
6.2.2 Background Knowledge . . . 113
6.3 Individual Gene Analysis . . . 115
6.3.1 Introduction . . . 115
6.3.2 Method . . . 115
6.3.3 Results . . . 116
6.3.4 Analysis . . . 119
6.4 Neighbouring Pairs . . . 121
6.4.1 Introduction . . . 121
6.4.2 Method . . . 121
6.4.3 Results . . . 123
6.4.4 Neighbouring Molecular Function Classes . . . 127
6.4.5 Analysis . . . 127
6.5 Clusters of Heterogeneous Gene Function . . . 128
6.5.1 Introduction . . . 128
6.5.2 Method . . . 129
6.5.3 Results . . . 131
6.5.4 Analysis . . . 137
6.6 Transitive Gene Sequences . . . 138
6.6.1 Introduction . . . 138
6.6.2 Method . . . 138
6.6.3 Results . . . 140
6.6.4 Analysis . . . 147
6.7 Discussion . . . 149
6.7.1 Main Discoveries . . . 149
6.8 Conclusions . . . 151

7 Pattern Mining: Molecular Phylogeny 153
7.1 Introduction . . . 153

7.2 Swiss-Prot Protein Sequence Data . . . 154
7.3 Superfamily Class Data . . . 155
7.4 Method . . . 157
7.4.1 Sequence Selection from the Swiss-Prot Database . . . 158
7.4.2 Generating the Phylogenetic Trees . . . 159
7.5 The Tree Database . . . 161
7.5.1 Knowledge Base . . . 161
7.5.2 Background Knowledge . . . 161
7.6 Parametric and Structural Queries . . . 162
7.7 General Queries . . . 166
7.7.1 Phylogenetic Pairs . . . 166
7.7.2 Method . . . 167
7.7.3 Results . . . 167
7.8 Protein Evolution . . . 176
7.9 Conclusion . . . 178

8 Phylogenetic Consensus Tree 180
8.1 Introduction . . . 180
8.2 Phylogenetic Triples . . . 183
8.3 Reconstructing Phylogenetic Trees . . . 184
8.4 New Knowledge Base . . . 184
8.4.1 Background Knowledge . . . 185
8.5 Phase I . . . 186
8.5.1 Phylogenetic Triple Miner . . . 188
8.5.2 Evaluation of Data . . . 190
8.5.3 Most Frequent Permutation . . . 194
8.5.4 Expectation and Confidence . . . 194
8.5.5 Analysis of Phase I Results . . . 197
8.6 Phase II . . . 199
8.6.1 Analysis of Phase II Results . . . 199
8.7 Results . . . 201
8.8 Conclusion . . . 208

9 Discussion 209
9.1 Introduction . . . 209
9.2 Review . . . 210
9.3 Key Findings and Results . . . 212
9.4 Significant Pattern Discovery in Epigenetics . . . 214
9.5 Phylogenetic Structure Mining . . . 215

9.5.1 Phylogenetic Consensus Tree 216

10 Conclusions **218**

A Supplementary Data: Gene Location **221**
A.1 Gene location 221
A.1.1 Key to tables 221

B Supplementary Data: Gene Neighbours **240**

References **244**

List of Figures

3.1 The model organism *Arabidopsis thaliana* 52
3.2 Graph of gene lengths for chromosome I of *A. thaliana* 55
3.3 Graph of gene lengths for chromosome II of *A. thaliana* 56
3.4 Graph of gene lengths for chromosome III of *A. thaliana* 57
3.5 Graph of gene lengths for chromosome IV of *A. thaliana* 58
3.6 Graph of gene lengths for chromosome V of *A. thaliana* 59
3.7 Pdf plots for gene gap lengths of *A. thaliana* 61
3.8 Common baker's yeast, *Saccharomyces cerevisiae*. 64
3.9 Histogram of the lengths of genes in *S. cerevisiae*. 68
3.10 Histogram of the gap lengths between genes in *S. cerevisiae*. 69

5.1 Graph of gene frequency on chromosome I 93
5.2 Box and whisker plots for the rankings of the molecular functional classes of *A. thaliana* without tandem duplicates 100
5.3 Box and whisker plots for the rankings of the molecular functional classes of *A. thaliana* including tandem duplicates 103

6.1 The SPD frequent pattern analysis system 111
6.2 Listing for the language bias file required for a statistical analysis of genes and their attributes using WARMR. 116
6.3 Listing for the language bias file required for frequent pattern discovery in neighbouring pairs of genes and their attributes using WARMR. 122
6.4 Frequency plots of the inter gene gap lengths of neighbouring pairs of the four transcription direction types; w-sequent, convergent, c-sequent and divergent. 125
6.5 Listing for the language bias required for the discovery of frequent patterns in clusters of heterogeneous gene function. 129
6.6 Listing for the *close_to_class*1 predicate function in the background knowledge file used in the frequent pattern mining search performed using WARMR. 130

6.7 Frequent pattern mining results for clusters of genes with heterogeneous molecular function showing the query pattern generated at each level and its relative frequency. 132
6.8 Listing for the *close_to_class* predicate function in the background knowledge file used in the frequent pattern mining search performed using WARMR 139
6.9 Frequent pattern mining results for transitive sequences of genes with heterogeneous molecular function for region length 5,000 bp . 141
6.10 Frequent pattern mining results for transitive sequences of genes with heterogeneous molecular function for region length 10,000 bp . 145
6.11 Frequent pattern mining results for transitive sequences of genes with heterogeneous molecular function for region length 20,000 bp . 146

7.1 Diagram of a phylogenetic pair 166

8.1 Diagram of a triple 183
8.2 Block diagram of triple mining procedure 187
8.3 Commonly accepted phylogeny of yeast 190
8.4 Frequent phylogenetic triples that conform to commonly accepted yeast phylogeny. 192
8.5 Frequent phylogenetic triples that do not conform with accepted yeast phylogeny. 193
8.6 Anomalous result in the phylogenetic relations of Man and two species of yeast 198
8.7 Anomalous result in the phylogenetic relations of Man, the Worm and the Fly 198
8.8 Accepted taxonomy of the subphyla of *N. crassa*, *S. cerevisiae* and *S. pombe* 200
8.9 Consensus tree for organisms in a clade to which *A. thaliana* does not belong. 203
8.10 Consensus tree result part 1 204
8.11 Consensus tree result part 2 205
8.12 Consensus tree result part 3 206
8.13 Consensus tree result part 4 207

List of Tables

2.1 Locations of G-protein complex genes. 15
2.2 Gene Ontology (GO) Data: The files and Datalog schema for the relational database. 17
2.3 List of all amino acids and corresponding codons 19
2.4 SCOP: Structural Classification of Proteins 23
2.5 List of all amino acids and probabilities 35
2.6 NCBI taxonomy statistics (August 2008). 49

3.1 Details of gene number and chromosome length in base pairs for the chromosomes of *A. thaliana*. 54
3.2 A definition for gene length classification for *A. thaliana*, which will be used in pattern mining. 54
3.3 A definition for gene gap length classification for *A. thaliana*, which will be used in frequent pattern mining. 60
3.4 Gene Ontology (GO) Data: Molecular function class density for level 1 classes for all five chromosomes of *A. thaliana*. 62
3.5 Level 2 molecular function classes for *A. thaliana* for all five chromosomes. 63
3.6 Details of the genome of *Saccharomyces cerevisiae* giving the base pair length of each of the 16 chromosomes and the number of genes on each chromosome. 66
3.7 Details of the location of centromeres on all 16 chromosomes of *S. cerevisiae*. 67
3.8 A definition for gene length classification for *S. cerevisiae*. 67
3.9 A definition for gene gap length classification for *S. cerevisiae*. . . . 69

5.1 Details of the centromeric regions excluded from the analysis showing the start and end locations of the centromeres determined by the method given in the text. 94
5.2 Ranking of the standard deviation in the distribution of the original genes against the mean of 1000 Monte Carlo simulations 97

5.3 Average ranking of all the functional classes analysed with and without tandem duplicates 98
5.4 Descriptions of the Gene Ontology annotations used in the boxplots in Figure 5.2 and Figure 5.3. 101

6.1 The files and formats for the gene location knowledge base for *Saccharomyces cerevisiae* 114
6.2 Individual gene analysis results (Warmr Level 2) showing the probabilities of genes of the attributes given in the second term in each clause. 117
6.3 Frequencies of genes of molecular function and length. 120
6.4 Relative frequency of the four types of neighbouring pairs related by locational and transcription direction. 124
6.5 Relative frequencies of different gap lengths between the four neighbouring pair types. 124
6.6 The molecular function classes of neighbouring pairs for the 10 most significant results 127
6.7 Significance results for the frequency of the WARMR frequent query at level 5 for an area $\pm 10,000$ bp 133
6.8 Significance results for the frequency of the WARMR frequent query at level 5 for an area $\pm 20,000$ bp 134
6.9 Significance results for the frequency of the WARMR frequent query at level 6 for an area $\pm 10,000$ bp 135
6.10 Significance results for the frequency of the WARMR frequent query at level 6 for an area $\pm 20,000$ bp 136
6.11 Significance results for the frequency of patterns of transitive sequences of genes of heterogeneous molecular function for a region of $5,000$ bp. 142
6.12 Significance results for the frequency of patterns of transitive sequences of genes of heterogeneous molecular function for a region of $10,000$ bp. 143
6.13 Significance results for the frequency of patterns of transitive sequences of genes of heterogeneous molecular function for a region of $20,000$ bp. 144
6.14 The most frequently populated region length for each chromosome of *S. cerevisiae* 147
6.15 Level 1 molecular function classes for *S. cerevisiae*, showing the Gene Ontology identifier and a description of the molecular function. 151

7.1 File names and datalog schema of the background knowledge files suitable for use with the trees knowledge base. 161

7.2 Organisms having a common ancestor with *H. sapiens* (part 1) . . . 169
7.3 Organisms having a common ancestor with *H. sapiens* (part 2) . . . 170
7.4 Human/polecat analysis results . . . 172

8.1 Background knowledge files and schema. . . . 185
8.2 Yeasts used as candidate organisms for validation of the results. . . 191
8.3 Eighteen of the most significant examples of phylogenetic triples containing *Homo sapiens* . . . 197
8.4 The top five most significant examples of phylogenetic triples from the *phase II* results . . . 200

A.1 Results for gene classes taken from level 1 of the Gene Ontology heirarchy on chromosome 1 using TIGR data with the tandem duplications removed. . . . 222
A.2 Results for gene classes taken from level 1 of the Gene Ontology heirarchy on chromosome 2 using TIGR data with the tandem duplications removed. . . . 223
A.3 Results for gene classes taken from level 1 of the Gene Ontology heirarchy on chromosome 3 using TIGR data with the tandem duplications removed. . . . 223
A.4 Results for gene classes taken from level 1 of the Gene Ontology heirarchy on chromosome 4 using TIGR data with the tandem duplications removed. . . . 224
A.5 Results for gene classes taken from level 1 of the Gene Ontology heirarchy on chromosome 5 using TIGR data with the tandem duplications removed. . . . 224
A.6 Results for gene classes taken from level 2 of the Gene Ontology heirarchy on chromosome 1 using TIGR data with the tandem duplications removed. . . . 225
A.7 Results for gene classes taken from level 2 of the Gene Ontology heirarchy on chromosome 2 using TIGR data with the tandem duplications removed. . . . 226
A.8 Results for gene classes taken from level 2 of the Gene Ontology heirarchy on chromosome 3 using TIGR data with the tandem duplications removed. . . . 227
A.9 Results for gene classes taken from level 2 of the Gene Ontology heirarchy on chromosome 4 using TIGR data with the tandem duplications removed. . . . 228
A.10 Results for gene classes taken from level 2 of the Gene Ontology heirarchy on chromosome 5 using TIGR data with the tandem duplications removed. . . . 229

A.11 Results for gene classes taken from level 3 of the Gene Ontology heirarchy on chromosome 1 using TIGR data with the tandem duplications removed. . . . 230
A.12 Results for gene classes taken from level 3 of the Gene Ontology heirarchy on chromosome 2 using TIGR data with the tandem duplications removed. . . . 231
A.13 Results for gene classes taken from level 3 of the Gene Ontology heirarchy on chromosome 3 using TIGR data with the tandem duplications removed. . . . 232
A.14 Results for gene classes taken from level 3 of the Gene Ontology heirarchy on chromosome 4 using TIGR data with the tandem duplications removed. . . . 233
A.15 Results for gene classes taken from level 3 of the Gene Ontology heirarchy on chromosome 5 using TIGR data with the tandem duplications removed. . . . 234
A.16 Results for gene classes taken from level 4 of the Gene Ontology heirarchy on chromosome 1 using TIGR data with the tandem duplications removed. . . . 235
A.17 Results for gene classes taken from level 4 of the Gene Ontology heirarchy on chromosome 2 using TIGR data with the tandem duplications removed. . . . 236
A.18 Results for gene classes taken from level 4 of the Gene Ontology heirarchy on chromosome 3 using TIGR data with the tandem duplications removed. . . . 237
A.19 Results for gene classes taken from level 4 of the Gene Ontology heirarchy on chromosome 4 using TIGR data with the tandem duplications removed. . . . 238
A.20 Results for gene classes taken from level 4 of the Gene Ontology heirarchy on chromosome 5 using TIGR data with the tandem duplications removed. . . . 239

B.1 The molecular function classes of neighbouring pairs (84 results). Table 1 of 3. . . . 241
B.2 The molecular function classes of neighbouring pairs (84 results). Table 2 of 3. . . . 242
B.3 The molecular function classes of neighbouring pairs (84 results) . Table 3 of 3. . . . 243

Chapter 1

Introduction

Central to our understanding of life is to understand how the cell and its DNA function to create living organisms. The immense complexity of the cell is on a scale beyond imagination. Consequently, the quantity of data produced from biological research is on a similar scale. This data alone does little to aid our understanding of life until we can extract the knowledge it contains.

This thesis investigates the acquisition of knowledge through the use of pattern mining in biological data. It is a new area in the larger field of bioinformatics, which is itself a relatively new field using computational methods to process and analyse biological data to produce new information.

Sequencing of genomes is now commonplace with improved sequencing methods resulting in new organism genome sequences being published frequently. A recently introduced method called *pyro sequencing*[1] has greatly increased the speed of genome sequencing such that now entire genomes can be sequenced within a day. This produces huge quantities of data that needs computational methods for analysis and knowledge acquisition. These methods include sequence matching, gene/protein prediction and classification.

There is an acknowledged requirement for the acquisition of more knowledge about genes and proteins than is presently discovered through sequence alignment alone.

[1] More information at www.pyrosequencing.com/

In this thesis several related knowledge discovery and acquisition methods are introduced:

1. A novel application of the Greenwood statistic and Monte Carlo methods to the genome of the flowering plant *Arabidopsis thaliana* reveals unexplained order in the location of genes classified by molecular function.
2. Frequent pattern mining methods combined with Monte Carlo methods and significance ranking are used to discover knowledge in the location of genes with respect to many different gene attributes in the fungus *Saccharomyces cerevisiae*.
3. The creation of a large phylogenetic database detailing the evolutionary histories of protein sequences from many organisms where the protein sequences are classified into groups of homologous protein sequences.
4. The use of frequent patterns of phylogenetic structure to extract data with high confidence from large phylogenetic databases.

These methods provide more information to aid the identification of unknown genes and proteins than the use of sequence alignment alone. Furthermore, elements from this research may provide information on cellular functions dependent on cooperating proteins by revealing which proteins cooperate from their gene location and phylogeny.

This thesis is organised as follows:

- Chapter 2 discusses the biological motivation for pattern mining in the physical locations of genes on the genome, followed by an introduction to proteins and protein structure. It then goes on to cover a general background to comparative genomics, logic programming, frequent pattern mining and then discusses phylogenetics.
- Chapter 3 introduces the model organisms *Arabidopsis thaliana* and *Saccharomyces cerevisiae* used in the research into gene distribution and location described in later chapters.
- Chapter 4 discusses the statistical and mathematical tools and methods used

in this research.

- Chapter 5 covers new work on the statistical analysis of the distribution of genes on the genome of *Arabidopsis thaliana*. A novel bioinformatics method was developed based on Monte Carlo methods and Greenwood's spacing statistic for the computational analysis of the distribution of individual functional classes of genes. This work has been published in BMC Bioinformatics (Riley *et al.* , 2007).

- Chapter 6 describes a significant frequent pattern mining system and its application to data mining in the genome of *Saccharomyces cerevisiae*. The system is called the SPD (Significant Pattern Discovery) system. Essentially, it is an extension of WARMR, which is a frequent pattern mining program, providing automated filtering and significance determination of discovered frequent patterns. WARMR is one of several tools incorporated within ACE, an ILP data mining package and it was applied to the physical locations of genes with respect to various gene attributes. Such attributes include molecular function, gene length, direction of transcription, location on strand and relative location between genes.

- Chapter 7 describes the creation of a knowledge base of phylogenetic trees for protein groups classified by SCOP, which is a protein domain database, and then discusses the research into frequent pattern mining of phylogenetic data. The data preparation for this analysis proved to be a challenge utilizing many bioinformatics techniques and methods including BLAST and ClustalW and some in depth statistical analysis. Essentially, this research involved selecting known protein sequences from Swiss-Prot, a well annotated protein database, for many organisms and classifying the protein sequences according to the SCOP superfamily classes using BLAST. The organisms from which the proteins from each class were obtained were organised into a phylogenetic tree database using ClustalW.

- Chapter 8 takes a more in depth look at the use of particular patterns of phylogenetic structure to deconstruct and reconstruct phylogenetic trees. Frequent pattern mining was used to find frequent phylogenetic triples in

these trees and those triples with a high statistical confidence were used to establish a new phylogenetic order for the constituent organisms.

- Chapter 9 is an in depth discussion on the results presented in this thesis and presents some future directions for research.
- Chapter 10 summarises the main conclusions and findings.

Chapter 2

Background

A general background to the research in this thesis is presented in this chapter. There are six main areas:

1. Genes, gene location and the classification of genes.
2. Proteins, protein structure and the classification of proteins according to structure.
3. Comparative genomics and proteomics, which discusses methods of gene and protein sequence alignment.
4. A computational methodology known as logic programming, which proved invaluable in the analysis of data.
5. Frequent pattern mining and the specific frequent pattern mining program WARMR.
6. Phylogenetics and phylogenetic trees, which can be used to trace the evolution of living organisms.

These areas are presented in six separate sections in this chapter. In addition to this, there are two further areas of background discussion: model organisms are introduced in Chapter 3; and the statistical tools and methods required in this research are outlined in Chapter 4. More specific background information appears

throughout this thesis.

2.1 Genes and Gene Location

In this section we discuss genes and the physical location of genes on the genome. Further, the factors that contribute to the randomizing of gene location and the factors that preserve order in gene location are presented. This lays down the background and motivation for the research discussed in Chapters 5 and 6.

2.1.1 The Genome

The genome is the name given to all of the genetic information within the cells of an organism and also refers to the DNA that carries that information (Alberts *et al.* , 2002). The genome may be a single chromosome or divided into several chromosomes, which are very long DNA molecules and associated proteins carrying all or part of the hereditary information of an organism. The genome, and consequently the chromosomes, consist of individual units of coding information called genes. A gene is a region of DNA that controls a discrete hereditary characteristic usually corresponding to a single protein or RNA. This definition includes the entire functional unit, encompassing coding DNA sequences, noncoding regulatory DNA sequences, and introns (Alberts *et al.* , 2002). Genes function by controlling the synthesis of proteins that act as biological catalysts in cellular pathways[1] and collectively, they can be regarded as the *inherited determinant of the phenotype* (Tamarin, 1999).

2.1.2 Genome Morphology

Genes and genomes are subject to change by several different kinds of mutation:

[1]In fact, there are many exceptions including genes for structural proteins such as those in hair, and genes that code for RNAs that function directly as RNAs (Tamarin, 1999).

- Single locus mutation.
- Tandem duplication of genes.
- Mobile genetic elements.
- Whole genome duplication.

With this variety of different mutations it is clear that in the absence of selective pressure on gene order, successive rearrangements will lead to the complete randomization of gene order (Durand & Sankoff, 2003).

Single locus mutation

A single locus mutation may occur during cell division where a single nucleotide in the daughter DNA strand is erroneously copied. DNA replication and DNA repair mechanisms are so efficacious that only 1 nucleotide pair in 1000 is randomly changed in every 200,000 years (Alberts *et al.* , 2002). Note that single locus mutation should not be confused with single nucleotide polymorphism (SNP). SNPs account for variation between individuals of the same species at certain nucleotide positions in the genome.

Tandem duplication

A tandem duplication or tandem repeat is an adjacent identical chromosome segment. If the duplicate segment is an entire gene then this frees one copy of each gene sequence to drift and potentially to acquire a new function. Tandem duplication is thought to occur by a phenomenon known as *unequal crossover* (Smith, 1976).

Mobile genetic elements

Mobile genetic elements are a type of DNA that can move around within the genome. They include:

- Transposons
- Organellar DNA
- Viral DNA

There are three types of transposon: retrotransposons, DNA transposons and insertion sequences. Retrotransposons (class I mobile genetic elements) move in the genome by being transcribed to RNA and then back to DNA by reverse transcriptase (SanMiguel & Bennetzen, 1998). DNA transposons (class II mobile genetic elements) move from one position to another within the genome using a transposase in a 'cut and paste' fashion. Insertion sequences (Mahillon & Chandler, 1998)(also known as an IS, an insertion sequence element, or an IS element) is a short DNA sequence that acts as a simple transposable element. Insertion sequences are small relative to other transposable elements (700 to 2500 bp in length) and only code for proteins that catalyse the enzymatic reaction in the transposition activity and for proteins that either stimulate or inhibit the transposition activity. They differ from transposons in that they do not carry accessory genes.

Organellar DNA is DNA within the cell that does not include the main genome or nuclear DNA. Unlike nuclear DNA, which is present as linear molecules, organellar DNA is present as circular DNA molecules of high copy number. Plasmids are also extra-nuclear strands of DNA. There are two types: non-integrating plasmids and episomes. Non-integrating plasmids replicate independently, whereas episomes integrate into the host chromosome and this will affect gene location.

Viral DNA or bacteriophage elements are passed into the cell by viruses. This DNA can also integrate into the host genome and thereby affect gene location.

Whole genome duplication

Whole genome duplication and gene duplication can be considered a major force in evolution since when a redundant gene locus is created by duplication it can then freely acquire mutations and emerge as a new gene with an entirely new function (Ohno, 1970).

A considerable synteny in the locations of genes between *Ashbya gossypii* and *Saccharomyces cerevisiae* has been documented in a paper by Altmann-Johl (Altmann-Johl & Philippsen, 1996) and this initiated the *Ashbya Gossypii Project* under Peter Philippsen of the University of Basel, Switzerland.

It is believed that more than 100 million years ago[2] *A. gossypii* and *S. cerevisiae* diverged after an ancestor of *S. cerevisiae* experienced a whole genome duplication (WGD) event and consequently the genomes differ substantially in GC content (52% for *A. gossypii* and 38% for *S. cerevisiae*), but for 95% of the protein coding sequences of *A. gossypii*, there exist homologs in the *S. cerevisiae* genome (Dietrich *et al.* , 2004).

It is believed that the WGD of *S. cerevisiae* opened new possibilities for functional divergence not available to *A. gossypii* and the WGD event created approximately 5000 twin ORFs[3] in *S. cerevisiae* of which 496 of these ancient twin ORFs can still be seen in the double synteny patterns produced by Dietrich *et al.* (2004).

The evidence presented for a WGD event is compelling and if this is the case, then we may find clusters of inter-dependent genes that have been entirely conserved in one chromosome and entirely deleted in the other.

2.1.3 Genome Order

In this section we look into the factors that preserve or possibly even create order in the location of genes on the genome. These are:

- Operons.
- Multi domain proteins.
- The structure of chromatin.

[2]The age of this event is estimated from the mutation rate of 16s ribosomal RNA.

[3]An Open Reading Frame (ORF) is a sequence of DNA that is considered to be a potential coding sequence. An ORF is recognised by having a start codon (ATG , and also TTG and GTG in bacteria) and includes all codons in the sequence (usually longer than 100 nucleotides) up to the stop codon (TAA, TAG or TGA). Unlike the start codon, the stop codon does not code for an amino acid and is not included in the ORF (Lackie, 2007).

- Chromosomal interaction.
- Protein complexes.
- Housekeeping genes.

Operons

Operon is a term that refers to a control mechanism, but generally, operons are understood to be a sequence of adjacent genes transcribed into a single messenger RNA, to encode proteins with related functions. They are characterized by having protein coding sequences very close to each other on the chromosome and in some cases the protein coding sequences overlap (Moreno-Hagelsieb & Collado-Vides, 2002). The advantage of this is that cooperating proteins can be expressed simultaneously and/or localized spatially in the cytoplasm.

Their existence has been known since the early 1960s after the work of Jacob and Monod on the lac operon in *Escherichia coli* (Jacob & Monod, 1961). Since then operons have been found predominantly in prokaryotes and were thought to be exclusive to prokaryotes until recently where they have now been found in eukaryotes such as the nematode worm (*Caenorhabditis elegans*) (Blumenthal & Gleason, 2003) (Blumenthal, 2004).

The reason for the tendency of genes to be tightly packed together in operons may be simply one of economy. Considering the spacial localization of genes, a more significant reason might be to protect the resulting mRNA from degradation by association with ribosomes (Moreno-Hagelsieb & Collado-Vides, 2002). However, in a published summary of the work by Schneider *et al.*, which was cited by Moreno-Hagelsieb, they claim that the existence of nearby ribosomes did not play any role in the degradation of lac operon mRNA, but they do acknowledge that mRNA degradation occurs (Schneider *et al.* , 1978).

The lac operon is involved in lactose metabolism in the bacteria *E. coli*. Lactose is a β galactoside that can be used by *E. coli* for energy and as a source of carbon when broken down by β galactosidase. However, two other enzymes are required

before *E. coli* can metabolize lactose correctly and they are β galactoside permease and β galactoside acetyltransferase. In wild type *E. coli* grown in the absence of lactose, there are very few molecules of β galactosidase, however, when lactose is added to the growth medium, β galactosidase, β galactoside permease and β galactoside acetyltransferase are produced in abundance within the bacterial cell (Elliot & Elliot, 1997) (Tamarin, 1999) (Jacob & Monod, 1961). This shows that the lac operon fulfils a temporal requirement of the cell, enabling the transcription of necessary enzymes only when they are required.

Moreno-Hagelsieb *et al.* (2002) declare a method to detect operons based on their inter-genic distances. Further, they predicted that inter-genic distances of operon genes are actually overlaps of 1 to 4 base pairs. This may not be such a good criteria for determining operons since there is no reason why non-operon genes should not overlap, particularly on the densely packed genomes of bacteria.

Operons can fulfil both a spatial requirement of cell activity where genes need to be localized to build, for example, a complex multi protein enzyme and a temporal requirement where genes are needed to be expressed simultaneously. Sequences on the chromosome that transcribe to operons will be tightly packed and so these sequences are most likely to be gene clusters (Moreno-Hagelsieb & Collado-Vides, 2002).

Proteolysis may also play an important role in occurrence of operons.

Proteolysis

The interior of a cell reflects the harsh survival environment you find in nature. Newly created proteins face a tough battle to reach their functional goal/purpose. They are prey to enzymes whose job it is to recycle unused proteins and fragments of polypeptides (long sequences of amino acids). This means that inter-dependent proteins do not have very much time to fulfil their function before being destroyed. This general 'housekeeping' work of a cell is called proteolysis. This is the degradation or recycling of proteins that are damaged, misfolded, unassembled or unused and also to regulate the concentrations of certain normal proteins promptly in

response to the state of the cell (Alberts *et al.* , 2002). Proteolysis may also provide an explanation for the existence of operons. The assembly of multi-protein structures such as ribosomes is likely to be more efficient if the individual proteins are translated locally and operons can facilitate this.

Multi domain proteins

Multi domain proteins are similar to operons in that they are thought to have been separate genes originally that have fused into one gene that transcribes into a single mRNA, which, unlike operons, translates into a single protein that has multiple molecular functions (Apic *et al.* , 2003). In the gene location analysis described in Chapters 5 and 6, multi domain proteins will appear as large single genes and so are unlikely to affect the distribution of the locations of genes. There is more information on protein domains in Section 2.2.1.

The physical structure of the chromatin

The degree of coiling of the chromatin, which is the name given to the molecular and physical structure of chromosomes, varies during the life cycle of the cell bringing a varying number of genes into expression at different times. More genes are available for expression during the phases required for cell division. The chromatin is more tightly coiled when a cell has 'matured' and is performing a specific duty so evidently less genes are available for expression. Clearly the genes required for the cell's specific function must be available and will need to be located on regions of the chromatin that are 'open' for expression. This could be another factor affecting the distribution of genes. The role of chromatin in gene expresion and phenotype development is a relatively new area of research in the field of *epigenetics* and is discussed in more detail in Chapter 6.

Chromosomal interaction

Recent work by Spilianakis *et al.* on human T helper (T_H) cells (Spilianakis *et al.* , 2005) has revealed that the location of genes on the chromosomes may even be functionally critical through interaction with other chromosomes. They found that certain regions from different chromosomes appear to 'communicate' with each other by bringing related genes together in the nucleus of naive T_H cells to coordinate their expression. Once the naive T_H cell has differentiated to either a T_H1 or a T_H2 cell the chromosomal regions move apart and it is believed that this represses the genes required to differentiate the cells. This phenomenon will have a bearing on the distribution of genes especially if this is later found to be a common method to control gene expression.

Protein complex

A protein complex is a group of two or more associated proteins. A complex formation often serves to activate or inhibit member proteins. There are presently 232 identified protein complexes in *S. cerevisiae*[4] (Gavin *et al.* , 2006a) (Gavin *et al.* , 2006b). Two examples of protein complexes are discussed below.

MRE11 complex

The MRE11 complex is a trimeric protein complex that possesses endonuclease activity and it is involved in DNA repair and checkpoint signaling. In Saccharomyces the complex comprises three proteins: Mre11p, Rad50p, and Xrs2p. Complexes identified in other species generally contain proteins related to these Saccharomyces proteins[5].

Also known as the MRX complex, the MRE11 complex in yeast performs a large range of functions in conjunction with many different sets of proteins. These range

[4]Genes involved in yeast protein complexes are listed at:
http://yeast.cellzome.com/.

[5]Definition taken from SGD website.

from helping (Spo11 topoisomerase) to create double-strand breaks during meiosis, to the maintenance of telomeres, to various types of repair and processing of double-strand breaks in both meiosis and mitosis. These proteins also play a role in the checkpoint responses of cells to the presence of a chromosome break. Many other roles in mammalian cells have been inferred but the absence of viable mutants of Mre11p or Rad50p has prevented a direct assessment of their functions.

The MRE11 complex comprises three proteins; XRS2, MRE11 and RAD50.

XRS2 (Yeast gene YDR369C) Protein required for DNA repair; component of the Mre11 complex, which is involved in double strand breaks, meiotic recombination, telomere maintenance, and checkpoint signalling.
Location; ch4 1215007 - 1217571 bp.

MRE11 (Yeast gene YMR224C) Subunit of a complex with Rad50p and Xrs2p (RMX complex) that functions in repair of DNA double-strand breaks and in telomere stability, exhibits nuclease activity that appears to be required for RMX function; widely conserved.
Location ch13 718574 - 720652 bp.

RAD50 (Yeast gene YNL250W) Subunit of MRX complex, with Mre11p and Xrs2p, involved in processing double-strand DNA breaks in vegetative cells, initiation of meiotic DSBs, telomere maintenance, and nonhomologous end joining.
Location; ch14 175411 - 179349 bp.

The three genes that code for the MRE11 proteins are widely dispersed throughout the genome[6].

Protein	Gene	Chr	Bp location	Description
GPA1	YHR005C	8	113k	alpha subunit
STE4	YOR212W	15	743k	beta subunit
STE18	YJR086W	10	586k	gamma subunit
GPR1	YDL035C	9	392k	GPCR
STE2	YFL026W	6	83k	GPCR
STE3	YKL178C	11	114k	GPCR
GPG1	YGL121C	7	281k	gamma mimic

Table 2.1: Locations of G-protein complex genes.

G-protein complex

The locations of G-protein complex genes are shown in Table 2.1. From this Table it can be seen that, like the MRE11 complex above, the genes for the G-protein complex are widely dispersed. The nature of this dispersion is worthy of consideration in the research of gene location.

Housekeeping Genes

Housekeeping genes are genes that are constitutively expressed in most, if not all cells. These genes encode proteins that provide the basic, essential functions that all cells need to survive. It is generally assumed that housekeeping genes express at the same level in all cells and tissues, but there are actually some variances, especially during cell growth and organism development. It is estimated that there are about 300-500 house keeping genes in humans. Many hundreds of housekeeping genes have been identified. One of the most commonly known genes, GAPDH (glyceraldehyde-3-phosphate dehydrogenase), codes for an enzyme that is vital to the glycolytic pathway. Since housekeeping genes are essential, their location on the genome is likely to be significant.

[6]There is more information on the MRE11 complex at:-
http://mips.gsf.de/proj/yeast/pathways/haber.html and
http://db.yeastgenome.org/cgi-bin/locus.pl?dbid=S000004837.

2.1.4 Gene Classification

The Gene Ontology Consortium (GO) (2000) is a consortium providing a vocabulary for the annotation of genes/proteins. It has three organizing principles: 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.

Cellular component A cellular component is a component of a cell, which is part of some larger object such as an anatomical structure or a gene product group.

Biological process A biological process is series of events accomplished by one or more ordered assemblies of molecular functions. Broad examples are cellular physiological process or signal transduction. Examples of more specific terms are pyrimidine metabolic process or alpha-glucoside transport. Generally, unlike molecular function, a biological process must have multiple distinct steps.

Molecular function Molecular function describes activities that occur at the molecular level. 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. To avoid confusing a gene product name with its molecular function many GO molecular functions are appended with the word "activity".

Throughout this research on genes the focus is on molecular function. The GO version used is revision 1.11. The files and Datalog schema of the Gene Ontology annotations used throughout this research are given in Table 2.2.

Another valuable annotation of genes/proteins is by structure. This is discussed in Section 2.2.3.

File	Format
m_class_level.pl	m_class_level(class, level)
clss_hier.pl	child(subclass, class)

Table 2.2: Gene Ontology (GO) Data: The files and Datalog schema for the relational database.

2.2 Proteins and Protein Structure

2.2.1 Proteins

Proteins are the essential molecular components of cells. They are linear polymers of amino acids linked together by peptide bonds in a specific sequence (Alberts *et al.* , 2002). There are 20 different amino acids involved in protein structure, which are listed in Table 2.3. Amino acids are also frequently referred to as *residues*[7] when incorporated into a protein sequence. Proteins are also referred to as *polypeptides* and *polypeptide chains*.

A textual representation of a protein sequence in FASTA format:

```
>Tc00.1047053503625.10 Trypanosoma_cruzi
MTSGDPAAFIQLQEQIVTVKQVFSSALAKELNLVEVQAPLLACCGDGTQDNLSGTEKAVQ
VHVKGIPDSKFEVVHSLAKWKRQTLGDHKFPVGQGIYVHMKALRVEEELDTTHSVFVDQW
DWELVMPPQERNLTFLKNTVQRLYAAIRQTEEAICSKYNLDRVLPANIQFLHAEHLLKMY
PEMNMKERERAIVKKYGAVFLIGIGGNLTSGEPHDLRAPDYDDWSSPVSAADITFPCGDP
TMNSLASLPGLNGDILVYNPVLDDVLELSSMGIRVDAETLRRQLTLLSNEDRLGYVWHKR
LLAGEFPQTIGGGIGQSRLLMLLLKKKHIGEVQCSVWPKEMRQNYPLL
```

Proteins are generally composed of one or more, smaller, stable, independent functional or structural units, which are of particular interest in the identification and classification of proteins. These independent units are known as *protein domains* (Wetlaufer, 1973).

Protein domains

A protein domain is often found as an independent substructure within a protein sequence. It is considered to be a functional and/or evolutionary unit, which can exist independently of the rest of the host protein. Each domain can fold

[7] A *residue* is actually a general term for the unit of a polymer. More specifically it is that portion of a sugar, amino acid, or nucleotide that is retained as part of the polymer chain during the process of polymerization.

Amino Acid	Sym.	Mne	Codons
alanine	A	Ala	GCA GCC GCG GCU
cysteine	C	Cys	UGC UGU
aspartic acid	D	Asp	GAC GAU
glutamic acid	E	Glu	GAA GAG
phenylalanine	F	Phe	UUC UUU
glycine	G	Gly	GGA GGC GGG GGU
histidine	H	His	CAC CAU
isoleucine	I	Ile	AUA AUC AUU
lysine	K	Lys	AAA AAG
leucine	L	Leu	UUA UUG CUA CUC CUG CUU
methionine	M	Met	AUG
asparagine	N	Asn	AAC AAU
proline	P	Pro	CCA CCC CCG CCU
glutamine	Q	Gln	CAA CAG
arginine	R	Arg	AGA AGG CGA CGC CGG CGU
serine	S	Ser	AGC AGU UCA UCC UCG UCU
threonine	T	Thr	ACA ACC ACG ACU
valine	V	Val	GUA GUC GUG GUU
tryptophan	W	Trp	UGG
tyrosine	Y	Tyr	UAC UAU
STOP			UAA UGA UAG

Table 2.3: A list of all amino acids (residues): the building blocks of proteins. The heading **Sym.** signifies a designatory letter used as a symbol in the text representation of each amino acid in protein sequences; **Mne** is a mnemonic for each amino acid and **Codons** lists all the sets of three nucleotides that translate to produce each amino acid.

autonomously into a stable, compact three-dimensional structure, which is known as the *tertiary structure* (see below). Any single domain may appear in a variety of evolutionarily related proteins.

Domains vary in length, but have limits on size (Savageau, 1986). The size of presently known structural domains varies from 36 residues in E-selectin to 692 residues in lipoxygenase-1 (Jones *et al.* , 1998). The average length of all domains is approximately 100 residues (Islam *et al.* , 1995), with the majority of approximately 90%, having less than 200 residues (Siddiqui & Barton, 1995).

Most proteins contain several domains forming multidomain proteins, which are consequently multifunctional (Chothia, 1992). Each domain may fulfil its own function independently, or it may function in cooperation with other domains within the host protein. Domains are units of *tertiary structure* within proteins.

2.2.2 Protein Structure

Primary structure

The primary structure refers to the string of amino acids or residues from which the protein is composed.

Secondary structure

Secondary structure is a term used to refer to a regular local folding pattern in polymers. In proteins there are two main types of secondary structure:

- α-helices: a linear sequence of amino acids that fold into a right-handed helix, which is stabilized by internal hydrogen bonding between the polypeptide backbone atoms.
- β-sheets: where different sections of the polypeptide chain run alongside each other, joined by hydrogen bonding between atoms of the polypeptide

backbone. Also known as a β-pleated sheet.

Combinations of secondary structure elements have been found to frequently occur in protein structure and are referred to as *secondary structure motifs* or simply *motifs*.

Tertiary structure

The overall 3-dimensional structure of the protein is referred to as the *tertiary structure*. Multiple secondary structure motifs pack together to form compact, local, semi-independent units called domains (Richardson, 1981). Domains are the fundamental units of tertiary structure.

2.2.3 Protein Classification

SCOP: Structural classification of proteins

SCOP is a database of protein domains (Murzin *et al.* , 1995) (Lo Conte *et al.* , 2002) (Andreeva *et al.* , 2004). SCOP classification is on hierarchical levels that embody the evolutionary and structural relationships. Those levels, going from the most specific to the most general are: *family*, *superfamily*, *common fold* and *class*.

Family

Proteins in families are classified based on two criteria that imply a common evolutionary origin. Those criteria are firstly, that all proteins have $> 30\%$[8] residue identity and secondly, a lower sequence identity of $> 15\%$[9] (nucleotide alignment), but with similar functions and structures.

[8]The definition of a protein family has been recently updated to contain proteins with residue identies of $> 35\%$ (Xiong, 2008).

[9]In the paper by Murzin et al (Murzin *et al.* , 1995), they give an example for globins having sequence identities of 15% so the sequence identity criterion of $> 15\%$ is an assumption based on their example.

Superfamily

The superfamily classification, as the name suggests, is a super class of the family classification. Superfamilies contain families whose proteins have low sequence identities, but whose structures and, in many cases, functional features suggest that a common evolutionary origin is probable.

Common fold

The definition for the common fold classification is that proteins have the same major secondary structures in the same arrangement with the same topological connections.

Class

There are presently seven classes, which are designated by the letters *a*–*g*. There are also four extra groups that are not true classes, which are designated by the letters *h*–*k*:

a *All alpha*: proteins having a structure essentially formed from α-helices.

b *All beta*: proteins having a structure essentially formed from β-sheets.

c *Alpha and beta*: proteins with largely interspersed α-helices and β-strands.

d *Alpha plus beta*: proteins with largely segregated α-helices and β-strands.

e *Multi-domain*: proteins with domains of different fold for which there are presently no known homologues.

f *Membrane and cell surface proteins and peptides*: does not include proteins in the immune system.

g *Small proteins*: usually dominated by ligand, heme, and/or disulfide bridges.

h *Coiled coil proteins*: not a true class.

i *Low resolution protein structures*: not a true class

Class	Folds	Superfamilies	Families
All alpha proteins	259	459	772
All beta proteins	165	331	679
Alpha and beta proteins (a/b)	141	232	736
Alpha and beta proteins (a+b)	334	488	897
Multi-domain proteins	53	53	74
Membrane and cell surface proteins	50	92	104
Small proteins	85	122	202
Total	1086	1777	3464

Table 2.4: SCOP: Structural Classification of Proteins (release 1.73, 26th September 2007).

j *Peptides*: peptides and fragments. Not a true class.

k *Designed proteins*: experimental structures of proteins with essentially non-natural sequences. Not a true class.

Note that only the first seven classes are relevant in this research.

Protein sequences that have been classified according to SCOP will have an alpha-numeric identifier in four parts separated by full stops. For example:

SCOP ID is *a.1.2.3* where:

a = Class

1 = Fold

2 = SuperFamily

3 = Family

Scop classification statistics

The classification statistics for the SCOP database used in this research, which was last updated on 27th September 2007, are given in Table 2.4. It is this version of SCOP that was used in the research described in Chapter 8.

Superfamily library

In 2001 Gough *et al.* constructed a library of hidden Markov models that represent all proteins of known structure (Gough *et al.* , 2001). This library is called Superfamily. The sequences of the domains in proteins of known structure, that have identities less than 95%, are used as seeds to build the models. The sequences used by Superfamily to generate the models are from the ASTRAL database (Brenner *et al.* , 2000). The ASTRAL database provides protein sequences categorised according to the SCOP domain definitions and are derived from the SEQRES entries in PDB files[10]. The SEQRES records in a PDB file contain the amino acid or nucleic acid sequence of residues in each chain of the macromolecule that was studied. The sequences used by Superfamily differ from the ASTRAL sequences in the following ways:

1. Superfamily sequence files have any sequence shorter than 30 residues removed rather than the limit of 20 in ASTRAL. Domains which are split across more than one chain had separate entries in ASTRAL, which had to be joined to make a single entry in Superfamily[11].

2. A small number of documented ASTRAL errors, which are considered significant, were corrected manually by Superfamily.

3. Some errors in domain definitions in the SCOP classification were detected and corrected in the Superfamily sequence files[12].

4. Sequences which are merely redundant shorter parts of other sequences are removed when filtering on sequence identity.

The methods Superfamily used identified many more superfamily classifications than SCOP, but there were problems with their classifications, which is explained later in Chapter 7.

[10]PDB is a database of three-dimensional structures of biological macromolecules (proteins, ribosomes, etc.). This database archives atomic coordinates determined by x-ray crystallography and NMR for each macromolecule along with experimental details, secondary structure, cofactors and author.

[11]Subsequent releases of ASTRAL however now include these joined domains.

[12]These errors have been corrected in a subsequent releases of SCOP.

The Superfamily model library is available from a public web server at http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/.

2.3 Comparative Genomics and Proteomics

Many researchers are interested in conserved homologues and the clustering of genes within these homologues. In this case researchers have used sequence comparison methods (Sankoff & Kruskal, 1983). This is not strictly cluster detection since the comparisons are made with clusters that have already been identified, usually by qualitative assessment or classical statistical methods.

2.3.1 Sequence Alignment

Essentially, there are two principal methodologies currently used for performing sequence alignment: methods based on the dynamic programming algorithm and methods based on probabilistic profiles. Although profile based methods are considered to be better for detecting remote homologies (Gough *et al.* , 2001), dynamic programming methods are by far the most popular. Present day pairwise methods such as ClustalW and BLAST are all based on optimizations and modifications of the dynamic programming algorithm.

2.3.2 The Dynamic Programming Algorithm

Elementary operations

An elementary operation is the simplest algorithmic operation in sequence comparison that can be performed resulting in a point differentiation of one sequence. Those operations are:-

1. Substitutions or replacements.
2. Deletions and insertions, which are collectively known as indels.
3. Compressions and expansions.
4. Transpositions or swaps.

Note that the second and third items in the list above are so similar that differences are often overlooked. Dealing with these four operations is the central theme of sequence comparison (Sankoff & Kruskal, 1983).

There are three main types of sequence comparison:-

1. Trace
2. Alignment or matching
3. Listing

A full description and an explanation of the implementation of all three types are given in a book by Sankoff and Kruskal (Sankoff & Kruskal, 1983). Briefly, alignment is the method most suited for fast protein/DNA sequence alignment algorithms as it applies elementary operations in order. Trace is essentially the same, but without distinctions in order and is more suited to text error correction. Interestingly, listings are said to correspond directly to the natural mechanisms by which sequences are believed to change (Sankoff & Kruskal, 1983).

As a practical mode of presentation, listings are awkward and have been little used. However, as a mode of analysis, listings have a theoretical importance, because it is possible to generalize them much more broadly than alignments and traces and because they correspond to the plausible underlying mechanisms in several major applications.

- Listings permit many successive changes compared to alignment, which allows only one.
- Listings can make distinctions based on the order in which the changes are made.
- Listings with more than one change in a position are not selected, in accordance with the parsimony principle[13]. This may scupper the detection of evolutionary changes (Sankoff & Kruskal, 1983).
- Listing order of substitutions can be many and varied to reach the same final

[13]Least number of changes

outcome so care is required when using many substitutions.

Levenshtein distance

Levenshtein distance is defined as the smallest number of elementary operations required to change the sequence under investigation to the sequence to which it is being compared. The two main definitions are:

1. Smallest number of substitutions and indels required to change A into B.
2. Smallest number of indels required to change A into B. No substitution permitted.

There are many more definitions for Levenshtein distance in sequence comparison.

The definition of Levenshtein distance is expressed more formally by taking two sequences:
Sequence **a** of the form

$$a_1, a_2, \cdots a_m \tag{2.1}$$

Sequence **b** of the form

$$b_1, b_2, \cdots b_n \tag{2.2}$$

whose terms belong to a given metric space:

$$\{-, a_1, a_2, \cdots, b_1, b_2, \cdots\} \tag{2.3}$$

which includes the null (-) as one of its elements and whose distance function is denoted by b. Note that the given values of $d(a_i, b_j)$,$d(a_i, -)$, etc., are assigned mutation and deletions costs as described below. Let $\bar{a}$ denote the set of all sequences of the form $\bar{a}_1, \bar{a}_2, \cdots \bar{a}_{m+n}$ formed by inserting n nulls into **a** and let **b** denote the set of all sequences of the form $\bar{b}_1, \bar{b}_2, \cdots \bar{b}_{m+n}$ formed by inserting m nulls into **b**. Then the evolutionary distance $d(a, b)$ is defined by

$$d(\mathbf{a}, \mathbf{b}) = min \sum_{i=1}^{n} d(\bar{a}_i, \bar{b}_i) \tag{2.4}$$

where the minimum is taken over all pairs of sequences: - $\bar{a}_1, \bar{a}_2, \cdots \bar{a}_{m+n}$ and $\bar{b}_1, \bar{b}_2, \cdots \bar{b}_{m+n}$ in $\bar{a}$ and $\bar{b}$ respectively.

Any two sequences of that format that achieve the minimum in the above expression $d(\mathbf{a}, \mathbf{b})$ constitute a 'metric alignment' of **a** and **b**.

The Algorithm

The advantages of dynamic programming: -

- Separation of evaluation from algorithm.
- Use of a simple evaluation score to select homologies.
- Global optimality will find the best homology. Whereas many algorithms do not explicitly link to the evaluation score so do not guarantee to find the best.
- Stable parameters and soft limits.
- Versatility and consistency.

The basic procedure is as follows:

1. Let the elementary operations be just substitutions and indels.
2. Consider all listings from a to b based on elementary operations.
3. Let the length of each listing be the number of elementary operations it contains.
4. Then the distance is the minimum length of any listing.

Another version of this procedure is performed as above, but the elementary operations are reduced to indels only. Alternatively a weighting w can be added so that indels score favourably over substitutions such that the score is given by:

$$Indels + w(Substitutions) \tag{2.5}$$

Where $w >= 2$. If $w > 2$, then it is always shorter for a listing to use an insertion/deletion pair in place of a substitution. Also, if $w = 2$ then it is as short to use an indel pair as it is to use a substitution.

2.3.3 ClustalW

ClustalW is a popular multiple sequence alignment package suitable for use with both DNA and protein sequences (Higgins & Sharp, 1988) (Thompson *et al.* , 1994) (Chenna *et al.* , 2003) (Larkin *et al.* , 2007). The basic information provided by multiple alignments of protein sequences is the identification of conserved sequence regions, which is useful in predicting the function and structure of proteins and in identifying new members of protein families. Clustal was originally written by Des Higgins, and later versions were developed by Julie Thompson, Toby Gibson and François Jeanmougin. Thompson and Jeanmogin maintain the recent versions. Recent features include the ability to detect and read different input formats (NBRF/PIR, Fasta, EMBL/Swissprot), align old alignments, produce phylogenetic trees after alignment (Neighbor Joining trees with a bootstrap option), write different alignment formats (Clustal, NBRF/PIR, GCG, PHYLIP) and the presence of a full command line interface.

The original idea is a 'quick and dirty' version of the Feng-Doolittle algorithm (Feng & Doolittle, 1987) where the assumption is that two sequences with a minimum Levenshtein distance are most likely to have been obtained from organisms that have most recently diverged. Pairwise alignment provides the most reliable information and so any spaces should be preserved in an overall multiple alignment.

ClustalW not only performs multiple alignments, it can also produce true phylogenetic trees in one of three output formats; NJ, Phylip and Dist. ClustalW is made available on web servers by the Genebee web server at the Belozersky Institute in Moscow, and at the European Bioinformatics Institute.

2.3.4 BLAST: Basic Local Alignment Search Tool

Protein sequence and DNA sequence databases became overwhelming large in the early 1980s and at that time there was a recognised requirement for efficient search algorithms for the global comparison of sequences in large databases that are similar to a given sequence.

Originally, global comparison methods, such as those by Fitch (Fitch, 1966), Dayhoff (Dayhoff, 1979), Needleman and Wunsch (Needleman & Wunsch, 1970), Sellars (Sellars, 1974), Smith *et al.* (Smith *et al.* , 1981) and Sankoff (Sankoff, 1972) aligned complete sequences. Although these methods were relatively accurate, they were computationally expensive, requiring computer time in the order of $N \times M$, where N and M and the lengths of the sequences compared. Wilbur and Lipman (Wilbur & Lipman, 1983) presented an algorithm for global comparison of sequences based on matching k-tuples of sequence elements. Computational time required is still in the order of $N \times M$, but the lengths of M and N are now very much reduced. Local search methods using fragments of two sequences had already been proposed at that time, such as those by Korn *et al.* (Korn *et al.* , 1977), Sellars (Sellars, 2000), Smith and Waterman (Smith & Waterman, 1981) and Goad and Kanehisa (Goad & Kanehisa, 1982).

BLAST (Altschul *et al.* , 1990) (Altschul *et al.* , 1997) is an extension of the k-tuple method of Wilbur and Lipman, using a heuristic that attempts to optimize a specific similarity measure. There are earlier heuristic methods by Pearson and Lipman (Pearson & Lipman, 1988), and Altschul and Gish (Altschul & Gish, 1996). BLAST is more suited to efficiently compare given sequences with large databases of sequences than ClustalW.

Expectation Values (E-values)

The expectation value (E-value) is a statistical measure of the significance of a database sequence match. The probability of getting a match by chance depends on the size of the database. This means that the E-value given for a pairwise match will vary between different databases. For this reason the E-value given by

BLAST, as a measure of similarity between sequences, should also be combined with the *bit score* for a true representation of sequence similarity (Xiong, 2008). In general, however, the E-value is frequently used on its own where a low E-value score indicates a high similarity to the model whereas a higher score is a lower similarity.

Formally, the E-value is the theoretically expected number of false hits per sequence query. It is calculated from the reverse score using the following formula:

$$E\text{-}value = \frac{library\ size}{1 + e^{-reverse\ score}} \tag{2.6}$$

The *library size* is the size of the model library. The *reverse score* is the simple score of the forward sequence with the simple score of the reverse sequence subtracted from it. See the Sequence Alignment and Modeling System (SAM) website[14] for further documentation on scoring.

Therefore, if the E-value is 1, it is likely that you would get one chance hit with this score to the query using the particular database that was searched.

The following general conclusions can be drawn from the E-values [15]:

- If the E-value is less than 1×10^{-50} ,the hit is very similar to the query sequence and is very likely to be evolutionarily related.
- If the E-value is between 1×10^{-50} and 1×10^{-2}, the hit has some similarity to the query sequence and may be related. When E-values in this range are obtained these values can indicate that the sample sequence is in the same family as the hit or it may have closely related functional domains. If the top hits (those sequences with the lowest E-values) all seem to be related, this makes it more likely that the query is of the same family/type.
- If the E-value is between 1×10^{-2} and 1, the hit has a slight possibility of being related to the query. This may indicate a distant evolutionary

[14]SAM website:
http://www.soe.ucsc.edu/research/compbio/sam.html.

[15]These definitions taken from http://www.swbic.org/origin/proc_man/Blast/BLAST_tutorial.html

relationship.

- If the E-value is above 1, the hit is not very closely related to any sequence in the database. This conclusion can also be made when no matches are found at all.

We found cases in example searches where non-zero E-values were exact matches. One example gave a top hit with an E-value of 1×10^{-132} which turned out to be an exact match to the query sequence. However, it should be further noted that sometimes exact and closely matched hits will have E-values of 0.

If the query sequence is short (less than 100 nucleotides or amino acids long), the top E-values may be larger than 1×10^{-50} even if there is an exact match. It is necessary to check the % identity of the top hits, not just the E-values. Hits with low E-values that only have similarity to short regions of the query sequence are more likely to indicate that the sequences have motif or functional domain similarities rather than that they represent related genes or proteins. This is very likely the case when all of the matches are from sequences whose names and descriptions do not seem to indicate that the hits are in any way related to each other. Hits with higher E-values, in the ranges of 1×10^{-50} to 1×10^{-5}, may still indicate that the query and hits are related if the hit has at least a 35% identity with the query over at least 80% of its length. Another indication of this is if several hits have names and/or descriptions that indicate that they are related to each other.

2.3.5 Profile Based Alignment

Profile based methods are said to perform with greater selectivity than pairwise methods and profile based methods using hidden Markov models are the most effective (Gough *et al.* , 2001). Recently hidden Markov models have been used for the detection of distant homologs to find multi domain proteins (Ekman *et al.* , 2005). One such proprietary system is a profile comparer called PRC 1.3.1[16], which is a stand-alone program for aligning and scoring two profile hidden Markov

[16]The profile comparer is downloadable from http://supfam.mrc-lmb.cam.ac.uk/PRC/.

models. This can be used to detect remote relationships between profiles more effectively than by doing simple profile-sequence comparisons.

Markov Chains

Given a set of events $\{S_0, S_1, ...S_{m-1}\}$ and a system for which event S_i follows event S_j with known probabilities $p(i, j)$, the system can be represented by the $m * m$ matrix of $p(i, j)$ values. This is known as a Markov chain and it can be used to work out the probability of a system being in a particular state at a particular time. If the probability of S_i at time t is $P_i(t)$, then the probability of being in state j at time $t + 1$ is

$$P_j(t + 1) = \sum P_i(t)p(i, j) \tag{2.7}$$

(Arthurs, 1965) (Aleksander & Morton, 1995).

If each amino acid is considered to be an event with a certain probability of occurrence (see Table 2.5) on the gene, it can be clearly realized how Markov chains can be used for effective sequence comparison of genes and the same technique can be applied to genes on the chromosome to detect homologous or paralogous sequences. However, it should be noted that the probability of the occurence of genes on the chromosome is vastly more difficult to calculate than that of amino acids.

Amino Acid	S	Prob.
alanine	A	0.0625
cysteine	C	0.03125
aspartic acid	D	0.03125
glutamic acid	E	0.03125
phenylalanine	F	0.03125
glycine	G	0.0625
histidine	H	0.03125
isoleucine	I	0.046875
lysine	K	0.03125
leucine	L	0.09375
methionine	M	0.015625
asparagine	N	0.03125
proline	P	0.0625
glutamine	Q	0.03125
arginine	R	0.09375
serine	S	0.09375
threonine	T	0.0625
valine	V	0.0625
tryptophan	W	0.015625
tyrosine	Y	0.03125
STOP		0.046875

Table 2.5: A list of all amino acids (residues) and probabilities. Where **S** signifies a designatory letter used in the text representation of protein sequences and **Prob.** is the probability of each of the amino acids occurring from a random selection of three nucleotides that constitute a codon.

2.4 Logic Programming

The basic idea of programming in logic was first presented by Robinson (Robinson, 1965). The key idea behind logic programming is that computation can be expressed as controlled deduction from declarative statements.

2.4.1 Prolog

Prolog is a programming language used extensively throughout this research. It is a declarative language in that it specifies *what* computation is to be done, unlike imperative languages, for example C++ or Java, that specify *how* a computation is to be done. Prolog was first devised by the Artificial Intelligence Group at Luminy, Marseille under Alain Colmerauer in 1972 and independently by Clocksin and Mellish in Edinburgh. It was the first practical embodiment of the concept of logic programming, which is due to Robert Kowalski (Ait-Kaci, 1991).

The motivation behind Prolog is summed up nicely in the following quote from Alain Colmerauer:

> "...Prolog was a response to a challenge of creating an extremely high level language and a response which, paradoxically, was seen as inefficient in computer science terms...The challenge was, therefore, to be able to write programs very rapidly, leaving the machine to carry out the laborious execution..."
>
> Alain Colmerauer 1986 (Giannesini *et al.* , 1986).

Prolog is a programming language designed for representing and making use of knowledge about a particular domain. More precisely, the domain is a set of objects and the knowledge is formalized by a set of relationships that describe simultaneously the properties of these objects and their interactions. The set of rules describing these objects and relations constitutes a Prolog program.

Prolog's syntax is that of first-order predicate logic formulas written in clause form, which is a conjunctive normal form in which quantifiers are not explicitly

written. It is further restricted to Horn clauses only, which are clauses that have at most one positive literal. In the early 80s, David H. D. Warren designed an abstract machine for execution of Prolog. It became known as the Warren Abstract Machine (Warren, 1983) and has become the *de facto* standard for implementing Prolog compilers. There is a good, in depth book on this by Ait-Kaci (Ait-Kaci, 1991).

The procedural meaning of Prolog is based on the *resolution principle* for mechanical theorem proving called SLD (Selection, Linear, Definite)(Robinson, 1965), but with some short comings. Prolog does not truly perform what is called *unification* in logic, instead it uses what is known as *matching*. However, for efficiency, most Prolog systems implement matching in a way that is perfectly adequate in practice (Bratko, 2001). For this reason, Prolog can be seen as a first step towards (but not actually) the ultimate goal of *programming in logic* (Clocksin & Mellish, 1984).

2.5 Frequent Pattern Mining

The purpose of data mining is to find valid, potentially useful and ultimately understandable patterns in data (Fayyad *et al.* , 1996). A pattern in this sense is a discernible arrangement or sequence found in comparable objects or events. Patterns are indicative of order or trend; a reduction of entropy; a suggestion of purpose or the result of intelligent intervention or design.

Discovered patterns are evaluated in accordance with the user's interest in those patterns. This is accomplished through either a user-driven or *subjective* approach, which is often very inefficient, or a data-driven or *objective* approach. The subjective approach is complicated and difficult to automate. Of the objective approaches, pattern frequency is by far the most popular.

There is a clear motivation for selecting the most interesting rules/patterns. The main challenge is to discover novel, useful patterns, going beyond accuracy and comprehensibility. User-driven and data-driven approaches have complementary advantages and disadvantages. Using a hybrid approach is a possible sensible solution and one such system using this approach is described in Chapter 6.

The frequency in the occurrence of patterns can be used as a measure of significance with regard to the purpose or meaning of the pattern. Frequent and, more importantly, significant patterns will often reveal information and this is one of the principal paradigms for knowledge discovery in databases. This is also considered in the system described in Chapter 6.

The basic methodology behind frequent pattern mining is to iteratively generate the set of candidate patterns of length $(k+1)$ from the set of frequent patterns of length k (for $k \geq 1$), and check their corresponding frequencies of occurrence in the database, pruning off infrequent branches of the search space.

The frequent pattern mining methodology has been applied to many areas in knowledge discovery in databases such as: association rules (Agrawal & Srikant, 1994) (Klemettinen *et al.* , 1994) (discussed below), correlations (Brin *et al.* , 1997), causality (Silverstein *et al.* , 1998), sequential patterns (Agrawal & Srikant,

1995), episodes (Mannila *et al.* , 1997), multi dimensional patterns (Lent *et al.* , 1997) (Kamber *et al.* , 1997), max-patterns (Bayardo, 1998), partial periodicity (Han *et al.* , 1999) and emerging patterns (Dong & Li, 1999).

The main problem with pattern mining is the huge number of potential candidates and the consequent computational expense. This is discussed in more detail in the following section using association rule mining as an example.

2.5.1 Apriori: an Association Rule Mining Algorithm

Association rule mining programs discover associations in data based on the premise that items occurring together frequently are in some way associated. Association rule mining is a well established field and several surveys of common algorithms exist (Hipp & Nakhaeizadeh, 2000) (Mueller, 1995). The Apriori Algorithm (see below) is a popular and relatively efficient association rule mining algorithm (Agrawal & Srikant, 1994).

Many previous studies by researchers such as Agrawal and Srikant (Agrawal & Srikant, 1994), Mannila (Mannila *et al.* , 1994), Lent (Lent *et al.* , 1997), Srikant (Srikant *et al.* , 1997), Ng (Ng *et al.* , 1998) and Grahne (Grahne *et al.* , 2000), have adopted an Apriori-like approach.

The Apriori algorithm achieves good performance gained by significantly reducing the size of candidate sets. However, a huge number of candidate sets are generated in situations with a large number of frequent patterns, long patterns, or quite low minimum support thresholds. This is computationally expensive. For example, if there are 10^4 frequent 1-itemsets, the Apriori algorithm will need to generate more than 10^7 length-2 candidates and accumulate and test their frequency of occurrence. Moreover, to discover a frequent pattern of size 100, such as $\{a1, \cdots, a100\}$, it must generate $2^{100} - 2$ or about 10^{30} candidates in total. Generating that many patterns on a computer capable of billions of instructions per second will take billions of years. This is the inherent cost of candidate generation, no matter what implementation technique is applied.

Terminology

The terms described use market basket analysis as an example, but also provided are examples from pattern mining in genetic information such as the research covered in Chapter 6.

Item A single datum such as an article in a basket or a single gene attribute.

Transaction All the articles in one basket or all attributes pertaining to a single gene (one dimensional array).

Transaction data set All data such as all articles in all baskets or all genes and all their respective attributes (two dimensional array).

Itemset A non-specific set of items or gene attributes.

k **itemset** A non-specific set of k items or k gene attributes.

In any transactional data set D (all data) the number of times that an itemset occurs is the count (frequency). [17]

Support (S) for itemset X in D is defined:-

$$support(x) = \frac{freq(X)}{|D|} \tag{2.8}$$

For association rule $X \rightarrow Y$:

$$support(X \rightarrow Y) = support(XY) = support(X \cup Y) \tag{2.9}$$

The confidence (C) in any association rule is given by:-

$$confidence(X \rightarrow Y) = \frac{support(X \cup Y)}{support(X)} \tag{2.10}$$

[17] An association rule about a relationship between two disjoint itemsets X and Y, is denoted by $X \rightarrow Y$. It represents the relation that, when X occurs, Y also occurs. $X \rightarrow Y$ does not mean X causes Y. $X \rightarrow Y$ can imply a different meaning than $Y \rightarrow X$

Frequent itemsets are those whose support is greater than the minimum support ($minS$).

Interesting association rules are those whose support and confidence are greater than the minimum support and minimum confidence ($minC$).

The Apriori downward closure property implies that any subsets of a frequent itemset are also frequent itemsets. This is also popularly known as the Apriori heuristic (Agrawal & Srikant, 1994): *if any length k pattern is not frequent in the database, its length (k + 1) super-pattern can never be frequent.*

There are two major steps in association rule mining/pattern mining:-

1 Frequent itemset generation

2 Rule derivation

Frequent itemset generation

The algorithm is as follows where C represents candidate itemsets; L represents frequent itemsets and k is the number of items or attributes in each itemset:

1 $k = 1$;

2 Find frequent itemset L_k from C_k, the set of all candidate itemsets;

3 Form C_{k+1} from L_k;

4 $k = k + 1$;

5 Repeat 2 - 4 until C_k is empty

Step **2** is the frequent itemset search stage. This is achieved by scanning D and counting each itemset C_k. If the count is greater than $minS$, then add that itemset to L_k.

Step **3** is the candidate itemset generation stage and follows this procedure:-

For $k = 1$, C_1 = all itemsets of length 1

For $k > 1$, generate C_k from L_{k-1} as follows:-

The join step

1 $C_k = k - 2$ way join of L_{k-1} with itself

2 If both $(a_1, ..., a_{k-2}, a_{k-1})$ and $(a_1, ..., a_{k-2}, a_k)$ are in L_{k-1}, then add $(a_1, ..., a_{k-2}, a_{k-1}, a_k)$ to C_k

3 Items are always stored in the sorted order

Prune step

1 Remove $(a_1, ..., a_{k-2}, a_{k-1}, a_k)$ if it contains a non frequent $(k-1)$ subset

Rule derivation

So far the algorithm has discovered frequent itemsets, but frequent itemsets do not mean association rules. Association rules can be found from every frequent itemset as follows:

For every non-empty subset X of D:

1 Let $Y = D - X$

2 $X \rightarrow Y$ is an association rule if the confidence of $(X \rightarrow Y) \geq minC$

where X is an itemset L_k and Y is an itemset L_{k+1}.

The Apriori algorithm lends itself well to the market basket analysis example since the information about the items in the basket is all that is required: the basket itself has no importance. The gene attribute example shows that we may find patterns in the attributes of genes, but no information about the genes themselves is available. Market basket analysis is an example of a 2-dimensional database, which can be represented by a single table and the Apriori algorithm works well with this type of data. The gene data used in Chapter 6 is multi-dimensional and is represented by many tables. This type of data requires first order pattern mining techniques.

2.5.2 First Order Pattern Mining

Candidate patterns in Apriori are effectively patterns of propositional statements or rules, which do not contain variables, whereas first order statements are propositional statements that do contain variables. The motivation for extending Apriori-like association rule mining to first order association rule mining is that first order statements are much more expressive (Mitchell, 1997).

In this research frequent pattern mining was performed using Prolog and a Prolog based frequent pattern mining program named WARMR.

WARMR

WARMR (Dehaspe *et al.* , 1998) is an ILP data mining program used to identify frequent patterns, where these patterns are represented as conjunctive queries in Datalog schema. It is a general purpose data mining algorithm suitable for discovering knowledge in structured data, where patterns reflect the one-to-many and many-to-many relationships of multiple tables. This is not possible with standard data mining programs (King *et al.* , 2001).

WARMR is an extension of the APRIORI algorithm (Agrawal *et al.* , 1993) (Agrawal *et al.* , 1996) using an efficient level-wise method to mine Association Rules in Multiple Relations (ARMRs) within huge datasets. The WARMR level-wise search algorithm (Mannila *et al.* , 1997) is based on a breadth first search of pattern space. This space is ordered by the generality of patterns. The level-wise method searches this space one level at a time starting with the most general. The method iterates between candidate generation and candidate evaluation phases:

Candidate generation uses the lattice structure for pruning non-frequent patterns from the next level.

Candidate evaluation computes the frequencies of the candidates with respect to the database.

Pruning is based on monotonicity of generality with respect to frequency. In other words, if a pattern is not frequent then none of its specializations is frequent. So while generating candidates for the next level, all the patterns that are specializations of infrequent patterns can be pruned.

The nature and structure of the candidate patterns are determined by the user through the *language bias*. WARMR has the advantage that background knowledge can be easily incorporated to refine searches and include previously discovered patterns to enable the discovery of increasing complex patterns.

WARMR was used in the research discussed in Chapter 6 for pattern mining in gene location. Pattern mining in the location of genes may reveal patterns in the molecular function which could assist in determining the function of unknown genes occurring in similar locational patterns.

2.6 Phylogeny

2.6.1 Introduction

Phylogeny is the study of evolutionary relationships among organisms. Traditionally the methods used to determine phylogeny are: morphology, anatomy, physiology and palaeontology. Nowadays molecular phylogeny is used exclusively in the study of evolutionary relationships.

The phylogeny of living organisms is best represented by a phylogenetic tree, which is a graphical representation of the evolutionary relationships between groups of organisms and is discussed later in Section 2.6.2.

Both taxonomy and phylogeny are referred to in the research described later in this thesis. Taxonomy is a systematic classification of living organisms, whereas phylogeny is a theoretical model of the sequence of evolutionary divergence of organisms from their common ancestors. However, the structures of a phylogenetic tree and a taxonomic tree are identical.

Linnaean Taxonomy

Swedish born Carl Linnaeus (13/05/1707 – 10/1/1778) established the first taxonomy of organisms, which is a method of classifying living things. He also invented the naming convention known as *Binomial nomenclature*, which is still in use. Organism names are given by their species name (genus or generic), followed by a specific name (specific descriptor or specific epithet). For example the binomial name for the Bottlenose Dolphin is *Tursiops truncatus*.

Molecular phylogeny

Molecular phylogeny specifically determines evolutionary relationships from nucleotide sequences and/or protein sequence data.

One of the first noteable applications of molecular phylogeny was performed by Carl Woese. He redrew the taxonomic tree of life in 1977 by introducing a new domain named *archaea*. By phylogenetic taxonomy of 16S ribosomal RNA, a technique pioneered by Woese, he showed that archaea are neither bacteria nor eukaryotes (Woese & Fox, 1977). Hitherto, archaea were considered to be bacteria, but although they are prokaryotes, they are as different from bacteria as they are different from eukaryota (Woese *et al.* , 1978). His three-domain system, based upon genetic relationships rather than obvious morphological similarities, divided life into 23 main divisions, all incorporated within three domains: bacteria, archaea, and eukaryota (Woese *et al.* , 1990).

2.6.2 Phylogenetic Trees

A phylogenetic tree is a graph of evolutionary relationships. It is composed of nodes and branches where only one branch connects any two adjacent nodes. Nodes represent the taxonomic units such as species, populations, individuals or genes. Branches define the ancestral relationships where their length may be proportional to the difference between nodes. The branching pattern is known as the topology. External nodes represent extant taxonomic units and are specifically referred to as operational taxonomic units (OTUs). Internal nodes represent ancestral units. A tree is additive if the distance between any two OTUs is equal to the sum of the lengths of all branches connecting them. A node is bifurcating if it has only two immediate descendant lineages, multifurcating if more than two. A clade is defined as a group of species that have a unique common ancestor, which is not shared by any other species.

There is a good background on the development of genetic distance measure and phylogenetic tree methods by Fitch (Fitch & Margoliash, 1967), Nei (Nei, 1975) and Felsenstein (Felsenstein, 1988) (Felsenstein, 2004).

Phylogenetic tree construction used in this research was performed using ClustalW, which has two methods: *Neighbor Joining* (Saitou & Nei, 1987), which repeatedly joins the "nearest neighbours" to build a tree; and *UPGMA* (Unweighted Pair

Group Method with Arithmetic Mean) (Sneath & Sokal, 1973), which clusters close taxa, assuming the rate of evolution is the same across lineages. The Neighbor Joining method is more accurate, but UPGMA is faster.

2.6.3 Broad Sampling and Consensus Trees

The accuracy of molecular phylogeny is sequence specific. Gene sequences from essential genes are not likely to vary much from species to species. Genes that do vary may not necessarily vary at a rate proportional to the rate of evolution of the host species. Of course, these same two problems also apply to proteins. Mitochondrial DNA sequences have been used to generate phylogenies in the past because they are less stressed by natural selection and can more freely acquire mutations roughly in proportion with time. However, we are more interested in the specific phylogeny of phenotypes and mitochondrial DNA plays an insignificant part in the development of the phenotype.

Character evolution, which is an expression used to refer specifically to the evolution of the phenotype, can be determined with a higher resolution by using *broad phylogenomic sampling* (Dunn *et al.* , 2008). This method uses a consensus of results from multiple phylogenies of organisms based on orthologous protein or gene sequences. A *consensus tree* represents the phylogeny of organisms determined by this method (Bryant, 2003). The procedure of broad phylogenomic sampling and corresponding consensus tree building is the central theme of Chapter 8.

2.6.4 Phylogenetic Databases

In general, web based phylogenetic databases can be used as a source of information on phylogenetic relationships through access to published phylogenetic studies and the corresponding data and trees that they contain. The information they contain is usually as follows:

- Information on the phylogeny of particular groups of interest.

- Datasets for studies of character evolution, including general patterns across many groups.
- Trees with representatives in particular geographical areas.
- Information on host and parasite phylogenies.
- Molecular and morphological phylogenies for particular groups.
- Sources and references for the data.

Three such databases were used in this research: TreeBASE; the NCBI taxonomy browser and the Tree of Life Web Project. All three should only be used as a guide and not an authoritative taxonomy without further reference to the individual contributors to these databases.

TreeBASE

TreeBASE is a relational database designed to manage and explore information on phylogenetic relationships (Piel *et al.* , 2002). Essentially, it is a store for published phylogenetic trees and data matrices. It also includes "...bibliographic information on phylogenetic studies, and some details on taxa, characters, algorithms used, and analyses performed."

> "The database is designed to allow retrieval and recombination of trees and data from different studies, and it can be explored interactively using trees included in the database. The primary data objects in TreeBASE are bibliographic references to published phylogenetic studies, taxon by character data matrices, and phylogenetic trees resulting from the analysis of such data matrices. Information is also available that links data matrices and trees, including types of analyses performed, software used, etc. The TreeBASE web site allows searches to be performed in terms of taxonomic names or bibliographic keywords. Data matrices can also be downloaded from TreeBASE in nexus file format for further analysis"

	higher taxa	genus	species	lower taxa	total
Archaea	88	105	496	99	788
Bacteria	984	1836	13665	4589	21074
Eukaryota	15011	44147	163594	12578	235330
Fungi	1092	3279	18424	1008	23803
Metazoa	10944	26474	70726	6289	114433
Viridiplantae	1850	12435	68252	4658	87195
Viruses	440	286	4755	31762	37243
All taxa	16544	46381	187077	49058	299060

Table 2.6: NCBI taxonomy statistics (August 2008).

NCBI taxonomy browser

The NCBI taxonomy browser is part of a large depository of biological data hosted by the National Center for Biotechnology Information (Wheeler *et al.* , 2000) (Benson *et al.* , 2000). Statistics on the content of this taxonomy are given in Table 2.6. The topology of this database and supporting files are downloadable from the NCBI ftp site (ftp://ftp.ncbi.nih.gov/pub/taxonomy/). The NCBI state that the NCBI taxonomy database is not a phylogenetic or taxonomic authority.

Tree of Life web project

The Tree of Life Web Project is a website (http://tolweb.org/tree/), which is the result of a collaborative effort of biologists worldwide and contains well over 9000 World Wide Web pages to date(Maddison *et al.* , 2007). The project is a work-in-progress and provides information about the diversity of organisms, their evolutionary history (phylogeny), and characteristics. The site is described thus:

> "The Tree of Life Web Project is a collection of information about biodiversity compiled collaboratively by hundreds of expert and amateur contributors. Its goal is to contain a page with pictures, text, and other information for every species and for each group of organisms, living or extinct. Connections between Tree of Life web pages follow

> phylogenetic branching patterns between groups of organisms, so visitors can browse the hierarchy of life and learn about phylogeny and evolution as well as the characteristics of individual groups."

The Tree of Life host website provides hyperlinks to independently hosted web pages and so the information content is entirely the responsibility of the individual contributors. This is important when considering the authority and the citation of any information obtained from the Tree of Life web project.

2.6.5 New Hampshire/Newick format

Previously called the New Hampshire Format, the Newick Standard is a method to describe trees by parentheses and commas. It describes trees by combining OTUs (operating taxonomic units) by nested parentheses. The branch lengths are written in numerals after OTU names followed by colons (:). The outermost parenthesis usually contains three elements indicating an unrooted tree. If the outermost parenthesis has only two elements then this indicates a rooted tree. A semicolon (;) is needed after the outermost parenthesis. Below is an example.

((Human:0.3, Chimpanzee:0.2):0.1, Gorilla:0.3, (Mouse:0.6, Rat:0.5):0.2);

This style is often used by many programs for sequence data analyses and phylogenetic tree representation. The details of this style is explained in a document attached to PHYLIP (ver. 3.572) developed by Felsenstein and colleagues.

Chapter 3

Model Organisms

Model organisms are intensively studied and consequently there is a considerable amount of background knowledge. They are chosen based largely on pragmatic reasons such as the ease with which they can be studied or their importance to commerce or medicine. Two model organisms feature in this research: the flowering plant *Arabidopsis thaliana*, which is easy to study having a rapid life cycle, and the yeast *Saccharomyces cerevisiae*, which is important to commerce being used in brewing and bread making industries.

3.1 The Model Organism *Arabidopsis thaliana*

The flowering plant *Arabidopsis thaliana* shown in Figure 3.1 is also known as wall cress, thale cress and mouse-ear cress, is a small weed often found in pavements and borders in the UK. It is a very hardy plant, which can found growing throughout the temperate regions of the world. It is a member of the mustard (Brassicaceae) family, a family which also includes cultivated species such as cabbage and radish. It has a rapid life cycle of about 6 weeks from germination to mature seed, prolific seed production and it is easy to cultivate in restricted space due to its small size (leaves 1-5 cm long). For these reasons *A. thaliana* is an attractive plant for scientists to study. Friedrich Laibach first summarized the potential of *A.*

Figure 3.1: The model organism *Arabidopsis thaliana* being grown in a laboratory. (Photo courtesy of Nicole Hanley Markelz of the Plant Genome Research Outreach Program at Cornell University.)

thaliana as a model organism for genetics in 1943 having studied it for many years previously. In fact, he first published correctly that it had five chromosomes as early as 1907. Since then a wealth of knowledge has accumulated and it has now become a model organism for studies of the cellular and molecular biology of flowering plants. It comes as little surprise that *A. thaliana* was the first complete genome of a plant to be sequenced and is now an important model system for identifying genes and determining their functions.

3.1.1 The *Arabidopsis thaliana* Genome

In 1996 many scientists collectively known as the Arabidopsis Genome Initiative (The Arabidopsis Genome Initiative, 2000) started sequencing the entire genome of *A. thaliana* and the results were published by the year 2000 (Theologis *et al.* , 2000; Lin *et al.* , 1999; Salanoubat *et al.* , 2000; Mayer *et al.* , 1999; Tabata *et al.* , 2000). They sequenced regions covering 115.4 megabases of the 125 megabase genome, extending the sequencing well into the centromeric regions and discovered

that the genome of *A. thaliana* contains 25,498 genes encoding proteins from 11,000 families, which is similar to the functional diversity of *Drosophila melanogaster*, the fruit fly and *Caenorhabditis elegans*, the nematode worm. More recent research reveals that the length of the genome of *A. thaliana* is now thought to be 157 Mbp (Bennett *et al.* , 2003). Further research by the Arabidopsis Genome Initiative revealed that an ancestor genome of *A. thaliana* experienced a whole genome duplication followed by subsequent gene loss and substantial local gene duplications. Roughly 17% of all genes are arranged in tandem arrays comprising 4140 tandem duplicate genes, most of which are in pairs. Altogether, there are 1528 tandem arrays and the two longest arrays have more than 21 adjacent tandemly repeated genes (The Arabidopsis Genome Initiative, 2000).

Research continues on the genome of *A. thaliana* and of note is a major re-annotation of the entire genome in 2005 (Haas *et al.* , 2005). The latest data from many contributors can be found on the TIGR[1] and TAIR[2] websites.

3.1.2 *Arabidopsis thaliana* Genome Statistics

This section details some statistics on chromosome length, gene frquency and density and then describes some preliminary work looking into gene lengths and the lengths of the gaps between genes. These statistics are used later in this research.

The number of genes in each chromosome and the nucleotide lengths of each chromosome are given in Table 3.1. Also in the same table are the number of genes and the length of the entire genome of *A. thaliana*. Note that this data falls more in line with the findings of the Arabidopsis Genome Initiative (AGI), rather than the more recent work of Haas *et al.* (2005). This is mainly due to the availability of reliable data from TIGR at the time of the experiments and that data was from the AGI.

[1] http://www.tigr.org
[2] http://www.arabidopsis.org

Chromosome	Number of genes	Length (bp)
I	6813	30,034,249
II	4181	19,845,587
III	5363	23,773,436
IV	3987	17,790,360
V	6096	26,990,441
All	26440	118,434,073

Table 3.1: Details of gene number and chromosome length in base pairs for the chromosomes of *A. thaliana*.

Classification	Definition
Small	$\leq 400\ bp$
Medium	$\geq 401\ bp$, $\leq 2000\ bp$
Large	$\geq 2001\ bp$

Table 3.2: A definition for gene length classification for *A. thaliana*, which will be used in pattern mining.

Gene lengths

Probability density plots for gene lengths of all five chromosomes are given in Figures 3.2, 3.3, 3.4, 3.5 and 3.6. The average mode gene length over all 5 chromosomes is 1238. The average gene length at the lower 2/3 count (where the frequency is 2/3 of the mode frequency) is 428 and upper 2/3 is 1986. The lower and upper 2/3 counts are simply an arbitrary way of catagorizing the data. For example, if the mode frequency is 75, then we say that all gene lengths from 0 to the gene length that has a frequency of 50 (2/3 of 75) are small. This is the lower 2/3 count. Then all gene lengths that occur more frequently than 50 up to the mode frequency and then falling to the next point in the graph where the frequency is roughly 50 again are considered medium gene lengths. This is the upper 2/3 count. Clearly all gene lengths larger than the largest gene length that has a frequency of 50 are classed as large gene lengths.

We use this information as guide for rough classifications on gene length in Table 3.2, which are used later in frequent pattern mining of gene location.

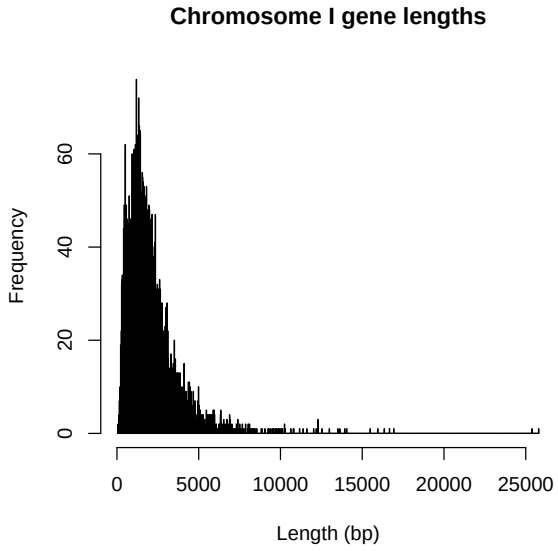

Figure 3.2: Graph of gene lengths for chromosome I of *A. thaliana*. The mode length is 1180 bp and there are 75 examples of genes of this length. There are 50 examples (lower 2/3 of the mode) of genes of length 440 bp, and 50 examples (upper 2/3 of the mode) of genes of length 1820 bp.

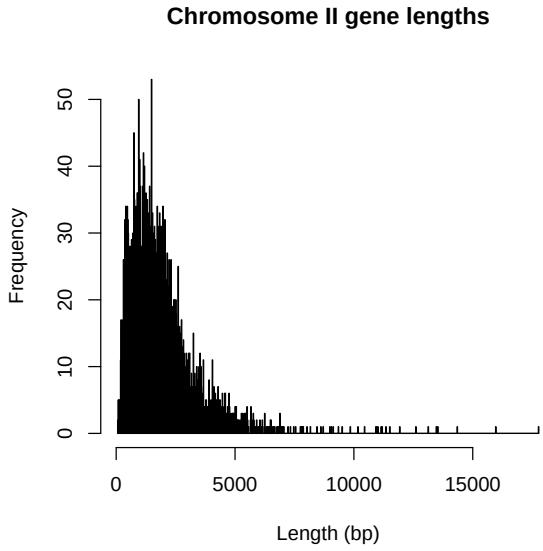

Figure 3.3: Graph of gene lengths for chromosome II of *A. thaliana*. The mode length is 1200 bp and there are 50 examples of genes of this length. There are 33 examples (lower 2/3 of the mode) of genes of length 400 bp, and 33 examples (upper 2/3 of the mode) of genes of length 1980 bp.

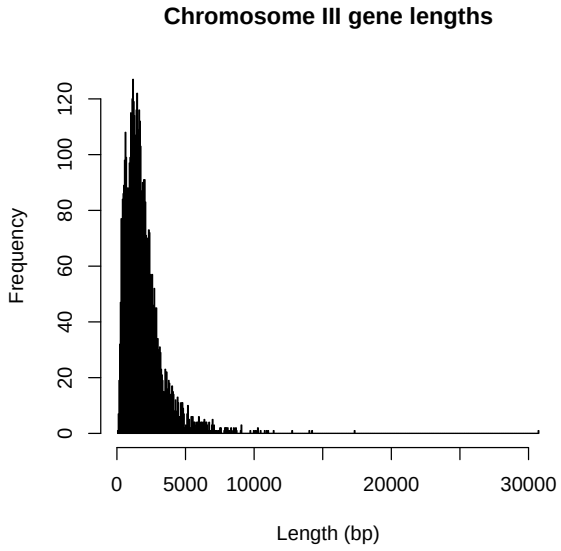

Figure 3.4: Graph of gene lengths for chromosome III of *A. thaliana*. The mode length is 1150 bp and there are 127 examples of genes of this length. There are 84 examples (lower 2/3 of the mode) of genes of length 400 bp, and 84 examples (upper 2/3 of the mode) of genes of length 2050 bp.

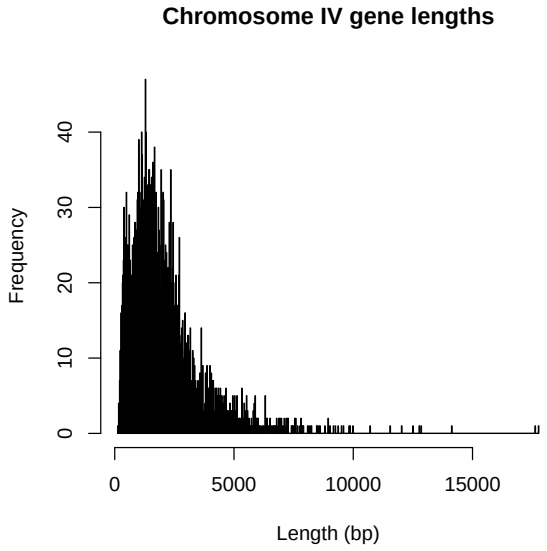

Figure 3.5: Graph of gene lengths for chromosome IV of *A. thaliana*. The mode length is 1280 bp and there are 47 examples of genes of this length. There are 31 examples (lower 2/3 of the mode) of genes of length 480 bp, and 31 examples (upper 2/3 of the mode) of genes of length 2040 bp.

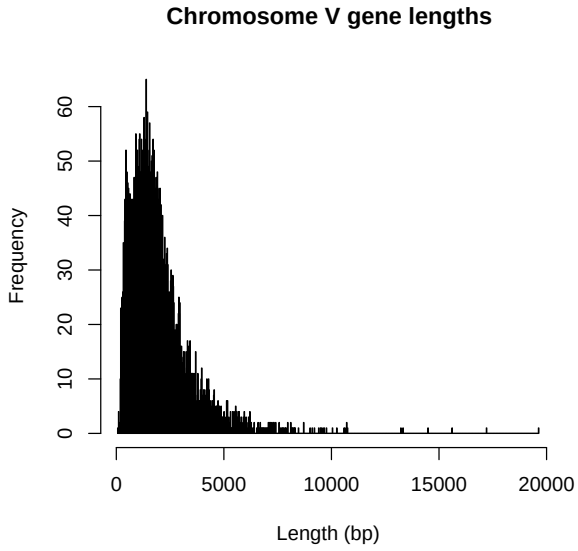

Figure 3.6: Graph of gene lengths for chromosome V of *A. thaliana*. The mode length is 1380 bp and there are 65 examples of genes of this length. There are 43 examples (lower 2/3 of the mode) of genes of length 420 bp, and 43 examples (upper 2/3 of the mode) of genes of length 2040 bp.

Classification	Definition
Small	$\leq 300\ bp$
Medium	$\geq 301\ bp$, $\leq 700\ bp$
Large	$\geq 701\ bp$

Table 3.3: A definition for gene gap length classification for *A. thaliana*, which will be used in frequent pattern mining.

Gap lengths

As part of the foundation for the analysis of the distribution of the locations of all known genes, the probability density plots of the gap lengths between genes for all five chromosomes are given in Figure 3.7. These plots reveal that the most frequently occurring gap lengths are between 300 and 700 base pairs (bp) on all five chromosomes, so for future use, we can use the gap length classifications given in Table 3.3. This classification of gap lengths in used later in Chapter 6.

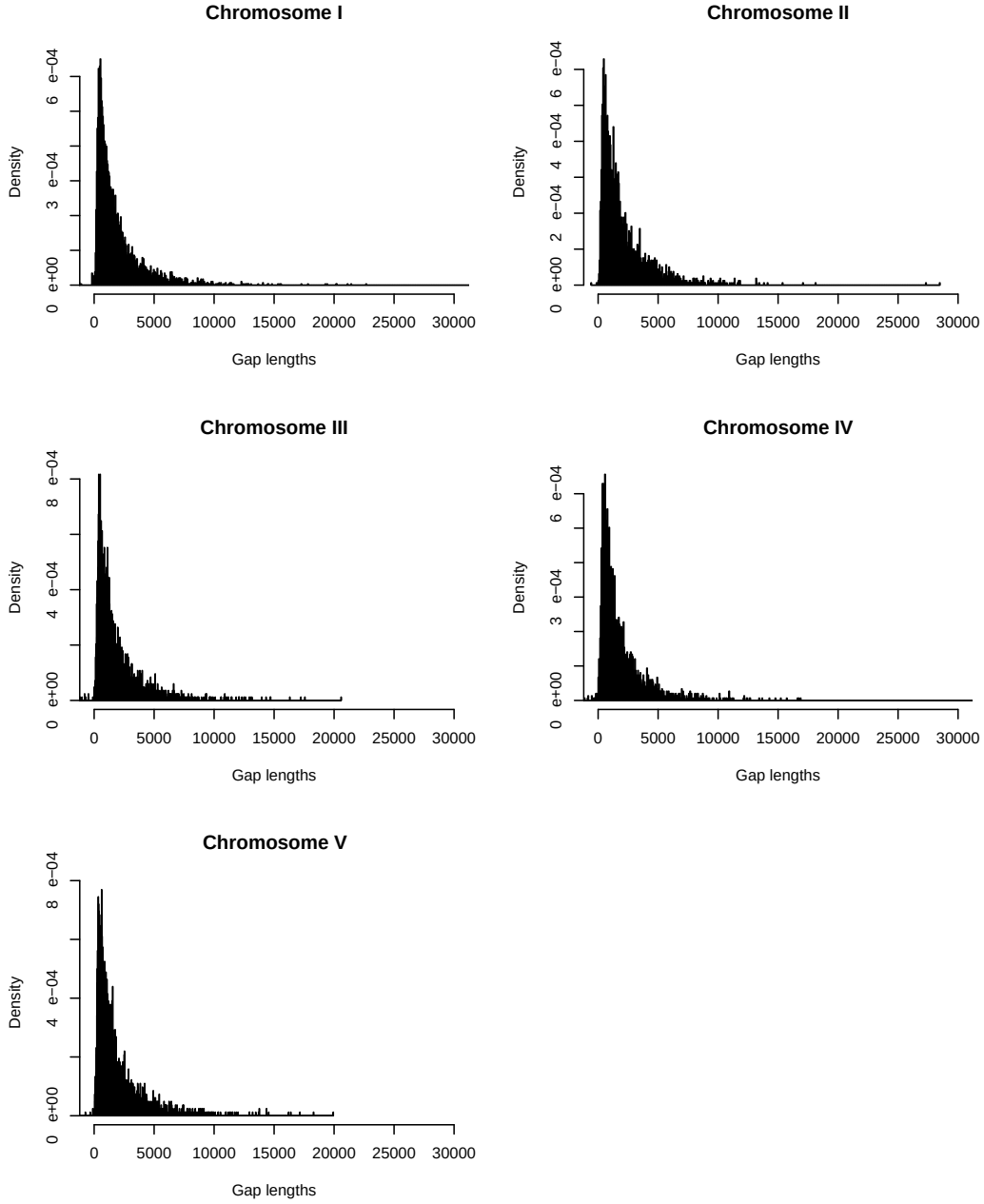

Figure 3.7: Probability density function plots of inter gene gap lengths for all five chromosomes of *A. thaliana*. The curves are not asymptotic to the Y axis; the peak occurs between 300–700 bp.

GO ID	Chr I	Chr II	Chr III	Chr IV	Chr V
go:3774	0.0030	0.0024	0.0039	0.0030	0.0033
go:3824	0.2292	0.2090	0.2172	0.2302	0.2162
go:4871	0.0060	0.0048	0.0054	0.0050	0.0059
go:5198	0.0126	0.0151	0.0188	0.0113	0.0138
go:5215	0.0419	0.0371	0.0362	0.0308	0.0376
go:5488	0.1352	0.1354	0.1372	0.1379	0.1550
unknown	0.5243	0.5508	0.5420	0.5345	0.5250
go:16209	0.0038	0.0045	0.0036	0.0060	0.0039
go:30188	0.0	0.0	0.0	0.0	0.0
go:30234	0.0067	0.0076	0.0071	0.0045	0.0069
go:30528	0.0311	0.0289	0.0252	0.0301	0.0277
go:30533	0.0	0.0	0.0	0.0	0.0
go:31386	0.0	0.0	0.0	0.0	0.0
go:31992	0.0	0.0	0.0	0.0	0.0
go:42056	0.0	0.0	0.0	0.0	0.0
go:45182	0.0054	0.0033	0.0030	0.0038	0.0033
go:45499	0.0	0.0	0.0	0.0	0.0
go:45735	0.0004	0.0009	0.0005	0.0027	0.0013

Table 3.4: Gene Ontology (GO) Data: Molecular function class density for level 1 classes for all five chromosomes of *A. thaliana*.

Gene classification by molecular function

The genes of *A. thaliana* were classified according to molecular function using the Gene Ontology (GO) annotations (The Gene Ontology Consortium, 2000) (see Section 2.1.4). These classes were arranged in levels of increasing specificity. There are 18 subclasses of the molecular function class. These have been designated as the level 1 classes. The subclasses of these level 1 classes were designated as level 2, and so on for levels 3 and 4. The density of genes of each of the level 1 molecular function classes is given in Table 3.4, and similarly, the density and quantity of level 2 classes is are given in Table 3.5. Similar data exists on file for the level 3 and level 4 classes used in this research, but the classes at these levels are too numerous to include in tables in this thesis.

Class	Chr I		Chr II		Chr III		Chr IV		Chr V	
	Freq	Qty	Freq	Qty	Freq	Qty	Freq	Qty	Freq	Qty
go:156	0.00117	8	0.00167	7	0.001118	6	0.001003	4	0.001476	9
go:166	0.01306	89	0.01195	50	0.014171	76	0.015048	60	0.014273	87
go:1871	0	0	0	0	0.000186	1	0	0	0	0
go:3676	0.03317	226	0.03731	156	0.037106	199	0.02784	111	0.042001	256
go:3682	0.00029	2	0.00023	1	0.001118	6	0.00025	1	0.001148	7
go:3700	0.05944	405	0.05381	225	0.047361	254	0.056182	224	0.059064	360
go:3701	0.00014	1	0.00023	1	0	0	0	0	0	0
go:3702	0.00044	3	0.00071	3	0.000932	5	0.000752	3	0	0
go:3711	0	0	0	0	0	0	0	0	0.000164	1
go:3712	0.00117	8	0	0	0	0	0.000501	2	0.000328	2
go:3735	0.01056	72	0.01315	55	0.016035	86	0.009029	36	0.011156	68
go:3777	0.0022	15	0.00167	7	0.003542	19	0.002257	9	0.002625	16
go:4133	0	0	0.00023	1	0	0	0	0	0.000164	1
go:4362	0	0	0	0	0.000372	2	0	0	0	0
go:4386	0.00484	33	0.00382	16	0.004475	24	0.002006	8	0.004593	28
go:4601	0.00278	19	0.00406	17	0.002424	13	0.005768	23	0.003609	22
go:4791	0	0	0.00047	2	0	0	0.00025	1	0	0
go:4857	0.00396	27	0.0055	23	0.004475	24	0.001755	7	0.004429	27
go:4872	0.00117	8	0.00023	1	0.000559	3	0.001003	4	0.000656	4
go:5057	0.00205	14	0.00143	6	0.000745	4	0.001504	6	0.000328	2
go:5102	0.0019	13	0.00071	3	0.000745	4	0.00025	1	0.000492	3
go:5199	0.00014	1	0	0	0	0	0	0	0	0
go:5200	0.00117	8	0.00119	5	0.001305	7	0.00025	1	0.001312	8
go:5201	0	0	0	0	0.000186	1	0	0	0	0
go:5212	0	0	0	0	0	0	0	0	0.000164	1
go:5275	0.00249	17	0.00287	12	0.002424	13	0.000752	3	0.002953	18
go:5319	0.00029	2	0.00047	2	0.000186	1	0.000752	3	0.000492	3
go:5342	0.00014	1	0.00023	1	0.000186	1	0.000501	2	0.000164	1
go:5344	0	0	0.00023	1	0.000186	1	0	0	0	0
go:5372	0.00088	6	0.00215	9	0.001305	7	0.002006	8	0.000656	4
go:5386	0.00557	38	0.00191	8	0.002796	15	0.002006	8	0.001968	12
go:5478	0.00088	6	0.00047	2	0.001678	9	0.000752	3	0.001148	7
go:5496	0.00014	1	0.00047	2	0.000186	1	0.001254	5	0.000656	4
go:5515	0.02832	193	0.03061	128	0.029461	158	0.030599	122	0.031829	194
go:8047	0.00014	1	0	0	0	0	0.00025	1	0	0
go:8135	0.00543	37	0.00334	14	0.002796	15	0.003762	15	0.003281	20
go:8265	0.00029	2	0	0	0	0	0	0	0.000328	2
go:8289	0.00293	20	0.00287	12	0.002983	16	0.007775	31	0.005742	35
go:8430	0.00014	1	0	0	0.000186	1	0.000501	2	0	0
go:8565	0.00366	25	0.00478	20	0.004102	22	0.001755	7	0.003117	19
go:8639	0.00205	14	0.00191	8	0.002051	11	0.001003	4	0.001476	9
go:8641	0.00014	1	0.00047	2	0	0	0.00025	1	0.00082	5
go:8686	0	0	0.00023	1	0	0	0	0	0.000328	2
go:9927	0.00029	2	0.00023	1	0.000372	2	0	0	0.000164	1
go:9975	0.00014	1	0.00023	1	0.001305	7	0.000501	2	0.000328	2
go:15075	0.01247	85	0.01219	51	0.01174	63	0.01103	44	0.01345	82
go:15144	0.00513	35	0.00406	17	0.00335	18	0.00225	9	0.00393	24
go:15197	0.00044	3	0.00023	1	0.00037	2	0.00125	5	0.00114	7
go:15238	0	0	0.00071	3	0.00018	1	0	0	0.00032	2
go:15457	0	0	0.00023	1	0	0	0	0	0	0
go:15646	0.00058	4	0.00095	4	0.00018	1	0.00025	1	0.00049	3
go:15665	0.00014	1	0	0	0	0	0	0	0	0
go:15932	0.00176	12	0.00047	2	0.00018	1	0.002	8	0.00032	2
go:16491	0.03772	257	0.02678	112	0.02778	149	0.03135	125	0.03199	195
go:16563	0	0	0	0	0.00018	1	0.0005	2	0.00016	1
go:16564	0	0	0	0	0.00018	1	0	0	0.00016	1
go:16740	0.08219	560	0.08108	339	0.07458	400	0.08352	333	0.07875	480
go:16787	0.06722	458	0.06433	269	0.07085	380	0.07148	285	0.06529	398
go:16829	0.00851	58	0.00908	38	0.01212	65	0.01003	40	0.00853	52
go:16853	0.00528	36	0.00406	17	0.00652	35	0.00476	19	0.00508	31
go:16874	0.00807	55	0.00693	29	0.00857	46	0.00852	34	0.00689	42
go:16986	0.00088	6	0.00023	1	0	0	0.001	4	0.00016	1
go:17084	0	0	0.00023	1	0.00018	1	0	0	0	0
go:17140	0	0	0.00023	1	0	0	0	0	0.00016	1
go:19207	0.00146	10	0.00119	5	0.00018	1	0.001	4	0.00098	6
go:19208	0.00044	3	0	0	0.00111	6	0	0	0.00065	4
go:19239	0	0	0.00047	2	0	0	0.00225	9	0	0
go:19825	0.00689	47	0.00837	35	0.01174	63	0.00953	38	0.00738	45
go:19842	0.00014	1	0	0	0	0	0.00025	1	0.00016	1
go:30246	0.00499	34	0.00047	2	0.00186	10	0.00175	7	0.00164	10
go:30695	0.00058	4	0.00071	3	0.0013	7	0.00125	5	0.00082	5
go:42277	0	0	0	0	0.00018	1	0	0	0.00016	1
go:42562	0	0	0	0	0	0	0.00025	1	0	0
go:42910	0.00014	1	0	0	0	0	0	0	0	0
go:43021	0	0	0.00023	1	0.00018	1	0	0	0.00016	1
go:43167	0.01203	82	0.011	46	0.01025	55	0.01254	50	0.01476	90
go:43176	0.00073	5	0.00047	2	0.00055	3	0.00075	3	0.00016	1
go:45174	0.00044	3	0	0	0	0	0	0	0.00016	1
go:46790	0.00014	1	0	0	0	0	0	0	0	0
go:46906	0.00102	7	0.00143	6	0.00111	6	0.0005	2	0.00032	2

Table 3.5: Level 2 molecular function classes for *A. thaliana* for all five chromosomes. The relative frequency (Freq) and the number of examples (Qty) are given for each of the five chromosomes.

3.2 The Model Organism *Saccharomyces cerevisiae*

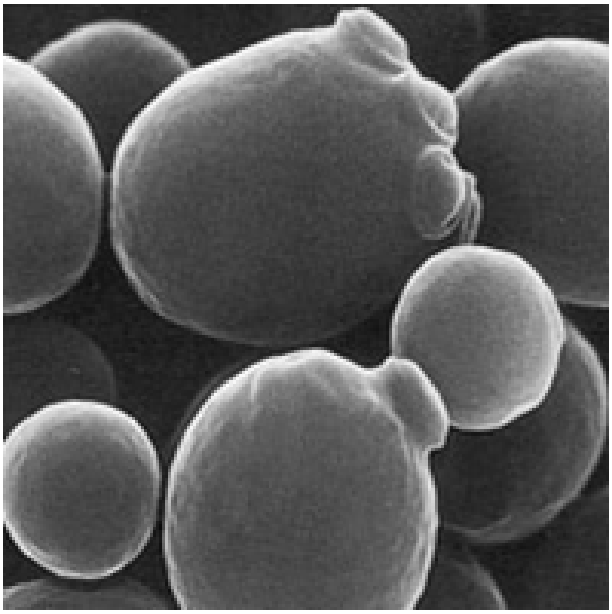

Figure 3.8: Common baker's yeast, *Saccharomyces cerevisiae*.
(Image courtesy Alan Wheals, University of Bath, UK)

The yeast *Saccharomyces cerevisiae* shown in Figure 3.8 is more commonly known as brewer's yeast, baker's yeast, budding yeast and top fermenting yeast. Its binomial name is made up from "saccharomyces", derived from Greek and means "sugar mould", and "cerevisiae", which comes from Latin and means "of beer". It is an ascomycetous fungus forming spores through sexual reproduction inside an ascus, a sac-like structure. It also reproduces by budding, which is a cloning process.

3.2.1 The *Saccharomyces cerevisiae* Genome

The genome of the *S. cerevisiae* has been completely sequenced through a world-wide collaboration (Goffeau *et al.* , 1996). At that time, the discovered sequence of 12,068,000 base pairs (12,068 kilobases) defined 5885 potential protein-encoding genes, approximately 140 genes specifying ribosomal RNA, 40 genes for small nuclear RNA molecules, and 275 transfer RNA genes. This organism is subject to

ongoing research and it is now generally considered that the genome is composed of about 13,000,000 base pairs and 6,275 genes, compactly organised on 16 chromosomes. Only about 5,800 of these are believed to be true functional genes. In addition, the complete sequence provides information about the higher order organization of yeast's 16 chromosomes and allows some insight into their evolutionary history. These figures vary according to the source of data on the genome of *S. cerevisiae*, but the figures presented here serve as a rough guide. Another outcome of the research mentioned above is that the genome shows a considerable amount of apparent genetic redundancy (Goffeau *et al.* , 1996).

Many proteins important in human biology were first discovered by studying their homologs in yeast; these proteins include cell cycle proteins, signalling proteins, and protein-processing enzymes. It is estimated that yeast shares about 23% of its genome with that of humans.

There is a *petite* mutation of *S. cerevisiae* which is particularly interesting. It has little or no mitochondrial DNA and forms small anaerobic colonies when grown on media. The *petite* mutation can be induced by certain mutagens, which have been linked with increased rates in degenerative disease and in aging (Ferguson & von Borstel, 1992).

The Saccharomyces Genome Database (SGD)[3] is a scientific database of the molecular biology and genetics of the yeast *S. cerevisiae*. This database is highly annotated and remains a very important tool for developing basic knowledge about the function and organization of eukaryotic cell genetics and physiology.

Another important database for *S. cerevisiae* is maintained by the Munich Information Center for Protein Sequences (MIPS) [4].

> The MIPS Comprehensive Yeast Genome Database (CYGD) aims to present information on the molecular structure and functional network of the entirely sequenced, well-studied model eukaryote, the budding yeast Saccharomyces cerevisiae. In addition the data of various projects on related yeasts are used for comparative analysis.

[3] http://www.yeastgenome.org/
[4] http://mips.gsf.de/genre/proj/yeast/

Chromosome	Length (bp)	Genes
1	230208	117
2	813178	456
3	316617	183
4	1531917	836
5	576869	324
6	270148	141
7	1090946	584
8	562643	321
9	439885	241
10	745667	398
11	666454	348
12	1078175	578
13	924429	505
14	784333	435
15	1091289	598
16	948062	510

Table 3.6: Details of the genome of *Saccharomyces cerevisiae* giving the base pair length of each of the 16 chromosomes and the number of genes on each chromosome.

3.2.2 *Saccharomyces cerevisiae* Genome Statistics

The data required for the statistics in this section were downloaded from the Saccharomyces Genome Database (SGD) on 21st May 2007. The genome of *Saccharomyces cerevisiae* has 16 chromosomes. Details of the sizes of these chromosomes and the number of genes on each chromosome are given in Table 3.6.

Also available was data on the location of the centromeres, which are shown in Table 3.7. Note that the locations of centromeres on the *A. thaliana* genome were identified using gene frequency plots in Chapter 5. The genome of *S. cerevisiae* is more compact than that of *A. thaliana* and consequently, the centromeres are very difficult to distinguish using gene frequency plots.

Also note an additional autonomous replicating sequence on Chromosome XII from 150946 to 151388 on the Watson strand (CEN12, ARS, ARS1208, ARS1208, CEN12, ARS, ARSXII-151), which is not included in Table 3.7. It is not thought

Chromosome	From	To	Strand
ChrI	151467	151584	w
ChrII	238209	238325	w
ChrIII	114383	114499	w
ChrIV	449708	449818	w
ChrV	151986	152103	w
ChrVI	148505	148622	w
ChrVII	497042	496924	c
ChrVIII	105698	105581	c
ChrIX	355626	355742	w
ChrX	436419	436301	c
ChrXI	439889	439772	c
ChrXII	150946	150827	c
ChrXIII	268031	268149	w
ChrXIV	628760	628877	w
ChrXV	326703	326585	c
ChrXVI	555954	556070	w

Table 3.7: Details of the location of centromeres on all 16 chromosomes of *S. cerevisiae*.

Classification	Definition
Small	$\leq 300\ bp$
Medium-small	$\geq 301\ bp$, $\leq 360\ bp$
Medium	$\geq 361\ bp$, $\leq 1220\ bp$
Large	$\geq 1221\ bp$

Table 3.8: A definition for gene length classification for *S. cerevisiae*.

to operate as a true centromere, but may have an affect on gene distribution.

Gene lengths

A histogram of the frequencies of gene lengths is given in Figure 3.9. Note that there is a sharp increase in frequency from approximately 60 to approximately 140 for a gene lengths over 300 bp and a subsequent drop in frequency from approximately 140 to approximately 80 for gene lengths over 360 bp. This was considered unusual and so an extra category of gene length, medium-small, was introduced for further research on *S. cerevisiae*. These categories or classifications

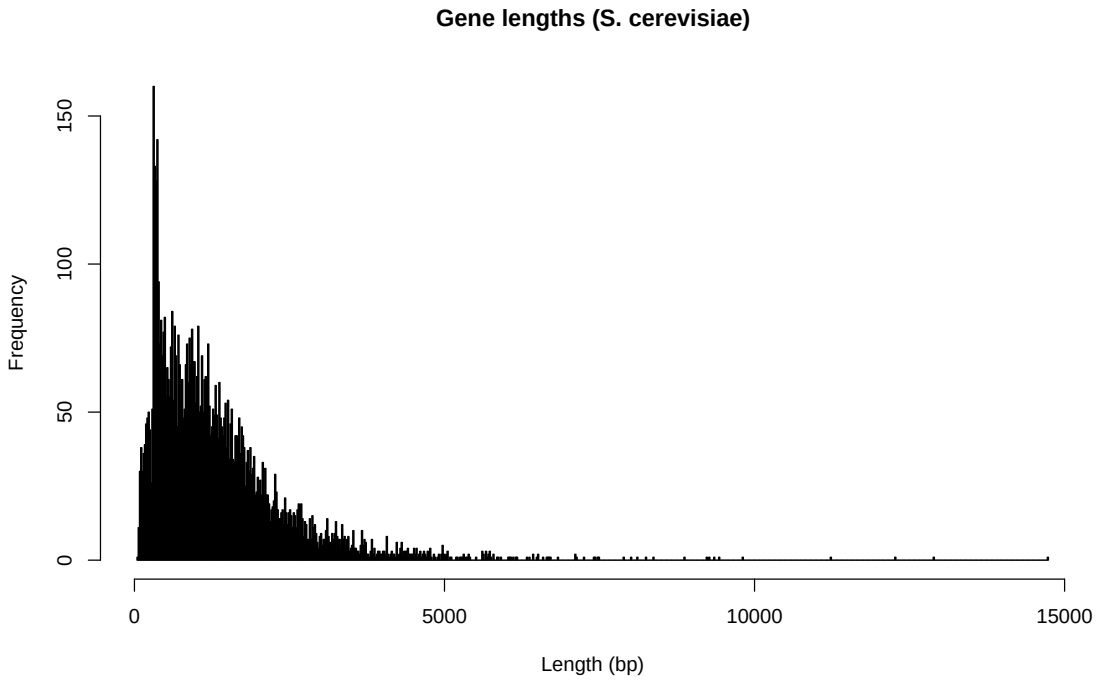

Figure 3.9: Histogram of the lengths of genes in *S. cerevisiae*. There is a sharp increase in frequency from approximately 60 to approximately 140 for a gene lengths over 300 bp and a subsequent drop in frequency from approximately 140 to approximately 80 for gene lengths over 360 bp.

are defined in Table 3.8.

Gap lengths

A histogram of the frequency of lengths of gaps between neighbouring genes is given in Figure 3.10. Note that there is a peak between 240 and 280 bp and another small local maximum at -280 to -320 bp. Categories for the lengths of gaps between neighbouring genes were derived using Figure 3.10. There are four categories: negative, small, medium and large. The definitions for these gap length categories are given in Table 3.9.

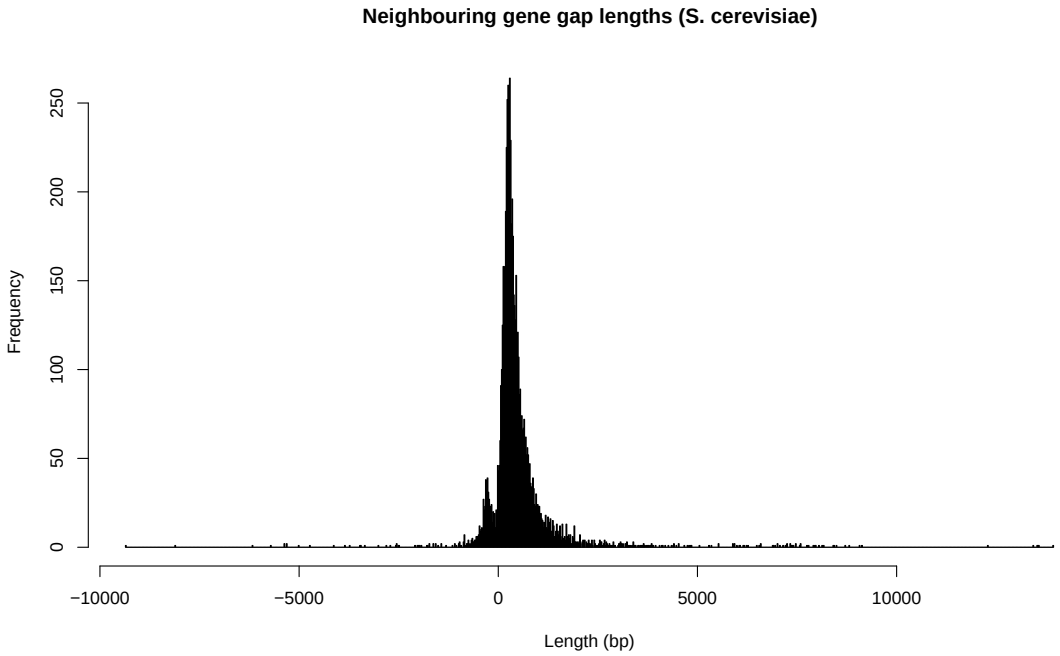

Figure 3.10: Histogram of the gap lengths between genes in *S. cerevisiae*. There are two peaks, one occurring between 240 bp and 280 bp and another small local maximum at -280 bp to -320 bp.

Classification	Definition
Negative	$\leq 0\ bp$
Small	$\geq 1\ bp$, $\leq 160\ bp$
Medium	$\geq 161\ bp$, $\leq 360\ bp$
Large	$\geq 361\ bp$

Table 3.9: A definition for gene gap length classification for *S. cerevisiae*.

3.3 Summary

This chapter introduced the model organisms *Arabidopsis thaliana* and *Saccharomyces cerevisiae* and presented some information on the distributions of genes, gene lengths and the lengths of the gaps between neighbouring genes. This information is used later in the analysis of locational clustering in Chapter 5 and in the data mining of gene location in Chapter 6.

Chapter 4

Statistical Tools and Methods

4.1 Statistical Methods

Originally *statistics* was a term used to refer to a collection of numbers, but the modern view is that *data* is the term to refer to a collection of numbers and the analysis and extraction of information from data is *statistics*, which is now popularly defined as "the science of decision making" (Dudewicz & Mishra, 1988). The *average* value of a random phenomenon or a set of data is referred to as one of the *moments* and subsequently the mean and the variance are known as moments. These moments are used throughout this research.

4.1.1 Moments

The *mean* ($\bar{m}$), which is also referred to as the *expectation* (E), can be thought of as a measure of location of the distribution of the data. The *standard deviation* can thought of as a measure of dispersion of the distribution of data. The mean can be seen as an average value of a set of data for most data sets. It is essentially the sum of all data divided by the size of the data so that, given a set of n numbers

$\{x_1, x_2, x_3, ..., x_n\}$ the mean value of all x_i, denoted by $\bar{x}$ is given by

$$\bar{x} = \frac{\sum x_i}{n} \tag{4.1}$$

$\bar{x}$ is sometimes called the *arithmetic mean* (Croft *et al.* , 1992).

A commonly used measure of the dispersion or spread of all x is the *standard deviation* often denoted by σ, which is the sum of the difference or *deviation* between x_i and the mean, $\bar{x}$ or $x_i - \bar{x}$. However, these deviations will be negative and positive and will always sum to zero so the squared deviation $(x_i - \bar{x})^2$ is summed giving the variance, which is often denoted by σ^2. Hence, the standard deviation is given by

$$\sigma = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}} \tag{4.2}$$

The standard deviation has the same units as x_i (Croft *et al.* , 1992)

There are two other moments of interest to this research and they are the third moment, which is called *skew* and the fourth moment, which is called *kurtosis*. Skew or skewness refers to the asymmetry of a distribution. There are a number of methods to determine skewness, but the most widely used one is:

$$Sk = \frac{\sum(x_i - \bar{x})^3}{n\sigma^3} \tag{4.3}$$

where the distribution of x is considered symmetrical if $Sk = 0$; positively skewed if $Sk > 0$; and negatively skewed if $Sk < 0$ (Kirk, 1999). However, there are likely to be slight fluctuations in the measure of skewness in any finite set of numbers and so a general rule of thumb is that the distribution is considered skewed if $Sk > \frac{\sigma}{2}$ or $Sk < -\frac{\sigma}{2}$.

Kurtosis refers to the peakedness or flatness of a distribution and is most commonly given by:

$$Kur = \frac{\sum(x_i - \bar{x})^4}{n\sigma^4} - 3 \tag{4.4}$$

where if $Kur = 0$, the peakedness is the same as a normal or Gaussian distribution and is referred to as *mesokurtic*. If $Kur < 0$ the distribution is flatter (broader

hump and thicker tails) than a normal distribution and this is referred to as *platykurtic*, and if $Kur > 0$, the distribution is more peaked (narrower hump and thinner tails) than a normal distribution and is referred to as *leptokurtic* (Kirk, 1999).

4.1.2 Standard Error

The statistical moments previously described give values that represent various features of a distribution, but the significance of these values depends on the size of the data set for a given distribution. For example, throwing a die four times to determine if it is fair would be a poor way to assess the fairness of the die since a four, a five and two sixes would not be unreasonable outcomes. But the data from these outcomes are skewed and indicates that the die is not fair. However, intuitively we know that one thousand throws of the die would give superior statistical data to assess the fairness of the die. The measure of significance of the statistical data is determined from the *standard error* and is inversely proportional to the size of the data. The standard error for standard deviation (se_σ), variance (se_{σ^2}), skew (se_{skew}) and kurtosis ($se_{kurtosis}$) can be calculated from equations 4.5, 4.6, 4.7 and 4.8 given below (Yule & Kendall, 1946).

$$se_\sigma = \frac{\sigma}{\sqrt{2n}} \tag{4.5}$$

$$se_{\sigma^2} = \sigma^2\sqrt{\frac{2}{n}} \tag{4.6}$$

$$se_{skew} = \sigma^3\sqrt{\frac{6}{n}} \tag{4.7}$$

$$se_{kurtosis} = \sigma^4\sqrt{\frac{96}{n}} \tag{4.8}$$

The form of these equations has been widely used to determine standard error,

but they strictly apply where the parent distribution is normal and should be used with caution for distributions that are not normal (Yule & Kendall, 1946). As a general 'rule of thumb' the statistical result of a moment can be considered significant if its absolute value exceeds two standard errors (Tabachnick & Fidell, 1996).

4.1.3 Mean Filtering

A mean filter is a simple method of smoothing using convolution that works by taking the mean of the data in a window, which moves along the data. This is a popular approach used to determine the degree of clustering in sequences, such as GC content in a nucleotide sequence or the clustering of genes in the genome. However, this approach does not work well on data with a high standard deviation or large transients between adjacent data.

The Gaussian smoothing operator is a convolution operator that acts like a filter to remove detail or noise. It is similar to a mean filter, but it uses a different kernel that represents the shape of a Gaussian curve. Gaussian smoothing is preferable because it reduces the windowing effect of large transients. This is because the centre of the sampling window has a high significance and the significance of the value of the data falls away as you move to the edges of the sampling window. This is the approach used in Chapter 5 in an analysis of gene clustering.

If μ is taken as the centre element of each sampling window and x represents the location of each element from the beginning to the end of the sampling window, then Gaussian smoothing can be implemented using equation 4.9.

$$G(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}} \tag{4.9}$$

where $G(x)$ is the Gaussian function of x and σ is the standard deviation of the required Gaussian curve and should take a value approximately one fifth of the number of elements in the sampling window, or less.

4.2 Probability

Probability is the scientific interpretation of chance and is in constant use in decision making (Ruelle, 1991). It is a measure of how likely an outcome will be and is commonly expressed as a percentage where 100% represents certainty and 0% represents no chance at all. However, in scientific research, probability is most frequently expressed as a number between 1 and 0, where 1 represents certainty and 0 represents no probability. Probability is also expressed as a ratio such as 1 in 25, meaning that there is only one chance of a particular outcome given that there are 25 possible outcomes. In this case a ratio of 1 in 1 represents certainty, but no chance at all is not well defined being 1 in (an infinitely large number). The representation should not be confused with *odds*, which is a ratio of a particular outcome compared with the number of alternative outcomes. For example the probability of throwing a six on a fair die is 1 in 6, but the odds are 1 to 5 (Bland & Altman, 2000).

There are three basic assertions of the mathematical representation of probabilities:

1. $P(Not\ A) = 1 - P(A)$
2. If A and B are incompatible then $P(A\ or\ B) = P(A) + P(B)$
3. If A and B are independent then $P(A\ and\ B) = P(A) * P(B)$

Two events are said to be *incompatible* if they cannot occur together (Mutually exclusive). Two events are said to be *independent* if they are unrelated.

Probability is used extensively in this research to determine significance and confidence in results. More specific applications of probability corresponding to various areas of the research covered in this thesis are described in more detail in following chapters. Some of the general concepts are presented here.

4.2.1 Factorial

Factorials are designated by the symbol ! and frequently feature in probability calculations. The factorial function is formally defined by: -

$$n! = \prod_{k=1}^{n} k \qquad \forall n \in N. \tag{4.10}$$

noting the special case where: -

$$0! = 1 \tag{4.11}$$

In other words, the product of no numbers at all is 1. This assertion for factorials is useful because the recursive relation: -

$$(n+1)! = n! \times (n+1) \tag{4.12}$$

works for $n = 0$ and so this definition makes many identities in combinatorics valid for zero sizes. In particular, the number of combinations or permutations of an empty set is, simply, 1.

A simple algorithm is given in Algorithm 1 for the solution of factorials and this works well for factorials of 170 of less.

Algorithm 1 Factorial(number)

Require: $result \Leftarrow 1$ and $i \Leftarrow 1$
  while $i \leq number$ **do**
    $result \Leftarrow result \times i$
    $i \Leftarrow i + 1$
  end while
  return($result$)

The problem with factorials is that they become very large numbers and quickly become intractable. E.g: -

$$450! = 1.73336873...10^{1,000} \tag{4.13}$$

This number exceeds the range available in most programming languages. In the

C programming language, double precision values type cast as *double* have 8 bytes, so the double type contains 64 bits; 1 for sign, 11 for the exponent, and 52 for the mantissa. Therefore, this type has a range of approximately $1.7 * 10^{-308}$ to $1.7 * 10^{308}$ and severely limits $n!$ such that $n \leq 170$.[1]

A simple solution is to return the factorial as a logarithm as demostrated in Algorithm 2, remembering to process the result as a logarithm in all further calculations.

Algorithm 2 Factorial(number). (*Result returned as a logarithm*)

Require: $result \Leftarrow 0$ and $i \Leftarrow 1$
  while $i \leq number$ **do**
    $result \Leftarrow result + log(i)$
    $i \Leftarrow i + 1$
  end while
  return($result$)

Stirling's approximation

The factorial definition given above works only for integers. Where n is not an integer we can use Stirling's approximation:

$$n! \sim n^n e^{-n} \sqrt{2\pi n} \tag{4.14}$$

where the sign $\sim$ indicates that the ratio of the two sides of the equation tend to unity as $n \to \infty$ (Feller, 1950).

4.2.2 Random Selection

Consider a large set of items of n different types. Then the following list summarizes the number of distinct ways in which k items can be selected:

[1] For GNU C++ on Windows XP and Solaris the highest factorial is 170!

1. Ordered samples selected with replacement:

$$n^k \tag{4.15}$$

2. Permutation, which is ordered samples selected without replacement:

$$nPk = \frac{n!}{(n-k)!} \tag{4.16}$$

   For example, if we pick $k = 3$ letters, (a, b, c) from n letters, then all permutations of those 3 letters will count.

3. Combination or Binomial coefficient, which is unordered samples selected without replacement:

$$nCk = \left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!} \tag{4.17}$$

   The order is irrelevant so (a, b, c) is the same as (a, c, b), (b, a, c) etc., i.e the permutations are not counted.

There are examples of the application of permutation and combination to establish significance in results, which are discussed in Chapters 6, 7 and 8.

4.2.3 Poisson Distribution

Poisson distribution describes the probability of the actual number of events occurring in any interval given an expected average. For example, if the average number of random occurrences per interval is given by λ, the probability P of k occurrences in the interval is given by equation 4.18 below.

$$P(k) = \frac{e^{-\lambda}\lambda^k}{k!} \tag{4.18}$$

where k is an integer (Papoulis, 1965) (Karr, 1993). The equation describes

the probabilities of random occurrences and is applicable to intervals in time or space.

As an example, consider a bus service that randomly despatches on average five buses per hour. This is an example of a Poisson distribution in time. Intuitively one would expect a bus to turn up once every 12 minutes, but the probability of exactly one bus arriving in that 12 minute interval is in fact:

$$P(k) = \frac{e^{-1}1^1}{1!} = e^{-1} = 0.368 \tag{4.19}$$

which is a less than evens chance. But, any passenger would be happy to take the first bus of 2 or more that turns up in the same interval so the passenger is more interested in the chances of any bus arriving in the 12 minute interval. The probability of any bus turning up is equal to one minus the probability that no buses turn up, as given by:

$$P(k = not\ 0) = 1 - [\frac{e^{-1}1^0}{0!}] = 1 - [e^{-1}] = 0.642 \tag{4.20}$$

Note: $0! = 1$ (see Section 4.2.1)

This will come as little surprise to bus passengers in busy city centres where the buses cannot run to a timetable due to the traffic congestion.

The Poisson distribution is ideal for describing events over time or space at a fixed rate on average, although occurring independently and at random (Altman, 1991, p66). It should be noted that Poisson distributions are asymmetric when the mean is small, but become symmetrical as the mean increases (Altman, 1991). This phenomenon is explored in more detail in Chapter 5.

The research described in Chapter 5 on gene distribution utilizes the Poisson distribution in space. An example of this can be modelled by taking say, one hundred small boxes and throwing 500 marbles at the collection of boxes. Assuming that all of the marbles fall into the boxes, there should be an average of five marbles in each box. The probability of finding five marbles in each box can be calculated

from equation 4.18. In this way it can be seen that for an average of 0.9 marbles per box ($\lambda = 0.9$), intuitively it would seem there is a higher chance of one marble in a particular box than none, but in fact there is a higher probability of finding none.

This application of the Poisson distribution has been used to determine the probability of finding a given number of genes in a selected contiguous sequence on the genome and this is discussed in Chapter 5.

4.2.4 Binomial Coefficient

In probability theory and statistics, a binomial coefficient given by equation 4.17 above, is a coefficient of the x^k terms in the expansion of the binomial $(1 + x)^n$. For example, given that there are n pizza toppings to select from, if one wishes to bake a pizza with exactly k toppings, then the binomial coefficient expresses how many different types of such k-topping pizzas are possible. A simple algorithm suitable for calculating the binomial coefficient is given below in algorithm 3.

Algorithm 3 The Binomial Coefficient - nCk(n, k)

if $(n \leq 0 || k < 0 || k > n)$ **then**
  return
end if
if $k < \frac{n}{2}$ **then**
  $k \Leftarrow n - k$
end if
$acc \Leftarrow 1$
for $i = 0$ to k **do**
  $acc \Leftarrow acc \times \frac{i+(n-k)}{i}$
end for
return(acc)

However, algorithm 3 is limited to smaller values for n. Algorithm 4 returns the binomial coefficient as a logarithm allowing a much larger range for n. It is this algorithm that is used extensively in the research presented in Chapters 7 and 8.

Algorithm 4 The Binomial Coefficient returned as a logarithm - log_nCk(n, k)

```
if (n ≤ 0||k < 0||k > n) then
  return
end if
if k < n/2 then
  k ⇐ n − k
end if
acc ⇐ 0
for i = 0 to k do
  acc ⇐ acc + log(i + (n − k)) − log(i)
end for
return(acc)
```

4.2.5 Binomial Distribution

In probability theory and statistics, the binomial distribution is the discrete probability distribution of the number of successes in a sequence of n independent tests. These independent tests are also called *Bernoulli experiments* or *Bernoulli trials*. In fact, when $n = 1$, the binomial distribution is a Bernoulli distribution. The binomial distribution is the basis for the popular binomial test of statistical significance, which is calculated using the probability mass function given in equation 4.22.

Probability mass function

Given that the expectation E of a result from a number of Bernoulli trials is given by:

$$E = np \tag{4.21}$$

where n is the number of trials and p is the probability of the positive outcome.

Then the probability of a result of k examples is given by:

$$f(k; n, p) = \begin{pmatrix} n \\ k \end{pmatrix} p^k (1-p)^{n-k} \tag{4.22}$$

The probability mass function is essential in the calculation of significance in Chapter 8.

Poisson approximation

The binomial distribution converges towards the Poisson distribution as the number of trials goes to infinity while the product np remains fixed. Therefore the Poisson distribution with parameter $\lambda = np$ can be used as an approximation to $B(n, p)$ of the binomial distribution if n is sufficiently large and p is sufficiently small. According to two rules of thumb[2], this approximation is good if n = 20 and p = 0.05, or if n = 100 and np = 10.

Note that, as above, the expectation $E = np$.

4.2.6 Multinomial Distribution

The multinomial distribution is a generalization of the binomial distribution.

For n trials where $n > 0$ and $p_1, \cdots, p_k$ event probabilities where $(\Sigma_{p_i} = 1)$, the probability mass function (abbreviated pmf) is a function that gives the probability that a discrete random variable is exactly equal to some value.

$$pmf = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k} \tag{4.23}$$

and $E(X_i) = np_i$

[2]From Counts Control Charts, eHandbook of Statistical Methods(6.3.3.1.) http://www.itl.nist.gov/div898/handbook/pmc/section3/pmc331.htm

4.3 Multiple Hypothesis Testing: Bonferroni Correction

One of the problems with multiple hypothesis testing is that, as more tests are made, there is a higher likelihood of finding a significant result by chance. The Bonferroni Correction is a safeguard against multiple tests of statistical significance on the same data, where any 1 in 20 hypotheses tested will appear be significant at the $P = 0.05$ level purely due to chance. This correction to significance is suited more to social science methods and clinical trials where the number of tests is generally small and significant results are often presented in isolation of the whole series of tests.

If testing n independent hypotheses on a set of data, then the statistical significance of each hypothesis should be reduced by a factor of n, by multiplying the P value by n. This approach is naive because it works on the premise that, given a statistical significance of $P = 0.05$, 1 test in 20 tests *will* be significant by chance when in fact 1 test in 20 tests *may* be significant by chance.

A disadvantage of Bonferroni is if a very large number of tests are done (e.g. 10,000) then the $P - value$ becomes so small after correction that it can become impossible for any result to be significant. This will lead to the loss of important results.

The research described in Chapter 5 has produced close to 1000 results and the significance of those results has been presented by way of ranking where a ranking of 1000 is equivalent to a $P - value \leq 0.001$. It has been suggested by reviewers of this work that Bonferroni correction should be applied to the significance results. However, due to the large number of results, Bonferroni correction produces uninformative significance figures and consequently, it has not been applied. Furthermore, the significance of some outlying results is directly affected by the general overall results, which would be obscured by Bonferroni correction (Explained in more detail in Chapter 5). However, some degree of consideration should be given to multiple hypothesis correction when assessing the significance of the results presented in Chapter 5.

4.4 The Greenwood Statistic

The Greenwood statistic is a spacing statistic suitable for the detection of uniformity of a locational distribution or, conversely, how clustered a distribution is (Greenwood, 1946). In preliminary testing we have found this statistic to be a more sensitive test of clustering than the window sampling method and produces more reliable results where data are sparse (Riley *et al.* , 2007).

In general, for a given sequence of events in time or space the statistic is given by:

$$G(n) = \sum_{i=1}^{n+1} D_i^2 \tag{4.24}$$

where D_i represents the interval between events and is a number between 0 and 1 such that the sum of all $D_i = 1$.

Where intervals are given by numbers that do not represent a fraction of the entire sequence, such as the base pair locations of genes, the Greenwood statistic can be modified (D'Agostino & Stephens, 1986) and is given by:

$$G(n) = \frac{\sum_{i=1}^{n+1} X_i^2}{T_n^2} \tag{4.25}$$

where:

$$T_n = \sum_{i=1}^{n+1} X_i \tag{4.26}$$

and X represents the base pair length of the interval between start loci of the genes.

The Greenwood statistic is a comparative measure that has a range of values between 0 and 1, which is inversely proportional to the number of points being analysed for a sequence of a given length. For example, applying the Greenwood statistic to a sequence of length 55 with eleven evenly spaced points each 5.5 units apart would give a result of 0.1. For a clustered sequence of six points 10 units

apart with a cluster of five points 1 unit apart the result is 0.167. The result for a random distribution of 11 points on the sequence will fall somewhere between these values[3].

We have used the Greenwood statistic to determine the nature of the locational distribution of genes in *Arabidopsis thaliana* and this work is described in Chapter 5.

4.5 Summary

The mathematical tools described in this chapter all prove to be very useful in the analysis of functional genomics, epigenetics and phylogenetics presented in this thesis. The use of sampling and Gaussian smoothing was essential in the preparatory work required for the research described in Chapter 5. A novel application of the Greenwood statistic is also presented in Chapter 5. The use of probability and expectation are used throughout Chapters 6, 7 and 8 to determine significance.

[3]This can be confirmed empirically

Chapter 5

The Locational Distribution of Genes in *Arabidopsis thaliana*

5.1 Introduction

This chapter details the work published by the author and others in BMC Bioinformatics (Riley *et al.* , 2007) and is essentially concerned with the statistical analysis of gene frequency and the locational distribution of genes classified by molecular function in the model organism *Arabidopsis thaliana*. For more information on this model organism see Section 3.1.

The research in this chapter focuses on the locational distribution of genes and their functions in genomes, as this distribution has both functional and evolutionary significance. Gene locational distribution is known to be affected by various evolutionary processes, with tandem duplication thought to be the main process producing clustering of homologous sequences. Recent research has found clustering of protein structural families in the human genome, even when genes identified as tandem duplicates have been removed from the data (Mayor *et al.* , 2004). However, this previous research was hindered as they were unable to analyse small sample sizes. This is a challenge for bioinformatics as more specific functional classes have fewer examples and conventional statistical analyses of these small

data sets often produces unsatisfactory results.

Further on in this chapter, a novel bioinformatics method based on Monte Carlo methods and Greenwood's spacing statistic is introduced. It is a method for the computational analysis of the distribution of individual functional classes of genes. We used this to make the first comprehensive statistical analysis of the relationship between gene functional class and location on a genome. Analysis of the distribution of all genes except tandem duplicates on the five chromosomes of *A. thaliana* reveals that the distribution on chromosomes I, II, IV and V is clustered at $P = 0.001$ (see Section 5.2.6). Many functional classes are clustered, with the degree of clustering within an individual class generally consistent across all five chromosomes. A novel and surprising result was that the locational distribution of some functional classes were significantly more evenly spaced than would be expected by chance.

5.1.1 The Locational Distribution of Genes

It was once thought that the distribution of genes on the chromosomes of eukaryotes was essentially locationally independent, i.e. knowledge of the position of n genes on the chromosome does not help you to find the $n + 1$th gene (just as knowledge of n tosses of a fair coin do not help you to predict the $n + 1$th toss). However, recent studies on the genomes of *Homo sapiens* and *Caenorhabditis elegans* have challenged this view (Mayor *et al.* , 2004; Blumenthal & Gleason, 2003; Blumenthal, 2004).

There has been considerable research into the location of genes in prokaryotes since the discovery of the operon in *Escherichia coli* (Jacob & Monod, 1961). The genome of *E. coli* has a heterogeneous gene frequency distribution overall (Riley *et al.* , 1978), but is divided into areas of homogeneous gene frequency (De Martelaere & Van Gool, 1981). Recent research has found scale invariant correlations (Audit & Ouzounis, 2003), convergence of coregulating regions (Warren & ten Wolde, 2004b), periodicity (Képès, 2004) and strong compositional asymmetries between leading and lagging strands (Rocha *et al.* , 1999). However, protein syn-

thesis and the structure of the genome in eukaryotes is altogether very different from prokaryotes and consequently the mechanisms affecting gene location in eukaryotes are likely to be very different.

An important consideration in the location of genes is the existence of operons (see Section 2.1.3). Operons afford an organism an evolutionary advantage by having co-operating or interdependent genes located in close proximity on the genome. This would be a contributing factor to a non-random distribution of genes on the genome. To date however, operons have not been found in *A. thaliana* and therefore cannot be considered when explaining the distribution of genes in the organism.

Among the many reasons why genes may not be located independently is the process of genetic mutation by tandem duplication. Tandem duplications (aka tandem repeats) are genetic mutations where a sequence of nucleotides becomes duplicated, with the duplicated sequence lying adjacent to the original sequence. Where tandem duplication extends to duplicating an entire gene, the resulting redundant gene can freely acquire mutations and emerge with a refined or entirely new function (Ohno, 1970). Tandem duplications that include complete genes may produce clusters of identical genes, which become mutated further through subsequent evolution to produce a cluster of similar genes. When considering gene function, it is likely that these genes will belong to the same functional class.

It is still not clear for eukaryotic genomes whether all gene clusters occur simply as a consequence of genetic mutations such as tandem duplication, or whether there is a functional benefit to gene clustering that conveys an evolutionary advantage. The persistence of intragenic repeats in certain classes of genes implies a possible evolutionary advantage. For example, in the genome of *Saccharomyces cerevisiae* most genes containing intragenic repeats encode cell-wall proteins (Verstrepen *et al.* , 2005). We may gain some insight by isolating the known causes of clustering and analysing the gene distributions that remain. Most research looking into the distribution of genes has focused attention on what are loosely described as clusters (Durand & Sankoff, 2003), and has largely involved analysing histograms of gene loci. In organisms with large genomes, such as *Homo sapiens*, dense clusters of

genes are clearly visible in the histograms (Venter *et al.* , 2001). However, in organisms with more compact genomes, such as *A. thaliana*, the distribution of genes is more difficult to analyse visually. Therefore, a more directly statistical approach is required.

5.1.2 Methodology Overview

In the first part of this study we analyse the locational distribution of all known genes after removing tandem duplicates and genes in the centromeric regions. We use a sliding window analysis where we take the standard deviation of the results as a measure of the degree of clustering and compare with randomly generated sequences of gene locations (see Section 5.2). If tandem duplication and the centromeres are the sole causes of clustering we would expect to obtain locationally independent distributions, which would be statistically related to distributions of genes placed at random on a simulated chromosome. However, the results reveal that, after the removal of the centromeres and tandem repeats, the distribution of all known genes is still locationally dependent.

Further in this study we analyse the locational distribution of genes classified by molecular function. Here we introduce Greenwood's spacing statistic which uses the distances between points or the time between events to give a comparative measure of clustering of those points or events. Low values tending to 0, are indicative of points being evenly spaced apart, whereas high values tending to 1, indicate that points are clustered (see also Section 4.4). Values roughly half way between indicate that the points are distributed at random. We compare the results with those of randomly selected gene locations on the original sequence (see Section 5.2). This gives us a relative measure of how clustered or how evenly spaced the distribution is compared to a locationally independent distribution. We establish the locationally independent distribution using Monte Carlo methods (Metropolis & Ulam, 1949) and by using this method we do not need to exclude genes in the centromere, but we do exclude tandem duplicates.

Again, the results reveal that the distribution of molecular functional classes of

genes is not locationally independent.

5.2 Methods

We first analysed the overall gene distribution using a standard statistical technique, then analysed individual functional class distribution using the Greenwood spacing statistic.

5.2.1 Data

The gene data to be analysed were downloaded from the MIPS[1] website in April 2005. This version of the data was dated 5/5/04. This data was used to extract the base pair (BP) start loci, end loci and BP lengths together with the gene identifiers (IDs). The Gene Ontology[2] molecular function annotations (version 3.230 - 31/3/2005) were downloaded from the TIGR[3] website. From this we extracted lists of gene IDs for each classification (The Gene Ontology Consortium, 2000). We examined all molecular functional classes that had at least 100 instances across the entire genome with any evidence code[4]. The classes were arranged in levels of increasing specificity. Excluding the obsolete and unknown classes, there are 10 subclasses of the molecular function class. These we have designated as the level 1 classes. The subclasses of these level 1 classes were designated as level 2, and so on for levels 3 and 4. This data was cross referenced with the loci data set to obtain a data set of the loci of each class of genes. This dataset was then used to analyse the distribution of genes on the chromosomes of *A. thaliana*. The molecular functional classes analysed are listed in Appendix A together with the results.

[1]MIPS website: http://mips.gsf.de
[2]GOC website: http://www.geneontology.org
[3]TIGR website: http://www.tigr.org
[4]For information on GO evidence codes see http://www.geneontology.org/GO.evidence.shtml

5.2.2 Removal of Tandem Duplicates

Previous research (Mayor *et al.* , 2004) has demonstrated that tandem duplicates have an impact on the degree of clustering. We were therefore interested in examining how tandem duplicates affect the gene distributions in *A. thaliana*. The Arabidopsis Genome Initiative (AGI) have published data on genes thought to be tandem duplicates. They identified these tandem duplicates using BLASTP (Altschul *et al.* , 1997) with a threshold of $E < 10^{-20}$ and one unrelated gene among cluster members was tolerated. By this method they identified 3737 tandem duplicates in 1456 tandem arrays. The latest data on tandem duplicates (release 5.0) was downloaded from the TIGR website.

To confirm these results we used BLAST version 2.2.13 (see 2.3.4) to identify tandem duplicates. We used the same threshold as the AGI of $E < 10^{-20}$, but we did not tolerate any unrelated genes within cluster members. We identified a similar number of genes to the data downloaded from TIGR. However, we chose to use the TIGR tandem duplicates data in our further analysis.

All of the genes identified as tandem duplicates were removed from the molecular function class data except for the first gene in each array. A total of 2281 tandem duplicate genes were removed. Clearly, the interval between the remaining gene marking the location of the tandem array and its nearest neighbour is marginally extended, but this has a negligible impact on the results.

5.2.3 Distribution of All Genes

To determine the distribution of all genes on each chromosome of *A. thaliana* we used a sampling window to sum the intergene gap lengths within each of the windows along the entire chromosome minus the centromere (see below). The length of the sampling window was chosen such that the mean for the number of genes in each window is 10. This is a compromise between Poisson asymmetry from smaller windows (see below) and clustering insensitivity from larger windows. Sampling windows were applied sequentially with no overlap. A test example using

a 10% overlap gave only a marginal improvement in clustering sensitivity, but at a tenfold cost in processing time.

We used the standard deviation of the results obtained from the above method as a measure of the clustering of the distribution; a high standard deviation would imply a higher degree of clustering. This is because the limiting case would be a constant intergene gap distance (0 standard deviation) which would give an evenly spaced distribution (minimum clustering). To determine how clustered the distributions are, the results are compared to a Monte Carlo simulation (Metropolis & Ulam, 1949) of locationally independent events. Each Monte Carlo trial involved creating a 'pseudo-chromosome' by randomly selecting a gene gap length from the original gene data and then randomly selecting a gene length from the original data. Once a gap length or gene length had been selected it was removed from the random selection procedure so that each datum is selected without replacement. The random selection of gap lengths and gene lengths continues for all the genes in the chromosome being analysed. We are therefore effectively scrambling the locations of the genes. Once the 'pseudo-chromosome' is created, the same statistical analysis is used to obtain the standard deviation of the number of genes in each window. The generation of 'pseudo-chromosomes' in this way is equivalent to a null model that states that all the clustering is due to the known first-order distribution of lengths of genes and gaps between genes. One thousand Monte Carlo trials were taken, producing one thousand values for the standard deviation. The mean value of the standard deviations was recorded and this gives a reliable measure of the clustering of the distribution of genes on a chromosome where the genes are randomly distributed, and so this value can be used for comparison to the original.

As it is well known that genes are depleted within the centromeric regions of eukaryotic chromosomes (Alberts *et al.* , 2002), inclusion of the centromeric region in this analysis would directly indicate clustering. Therefore, since we were more interested in the distribution of genes in the 'main' sequence of the chromosome it was necessary to exclude the data from the centromere of each chromosome. From a gene frequency plot the approximate centre of the centromeres could easily be identified. An example of one of these plots for chromosome I is shown in Figure

5.1. The centromeric regions were then identified as regions where the average gene frequency for a sampling window of 31,000 bp fell below 6.5 for all contiguous sampling windows about the approximate centre of the centromere. A total of 6200 genes were excluded, which is a fairly large number, but ensures we have excluded all centromeric gene depletion. Details of the beginning and end of each centromere and the genes excluded are shown in Table 5.1.

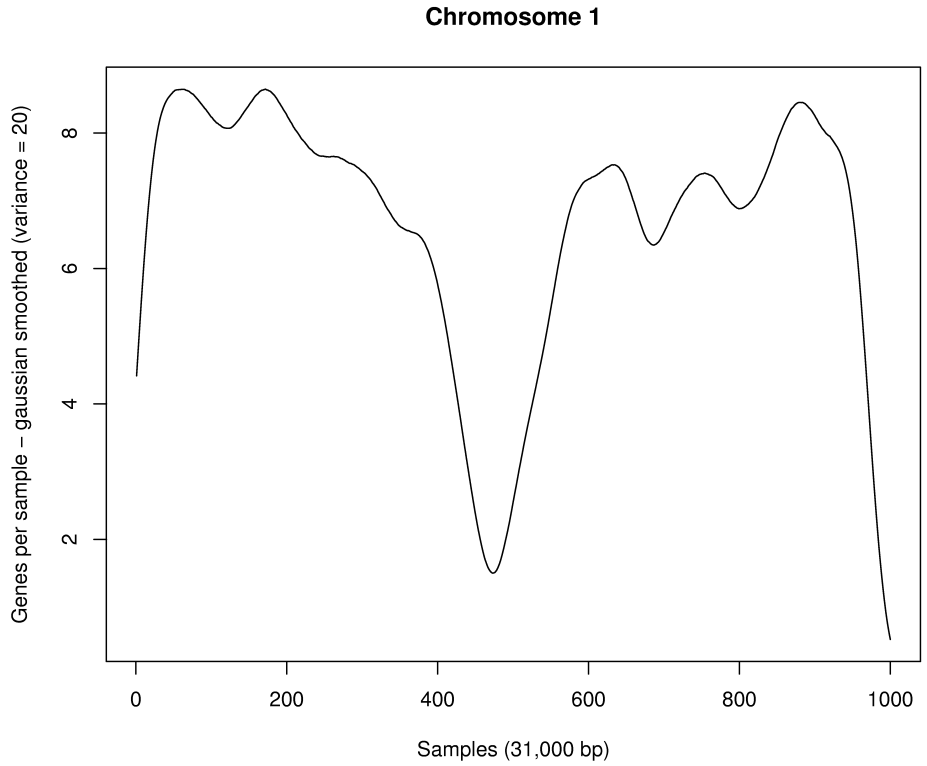

Figure 5.1: An example of a plot of gene frequency smoothed to reveal the extent of the centromeric region in chromosome I. Similar plots for the remaining chromosomes were also used to produce the areas of the chromosome to be excluded, which are displayed in Table 5.1. Note that the approximate base pair locations are taken from the x-axis multiplied by the number of samples (31,000 for this example).

Chromosome	Start (Mbp)	End (Mbp)	Genes excluded
1	11.5	18.5	At1g32000 - At1g50919
2	0.0	7.2	At2g01050 - At2g16160
3	9.1	17.1	At3g25100 - At3g47090
4	0.0	6.0	At4g00010 - At4g11240
5	5.4	16.9	At5g16500 - At5g42320

Table 5.1: Details of the centromeric regions excluded from the analysis showing the start and end locations of the centromeres determined by the method given in the text.

5.2.4 The Locational Distribution of Functional Classes of Genes

The locational distribution of genes on both W and C strands of each chromosome classified by molecular function was also considered. Mayor *et al.* (2004) have previously used a symmetric Poisson distribution to study the related problem of the locational distribution of structural classes of proteins in the human genome. This Poisson distribution based approach has the disadvantage that as the expectation or mean decreases the Poisson distribution becomes asymmetric (Altman, 1991). As some of the classes have less than ten examples on some strands this approach is therefore problematic.

By plotting a series of graphs of the Poisson distribution for a range of expectations from 0 to 10 in increments of 0.5, it can be clearly seen that expectations below 4.5 produce a significantly asymmetric Poisson distribution, resulting in unreliably skewed results. Sampling with an expectation above 4.5 results in there possibly being too few samples for analysis in the smaller data sets such as the molecular function classes at more specific levels in the Gene Ontology hierarchy. The standard error calculated from equation (5.1) where n is the number of samples and σ is the standard deviation (Yule & Kendall, 1946), means that for a set of data of just two or three samples the standard error is thus about 40 - 50%.

$$se_\sigma = \frac{\sigma}{\sqrt{2n}} \tag{5.1}$$

As a general 'rule of thumb' any statistic should only be considered significant

if it exceeds two standard errors (Tabachnick & Fidell, 1996) and consequently, we would be looking for a standard deviation to vary by 80 - 100% to be significant. This is unlikely to be informative and so an alternative approach was considered.

5.2.5 The Greenwood Statistic

The Greenwood statistic is a more sensitive measure of clustering than using the variance in the number of samples found within a sampling window. This increased sensitivity comes at the expense of location information or, put more simply, we can detect the existence of clusters but we cannot tell where the clusters are. At this stage in this research we are more interested in the existence of clustering so the Greenwood statistic is more suitable. Details of the Greenwood statistic can be found in Section 4.4.

To determine significance levels for the Greenwood statistic on gene function we used a Monte Carlo approach based on comparing the Greenwood statistic for a particular functional class of genes, with the Greenwood statistic for a thousand simulated chromosomes. These simulated chromosomes are created by randomly selecting the same number of genes as the class under investigation, from any class of genes on the chromosome. In this way we are using the distribution of genes on the existing chromosome as a null model from which we can make a comparison and thereby alleviating the need to exclude genes in the centromeres. By evaluating the Greenwood statistic for one thousand simulated chromosomes we obtained an empirical distribution of the probability of the evenness or clustering of a random distribution. The results of the Greenwood statistic for one thousand simulated chromosomes are arranged by order of value giving us a ranking by which we can compare the Greenwood statistic of the molecular function class under investigation.

To apply the Greenwood statistic accurately to the distances between genes (or intervals on the chromosomes) it is important that the simulated chromosomes generated are exactly the same length as the original chromosome. Also, the

interval from the beginning of the chromosome to the start of the first gene and the interval from the end of the last gene to the end of the chromosome must be included in the data.

The random selection algorithm used for the Monte Carlo trials utilized Park and Miller's minimal standard congruential multiplicative random number generator (Park & Miller, 1988) ensuring good properties of a random number generator.

5.2.6 Ranking and P-values

Note that rankings used throughout this chapter range from 1 to 1000 and a ranking of 500 represent the results we would expect from a locationally independent distribution. Rankings below 500 are increasingly evenly spaced distributions and rankings above 500 are increasingly clustered distributions.

Note that the P-values (P) given in the introduction are obtained from the ranking thus:-

$$P = \frac{(1000 - ranking) + 1}{1000} \tag{5.2}$$

5.3 Results

5.3.1 Distributions of All Genes

The results for the distribution of all genes without tandem duplicates are briefly summarized in Table 5.2, which shows that the genes on all five chromosomes of *A. thaliana* are significantly more clustered than would be expected from a locationally independent distribution.

We can use the standard deviation as a measure of clustering, as explained in the methods section, and we can use the standard error as a measure of the significance

Chr	Rank	Original SD	Mean MC SD	Std Err
1	1000	2.71	2.43	0.054
2	1000	2.67	2.31	0.052
3	957	2.42	2.29	0.051
4	1000	2.74	2.44	0.054
5	1000	2.51	2.19	0.049

Table 5.2: Table detailing the ranking (see Section 5.2.6), the standard deviation in the distribution of the original genes (Original SD), the mean of 1000 standard deviations from the Monte Carlo simulations (Mean MC SD) and the standard error (Std Err) on all five chromosomes (Chr). The standard deviation gives us a measure of clustering. The significance of these results can be determined from the difference between original standard deviation (Original SD) and the mean standard deviation for all Monte Carlo simulations (Mean MC SD), divided by standard error (Std Err).

of the result. We establish the null hypothesis from the mean standard deviation of 1000 Monte Carlo trials of randomly generated chromosomes. Refering to Table 5.2, we can see that the standard deviation (Original SD) for chromosome I is 2.71 and the mean standard deviation for 1000 Monte Carlo trials of randomly generated chromosomes (Mean MC SD) is 2.43. The standard error for the size of this data set (Std Err) is 0.054. The difference between the standard deviations divided by the standard error is 5.18; i.e. the standard deviation for chromosome I is 5.18 standard errors from the null hypothesis. Any result greater than two standard errors should be considered significant (Tabachnick & Fidell, 1996) so we can see that this result is very significant.

The standard deviation of the distribution of genes on chromosomes I, II, IV and V ranked 1000 out of 1000 Monte Carlo simulations of a random chromosome. The standard deviations for these chromosomes exceeded 5 standard errors of the mean standard deviation for the Monte Carlo simulations. The standard deviation of chromosome III ranked 957 out of 1000 and had a value of 2.34 standard errors from the mean, which indicates that this result is significant, but there is a small probability that this distribution could occur by chance.

Level	Ave. ranking (TD removed)	Ave. ranking (all)
1	713	796
2	705	779
3	652	725
4	675	745

Table 5.3: Average ranking of all the functional classes analysed with and without tandem duplicates (TD) on all five chromosomes of *A. thaliana* from four levels of the Gene Ontology hierarchy showing that the degree of clustering of the distribution of broadly classified genes is similar to that of the more specific classifications.

5.3.2 The Locational Distribution of Functional Classes of Genes

The full results for the distribution of individual functional classes are listed in 20 tables in Appendix A. The tables are arranged so that each table lists the results for each of the five chromosomes over four levels of the Gene Ontology hierarchy (explained in more detail in the methods section) making 20 tables in total.

The Greenwood statistic of each functional class was compared to 1000 Monte Carlo simulations of a random distribution of the same number of genes as found in each functional class. The average rankings of the Greenwood statistic for all classes in all four levels of the Gene Ontology hierarchy across all five chromosomes are listed in Table 5.3. These show that, in general, the functional classes are more clustered than would be expected from a locationally independent distribution. Furthermore, referring to the supplementary tables in Appendix A, we can see that 12% of functional classes in level 1 were super-clustered having a ranking of 1000 out of 1000.

For each class there are ten results representing the relative ranking of the Greenwood statistic compared to the null hypothesis, one for each strand on each of the five chromosomes. The individual results can be found in Tables A.1, A.2, A.3, A.4 and A.5 in Appendix A. To better visualize these results for the 10 most populated functional classes at level 1 we used the R statistics software package (R Development Core Team, 2005) to create box and whisker plots (aka boxplots)

(Tukey, 1977) and these are displayed in Figure 5.2 [5]. The circles represent outliers as interpreted by the default boxplot parameters of the R statistics software.

[5]A second boxplot is displayed in Figure 5.3 where the data includes tandem duplicates and shows a marked increase in clustering over all functional classes.

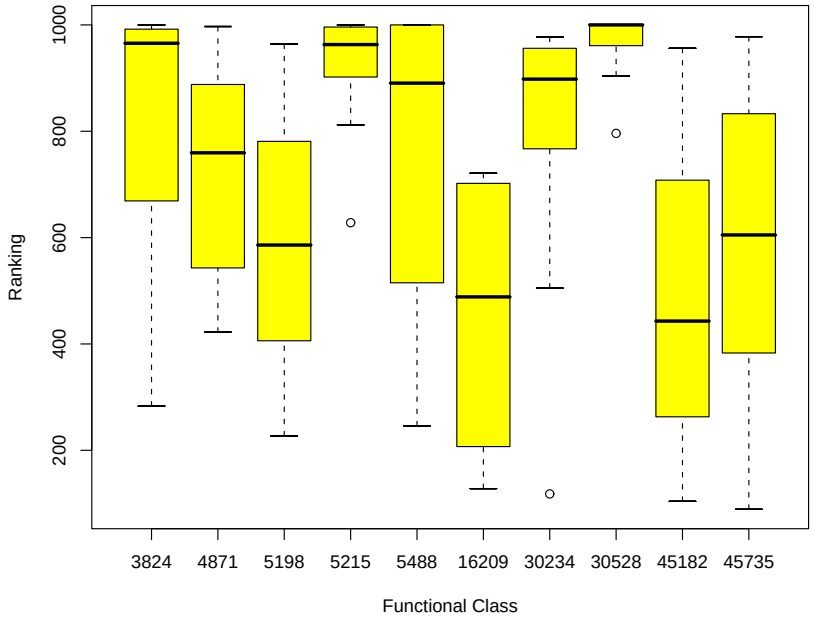

Figure 5.2: Distribution of rankings of the functional classes without tandem duplicates at level 1, the ten most general functional classes of the GO hierarchy of both W and C strands across all five chromosomes of *Arabidopsis thaliana*. The labels on the x axis refer to the Gene Ontology classifications described in Table 5.4. The y axis is representative of the relative degree of clustering of genes, where 500 indicates what we would expect if the genes are located at random, above 500 is increasingly clustered and below 500 the genes are increasingly evenly spaced apart. This plot demonstrates that different functional classes have remarkably different degrees of clustering.

Class No.	Description
GO:0003824	Catalytic activity
GO:0004871	Signal transducer activity
GO:0005198	Structural molecule activity
GO:0005215	Transporter activity
GO:0005488	Binding
GO:0016209	Anti oxidant activity
GO:0030234	Enzyme regulator activity
GO:0030528	Transcription regulator activity
GO:0045182	Translation regulator activity
GO:0045735	Nutrient reservoir

Table 5.4: Descriptions of the Gene Ontology annotations used in the boxplots in Figure 5.2 and Figure 5.3.

5.3.3 Clustered Distributions

The functional classifications at level 1 are very broad. It is therefore surprising that there is a marked difference in the degree of clustering among the functional classes. The plots of the genes associated with structural molecule activity (GO:0005198), anti oxidant activity (GO:0016209), translation regulator activity (GO:0045182) and nutrient reservoir classification (GO:0045735) are examples of the distributions that might be expected from these broad classifications, as they show no significant clustering on all five chromosomes for these functional classes. However, most of the functional classes show a high degree of clustering that prevails across all five chromosomes. The plots for genes associated with catalytic activity (GO:0003824), transporter activity (GO:0005215), enzyme regulator activity (GO:0030234), transcription regulator activity (GO:0030528) and binding (GO:0005488) indicate that these functional classes are consistently and very highly clustered throughout the genome.

A number of molecular function subclasses of the five main clustered classes mentioned above are also super-clustered having a ranking of 1000 out of 1000. Referring to the results in the tables in Appendix A it can be seen that at level 2 we found five out of ten super-clustered instances of transcription factor activity (GO:0003700), which is a subclass of transcription regulator activity. For

the binding class we found 3 out of 10 super-clustered instances of nucleic acid binding (GO:0003676), one of nucleotide binding (GO:0000166), one of protein binding (GO:0005515) and one of lipid binding (GO:0008289) and at level 3 we have one instance of DNA binding (GO:0003677) and one of purine nucleotide binding (GO:0017076). Finally, there are 8 super-clustered subclasses of catalytic activity, which can be found on levels 2, 3 and 4.

With catalytic activity class members displaying such a consistency in clustering it was surprising to find that there was one class member at level 4, calcium ion binding (GO:0005509), that had one instance displaying a very evenly spaced distribution with a ranking of 0 out of 1000. Looking at molecular function classes from all levels in the GO hierarchy we found 9 instances of evenly spaced distributions with a ranking of 25 or less out of 1000, which were all members of three of the five main clustered classes, with just two exceptions that belonged to the signal transducer activity class (GO:0004871).

We repeated these statistical analyses without removal of tandem duplicates. This resulted in slightly more evidence for clustering but did not affect any major conclusion. The results of this analysis are summarised in Figure 5.3.

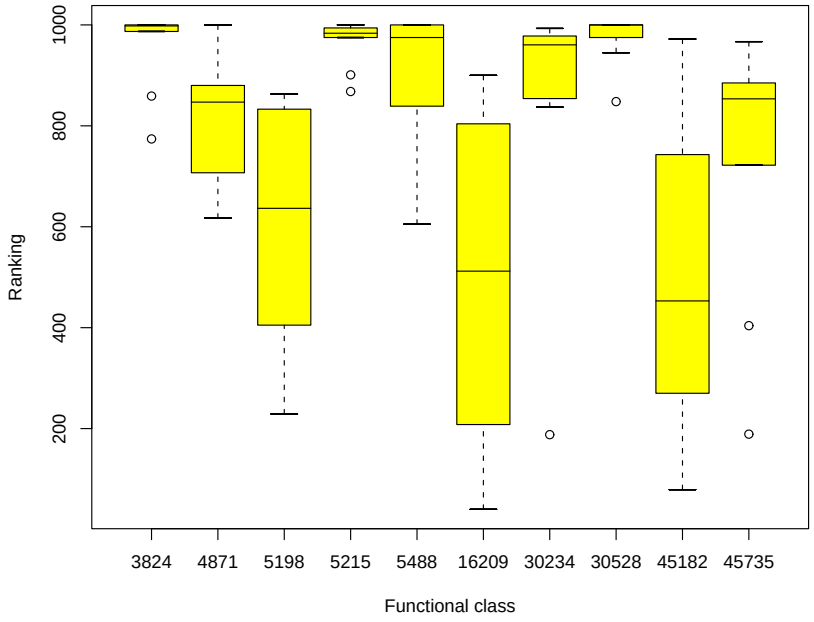

Figure 5.3: Distribution of rankings of the functional classes including tandem duplicates at level 1 of the GO hierarchy of both W and C strands across all five chromosomes of *Arabidopsis thaliana*. The labels on the x axis refer to the Gene Ontology classifications. Refer to Table 5.4 for a description of these annotations. This plot is the same as Figure 5.2, but with the tandem duplicates included. This demonstrates that tandem duplicates increase clustering by a small degree in all of the most general functional classes. Note that we found some more specific classes at level 4 that were much less susceptible to tandem duplication (see main text).

5.3.4 Evenly Spaced Distributions

We also took a closer look at three specific molecular function classes at level 4 in the GO hierarchy which showed very evenly spaced distributions. These were calcium ion binding activity, G-protein receptor activity and metallopeptidase activity.

Calcium ion binding activity

Genes associated with calcium ion binding activity (GO:0005509) have a very evenly spaced distribution on the W strand on chromosome IV, having a Greenwood statistic ranking of 0 out of 1000. Closer analysis of these 275 genes shows that 9% of these genes are tandem duplicated compared to the average of 17% for all genes. Using the AGI data for tandem duplicates, 12 tandem arrays were identified, 11 tandem pairs and one tandem triplet. There were no observed tandem duplications on the W strand of chromosome IV.

G protein coupled receptor activity

Genes associated with G-protein coupled receptor activity (GO:0004930) displayed more evenly spaced distributions on both W and C strands on chromosome IV with statistic rankings falling in the lowest 4%. There are 157 genes associated with G-protein receptor activity (GO:0004930) in *A. thaliana*, but only eight tandem duplicates have been identified. Furthermore, there were no tandem duplications on chromosomes II and IV. This class was particularly interesting because we found evenly spaced distributions and no tandem duplications on both strands of chromosome IV. However, there are also no tandem duplications on chromosome II, which has a highly clustered distribution. N.B. The location of G protein coupled receptor activity genes in the human genome are frequently distributed in tandem arrays.

Metallopeptidase activity

Of the 172 genes associated with metallopeptidase activity (GO:0008237) only 10 were tandem duplications with one pair on chromosome I and two pairs and an array of four tandem duplications on chromosome V. This functional class has an average ranking for chromosomes I, II, III and V that is similar to the average ranking for all functional classes, but this class on chromosome IV ranks in the bottom 10% indicating a very evenly spaced distribution. This would indicate that evenly spaced distributions are not necessarily dependent on gene molecular function class.

These three molecular function classes where we have found evenly spaced distributions all have a lower than average frequency of tandem duplications.

5.4 Discussion

We have seen evidence of very high levels of clustering even after the removal of tandem duplicates for half of the number of molecular function classes at level 1. The remaining half showed higher than average levels of clustering compared to the Monte Carlo simulation with just one exception. Throughout the subclass levels 2, 3 and 4 we find both extremes in that there are frequent occurrences of super-clustered distributions and a number of distributions that are more evenly spaced than we would expect. Although it must be considered that the evenly spaced distributions could just possibly have occurred by chance, this seems unlikely and we consider these anomalous distributions to be worthy of more research.

5.4.1 Tandem Duplicates

Tandem duplication is thought to be one of the principal mechanisms of gene proliferation and is also thought to be the main cause of clustering. Our results confirm that tandem duplication is a cause of clustering, but is unlikely to be the sole cause. The results of the further analysis of genes associated with G protein

coupled receptor activity in *A. thaliana* indicate clearly that tandem duplications are not the only process that generate gene clustering since the distribution of this class on chromosome II is clustered, but contains no tandem duplications.

Another observation regarding tandem duplications is that genes of many individual classes show roughly the same degree of clustering across both strands on all five chromosomes, and this indicates that clustering is in some way dependent on gene molecular function. This may further imply that tandem duplications are gene molecular function dependent.

5.4.2 Evenly Distributed Classes of Genes

There are many reasons to expect clustered gene functional distributions as we have already discussed. There is also strong evidence for clustering of structurally related genes in the human genome (using a different statistical approach) (Mayor *et al.* , 2004). It was therefore surprising to find that some functional classes on some chromosomes were significantly more evenly spaced than would be expected by chance. The evenly spaced distribution of some functional classes would imply something about the nature of genes of that molecular function. We have found that the classes displaying even distributions have fewer than average tandem repeats. It would seem that some gene functional classes do not appear to be so prone to tandem duplication. But, since tandem duplication is not the only cause of clustering there are likely to be other factors involved. For example, there maybe an evolutionary advantage in distributing essential genes evenly across the genome.

Other factors affecting the locational distribution of gene functional classes may include the 3 dimensional structure of the chromosome itself. The degree of coiling of the chromatin varies during the life cycle of the cell. When the chromatin is tightly coiled or highly condensed the number of genes physically available for expression is low. More genes are available for expression during the phases required for cell division when the chromatin is decondensed. The chromatin exists in a partially condensed state when a cell has matured. Evidently, in the matured

state, less genes are physically available for expression and clearly the genes required for the specific functions of the matured cell must be available. These genes will need to be located in regions of the chromosome that are available for expression and this could lead to both clustering and even spacing. Clustering because essential genes available for expression will occur in the physically accessible areas. Even spacing because the coiling/structure of the chromatin will lead to physically accessible regions having an inherent cyclic nature and essential genes located in these areas will have an evenly spaced distribution on the primary structure of the genome.

5.5 Conclusions

The distribution of all genes and the distribution of individual functional classes of genes in *Arabidopsis thaliana* were found to be more clustered than we would expect from a locationally independent distribution. Although tandem duplications contribute considerably to clustering, they are clearly not the only factor affecting the observed clustered distributions. This result is consistent with the observations of Mayor *et al.* (2004) on the distribution of protein structural domains in the human genome. We found three molecular function classes in *A. thaliana* that are significantly more evenly distributed than would be expected from a locationally independent distribution. The mechanism for this evenness is unknown. Both the evidence of clustering and the evidence of evenness implies that there are unexplained elements of order in the locational distribution of genes in *A. thaliana*.

Chapter 6

Pattern Mining: Gene Location

6.1 Introduction

This chapter continues the theme from the last chapter, but focuses on a more specific analysis of the location of genes on the genome. By using pattern mining we expect to obtain more detailed information on gene location. The patterns can be represented as queries and one such example query could be that, where we find genes of class A, do we find genes of class B close by? If so, is the expression of genes affected by their neighbours? Answers to these questions may reveal simple, or even highly complex systems of gene expression of some other form of gene 'communication'.

A large number of sources of research have concluded that the complex biology of an organism arises from more information than is contained in the DNA sequence alone (Goldberg *et al.* , 2007). On the basis of this the view of chromatin has broadened to more than just DNA packaging. Chromatin is the name given to all the supporting proteins surrounding the DNA that were thought only to control the coiling and packing of the DNA when it is in a condensed state. Chromatin is now seen as instrumental in the regulation of gene expression and further, it is a complex network central to regulation of different genome functions. This network may be the determinant of gene activity by the maintenance and inheritance

of active and inactive chromatin states. Higher order chromatin structures are important for replication and 'faithful' separation of chromosomes and for spatial organization of genes within the nucleus.

To date, three principal specific 'non genetic' biochemical mechanisms have been identified: DNA methylation; histone modifications; and the binding of non-histone proteins such as polycomb 2 and trithorax group complexes (Bock & Lengauer, 2008). 'Non genetic' functions and processes such as these are considered to be the bridge between genotype and phenotype, and collectively they are known as *epigenetics* (Goldberg *et al.* , 2007).

6.1.1 Epigenetics

Conrad Waddington first introduced the term *epigenetics* in 1942 to mean "...the branch of biology which studies causal interactions between genes and their products, which bring the phenotype into being" (Waddington, 1942). The modern interpretation is that epigenetics is the field concerned with the molecular mechanisms that influence the phenotypic outcome of the gene or genome, in the absence of changes to the underlying DNA sequence.

The field of epigenetics is attracting increasing interest from many areas and particularly from cancer research. Epigenetic inheritance is encoded in modifications of the covalent bonding of the DNA and the chromatin proteins attached to it. There is now growing evidence that epigenetic 'errors' are more likely than genetic 'errors' and this is of particular interest in the study of cancers (Jones & Baylin, 2007) (Schlesinger *et al.* , 2007) (Ohm *et al.* , 2007). Recently the epigenetic analysis of stem cells has started to unveil the basic circuitry of mammalian development (Bock & Lengauer, 2008), (Surani *et al.* , 2007).

Epigenetic factors are clearly important in gene expression and so it is clear that the physical locations of genes on the genome will have an impact on gene expression.

6.1.2 Gene Location

This chapter looks into the nature of genes with respect to their physical location on the genome. This is an area of epigenetics which has hitherto attracted very little attention.

The frequent pattern mining program, WARMR (see Section 2.5.2) was used to search for patterns in gene location. WARMR is a first order pattern mining algorithm and is required specifically for the analyses described in Sections 6.5 and 6.6. The analyses performed in Sections 6.3 and 6.4 could easily be performed by simpler relational mining algorithms such as Apriori. Furthermore, in some instances the data mining, such as in Section 6.3, could be achieved easily using non-optimised Prolog programs. However, WARMR is capable of, and was in fact used for performing the analyses required in all sections with no noticeable computational overhead. In this way the results are all produced in files of the same format and can be further analysed by the same Prolog programs for each section.

As previously discussed, supporting Prolog programs were also required to refine the results produced by WARMR and this resulted in the construction of a data mining system designed to find significantly frequent patterns in the location of genes on the genome. The system is named SPD (Significant Pattern Discovery).

The SPD system is used to find significant patterns in individual genes and their attributes in Section 6.3. In Section 6.4 the system is employed in the discovery of patterns in the nature of neighbouring pairs of genes. The SPD system is used in its fullest capacity in Sections 6.5 and 6.6 where Monte Carlo methods are employed to establish a null hypothesis in gene locations. This method was used to determine the significance of discovered patterns in localized and dispersed clusters of genes classified by molecular function.

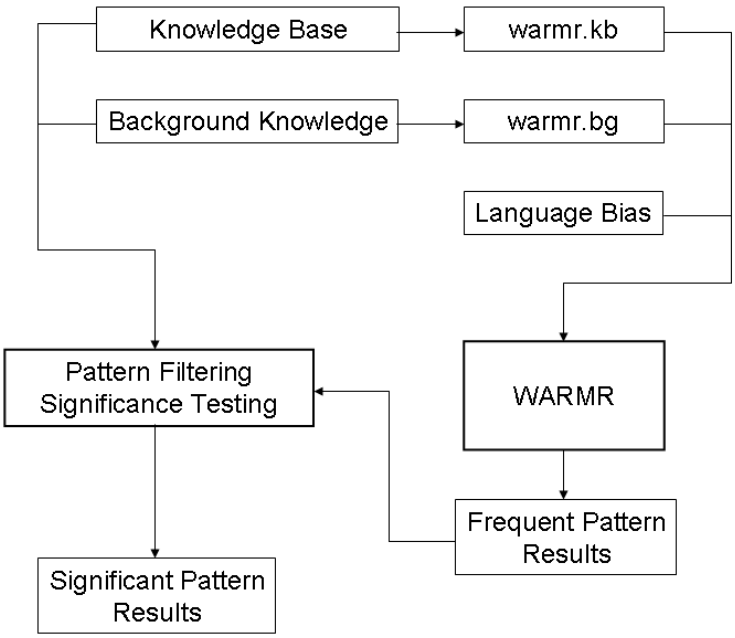

Figure 6.1: The SPD frequent pattern analysis system

6.2 The SPD System

The SPD (Significant Pattern Discovery) system is an extension of the WARMR frequent pattern mining program enabling the user to sort and filter the results produced by WARMR and to easily determine patterns of significant frequency.

Many preliminary frequent pattern mining experiments were conducted during this research, which have not been included in this thesis. These experiments clarified many inherent problems with frequent pattern mining and showed that a more flexible system than the Apriori approach (see Section 2.5.1) was required.

Frequent pattern mining is computationally expensive in both computer time and in computer memory usage. Furthermore, searches need to be carefully designed, because throwing all data available at a frequent pattern mining algorithm

without forethought will result in many uninformative patterns. Setting the frequency threshold high reduces the time taken for searches. However, due to the monotonicity of the frequency of patterns discovered by Apriori based search algorithms, potentially interesting patterns are lost. More simply, some patterns of significant frequency may be lost because they fall below the frequency threshold.

Total reliance on the WARMR frequent pattern mining algorithm alone was discarded in favour of the more versatile SPD pattern mining system capable of being adapted to suit the nature of different frequent pattern enquiries. The WARMR frequent pattern mining algorithm is still used as a component of the SPD system with the addition of a series of analytical programs used to analyse the results produced by WARMR. A block diagram of the SPD system can be seen in Figure 6.1. The analysis section of the SPD system utilized both Prolog and C++ programming languages. The output from WARMR is in Datalog format and therefore a Prolog based analysis system can immediately access and process the WARMR results without pre-processing or any external intervention. The C++ programming language was better suited for fast applications of the Monte Carlo procedure used in Section 6.6.

The SPD system has two principal functions:

1. Frequent pattern discovery.
2. Determination of the significance of discovered frequent patterns.

The SPD system requires two basic sources of data:

1. *Knowledge base*: all available data on genes.
2. *Background knowledge*: knowledge relevant to, but not specific to, genes.

The file system in the SPD system is specifically designed so that all changes or updates to data are made only to the knowledge base files and the background knowledge files. There is no need to change the WARMR specific files, which are labelled *warmr.kb* and *warmr.bg* in Figure 6.1. By using this method there is no need to track changes between the files required for the significance testing

performed using Prolog and the files required by WARMR for frequent pattern discovery. A further advantage in this method is that the *warmr.kb* file need consult only those knowledge base files required for the search by simply commenting out the unnecessary *consult*/1 predicates. This allows us to reduce the search space to only contain data relevant to specific searches. The frequent pattern searches are 'designed' by the language bias, which is a WARMR specific file (*warmr.s*).

6.2.1 Knowledge Base

The principle source of data for this work on *Saccharomyces cerevisiae* was the Saccharomyces Genome Database (SGD) from which the *GFF* file (s_cere_2a.gff) was downloaded on the 21st May 2007. GFF is a format for describing genes and other features associated with DNA, RNA and protein sequences. Using this file and supporting data from Gene Ontology (GO) (The Gene Ontology Consortium, 2000), fourteen files of specific gene data were extracted and these are listed in Table 6.1 along with the respective Datalog schema for the data in each file. These files are selectively consulted as required by Prolog programs, and also for the WARMR analysis by the WARMR specific knowledge base file (yeast.kb).

6.2.2 Background Knowledge

The background knowledge in the file *yeast_bg.pl* is more general knowledge required for the analysis and not necessarily specific in this case to *Saccharomyces cerevisiae*. For example, a definition of small, medium and large gene lengths is provided from the data discussed in Chapter 3. This knowledge could apply to any genome. The background knowledge used in each part of this research is discussed in more depth in each corresponding section. The background knowledge file is consulted by the WARMR specific background knowledge file *warmr.bg*.

File	Schema
yst_g_gene.pl	gene(gene id).
yst_g_strand.pl	strand(gene id, strand(w/c)).
yst_g_chromo.pl	chromosome(gene id, chromosome id).
yst_g_class.pl	class(gene id, GO class).
yst_g_gaplength.pl	gap_length(gene id, gene id, length).
yst_g_length.pl	gene_length(gene id, length).
yst_g_next.pl	is_next_to(gene id, gene id).
yst_g_notes.pl	sgd_notes(gene id, notes).
yst_g_start.pl	start_locus(gene id, location).
yst_g_centre.pl	centre_point(gene id, location).
yst_g_class_1.pl	class_1(gene id, GO class).
yst_g_class_2.pl	class_2(gene id, GO class).
yst_g_class_3.pl	class_3(gene id, GO class).
yst_g_class_4.pl	class_4(gene id, GO class).

Table 6.1: The files and formats for the gene location knowledge base for *Saccharomyces cerevisiae*. Each *location* is specifically the base pair location given by the number of nucleotides or base pairs from the 5' end of the W strand and applies to the genes on both W and C strands. Each *length* is given by the number of nucleotides.

6.3 Individual Gene Analysis

6.3.1 Introduction

The first part of the research of frequent patterns in gene location was to perform a statistical analysis to determine the relative frequencies or probabilities of individual genes with respect to their attributes. This can be done using WARMR since the required statistics are returned in the results for level 2 giving the relative frequency of the gene attributes, including the molecular function classes of genes. The attributes used in this search are the molecular function classes, gene length, on which strand of the DNA and on which chromosome the genes are located. This can be used later to determine expectations for the frequency of discovered patterns in order that the significance of the discovered patterns can be calculated.

Although, at this stage the main focus is on the level 2 search, the WARMR search was allowed to continue through all levels where frequent patterns occurred.

6.3.2 Method

Language bias

A listing for the language bias required for the statistical analysis of genes and their attributes is given in Figure 6.2. Using this language bias, the results produced by WARMR at level 2 should provide the following:

- Relative frequency/probability of genes on each strand.
- Relative frequency/probability of genes in each molecular function class.
- Relative frequency/probability of genes in each of the four length categories.
- Relative frequency/probability of genes on each chromosome.

The frequent pattern search was performed using WARMR with the language bias described above (see Table 6.2) and the knowledge base and background knowledge

```
warmode_key(gene(-GeneA)).

rmode(1:strand(+GeneA, w)).
rmode(1:strand(+GeneA, c)).
rmode(1:class_1(+GeneA, #[go:3774,go:3824,go:4871,go:5198,go:5215,
    go:5488,unknown,go:16209,go:30188,go:30234,go:30528,go:30533,
    go:31386,go:31992, go:42056,go:45182,go:45499,go:45735])).
rmode(1:small_gene(+GeneA)).
rmode(1:med_sml_gene(+GeneA)).
rmode(1:medium_gene(+GeneA)).
rmode(1:large_gene(+GeneA)).
rmode(1:chromosome(+GeneA, #)).
```

Figure 6.2: Listing for the language bias file required for a statistical analysis of genes and their attributes using WARMR.

described in Sections 6.2.1 and 6.2.2.

6.3.3 Results

The results from the WARMR level 2 search are given in Table 6.2. These results indicate that:

1. The percentage of genes on each strand of the DNA throughout the entire genome is approximately 50% on each strand as would be expected from a uniform probability distribution.

2. The top three results for the most frequent number of genes in each molecular function class indicate that almost 44% are unknown; 26% are involved in catalytic activity (GO:3824) and roughly 11% are associated with binding activity (GO:5488). The remaining results for molecular function class frequencies are all in the order of 5% or less.

3. There is a disparity in the frequencies of the 4 categories of gene length. This is not significant because the boundaries on each category were not chosen to have roughly equivalent frequencies (see Chapter 3).

4. The frequencies of genes on each chromosome are in proportion with the

Clause	Probability
gene(A),strand(A,w)	0.504312301407172
gene(A),strand(A,c)	0.495687698592828
gene(A),class_1(A,go:3774)	0.00211832349825995
gene(A),class_1(A,go:3824)	0.258284157966409
gene(A),class_1(A,go:4871)	0.00680889695869269
gene(A),class_1(A,go:5198)	0.0485701316386745
gene(A),class_1(A,go:5215)	0.0556816462399758
gene(A),class_1(A,go:5488)	0.114692086548646
gene(A),class_1(A,unknown)	0.436223331820245
gene(A),class_1(A,go:16209)	0.00257224996217279
gene(A),class_1(A,go:30188)	0.00121047057043426
gene(A),class_1(A,go:30234)	0.022545014374338
gene(A),class_1(A,go:30528)	0.0434256317143289
gene(A),class_1(A,go:31386)	0.00136177939173854
gene(A),class_1(A,go:45182)	0.00650627931608413
gene(A),small_gene(A)	0.0676350431230141
gene(A),med_sml_gene(A)	0.0630957784838856
gene(A),medium_gene(A)	0.434105008321985
gene(A),large_gene(A)	0.435164170071115
gene(A),chromosome(A,chrI)	0.017703132092601
gene(A),chromosome(A,chrII)	0.0689968225147526
gene(A),chromosome(A,chrIII)	0.0276895142986836
gene(A),chromosome(A,chrIV)	0.126645483431684
gene(A),chromosome(A,chrIX)	0.036465425934332
gene(A),chromosome(A,chrMito)	0.0042366469965199
gene(A),chromosome(A,chrV)	0.0490240581025874
gene(A),chromosome(A,chrVI)	0.0213345438039038
gene(A),chromosome(A,chrVII)	0.0883643516417007
gene(A),chromosome(A,chrVIII)	0.0485701316386745
gene(A),chromosome(A,chrX)	0.0602209108791043
gene(A),chromosome(A,chrXI)	0.0526554698138902
gene(A),chromosome(A,chrXII)	0.087456498713875
gene(A),chromosome(A,chrXIII)	0.0764109547586624
gene(A),chromosome(A,chrXIV)	0.0658193372673627
gene(A),chromosome(A,chrXV)	0.0904826751399607
gene(A),chromosome(A,chrXVI)	0.0773188076864881
gene(A),chromosome(A,2-micron)	0.000605235285217128

Table 6.2: Individual gene analysis results (Warmr Level 2) showing the probabilities of genes of the attributes given in the second term in each clause. These probabilities are used to determine the expectations for superclauses involving multiples of attribute terms. See Table 6.15 for descriptions of the molecular functions designated by the *go:* numbers.

length of each chromosome, which is not an unexpected result.

The WARMR search continued up to level 7 and the number of results produced was enormous. These results are available on file, but have not been included in this thesis. Using the WARMR query section of the SPD system the WARMR results file was filtered to remove non-relevant and uninformative results. Further analysis revealed two possibly interesting results at level 3.

The first interesting result is a single clause:

$$gene(A), strand(A, c), chromosome(A, chrMito), 0.0$$

This result is interesting because it indicates that there are no genes on the Crick strand of the mitochondria DNA. This phenomenon has been reported by Foury *et al.* (1998).

The second interesting result is revealed in the lengths of genes classified by molecular function. These results are reported in Table 6.3. From this table we can see that genes associated with catalytic activity (GO:3824) are the most numerous having 1707 examples, but only 17 are 360 bp or less in length. The *small_gene* and *med_sml_gene* sets for all genes annotated with molecular function are much smaller than the *medium_gene* and *large_gene* sets, but the results obtained are not in proportion with the sizes of the sets. The clause describing small and medium small genes of catalytic activity is:

$$class\_1(A, go : 3824) \wedge (small\_gene(A) \vee med\_sml\_gene(A))$$

The expectation E for this rule is given by:

$$E = P_{class} \times (P_X + P_Y) \times F_{all\_genes}$$

Where P_{class} is the probability of $class\_1(A, go : 3824)$; P_X is the probability of $small\_gene(A)$; P_Y is the probability of $med\_sml\_gene(A)$ and F_{all_genes} is the frequency of all genes in the data. The probabilities can be read straight from Table 6.2 giving the expectation E:

$$E = 0.2583 \times (0.0676 + 0.0631) \times 6609 = 223.12$$

Allowing for rounding errors this means we should expect to find 223 small and medium-small genes of catalytic activity, so the probability of finding only 17 is very low and so this should be considered significant. This probability in the occurrence of small genes prevails throughout all classified genes with the possible exception of *protein tag* genes (GO:31386) and suggests that classified genes tend to have longer nucleotide sequences than average. However, looking at the result for genes of unknown molecular function we find that of the total of 447 of all small genes, 375 are unknown and similarly of all 417 medium-small genes, 375 are unknown. This means that 84% of all small genes and 90% of all medium-small genes are unknown.

6.3.4 Analysis

The work described in this section has provided useful information on the probabilities of genes with certain attributes, which are given in Table 6.2. This information is necessary in determining significance of future results described in the following sections.

The result showing that all genes in the DNA of the mitochondrion in the cells of *S. cerevisiae* are on the Watson strand and none on the Crick strand, is merely a curiosity in this research and has been previously reported (Foury *et al.* , 1998). The main concern is with the location of genes and their possible interaction in the genome of *S. cerevisiae*. The genes in the mitochondrion are not able to interact with genes in the nucleus so, in this respect, the mitochondrion can be considered a separate organism.

Regarding the final result of interest, although it is entirely conceivable that gene length and molecular function may be related, with such a significant number of genes shorter than 360 bp with unknown molecular function, no conclusion can be drawn. It is curious that research into gene identification and classification favours larger genes.

Clause	Probability	Examples
gene(A),class_1(A,go:3774),medium_gene(A)	0.00030	2.0
gene(A),class_1(A,go:3774),large_gene(A)	0.00181	12.0
gene(A),class_1(A,go:3824),small_gene(A)	0.00121	8.0
gene(A),class_1(A,go:3824),med_sml_gene(A)	0.00136	9.0
gene(A),class_1(A,go:3824),medium_gene(A)	0.10153	671.0
gene(A),class_1(A,go:3824),large_gene(A)	0.15418	1019.0
gene(A),class_1(A,go:4871),small_gene(A)	0.00030	2.0
gene(A),class_1(A,go:4871),med_sml_gene(A)	0.00015	1.0
gene(A),class_1(A,go:4871),medium_gene(A)	0.00227	15.0
gene(A),class_1(A,go:4871),large_gene(A)	0.00408	27.0
gene(A),class_1(A,go:5198),small_gene(A)	0.00197	13.0
gene(A),class_1(A,go:5198),med_sml_gene(A)	0.00121	8.0
gene(A),class_1(A,go:5198),medium_gene(A)	0.03435	227.0
gene(A),class_1(A,go:5198),large_gene(A)	0.01104	73.0
gene(A),class_1(A,go:5215),small_gene(A)	0.00272	18.0
gene(A),class_1(A,go:5215),med_sml_gene(A)	0.00076	5.0
gene(A),class_1(A,go:5215),medium_gene(A)	0.01861	123.0
gene(A),class_1(A,go:5215),large_gene(A)	0.03359	222.0
gene(A),class_1(A,go:5488),small_gene(A)	0.00303	20.0
gene(A),class_1(A,go:5488),med_sml_gene(A)	0.00197	13.0
gene(A),class_1(A,go:5488),medium_gene(A)	0.04373	289.0
gene(A),class_1(A,go:5488),large_gene(A)	0.06597	436.0
gene(A),class_1(A,unknown),small_gene(A)	0.05674	375.0
gene(A),class_1(A,unknown),med_sml_gene(A)	0.05674	375.0
gene(A),class_1(A,unknown),medium_gene(A)	0.20517	1356.0
gene(A),class_1(A,unknown),large_gene(A)	0.11757	777.0
gene(A),class_1(A,go:16209),medium_gene(A)	0.00212	14.0
gene(A),class_1(A,go:16209),large_gene(A)	0.00045	3.0
gene(A),class_1(A,go:30188),medium_gene(A)	0.00060	4.0
gene(A),class_1(A,go:30188),large_gene(A)	0.00060	4.0
gene(A),class_1(A,go:30234),small_gene(A)	0.00106	7.0
gene(A),class_1(A,go:30234),med_sml_gene(A)	0.00030	2.0
gene(A),class_1(A,go:30234),medium_gene(A)	0.00777	51.0
gene(A),class_1(A,go:30234),large_gene(A)	0.01347	89.0
gene(A),class_1(A,go:30528),small_gene(A)	0.00015	1.0
gene(A),class_1(A,go:30528),med_sml_gene(A)	0.00015	1.0
gene(A),class_1(A,go:30528),medium_gene(A)	0.01468	97.0
gene(A),class_1(A,go:30528),large_gene(A)	0.02845	188.0
gene(A),class_1(A,go:31386),small_gene(A)	0.00030	2.0
gene(A),class_1(A,go:31386),med_sml_gene(A)	0.00030	2.0
gene(A),class_1(A,go:31386),medium_gene(A)	0.00077	5.0
gene(A),class_1(A,go:45182),small_gene(A)	0.00015	1.0
gene(A),class_1(A,go:45182),med_sml_gene(A)	0.00015	1.0
gene(A),class_1(A,go:45182),medium_gene(A)	0.00227	15.0
gene(A),class_1(A,go:45182),large_gene(A)	0.00393	26.0

Table 6.3: Frequencies of genes classified by molecular function and gene length attributes. See Table 6.15 for descriptions of the molecular functions designated by the *go:* numbers.

6.4 Neighbouring Pairs

6.4.1 Introduction

This section describes research on neighbouring pairs of genes. There is evidence that neighbouring pairs of genes co-operate in prokaryotes to form simple biochemical networks (Warren & ten Wolde, 2004a). These co-operating gene pairs are characterized by having overlapping regulatory domains and this results in what has been described as *correlated* and *anti-correlated* gene expression. Correlated gene expression is where the product of one gene promotes the expression of another gene and anti-correlated gene expression inhibits the expression of another gene. Evidence for correlated and anti-correlated gene expression in eukaryotes has been reported (Willy & Kobayashi, 2000) (Szallasi, 2001). Evidence for co-operating gene pairs in *S. cerevisiae* may be found in the attributes of neighbouring genes and the lengths of the gaps between them.

6.4.2 Method

Language bias

The language bias for this search in Figure 6.3, is similar to the one used for the individual gene analysis, but with the following modifications. The neighbouring pair specific predicate $is\_next\_to/2$ and the gap length predicates $neg\_gap/2$, $small\_gap$, $medium\_gap/2$ and $large\_gap$ are added. The predicate $is\_next\_to/2$ takes $GeneA$ and returns $GeneB$ if $GeneB$ is the immediate neighbour of $GeneA$ on either strand. Note that $GeneA$ is downstream of $GeneB$ in the knowledge base. The gap length predicates are defined in the background knowledge according to the information given in Chapter 3. A further modification is that the predicate $chromosome/2$ is not required, since no gene can have a neighbour on a different chromosome. Removing unnecessary queries from the language bias improves efficiency.

The frequent pattern search was performed using WARMR with the language bias

```
rmode(1:is_next_to(+GeneA, -GeneB)).
rmode(1:neg_gap(+GeneA, +GeneB)).
rmode(1:small_gap(+GeneA, +GeneB)).
rmode(1:medium_gap(+GeneA, +GeneB)).
rmode(1:large_gap(+GeneA, +GeneB)).
rmode(2:strand(+GeneA, w)).
rmode(2:strand(+GeneA, c)).
rmode(2:class_1(+GeneA, #[go:3774,go:3824,go:4871,go:5198,go:5215,
    go:5488,go:16209,go:30188,go:30234,go:30528,go:30533,go:31386,
    go:31992,go:42056,go:45182,go:45499,go:45735])).
rmode(2:small_gene(+GeneA)).
rmode(2:medium_gene(+GeneA)).
rmode(2:large_gene(+GeneA)).
rmode(2:med_sml_gene(+GeneA)).
```

Figure 6.3: Listing for the language bias file required for frequent pattern discovery in neighbouring pairs of genes and their attributes using WARMR.

described above and the knowledge base and background knowledge described in Sections 6.2.1 and 6.2.2.

Background knowledge

There are four ways in which each neighbouring pair can be related by their location and by their direction of transcription. These ways have been labelled w-sequent, convergent, divergent and c-sequent, where the w in w-sequent represents the Watson strand and the c in c-sequent represents the Crick strand. The w-sequent neighbours have both genes on the Watson strand and are transcribed downstream. This relation is given by the clause:

$$w\_sequent(A, B) : -is\_next\_to(A, B), strand(B, w), strand(A, w)$$

Similarly, the c-sequent neighbours have both genes on the Crick strand and clearly, they must be transcribed upstream. This relation is given by:

$$c\_sequent(A, B) : -is\_next\_to(A, B), strand(B, c), strand(A, c)$$

Convergent neighbours have genes on opposite strands where the gene on the Crick strand is downstream of the gene on the Watson strand so that the direction of

transcription converges towards the gap between the neighbouring genes. This relation is given by:

$$convergent(A, B) : -is\_next\_to(A, B), strand(B, w), strand(A, c)$$

Divergent neighbours have genes on opposite strands where the gene on the Crick strand is upstream of the gene on the Watson strand so that the direction of transcription diverges away from the gap between the neighbouring genes. This relation is given by:

$$divergent(A, B) : -is\_next\_to(A, B), strand(A, w), strand(B, c)$$

These four clauses are added to the background knowledge.

6.4.3 Results

It can be seen from the results in Table 6.4 that there are more convergent and divergent neighbouring pairs occurring with a frequency at about 55%. We would expect this figure to be about 50% in a uniform probability distribution. However, the interesting results can be seen in Table 6.5. Divergent and convergent neighbours overlap four times more frequently than sequent neighbours as indicated by the results for the $neg\_gap/2$ term. Looking further at the $small\_gap/2$ term we see that convergent neighbours are significantly more frequent. Neighbours with medium sized gaps between them all have a similar frequency with the exception of convergent neighbours whose frequency is elevated by 28% over the average of the other three types. For large gap lengths between neighbours there is a notable increase in the number of divergent neighbours and a surprising decrease in the number of convergent neighbours. In light of these results it was thought that graphs of the frequencies of gap lengths might be of further interest.

From the frequency plots in Figure 6.4 it can be seen clearly that the frequency distribution of gap lengths in the w-sequent pairs and the c-sequent pairs is roughly the same with the peak for w-sequent pairs occurring at ≈ 300 bp and the peak for c-sequent pairs occurring at ≈ 250 bp. There is a significant number of negative gap lengths for convergent and divergent pairs, which is interesting because they

Neighbouring pairs	Rel. freq.
w_sequent(A,B)	0.23059
convergent(A,B)	0.27311
divergent(A,B)	0.27235
c_sequent(A,B)	0.22121

Table 6.4: Relative frequency of the four types of neighbouring pairs related by locational and transcription direction.

Level	No.	Pair type defn.	Rel. freq.
5	1	w_sequent(A,B),neg_gap(B,A)	0.01029
5	2	convergent(A,B),neg_gap(B,A)	0.03465
5	28	divergent(A,B),neg_gap(B,A)	0.04554
5	55	c_sequent(A,B),neg_gap(B,A)	0.00756
5	219	w_sequent(A,B),small_gap(B,A)	0.01528
5	220	convergent(A,B),small_gap(B,A)	0.07369
5	251	divergent(A,B),small_gap(B,A)	0.01210
5	275	c_sequent(A,B),small_gap(B,A)	0.01574
5	490	w_sequent(A,B),medium_gap(B,A)	0.07747
5	491	convergent(A,B),medium_gap(B,A)	0.10017
5	523	divergent(A,B),medium_gap(B,A)	0.07429
5	555	c_sequent(A,B),medium_gap(B,A)	0.08261
5	841	w_sequent(A,B),large_gap(B,A)	0.12755
5	842	convergent(A,B),large_gap(B,A)	0.06461
5	874	divergent(A,B),large_gap(B,A)	0.14041
5	907	c_sequent(A,B),large_gap(B,A)	0.11530

Table 6.5: Relative frequencies (Rel. freq.) of different gap lengths between the four locational types (Pair type defn.) of neighbouring pairs. The level refers to the WARMR level and the number (No.) refers to the line number of the clause in the WARMR frequent queries file.

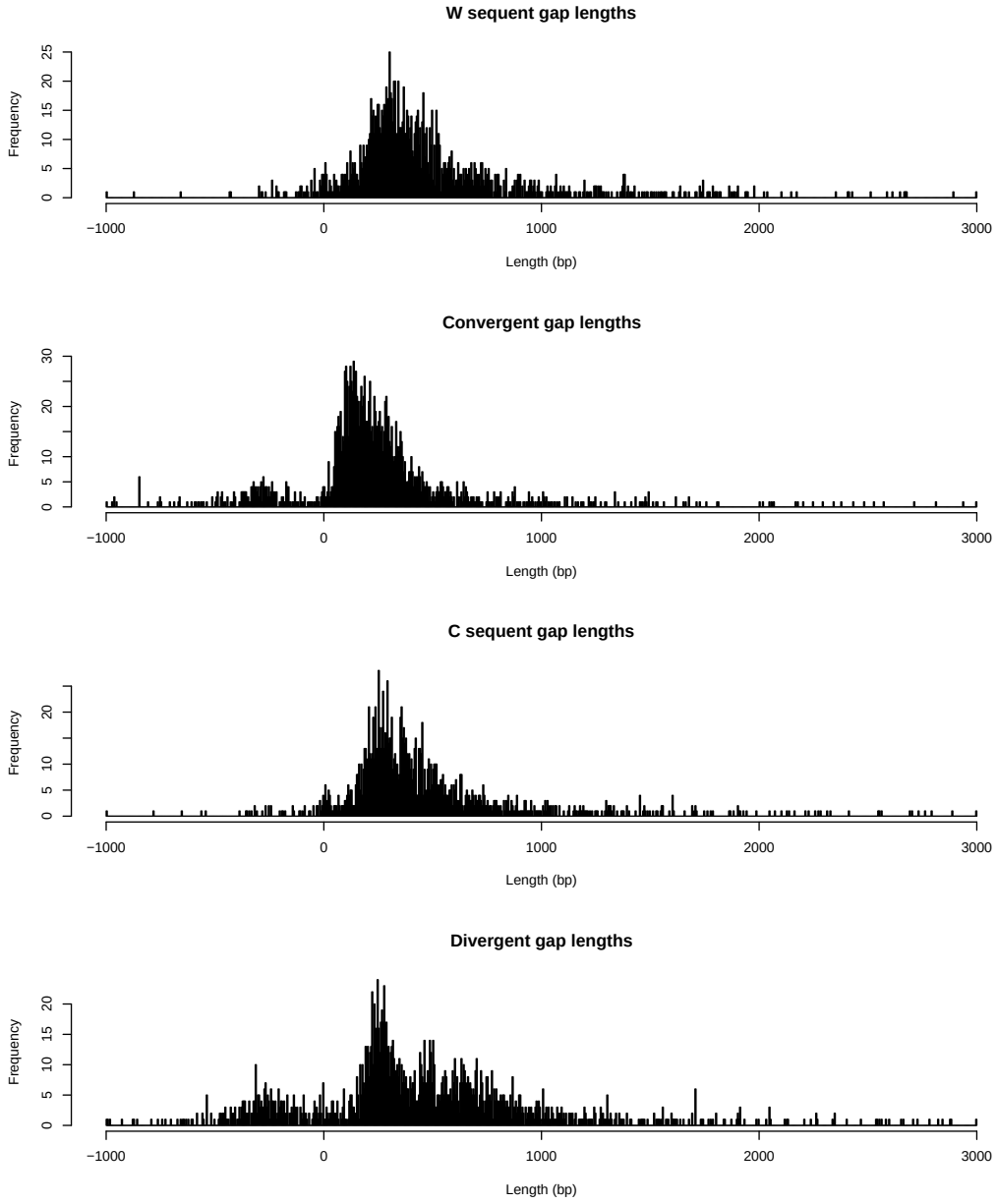

Figure 6.4: Frequency plots of the inter gene gap lengths of neighbouring pairs of the four transcription direction types; w-sequent, convergent, c-sequent and divergent.

both have local maxima at ≈ -250 bp. The convergent pairs have a higher frequency of short gap lengths than any other type and there is a peak at ≈ 100

bp. By contrast to the convergent pairs, the divergent pairs tend to have higher relative frequencies of longer gap lengths and a peak of ≈ 250 bp, similar to the sequent pairs. Also of possible interest is a secondary peak at ≈ 500 bp, which is twice the length of the most frequent gap lengths.

Class A	Class B	Examples	Expectation	Pmf
go:5198	go:5198	30.0	15.59	0.00379484
go:16209	go:16209	1.0	0.04	0.0400002
go:3774	go:30234	2.0	0.32	0.0511972
go:31386	go:4871	1.0	0.06	0.0600005
go:3774	go:4871	1.0	0.10	0.100002
go:5198	go:3824	64.0	82.91	0.106225
go:16209	go:3824	9.0	4.39	0.107706
go:45182	go:3824	18.0	11.11	0.129862
go:30234	go:4871	3.0	1.01	0.169991
go:5488	go:31386	3.0	1.03	0.176792

Table 6.6: The molecular function classes of neighbouring pairs for the 10 most significant results (Full results in Appendix B). See Table 6.15 for descriptions of the molecular functions designated by the *go:* numbers.

6.4.4 Neighbouring Molecular Function Classes

The 10 most significant results for analysis of neighbouring pairs and their respective molecular function are given in Table 6.6. These results show that only the top single result is of any real significance. Neighbouring pairs of genes of *structural molecule activity* (GO:5198) are more frequent than expected with a probability of 0.0038.

6.4.5 Analysis

The main discovery from this work is in the nature of gap lengths between neighbouring pairs. There is a clear difference between the distribution of divergent neighbouring pair gap lengths and the distribution of gap lengths for the remaining three types: w-sequent, c-sequent and convergent.

The secondary result indicating a significance in the number of neighbouring genes of structural molecule activity is most likely due to tandem duplication.

6.5 Clusters of Heterogeneous Gene Function

6.5.1 Introduction

The research discussed in this section and the next is motivated from the findings in Chapter 5, where we found unknown elements of order in the locations of genes classified by molecular function in the genome of *A. thaliana*. Although we are investigating the genome of *S. cerevisiae* in this chapter, we expect that the more in depth analysis in gene location described here may shed some light on previous findings.

The analysis in this section is concerned with regions of specific size around a gene of a specific class on the genome of *Saccharomyces cerevisiae* to determine if genes from other classes are frequently located nearby. Put simply, if we find a gene of, for example class A, do we often find genes of class B or C nearby? From the background research presented in Chapter 2 we might expect that, for genes classified by molecular function, certain classes will be more local to others than we would expect from a random distribution. This may imply a localised influence or co-operation between certain molecular function classes of genes.

From the research presented in Chapter 5 we know that there are likely to be frequent patterns in genes of homogeneous molecular function due to the proliferation of genes through tandem duplication. Clusters of genes of heterogeneous molecular function are more interesting because, should they exist, it is likely that they have been brought together through natural selection to serve a functional benefit to the host organism. In this case, it would be interesting to discover which classes are brought together. However, even with only 13 molecular function classes annotated at GO level 1 in the genome of *S. cerevisiae*, there could be a potential for $> 13^{10}$ candidate queries describing all patterns of up to 10 genes of different molecular function.

The solution is to search for increasingly larger patterns of genes of different molecular function without actually specifying the molecular function class. Effectively, we are searching for clusters of genes of heterogeneous molecular function. If we

do not find significant patterns in this search, then we will not find significant patterns in any specific molecular function search.

6.5.2 Method

Language Bias Settings

```
warmode_key(location(-Gene, -L)).
rmode(1: class1(+Gene, -Class)).
rmode(close_to_class1(+Class, \ClassA, +Gene)).

typed_language(yes).
type(close_to_class1(class,class,gene)).
type(class1(gene,class)).
type(location(gene,location)).
```

Figure 6.5: Listing for the language bias required for the discovery of frequent patterns in clusters of heterogeneous gene function.

The language bias required in this experiment is listed in Figure 6.5. It is a simple search for all possible genes of different functional class, which are local to the subject gene. The size of the local region is defined in the background knowledge.

Background Knowledge

The background knowledge required for this search is just one predicate function named *close_to_class*1/3, which is detailed in the listing in Figure 6.6. This predicate function returns the molecular function class ($ClassA$) of a given gene ($GeneA$) and another class ($ClassB$), being the molecular function class of any gene located within the specified region ($\pm 10,000bp$ in this example) surrounding the given query gene ($GeneA$). Successive function calls to this predicate function will return all classes within the specified region. Note that class $go : 5554$ designates genes of presently unknown molecular function and so these genes are omitted from the frequent pattern search.

```
close_to_class1(ClassA, ClassB, GeneA):-
    class1(GeneA, ClassA),
    class1(GeneB, ClassB),
    not(ClassB = go:5554),
    not(ClassA = go:5554),
    location(GeneA, B),  /* B & C are bp loci */
    location(GeneB, C),
    J is B - C,
    J < 10000,
    J > -10000,
    J =\= 0,
    chromosome(GeneA, X),
    chromosome(GeneB, Y),
    X = Y.
```

Figure 6.6: Listing for the *close_to_class*1 predicate function in the background knowledge file used in the frequent pattern mining search performed using WARMR.

The frequent pattern mining search was performed using WARMR with the language bias listed in Figure 6.5 and the background knowledge listed in Figure 6.6

Permutation Testing

For the previous experiments described in this chapter it has been possible to ascertain the significance of the results using traditional statistical methods. The significance of the patterns found in clusters of genes with heterogeneous molecular function is not easy to determine in this way because the molecular function classifications of each gene in the patterns discovered are unknown; we simply know that they are all different. So instead we can establish a null hypothesis in the locations of genes using Monte Carlo methods and draw a comparison with that null hypothesis. This technique is similar to the one used to determine the significance in the clustering and uniformity in the locations of genes of *A. thaliana* (see Chapter 5). The frequent patterns discovered in the WARMR search on the original gene sequence are isolated and then applied to a series of 1000 gene sequences where the locations of member genes have been scrambled. The frequency

of the query pattern is recorded for each one of the 1000 trials and from this data we calculate the mean and the standard deviation and see how this compares to the original pattern frequency. The significance can be determined quite easily using this method by sorting all 1000 results from the Monte Carlo trials in order of increasing frequency and then ranking the original pattern frequency with the sorted data.

6.5.3 Results

The results from the WARMR frequent pattern search are given in Figure 6.7 showing frequent patterns up to level 8. Level 8 describes a pattern or query where there are 7 different molecular function classes in an area 20,000 bp either side of the location of the subject gene designated by the variable A. The variable C represents the molecular function class of the subject gene and the variables D, E, F, G, H, I are the classes of genes in the designated area that are all different. The variable B is the base pair location of the subject gene A.

For further analysis into the significance of these frequent queries, patterns from two levels were chosen; level 5, where 48% of subject genes belong to a set of four genes of heterogeneous molecular function and level 6, where 37% of subject genes belong to a set of five genes of heterogeneous molecular function. Levels 5 and 6 were chosen because there were many examples of patterns at these levels and so there is a good chance that there will be a sufficient number of these patterns on the smaller chromosomes for the frequency results to have a reliable significance.

The results for the level 5 query are given in Table 6.7 for an area $\pm 10,000$ bp and Table 6.8 for an area $\pm 20,000$ bp. The results for the level 6 query are given in Table 6.9 for an area $\pm 10,000$ bp and Table 6.10 for an area $\pm 20,000$ bp.

```
level(1).
location(A,B).
1.0

level(2).
location(A,B),class1(A,C).
1.0

level(3).
location(A,B),class1(A,C),
close_to_class1(C,D,A),not(D=C).
0.542056074766355

level(4).
location(A,B),class1(A,C),
close_to_class1(C,D,A),not(D=C),
close_to_class1(C,E,A),not(E=C),not(E=D).
0.526479750778816

level(5).
location(A,B),class1(A,C),
close_to_class1(C,D,A),not(D=C),
close_to_class1(C,E,A),not(E=C),not(E=D),
close_to_class1(C,F,A),not(F=C),not(F=D),not(F=E).
0.482866043613707

level(6).
location(A,B),class1(A,C),
close_to_class1(C,D,A),not(D=C),
close_to_class1(C,E,A),not(E=C),not(E=D),
close_to_class1(C,F,A),not(F=C),not(F=D),not(F=E),
close_to_class1(C,G,A),not(G=C),not(G=D),not(G=E),not(G=F).
0.370716510903427

level(7).
location(A,B),class1(A,C),
close_to_class1(C,D,A),not(D=C),
close_to_class1(C,E,A),not(E=C),not(E=D),
close_to_class1(C,F,A),not(F=C),not(F=D),not(F=E)
close_to_class1(C,G,A),not(G=C),not(G=D),not(G=E),not(G=F),
close_to_class1(C,H,A),not(H=C),not(H=D),not(H=E),not(H=F),not(H=G).
0.152647975077882

level(8).
location(A,B),class1(A,C),
close_to_class1(C,D,A),not(D=C),
close_to_class1(C,E,A),not(E=C),not(E=D),
close_to_class1(C,F,A),not(F=C),not(F=D),not(F=E),
close_to_class1(C,G,A),not(G=C),not(G=D),not(G=E),not(G=F),
close_to_class1(C,H,A),not(H=C),not(H=D),not(H=E),not(H=F),not(H=G),
close_to_class1(C,I,A),not(I=C),not(I=D),not(I=E),not(I=F),not(I=G),not(I=H).
0.0186915887850467
```

Figure 6.7: Frequent pattern mining results for clusters of genes with heterogeneous molecular function showing the query pattern generated at each level and its relative frequency.

Chr	No. found	Mean	Std Dev	Rank (1000)	Ave rank
I	19	16.54	5.29	634–696	665.0
II	140	151.14	11.40	164–194	179.0
III	43	46.13	6.48	278–335	306.5
IV	252	282.61	16.10	33–34	33.5
V	68	92.11	9.02	5	5.0
VI	39	33.14	6.56	797–843	820.0
VII	202	201.12	12.01	499–528	514.5
VIII	108	95.46	9.434	896–917	906.5
IX	67	72.39	7.75	210–258	234.0
X	67	92.11	10.84	10–13	11.5
XI	128	119.28	9.63	812–838	825.0
XII	169	160.29	12.35	763–782	772.5
XIII	167	151.88	11.92	894–908	901.0
XIV	144	147.22	11.39	362–401	381.5
XV	171	174.68	13.34	371–387	379.0
XVI	194	185.09	11.95	761–784	772.5
All					481.7

Table 6.7: **Level 5, region** $\pm 10,000$ **bp.** Significance results for the frequency of the WARMR frequent query at level 5 (See listing in Figure 6.7), which is a pattern describing four different classes from GO level 1 in an area $\pm 10,000$ bp (approx 11 genes) either side of the location of each subject gene A. **Chr** represents the chromosome number in Roman numerals. **No. found** is the frequency of the pattern in the original gene sequence. The **Mean** and **Std Dev** represent the mean and the standard deviation of the pattern frequency in 1000 gene sequence permutations. The **Rank** shows where the original pattern frequency ranks alongside the ordered list of pattern frequencies in 1000 gene sequence permutations and a single central figure for this ranking is given in the column headed **Ave Rank**. The bottom line in the table gives an average rank for the whole genome of *S. cerevisiae*.

Chr	No. found	Mean	Std Dev	Rank (1000)	Ave rank
I	43	36.55	4.72	894–934	914.0
II	234	233.58	8.44	475–510	492.5
III	82	76.40	5.39	819–885	852.0
IV	434	445.19	13.05	179–195	187.0
V	148	154.50	7.82	195–226	210.5
VI	65	61.92	5.96	647–706	676.5
VII	314	308.08	8.66	726–768	747.0
VIII	155	153.62	8.92	528–563	545.5
IX	115	114.79	5.46	432–496	464.0
X	160	169.38	11.01	193–207	200.0
XI	182	185.06	7.05	303–334	318.5
XII	267	265.23	10.30	528–564	546.0
XIII	248	250.09	9.85	378–417	398.5
XIV	230	229.26	8.82	484–533	508.5
XV	295	297.03	11.84	390–420	405.0
XVI	269	274.27	8.28	218–258	238.0
All					481.5

Table 6.8: **Level 5, region** $\pm 20,000$ **bp.** Significance results for the frequency of the WARMR frequent query at level 5 (See listing in Figure 6.7), which is a pattern describing four different classes from GO level 1 in an area $\pm 20,000$ bp (approx 22 genes) either side of the location of each subject gene A. **Chr** represents the chromosome number in Roman numerals. **No. found** is the frequency of the pattern in the original gene sequence. The **Mean** and **Std Dev** represent the mean and the standard deviation of the pattern frequency in 1000 gene sequence permutations. The **Rank** shows where the original pattern frequency ranks alongside the ordered list of pattern frequencies in 1000 gene sequence permutations and a single central figure for this ranking is given in the column headed **Ave Rank**. The bottom line in the table gives an average rank for the whole genome of *S. cerevisiae*.

Chr	No. found	Mean	Std Dev	Rank (1000)	Ave rank
I	6	*4.88	*3.97	587–665	626.0
II	50	66.15	11.98	86–98	92.0
III	21	17.61	6.24	660–719	689.5
IV	104	119.07	15.75	158–174	166.0
V	18	35.42	9.30	24–31	27.5
VI	10	11.97	5.99	357–433	395.0
VII	100	93.22	13.32	681–710	695.5
VIII	44	38.99	9.38	679–719	699.0
IX	21	34.04	8.16	42–61	51.5
X	26	29.96	9.17	316–346	331.0
XI	55	52.31	10.01	357–433	395.0
XII	73	65.42	11.90	724–759	741.5
XIII	38	57.18	11.88	42–56	49.0
XIV	41	63.03	11.51	26–32	29.0
XV	82	65.05	12.40	905–916	910.5
XVI	88	86.14	13.15	538–565	551.5
All					403.1

Table 6.9: **Level 6, region** $\pm 10,000$ **bp.** Significance results for the frequency of the WARMR frequent query at level 6 (See listing in Figure 6.7), which is a pattern describing five different classes from GO level 1 in an area $\pm 10,000$ bp (approx 11 genes) either side of the location of each subject gene A. **Chr** represents the chromosome number in Roman numerals. **No. found** is the frequency of the pattern in the original gene sequence. The **Mean** and **Std Dev** represent the mean and the standard deviation of the pattern frequency in 1000 gene sequence permutations. The **Rank** shows where the original pattern frequency ranks alongside the ordered list of pattern frequencies in 1000 gene sequence permutations and a single central figure for this ranking is given in the column headed **Ave Rank**. The bottom line in the table gives an average rank for the whole genome of *S. cerevisiae* (* The distribution of these frequencies was asymmetric so the mean and standard deviation given are not accurate).

Chr	No. found	Mean	Std Dev	Rank (1000)	Ave rank
I	29	21.95	6.014	855–889	872.0
II	179	175.94	13.97	565–587	576.0
III	56	51.91	8.34	655–701	678.0
IV	289	326.39	20.35	33–34	33.5
V	86	107.95	11.53	25–27	26.0
VI	40	39.06	8.19	522–560	541.0
VII	230	242.97	15.04	188–200	194.0
VIII	119	105.32	12.13	858–881	869.5
IX	90	90.36	8.85	450–499	474.5
X	105	97.68	13.62	689–719	704.0
XI	157	141.71	12.30	888–903	895.5
XII	202	188.07	15.49	801–819	810.0
XIII	197	173.46	15.49	934–939	936.5
XIV	164	168.83	14.40	334–358	346.0
XV	210	201.79	18.00	659–670	664.5
XVI	220	214.35	14.94	614–644	629.0
All					578.1

Table 6.10: **Level 6, region** $\pm 20,000$ **bp.** Significance results for the frequency of the WARMR frequent query at level 6 (See listing in Figure 6.7), which is a pattern describing five different classes from GO level 1 in an area $\pm 20,000$ bp (approx 22 genes) either side of the location of each subject gene A. **Chr** represents the chromosome number in Roman numerals. **No. found** is the frequency of the pattern in the original gene sequence. The **Mean** and **Std Dev** represent the mean and the standard deviation of the pattern frequency in 1000 gene sequence permutations. The **Rank** shows where the original pattern frequency ranks alongside the ordered list of pattern frequencies in 1000 gene sequence permutations and a single central figure for this ranking is given in the column headed **Ave Rank**. The bottom line in the table gives an average rank for the whole genome of *S. cerevisiae*.

6.5.4 Analysis

The main and somewhat unexpected conclusion that can be drawn from the results is that, for any particular gene location, there are no frequent patterns of heterogeneity in the most general classification of molecular function (level 1) of localized genes. However, the next section covers work which revealed that there are frequent patterns of heterogeneity in the most general classification of molecular function (level 1) in genes that are more dispersed along the genome.

6.6 Transitive Sequences of Genes of Heterogeneous Molecular Function

6.6.1 Introduction

In this section the research covered is very similar to the work in the last section. However, the patterns differ in that now we are searching for patterns in genes of heterogeneous molecular function that follow in sequence, but may have many genes interspersed between pattern members. In this way we may find clustering of genes of different molecular function classes at certain points on the primary (single dimensional) structure of the genome. This may be an indication of physically localised clustering within the 3 dimensional structure of the genome.

6.6.2 Method

Language bias

```
warmode_key(gene(-GeneA)).
rmode(close_to_class(+GeneA, -Class, \ClassB, -GeneB, 10000)).
```

The language bias for this search shown in the listing above is quite simple. It uses only the predicate fact $gene(GeneA)$, identifying each subject gene to be counted, and the predicate function:

$$close\_to\_class(GeneA, ClassA, ClassB, GeneB, Region)$$

This predicate identifies the molecular function classification of object genes located within a predetermined region surrounding the subject gene. The language bias constrains the predicate to find only those object genes whose molecular function class is different to the subject gene and previously discovered object genes by using the backslash operator on the $\backslash ClassB$ term. In this way the WARMR program should find frequently occurring clusters of genes with heterogeneous molecular function. The predicate is defined in the background knowledge and is

discussed below. Although it is clear that this search can easily be performed using Prolog, it should be noted that any WARMR search could be performed using Prolog. The advantage of using WARMR is that it optimises the search.

Background knowledge

```
close_to_class(GeneA, ClassA, ClassB, GeneB, Region):-
    centre_point(GeneA, Locus_A),
    centre_point(GeneB, Locus_B),
    Distance is Locus_B - Locus_A,
    Distance < Region,
    Distance =\= 0,
    class_1(GeneA, ClassA),
    class_1(GeneB, ClassB),!.
```

Figure 6.8: Listing for the *close_to_class* predicate function in the background knowledge file used in the frequent pattern mining search performed using WARMR

The background knowledge defines the *close_to_class* rule with the predicate function:

$$close\_to\_class(GeneA, ClassA, ClassB, GeneB, Region)$$

where the object gene, $GeneB$, is considered *close to* the subject gene, $GeneA$, if and only if its location, given by the predicate fact:

$$centre\_point(GeneB, Locus\_B)$$

lies within a region surrounding the location of $GeneA$ given by:

$$centre\_point(GeneA, Locus\_A)$$

and also lies on the same chromosome. The region is defined by the variable $Region$, which is a length in base pairs or nucleotides either side of the location of $GeneA$. The predicate function returns the object gene and the molecular function classes of both the subject gene and the object gene. The cut symbol (!) ensures that given $GeneA$, once $GeneB$ is found $GeneA$ will not be searched again. This forces WARMR to unify $GeneA$ with the present literal $GeneB$ on

the next instantiation of the *close_to_class* predicate function. In this way we obtain a pattern of a sequence of genes whose molecular functions are all different classifications.

Permutation Testing

In order to determine the significance in the frequency of the discovered gene sequence patterns, permutation testing was used in the same way as performed in the last section. However, this work followed on from the previous section and an improvement was made to the permutation testing procedure. The permutation testing algorithm was a hybrid of Prolog and C++. The Prolog part was used to search for all occurrences of the clause detailing the gene sequence patterns discovered by WARMR in the original data and 1000 data sets of scrambled gene location. The C++ part was used to scramble the locations assigned to each gene and present a new file for the Prolog part to consult for each search. The files were transferred in virtual disk space for fast execution. This method reduced the permutation testing to a few hours whereas the previous all Prolog method used in the previous section (see Section 6.5) took days.

6.6.3 Results

The listings in Figures 6.9, 6.10 and 6.11 show the WARMR results for sequences of genes with heterogeneous molecular function for region $5,000$, $10,000$ and $20,000$ base pairs in length for all chromosomes. This shows that there are sequences of up to six genes of heterogeneous molecular function in regions $5,000$ and $10,000$, and up to seven genes in regions of $20,000$ base pairs. From further analysis it was found that for the smaller chromosomes I and VI, only sequences of four genes of different functional classes occurred in significant numbers. Therefore, further analysis was performed on frequent patterns of sequences of four genes of heterogeneous molecular function class. This pattern is given by:

```
gene(A),
close_to_class(A,B,C,D,Region),not(C=B),
```

```
gene(A),
close_to_class(A,B,C,D),not(C=B).
0.728971962616822

gene(A),
close_to_class(A,B,C,D),not(C=B),
close_to_class(D,E,F,G),not(F=B),not(F=C),not(F=E).
0.342679127725857

gene(A),
close_to_class(A,B,C,D),not(C=B),
close_to_class(D,E,F,G),not(F=B),not(F=C),not(F=E),
close_to_class(G,H,I,J),not(I=B),not(I=C),not(I=E),not(I=F),not(I=H).
0.130841121495327

gene(A),
close_to_class(A,B,C,D),not(C=B),
close_to_class(D,E,F,G),not(F=B),not(F=C),not(F=E),
close_to_class(G,H,I,J),not(I=B),not(I=C),not(I=E),not(I=F),not(I=H),
close_to_class(J,K,L,M),not(L=B),not(L=C),not(L=E),not(L=F),not(L=H),not(L=I),not(L=K).
0.0373831775700935

gene(A),
close_to_class(A,B,C,D),not(C=B),
close_to_class(D,E,F,G),not(F=B),not(F=C),not(F=E),
close_to_class(G,H,I,J),not(I=B),not(I=C),not(I=E),not(I=F),not(I=H),
close_to_class(J,K,L,M),not(L=B),not(L=C),not(L=E),not(L=F),not(L=H),not(L=I),not(L=K),
close_to_class(M,N,O,P),not(O=B),not(O=C),not(O=E),not(O=F),not(O=H),not(O=I),not(O=K),
not(O=L),not(O=N).
0.00934579439252336
```

Figure 6.9: WARMR result for region length 5,000 bp: Frequent pattern mining results for transitive sequences of genes with heterogeneous molecular function. Note that the *region* term has been edited out of the *close_to_class/5* atom.

```
close_to_class(D,E,F,G,Region),not(F=B),not(F=C),not(F=E),
close_to_class(G,H,I,J,Region),not(I=B),not(I=C),not(I=E),not(I=F),not(I=H).
```

The significance of transitive sequence patterns of four genes of different molecular function class was determined using permutation testing as described above. The results are presented in Tables 6.11, 6.12 and 6.13.

Chr	No. found	Mean	Std Dev.	Rank	E_p max
I	22	1.43	1.49	1000	9
II	68	11.05	4.29	1000	28
III	15	2.53	2.11	999	15
IV	121	19.77	5.79	1000	48
V	26	5.28	2.92	1000	18
VI	13	2.55	2.04	1000	12
VII	100	15.52	5.23	1000	34
VIII	42	7.28	3.52	1000	19
IX	31	5.23	3.09	1000	20
X	46	7.35	3.57	1000	22
XI	61	8.60	3.68	1000	25
XII	80	11.47	4.30	1000	29
XIII	75	12.32	4.39	1000	29
XIV	67	11.24	4.35	1000	27
XV	79	11.95	4.42	1000	30
XVI	76	13.10	4.50	1000	29

Table 6.11: Significance results for the frequency of patterns of transitive sequences of genes of heterogeneous molecular function for a region of $5,000$ bp. These are patterns where there is a gene B within the region of $5,000$ bp of the subject gene A; a gene C within the region of $5,000$ bp of gene B and a gene D within the region of $5,000$ bp of gene C. **Chr** represents the chromosome number in Roman numerals. **No. found** is the frequency of the pattern in the original gene sequence. The **Mean** and **Std Dev** represent the mean and the standard deviation of the pattern frequency in 1,000 gene sequence permutations. The **Rank** shows where the original pattern frequency ranks alongside the ordered list of pattern frequencies in 1,000 gene sequence permutations. The E_p **max** column is the maximum expectation figure found in 1,000 Monte Carlo trials, which demonstrates the extremity of some of the number of examples found.

Chr	No. found	Mean	Std Dev.	Rank	E_p max
I	23	1.29	1.75	1000	11
II	58	10.57	5.38	1000	29
III	17	3.12	2.79	999	17
IV	136	21.12	7.81	1000	51
V	47	5.17	3.42	1000	18
VI	11	2.37	2.36	995	12
VII	102	15.94	6.54	1000	40
VIII	35	7.69	4.45	1000	27
IX	31	5.88	3.91	1000	22
X	60	7.76	4.33	1000	30
XI	62	8.67	4.57	1000	29
XII	106	12.68	5.97	1000	43
XIII	89	13.56	5.81	1000	42
XIV	81	12.55	5.56	1000	31
XV	114	13.03	6.09	1000	42
XVI	99	13.24	5.90	1000	42

Table 6.12: Significance results for the frequency of patterns of transitive sequences of genes of heterogeneous molecular function for a region of $10,000$ bp. These are patterns where there is a gene B within the region of $10,000$ bp of the subject gene A; a gene C within the region of $10,000$ bp of gene B and a gene D within the region of $10,000$ bp of gene C. **Chr** represents the chromosome number in Roman numerals. **No. found** is the frequency of the pattern in the original gene sequence. The **Mean** and **Std Dev** represent the mean and the standard deviation of the pattern frequency in 1,000 gene sequence permutations. The **Rank** shows where the original pattern frequency ranks alongside the ordered list of pattern frequencies in 1,000 gene sequence permutations. The E_p **max** column is the maximum expectation figure found in 1,000 Monte Carlo trials, which demonstrates the extremity of some of the number of examples found.

Chr	No. found	Mean	Std Dev.	Rank	E_p max
I	12	0.58	1.40	998	13
II	49	8.56	6.17	1000	31
III	25	3.76	4.15	1000	23
IV	134	23.62	10.22	1000	64
V	61	4.09	3.85	1000	24
VI	10	1.27	2.12	992	14
VII	118	15.89	8.42	1000	53
VIII	38	5.99	5.26	1000	36
IX	45	4.64	4.26	1000	24
X	42	6.78	5.24	1000	30
XI	42	7.61	5.53	999	43
XII	81	12.79	7.42	1000	41
XIII	66	12.69	7.51	1000	53
XIV	52	11.56	7.02	1000	37
XV	71	13.49	7.58	1000	45
XVI	87	12.71	7.60	1000	50

Table 6.13: Significance results for the frequency of patterns of transitive sequences of genes of heterogeneous molecular function for a region of $20,000$ bp. These are patterns where there is a gene B within the region of $20,000$ bp of the subject gene A; a gene C within the region of $20,000$ bp of gene B and a gene D within the region of $20,000$ bp of gene C. **Chr** represents the chromosome number in Roman numerals. **No. found** is the frequency of the pattern in the original gene sequence. The **Mean** and **Std Dev** represent the mean and the standard deviation of the pattern frequency in 1,000 gene sequence permutations. The **Rank** shows where the original pattern frequency ranks alongside the ordered list of pattern frequencies in 1,000 gene sequence permutations. The E_p **max** column is the maximum expectation figure found in 1,000 Monte Carlo trials, which demonstrates the extremity of some of the number of examples found.

```
gene(A),
close_to_class(A,B,C,D),not(C=B).
0.731578151006204

gene(A),
close_to_class(A,B,C,D),not(C=B),
close_to_class(D,E,F,G),not(F=B),not(F=C),not(F=E).
0.404145861703737

gene(A),
close_to_class(A,B,C,D),not(C=B),
close_to_class(D,E,F,G),not(F=B),not(F=C),not(F=E),
close_to_class(G,H,I,J),not(I=B),not(I=C),not(I=E),not(I=F),not(I=H).
0.150703586019065

gene(A),
close_to_class(A,B,C,D),not(C=B),
close_to_class(D,E,F,G),not(F=B),not(F=C),not(F=E),
close_to_class(G,H,I,J),not(I=B),not(I=C),not(I=E),not(I=F),not(I=H),
close_to_class(J,K,L,M),not(L=B),not(L=C),not(L=E),not(L=F),not(L=H),not(L=I),not(L=K).
0.0481162051747617

gene(A),
close_to_class(A,B,C,D),not(C=B),
close_to_class(D,E,F,G),not(F=B),not(F=C),not(F=E),
close_to_class(G,H,I,J),not(I=B),not(I=C),not(I=E),not(I=F),not(I=H),
close_to_class(J,K,L,M),not(L=B),not(L=C),not(L=E),not(L=F),not(L=H),not(L=I),not(L=K),
close_to_class(M,N,O,P),not(O=B),not(O=C),not(O=E),not(O=F),not(O=H),not(O=I),not(O=K),
not(O=L),not(O=N).
0.0075654410652141
```

Figure 6.10: WARMR result for region length 10,000 bp: Frequent pattern mining results for transitive sequences of genes with heterogeneous molecular function. Note that the *region* term has been edited out of the *close_to_class/5* atom.

```
gene(A),
close_to_class(A,B,C,D),not(C=B).
0.735966106824028

gene(A),
close_to_class(A,B,C,D),not(C=B),
close_to_class(D,E,F,G),not(F=B),not(F=C),not(F=E).
0.372219700408534

gene(A),
close_to_class(A,B,C,D),not(C=B),
close_to_class(D,E,F,G),not(F=B),not(F=C),not(F=E),
close_to_class(G,H,I,J),not(I=B),not(I=C),not(I=E),not(I=F),not(I=H).
0.151460130125586

gene(A),
close_to_class(A,B,C,D),not(C=B),
close_to_class(D,E,F,G),not(F=B),not(F=C),not(F=E),
close_to_class(G,H,I,J),not(I=B),not(I=C),not(I=E),not(I=F),not(I=H),
close_to_class(J,K,L,M),not(L=B),not(L=C),not(L=E),not(L=F),not(L=H),not(L=I),not(L=K).
0.0476622787108488

gene(A),
close_to_class(A,B,C,D),not(C=B),
close_to_class(D,E,F,G),not(F=B),not(F=C),not(F=E),
close_to_class(G,H,I,J),not(I=B),not(I=C),not(I=E),not(I=F),not(I=H),
close_to_class(J,K,L,M),not(L=B),not(L=C),not(L=E),not(L=F),not(L=H),not(L=I),not(L=K),
close_to_class(M,N,O,P),not(O=B),not(O=C),not(O=E),not(O=F),not(O=H),not(O=I),not(O=K),
not(O=L),not(O=N).
0.0172492056286882

gene(A),
close_to_class(A,B,C,D),not(C=B),
close_to_class(D,E,F,G),not(F=B),not(F=C),not(F=E),
close_to_class(G,H,I,J),not(I=B),not(I=C),not(I=E),not(I=F),not(I=H),
close_to_class(J,K,L,M),not(L=B),not(L=C),not(L=E),not(L=F),not(L=H),not(L=I),not(L=K),
close_to_class(M,N,O,P),not(O=B),not(O=C),not(O=E),not(O=F),not(O=H),not(O=I),not(O=K),
not(O=L),not(O=N),
close_to_class(P,Q,R,S),not(R=B),not(R=C),not(R=E),not(R=F),not(R=H),not(R=I),not(R=K),
not(R=L),not(R=N),not(R=O),not(R=Q).
0.0033287940686942
```

Figure 6.11: WARMR result for region length 20,000 bp: Frequent pattern mining results for transitive sequences of genes with heterogeneous molecular function. Note that the *region* term has been edited out of the *close_to_class/5* atom.

Chr	X	Y
I	10,000	5,000
II	5,000	5,000
III	20,000	20,000
IV	10,000	20,000
V	20,000	5,000
VI	5,000	5,000
VII	20,000	10,000
VIII	5,000	10,000
IX	20,000	10,000
X	10,000	10,000
XI	10,000	10,000
XII	10,000	20,000
XIII	10,000	10,000
XIV	10,000	10,000
XV	10,000	20,000
XVI	10,000	10,000

Table 6.14: The most frequently populated region length for each chromosome. Column X lists the length of region out of the three lengths 5,000 bp, 10,000 bp and 20,000 bp that had the highest relative frequency of patterns. Column Y lists the length of regions out of the three lengths 5,000 bp, 10,000 bp and 20,000 bp that had the highest relative expectation of patterns.

6.6.4 Analysis

From the results in Tables 6.11, 6.12 and 6.12 it can seen clearly that the frequencies of all patterns rank very highly, indicating that transitive sequences of genes of heterogeneous molecular function occur more frequently than we would expect from genes distributed at random. This result prevails across all 16 chromosomes.

There is an interesting result in the relative number of patterns found in each region length for each chromosome. Intuitively, we would expect to find a higher frequency of patterns in larger regions such as the region of 20,000 base pairs. This is because there is a higher probablity of finding genes of different functional classes if there are more genes in the search. However, the results in Table 6.14 present two anomalies. Firstly, 9 out of 16 chromosomes have more patterns in a 10,000

base pair search space. This might suggest a cyclic nature to the distribution of different functional classes, but this assumption comes into question due to the second anomaly. The second anomaly is that the expectations for the pattern frequencies for each region length do not confirm the initial intuitive expectation and only 50% of the expectations correlate with the actual frequency of discovered patterns.

6.7 Discussion

The research described in this chapter is mainly concerned with the discovery of significant frequent patterns in the physical location of genes on the genome of *S. cerevisiae*. This was achieved using Datalog for the representation of the patterns and using both coventional statistical methods and the novel application of Monte Carlo methods to determine significance. These methods were incorporated into the SPD data mining system.

The selection of "interesting" rules or patterns is central to knowledge discovery in databases. Discovering patterns that are truly interesting to the user without using a lot of user-specified domain knowledge is an open problem. In this research interesting patterns are considered to be those patterns with frequencies that deviate significantly from the expectation. This is an objective approach in that it is essentially data-driven.

In general, data mining algorithms can easily discover too many patterns and many discovered patterns are either irrelevant or uninformative. The solution to this problem in this research was to use a subjective approach. Firstly, searches can be optimized to reduce time at the language bias stage by focussing on the sort of patterns that are most likely to be interesting. Secondly, the results can be sorted in order of frequency or significance enabling the user to reject all results in the ordered list of results whose frequency or significance parameters fall beyond a chosen threshold.

The SPD system effectively optimized searches and assisted the user in discovering interesting patterns.

6.7.1 Main Discoveries

The main findings from the work presented in this chapter are listed below:

1. 87% of genes of 360 base pairs or less in length have unknown molecular function.

2. There are more neighbouring gene pairs with diverging or converging directions of transcription than pairs with consequent directions of transcription.
3. There is a marked trend in the lengths of the gaps between neighbouring genes with peaks in the frequency of lengths between 250–300 base pairs.
4. There are no significant patterns in the molecular function classes of neighbouring genes with just one exception: neighbouring pairs of structural molecule activity are frequent.
5. Of all four types of neighbouring gene pairs typified by direction of transcription, the nature of the gap lengths between diverging pairs is very different to the nature of the gap lengths between pairs of the other three types.
6. The frequency of localized patterns of genes of different molecular function are no more frequent than we would find in a random, or locationally independent distribution of genes.
7. The frequency of patterns of dispersed genes of different classes of molecular function are very much more frequent than would be found in a locatioanlly independent distribution of genes.

The discrepencies in the results of the last two sections reveal a structure in the locations of genes of different functional classes. The nature of this structure can be analysed using conventional statistical and graphical methods, but using these methods alone, it is unlikely that this structure would ever have been revealed in the first place.

Class No.	Description
GO:0003774	Motor activity
GO:0003824	Catalytic activity
GO:0004871	Signal transducer activity
GO:0005198	Structural molecule activity
GO:0005215	Transporter activity
GO:0005488	Binding
GO:0016209	Anti oxidant activity
GO:0030188	Chaperone regulator activity
GO:0030234	Enzyme regulator activity
GO:0030528	Transcription regulator activity
GO:0031386	Protein tag
GO:0045182	Translation regulator activity
GO:0045735	Nutrient reservoir

Table 6.15: Level 1 molecular function classes for *S. cerevisiae*, showing the Gene Ontology identifier and a description of the molecular function.

6.8 Conclusions

One of the main problems with frequent pattern mining is the length of computer time required, especially with the large databases associated with bioinformatics. This problem is exacerbated by the use of Monte Carlo methods for the determination of significance.

Particular attention needs to be paid to the use of the frequency thresholds and the language bias in order to reduce computational time, but still discover complex significant patterns. This problem is universal throughout the field of knowledge discovery in databases.

The time required for significance testing was optimized by starting with a small number of Monte Carlo trials and only increasing the number of trials when significance was detected.

The initial approach used prior to the work described in this chapter was largely heuristic. This approach is extremely time consuming, which demonstated the need for a versatile significant pattern mining system. The system developed was named the SPD system. It improved the efficiency in the discovery of signifi-

cant frequent Datalog patterns in the location of genes in *S. cerevisiae* and has discovered new information.

Chapter 7

Pattern Mining: Molecular Phylogeny

7.1 Introduction

Molecular phylogenetics is an area of research which is particularly concerned with the evolution of organisms with respect to the nature of the DNA, genes or proteins belonging to each organism. There are other methods of determining phylogeny or taxonomy of organisms which are discussed in Chapter 2. In this chapter the generation of a broad database of phylogenetic trees of protein structure is discussed and effective query methods for use with databases of this nature are presented. This database is used to investigate methods for the determination of species phylogeny with high confidence, which is discussed in Chapter 8, and may also be used to investigate protein evolution, which is discussed in Section 7.8 in this chapter.

There are various formats for representing phylogenetic relations and phylogenetic trees such as the New Hampshire/Newick format discussed in Chapter 2. However, Nakleh *et al.* extol the virtues of using Datalog in the representation of phylogenetic relationships in a manifesto paper of 2003 (Nakhleh *et al.* , 2003), but appears not to have pursued this approach empirically. In this research we

adopt the Datalog approach, which simplifies the use of Prolog as the programming language for analyses.

This chapter describes the generation of a database of many phylogenetic trees where each tree represents a taxonomy of organisms determined by member protein sequences within the genome of each organism. The protein sequences used to create each tree belong to a specific protein structural classification (see Section 2.2.3). This database can be used to trace both the evolution of organisms and the evolution of classes of proteins within this database. More information can be obtained from this database using various queries detailed in Section 7.6, which are relevant to research in phylogenetics in general. Further work on using an example query to determine the closest relatives to *Homo sapiens* reveal some anomalies, which highlight some problems with the database and presents some possible solutions. Also in this section the database generated is evaluated to consider the possibility of using the data to build a phylogenetic consensus tree to establish an overall taxonomy of all member organisms.

We chose Swiss-Prot (Boeckmann *et al.* , 2003) for the protein sequence data and Superfamily (Gough *et al.* , 2001) for the protein classifications.

7.2 Swiss-Prot Protein Sequence Data

Swiss-Prot is a manually curated biological database of protein sequences, which was created in 1986 by Amos Bairoch. Swiss-Prot was chosen because it provides reliable protein sequences associated with a high level of annotation, a minimal level of redundancy and high level of integration with other databases. The Swiss-Prot data required for this part of the research was downloaded on 4th December 2007 from the FTP site at ExPASy[1] in a file named uniprot_sprot.fasta release 54.5, dated 13th November 2007. The file contains 289,473 protein sequences in FASTA format (Pearson, 1990) where the header contains the protein sequence

[1]The ExPASy (Expert Protein Analysis System) proteomics server of the Swiss Institute of Bioinformatics (SIB) is dedicated to the analysis of protein sequences and structures: www.expasy.org/.

identifier, data on the protein function and the organism from which the sequence was obtained.

The data downloaded from Swiss-Prot was first pre-processed to re-annotate the header of each sequence with the organism name and a unique number, which is referred to as the sequence identifier or simply the sequence ID. This was to facilitate ClustalW (see Section 2.3.3), which identifies each sequence from the first 30 characters in the fasta header and will not operate with duplicate names. Furthermore, subsequent BLAST (see Section 2.3.4) and ClustalW operations are confused by some characters used in the complete headers so the remaining fasta header data for the protein sequences were removed leaving only the organism name from which the protein sequence was obtained and the sequence ID identifying the protein sequence.

The re-annotated file is named aspro.fa and has the schema:

>Binomial_name sequence_ID
PROTEIN SEQUENCE

The remaining header information extracted from uniprot_sprot.fasta was saved in a separate file named aspro_dat.pl:

aspro(sequence ID, organism, original sequence/protein ID, description).

The file aspro_dat.pl contains the sequence ID, by which the information can be cross referenced with the protein sequence data, the binomial organism name, the protein sequence identification and the protein function description in Datalog format.

7.3 Superfamily Class Data

Gough *et al.* have constructed a library of hidden Markov models called Superfamily, that represent all proteins of known structure (Gough *et al.* , 2001). The sequences of the domains in proteins of known structure, that have identities less than 95%, are used as seeds to build the models. The sequences used by Gough *et*

al. to generate the models are from the ASTRAL database (Brenner *et al.* , 2000). The ASTRAL database provides protein sequences categorised according to the SCOP domain definitions and are derived from the SEQRES entries in PDB files (see Section 2.2.3). The sequences used by Gough *et al.* differ from the ASTRAL sequences in several ways. These differences are explained in Section 2.2.3.

The methods Gough *et al.* used identified many more superfamily classifications than SCOP, but there were problems with their classifications, which is explained later in this chapter.

The superfamily classification data was downloaded from the Superfamily website[2] (Gough *et al.* , 2001) on the 27th September 2007. This data consists of many files of which the two used in this research are detailed below: -

765ass.tab is the main file containing data on the seed sequences representing the superfamily classes in the schema: -

organism mnemonic, protein id, model id, location, A_e, A_n, A_d, B_e, B_n, B_d

where the *organism mnemonic* consists of two alphanumeric characters identifying the organism given by its *binomial name*. The *location* is the bp location of seed in protein sequence. A and B represent further homologous sequences that maybe included in the file where the subscripts e, n and d represent the E-value, protein ID[3] and protein sequence description respectively.

genome.tab is a list of all 603 organisms involved in the Superfamily data in the schema:-

organism mnemonic, binomial name, domain, ftp location

where the *organism mnemonic* consists of two alphanumeric characters identifying the organism given by its *binomial name*. The *domain* is given by

[2]Superfamily download site:
http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/.

[3]This is the protein identifier given by the repository for the sequence (e.g. Genbank) and not to be confused with protein identifiers assigned in this work.

a single letter; A for archaea; B for bacteria and E for eukaryota. The *ftp location* represents the URL for the FTP site from which the original protein sequence data was obtained. There are many duplicate organisms in this data and so this file was manually sorted to remove the duplicates and saved as *sorted_genome.tab*.

sorted_genome.tab is a list of 467 unique organisms filtered from *genome.tab* above and takes exactly the same schema.

7.4 Method

The protein sequence data from Superfamily were sorted placing the data on each protein sequence into separate files. Each file lists all protein sequences that were considered homologous by Superfamily to their models. The files generated take the following schema:

protein(protein ID, organism mnemonic, model ID, E-value, organism name)

The *organism name* now takes the form of the binomial name, but with the space between names replaced with an underscore character so this name can be treated as a single variable in subsequent analysis. This data is stored in a database named *a_phylo*.

In the paper that introduced Superfamily by Gough *et al.* (2001) they identified 4894 models. However, there were 10,894 models listed in the data downloaded so it is assumed that ongoing work by Superfamily had identified a further 6000 potential models. The procedure above to identify Superfamily's homologues only found one or more identified homologues for only 8,911 models. Again, it is assumed that Superfamily will in time identify homologues for the outstanding models. However, there is sufficient data to proceed with this research.

Three files were generated containing lists of organisms from each of the three domains; archaea.pl listing all archaea, bacteria.pl listing all bacteria and eukaryot.pl listing all eukaryota in the Superfamily data.

All the protein sequence data for all organisms listed in sorted_genome.tab were downloaded from their respective ftp sites as given in the same file and stored in a protein sequence database named *prot_db*.

From the *a_phylo* database only the eukaryote protein identifiers from eukaryot.pl were selected that were most similar to each Superfamily model sequence by having the lowest E-value. Each one of these protein identifiers was used to search and obtain the protein sequence for that identifier from the *prot_db* database. The results from this filtering process are 10,894 separate files in FASTA format containing one complete protein sequence considered to be most homologous to each of the Superfamily models. These sequences are referred to as the *model protein sequences* and are identified by the *model protein ID* or the *model protein identifier* from here on. The schema for the FASTA header for this data is: -

model_protein_ID protein_ID organism_name

where the *model_protein_ID* is the same identifying number used to represent all 10,894 Superfamily models. This is a number from 34782 to 45675. These data are stored in a database named *model_proteins*.

7.4.1 Sequence Selection from the Swiss-Prot Database

BLAST was used to produce files of proteins that are homologous to each model protein sequence in the *model_proteins* database (see Section 2.3.4 for more information about BLAST). The E-value threshold parameter was set to 1, which produces a fairly broad selection of homologous sequences. Also, the inclusion of the sequences in the results files was supressed by setting the $-b$ parameter to 0. This speeds up the processing and simplifies parsing of the results.

Many of the files in the *model_proteins* database were empty and would cause errors in BLAST if these empty files were used. A process list was created listing the 8,911 model protein identifiers that contained valid sequences. This list was used in a script to select only files containing a valid protein sequence.

7.4.2 Generating the Phylogenetic Trees

The multiple sequence alignment program ClustalW was used to create the phylogenetic trees from the homologous sequence data produced by BLAST.

At this stage it was necessary to update the fasta format header for each protein sequence for three reasons: -

1. ClustalW will not operate with duplicate names. We are identifying our sequences by the name of the organism from which the protein sequences were obtained and there are multiple instances of each organism. There is already a unique sequence ID in the header so each header should be unique. However, there is another limitation.

2. ClustalW identifies each sequence from the first 30 characters only in the fasta header. Since many organism names exceed 30 characters the unique sequence ID coming after the organism name will not be registered. These headers will be seen as duplicates. This problem was remedied by preceding the organism name with the sequence ID, but this leads to another difficulty.

3. If the protein sequence identifier begins with a number further analysis using Prolog and WARMR would be complicated since Prolog would see each protein sequence identifier as a real number instead of a variable. Placing the protein sequence ID in inverted commas confuses certain WARMR processes so the final solution was to re-annotate each fasta header by having a lower case 'n' followed by a unique sequence ID prefixed to the organism name.

Having re-annotated the fasta headers the tree creation process can begin. A script was used to select model protein identifiers from the process list from before to perform a multiple sequence alignment and a corresponding tree generation using ClustalW for all sequences homologous to each of the model protein sequences. In this way a tree in the Phylip/New Hampshire format (see Chapter 2) for each model protein sequence was produced. The Phylip/New Hampshire format was converted into Datalog format such that the tree structure is represented by terms describing each edge and its respective nodes explained in more detail in Section 7.5. Each node represents an Operational Taxonomic Unit (OTU) and has either

the name of an organism or is labelled as 'iNodeXX' where XX is a unique number identifying that node. Each iNode is an internal node and is representative of extinct OTUs.

File	Format
organism.pl	organism(organism_name)
aspro_dat.pl	aspro(sequence ID, organism_name, orig. protein ID, description).
scop_data.pl	scop_data(model protein ID, 'SCOP class', A, B, C, D).
model_dat.pl	model_data(model protein ID, 'description').
eukaryot.pl	eukaryote(organism_name).
archaea.pl	archaea(organism_name).
bacteria.pl	bacteria(organism_name).

Table 7.1: File names and datalog schema of the background knowledge files suitable for use with the trees knowledge base. A more detailed explanation of these files is given in the text.

7.5 The Tree Database

7.5.1 Knowledge Base

The tree knowledge base contains data describing 2,216,415 edges defining the structure of 8192 phylogenetic trees, where each tree maps the potential evolutionary history of each model protein sequence. The database contains examples from 6531 different organisms using 170,872 protein sequences out of all 289,473 protein sequences from Swiss-Prot.

The tree data was produced in the key format and the model format both suitable for WARMR (see Section 2.5.2). However, the model format is not suitable for Prolog, but allows WARMR to process much larger data sets. The schema of predicates in the knowledge base is: -

edge(Organism/iNode, ancestral node, E-value, tree ID, sequence ID).

7.5.2 Background Knowledge

The background knowledge files are listed in Table 7.1 showing the datalog schema associated with each file. These files are described in more detail below:

organism.pl List of all organisms

aspro_dat.pl Data on all 290,484 protein sequences taken from Swiss-Prot.

scop_data.pl Gives the SCOP protein structure classification of the model protein used to generate each tree identified by the model protein ID. The SCOP class annotation is a four part alphanumeric separated by full stops and this is given by the term *SCOP class*. The four parts of the SCOP class are divided up into terms *A, B, C, D* in Table 7.1 as shown in this example: *scop_data(model protein ID, 'b.47.1.2', b, 47, 1, 2).*

model_dat.pl Gives a description of the model protein used to generate each tree identified by the model protein ID.

eukaryot.pl List of 121 eukaryota from the Superfamily data.

archaea.pl List of 37 archaea from the Superfamily data.

bacteria.pl List of 309 bacteria from the Superfamily data.

7.6 Parametric and Structural Queries

Here we present queries suitable for the extraction of information from the phylogenetic database. Some of the following queries have been adapted from those given by Nakhleh *et al.* (2003), but it should be noted that the derivation of queries such as these are well known in the logic programming community.

The queries in this section are concerned with the internal structure and parametric details of phylogenetic trees in general. The term T used throughout these queries refers to the model protein identifier, which is used to discriminate the trees since each tree is generated from sequences thought to be homologous to the model protein sequence. Although the atomic predicates include the variable *E-distance*, it is not used in these queries. The distance measure used in these queries in the topological distance, which is the number of edges between OTUs.

Atomic predicates

Given that the knowledge base has the datalog format:

```
edge(node, ancestor node, E-distance, model protein ID)
```

Transitive closure of predicate *edge*

Transitive closure of the *edge* predicate can be used to find all ancestors or confirm an ancestral relation between nodes A and B. *ancestor(A, B, T)* is true iff there is a path from A to an ancestor B within the phylogenetic tree T:

```
ancestor(A,B,T):-
   edge(A,B,_,T).
ancestor(A,B,T):-
   edge(A,X,_,T),
   ancestor(X,B,T).
```

Common ancestor

The predicate $common\_ancestor(A, B, C, T)$ is true where C is an ancestor of both A and B within a tree T:

```
common_ancestor(A,A,A,T).
common_ancestor(A,B,C,T):-
   ancestor(A,C,T),
   ancestor(B,C,T).
```

Most recent common ancestor

Most Recent Common Ancestor given by $mrca(A, B, C, T)$ is true iff C is the most recent common ancestor of A and B within tree T:

```
not_mrca(A,B,C,T):-
   common_ancestor(A,B,C,T),
```

```
    common_ancestor(A,B,D,T),
    ancestor(D,C,T).
mrca(A,B,C,T):-
    common_ancestor(A,B,C,T),
    \+not_mrca(A,B,C,T).
```

Most distant common ancestor

Most distant common ancestor given by $mdca(A, B, C, T)$ is true iff C is the root node for all ancestral nodes of both A and B within tree T:

```
not_mdca(A,B,D,T):-
    common_ancestor(A,B,C,T),
    common_ancestor(A,B,D,T),
    ancestor(D,C,T).
mdca(A,B,C,T):-
    common_ancestor(A,B,C,T),
    \+not_mdca(A,B,C,T).
```

Minimum spanning clade of two leaf nodes

Minimum spanning Clade returns C being all nodes within the clade that has the most recent common ancestor of A and B at its root:

```
msc(A,B,C,T):-
    mrca(A,B,C,T).
msc(A,B,C,T):-
    mrca(A,B,D,T),
    ancestor(C,D,T).
```

Basal node of a minimum spanning clade of two leaf nodes

Basal node is explicitly the most recent common ancestor:

```
basal(A,B,C,T):-
```

```
    mrca(A,B,C,T).
```

Length of path between two nodes

The topological length of the path between two nodes A and B within tree T is returned in P from the predicate $path\_length(A, B, P, T)$:

```
path_length(A,B,1,T):-
    edge(A,B,_,T).
path_length(A,B,P,T):-
    edge(A,X,_,T),
    path_length(X,B,M,T),
    P is M + 1.
```

Distance between two leaves

The topological distance between two leaves A and B within tree T is returned in P from predicate $distance(A, B, P, T)$:

```
distance(A,B,P,T):-
    mrca(A,B,C,T),
    path_length(A,C,M,T),
    path_length(B,C,N,T),
    P is M + N.
```

Comparing distance between 3 leaf nodes

Given 3 leaf nodes A, B and C, the predicate $closer(A, B, C, T)$ is true if the topological distance between A and C is shorter than between A and B:

```
closer(A,B,C,T):-
    distance(A,C,M,T),
    distance(B,C,N,T),
    M < N.
```

7.7 General Queries Incorporating Background Knowledge

In this section some example general queries incorporating background knowledge are presented demonstrating how new information can be extracted from the phylogenetic tree database.

7.7.1 Phylogenetic Pairs

Phylogenetic pairs are pairs of extant OTUs that share a common ancestor. Searching for frequent phylogenetic pairs for all organisms can be used to determine evolutionary relationships.

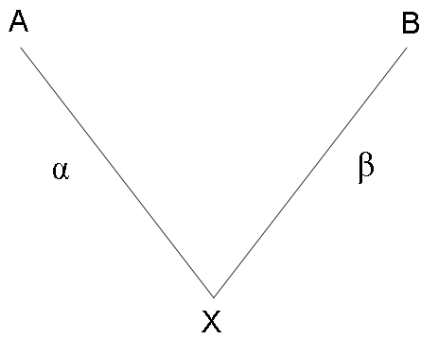

Figure 7.1: Diagram of a phylogenetic pair where A and B represent OTUs that have paths α and β respectively to a common ancestor represented by node X.

Referring to Figure 7.1, A and B represent OTUs that have paths α and β respectively to a common ancestor represented by node X. The paths α and β consist

of one or more edges, but it should be noted that long paths are of arguably limited use. This is because sequence alignment algorithms give increasingly reliable results with increasingly homologous sequences and those sequences will belong to OTUs that have shorter paths between them. Furthermore, searching for relationships with long paths is computationally expensive. For this reason there is often a constraint on the number of edges in a single path.

7.7.2 Method

The method adopted in this research has been previously referred to as *broad phylogenomic sampling* (Dunn *et al.* , 2008). This method relies on the most frequent phylogenetic results from many trees to establish evolutionary relationships between *H. sapiens* and other organisms. The *most recent common ancestor* query from above is used to determine phylogenetic pairs. However, we did not use the topological distance metric above. Instead, the E-distance measure E is given by:

$$E = 1 - \sum_{k=1}^{N} Ev_k \tag{7.1}$$

In equation 7.1 N is the number of edges linking the pair of OTUs via the most recent common ancestor and Ev is the evolutionary distance given by the term *E-value* in the *edge* predicate facts in the knowledgebase[4]. In this experiment the path lengths were limited to 8 edges.

7.7.3 Results

The results presented in Tables 7.2 and 7.3 are encouraging in that they indicate that all but one of the closest relatives to *H. sapiens* are primates. Furthermore, at the other extreme, the most distant relatives include blue-green algae *Anabaena* and fungi *C. glabrata, C. albicans and Y. lipolytica*. In fact there are organisms

[4]This was discovered later to an unsound approach and is discussed in more detail in Chapter 9.

from a variety of different phyla, which is to be expected. However, there are some important anomalies that warrant further consideration:-

1. The second closest organism to *H. sapiens* is *Mustela putorius furo*, the polecat, even though several ape species are included in the analysis.
2. The chimpanzee, *Pan troglodytes* is considered one of the closest relatives to humans, but falls in eighth place in the results.
3. In the results the fruit fly, *D. melanogaster* is a more distant relative than the yeast, *S. cerevisiae* even though the fruit fly, *D. erecta* is a closer relative than yeast. This problem is exacerbated by the large number of examples for both *D. melanogaster* and *S. cerevisiae*, implying a high degree of confidence in this anomalous result.

The placing of the polecat in the results Table 7.2 was particularly perplexing. We referred to this as the 'polecat problem' and it is studied in more detail in the following Section 7.7.3. It is hoped that this will shed some light on the nature of the problems with these results that may also explain the remaining anomalies.

Organism	E-distance	Examples
Pan paniscus	0.995364	86
Mustela putorius furo	0.983112	13
Saguinus oedipus	0.981555	63
Aotus nancymaae	0.980348	154
Macaca fuscata fuscata	0.980249	28
Theropithecus gelada	0.978452	11
Gorilla gorilla gorilla	0.971265	426
Pan troglodytes	0.961599	2855
Pongo pygmaeus	0.96042	7276
Papio hamadryas	0.951834	46
Cercopithecus aethiops	0.947222	97
Macaca fascicularis	0.932491	2318
Saimiri sciureus	0.920972	62
Callithrix jacchus	0.914683	125
Macaca mulatta	0.909613	458
Dama dama	0.901702	12
Tupaia glis belangeri	0.876846	30
Bos taurus	0.844367	7080
Papio anubis	0.840821	98
Spermophilus tridecemlineatus	0.830421	27
Canis familiaris	0.826853	1260
Sus scrofa	0.819325	1081
Oryctolagus cuniculus	0.816233	1105
Mesocricetus auratus	0.791056	179
Thalassophryne nattereri	0.786867	41
Equus caballus	0.770369	190
Bufo marinus	0.754007	15
Rattus norvegicus	0.730419	4102
Cricetulus griseus	0.710609	78
Vicia faba	0.709323	11
Mus musculus	0.677853	30532
Ovis aries	0.651653	34
Cavia porcellus	0.649648	141
Drosophila erecta	0.640529	35
Felis silvestris catus	0.63979	97
Coturnix coturnix japonica	0.59183	50
Ictalurus punctatus	0.557048	12
Halobacterium salinarium	0.53434	16
Gallus gallus	0.499774	1021

Table 7.2: Organisms having a common ancestor with *H. sapiens* (part 1). **E-distance** is the evolutionary distance and **Examples** is the number of examples of the common ancestor pairing found.

Organism	E-distance	Examples
Anthocidaris crassispina	0.493827	11
Strongylocentrotus purpuratus	0.487301	21
Dugesia tigrina	0.450764	39
Hydra attenuata	0.445304	118
Xenopus laevis	0.429699	568
Xenopus tropicalis	0.412083	73
Danio rerio	0.361138	440
Pleurodeles waltlii	0.359499	17
Fasciola hepatica	0.358523	12
Ustilago maydis	0.306608	36
Agaricus bisporus	0.303682	30
Discopyge ommata	0.303602	13
Escherichia coli	0.292745	22
Fowlpox virus	0.283692	13
Medicago sativa	0.279168	13
Dictyostelium discoideum	0.274396	41
Arabidopsis thaliana	0.266403	344
Neurospora crassa	0.256251	12
Ashbya gossypii	0.255217	24
Oryza sativa subsp japonica	0.235131	74
Schizosaccharomyces pombe	0.228683	280
Manduca sexta	0.222518	22
Caenorhabditis elegans	0.222162	471
Oxalobacter formigenes	0.216282	18
Bombyx mori	0.203593	11
Fugu rubripes	0.194571	27
Acanthamoeba polyphaga mimivirus	0.194415	35
Emericella nidulans	0.192196	28
Zea mays	0.191359	15
Saccharomyces cerevisiae	0.188163	334
Methanococcus jannaschii	0.183786	15
Bacillus subtilis	0.178216	20
Yarrowia lipolytica	0.172073	23
Candida albicans	0.151799	19
Candida glabrata	0.143612	24
Drosophila melanogaster	0.141827	655
Dugesia japonica	0.137991	11
Sturnus vulgaris	0.0966652	29
Lymnaea stagnalis	0.0852633	18
Anabaena sp	0.08023	11

Table 7.3: Organisms having a common ancestor with *H. sapiens* (part 2). **E-distance** is the evolutionary distance and **Examples** is the number of examples of the common ancestral pairing found.

The Polecat problem

In the results there was a problem in that the polecat showed up as being one of the closest relatives to humans. The database was interrogated to find all trees where the pair relation for *Homo sapiens* and *Mustela putorius furo* occurred and the following data extracted:

- Identity of the two protein sequences involved.
- The evolutionary distance between the protein sequences of *Homo sapiens* and *Mustela putorius furo*.
- Identity of the model protein sequence tree.
- A description of the function of the model protein from which the tree was derived.

These results are given in Table 7.4.

From Table 7.4 it can be seen that there are 13 examples of the pair relation for *Homo sapiens* and *Mustela putorius furo* and it is clear that the same two protein sequences have been detected in all 13 model protein trees. Both of the protein sequences n107304 and n107306 are annotated in the Swiss-prot data as being associated with *Kv channel* function.

As a side note it is interesting to see that the same two sequences n107304 and n107306 have different evolutionary distances varying from 0.00271 to 0.0226 for different model protein trees. These values assigned by ClustalW are inversely proportional to sequence homology, i.e. smaller values indicate greater sequence similarity. This is evidence that the E-distances produced by ClustalW are tree specific.

Although it is clear that all 13 results involve the same two protein sequences, we needed to know why these sequences fell into 13 different superfamily classifications. It was hypothesized that these were multi-domain proteins and each domain had one or more similarities in each of the 13 model protein sequences.

Prot. A	Prot. B	E-distance	Tree ID	Tree description
n107304	n107306	0.02260	m35206	Calcineurin regulatory subunit (B-chain).
n107304	n107306	0.01942	m35583	Neurocalcin (is a neuronal calcium-binding protein).
n107304	n107306	0.02105	m37732	Frequenin (neuronal calcium sensor 1).
n107304	n107306	0.02106	m37987	Frequenin (neuronal calcium sensor 1).
n107304	n107306	0.01852	m38680	Apoptosis linked protein alg-2.
n107304	n107306	0.01960	m39467	Guanylate cyclase activating protein 2, GCAP 2.
n107304	n107306	0.02094	m40067	Calpain small (regulatory) subunit (domain VI).
n107304	n107306	0.00636	m42852	Calcyclin (S100).
n107304	n107306	0.01482	m43209	Recoverin (neuronal calcium-binding protein (optic)).
n107304	n107306	0.01715	m43210	Translational regulator protein regA.
n107304	n107306	0.00271	m43360	Oncomodulin.
n107304	n107306	0.01762	m43503	Kchip1, Kv4 potassium channel interacting protein.
n107304	n107306	0.01770	m44365	Calcineurin B-like protein.

Table 7.4: Analysis of the apparent evolutionary relationship between Humans (*H. sapiens*) and the polecat (*M. putorius furo*). **Prot. A** is the protein identifier for *H. sapiens* and **Prot. B** is the protein identifier for *M. putorius furo*. **E-distance** refers to the evolutionary distance between the two identified protein sequences. **Tree ID** refers to the model protein tree identifier and **Tree description** describes the function of the model protein from which the tree was derived. It is clear that the same two protein sequences have been detected in all 13 model protein trees indicating a potential problem with the database. Note also that the E-distances vary even though the two protein sequences are the same in each case.

A sequence alignment was performed on the phylogenetic pair sequences for *Homo sapiens* and *Mustela putorius furo* and all thirteen model protein sequences. The most interesting 300 residue sites (points that may contain a residue or a gap) are listed below:

```
Homo_sapiens              SYDQLTGHPPGPTKKALKQRFLKLLPCCGPQVLPSVSETLAAPASLRPHR
Mustela_putorius_furo     SYDQLTGHPPGPTKKALKQRFLKLLPCCGPQALPSVSETLAVPASLRPHR
0043209                   ----ISGNLPSQNKKAFKQRFLKLLPCCGPESTPSVSESFALNPSLS---
0043503                   ----MSG-----CSKRCRLGFVKFAQTIFKLITGTLSKDIAWWYYQYQR-
0037732                   --------------------------------MGNAKSKLS---------
0035583                   --------------------------------------------------
0039467                   --------------------------------------------------
0037987                   --------------------------------MGALVSKIGFSCRKKK--
0035206                   --------------------------------------------------
0043210                   HLRMEQIYRFIYDKKTSNKQKYSMQSKINQQALPVASGSISIMLNVARKD
0044365                   ------------------------MVDSSEGLRRLAALLFKCCSLDS--
0042852                   --------------------------------------------------
0038680                   ----------------------------MQLIKKLSPKRWFSSKKDR--
0040067                   --------------------------------------------------
0043360                   RLVKVKFGPEEAFYTVNRKGEKLRKVNFKGVELLSIEVTQDARKKPMLLV


Homo_sapiens              PRPLDPDSVDDEFELSTVCHRPEGLEQLQEQTKFTRKELQVLYRGFKNEC
Mustela_putorius_furo     PRPLDPDSVEDEFELSTVCHRPEGLEQLQEQTKFTRKELQVLYRGFKNEC
0043209                   ------DSVEDDFELSTVCHRPEGLEQLQEQTKFTKKELQVLYRGFKNEC
0043503                   ------DKIEDDLEMTMVCHRPEGLEQLEAQTNFTKRELQVLYRGFKNEC
0037732                   --------------------PEQLNELQKSTNFEKKELQQWYKGFLKDC
0035583                   --------------------MGQTATLPCRKGGTYVTELYEWFRKFLNEC
0039467                   ----------------------------------------EWYKKFLEEC
0037987                   ---------------LLGRDPKRTIVRLVRVTNFTEGEVKKWMKEFEKDC
0035206                   ----------------------------MGGKLTKKDVERLQRRFQRLA
0043210                   YGFNLQSFKKKIAPKSKFKMKWKTIQWLLKNRKDAIRQIFQNYQSIVKQA
0044365                   ---------------SNRPNGLQDPERLARETVFNVNEIEALYELFKKIS
0042852                   ----------------------------------MSPWTHHLKNIDKGC
0038680                   ------------------SELSRSEPSSFSSGTASSDASDSSISNVKANS
0040067                   ----------------------MAYHYQQPQGGYDPNYLSGIFQRVDKDR
0043360                   RVPRDYDLVLEFDSTASRNRFLHKLETFLTSHNKHLEQIPTYREQMLSNA
                                                                          .
```

```
Homo_sapiens              P---SGIVNEENFKQIYSQFFPQGDSST----------------------
Mustela_putorius_furo     P---SGIVNEENFKQIYSQFFPQGDSST----------------------
0043209                   P---SGIVNEENFKQIYSQFFPQGDSSM----------------------
0043503                   P---SGVVNEETFKQIYAQFFPHGDAST----------------------
0037732                   P---SGQLDKTEFQKIYKQFFPFGDPSR----------------------
0035583                   P---SGLITLHEFRRHFCNGTVGKESAE----------------------
0039467                   P---SGTLFMHEFKR-FFKVPDNEEATQ----------------------
0037987                   P---DGFLREDEFVAHYSDYFSAGNQRRKE--------------------
0035206                   NN--TGKVQIATFQTMVELG---GNP-------------------------
0043210                   KDF-PEGLNREQFQGLLISFGLGADKN----------------------
0044365                   S----AVVDDGLINKEEFQLALFKTNRKDS--------------------
0042852                   P-------------KPLTSASPLSEEQ----------------------
0038680                   AAA-NAGFRTPTSVLPQISGDWSDMSTDFY--------------------
0040067                   S----GSISSNELQQALSNGTWTPFNPET----------------------
0043360                   ETRERREMRLEHFFREAYALTFGPKPGEKRKMEEVTGDAIIVMRTSLSRS


Homo_sapiens              -------------YATFLFNAFDTNHDGSVSFEDFVAGLSVILRG-TVDD
Mustela_putorius_furo     -------------YATFLFNAFDTNHDGSVSFEDFVAGLSVILRG-TIDD
0043209                   -------------YAHFLFNAFDTDHSGSVSFEDFVAGLSVILRG-TIDD
0043503                   -------------YAHYLFNAFDTTQTGSVKFEDFVTALSILLRG-TVHE
0037732                   -------------FADYVFNVFDGDKNGFIDFKEFICALSVTSRG-RVDE
0035583                   -------------YAEQIFRTLDNNGDGVVDFREYVTAISMLIEG-STVE
0039467                   -------------YVEAMFRAFDTNGDNTIDFLEYVAALNLVLRG-TLEH
0037987                   ------------ALAKQIFRTFDKDASGCVDFWEFMCGMSALLRG-TTVE
0035206                   -------------FVPHLFKLFDSSGDGSLNLEEFTRALEYFGQLDNEEE
0043210                   -------------LAEKLFYVFDEDSSGTVDYKELIVGLEVLKDD-TIDE
0044365                   ------------MFADRVFDLFDTKHNGILGFEEFARALSVFHPNAPIDD
0042852                   --------------LNKIFNRYDTNGDGHLSWEELKSAYNILGMSFPGLR
0038680                   ------------FELTQAFKVIDRDNDGLVSRNELEALLTRLGAEPPSSQ
0040067                   --------------VRLMIGMFDRDNNGTINFQEFSSLWKYITDW-----
0043360                   EFASALGMKGDDVFVKMMFNIVDKDGDGRISFQEFLDTVVLFSKG-RTED
                                          :   *    . :   :
```

```
Homo_sapiens              RLNWAFNLYDLNKDGCITKEEMLDIMKSIYDMMG---KYTYPALREEAPR
Mustela_putorius_furo     RLNWAFNLYDLNKDGCITKEEMLDIMKSIYDMMG---KYTYPALREEAPR
0043209                   KLNWAFNLYDLNKDGCITKEEMLDIMKSIYDMMG---KYTYPNMREEAPR
0043503                   KLRWTFNLYDINKDGYINKEEMMDIVKAIYDMMG---KYTYPVLKEDTPR
0037732                   KLYWAFQLYDIDNDGYITREEMLNIVDAIYKMVG---SMVKLPPDEDTPE
0035583                   KLRWSFKLYDKDKDGAITRSEMLEIMQAVYKMSV---AASLTKPDPLTAE
0039467                   KLKWTFKIYDKDRNGCIDRQELLDIVEALESFRV---CFPTT----MKPE
0037987                   KLKWAFSMYDLDGNGYICTTELLNVLKLMHELRYPSATEEELEKVQAPLE
0035206                   QYKFAFRIYDQDGDGFISSEELFNVLQTLMGAAVP----------DSQLE
0043210                   KLKIFFDLCDEDGSGKVSEKEIFNILKQNIINEN----------DKYQLK
0044365                   KIDFAFKLYDLKQQGFIEKQEVKQMVVATLAESG-------MNLSDEIIE
0042852                   ALK-ALCVADENRDGYISQKEFIKLMRKKYRK------------------
0038680                   EMAVMLGEVDLISVEELASRLGSACEPAGGDELRDAFVFFDSDRDGKITA
0040067                   --QTTFRNYDRDSSGTIDKNELQNALTSFGYRLS---------------D
0043360                   KLRIIFDMCDKDHNGVIDKGELSEMLRSLVEIAR-----TNNSLNDDQVT
                               :   *      :

Homo_sapiens              EHVESFFQKMDRNKDGVVTIEEFIESCQKDENIMRSMQL--FDNVI----
Mustela_putorius_furo     EHVESFFQKMDRNKDGVVTIEEFIESCQKDENIMRSMQL--FDNVI----
0043209                   EHVENFFQKMDRNKDGVVTIEEFIESCQKDENIMRSMQL--FDNVI----
0043503                   QHVDVFFQKMDKNKDGIVTLDEFLESCQEDDNIMRSLQL--FQNVM----
0037732                   KRVKKIFDLMDNDKDGRLSMEEFKEGSKKDPTILQALNL--YDGMV----
0035583                   ECTNRIFVRLDKDNNAIISQDEFIEGALNDEWIREMLEC--DPNTVKMER
0039467                   EVVDRIFLLVDENGDGQLSLNEFVEGARRDKVLL----------------
0037987                   KVRDRVFNELDRDGDGRLELREFVEGVRKNPALLKMVEE--GGDDDCTVQ
0035206                   QVVYNTMSEFDRDGDNKLDMQEFKALLSRDDLANKFSMS--M--------
0043210                   MVIREMIKQVDQDGDGELNKEEILQAASKNPILRRLLEQ--TISNVRRID
0044365                   GIIDKTFEEADTKHDGKIDKEEWRNLVLRHPSLLKNMTLPYLRDITTTFP
0042852                   --------------------------------------------------
0038680                   EELLNVYKAFGDEKCTLEDCKGMIAVVDKNGDGFVCFEDFCRMMELQR--
0040067                   KFYSILIKKFDRSGRGVVNFDDFIQCCVVIQMLTNAFQAYDNNRNGWISI
0043360                   ELIDGMFQSAGLEHKDALTYDDFKLMMREYHGDFIAIGLDCKGAKQNYLD
```

Although the alignment falls into two regions there is clearly a sequence similarity throughout both regions in all 15 sequences so it can be concluded there are not multiple domain homologies in these results. We conclude from this that the sequences of the model proteins representing the superfamily classifications in these results are too closely related. This means that many of the superfamily classifications derived by Gough *et al.* (2001) in the Superfamily project are likely to result in overlapping homologies. This will result in over representation of many phylogenetic relationships such as we have seen with our polecat example.

7.8 Protein Evolution

As previously stated in the introduction, the protein tree database discussed in this chapter can be used to study protein evolution.

Protein evolution is becoming of increasing interest to researchers in many areas including resolving problems and inaccuracies in the Tree of Life, which is discussed further in Section 8.1 in Chapter 8, and in pathogen recognition and drug discovery (Zhang *et al.* , 2006).

Proteins can be viewed as the active elements in cells (see Section 2.2.1) and so it is generally considered that protein evolution must in some way affect species evolution. There are 1210 trees in the protein tree database, which plot the evolutionary history of organisms that contain proteins from each family of known protein orthologues found in the Swiss-Prot database. This protein tree database can be used for knowledge discovery of proteins and protein function. For example, where a protein tree contains many diverse species we may conclude that this protein family is highly conserved and is therefore most likely to be an essential protein for cells such as 'house keeping' proteins. Cytochrome c and cytochrome oxidase are two essential proteins used in respiration and indeed, there are several trees of the cytochrome family (m45394, m41780, m44721 and m42845) in the protein tree database that include many diverse species. Also, where a protein tree is constrained to a clade of closely related organisms we may conclude that this protein family has a function very specific to the organisms in this clade. An

example of this is the auxin binding protein tree (m40629), which contains only plants.

Where protein trees are specific to peptide sequences, gene trees are specific to nucleotide sequences. The nature of protein evolution can be determined by comparing the nucleotide sequences from gene trees to the peptide sequences from protein trees[5].

Naturally occurring substitutions in nucleotides that do not affect the peptide sequence are known as *synonymous substitutions* and are commonly designated by the symbol K_s. Substitutions in the nucleotide sequence that lead to a change in the peptide sequence are known as *non-synonymous substitutions* and are commonly referred to by the symbol K_a. The nature of protein evolution can be ascertained from the ratio of non-synonymous substitutions and synonymous substitutions.

$$\frac{K_a}{K_s} = 1 \tag{7.2}$$

Equation 7.2 indicates that both synonymous and non-synonymous substitutions are occurring randomly and it is most likely that these proteins are not undergoing evolutionary selection, which further suggests the possibility that they are no longer expressed or that they have been *silenced*.

$$\frac{K_a}{K_s} > 1 \tag{7.3}$$

Where there are more non-synonymous substitutions than synonymous substitutions as indicated by Equation 7.3, it can be concluded that the protein is undergoing *positive* selection. Positive selection is of considerable interest in protein evolution research since it is popularly thought that the genetic differences between closely related organisms will be found in proteins undergoing positive selection (Hurst, 2009). However, Hurst (2009) challenges this assumption based

[5]Nucleotide sequences are usually available from the same download site as the corresponding protein or peptide sequences for each gene.

on recent evidence from research into differences between the genomes of Humans and Chimpanzees (Berglund *et al.* , 2009) (Galtier *et al.* , 2009).

$$\frac{K_a}{K_s} < 1 \tag{7.4}$$

Finally, Equation 7.4 indicates *purifying* selection, where the protein is highly optimised and further substitutions are deleterious. Most expressed proteins will fall into this category (Xiong, 2008).

This section has described just a few examples of the methods used in the analysis of protein trees and their comparison with gene trees, but from the point of view of this thesis, it remains an area for future research.

7.9 Conclusion

The research presented in this chapter has shown two main difficulties for the creation of a phylogenetic consensus tree. The superfamily classifications we have used in the database are unsuitable and the phylogenetic pair relation is not sufficiently informative.

The superfamily classifications we have used to generate the database are not sufficiently distinct. The superfamily classifications derived by Gough *et al.* (2001) represent clusters of protein sequence homology, which is in itself a very interesting line of research. However, we require classifications that avoid overlapping homologies where, ideally, each protein sequence belongs to only one classification[6].

The problem with phylogenetic pairs is that the ancestral relationship is not defined so we can only determine those OTUs that are close relatives to other OTUs by direct sequence similarity, but not by sequence phylogeny. One solution is to use phylogenetic triples, which detail a phylogenetic relationship between three

[6]This is, in fact, impossible since multi-domain proteins are very likely to have multiple classifications.

OTUs where all three OTUS form a clade, but two OTUs are in a separate subclade to which the third does not belong. This concept is extremely useful in phylogenetics and is described in more detail in Chapter 8.

Chapter 8

Phylogenetic Consensus Tree

8.1 Introduction

The Tree of Life (a taxonomy of all living organisms) is considered by some to be a modern day 'Holy Grail'. A good example of this is in the creation of the Tree of Life Web Project (ToL) (Maddison *et al.* , 2007). The Tree of Life Web Project is a website (Maddison & (eds.), 2007), which is the result of a collaborative effort of biologists worldwide and contains well over 9000 World Wide Web pages to date. The project is a work-in-progress and provides information about the diversity of organisms, their evolutionary history (phylogeny), and characteristics. The sources of information for the ToL project are many and varied, and consequently the accuracy and resolution of the phylogeny is open to debate.

It may soon be possible to reconstruct a molecular phylogenetic tree (See Section 2.6) for a large sample of all living organisms since Genbank[1] has recently announced that it now holds sequence data on 10% of species 'presently known' (Sanderson, 2008) (Benson *et al.* , 2000) (Benson *et al.* , 2007). This is mainly because there is a considerable input of new evidence from computer science with higher performance and improved algorithms. Data from multiple whole genome

[1]Genbank is the name of a popular online protein sequence repository http://www.ncbi.nlm.nih.gov/Genbank/

sequences from a single species (Drosophila 12 Genomes Consortium, 2007); expressed sequence tag libraries (Hughes *et al.* , 2006) (Steinke *et al.* , 2006); and barcode sequences (Hajibabaei *et al.* , 2007) are all contributing to this pool of evidence, which is continually increasing the scale of feasible phylogenetic inference (Sanderson, 2007). Recently, it has been suggested that for well studied organisms (mostly mammals) nearly complete phylogenetic trees can already be constructed (Bininda-Emonds *et al.* , 2007). Sanderson summed up the present situation well in the following quote:

> "Construction of a high resolution phylogenetic tree containing all eukaryota in the database [Genbank] is a grand challenge that is substantially more tractable than inferring the entire tree of life, but to succeed, strategies will have to overcome serious sampling impediments. Quantifying the distribution and strength of phylogenetic evidence currently in the database is a prerequisite for this effort"

Indeed, as with nearly all data in bioinformatics, the molecular sequence data used to create phylogenetic trees is noisy, as Sanderson suggests, and this results in substantial inaccuracies. One method to ameliorate this is to sample from many phylogenetic model protein trees and draw a consensus of information to use in building the Tree of Life for organisms.

This chapter describes a method to generate a consensus tree representing a phylogeny of organisms from the model protein tree database described in Chapter 7. This method has been named *broad phylogenomic sampling* by Dunn *et al.* and it is a method that has been suggested to overcome the problems with noisy data and consequently, to improve the resolution of the Tree of Life (Dunn *et al.* , 2008). The work in Chapter 7 revealed that the model protein database needed to be modified to make it suitable for this specific consensus tree building application and this is discussed later in this chapter. Two methods titled *phase I* and *phase II* are described, using phylogenetic triples to deconstruct phylogenetic trees. The procedures in *phase I* and *phase II* are identical except that a more constrained rule for the phylogenetic triple is used in *phase II*. The results from both phases are evaluated before Aho's algorithm is used to reconstruct an example consensus

tree and a final consensus tree. Finally, an evaluation of the final consensus tree is reported highlighting further problems with the database and also, the initial methods used are discussed, which further clarifies the research required to create accurate phylogenetic trees.

8.2 Phylogenetic Triples

Phylogenetic triples are used to extract information about the evolutionary relatedness of specific organisms within a larger phylogenetic tree.

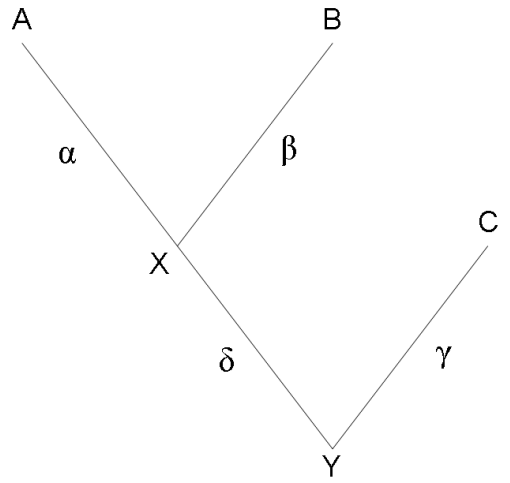

Figure 8.1: Diagram of a triple where A, B and C are external nodes that represent OTUs or extant organisms, and X and Y are internal nodes representing possibly extinct organisms. The characters α, β, γ and δ represent paths between nodes that consist of one or more edges.

A triple is essentially a set of three related variables. A phylogenetic triple describes the phylogenetic relationship between a set of three variables, where those variables are operational taxonomic units (OTUs) or extant organisms in nature. A diagram of a phylogenetic triple is shown in Figure 8.1 which clearly describes the relationship between three OTUs or organisms represented by external nodes A, B and C. A rule is applied to the external nodes such that A and B have a common ancestor represented by the internal node X that C does not, but there is a more distant ancestor represented by the internal node Y shared by all three OTUs. All five nodes A, B, C, X and Y are connected by edges α, β, γ and δ, which have values proportional to the evolutionary distance between nodes.

In essence, all three OTUs form a clade with the immediate ancestor of C being

the basal node Y of that clade, but A and B form a sub-clade where the most recent common ancestor of A and B is the basal node X, and C is *not* a member of that sub-clade. The relationship this determines, simply and reliably, is that C diverged from the ancestral line before A and B. It is this information that is essential in determining the structure of a phylogenetic tree.

8.3 Reconstructing Phylogenetic Trees

A consensus method takes a collection of disparate trees representing phylogenies of a set of taxa and returns a single consensus tree of the same set of taxa. One of the first consensus tree building algorithms was created by Adams (1972) and now this is only one of many available consensus tree methods (Bryant, 2003).

For this research a computer program specifically designed to work with phylogenetic triples was written by Amanda Clare. It was written in the functional programming language Haskell and is a realisation of an algorithm presented in a paper by Bryant (2003), which is based on an algorithm written by Aho *et al* (1981).

8.4 New Knowledge Base

The knowledge base used in Chapter 7 details trees created from the Superfamily data describing 10,891 sets of homologous sequences. However, SCOP have defined only 1538 superfamilies. The Superfamily project were more concerned with clusters of homologous sequences and they found multiple distinct clusters within the superfamily classifications defined by SCOP. For each cluster, they defined a new superfamily classification represented by a model protein sequence. Protein sequences that are homologous to the model protein sequences are classified as members of that superfamily set. Using a fixed threshold of sequence similarity, such as the E-value used in BLAST searches, there would be an overlap between sets. We wished to avoid overlapping sets because this over represents many super-

File	Format
the_trees.pl	edge(OTU/node, ancestor node, E-value, model ID).
org_freq.txt	frequency, organism
org_tre_freq.pl	org_tre_freq(organism, number of trees)
org.pl	org(organism)
query_org.pl	org(organism)
pop_orgs.pl	organism(organism)
tree_data.pl	model_data(model ID, protein_function)
tree_size.pl	pTreeSize(model ID, number of edges)

Table 8.1: File names and schema for the background knowledge files used in triple mining. The term **OTU** refers to the Operational Taxanomic Unit, which is the name of an organism; **E-value** is the evolutionary distance determined by ClustalW and **model ID** refers to the model protein identifier.

family classes. With this in mind, a new knowledge base was extracted from the knowledge base used in Chapter 7 by selecting the most populated phylogenetic model protein tree from each of the 1538 SCOP superfamily classes. This resulted in a database containing 397795 facts (edges in trees) on 5302 organisms in 1211 trees representing the phylogeny of proteins from the SCOP superfamily classes. Note that the trees for the outstanding 327 SCOP superfamily classes contained less than 3 organisms. At least 3 organisms are required in each tree for analysis and consequently, these trees were not included in the database.

8.4.1 Background Knowledge

The background knowledge files are described below and the Datalog schema for the data contained in those files is detailed in Table 8.1, which can be used as a reference for later Prolog query design.

the_trees.pl lists 397,795 facts representing edges, giving the evolutionary distance between two given nodes and the model protein identifier, which identifies the protein sequence used to represent each of the 1,211 different SCOP superfamily classes

org_freq.txt lists all 5,302 organisms and their frequency in the knowledge base.

Note that there are 933 trees that contain at least one example of *Homo sapiens*.

org_tre_freq.pl lists the top 906 most frequent organisms that have at least one example in more than 10 trees and the number of trees where there is at least one example of that organism.

org.pl all 5,302 organisms listed as predicate facts.

query_org.pl is a subset of org.pl used by the phylogenetic triple mining procedure.

pop_orgs.pl contains 906 predicate facts which represent all organisms that have at least one example in more than 10 trees.

tree_data.pl lists all 10,894 Superfamily models by the model number prefixed by 'm' and the description of the protein function of the sequences in the set derived from that model.

tree_size.pl lists of 8,192 trees/models and the number of edge predicates for each tree.

8.5 Phase I

This section describes the method used to find frequent phylogenetic triples in the new knowledge base. Note that the terms *model protein ID* and *tree ID* are synonymous throughout this chapter.

Referring to Figure 8.2, the first step labelled 'triple miner', extracts all triples using the triple structure definition given in Section 8.5.1. Each organism in each triple is drawn from the subset of 906 organisms that have at least one example in more than 10 trees. The choice of setting a threshold at 10 is arbitrary, but there will not be significant frequent patterns found involving organisms that are scarce within the database. All triples having less than 3 examples are removed from the results. For each triple there are 6 possible permutations of the member organisms. The second step finds the most frequent permutation of each triple. The next

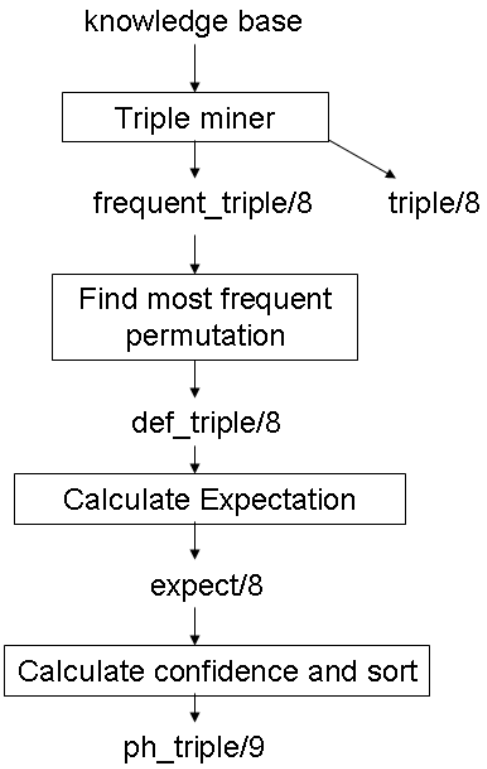

Figure 8.2: Frequent triple mining procedure and the predicate facts produced by each step. The first step extracts all triples and removes non-frequent triples. The second step finds the most frequent permutation of each triple. The next step calculates the expectation of the frequency of each triple and using this result, the final step calculates the confidence and sorts the results in order of confidence.

step calculates the expectation of the frequency of each triple using conventional statistical methods. Finally, the confidence is calculated from the deviation of the frequency of examples found from the expected frequency calculated in the previous step. The higher the frequency of the examples found above the expected frequency, the more confidence we have in the result. The final results are sorted in order of confidence. This procedure is discussed in more detail in the remainder of this section.

8.5.1 Phylogenetic Triple Miner

From the diagram in Figure 8.2 we see that the first step *triple miner* is required to find all phylogenetic triples in the knowledge base. The phylogenetic triples are defined by the following rule:

```
triple_function(A, B, C, Eva, Evb, Evc, Evd, Tree, Candidate):-
    A = Candidate,
    edge(A, X, Eva, Tree),
    ancestor(B, X, Evb, Tree),
    edge(C, Y, Evc, Tree),
    ancestor(X, Y, Evd, Tree),
    not(A = B),
    not(A = C),
    not(B = C),
    organism(B),
    organism(C).
```

Referring to the diagram of the phylogenetic triple in Figure 8.1 this rule selects three different organisms A, B and C from the subset of 906 popular organisms where all three are members of a single model protein tree. These three organisms form a phylogenetic triple where A has a direct edge to ancestor X; B has a transistive edge to ancestor X; X has a transitive edge to ancestor Y and C has a direct edge to ancestor Y. The results from applying this rule to search the knowledge base are saved along with a subset where triples with less than three examples are filtered out. This gives us the following results:

triple.pl contains 678,206 predicate facts describing all triples found in the knowledge base. Schema:

$$triple(A, B, C, \alpha, \beta, \gamma, \delta, tree\_id).$$

This data is necessary so that the tree ID/model protein ID can be traced to any frequent set. For example if we find a frequent triple, we can search this data set to identify all the model proteins associated with this frequent triple. The terms α, β, γ and δ represent the evolutionary distances between

the nodes (see Section 8.2).

freq_triple.pl contains 50,516 predicate facts for every triple having 3 or more examples in the knowledge base. Schema:

$$frequent\_triple(A, B, C, \bar{\alpha}, \bar{\beta}, \bar{\gamma}, \bar{\delta}, Number).$$

The terms $\bar{\alpha}, \bar{\beta}, \bar{\gamma}$ and $\bar{\delta}$ represent the mean of the evolutionary distances between the nodes (see Section 8.2) and $Number$ is simply the frequency of that triple.

This data lists all phylogenetic triples and frequent triples, but the frequent pattern mining criterion is based on the confidence we have in the frequency of the example phylogenetic triples given the expectation of those triples. This is described later after an evaluation of the data obtained so far.

8.5.2 Evaluation of Data

At this point it is necessary to determine if the results so far are useful. The ancestral relationships described by the phylogenetic triples can be compared with an existing commonly accepted phylogeny. A small set of eight yeasts listed in Table 8.2 are used as candidate organisms in the phylogenetic triple rule given above in Section 8.5.1. The results are compared with the commonly accepted phylogeny of yeasts produced by Cliften *et al.* (2003) given in Figure 8.3. From the comparison, those triples that are confirmed by the commonly accepted phylogeny are given in Figure 8.4 and those do not are given in Figure 8.5.

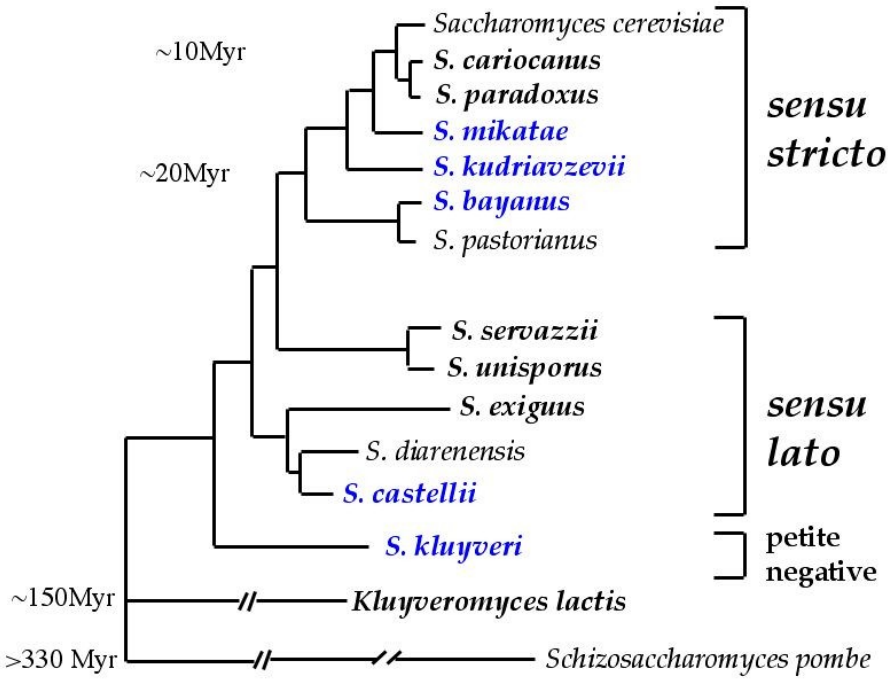

Figure 8.3: Commonly accepted phylogeny of yeast (Cliften *et al.* , 2003).

From the results there are 54 examples of phylogenetic triples that conform to accepted phylogeny and 27 that do not. All positive examples correctly place *S. pombe* on the C node implying that *S. pombe* branched first. Of all permutations of the three organisms *K. lactis, S. cerevisiae* and *S. pombe* the most frequent permutations conform to accepted phylogeny. On the whole these results are

Organisms
Kluyveromyces lactis
Saccharomyces bayanus
Saccharomyces castellii
Saccharomyces cerevisiae
Saccharomyces exiguus
Saccharomyces kluyveri
Schizosaccharomyces pombe
Kluyveromyces marxianus

Table 8.2: Yeasts used as candidate organisms for validation of the results.

positive and with further filtering by selecting those triples in which we have the most confidence, it should be possible to extract reliable information.

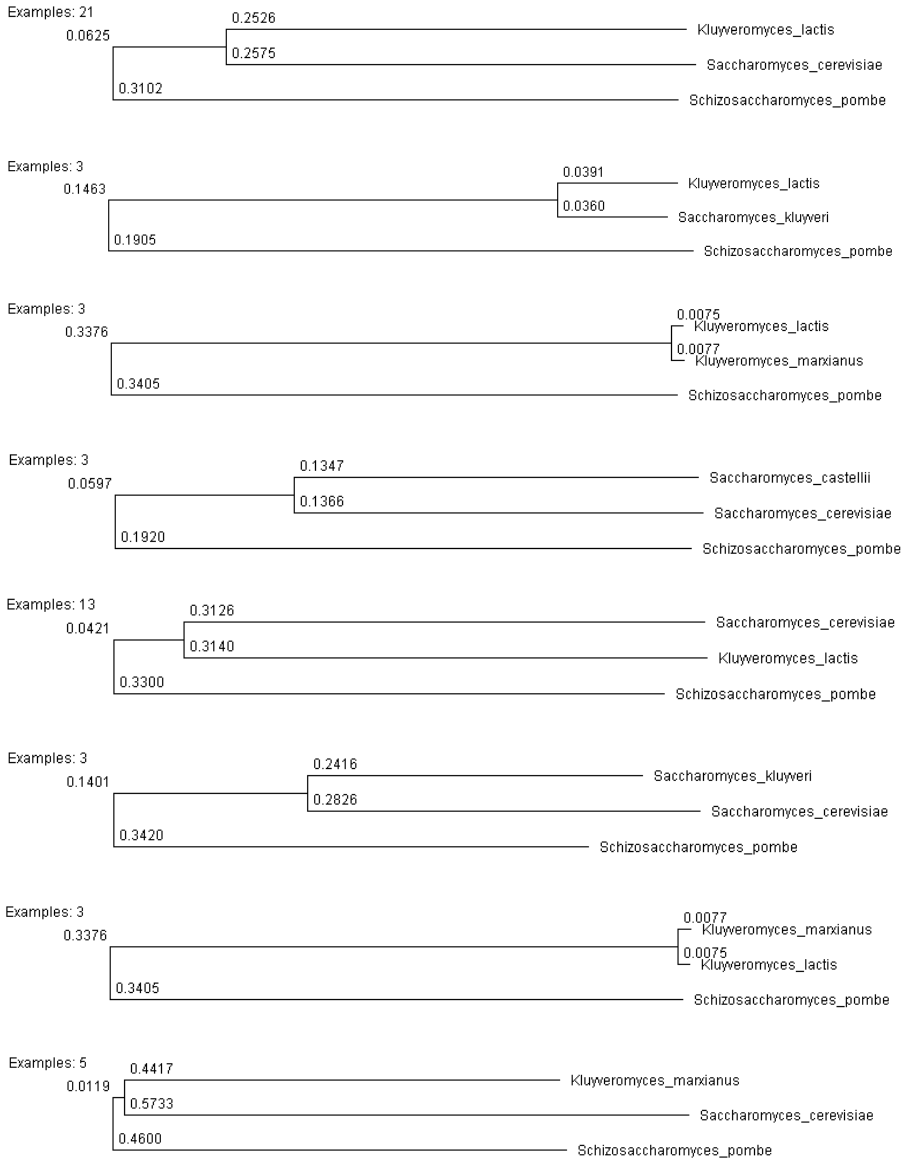

Figure 8.4: Frequent phylogenetic triples that conform to commonly accepted yeast phylogeny. The 'Examples' number refers to the frequency of each triple and the branch lengths are proportional to the relative evolutionary distance between member organisms of each triple.

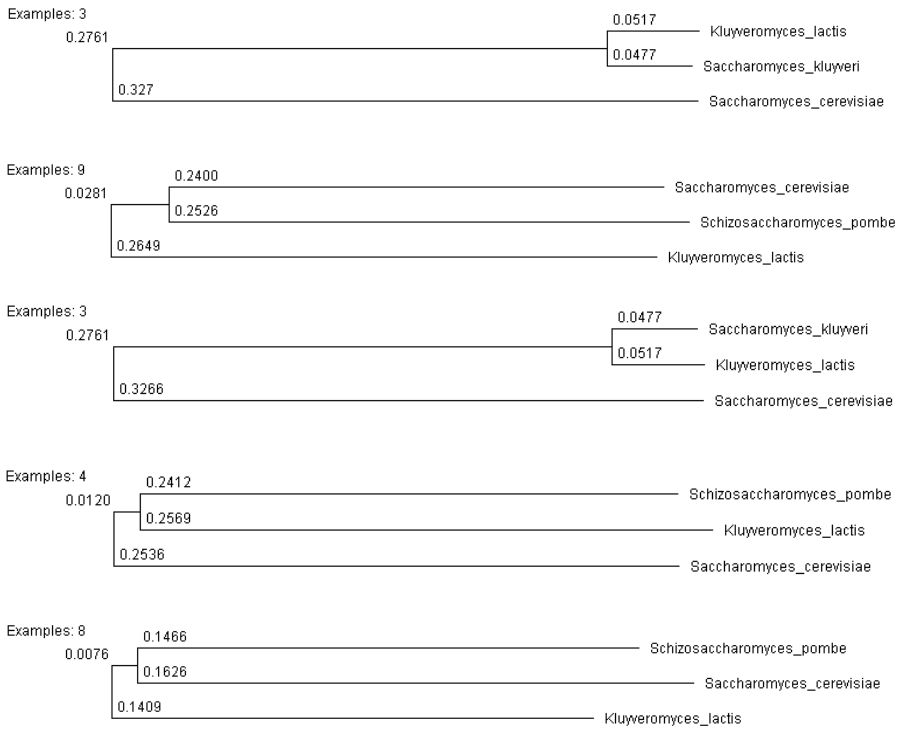

Figure 8.5: Frequent phylogenetic triples that do not conform with accepted yeast phylogeny. The 'Examples' number refers to the frequency of each triple and the branch lengths are proportional to the relative evolutionary distance between member organisms of each triple. Where more frequent permutations exist, then the triples in this figure can be filtered from the results.

8.5.3 Most Frequent Permutation

For any three organisms there are six permutations of each phylogenetic triple. The results obtained so far are filtered by selecting only the most frequent permutation. This gives us the following results file:

full_res.pl A file containing 5594 results of the following schema:

$$def\_triple(A, B, C, \bar{\alpha}, \bar{\beta}, \bar{\gamma}, \bar{\delta}, N).$$

The terms $\bar{\alpha}, \bar{\beta}, \bar{\gamma}$ and $\bar{\delta}$ represent the mean of the evolutionary distances between the nodes (see Section 8.2) and $Number$ is simply the frequency of that triple.

This step is really only necessary to reduce the quantity of data for further processing and thus reduce the computational processing time.

8.5.4 Expectation and Confidence

We can determine a measure of the confidence in the selected triple by comparing the frequency of the selected triple with the frequency of the same triple selected from a randomized dataset having the same relative frequency of the organisms comprising the triple.

Given n examples, the number of combinations of a subset of r examples is given by

$$nPr = \frac{n!}{(n-r)!} \equiv n(n-1)(n-2)\cdots(n-(r-1)) \tag{8.1}$$

From this the number of all triples N in a set of n examples is given by

$$N = n(n-1)(n-2) \tag{8.2}$$

The number of phylogenetic triples M of three organisms whose frequencies are m_1, m_2 and m_3 is given by

$$M = m_1 m_2 m_3 \tag{8.3}$$

The expectation E of a given phylogenetic triple permutation from M in a new subset K of samples selected at random from the set N is given by

$$E = K\frac{M}{N} \tag{8.4}$$

Taking as an example the triples of *Saccharomyces cerevisiae*, *Kluyveromyces lactis* and *Schizosaccharomyces pombe*. The frequency of *Saccharomyces cerevisiae* is $m_1 = 2,739$, the frequency of *Kluyveromyces lactis* is $m_2 = 348$ and the frequency of *Schizosaccharomyces pombe* is $m_3 = 2,412$. The number of all OTUs in the data is $187,617$. The number of all phylogenetic triples found $K = 678,206$. So the expectation is given by:

$$E = 678,206\frac{2,739 \times 348 \times 2,412}{187,617 \times 187,616 \times 187,615} = 0.236 \tag{8.5}$$

With a frequency of 21 the most frequent phylogenetic triple is:

$$\begin{aligned} \text{A} &= \textit{Kluyveromyces lactis} \\ \text{B} &= \textit{Saccharomyces cerevisiae} \\ \text{C} &= \textit{Schizosaccharomyces pombe} \end{aligned}$$

To determine the significance explicitly the cumulative density function should be used. However, we are only interested in comparative significance and this can be determined more easily using the probability mass function of the binomial distribution:

$$f(k; n, p) = \begin{pmatrix} n \\ k \end{pmatrix} p^k (1-p)^{n-k} \tag{8.6}$$

Where n is the number of all phylogenetic triples, k is the number of examples of the subject phylogenetic triple and p is the probability of the expectation of the subject phylogenetic triple given by:

$$p = \frac{E}{n} \tag{8.7}$$

The probability mass function gives us the probability of obtaining exactly k examples of each triple. The higher the number of examples found over and above the expectation, the lower the probability. We can use this as an inverse measure of confidence, because the lower the probability, the higher the confidence.

The equations for the expectation and the confidence are implemented in an algorithm to produce the following results:

exp_res.pl This file contains 4,295 results where the number of examples exceed the expectation (Note that only 1,299 results fell below expectation). Schema:

$$expect(A, B, C, freq.A, freq.B, freq.C, examples, expectation).$$

where $freq.A, freq.B$ and $freq.C$ are the individual frequencies of the organisms in each phylogenetic triple and can used to check the value given by *expectation*.

ph_triple.pl This file has 4,295 results arranged in ascending order of probability, where the highest deviation from the expectation is at the beginning.

$$ph\_triple(A, B, C, examples, expectation, probability).$$

where *probability* is the probability mass function (pmf) for the number of examples obtained given the expectation. Note that the lower the probability, the more significant is the frequency of that phylogenetic triple and we have a higher confidence due to the higher significance. Therefore, lower values for the probability imply a higher confidence.

These are the results files that would be used to generate the final consensus tree, however, the analysis in the following section reveals some problems with this data.

A	B	C	N	E	Pmf
H. sapiens	M. musculus	D. melanogaster	1755	854.10	4.14e-11
H. sapiens	R. norvegicus	D. melanogaster	654	408.65	2.95e-6
H. sapiens	R. norvegicus	G. gallus	591	397.26	3.86e-5
H. sapiens	M. musculus	C. elegans	1117	827.72	0.00067
H. sapiens	M. musculus	S. pombe	1130	843.43	0.00081
H. sapiens	M. musculus	G. gallus	1080	835.45	0.00192
H. sapiens	R. norvegicus	C. elegans	467	392.56	0.00960
H. sapiens	M. musculus	D. rerio	857	771.98	0.02352
S. pombe	H. sapiens	S. cerevisiae	393	246.51	3.41e-6
P. pygmaeus	H. sapiens	B. taurus	353	263.22	0.00079
C. elegans	H. sapiens	D. melanogaster	297	234.49	0.00300
G. gallus	H. sapiens	D. melanogaster	299	237.82	0.00350
B. taurus	H. sapiens	G. gallus	310	278.46	0.02287
B. taurus	H. sapiens	D. melanogaster	307	287.46	0.03226
G. gallus	H. sapiens	D. rerio	209	203.42	0.03844
B. taurus	H. sapiens	M. musculus	963	947.66	0.03935
R. norvegicus	M. musculus	H. sapiens	3890	1205.73	5.0e-23
S. scrofa	B. taurus	H. sapiens	256	243.30	0.03509

Table 8.3: Eighteen of the most significant examples of phylogenetic triples containing *Homo sapiens*. A and B share a common ancestor not shared by C. The number of examples of the triple found is denoted by N and E is the expectation. The Pmf figure in the last column is the probability mass function, which is used as a measure of significance of the number of examples found given the expectation.

8.5.5 Analysis of Phase I Results

A small set of results, being all results containing *Homo sapiens* in A, B or C positions on a phylogenetic triple, were isolated for further analysis and these are shown in Table 8.3. The majority of the results follow the accepted thinking behind the taxonomy of these organisms, but two results are curious and require further investigation.

The result for *S. pombe, H. sapiens, S. cerevisiae* is presented as a graph in Figure 8.6 and suggests that *S. pombe* is a closer relation to *H. sapiens* than to *S. cerevisiae* and that *S. pombe* and *H. sapiens* share a common ancestor that *S. cerevisiae* does not. This seems unlikely. However, the pattern search may have selected only a

specific sub set of proteins from humans and yeast that are similar merely by coincidence. Since these organisms are so diverse, these selected proteins would be essential proteins to both organism types (Eg. essential cell function) and so are not likely to have changed much through time.

The second anomolous result is for the triple comprising *C. elegans, H. sapiens, D. melanogaster*. Referring to Figure 8.7, it seems unlikely that *C. elegans* (nematode worm) shares a common ancestor with *H. sapiens* (human), but not with *D. melanogaster* (fruit fly). It is currently thought that *C. elegans* and *D. melanogaster* both being members of the superphylum ecdysozoa, share a common ancestor not shared by *H. sapiens* (Maddison & (eds.), 2007).

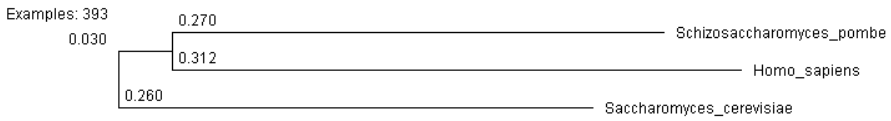

Figure 8.6: Anomalous triple I: It is generally accepted that *S. cerevisiae* and *S. pombe* would share a common ancestor not shared by *H. sapiens*, which is not the indication from this result.

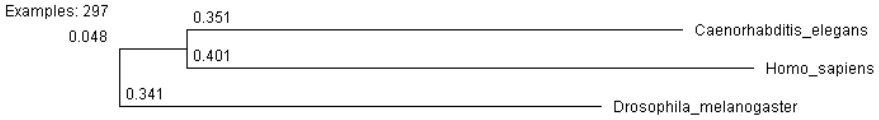

Figure 8.7: Anomalous triple II: This result is contrary to current thought that *C. elegans* and *D. melanogaster* share a common ancestor not shared by *H. sapiens*.

It is clear from Figures 8.6 and 8.7 that the evolutionary distances from the OTUs and their respective common ancestors are relatively large indicating a distant evolutionary divergence. The resolution of distant phylogenies such as this are less reliable (see Section 2.3.3) so the pattern mining criteria were modified in phase II.

8.6 Phase II

Phase II employs a search for the pattern described by a phylogenetic triple where A and B both have a common immediate ancestor. In phase I, B simply had an ancestor that was the immediate ancestor of A. The new pattern or structure (see Figure 8.1) is defined by the following rule:

```
triple(A, B, C, Eva, Evb, Evc, Evd, Tree, Candidate):-
   A = Candidate,
   edge(A, X, Eva, Tree),
   edge(B, X, Evb, Tree),
   edge(C, Y, Evc, Tree),
   ancestor(X, Y, Evd, Tree),
   not(A = B),
   not(A = C),
   not(B = C),
   organism(B),
   organism(C).
```

This phylogenetic triple is more constrained than that in phase I and should reduce or possibly even eliminate unreliably distant relationships. Exactly the same methods used in phase I were used in phase II with this modified phylogenetic triple.

8.6.1 Analysis of Phase II Results

It was thought that reducing the confidence threshold, removing triples in which we have less confidence, would produce a set more consistent with accepted phylogeny. However, by taking the evolutionary relationship between the fungi, *S. cerevisiae* and *S. pombe* as an example, we find that the most significant result in Table 8.4 is the one considered to be anomalous.

Presently, it is commonly accepted that *S. pombe* diverged from the ancestral line before *S. cerevisiae*. If that is true, then the first entry in Table 8.4, the one in

A	B	C	N	Pmf
N. crassa	S. pombe	S. cerevisiae	20	9.768e-011
S. cerevisiae	C. albicans	S. pombe	15	5.24333e-007
S. cerevisiae	A. gossypii	S. pombe	13	9.39519e-006
S. cerevisiae	C. elegans	S. pombe	19	5.48882e-005
C. glabrata	S. cerevisiae	S. pombe	10	0.000545159

Table 8.4: The top five most significant examples of phylogenetic triples from the *phase II* results that include both *S. pombe* and *S. cerevisiae*. A and B share a common ancestor not shared by C. The number of examples of the triple found is denoted by N . The Pmf figure in the last column is the probability mass function, which is used as a measure of significance.

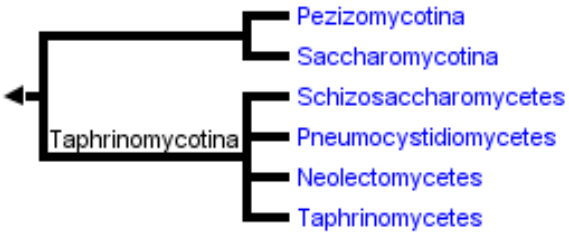

Figure 8.8: The taxonomy of the subphyla of *N. crassa*, *S. cerevisiae* and *S. pombe* taken from the Tree of Life Project. This subtree shows that *pezizomycotina*, the subphylum of *N. crassa*, and *saccharomycotina*, the subphylum of *S. cerevisiae*, are grouped together in a clade to which *taphrinomycotina*, the subphylum of *S. pombe*, does not belong.

which we have most confidence, presents a problem. All three organisms are classified under the phylum *ascomycota*, but all three belong to different subphyla. Using information from the NCBI taxonomy browser[2] we find *N. crassa* belongs to the subphylum *pezizomycotina*; *S. cerevisiae* belongs to the subphylum *saccharomycotina* and *S. pombe* belongs to the subphylum *taphrinomycotina*. The phylogeny of these subphyla are given by the Tree of Life project (Maddison *et al.* , 2007) as shown in Figure 8.8 and reveal that *pezizomycotina* and *saccharomycotina* are grouped together in a clade to which *taphrinomycotina* does not belong. From this we can conclude that the taxonomy of these three organisms should be

[2]The NCBI taxonomy database is not a phylogenetic or taxonomic authority and is used in this context as merely a guide.

correctly represented by the triple {{*N. crassa, S. cerevisiae*} *S. pombe*}.

Further investigation of the anomalous triple {{*N. crassa, S. pombe*}*S. cerevisiae*} in the results revealed that all 20 examples of this triple, which are all from separate trees, comprised the same protein sequences from each organism. This is a problem which has arisen because each example of this triple was selected due to the small evolutionary distance between each organism, but none of the protein sequences representing the organism in these examples may have been necessarily the most similar protein sequence to the seed sequence that was used by the BLAST search to isolate the members of the tree. A refinement to the selection procedure for tree members would be necessary, whereby each tree should contain only one protein sequence for each representative organism and that should be the sequence most similar to the seed sequence. Although we lose interesting evolutionary relationships between many potentially homologous proteins, this method should suffice for determining the taxonomy of organisms.

8.7 Results

The resulting triples from phase II were combined into a consensus tree using Aho's algorithm. The resulting phylogenetic tree diagrams in the following Figures 8.9, 8.10, 8.11, 8.12 and 8.13 were produced using a phylogenetic tree graphics package called Dendroscope (version 1.4, 28th July 2008) (Huson *et al.* , 2007). These diagrams are a graphic representation of the phylogenetic consensus trees generated using Aho's algorithm (Aho *et al.* , 1981).

The phylogenetic consensus tree in Figure 8.9 was generated from a subset of triples where the C node was represented by the flowering plant *A. thaliana* and the confidence threshold was set to 1×10^{-3}. By doing this the consensus tree should contain all those organisms that form a clade to which *A. thaliana* does not belong and so we would expect that the tree should contain mostly plants. However, at the top of the figure we see a relatively large group of animals and fungi. Further down there are groups of plants interspersed with fungi and bacteria. This is not very informative and is most likely due to smaller trees in the knowledge base

where the immediate ancestor of *A. thaliana* is a node representing a large clade such as a domain or kingdom.

The results for all organisms given in Figures 8.10, 8.11, 8.12 and 8.13, which represent a single consensus tree split into 4 pages for easier viewing. The results could only be obtained where the confidence threshold[3] was set to 1×10^{-19} resulting in a consensus tree containing only 209 OTUs. It is encouraging that most genus names and some known close relatives have been grouped together, but the lack of depth in the consensus tree renders it largely uninformative.

[3]Note that we are selecting those triples where the PMF value is lower than the threshold indicating a high confidence.

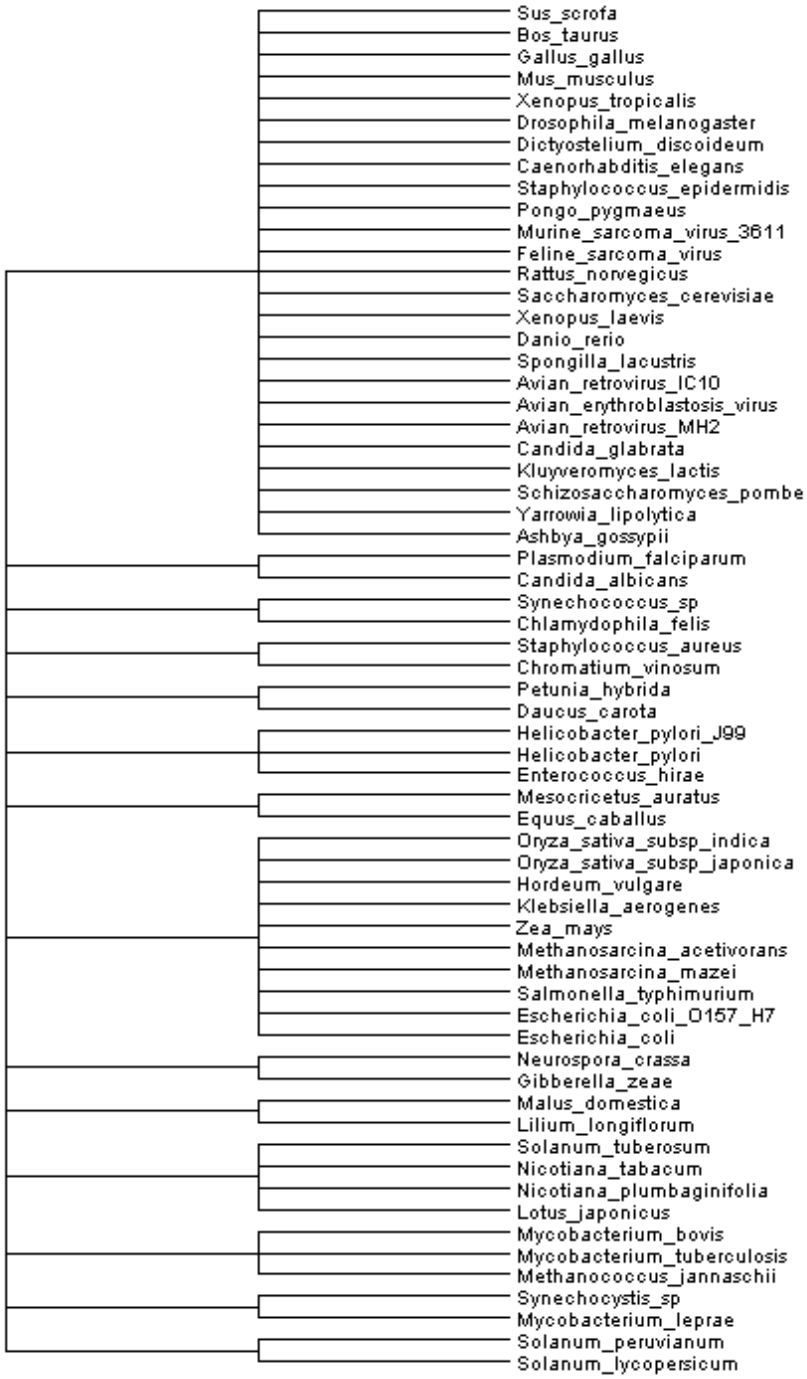

Figure 8.9: A consensus tree drawn from all organisms (63 in total) in a clade to which *A. thaliana* does not belong. It was expected that the tree should contain mostly plants. However, at the top of the figure we see a relatively large group of animals and fungi. Further down there are groups of plants interspersed with fungi and bacteria. This is most likely due to smaller trees in the knowledge base where the immediate ancestor of *A. thaliana* is a node representing a large clade such as a domain or kingdom.

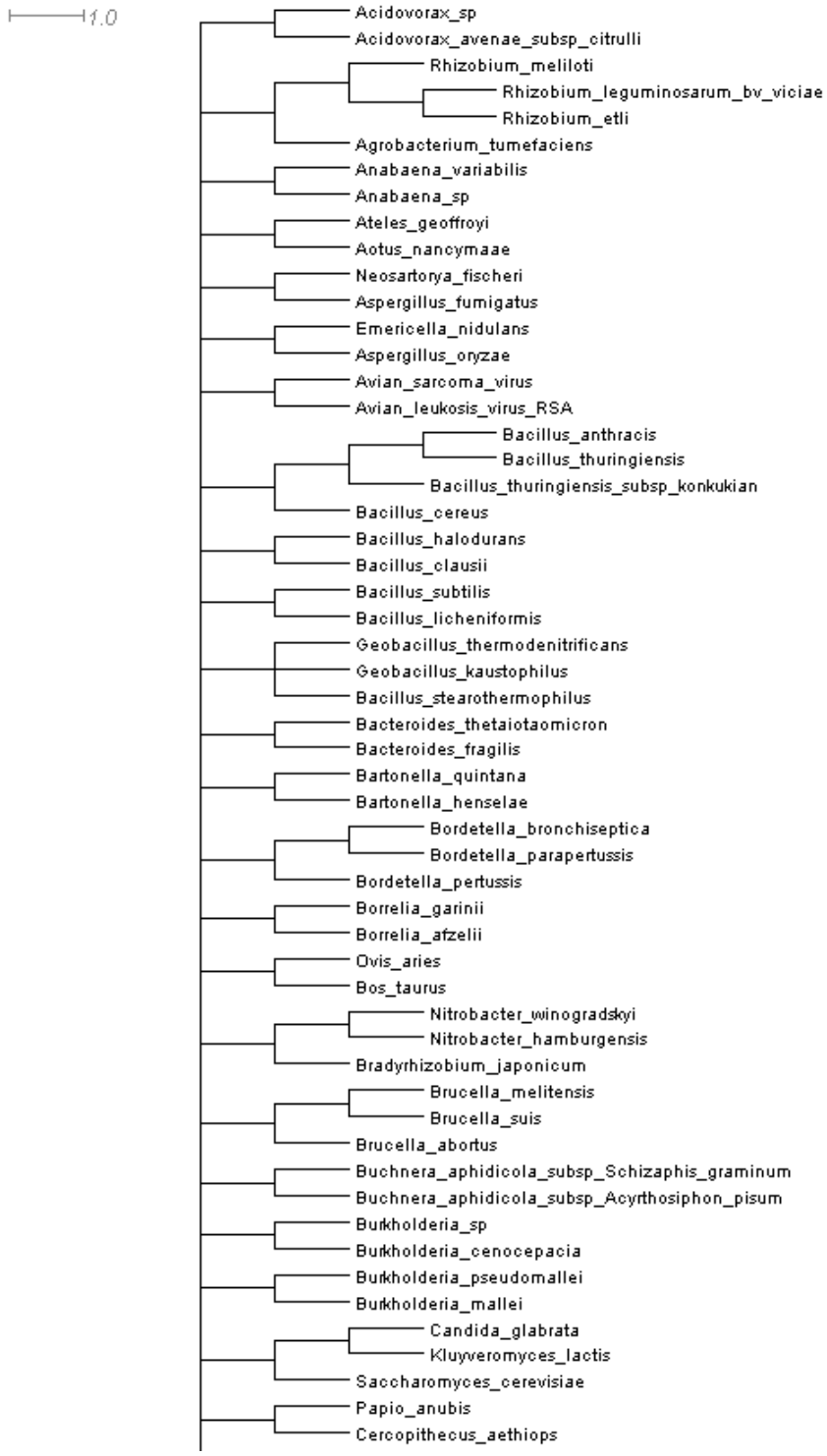

Figure 8.10: Final consensus tree with 209 organisms in total where the confidence threshold is 1×10^{-19} (Part I).

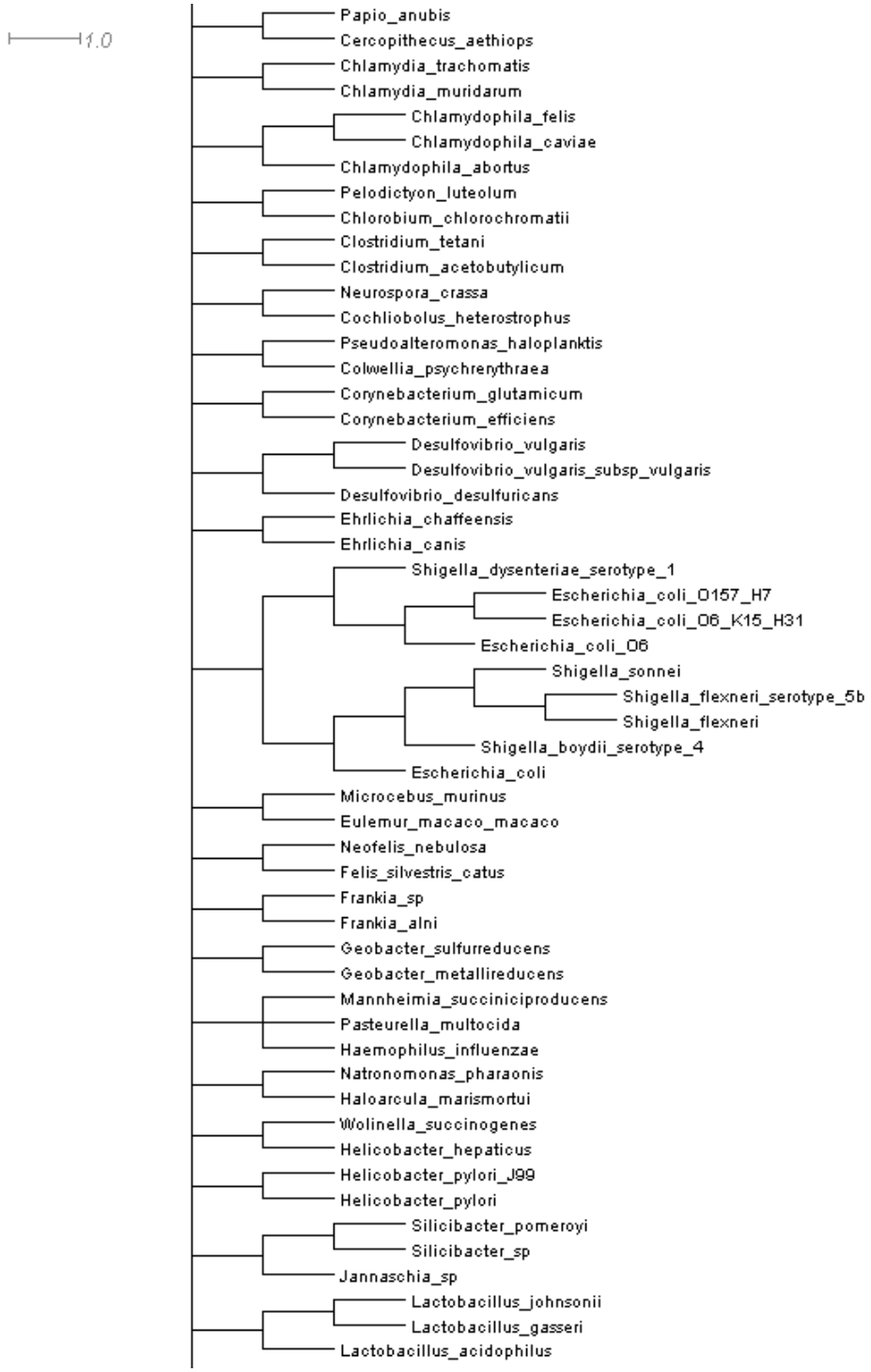

Figure 8.11: Final consensus tree with 209 organisms in total where the confidence threshold is 1×10^{-19} (Part II).

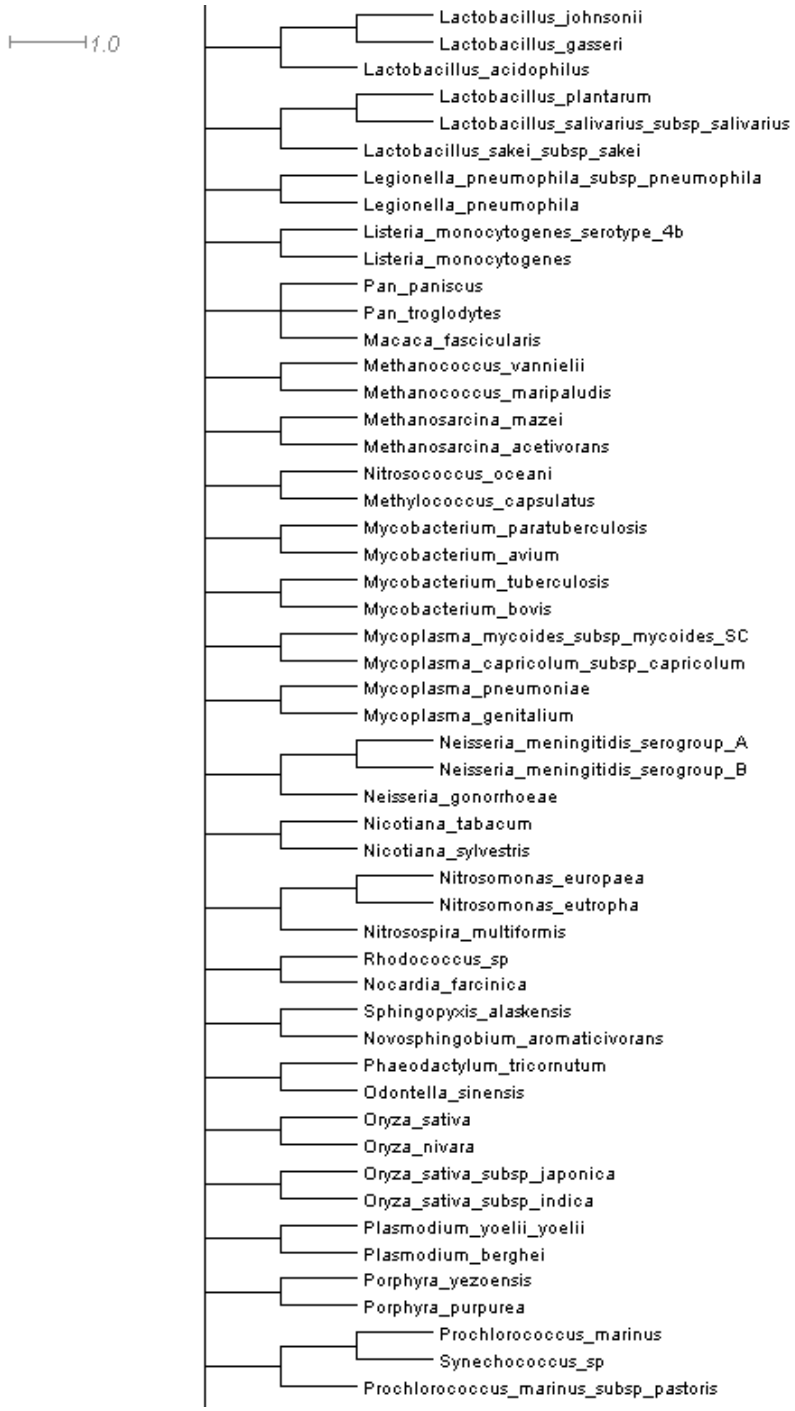

Figure 8.12: Final consensus tree with 209 organisms in total where the confidence threshold is 1×10^{-19} (Part III).

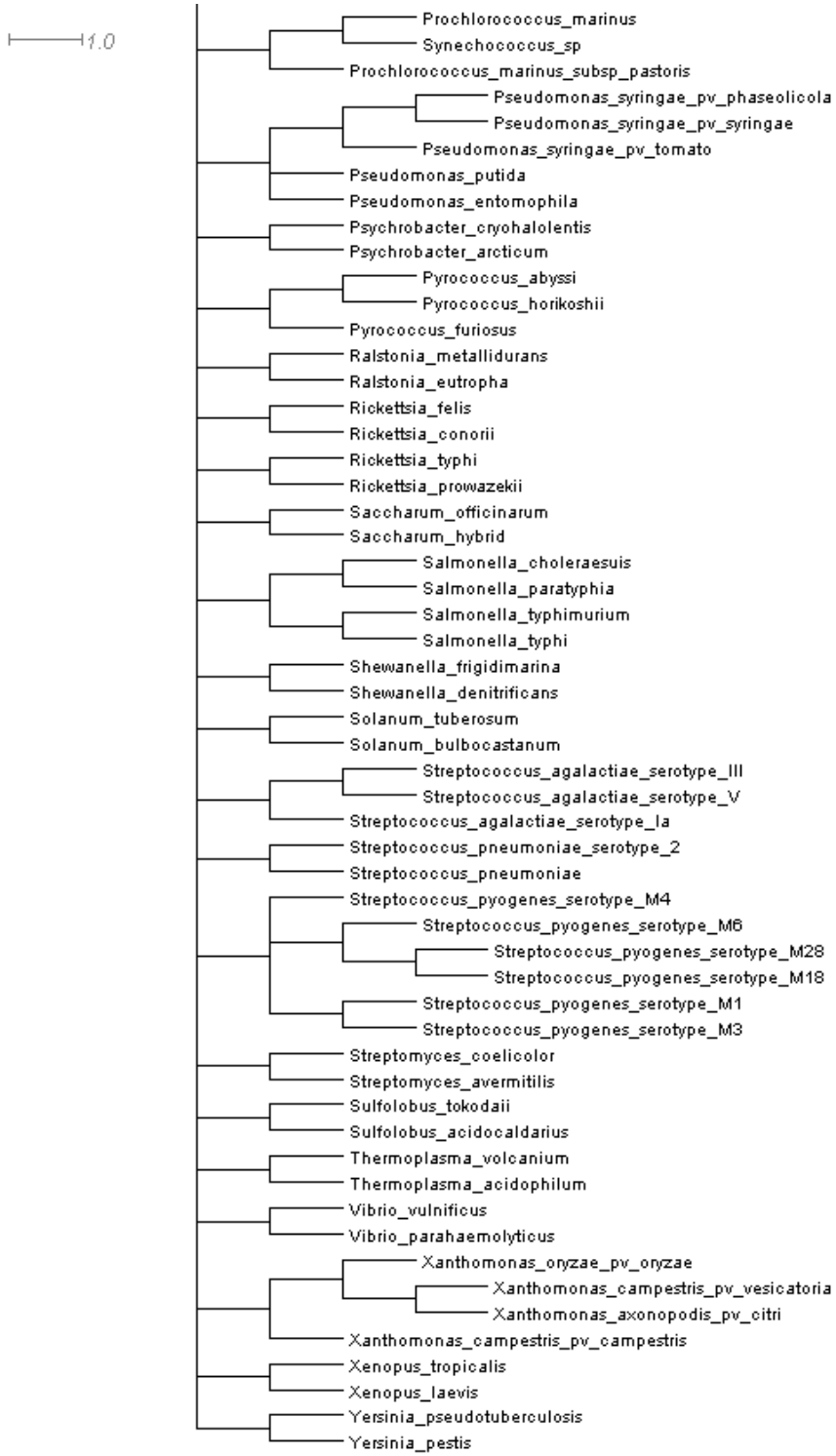

Figure 8.13: Final consensus tree with 209 organisms in total where the confidence threshold is 1×10^{-19} (Part IV).

8.8 Conclusion

The knowledge base contains a considerable quantity of data, but the filtering process used to produce confident results has eliminated a relatively large quantity of potentially useful data. The lack of depth and the number of unassigned organisms in the resulting consensus tree is due to an insufficient number of phylogenetic triples. Future work will need to address these problems and this is discussed in Chapter 9.

Chapter 9

Discussion

9.1 Introduction

The whole process of knowledge acquisition in bioinformatics through significant pattern discovery relies heavily on combinations of statistical analysis and various methods from the field of computer science. This process has been applied to two principal areas in bioinformatics: gene location, which is a new area of research in epigenetics, and phylogenetics, which is an area within bioinformatics of rapidly increasing popularity.

There are two fundamental areas of research in bioinformatics: dealing with very large databases and making sense of the data.

Dealing with very large databases relies on many techniques from the computer science field to address the problems associated with them. Essentially, the problems are centred around the computational time required for processing. Frequent pattern mining is a case in point in that even with fast computers the quantity of data to be mined makes frequent pattern mining computationally time consuming. For example, the frequent pattern search of the topological distance between homo sapiens and all other eukaryote organisms in the phylogenetic database outlined in Chapter 7 has taken 2–3 days on 3GHz, 500Mb personal computer. This was considered a fairly simple search.

Datalog and logic programming have been employed in this research as this methodology has proved to be efficient with databases in general. However, for large number crunching procedures such as Monte Carlo methods, C++ was found to be a more efficient language to work with. Java and R were better suited for graphical representation for the visualization of results. In short, we chose the tool that best suited the job.

Making sense of the data or knowledge acquisition is the main purpose of bioinformatics. There are many methods to achieve knowledge acquisition, which all depend on what is *interesting* to the researcher. In this research we have used significant frequent pattern mining methods, because significance in the frequency of structures or patterns is what we deem to be interesting. Frequency alone does not necessarily correspond with patterns of interest in bioinformatics. We have also used the degree of significance as a measure of confidence in the interesting patterns discovered.

We continue this chapter with a review of the research detailed in Chapters 5, 6, 7 and 8. This is followed by a summary of the main findings, discoveries and results. The final two sections discuss the problems and future directions for significant pattern discovery in epigenetics and improving resolution in phylogenetics using frequent structure mining.

9.2 Review

From the research described in Chapter 5 the location of genes on the genome of *Arabidopsis thaliana* are found to have unknown elements or order. Both a conventional statistical sampling method and Greenwood's statistic were used in this analysis. Greenwood's statistic was used to detect clustering and uniformity in the spatial location of genes because it is a very sensitive measure and works well with smaller numbers of sample data. However, the extra sensitivity of this measure is a compromise in that it can tell us very little about the nature of the discovered clusters or uniformity in gene location. Greenwood's statistic is a relative metric and to be effective we used Monte Carlo methods to establish a null

hypothesis. By comparison of the resulting statistic to the Greenwood statistic for the null hypothesis, we can determine the significance.

The development and application of a system for determining the significance of frequent patterns, called the SPD (Significant Pattern Discovery) system, was the focus of Chapter 6. The system was capable of determining significance using both conventional statistical methods and Monte Carlo methods. The SPD system should be capable of revealing much more information about the nature of the distributions of locations of genes from Chapter 5. However, the change of model organism may have complicated this research. This is explained in more detail below.

In Chapter 7 we created a large database of phylogenetic trees. Each tree plotting the evolutionary history of a single superfamily class of proteins. The database was queried using a phylogenetic pair structure returning data on the relationships between organisms. There were anomalies that revealed necessary enhancements to the database. An in-depth analysis of one particular anomaly revealed the necessity of reducing the E-value threshold in the initial BLAST sequence alignment procedure for future databases.

In Chapter 8 a new database was created based on the information obtained in the last chapter. One type of phylogenetic triple was used to extract phylogenetic relational data. This data was evaluated by comparison to an existing commonly accepted phylogeny for yeast species. The discovered anomalies revealed the need to constrain the phylogenetic triple parameters. The evaluation of the corresponding results exposed further anomalies indicating the need for more enhancements. However, using the existing results a trial consensus tree was constructed to evaluate the performance of the consensus tree building algorithm.

The continuous process of evaluation, modification and refinement of data throughout the research described in Chapters 7 and 8 has been shown to be effective in improving the existing results and providing foundations for further improvements.

9.3 Key Findings and Results

In Chapter 5, the statistical analysis of the distribution of gene location on the genome of *Arabidopsis thaliana* revealed the following:

- The locations of genes on the genome of *Arabidopsis thaliana* are more clustered than would be expected from a locationally independent distribution (also referred to as a uniform probability distribution).
- Tandem duplications were thought to be the main cause of gene clustering, but we found that significant clustering was still present after removal of all tandem duplicates.
- There are marked differences in the degree of clustering of genes of certain molecular functions. Genes associated with catalytic activity, transporter activity, binding, enzyme regulator activity and transcription regulator were all significantly clustered. Less clustered are genes of signal transducer activity, and genes associated with structural molecule activity, anti oxidant activity, translation activity and nutrient reservoir showed no clustering at all.
- There is evidence of significantly even distributions in genes associated with calcium ion binding, G-protein coupled receptor activity and metallopeptidase activity.

In Chapter 6, a first order pattern discovery system is used to discover significant patterns in the location of genes. In this chapter we find the following:

- 87% of all genes smaller than 360 bp in length have unknown molecular function.
- Converging and diverging gene pairs are more frequent than consequent gene pairs.
- The nature of the gap lengths between diverging gene pairs are very different to any other type of neighbouring gene pairs.
- There is some significance in the length of 250–300 nucleotides as there are

frequent peaks of this length in the graphs of gap length between neighbouring pairs.

- There are no significant patterns in the molecular functions of neighbouring gene pairs except for pairs both of structural molecule activity. The frequency of these pairs is significant.
- Localized clusters of genes of different molecular function are no more frequent than we would expect from a locationally independent distribution.
- There are very significant patterns in locations of genes of different molecular function that are dispersed rather than localized.

In Chapter 7 we create a large protein structural phylogenetic database and introduce methods for pattern mining within the database. From this we learned:

- Evolutionary distances used in tree mining algorithms are tree specific and cannot be used for comparisons across different trees.
- The setting of the BLAST E-value threshold is critical in order to extract sufficient samples of protein sequences for analysis, but avoiding sequences that are too distantly related.
- Data mining using phylogenetic substructures representing the phylogenetic relationship between pairs of organisms showed great promise, but revealed problems with the database.

In Chapter 8, we refined the database outlined in Chapter 7 and then applied frequent phylogenetic substructure mining with a view to generate a consensus tree of high resolution. We found:

- Frequent substructure mining using substructures of phylogenetic triples extracted reliable elements for phylogenetic determination, but further research on the precise structure of the triples used is required.
- Although the database had been refined, a further refinement is required to ensure that each organism is represented by only one protein sequence that is the most homologous to the model protein sequence for each structural

classification.

- Distant evolutionary relationships in phylogenetic trees are unreliable. This problem is inherent in most, if not all, tree building algorithms.

No reliable biological knowledge could be obtained from the research in either Chapters 7 or 8, but this ground work may prove invaluable in the future for research in phylogenetics.

9.4 Significant Pattern Discovery in Epigenetics

The main motivation for the SPD pattern mining system outlined in Chapter 6 is in the need to determine significance from the frequent pattern results produced by WARMR. Inspired by the results from the previous research presented in Chapter 5, it was hypothesized that frequent pattern mining would reveal more information about the nature of the order in gene location.

The previous research in Chapter 5 indicated elements of order in the locations of genes in general, and also order in the locations of genes classified by their molecular function in the genome of the flowering plant *Arabidopsis thaliana*. The model organism chosen for the significant pattern mining research was the yeast fungus *Saccharomyces cerevisiae*. The obvious question at this point is, why switch to a different model organism?

At the time of this research (2006–2007) there was considerable excitement about a new Robot Scientist project at this University's Computer Science Department (King *et al.* , 2004). This project focussed on very high throughput of biological experiments on *Saccharomyces cerevisiae*. By switching the model organism in this research to the same model organism used in the Robot Scientist project, there was a potential for interesting and rewarding collaboration. Although this opportunity never came to fruition, the switch of model organisms may have unintentionally revealed a correlation between gene location and phenotype development in that

the nature of gene location in *Arabidopsis thaliana* , a multi cellular organism, is very different from *Saccharomyces cerevisiae*, a single celled organism.

In Chapter 6 the results presented firmly indicated that there are no significant patterns in the localized clustering of genes all having different molecular functions. Although this is not entirely a contradiction of the previous research in Chapter 5, it is, however, an unexpected result. Succeeding research in gene location in Chapter 6 then showed significant frequent patterns in the locations of genes of different molecular functions dispersed along the genome. This is curious and warrants a further study which is likely to go beyond pattern mining and/or involve advanced novel pattern mining techniques.

The SPD system has made some interesting discoveries, but it is still unclear what exactly the Greenwood statistic has detected in the distribution of genes on the genome of *Arabidopsis thaliana*. It would seem that a comparative study of the epigenetics of *Arabidopsis thaliana* and *Saccharomyces cerevisiae* is required to clarify this.

9.5 Improving Resolution in Phylogenetics using Frequent Structure Mining

Higher resolution in protein structural phylogenetic trees is popularly achieved by using as much available data as possible to create them. These data are remarkably noisy and frequently contain contradictions. The research discussed in Chapters 7 and 8 sought to ameliorate this through the use of frequent phylogenetic substructure mining. Two substructures were subjected to experimentation. Phylogenetic pairs represent the evolutionary relationship between two organism and phylogenetic triples represent the relationship between three organisms where two organisms are members of a clade to which the third does not belong. Phylogenetic pairs and triples serve as a strategy to overcome data sampling impediments by precisely defining discrete phylogenetic relationships.

There were problems with the determination of the precise structures of the sub-

structures used for the most efficient data extraction. This area would benefit from further research. However, using the preliminary substructures we did obtain some data, which was considered at the time to be sufficient for the construction of a consensus tree.

9.5.1 Phylogenetic Consensus Tree

In Chapter 7 we discuss the creation of a large phylogenetic database of protein structural phylogenetic trees. Although this database is intrinsically of great value in protein research, the eventual objective was to generate a potentially highly reliable consensus tree from this database. This is the work covered in Chapter 8. A consensus tree was created, but it was clear in the early stages of this research that there would be problems and that, for the creation of a consensus tree, a very different approach in the generation of the original phylogenetic database would be required.

The evolutionary distances generated by ClustalW are specific to each tree and cannot be used as a universal measure of evolutionary distance. This is a consequence of the *Neighbor Joining* method used by ClustalW and there is no evidence to suggest that any other tree generating algorithm does not suffer from this short coming. Therefore, the inclusion of this data in the database is largely superfluous, but it can be used for intrinsic comparisons within each tree. Unfortunately, before this fact was discovered, the evolutionary distances had been used to determine phylogenetic triples most representative of each protein structural classification. This approach also presented a further problem where the members of each representative triple may not necessarily be most representative of the classification.

The database required for consensus tree generation should contain only one protein sequence from each organism that best represents the protein structural classification.

The choice of protein structural classification used in this research was the *superfamily* classification, but this may result in the inclusion of very distant homologs

representing less well studied organisms. This could produce erroneous positioning on the phylogenetic trees. The more specific *family* classification may produce more reliable trees in this respect.

Another potential refinement is in the detection of homologous sequences. We used BLAST, which more suited to sequence similarity rather than sequence homology. Using more homology based search algorithms such as PSI-BLAST (Altschul *et al.* , 1997) or a 'home grown' methodology, HI (Homology Induction) (Karwath & King, 2002), could prove to be beneficial.

A further refinement could be to select protein structural families that have a higher phylogenetic resolution. Comparing the phylogenies of different families of proteins from well known organisms to the putative phylogeny of these organisms could determine those protein families that more phenotype specific. These structural protein classifications should provide much more reliable results.

Chapter 10

Conclusions

The principal theme throughout this research has been the determination of the significance of the frequency of discovered patterns. Conventional statistical methods, which determine expectation and then use the probability of the extent of the deviation from the expectation as a measure of significance, have proved to be effective in many cases. However, where this statistical method fails, the application of Monte Carlo methods to establish a null hypothesis and a system of ranking the frequency of the discovered pattern in the original data against the frequencies of the same pattern found in many random distributions, has proved very effective.

The principal tools have been drawn from the field of logic programming. Representing the database and the candidate patterns in Datalog allows the use of logic programming for the analysis. The benefits are higher processing speeds of data mining, because of the highly optimized search algorithms, which are central to the logic programming methodology. Another benefit is the ease with which discovered structures or patterns represented in Datalog can be incorporated into the background knowledge.

The requirements for efficient significant frequent pattern mining of databases in Datalog schema necessitated the development of the Significant Pattern Discovery (SPD) system. This is a hybrid system centred around the WARMR frequent

pattern mining program. The filtering of results and the significance of frequent patterns along with ad hoc analyses are all provided in the Prolog and C++ programming languages in order to fully utilize the strengths of each programming language. The SPD system is a novel approach which enhances frequent pattern mining by providing fast efficient procedures for the discovery of interesting patterns.

Other novel methods used in this research are the effective application of the Greenwood statistic on sparse data together with the use of Monte Carlo methods and the use of logic programming in frequent substructure mining in phylogenetic tree structures represented in Datalog schema.

There are several key discoveries from the research presented in this thesis. This research has revealed elements of order in the physical location of genes in the genome of the flowering plant *Arabidopsis thaliana*. Further research, using the SPD system for significant pattern discovery, on the location of genes on the genome of the fungus *Saccharomyces cerevisiae*, showed an unexpected nature to the order in gene location. The SPD system also efficiently discovered significant patterns in neighbouring pairs of genes suggesting the possibility of a system of co-operative gene expression in *Saccharomyces cerevisiae*.

Research in the increasingly popular field of phylogenetics resulted in the creation of a protein structural phylogenetic database, which revealed previously unseen requirements. One such requirement would be a refinement in the selection or generation of model protein sequences representing the same protein domain over many disparate clades.

Moreover, the application of logic programming methods and significant pattern mining in the creation of phylogenetic consensus trees has demonstrated further requirements of databases of protein phylogeny. One consideration is the limitation of the number of organisms to clades with more recent common ancestry. Analysis of large phylogenetic trees quickly becomes intractable for large numbers of organisms (presently 200 on a personal computer). The overall phylogenetic relationship between these clades could be determined using maximum likelihood methods to determine the most likely ancestral protein sequences for

each clade.

The research described in this thesis has brought together various methods and techniques from both mathematics and computer science for the discovery of knowledge in bioinformatics. Several of these methods and techniques are novel and have proved their worth in various new discoveries. This research has achieved its main objective in unveiling new directions for future research to unravel the complex network of information contained within the molecular biology of living organisms.

Appendix A

Supplementary Tables for the Location of Genes in *A. thaliana*

A.1 Locational Distribution of Gene Functional Classes in *Arabidopsis thaliana*

A.1.1 Key to tables

The following list is a full description of the column headings in the tables used in this document.

Chr: Chomosome number.

Class: A number representing the molecular function classification annotated by Gene Ontology.

Std: Strand (Watson/Crick).

Orig. Grnwd: The Greenwood statistic for the distribution of genes on the original chromosomes.

MC Grnwd: The mean Greenwood statistic for the distributions of genes from 1000 pseudo-randomly generated chromosomes.

SD MC grnwd: The standard deviation for the above statistic.

Ranking: The ranking of the original result compared with 1000 simulated results.

Examples: The number of examples of molecular function class used.

Chr	Class	Std	Orig. Grnwd	MC Grnwd	SD MC grnwd	Ranking	Examples
1	03824	W	0.00461086	0.00345593	0.000321756	1000	872
1	03824	C	0.00428771	0.0037403	0.000265304	940	884
1	04871	W	0.0266987	0.0272012	0.00426267	543	82
1	04871	C	0.0359752	0.0339515	0.00627094	724	67
1	05198	W	0.030008	0.0320846	0.00512607	406	68
1	05198	C	0.034087	0.039007	0.00706869	265	58
1	05215	W	0.013537	0.00962302	0.00102992	997	251
1	05215	C	0.0139946	0.0115202	0.0014391	939	217
1	05488	W	0.00517308	0.00328976	0.00033366	1000	888
1	05488	C	0.00361528	0.00365203	0.000259156	515	916
1	16209	W	0.123493	0.15363	0.0390906	204	12
1	16209	C	0.194071	0.182619	0.0493941	702	10
1	30234	W	0.125975	0.0894086	0.0189708	956	22
1	30234	C	0.102725	0.0681844	0.0140613	977	31
1	30528	W	0.0167562	0.0102226	0.0010794	1000	236
1	30528	C	0.0132615	0.0113704	0.00138724	904	220
1	45182	W	0.0651981	0.0720028	0.0143043	360	28
1	45182	C	0.0775567	0.100828	0.0226898	104	20
1	45735	W	0.282141	0.2374	0.0656962	805	7
1	45735	C	0.109194	0.1462	0.0382068	89	13

Table A.1: Results for gene classes taken from level 1 of the Gene Ontology heirarchy on chromosome 1 using TIGR data with the tandem duplications removed.

Chr	Class	Std	Orig. Grnwd	MC Grnwd	SD MC grnwd	Ranking	Examples
2	03824	W	0.00493668	0.00518015	0.000450868	283	505
2	03824	C	0.00599414	0.00461907	0.00037109	992	565
2	04871	W	0.0845108	0.0672586	0.0153843	888	32
2	04871	C	0.127841	0.0590474	0.0134551	997	37
2	05198	W	0.0453661	0.045425	0.00901902	587	50
2	05198	C	0.0466727	0.0553937	0.0116024	227	40
2	05215	W	0.0200676	0.0161908	0.00225987	946	147
2	05215	C	0.0302128	0.018107	0.00269904	996	131
2	05488	W	0.00425628	0.00453212	0.000386012	246	608
2	05488	C	0.00436702	0.00445364	0.00034135	449	589
2	16209	W	0.192492	0.197252	0.0474917	522	9
2	16209	C	0.228358	0.212194	0.0536405	703	8
2	30234	W	0.185141	0.123328	0.0302565	956	16
2	30234	C	0.142932	0.0981309	0.0229557	948	21
2	30528	W	0.0393378	0.0172058	0.0023468	1000	137
2	30528	C	0.0558571	0.0189399	0.00290322	1000	125
2	45182	W	0.155136	0.166668	0.0403859	440	11
2	45182	C	0.138527	0.166958	0.0398796	263	11
2	45735	W	0.434175	0.3492	0.101882	846	4
2	45735	C	0.233419	0.192833	0.0493644	833	9

Table A.2: Results for gene classes taken from level 1 of the Gene Ontology heirarchy on chromosome 2 using TIGR data with the tandem duplications removed.

Chr	Class	Std	Orig. Grnwd	MC Grnwd	SD MC grnwd	Ranking	Examples
3	03824	W	0.00746425	0.00583099	0.000814598	977	643
3	03824	C	0.00672773	0.00540192	0.000715543	954	678
3	04871	W	0.0551221	0.0482816	0.010014	811	49
3	04871	C	0.0960669	0.0520379	0.0117095	996	47
3	05198	W	0.0425475	0.0380615	0.00816041	781	64
3	05198	C	0.055678	0.0371552	0.00804732	964	69
3	05215	W	0.0222488	0.0185479	0.00321253	902	146
3	05215	C	0.0417299	0.0156211	0.00277516	1000	183
3	05488	W	0.00846204	0.00574346	0.000765179	1000	659
3	05488	C	0.00760123	0.00505556	0.000490969	1000	689
3	16209	W	0.126566	0.171025	0.0473813	128	11
3	16209	C	0.149328	0.192084	0.0553073	207	10
3	30234	W	0.149969	0.101144	0.0242922	948	21
3	30234	C	0.106659	0.0906837	0.0234184	812	25
3	30528	W	0.0491048	0.0168493	0.00263517	1000	162
3	30528	C	0.0535272	0.0197254	0.0039783	1000	141
3	45182	W	0.281498	0.225753	0.0659261	849	8
3	45182	C	0.198122	0.129569	0.0353808	956	16
3	45735	W	0.298559	0.277851	0.0772352	689	6
3	45735	C	0.296933	0.16614	0.0488654	978	12

Table A.3: Results for gene classes taken from level 1 of the Gene Ontology heirarchy on chromosome 3 using TIGR data with the tandem duplications removed.

Chr	Class	Std	Orig. Grnwd	MC Grnwd	SD MC grnwd	Ranking	Examples
4	03824	W	0.0050018	0.00498841	0.000377788	585	500
4	03824	C	0.00538837	0.00519956	0.00043766	669	537
4	04871	W	0.0483991	0.0484439	0.00884644	589	43
4	04871	C	0.0407367	0.0435496	0.00878347	441	51
4	05198	W	0.0552778	0.0554576	0.0102561	585	37
4	05198	C	0.0598803	0.0602493	0.0118551	548	36
4	05215	W	0.0194584	0.0190879	0.00244008	629	117
4	05215	C	0.0258559	0.0175382	0.00277058	980	137
4	05488	W	0.00509508	0.00471058	0.000390767	831	545
4	05488	C	0.00582023	0.00542012	0.000408192	833	514
4	16209	W	0.130003	0.140046	0.0342328	455	13
4	16209	C	0.234765	0.237411	0.0604259	572	7
4	30234	W	0.114209	0.118632	0.0260909	505	16
4	30234	C	0.159594	0.211018	0.0552698	118	8
4	30528	W	0.0216402	0.0173024	0.00228024	961	130
4	30528	C	0.0293432	0.0203354	0.00323235	979	117
4	45182	W	0.149115	0.161546	0.0400584	446	11
4	45182	C	0.148636	0.154753	0.0391065	552	12
4	45735	W	0.234971	0.260521	0.071213	442	6
4	45735	C	0.231433	0.26178	0.0668359	383	6

Table A.4: Results for gene classes taken from level 1 of the Gene Ontology heirarchy on chromosome 4 using TIGR data with the tandem duplications removed.

Chr	Class	Std	Orig. Grnwd	MC Grnwd	SD MC grnwd	Ranking	Examples
5	03824	W	0.00470506	0.00374654	0.000359107	980	756
5	03824	C	0.00741941	0.00368441	0.000372623	1000	783
5	04871	W	0.0327333	0.0290924	0.00574659	795	81
5	04871	C	0.0362551	0.0398035	0.00895303	422	61
5	05198	W	0.0455819	0.0392262	0.00769359	833	58
5	05198	C	0.0386574	0.0384007	0.00822679	628	63
5	05215	W	0.0144524	0.0130435	0.0022869	812	191
5	05215	C	0.0217139	0.0126649	0.00248345	987	209
5	05488	W	0.00420059	0.00341291	0.000251068	979	862
5	05488	C	0.00400692	0.00338991	0.000343469	948	866
5	16209	W	0.212407	0.244203	0.076583	408	7
5	16209	C	0.122518	0.113782	0.0272106	721	18
5	30234	W	0.140406	0.117355	0.0280249	848	17
5	30234	C	0.10396	0.0925238	0.0203716	767	23
5	30528	W	0.0282562	0.0116901	0.00161739	1000	213
5	30528	C	0.0144269	0.0125376	0.00244378	796	211
5	45182	W	0.111221	0.103659	0.0252331	708	20
5	45182	C	0.0835833	0.0994304	0.0217667	240	21
5	45735	W	0.205942	0.218595	0.0618135	521	8
5	45735	C	0.109241	0.128322	0.033335	320	16

Table A.5: Results for gene classes taken from level 1 of the Gene Ontology heirarchy on chromosome 5 using TIGR data with the tandem duplications removed.

Chr	Class	Std	Orig. Grnwd	MC Grnwd	SD MC grnwd	Ranking	Examples
1	00166	W	0.0107258	0.00937975	0.00102211	891	258
1	00166	C	0.0129371	0.00863453	0.000971031	1000	301
1	03676	W	0.00883573	0.00583013	0.000557914	1000	441
1	03676	C	0.00627288	0.00602496	0.000603791	719	461
1	03700	W	0.0170799	0.0107792	0.00131015	1000	224
1	03700	C	0.0134604	0.0121163	0.00160345	840	207
1	03735	W	0.0396305	0.0448146	0.00736639	249	48
1	03735	C	0.0382082	0.0501455	0.00941507	46	43
1	03754	W	0.082164	0.07455	0.013918	750	27
1	03754	C	0.0564661	0.0556431	0.0106559	611	39
1	03793	W	0.139116	0.0967526	0.0197183	964	20
1	03793	C	0.100602	0.110653	0.0255398	409	18
1	04386	W	0.154128	0.102456	0.0217551	967	19
1	04386	C	0.0724309	0.0646965	0.0128912	775	33
1	04857	W	0.189537	0.142903	0.0363572	917	13
1	04857	C	0.151728	0.111908	0.0252016	926	18
1	04872	W	0.0542422	0.0447582	0.00787169	901	48
1	04872	C	0.0502293	0.0521483	0.0107102	509	42
1	05386	W	0.0317418	0.0232466	0.00346521	970	98
1	05386	C	0.0274814	0.0261488	0.00443165	710	89
1	05489	W	0.0396485	0.0400886	0.00665184	535	54
1	05489	C	0.0435345	0.0468206	0.00883886	409	47
1	05515	W	0.0237995	0.0216625	0.00311658	798	105
1	05515	C	0.0200567	0.0196699	0.00295315	629	121
1	08135	W	0.0651981	0.0726647	0.0142764	327	28
1	08135	C	0.0775567	0.100526	0.0229796	105	20
1	08289	W	0.0839434	0.117092	0.0263293	29	16
1	08289	C	0.16031	0.145564	0.0377233	736	13
1	08565	W	0.126357	0.0942259	0.0200566	937	21
1	08565	C	0.139498	0.135878	0.0348166	651	14
1	15034	W	0.112825	0.0976714	0.0212671	826	20
1	15034	C	0.0914775	0.111378	0.027302	226	18
1	15075	W	0.0602414	0.0434953	0.00752369	960	49
1	15075	C	0.0680512	0.0493716	0.00901085	959	44
1	15144	W	0.132551	0.125244	0.0285356	688	15
1	15144	C	0.2241	0.0964325	0.0217541	998	21
1	15267	W	0.120784	0.14178	0.0360506	296	13
1	15267	C	0.169791	0.127789	0.0307377	921	15
1	16491	W	0.0150151	0.0148735	0.00193802	622	156
1	16491	C	0.0202212	0.0181181	0.00266202	787	132
1	16740	W	0.0145807	0.00834435	0.000850983	1000	293
1	16740	C	0.0130997	0.00919653	0.00102678	998	279
1	16787	W	0.0119745	0.00811257	0.000836808	1000	303
1	16787	C	0.00959217	0.0081276	0.000885949	938	323
1	16829	W	0.0637395	0.0515665	0.00934176	904	41
1	16829	C	0.056659	0.0441453	0.00802875	923	50
1	16853	W	0.0598458	0.056515	0.010327	708	37
1	16853	C	0.0540784	0.063066	0.0133762	264	34
1	16874	W	0.0488572	0.0579512	0.0107939	169	36
1	16874	C	0.0669417	0.0742597	0.0162767	357	28
1	19825	W	0.097452	0.097467	0.0207345	584	20
1	19825	C	0.0892372	0.109385	0.0248307	200	18

Table A.6: Results for gene classes taken from level 2 of the Gene Ontology heirarchy on chromosome 1 using TIGR data with the tandem duplications removed.

Chr	Class	Std	Orig. Grnwd	MC Grnwd	SD MC grnwd	Ranking	Examples
2	00166	W	0.0120941	0.0138655	0.00194106	137	174
2	00166	C	0.0145925	0.0138498	0.00194909	719	173
2	03676	W	0.00666209	0.00746938	0.000772488	107	337
2	03676	C	0.00708366	0.00780332	0.000922006	200	326
2	03700	W	0.0394401	0.0181264	0.00274863	999	130
2	03700	C	0.0561825	0.0199386	0.0031745	1000	118
2	03735	W	0.0510885	0.0535256	0.0113382	498	41
2	03735	C	0.0748701	0.0731439	0.0160671	651	29
2	03754	W	0.21383	0.117079	0.0285792	991	17
2	03754	C	0.136785	0.131422	0.0321473	653	15
2	03793	W	0.440455	0.345091	0.0962091	852	4
2	03793	C	0.311615	0.33592	0.0848267	449	4
2	04386	W	0.108975	0.139285	0.0344028	168	14
2	04386	C	0.274744	0.167198	0.0403137	977	11
2	04857	W	0.24701	0.196816	0.0498654	862	9
2	04857	C	0.179946	0.146809	0.0376253	843	13
2	04872	W	0.128379	0.106413	0.0245292	833	19
2	04872	C	0.142208	0.0931953	0.0227478	959	22
2	05386	W	0.0498991	0.0444904	0.00895837	814	50
2	05386	C	0.0595119	0.0399326	0.00800476	975	56
2	05489	W	0.0627677	0.0681163	0.0151927	421	32
2	05489	C	0.0825069	0.0798037	0.0183442	645	27
2	05515	W	0.068128	0.0338404	0.00592966	1000	67
2	05515	C	0.0725999	0.0347042	0.00671814	999	66
2	08135	W	0.155136	0.165098	0.0416062	484	11
2	08135	C	0.138527	0.166002	0.0417856	268	11
2	08289	W	0.192819	0.198526	0.0498995	542	9
2	08289	C	0.311287	0.297822	0.0791541	653	5
2	08565	W	0.242899	0.181196	0.042781	914	10
2	08565	C	0.164953	0.137176	0.0329595	825	14
2	15034	W	0.211733	0.136878	0.0315354	974	14
2	15034	C	0.155726	0.123769	0.0296685	866	16
2	15075	W	0.0647434	0.0740739	0.0162051	320	29
2	15075	C	0.124802	0.0875377	0.0202259	940	24
2	15144	W	0.248186	0.267142	0.0687964	473	6
2	15144	C	0.0932863	0.145517	0.0365054	11	13
2	15267	W	0.121633	0.131592	0.0304631	439	15
2	15267	C	0.273804	0.197277	0.0504965	930	9
2	16491	W	0.0429582	0.0347538	0.00686123	903	66
2	16491	C	0.0394562	0.0321105	0.0059126	902	71
2	16740	W	0.0140019	0.0125648	0.00167913	847	193
2	16740	C	0.0100147	0.0117611	0.00146531	72	207
2	16787	W	0.0109255	0.0126858	0.00169714	88	191
2	16787	C	0.0158142	0.0112878	0.00146346	987	214
2	16829	W	0.0929941	0.0806471	0.018556	801	26
2	16829	C	0.142105	0.0678035	0.014861	997	32
2	16853	W	0.138935	0.139434	0.0340804	597	14
2	16853	C	0.14666	0.101124	0.0243464	940	20
2	16874	W	0.126924	0.0974982	0.0229079	881	21
2	16874	C	0.0955882	0.106161	0.0247493	397	19
2	19825	W	0.211733	0.13867	0.0341782	970	14
2	19825	C	0.133188	0.122226	0.0288378	715	16

Table A.7: Results for gene classes taken from level 2 of the Gene Ontology heirarchy on chromosome 2 using TIGR data with the tandem duplications removed.

Chr	Class	Std	Orig. Grnwd	MC Grnwd	SD MC grnwd	Ranking	Examples
3	00166	W	0.0150251	0.0128967	0.00184559	892	223
3	00166	C	0.0169452	0.0125978	0.00212673	968	215
3	03676	W	0.0193296	0.00996196	0.00130052	1000	307
3	03676	C	0.0170507	0.00925041	0.00134523	1000	337
3	03700	W	0.0492682	0.0179707	0.00294462	1000	151
3	03700	C	0.0544014	0.0206394	0.00416924	1000	135
3	03735	W	0.0821085	0.0498291	0.0108361	983	47
3	03735	C	0.0614972	0.0492735	0.0109175	876	50
3	03754	W	0.128367	0.0848729	0.0216931	956	26
3	03754	C	0.174207	0.100247	0.0249277	983	22
3	03793	W	0.692063	0.698654	0.1467	520	1
3	03793	C	0.208939	0.255625	0.0750361	320	7
3	04386	W	0.242068	0.120054	0.0304329	995	17
3	04386	C	0.10462	0.125315	0.0337524	299	17
3	04857	W	0.198285	0.162657	0.0453965	829	12
3	04857	C	0.240999	0.154312	0.0420038	956	13
3	04872	W	0.110363	0.0793271	0.0188017	929	28
3	04872	C	0.11981	0.100426	0.0261102	833	22
3	05386	W	0.0666984	0.0535273	0.0124187	880	44
3	05386	C	0.0513102	0.0356247	0.00771152	952	72
3	05489	W	0.0580403	0.0546816	0.0133442	717	43
3	05489	C	0.0924093	0.054163	0.0121947	990	44
3	05515	W	0.0343056	0.0272103	0.00492473	914	93
3	05515	C	0.0306511	0.0271684	0.00599417	795	99
3	08135	W	0.281498	0.222765	0.0624056	855	8
3	08135	C	0.198122	0.12951	0.0329998	963	16
3	08289	W	0.49036	0.362963	0.102025	889	4
3	08289	C	0.165608	0.145065	0.0405642	775	14
3	08565	W	0.193263	0.161309	0.040286	814	12
3	08565	C	0.127085	0.112153	0.0289605	783	19
3	15034	W	0.178703	0.150349	0.0423305	817	13
3	15034	C	0.116053	0.146729	0.0411603	223	14
3	15075	W	0.0983476	0.0977602	0.0252346	604	22
3	15075	C	0.0785358	0.0527191	0.0122656	958	46
3	15144	W	0.27218	0.220952	0.0626298	840	8
3	15144	C	0.23447	0.145341	0.0399757	967	14
3	15267	W	0.113042	0.150383	0.0405381	130	13
3	15267	C	0.248785	0.22389	0.0625603	733	8
3	16491	W	0.0357018	0.0264013	0.00489884	953	96
3	16491	C	0.0299024	0.0299587	0.00636557	597	86
3	16740	W	0.0251587	0.0137798	0.00199968	996	201
3	16740	C	0.0192089	0.0147304	0.00255908	934	195
3	16787	W	0.0136099	0.0118464	0.00167866	878	245
3	16787	C	0.0130666	0.0106046	0.00140157	943	261
3	16829	W	0.105143	0.0592663	0.0139266	990	39
3	16829	C	0.0936301	0.0590276	0.0139352	974	41
3	16853	W	0.153735	0.0841485	0.0200694	992	26
3	16853	C	0.0974768	0.0767759	0.0188336	876	30
3	16874	W	0.159735	0.0964863	0.0236703	979	22
3	16874	C	0.0838723	0.0733686	0.0186595	786	32
3	19825	W	0.113484	0.140931	0.0376668	226	14
3	19825	C	0.120244	0.15608	0.0451583	187	13

Table A.8: Results for gene classes taken from level 2 of the Gene Ontology heirarchy on chromosome 3 using TIGR data with the tandem duplications removed.

Chr	Class	Std	Orig. Grnwd	MC Grnwd	SD MC grnwd	Ranking	Examples
4	00166	W	0.0162266	0.0141836	0.00153174	891	161
4	00166	C	0.0133603	0.015613	0.00236995	114	156
4	03676	W	0.0100103	0.00865673	0.000769678	946	271
4	03676	C	0.0101189	0.00982037	0.00128182	710	258
4	03700	W	0.0231701	0.0185406	0.00226592	959	120
4	03700	C	0.0312043	0.0216811	0.00358336	974	109
4	03735	W	0.0967996	0.0827177	0.0187241	841	24
4	03735	C	0.114467	0.0891891	0.0198727	899	23
4	03754	W	0.132303	0.113325	0.0270486	816	17
4	03754	C	0.0964538	0.137028	0.0317815	42	14
4	03793	W	0.176233	0.150833	0.0362954	809	12
4	03793	C	0.142791	0.136258	0.0319889	663	14
4	04386	W	0.128754	0.163232	0.0412364	167	11
4	04386	C	0.287412	0.263527	0.0708856	725	6
4	04857	W	0.254984	0.256732	0.0668623	588	6
4	04857	C	0.381647	0.403214	0.107171	506	3
4	04872	W	0.054522	0.0650309	0.0125094	188	31
4	04872	C	0.0553316	0.0608396	0.0122822	385	35
4	05386	W	0.0455547	0.0477932	0.00846805	455	44
4	05386	C	0.0664276	0.0480133	0.00972745	946	46
4	05489	W	0.0509371	0.0535	0.0101575	463	39
4	05489	C	0.0595786	0.0559469	0.0112029	716	39
4	05515	W	0.0498539	0.0395684	0.00643584	937	53
4	05515	C	0.0390918	0.0334943	0.00603047	860	67
4	08135	W	0.149115	0.161563	0.0379221	433	11
4	08135	C	0.148636	0.154628	0.0385351	522	12
4	08289	W	0.152105	0.133703	0.0302301	778	14
4	08289	C	0.227056	0.151784	0.0374692	960	12
4	08565	W	0.16519	0.210706	0.0547764	186	8
4	08565	C	0.14776	0.191958	0.0472044	133	9
4	15034	W	0.199733	0.125047	0.0309225	972	15
4	15034	C	0.226662	0.176564	0.0414653	895	10
4	15075	W	0.0698339	0.0890899	0.0184024	109	22
4	15075	C	0.115726	0.0849851	0.0178984	939	24
4	15144	W	0.183592	0.228376	0.05971	215	7
4	15144	C	0.16285	0.210408	0.0542087	143	8
4	15267	W	0.224898	0.150683	0.0367392	965	12
4	15267	C	0.163498	0.144828	0.0341622	752	13
4	16491	W	0.0225098	0.0257414	0.00372175	156	85
4	16491	C	0.0308966	0.0272853	0.00454996	814	85
4	16740	W	0.0149445	0.0130421	0.00128686	932	173
4	16740	C	0.0166462	0.0157739	0.00238388	730	154
4	16787	W	0.0127713	0.0128543	0.00141491	540	180
4	16787	C	0.0194318	0.0122846	0.00175734	990	202
4	16829	W	0.0712519	0.0895783	0.0188815	135	22
4	16829	C	0.149617	0.0684695	0.0145365	1000	31
4	16853	W	0.173571	0.141627	0.0364694	854	13
4	16853	C	0.0992826	0.0964113	0.0211939	646	21
4	16874	W	0.0814678	0.0932682	0.0201482	294	21
4	16874	C	0.0740432	0.0891636	0.0208811	230	23
4	19825	W	0.200097	0.149703	0.0345069	921	12
4	19825	C	0.227205	0.194447	0.0479438	805	9

Table A.9: Results for gene classes taken from level 2 of the Gene Ontology heirarchy on chromosome 4 using TIGR data with the tandem duplications removed.

Chr	Class	Std	Orig. Grnwd	MC Grnwd	SD MC grnwd	Ranking	Examples
5	00166	W	0.0171599	0.00963764	0.00178565	987	267
5	00166	C	0.0159983	0.00972401	0.00165804	992	277
5	03676	W	0.00784552	0.006113	0.000996244	933	439
5	03676	C	0.00583458	0.00641422	0.000855619	295	426
5	03700	W	0.0283576	0.0122214	0.00215929	1000	205
5	03700	C	0.0188692	0.0138402	0.00253511	948	191
5	03735	W	0.0748125	0.0550829	0.0118006	926	41
5	03735	C	0.0451762	0.0486394	0.0102796	445	48
5	03754	W	0.076464	0.0608605	0.0129899	883	36
5	03754	C	0.0630745	0.0653014	0.0139664	521	35
5	03793	W	0.104764	0.0720973	0.0164544	958	30
5	03793	C	0.112627	0.0931866	0.0215842	839	23
5	04386	W	0.137231	0.101844	0.0233527	918	20
5	04386	C	0.0921482	0.0935801	0.0217333	538	23
5	04857	W	0.190225	0.19936	0.0555256	540	9
5	04857	C	0.196175	0.200091	0.0538125	553	9
5	04872	W	0.043182	0.041523	0.00849614	661	55
5	04872	C	0.102719	0.0648648	0.0140313	986	35
5	05386	W	0.034037	0.0328198	0.00646322	670	71
5	05386	C	0.0446296	0.0312793	0.00653988	953	79
5	05489	W	0.0853934	0.0452741	0.00922753	998	50
5	05489	C	0.0719073	0.0549533	0.0124121	903	42
5	05515	W	0.0354366	0.019366	0.00366596	996	126
5	05515	C	0.0421771	0.0230558	0.00487118	996	111
5	08135	W	0.111221	0.103082	0.0258864	722	20
5	08135	C	0.0835833	0.102263	0.0263908	246	21
5	08289	W	0.119533	0.102295	0.024839	815	20
5	08289	C	0.314204	0.134596	0.0335815	1000	15
5	08565	W	0.130795	0.117804	0.0282588	751	17
5	08565	C	0.153383	0.160563	0.043315	534	12
5	15034	W	0.13216	0.147834	0.0400977	409	13
5	15034	C	0.113269	0.0893003	0.0210844	883	24
5	15075	W	0.0638186	0.0786285	0.016819	170	27
5	15075	C	0.0835628	0.0510038	0.0114056	979	46
5	15144	W	0.124376	0.120329	0.0310194	660	17
5	15144	C	0.166513	0.186392	0.0522273	426	10
5	15267	W	0.108888	0.148349	0.038589	79	13
5	15267	C	0.0868651	0.0858772	0.0192674	617	25
5	16491	W	0.0260669	0.0209664	0.004027	899	116
5	16491	C	0.0233869	0.020247	0.00396904	830	127
5	16740	W	0.0121352	0.0104459	0.00191743	868	244
5	16740	C	0.0184173	0.0111379	0.00237493	982	241
5	16787	W	0.00999564	0.00903277	0.00114666	815	281
5	16787	C	0.0130848	0.0097178	0.00207321	922	280
5	16829	W	0.0711863	0.0622255	0.0126386	802	35
5	16829	C	0.0479351	0.0577829	0.0121708	211	40
5	16853	W	0.122868	0.0832939	0.0191864	965	25
5	16853	C	0.0695057	0.0861159	0.0201013	187	25
5	16874	W	0.070515	0.0739976	0.0161408	487	29
5	16874	C	0.0662705	0.0739378	0.0158829	371	30
5	19825	W	0.13216	0.146233	0.0362879	407	13
5	19825	C	0.105358	0.0865355	0.0193606	869	25

Table A.10: Results for gene classes taken from level 2 of the Gene Ontology heirarchy on chromosome 5 using TIGR data with the tandem duplications removed.

Chr	Class	Std	Orig. Grnwd	MC Grnwd	SD MC grnwd	Ranking	Examples
1	03677	W	0.0112648	0.00743769	0.000775259	999	334
1	03677	C	0.00877793	0.00811889	0.00123578	809	335
1	03723	W	0.0432581	0.0414197	0.00742404	691	52
1	03723	C	0.0344318	0.0384559	0.00749782	337	59
1	04888	W	0.061648	0.0592349	0.010758	674	35
1	04888	C	0.0671268	0.0771114	0.0159577	286	27
1	05516	W	0.118875	0.142978	0.0359628	245	13
1	05516	C	0.091567	0.0935347	0.0212433	546	22
1	08026	W	0.191664	0.132707	0.0312037	937	14
1	08026	C	0.0921932	0.100466	0.0222138	409	20
1	08135	W	0.0651981	0.0733003	0.0151423	320	28
1	08135	C	0.0775567	0.10067	0.0224354	92	20
1	08233	W	0.0233759	0.0267963	0.00391656	178	83
1	08233	C	0.0206304	0.0253335	0.00434446	70	93
1	08324	W	0.0892873	0.0643345	0.0122961	957	32
1	08324	C	0.0985978	0.0693797	0.0156099	955	31
1	08509	W	0.122651	0.118825	0.0287947	649	16
1	08509	C	0.251104	0.155078	0.0392027	974	12
1	15036	W	0.0889079	0.0901678	0.0196119	564	22
1	15036	C	0.109883	0.110495	0.0246957	575	18
1	15268	W	0.160893	0.165289	0.0441906	555	11
1	15268	C	0.169791	0.13006	0.0337903	888	15
1	15290	W	0.0543088	0.0471399	0.00858893	863	45
1	15290	C	0.0558487	0.0449885	0.00837621	913	49
1	15399	W	0.071959	0.0500326	0.00921786	975	42
1	15399	C	0.0715982	0.0730087	0.0156018	552	29
1	16614	W	0.109443	0.0727909	0.0138139	981	28
1	16614	C	0.0704166	0.0725814	0.0154577	525	29
1	16684	W	0.101995	0.1429	0.0357926	46	13
1	16684	C	0.239059	0.24122	0.0747899	618	7
1	16705	W	0.084949	0.0972554	0.0203405	299	20
1	16705	C	0.190129	0.145093	0.0353034	905	13
1	16741	W	0.0598148	0.0900472	0.0191944	3	22
1	16741	C	0.0751853	0.0883928	0.0193237	240	23
1	16746	W	0.136794	0.107688	0.0236941	892	18
1	16746	C	0.091873	0.0729661	0.0151366	901	29
1	16757	W	0.0667384	0.0439505	0.00771198	981	49
1	16757	C	0.0464788	0.0409976	0.00761115	811	54
1	16765	W	0.107635	0.0975849	0.0212669	767	20
1	16765	C	0.109202	0.129399	0.0330854	284	15
1	16772	W	0.0199833	0.0143497	0.0020718	965	164
1	16772	C	0.0222955	0.0169544	0.0026394	960	143
1	16788	W	0.0373592	0.0252128	0.00383814	989	89
1	16788	C	0.0220473	0.024448	0.00390301	272	96
1	16798	W	0.0545562	0.051458	0.00898214	705	41
1	16798	C	0.0345892	0.0385588	0.00729644	332	58
1	16817	W	0.0340734	0.0316031	0.00527004	730	70
1	16817	C	0.0434497	0.0301946	0.00521785	976	76
1	16830	W	0.288981	0.2113	0.0572581	907	8
1	16830	C	0.114107	0.121327	0.0291212	481	16
1	16835	W	0.114373	0.119055	0.0280642	530	16
1	16835	C	0.11854	0.106071	0.0268915	780	19
1	16879	W	0.0965149	0.0946107	0.0201971	620	21
1	16879	C	0.0932646	0.145352	0.037394	6	13
1	17076	W	0.0114333	0.0102652	0.00112816	868	234
1	17076	C	0.0131537	0.00902253	0.00102379	996	285
1	46872	W	0.024028	0.0180166	0.00237599	978	127
1	46872	C	0.0230331	0.0180724	0.00275136	952	133
1	46873	W	0.182189	0.153105	0.0401475	829	12
1	46873	C	0.113717	0.110653	0.0287141	638	18

Table A.11: Results for gene classes taken from level 3 of the Gene Ontology heirarchy on chromosome 1 using TIGR data with the tandem duplications removed.

Chr	Class	Std	Orig. Grnwd	MC Grnwd	SD MC grnwd	Ranking	Examples
2	03677	W	0.0104227	0.0103567	0.00121916	617	236
2	03677	C	0.0121405	0.0111161	0.0014223	815	218
2	03723	W	0.0346826	0.0377351	0.00711278	388	60
2	03723	C	0.0314191	0.0356429	0.00747756	298	64
2	04888	W	0.186446	0.136962	0.0323555	923	14
2	04888	C	0.189256	0.129371	0.0332084	949	15
2	05516	W	0.174792	0.156617	0.0385985	738	12
2	05516	C	0.173693	0.146776	0.0366692	800	13
2	08026	W	0.13553	0.157857	0.0403405	327	12
2	08026	C	0.313679	0.233308	0.0580832	907	7
2	08135	W	0.155136	0.167164	0.0429292	452	11
2	08135	C	0.138527	0.165157	0.0399058	286	11
2	08233	W	0.0456441	0.0501534	0.0101662	387	44
2	08233	C	0.0329689	0.0303633	0.00575822	761	77
2	08324	W	0.0783946	0.0940081	0.0219553	240	22
2	08324	C	0.142701	0.13091	0.0331694	728	15
2	08509	W	0.222295	0.215237	0.0573555	652	8
2	08509	C	0.258681	0.211388	0.0567755	842	8
2	15036	W	0.538734	0.265629	0.0691196	995	6
2	15036	C	0.182638	0.211328	0.0530256	336	8
2	15268	W	0.121633	0.12867	0.0291115	478	15
2	15268	C	0.274044	0.211519	0.054928	890	8
2	15290	W	0.105731	0.0943084	0.0227199	758	22
2	15290	C	0.0749605	0.0676848	0.0152143	752	32
2	15399	W	0.110104	0.0870072	0.0192825	894	24
2	15399	C	0.10166	0.0968783	0.0227527	672	21
2	16614	W	0.145478	0.135955	0.031394	690	14
2	16614	C	0.167532	0.145675	0.0346953	777	13
2	16684	W	0.232707	0.21529	0.0546774	708	8
2	16684	C	0.223316	0.196067	0.0473663	780	9
2	16705	W	0.240199	0.237971	0.0627918	608	7
2	16705	C	0.195492	0.166382	0.0439648	806	11
2	16741	W	0.20979	0.196969	0.049474	685	9
2	16741	C	0.160775	0.168885	0.043349	494	11
2	16746	W	0.104523	0.130026	0.0325353	196	15
2	16746	C	0.0890638	0.101596	0.0247075	363	20
2	16757	W	0.0688639	0.0692211	0.0153597	598	31
2	16757	C	0.114446	0.0736175	0.0176818	973	29
2	16765	W	0.194296	0.156408	0.0404844	856	12
2	16765	C	0.287522	0.233806	0.0585521	856	7
2	16772	W	0.0176957	0.0202128	0.0032316	196	117
2	16772	C	0.0166322	0.0190175	0.00289368	190	124
2	16788	W	0.0323267	0.0327083	0.00606734	576	70
2	16788	C	0.0282482	0.0312208	0.00577268	347	75
2	16798	W	0.0909692	0.089414	0.0202567	627	23
2	16798	C	0.136232	0.0777297	0.0177416	986	27
2	16817	W	0.0403907	0.0417495	0.00835943	534	54
2	16817	C	0.0561158	0.0573333	0.0135947	570	39
2	16830	W	0.142254	0.195474	0.0489293	102	9
2	16830	C	0.193914	0.180522	0.0444134	706	10
2	16835	W	0.198481	0.145977	0.0347089	927	13
2	16835	C	0.204152	0.147749	0.038601	923	13
2	16879	W	0.203086	0.138168	0.0323754	958	14
2	16879	C	0.104247	0.116908	0.0281578	390	17
2	17076	W	0.0124449	0.0150802	0.00213807	42	159
2	17076	C	0.0149238	0.0148545	0.00211246	597	160
2	46872	W	0.0562068	0.0275346	0.00497476	999	84
2	46872	C	0.0503996	0.031712	0.006277	982	73
2	46873	W	0.20553	0.180018	0.0451437	768	10
2	46873	C	0.239438	0.212352	0.0505511	749	8

Table A.12: Results for gene classes taken from level 3 of the Gene Ontology heirarchy on chromosome 2 using TIGR data with the tandem duplications removed.

Chr	Class	Std	Orig. Grnwd	MC Grnwd	SD MC grnwd	Ranking	Examples
3	03677	W	0.0263911	0.012224	0.00165541	1000	237
3	03677	C	0.0196053	0.0123025	0.00198502	996	241
3	03723	W	0.0741158	0.0598415	0.0138504	862	38
3	03723	C	0.101448	0.049219	0.011411	996	49
3	04888	W	0.184884	0.142813	0.038696	874	14
3	04888	C	0.1849	0.137107	0.0369359	905	15
3	05516	W	0.100107	0.128727	0.0344158	164	16
3	05516	C	0.210285	0.13761	0.0367163	955	15
3	08026	W	0.293636	0.173373	0.0468613	973	11
3	08026	C	0.187574	0.163889	0.0433285	779	12
3	08135	W	0.281498	0.223907	0.0629609	842	8
3	08135	C	0.198122	0.128622	0.0333265	952	16
3	08233	W	0.0306077	0.0401489	0.00879667	65	60
3	08233	C	0.0312083	0.0385205	0.00879191	173	67
3	08324	W	0.130001	0.151095	0.0409121	351	13
3	08324	C	0.0967897	0.0816011	0.0199545	833	28
3	08509	W	0.189316	0.225047	0.0648768	335	8
3	08509	C	0.129469	0.125406	0.035697	650	17
3	15036	W	0.223256	0.203921	0.0590748	733	9
3	15036	C	0.163199	0.152665	0.0408646	705	13
3	15268	W	0.113042	0.151833	0.0421542	128	13
3	15268	C	0.248785	0.225787	0.0671843	722	8
3	15290	W	0.115692	0.102749	0.0275028	777	21
3	15290	C	0.0955248	0.0692983	0.015565	930	34
3	15399	W	0.134115	0.12135	0.0314217	736	17
3	15399	C	0.109014	0.0662885	0.0157374	979	36
3	16614	W	0.15114	0.0876905	0.0207863	984	25
3	16614	C	0.102012	0.131898	0.0363948	200	16
3	16684	W	0.135366	0.203612	0.0569976	43	9
3	16684	C	0.290516	0.317298	0.0961902	499	5
3	16705	W	0.29337	0.162105	0.0437922	981	12
3	16705	C	0.27373	0.377635	0.113235	178	4
3	16741	W	0.137989	0.144688	0.0398418	530	14
3	16741	C	0.167095	0.136838	0.0371045	841	15
3	16746	W	0.158679	0.121109	0.0312022	885	17
3	16746	C	0.272176	0.176903	0.0477746	952	11
3	16757	W	0.100481	0.0767234	0.0181756	904	29
3	16757	C	0.0519794	0.0665585	0.0168971	137	35
3	16765	W	0.201503	0.245143	0.0698818	302	7
3	16765	C	0.164637	0.144875	0.0388722	761	14
3	16772	W	0.0314673	0.0214768	0.00359096	984	121
3	16772	C	0.0463411	0.0251376	0.00507941	993	106
3	16788	W	0.0556525	0.0277779	0.00528946	998	89
3	16788	C	0.0717801	0.0296899	0.0062334	1000	87
3	16798	W	0.0887334	0.0665127	0.0155181	916	34
3	16798	C	0.0746403	0.0575421	0.0128867	897	41
3	16817	W	0.0523847	0.0458081	0.00969333	814	52
3	16817	C	0.0585609	0.0418498	0.00961855	936	59
3	16830	W	0.247403	0.162812	0.0459642	943	12
3	16830	C	0.139028	0.148827	0.0423823	506	14
3	16835	W	0.130277	0.106633	0.0282779	842	20
3	16835	C	0.263629	0.138116	0.0370122	992	15
3	16879	W	0.20492	0.143483	0.039062	929	14
3	16879	C	0.107115	0.117466	0.030972	449	18
3	17076	W	0.0160823	0.0133698	0.00166913	958	204
3	17076	C	0.0184426	0.0145995	0.00258743	920	199
3	46872	W	0.0282059	0.0264356	0.00475089	742	96
3	46872	C	0.0331502	0.028703	0.00600743	816	92
3	46873	W	0.280883	0.279438	0.0824004	613	6
3	46873	C	0.164473	0.131526	0.0365733	855	16

Table A.13: Results for gene classes taken from level 3 of the Gene Ontology heirarchy on chromosome 3 using TIGR data with the tandem duplications removed.

Chr	Class	Std	Orig. Grnwd	MC Grnwd	SD MC grnwd	Ranking	Examples
4	03677	W	0.0129923	0.0121034	0.00130761	815	192
4	03677	C	0.0169025	0.0130652	0.00182494	960	188
4	03723	W	0.0417018	0.047902	0.00884043	254	44
4	03723	C	0.0540145	0.0599444	0.0128247	361	36
4	04888	W	0.0644233	0.0888302	0.0184649	21	22
4	04888	C	0.0836343	0.0965266	0.0221236	300	21
4	05516	W	0.118923	0.149038	0.0320399	154	12
4	05516	C	0.235037	0.178224	0.0437043	924	10
4	08026	W	0.167338	0.191652	0.0510272	354	9
4	08026	C	0.287412	0.255354	0.0655557	761	6
4	08135	W	0.149115	0.161074	0.0389979	433	11
4	08135	C	0.148636	0.157594	0.0372392	465	12
4	08233	W	0.0285846	0.0324468	0.00490406	200	66
4	08233	C	0.0462363	0.0425322	0.00769399	741	52
4	08324	W	0.0989273	0.141374	0.0336064	25	13
4	08324	C	0.150823	0.105121	0.0233651	948	19
4	08509	W	0.179975	0.209712	0.0523687	328	8
4	08509	C	0.324323	0.264871	0.0720354	837	6
4	15036	W	0.181255	0.193223	0.052831	483	9
4	15036	C	0.262949	0.212784	0.0547736	851	8
4	15268	W	0.224898	0.150251	0.0353664	953	12
4	15268	C	0.163498	0.142327	0.0325918	787	13
4	15290	W	0.0952999	0.113607	0.0277873	263	17
4	15290	C	0.11193	0.1119	0.0265576	586	18
4	15399	W	0.121387	0.118155	0.0281176	647	16
4	15399	C	0.100686	0.082736	0.0189495	863	25
4	16614	W	0.0975486	0.132796	0.0320763	75	14
4	16614	C	0.116744	0.096415	0.0206631	853	21
4	16684	W	0.110852	0.139844	0.0329314	156	13
4	16684	C	0.197149	0.177635	0.0449469	752	10
4	16705	W	0.178822	0.205722	0.0504966	341	8
4	16705	C	0.303436	0.293133	0.0760366	654	5
4	16741	W	0.15553	0.140792	0.0347722	749	13
4	16741	C	0.146812	0.135301	0.0315504	699	14
4	16746	W	0.600006	0.175414	0.0410942	1000	10
4	16746	C	0.185573	0.209597	0.0520227	370	8
4	16757	W	0.0659497	0.0615748	0.0117633	736	33
4	16757	C	0.0815639	0.0685936	0.0149952	841	31
4	16765	W	0.225286	0.207875	0.0504392	725	8
4	16765	C	0.248316	0.208959	0.0514157	833	8
4	16772	W	0.0270851	0.0234177	0.00327129	892	94
4	16772	C	0.0235973	0.0271343	0.0046604	205	85
4	16788	W	0.0459036	0.0499351	0.00902391	378	42
4	16788	C	0.0611591	0.0349187	0.00629433	996	66
4	16798	W	0.0842368	0.0932126	0.0213874	400	21
4	16798	C	0.0635588	0.0585905	0.0117683	752	37
4	16817	W	0.0564313	0.045103	0.00773016	925	46
4	16817	C	0.0596262	0.0560577	0.011947	700	39
4	16830	W	0.289936	0.290194	0.0766889	583	5
4	16830	C	0.309101	0.23276	0.0586496	906	7
4	16835	W	0.168894	0.209922	0.0530881	220	8
4	16835	C	0.160676	0.109992	0.0250202	957	18
4	16879	W	0.158204	0.14082	0.0336505	761	13
4	16879	C	0.213312	0.128431	0.0304347	982	15
4	17076	W	0.0171887	0.0155604	0.00191376	843	146
4	17076	C	0.0156878	0.0171109	0.0026329	312	140
4	46872	W	0.0291204	0.0293612	0.00445817	573	73
4	46872	C	0.0268637	0.0300383	0.00534075	307	76
4	46873	W	0.159174	0.208489	0.0552013	136	8
4	46873	C	0.138872	0.156184	0.0391834	387	12

Table A.14: Results for gene classes taken from level 3 of the Gene Ontology heirarchy on chromosome 4 using TIGR data with the tandem duplications removed.

Chr	Class	Std	Orig. Grnwd	MC Grnwd	SD MC grnwd	Ranking	Examples
5	03677	W	0.01122	0.00813623	0.00155813	959	322
5	03677	C	0.00895194	0.0086028	0.00193718	762	321
5	03723	W	0.0517797	0.0422362	0.00833106	873	54
5	03723	C	0.0431566	0.0465661	0.00977175	447	51
5	04888	W	0.0603689	0.0591141	0.0122314	617	37
5	04888	C	0.139476	0.0932902	0.0225263	965	23
5	05516	W	0.122023	0.122446	0.0302002	586	16
5	05516	C	0.208923	0.103458	0.022912	999	20
5	08026	W	0.169112	0.131586	0.0336958	880	15
5	08026	C	0.115781	0.14224	0.0376127	238	14
5	08135	W	0.111221	0.100929	0.023443	743	20
5	08135	C	0.0835833	0.100647	0.0245403	242	21
5	08233	W	0.0240237	0.0278832	0.0053231	231	85
5	08233	C	0.0345186	0.0346533	0.00784115	613	71
5	08324	W	0.112969	0.118477	0.0314447	523	17
5	08324	C	0.106023	0.069281	0.0155078	965	32
5	08509	W	0.140147	0.170098	0.0456754	255	11
5	08509	C	0.218768	0.151233	0.0424222	922	13
5	15036	W	0.197224	0.199226	0.0533452	578	9
5	15036	C	0.137191	0.115361	0.0277765	817	18
5	15268	W	0.109074	0.15834	0.0413776	41	12
5	15268	C	0.0868651	0.08578	0.0189725	595	25
5	15290	W	0.0407779	0.0485806	0.00972985	198	46
5	15290	C	0.0580766	0.0588008	0.0131948	575	39
5	15399	W	0.0886392	0.0846687	0.0196059	673	25
5	15399	C	0.102243	0.0626415	0.0138766	981	36
5	16614	W	0.0600013	0.0739093	0.0166394	167	29
5	16614	C	0.0748891	0.0887599	0.0202041	256	24
5	16684	W	0.263056	0.307539	0.0866888	364	5
5	16684	C	0.131215	0.14284	0.0380675	459	14
5	16705	W	0.260239	0.156134	0.0396956	973	12
5	16705	C	0.167615	0.149306	0.0365214	745	13
5	16741	W	0.0936445	0.0818252	0.0199275	817	26
5	16741	C	0.11064	0.120955	0.0300102	447	17
5	16746	W	0.171211	0.12504	0.0331477	908	16
5	16746	C	0.206059	0.121051	0.0320166	979	17
5	16757	W	0.0543074	0.0594586	0.012272	382	37
5	16757	C	0.100092	0.0616616	0.0134518	988	37
5	16765	W	0.152023	0.181115	0.0522266	327	10
5	16765	C	0.161363	0.140091	0.0361292	795	14
5	16772	W	0.0214387	0.0182991	0.00314096	864	133
5	16772	C	0.0235594	0.0195708	0.00407872	880	132
5	16788	W	0.0277638	0.0287146	0.00545909	525	82
5	16788	C	0.0290335	0.0287027	0.00615458	633	87
5	16798	W	0.0906331	0.0611611	0.0133093	961	36
5	16798	C	0.0457761	0.0539012	0.0126032	276	43
5	16817	W	0.0605538	0.0351744	0.00695328	994	66
5	16817	C	0.041682	0.0314016	0.00678317	909	79
5	16830	W	0.170408	0.155443	0.0402545	746	12
5	16830	C	0.156601	0.119835	0.0285022	901	17
5	16835	W	0.145995	0.138652	0.0374562	687	14
5	16835	C	0.14122	0.161296	0.0455584	389	12
5	16879	W	0.245606	0.171062	0.0489063	920	11
5	16879	C	0.0937029	0.113763	0.0281602	237	18
5	17076	W	0.0174472	0.0104488	0.00191839	987	244
5	17076	C	0.0161141	0.0111795	0.0015269	1000	257
5	46872	W	0.0211967	0.0168163	0.00295392	927	146
5	46872	C	0.0202159	0.0216254	0.00435813	432	118
5	46873	W	0.140386	0.170474	0.0453658	272	11
5	46873	C	0.16261	0.131026	0.0320155	870	15

Table A.15: Results for gene classes taken from level 3 of the Gene Ontology heirarchy on chromosome 5 using TIGR data with the tandem duplications removed.

Chr	Class	Std	Orig. Grnwd	MC Grnwd	SD MC grnwd	Ranking	Examples
1	04175	W	0.0513475	0.0562593	0.00995227	345	37
1	04175	W	0.0513475	0.0563434	0.0101552	350	37
1	04175	C	0.0449739	0.0478894	0.00929813	444	46
1	04175	W	0.0513475	0.05688	0.0106956	340	37
1	04175	W	0.0513475	0.0557699	0.00967406	361	37
1	04175	C	0.0449739	0.0475833	0.00890772	445	46
1	04518	W	0.109605	0.0745582	0.014036	972	27
1	04518	C	0.0995586	0.0883178	0.0189466	783	23
1	04553	W	0.0557008	0.0538762	0.00932359	648	39
1	04553	C	0.0374352	0.0410777	0.00773694	355	54
1	04930	W	0.108466	0.0990071	0.0226273	747	20
1	04930	C	0.0739106	0.0846932	0.0180843	310	24
1	05216	W	0.304123	0.212347	0.0604464	920	8
1	05216	C	0.19828	0.156097	0.0403301	877	12
1	05351	W	0.15101	0.16419	0.042401	452	11
1	05351	C	0.232256	0.103538	0.0229448	999	19
1	05509	W	0.0679089	0.060738	0.0110737	791	34
1	05509	C	0.125761	0.0705349	0.0142229	994	30
1	08168	W	0.0598148	0.0902619	0.0196243	1	22
1	08168	C	0.0757883	0.0931364	0.020261	166	22
1	08234	W	0.140658	0.118072	0.0264722	838	16
1	08234	C	0.0641682	0.0727113	0.0149595	307	29
1	08236	W	0.141556	0.0963499	0.0194063	971	20
1	08236	C	0.0781281	0.0852379	0.0184905	413	24
1	08237	W	0.0953594	0.120607	0.0299271	151	16
1	08237	C	0.118285	0.0995007	0.0212722	839	20
1	08238	W	0.142008	0.134512	0.0342668	680	14
1	08238	C	0.136302	0.144898	0.0382489	508	13
1	15077	W	0.182286	0.152211	0.0353606	838	12
1	15077	C	0.126123	0.128138	0.0306994	572	15
1	15291	W	0.0543088	0.0471359	0.00802181	844	45
1	15291	C	0.0558487	0.0453071	0.00905796	903	49
1	15405	W	0.09005	0.0586559	0.0111785	983	36
1	15405	C	0.0855052	0.083012	0.0174757	624	25
1	16301	W	0.0241956	0.0144615	0.00181786	999	162
1	16301	C	0.0261603	0.0159654	0.00235532	998	153
1	16616	W	0.109443	0.0724273	0.0135826	985	28
1	16616	C	0.0929066	0.0859	0.0191422	742	24
1	16747	W	0.141099	0.117739	0.027828	846	16
1	16747	C	0.0925472	0.0768066	0.0153089	850	27
1	16758	W	0.105398	0.0868072	0.0190894	867	23
1	16758	C	0.0805193	0.0593405	0.0119754	949	36
1	16773	W	0.0265585	0.0175763	0.00232149	995	130
1	16773	C	0.0307395	0.021999	0.00375021	970	108
1	16779	W	0.101478	0.0779168	0.0152251	923	26
1	16779	C	0.0827133	0.0792753	0.0169494	670	26
1	16789	W	0.116849	0.0808929	0.0164428	961	25
1	16789	C	0.0929107	0.0799874	0.017765	825	26
1	16818	W	0.0420057	0.0443335	0.0074675	439	49
1	16818	C	0.0471878	0.0373437	0.00644888	923	60
1	16820	W	0.1189	0.0783641	0.0149581	982	26
1	16820	C	0.120898	0.110661	0.0271952	732	18
1	16836	W	0.126712	0.153157	0.0407151	258	12
1	16836	C	0.214194	0.155979	0.0428392	910	12
1	16881	W	0.138938	0.143282	0.037581	550	13
1	16881	C	0.283063	0.216109	0.0593664	877	8
1	19001	W	0.0758917	0.0769395	0.0155041	546	26
1	19001	C	0.0801111	0.0646556	0.0134908	882	33
1	30554	W	0.0116206	0.0106348	0.00124603	823	225
1	30554	C	0.0131895	0.00942394	0.00107213	994	275
1	42578	W	0.0626789	0.0600429	0.011554	665	35
1	42578	C	0.0492096	0.0595325	0.0117279	177	36
1	46914	W	0.0335562	0.0258415	0.00373338	962	86
1	46914	C	0.0269117	0.0240134	0.00383419	808	97

Table A.16: Results for gene classes taken from level 4 of the Gene Ontology heirarchy on chromosome 1 using TIGR data with the tandem duplications removed.

Chr	Class	Std	Orig. Grnwd	MC Grnwd	SD MC grnwd	Ranking	Examples
2	04175	W	0.130923	0.158283	0.0391246	266	12
2	04175	C	0.0633672	0.0832933	0.0176976	83	25
2	04518	W	0.0632686	0.067484	0.0147448	462	32
2	04518	C	0.0995639	0.0788869	0.0182116	888	27
2	04553	W	0.10141	0.0959823	0.0237089	696	22
2	04553	C	0.137122	0.0846905	0.0209282	970	25
2	04930	W	0.216652	0.166296	0.0408891	900	11
2	04930	C	0.232543	0.16621	0.0395209	933	11
2	05216	W	0.184078	0.19591	0.0494723	479	9
2	05216	C	0.386627	0.294625	0.0770053	894	5
2	05351	W	0.248186	0.266352	0.0700611	475	6
2	05351	C	0.119458	0.166081	0.0400158	69	11
2	05509	W	0.130467	0.109584	0.0269941	826	19
2	05509	C	0.114574	0.118733	0.0302282	539	17
2	08168	W	0.20979	0.19595	0.0497701	677	9
2	08168	C	0.160775	0.16835	0.0429306	515	11
2	08234	W	0.458523	0.169967	0.0407021	999	11
2	08234	C	0.0655755	0.0729867	0.0155429	378	29
2	08236	W	0.189919	0.193229	0.0483657	564	9
2	08236	C	0.137261	0.147925	0.0382315	479	13
2	08237	W	0.12911	0.123147	0.0295054	664	16
2	08237	C	0.106069	0.129361	0.0302272	232	15
2	08238	W	0.364146	0.235382	0.0625365	961	7
2	08238	C	0.118795	0.154172	0.036845	130	12
2	15077	W	0.135514	0.158841	0.0405254	312	12
2	15077	C	0.181782	0.261148	0.0669092	49	6
2	15291	W	0.105731	0.0940487	0.023445	782	22
2	15291	C	0.0749605	0.0681076	0.015131	741	32
2	15405	W	0.117556	0.0970187	0.0221987	848	21
2	15405	C	0.142166	0.107537	0.0254887	891	19
2	16301	W	0.0291628	0.0267498	0.00448018	788	86
2	16301	C	0.0467824	0.0233949	0.00390525	999	100
2	16616	W	0.201552	0.146344	0.0353168	924	13
2	16616	C	0.172182	0.167402	0.0409736	642	11
2	16747	W	0.109136	0.137495	0.0312844	180	14
2	16747	C	0.0926203	0.106723	0.0251298	331	19
2	16758	W	0.0754195	0.0836011	0.0192207	397	25
2	16758	C	0.179883	0.131982	0.0333749	907	15
2	16773	W	0.0354505	0.03743	0.00784033	504	61
2	16773	C	0.0654902	0.029929	0.00546081	1000	77
2	16779	W	0.0429206	0.052901	0.0114393	154	42
2	16779	C	0.052084	0.0485073	0.00992722	735	46
2	16789	W	0.13646	0.0970424	0.0217254	930	21
2	16789	C	0.188477	0.155326	0.0396634	842	12
2	16818	W	0.0517624	0.0526443	0.0107515	564	42
2	16818	C	0.0827753	0.0713228	0.0161384	821	30
2	16820	W	0.139369	0.125466	0.0288928	735	16
2	16820	C	0.217831	0.167474	0.040854	891	11
2	16836	W	0.206834	0.165491	0.0399289	864	11
2	16836	C	0.20676	0.179301	0.0425763	781	10
2	16881	W	0.264655	0.183735	0.047894	942	10
2	16881	C	0.243628	0.179358	0.0422828	931	10
2	19001	W	0.119487	0.132036	0.0327903	414	15
2	19001	C	0.11199	0.111385	0.0266912	621	18
2	30554	W	0.0125171	0.0153304	0.00214348	46	156
2	30554	C	0.0166748	0.015771	0.00235116	726	151
2	42578	W	0.166253	0.12504	0.0299752	903	16
2	42578	C	0.144693	0.0767527	0.018392	989	28
2	46914	W	0.0598407	0.0366249	0.00757244	982	62
2	46914	C	0.073084	0.0421166	0.00823319	991	53

Table A.17: Results for gene classes taken from level 4 of the Gene Ontology heirarchy on chromosome 2 using TIGR data with the tandem duplications removed.

Chr	Class	Std	Orig. Grnwd	MC Grnwd	SD MC grnwd	Ranking	Examples
3	04175	W	0.0828367	0.0961691	0.0232794	330	22
3	04175	C	0.0582969	0.0765958	0.0172947	92	30
3	04518	W	0.0949133	0.0869089	0.0207863	740	25
3	04518	C	0.181452	0.15358	0.0414696	800	13
3	04553	W	0.0921558	0.0673205	0.0149732	936	33
3	04553	C	0.0837389	0.0658764	0.0161823	872	36
3	04930	W	0.184884	0.141904	0.0363811	886	14
3	04930	C	0.205915	0.155036	0.0419373	894	13
3	05216	W	0.154022	0.205372	0.0581426	158	9
3	05216	C	0.482657	0.322519	0.098263	925	5
3	05351	W	0.278424	0.246291	0.0697688	740	7
3	05351	C	0.255614	0.164705	0.0466587	945	12
3	05509	W	0.0729502	0.0832001	0.0195152	341	26
3	05509	C	0.0812411	0.090279	0.0237967	430	25
3	08168	W	0.137989	0.142825	0.0374344	559	14
3	08168	C	0.167222	0.145615	0.039742	779	14
3	08234	W	0.0722828	0.110967	0.0300086	12	19
3	08234	C	0.108186	0.117789	0.0295892	449	18
3	08236	W	0.147753	0.141818	0.0379138	663	14
3	08236	C	0.195742	0.145048	0.0383188	908	14
3	08237	W	0.179706	0.161171	0.0455517	746	12
3	08237	C	0.228625	0.176214	0.0493472	868	11
3	08238	W	0.177507	0.162559	0.047588	718	12
3	08238	C	0.279828	0.146712	0.0420409	987	14
3	15077	W	0.25912	0.247725	0.0680434	654	7
3	15077	C	0.230877	0.207613	0.0585072	731	9
3	15291	W	0.115692	0.101581	0.0278023	783	21
3	15291	C	0.0955248	0.0690029	0.0164043	925	34
3	15405	W	0.135881	0.135328	0.0379315	624	15
3	15405	C	0.112491	0.0720566	0.0182815	966	33
3	16301	W	0.0434729	0.0212459	0.00384822	999	124
3	16301	C	0.0424516	0.0230334	0.00454562	994	117
3	16616	W	0.155618	0.110908	0.0281267	920	19
3	16616	C	0.107058	0.138254	0.0370709	180	15
3	16747	W	0.162427	0.12691	0.0294896	887	16
3	16747	C	0.290047	0.208166	0.0595817	912	9
3	16758	W	0.157957	0.115376	0.0294689	912	18
3	16758	C	0.0819283	0.0999651	0.0257017	264	22
3	16773	W	0.0449781	0.0257899	0.00488935	997	98
3	16773	C	0.048904	0.0325901	0.00706568	967	80
3	16779	W	0.147724	0.121613	0.0309004	847	17
3	16779	C	0.135605	0.130562	0.0337814	637	16
3	16789	W	0.0941813	0.0721643	0.0170446	914	31
3	16789	C	0.120314	0.0892498	0.0219505	911	25
3	16818	W	0.0641669	0.0622456	0.0145049	658	36
3	16818	C	0.0628488	0.0493296	0.0112963	884	50
3	16820	W	0.226331	0.204136	0.0557434	735	9
3	16820	C	0.133456	0.0896865	0.0219195	964	25
3	16836	W	0.194999	0.150641	0.0417673	864	13
3	16836	C	0.484423	0.229455	0.0672273	990	8
3	16881	W	0.207774	0.186377	0.0529435	730	10
3	16881	C	0.166702	0.165997	0.0460357	610	12
3	19001	W	0.197903	0.131932	0.0325161	954	15
3	19001	C	0.192229	0.110521	0.0294331	981	20
3	30554	W	0.0162237	0.0141116	0.00208803	880	198
3	30554	C	0.0185255	0.0151088	0.0027518	898	191
3	42578	W	0.109951	0.0567932	0.0124646	997	40
3	42578	C	0.0882181	0.0592558	0.0142728	955	40
3	46914	W	0.0405973	0.0352331	0.00700902	828	69
3	46914	C	0.0577903	0.0420076	0.00890319	943	60

Table A.18: Results for gene classes taken from level 4 of the Gene Ontology heirarchy on chromosome 3 using TIGR data with the tandem duplications removed.

Chr	Class	Std	Orig. Grnwd	MC Grnwd	SD MC grnwd	Ranking	Examples
4	04175	W	0.0809505	0.0859348	0.0179553	472	23
4	04175	C	0.0903446	0.116667	0.0293574	127	17
4	04518	W	0.130448	0.163497	0.0415675	187	11
4	04518	C	0.14075	0.154779	0.0353639	394	12
4	04553	W	0.0978304	0.0976322	0.024293	628	20
4	04553	C	0.0641654	0.0596606	0.0117898	730	36
4	04930	W	0.0852606	0.132234	0.0334619	7	14
4	04930	C	0.114211	0.163782	0.0407241	34	11
4	05216	W	0.282537	0.210099	0.0537041	908	8
4	05216	C	0.378377	0.194926	0.0498971	992	9
4	05351	W	0.185814	0.258021	0.0673413	75	6
4	05351	C	0.193436	0.235765	0.0626581	248	7
4	05509	W	0.0643465	0.105155	0.0235894	0	18
4	05509	C	0.110167	0.10543	0.0250035	689	19
4	08168	W	0.15553	0.142188	0.0357029	742	13
4	08168	C	0.146812	0.135734	0.0302899	693	14
4	08234	W	0.0870523	0.101719	0.0214013	264	19
4	08234	C	0.0955579	0.0905114	0.0200065	686	23
4	08236	W	0.175203	0.162106	0.0395191	708	11
4	08236	C	0.156771	0.137068	0.0334364	775	14
4	08237	W	0.0932976	0.100795	0.0224498	439	19
4	08237	C	0.176593	0.259332	0.0670985	32	6
4	08238	W	0.124098	0.111082	0.0252676	769	17
4	08238	C	0.271337	0.233592	0.0573532	800	7
4	15077	W	0.210815	0.294016	0.0815613	80	5
4	15077	C	0.228909	0.165362	0.0389922	930	11
4	15291	W	0.0952999	0.112246	0.0241634	262	17
4	15291	C	0.11193	0.112212	0.0286301	592	18
4	15405	W	0.122235	0.123266	0.0265472	569	15
4	15405	C	0.103456	0.0926053	0.0198511	763	22
4	16301	W	0.0337529	0.0238267	0.00324817	987	91
4	16301	C	0.0284825	0.0258301	0.00453163	793	90
4	16616	W	0.0987236	0.139429	0.0321978	45	13
4	16616	C	0.131379	0.109607	0.0247594	845	18
4	16747	W	0.608741	0.229102	0.0620424	1000	7
4	16747	C	0.312581	0.231202	0.0601837	919	7
4	16758	W	0.0892111	0.107639	0.0254762	232	18
4	16758	C	0.108081	0.0997762	0.0216606	712	20
4	16773	W	0.0472746	0.0328362	0.00528212	981	66
4	16773	C	0.0396203	0.036579	0.00688912	759	62
4	16779	W	0.0808354	0.0886275	0.0177636	379	22
4	16779	C	0.0886001	0.122214	0.0282606	51	16
4	16789	W	0.196734	0.159638	0.0390374	867	11
4	16789	C	0.130747	0.121507	0.0281815	699	16
4	16818	W	0.0707813	0.0636646	0.0121313	770	32
4	16818	C	0.0999115	0.0801637	0.0163089	887	26
4	16820	W	0.294318	0.206485	0.0526063	933	8
4	16820	C	0.17155	0.120918	0.0272221	951	16
4	16836	W	0.28153	0.25623	0.0713785	730	6
4	16836	C	0.174081	0.145736	0.0346181	823	13
4	16881	W	0.284257	0.298884	0.0815264	535	5
4	16881	C	0.252904	0.151656	0.0347388	981	12
4	19001	W	0.0891474	0.0934613	0.0214568	515	21
4	19001	C	0.159834	0.144245	0.0326452	748	13
4	30554	W	0.0175082	0.0160527	0.0019055	809	142
4	30554	C	0.0165178	0.0178253	0.00315113	402	135
4	42578	W	0.0751362	0.101486	0.0225922	56	19
4	42578	C	0.121264	0.0733666	0.015393	986	29
4	46914	W	0.0469888	0.0406314	0.00644743	855	52
4	46914	C	0.0415339	0.0428242	0.0079099	511	52

Table A.19: Results for gene classes taken from level 4 of the Gene Ontology heirarchy on chromosome 4 using TIGR data with the tandem duplications removed.

Chr	Class	Std	Orig. Grnwd	MC Grnwd	SD MC grnwd	Ranking	Examples
5	04175	W	0.0602489	0.0578787	0.0120805	653	39
5	04175	C	0.049675	0.0620271	0.0137616	163	37
5	04518	W	0.0794314	0.0924232	0.0228543	310	23
5	04518	C	0.0785277	0.103793	0.0255301	93	20
5	04553	W	0.0927926	0.0657474	0.0137978	952	33
5	04553	C	0.0488214	0.0572722	0.012288	254	40
5	04930	W	0.0925231	0.117402	0.0302261	182	17
5	04930	C	0.174387	0.201923	0.0559239	360	9
5	05216	W	0.129782	0.184684	0.0522799	61	10
5	05216	C	0.106138	0.096063	0.022006	762	22
5	05351	W	0.162189	0.138181	0.0345115	815	14
5	05351	C	0.205858	0.243485	0.0697066	346	7
5	05509	W	0.0671803	0.0673656	0.0146473	604	32
5	05509	C	0.0688808	0.0751554	0.0160773	404	29
5	08168	W	0.0936445	0.0827125	0.0185907	771	26
5	08168	C	0.11064	0.118881	0.0275203	455	17
5	08234	W	0.0963299	0.100237	0.0258597	546	21
5	08234	C	0.109508	0.153285	0.0423915	89	13
5	08236	W	0.0787365	0.076426	0.0172648	648	28
5	08236	C	0.106795	0.0959824	0.0233556	746	22
5	08237	W	0.135089	0.112475	0.0284624	826	18
5	08237	C	0.137204	0.111444	0.0268743	864	19
5	08238	W	0.172127	0.148829	0.0381274	803	13
5	08238	C	0.143806	0.151473	0.0415583	523	13
5	15077	W	0.179947	0.201872	0.0586895	431	9
5	15077	C	0.191129	0.120371	0.0301801	971	17
5	15291	W	0.0407779	0.0488679	0.0105603	202	46
5	15291	C	0.0580766	0.0585064	0.012613	564	39
5	15405	W	0.108183	0.102527	0.0240083	674	20
5	15405	C	0.10619	0.0783856	0.0175298	936	28
5	16301	W	0.0306068	0.018271	0.00324143	992	133
5	16301	C	0.0220163	0.0173014	0.00340377	910	151
5	16616	W	0.0801271	0.0833716	0.0183193	503	25
5	16616	C	0.0957847	0.100318	0.0237078	514	21
5	16747	W	0.174295	0.140148	0.0360842	860	14
5	16747	C	0.209899	0.133287	0.0345469	970	15
5	16758	W	0.116678	0.0937626	0.0214066	885	22
5	16758	C	0.20706	0.0861903	0.0188635	1000	25
5	16773	W	0.0374379	0.0236634	0.00432713	992	101
5	16773	C	0.0251962	0.0227385	0.00478526	781	112
5	16779	W	0.0791245	0.0899414	0.0204316	337	23
5	16779	C	0.14305	0.118783	0.0294427	828	17
5	16789	W	0.0927072	0.0763825	0.0170208	850	28
5	16789	C	0.0679968	0.0860694	0.019992	134	25
5	16818	W	0.067962	0.0487938	0.00959897	956	46
5	16818	C	0.0499645	0.0406861	0.00890985	867	59
5	16820	W	0.144629	0.146245	0.0387674	599	13
5	16820	C	0.113412	0.0921847	0.0211511	866	23
5	16836	W	0.152897	0.169911	0.045352	415	11
5	16836	C	0.191063	0.223132	0.065604	356	8
5	16881	W	0.389263	0.30803	0.0899876	837	5
5	16881	C	0.1838	0.17098	0.0478417	704	11
5	19001	W	0.0920855	0.074159	0.0167015	876	29
5	19001	C	0.138827	0.0829745	0.0194841	984	26
5	30554	W	0.0176474	0.0106649	0.00150047	998	235
5	30554	C	0.0162879	0.0107516	0.00193222	980	250
5	42578	W	0.0690198	0.0755282	0.0161162	392	28
5	42578	C	0.0603234	0.0627738	0.0136233	515	36
5	46914	W	0.0358101	0.0241335	0.00444273	979	99
5	46914	C	0.0382661	0.0304407	0.00663895	880	82

Table A.20: Results for gene classes taken from level 4 of the Gene Ontology heirarchy on chromosome 5 using TIGR data with the tandem duplications removed.

Appendix B

Supplementary Tables on Neighbouring Genes

Class A	Class B	Examples	Expectation	Pmf
go:5198	go:5198	30.0	15.59	0.00379484
go:16209	go:16209	1.0	0.04	0.0400002
go:3774	go:30234	2.0	0.32	0.0511972
go:31386	go:4871	1.0	0.06	0.0600005
go:3774	go:4871	1.0	0.10	0.100002
go:5198	go:3824	64.0	82.91	0.106225
go:16209	go:3824	9.0	4.39	0.107706
go:45182	go:3824	18.0	11.11	0.129862
go:30234	go:4871	3.0	1.01	0.169991
go:5488	go:31386	3.0	1.03	0.176792
go:30188	go:30234	1.0	0.18	0.180005
go:5488	go:5488	104.0	86.94	0.18552
go:30528	go:3774	2.0	0.61	0.186056
go:3774	go:3824	7.0	3.62	0.204326
go:5215	go:5488	31.0	42.21	0.223138
go:3824	go:4871	6.0	11.62	0.260977
go:45182	go:4871	1.0	0.29	0.290013
go:5215	go:3774	2.0	0.78	0.304226
go:4871	go:30528	4.0	1.95	0.308946
go:5215	go:5215	27.0	20.49	0.337891
go:16209	go:5198	2.0	0.83	0.344484
go:30234	go:16209	1.0	0.38	0.380022
go:5198	go:4871	4.0	2.19	0.399637
go:5198	go:5215	12.0	17.87	0.406288
go:4871	go:5488	2.0	5.16	0.436291
go:5488	go:30528	40.0	32.92	0.443447
go:5215	go:30188	1.0	0.45	0.450031
go:30188	go:5215	1.0	0.45	0.450031

Table B.1: The molecular function classes of neighbouring pairs (84 results). Table 1 of 3.

Class A	Class B	Examples	Expectation	Pmf
go:30528	go:5198	9.0	13.94	0.453482
go:5488	go:5215	34.0	42.21	0.46957
go:5215	go:31386	1.0	0.50	0.500038
go:31386	go:5488	2.0	1.03	0.515002
go:30528	go:30234	9.0	6.47	0.537252
go:5198	go:5488	43.0	36.82	0.563594
go:5488	go:45182	7.0	4.93	0.570568
go:3824	go:45182	14.0	11.11	0.627645
go:3824	go:5198	74.0	82.91	0.639355
go:30528	go:30528	9.0	12.46	0.681607
go:3824	go:30188	3.0	2.07	0.690007
go:5215	go:30528	19.0	15.98	0.700769
go:3824	go:3824	457.0	440.89	0.720026
go:16209	go:30528	1.0	0.74	0.740083
go:31386	go:3824	3.0	2.32	0.773371
go:3824	go:5488	186.0	195.78	0.793428
go:5215	go:45182	3.0	2.39	0.796714
go:45182	go:5215	3.0	2.39	0.796714
go:5488	go:3774	2.0	1.61	0.805074
go:4871	go:3824	9.0	11.62	0.814144
go:30234	go:3824	42.0	38.48	0.815667
go:5488	go:30234	14.0	17.09	0.816622
go:3824	go:30528	79.0	74.13	0.826535
go:30528	go:5488	36.0	32.92	0.830468
go:5215	go:4871	3.0	2.51	0.836731
go:5488	go:5198	40.0	36.82	0.837617
go:30234	go:30234	4.0	3.36	0.840046
go:3824	go:5215	99.0	95.05	0.902646

Table B.2: The molecular function classes of neighbouring pairs (84 results). Table 2 of 3.

Class A	Class B	Examples	Expectation	Pmf
go:5198	go:30234	8.0	7.24	0.905033
go:5215	go:30234	9.0	8.30	0.922264
go:30234	go:5215	9.0	8.30	0.922264
go:5198	go:30528	15.0	13.94	0.925474
go:30234	go:30528	5.0	6.47	0.927151
go:5215	go:3824	91.0	95.05	0.934617
go:30528	go:5215	17.0	15.98	0.938962
go:5215	go:16209	1.0	0.95	0.950137
go:5215	go:5198	16.0	17.87	0.951045
go:45182	go:30234	1.0	0.97	0.970142
go:5488	go:3824	192.0	195.78	0.971682
go:30528	go:4871	2.0	1.95	0.97514
go:3824	go:30234	37.0	38.48	0.987304
go:30234	go:5488	16.0	17.09	0.994569
go:30528	go:3824	73.0	74.13	0.998074
go:3774	go:5488	1.0	1.61	1
go:3824	go:3774	3.0	3.62	1
go:3824	go:16209	4.0	4.39	1
go:3824	go:31386	2.0	2.32	1
go:4871	go:5198	2.0	2.19	1
go:4871	go:5215	2.0	2.51	1
go:4871	go:30234	1.0	1.01	1
go:5488	go:4871	5.0	5.16	1
go:16209	go:5488	1.0	1.95	1
go:30188	go:3824	2.0	2.07	1
go:30234	go:5198	7.0	7.24	1
go:45182	go:5488	4.0	4.93	1
go:45182	go:30528	1.0	1.87	1

Table B.3: The molecular function classes of neighbouring pairs (84 results) . Table 3 of 3.

References

Adams, E. N. (III). 1972. Consensus techniques and the comparison of taxonomic trees. *Systematic Zoology*, **21**(4), 390–397.

Agrawal, R., & Srikant, R. 1994. Fast algorithms for mining association rules. *In: Proc. of the 20th Int'l Conference on Very Large Databases.*

Agrawal, R., & Srikant, R. 1995. Mining sequential patterns. *Pages 3–14 of: Proc. 1995 Int. Conf. Data Engineering (ICDE95).*

Agrawal, R., Imielinski, T., & Swami, A. 1993. Mining association rules between sets of items in large databases. *In: Proc. of the ACM SIGMOD Conference on Management of Data.*

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., & Verkamo, A. I. 1996. *Fast discovery of association rules.* Menlo Park, CA, USA: American Association for Artificial Intelligence.

Aho, A. V., Sagiv, Y., Szymanski, T. G., & Ullman, J. D. 1981. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM J. Comput.*, **10**(3), 405–421.

Ait-Kaci, H. 1991. *Warren's abstract machine: A tutorial reconstruction.* MIT Press, Cambridge MA.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. 2002. *Molecular Biology of the Cell (4th Ed.).* Garland Publishing.

Aleksander, I., & Morton, H. 1995. *An introduction to neural computing (2nd ed.).* International Thomson Computer Press.

Altman, D. 1991. *Practical statistics for medical research.* Chapman and Hall, London.

Altmann-Johl, R, & Philippsen, P. 1996. AgTHR4, a new selection marker for transformation of the filamentous fungus *Ashbya gossypii*, maps in a four-gene cluster that is conserved between *A. gossypii* and *Saccharomyces cerevisiae*. *Molecular Genetics and Genomics*, **250**, 69–80.

Altschul, S. F., & Gish, W. 1996. Local alignment statistics. *Methods in enzymology*, **266**, 460–480.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. 1990. Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W, & Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25(17)**, 3389.

Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T., Chothia, C., & Murzin, A. G. 2004. Scop database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Research*, **32**, D226–D229.

Apic, G., Huber, W., & Teichmann, S. A. 2003. Multi-domain protein families and domain pairs: comparison with known structures and a random model of domain recombination. *J. Struct. Funct. Genomics*, **4(2–3)**, 67–78.

Arthurs, A. M. 1965. *Probability theory, library of mathematics.* Routledge & Kegan Paul.

Audit, B, & Ouzounis, C. A. 2003. From genes to genomes: Universal scale-invariant properties of microbial chromosome organisation. *J. Mol. Biol.*, **332**, 617–633.

Bayardo, R. J. 1998. Efficiently mining long patterns from databases. *Pages 85–93 of: ACM-SIGMOD Int. Conf. Management of Data (SIGMOD98).*

Bennett, M. D., Leitch, I. J., Price, H. J., & Johnston, J. S. 2003. Comparisons with caenorhabditis (approximately 100 mb) and drosophila (approximately 175 mb) using flow cytometry show genome size in arabidopsis to be approximately 157 mb and thus approximately 25 % larger than the arabidopsis genome initiative estimate of approximately 125 mb. *Annals of Botany*, **91**, 547.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Rapp, B. A., & Wheeler, D. L. 2000. GenBank. *Nucl. Acids Res.*, **28**(1), 15–18.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Wheeler, D. L. 2007. GenBank. *Nucl. Acids Res.*, **35**(suppl_1), D21–25.

Berglund, J., Pollard, K. S., & Webster, M. T. 2009. Hotspots of biased nucleotide substitutions in human genes. *PLoS Biology*, **7**, e1000026.

Bininda-Emonds, O. R. P., M., Cardillo, Jones, K. E., MacPhee, R. D. E., Beck, R. M. D., Grenyer, R., Price, S. A., Vos, R. A., Gittleman, J. L., & Purvis, A. 2007. The delayed rise of present-day mammals. *Nature*, **446**, 507–512.

Bland, J. M., & Altman, D. G. 2000. Statistics notes: The odds ratio. *BMJ*, **320**, 1468.

Blumenthal, T. 2004. Operons in eukaryotes. *Brief Funct Genomic Proteomic*, **3**, 199–211.

Blumenthal, T., & Gleason, K. S. 2003. *Caenorhabditis elegans* operons: Form and function. *Nature Reviews Genetics*, **4**, 110–118.

Bock, C., & Lengauer, T. 2008. Computational epigenetics. *Bioinformatics*, **24**, 1–10.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., & Schneider, M. 2003. The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic Acids Res.*, **31**, 365–370.

Bratko, I. 2001. *Prolog: Programming for artificial intelligence (3rd ed.)*. Addison Wesley.

Brenner, S. E., Koehl, P., & Levitt, M. 2000. The ASTRAL compendium for sequence and structure analysis. *Nucl. Acids Res.*, **28**, 254–256.

Brin, S., Motwani, R., & Silverstein, C. 1997. Beyond market basket: Generalizing association rules to correlations. *Pages 265–276 of: In Proc. 1997 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD97)*.

Bryant, D. 2003. A classification of consensus methods for phylogenetics. *Bio-Consensus DIMACS AMS*, **-**, 163–184.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G., & Thompson, J. D. 2003. Multiple sequence alignment with the clustal series of programs. *Nucleic Acids Res*, **31(13)**, –.

Chothia, C. 1992. Proteins. one thousand families for the molecular biologist. *Nature*, **357**, 543–544.

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B. A., & Johnston, M. 2003. Finding functional features in saccharomyces genomes by phylogenetic footprinting. *Science*, **301**, 71–76.

Clocksin, W. F., & Mellish, C. S. 1984. *Programming in Prolog (2nd ed)*. Springer Verlag, Berlin Heidelberg.

Croft, A., Davison, R., & Hargreaves, M. 1992. *Engineering mathematics: A modern foundation for electronic, electrical and control engineers*. Addison Wesley.

D'Agostino, Ralph B., & Stephens, Michael A. 1986. *Goodness-of-fit techniques*. Marcel Dekker, Inc., New York.

Dayhoff, M. O. 1979. *Atlas of protein sequence and structure*. Vol. 5. Washington: National Biomedical Research Foundation. Pages 345–352.

De Martelaere, D. A., & Van Gool, A. P. 1981. The density distribution of gene

loci over the genetic map of *Escherichia coli*: Its structural, functional and evolutionary implications. *J. Mol. Evol.*, **17(6)**, 354–360.

Dehaspe, L., Toivonen, H., & King, R. D. 1998. Finding frequent substructures in chemical compounds. *Pages 30–36 of: In the Fourth International Conference on Knowledge Discovery and Data Mining, AAAI Press.*

Dietrich, F. S., Voegeli, S., Brachat, S., Lerch, A., Gates, K., Steiner, S., Mohr, C., Pöhlmann, R, Luedi, P., Choi, S., Wing, R. A., Flavier, A., Gaffney, T. D., & Philippsen, P. 2004. The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science*, **304**, 304–307.

Dong, G., & Li, J. 1999. Efficient mining of emerging patterns: Discovering trends and differences. *Pages 43–52 of: In Proc. 1999 Int. Conf. Knowledge Discovery and Data Mining (KDD99).*

Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, **450**, 203–218.

Dudewicz, E. J., & Mishra, S. N. 1988. *Modern mathematical statistics.* John Wiley & Sons.

Dunn, C. W., Hejnol, A., Matus, D. Q., Pang, K., Browne, W. E., Smith, S. A., Seaver, E., Rouse, G. W., Obst, M., Edgecombe, G. D., Sørensen, M. V., Haddock, S. H. D., Schmidt-Rhaesa, A., Okusu, A., Møbjerg Kristensen, R., Wheeler, W. C., Martindale, M. Q., & Giribet, G. 2008. Broad phyolgenomic sampling improves resolution of the animal tree of life. *Nature*, **452**, 745.

Durand, D., & Sankoff, D. 2003. Tests for gene clustering. *Journal of Computational Biology*, **10(3-4)**, 453–482.

Ekman, D., Bjrklund, A. K., Frey-Sktt, J., & Elofsson, A. 2005. Multi-domain proteins in the three kingdoms of life: Orphan domains and other unassigned regions. *Journal of Molecular Biology*, **348(1)**, 231–243.

Elliot, H. C., & Elliot, D. C. 1997. *Biochemistry and molecular biology.* Oxford University Press.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (eds.). 1996. *Advances in knowledge discovery and data mining*. AAAI/MIT Press.

Feller, W. 1950. *An introduction to probability theory and its applications*. John Wiley & Sons, Inc.

Felsenstein, J. 1988. Phylogenies from molecular sequences:inference and reliability. *Annu. Rev. Genet.*, **22**, 521–565.

Felsenstein, J. 2004. *Inferring phylogenies*. Sinauer Associates Inc. Sunderland, Massachusetts.

Feng, D., & Doolittle, R. F. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol*, **60**, 351–360.

Ferguson, L.R., & von Borstel, R. C. 1992. Induction of the cytoplasmic 'petite' mutation by chemical and physical agents in *Saccharomyces cerevisiae*. *Mutation Research*, **265**, 103–48.

Fitch, W. M. 1966. An improved method of testing for evolutionary homology. *J. Mol. Biol.*, **16**, 9–16.

Fitch, W. M., & Margoliash, E. 1967. Construction of phylogenetic trees. a method based on mutation distances as estimated from cytochrome c sequences is of general applicability. *Science*, **155**, 279–284.

Foury, F., Roganti, T., Lecrenier, N., & Purnelle, B. 1998. The complete sequence of the mitochondrial genome of *saccharomyces cerevisiae*. *Febs Letts.*, **440(3)**, 325–331.

Galtier, N., Duret, L., Glémen, S., & Ranwez, V. 2009. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends in Genetics*, **25**, 1–5.

Gavin, A. C., *et al.* . 2006a. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Gavin, A. C., *et al.* . 2006b. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**, 631–636.

Giannesini, F., Kanoui, H., Pasero, R., & van Caneghem, N. 1986. *Prolog.* Addison Wesley.

Goad, W. B., & Kanehisa, M. I. 1982. Pattern recognition in nucleic acid sequences. I. A general method for finding local homologies and symmetries. *Nucleic Acids Research*, **10(1)**, 247–263.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., & Oliver, S. G. 1996. Life with 6000 genes. *Science*, **274(5287)**, 546–567.

Goldberg, A. D., Allis, C. D., & Bernstein, E. 2007. Epigenetics: a landscape takes shape. *Cell*, **128**, 635–638.

Gough, J., Karplus, K., Hughey, R., & Chothia, C. 2001. Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure. *J. Mol. Biol.*, **313(4)**, 903–919.

Grahne, G., Lakshmanan, L., & Wang, X. 2000. Efficient mining of constrained correlated sets. *Pages 512–521 of: In Proc. 2000 Int. Conf. Data Engineering (ICDE-00).*

Greenwood, Major. 1946. The statistical study of infectious diseases. *Journal of the Royal Statistical Society*, **109(2)**, 85–110.

Haas, B. J., Wortman, J. R., Ronning, C. M., Hannick, L. I., Smith Jr., R. K., Maiti, R., Chan, A. P., Yu, C., Farzad, M., Wu, D., White, O, & Town, C. D. 2005. Complete reannotation of the *Arabidopsis* genome: methods, tools, protocols and the final release. *BMC Biology*, **3(7)**.

Hajibabaei, M., Singer, G. A. C., Hebert, P. D. N., & Hickey, D. A. 2007. DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends in Genetics*, **23**, 167–172.

Han, J., Dong, G., & Yin, Y. 1999. Efficient mining of partial periodic patterns in time series database. *Pages 106–115 of: In Proc. 1999 Int. Conf. Data Engineering (ICDE-99).*

Higgins, D.G., & Sharp, P.M. 1988. Clustal: a package for performing multiple sequence alignment on a microcomputer. *Gene*, **73**, 237–244.

Hipp, J.and Güntzer, U., & Nakhaeizadeh, G. 2000. Algorithms for association rule mining — a general survey and comparison. *SIGKDD Explor. Newsl.*, **2(1)**, 58–64.

Hughes, J., Longhorn, S. J., Papadopoulou, A., Theodorides, K., de Riva, A., Mejia-Chang, M., Foster, P. G., & Vogler, A. P. 2006. Dense taxonomic EST sampling and its applications for molecular systematics of the *coleoptera* (beetles). *Mol. Biol. Evol.*, **23**, 268–278.

Hurst, L. D. 2009. A positive becomes a negative. *Nature*, **457**, 543–544.

Huson, D.H., Richter, D.C., Rausch, C., Dezulian, T., Franz, M., & Rupp, R. 2007. Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics*, **8**, 460.

Islam, S. A., Luo, J., & Sternberg, M. J. E. 1995. Identification and analysis of domains in proteins. *Prot. Eng.*, **8**, 513–525.

Jacob, F., & Monod, J. 1961. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, **3**, 318–356.

Jones, P. A., & Baylin, S. B. 2007. The epigenomics of cancer. *Cell*, **128**, 683–692.

Jones, S., Stewart, M., Michie, A., Swindells, M. B., Orengo, C., & Thornton, J. M. 1998. Domain assignment for protein structures using a consensus approach: characterization and analysis. *Protein Sci.*, **7**, 233–242.

Kamber, M., Han, J., & Chiang, J. Y. 1997. Metarule-guided mining of multi-dimensional association rules using data cubes. *Pages 207–210 of: In Proc. 1997 Int. Conf. Knowledge Discovery and Data Mining (KDD97).*

Karr, A. F. 1993. *Probability.* Springer Verlag.

Karwath, A., & King, R. D. 2002. Homology induction: the use of machine learning to improve sequence similarity searches. *BMC Bioinformatics*, **3**, 11.

Képès, F. 2004. Periodic transcriptional organisation of the *e. coli* genome. *J. Mol. Biol.*, **340**, 957–964.

King, R. D., Srinivasan, A., & Dehaspe, L. 2001. Warmr: a data mining tool for chemical data. *J. Comput. Aided Mol. Des.*, **15(2)**, 173–181.

King, R. D., Whelan, K. E., Jones, F. M., Reiser, P. G. K., Bryant, C. H., Muggleton, S. H., & Kell, D. B. 2004. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, **427**, 247–252.

Kirk, R. E. 1999. *Statistics: An introduction (fourth ed.)*. Harcourt Brace College Publishers.

Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H., & Verkamo, A. I. 1994. Finding interesting rules from large sets of discovered association rules. *Pages 401–408 of: Proc. 3rd int. conf. information and knowledge management.*

Korn, L. J., Queen, C. L., & Wegman, M. N. 1977. Computer analysis of nucleic acid regulatory sequences. *Proc. Natl. Acad. Sci. U S A*, **74(10)**, 4401–4405.

Lackie, J. M. 2007. *Dictionary of Cell and Molecular Biology (4th Ed.)* . Academic Press.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., & Higgins, D.G. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**(21), 2947–2948.

Lent, B., Swami, A., & Widom, J. 1997. Clustering association rules. *Pages 220–231 of: Int. Conf. Data Engineering (ICDE-97).*

Lin, X., Kaul, S., Rounsley, S., Shea, T. P., Benito, M. I., Town, C. D., Fujii, C. Y., Mason, T., Bowman, C. L., Barnstead, M., *et al.* . 1999. Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature*, **402**, 761–768.

Lo Conte, L., Brenner, S. E., Hubbard, T., Chothia, C., & Murzin, A. G. 2002.

Scop database in 2002: refinements accommodate structural genomics. *Nucleic Acids Research*, **30**, 264–267.

Maddison, D. R., & (eds.), K.-S. Schulz. 2007. *The tree of life web project.*

Maddison, D. R., Schulz, K. S., & Maddison, W. P. 2007. The tree of life web project. *Zootaxa*, **1668**, 19–40.

Mahillon, J., & Chandler, M. 1998. Insertion sequences. *Microbiology and Molecular Biology Reviews*, **62(3)**, 725–774.

Mannila, H., Toivonen, H., & Verkamo, A. I. 1994. Efficient algorithms for discovering association rules. *Pages 181–192 of: Proceedings of the AAAI Workshop on Knowledge Discovery in Databases.* AAAI Press.

Mannila, H., Toivonen, H., & Verkamo, A. I. 1997. Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery*, **1**, 259–289.

Mayer, K., Schüller, C., Wambutt, R., Murphy, G., Volckaert, G., Pohl, T., Düsterhöft, A., Stiekema, W., Entian, K. D., Terryn, N., *et al.* . 1999. Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. *Nature*, **402**, 769–777.

Mayor, L. R., Fleming, K. P., Muller, A., Balding, D. J., & Sternberg, M. J. E. 2004. Clustering of protein domains in the human genome. *J. Mol. Biol.*, **340**, 991–1004.

Metropolis, N., & Ulam, S. 1949. The Monte Carlo method. *J. Amer. Stat. Assoc.*, **44**, 335–341.

Mitchell, T. M. 1997. *Machine learning.* McGraw-Hill.

Moreno-Hagelsieb, G., & Collado-Vides, J. 2002. A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics*, **18(Suppl.)**, S329–S336.

Mueller, A. 1995. *Fast sequential and parallel algorithms for association rule mining: A comparison.* Tech. rept. Department of Computer Science, University of Maryland.

Murzin, A. G., Brenner, S. E., Hubbard, T., & Chothia, C. 1995. Scop: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Nakhleh, L., Miranker, D., Barbancon, F., Piel, W. H., & Donoghue, M. 2003. Requirements of phylogenetic databases. *Page 141 of: BIBE '03: Proceedings of the 3rd IEEE Symposium on BioInformatics and BioEngineering*. Washington, DC, USA: IEEE Computer Society.

Needleman, S. B., & Wunsch, C. D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.

Nei, M. 1975. *Molecular population genetics and evolution*. North-Holland, Amsterdam.

Ng, R., Lakshmanan, L. V. S., Han, J., & Pang, A. 1998. Exploratory mining and pruning optimizations of constrained associations rules. *Pages 13–24 of: In Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD98)*.

Ohm, J. E., McGarvey, K. M., Yu, X., Cheng, L., Schuebel, K. E., Cope, L., Mohammad, H. P., Chen, W., Daniel, V. C., Yu, W., Berman, D. M., Jenuwein, T., Pruitt, K., Sharkis, S. J., Watkins, D. N., Herman, J. G., & Baylin, S. B. 2007. A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. *Nat. Genet.*, **39**, 237–242.

Ohno, S. 1970. *Evolution by gene duplication*. George Allen and Unwin, London.

Papoulis, A. 1965. *Probability, random variables, and stochastic processes*. McGraw-Hill.

Park, S. K., & Miller, K. W. 1988. Random number generators: Good ones are hard to find. *Commun. ACM*, **31(10)**, 1192–1201.

Pearson, W. R. 1990. Rapid and sensitive sequence comparison with fastp and fasta. *Methods enzymol*, **183**, 63–98.

Pearson, W. R., & Lipman, D. J. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, **85**, 2444–2448.

Piel, W. H., Donoghue, M. J., & Sanderson, M. J. 2002. Treebase: A database of phylogenetic information. *Research Report from the National Institute for Environmental Studies, Japan*, **171**, 41–47.

R Development Core Team. 2005. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.* .

Richardson, J. S. 1981. The anatomy and taxonomy of protein structure. *Adv. Protein Chem.*, **34**, 167–339.

Riley, M., Solomon, L., & Zipkas, D. 1978. Relationship between gene function and gene location in *escherichia coli*. *J. Mol. Evol.*, **11(1)**, 47–56.

Riley, M. C., Clare, A., & D., King R. 2007. Locational distribution of gene functional classes in *Arabidopsis thaliana*. *BMC Bioinformatics*, **8**, 112.

Robinson, J. A. 1965. A machine-oriented logic based on the resolution principle. *Journal of the Association for Computing Machinery*, **12**(1), 23–41.

Rocha, E. P. C., Danchin, A., & Viari, A. 1999. Universal replication biases in bacteria. *Molecular Microbiology*, **32(1)**, 11–16.

Ruelle, D. 1991. *Chance and chaos*. Penguin.

Saitou, N., & Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.

Salanoubat, M., Lemcke, K., Rieger, M., Ansorge, W., Unseld, M., Fartmann, B., Valle, G., Blöcker, H., Perez-Alonso, M., Obermaier, B., *et al.* . 2000. Sequence and analysis of chromosome 3 of the plant *Arabidopsis thaliana*. *Nature*, **408**, 820–822.

Sanderson, M. J. 2008. Phylogenetic signal in the eukaryotic tree of life. *Science*, **321**, 121–123.

Sanderson, M.J. 2007. (L. A. S. JOHNSON REVIEW No. 9) construction and

annotation of large phylogenetic trees. *Australian systematic botany*, **20**, 287–301.

Sankoff, D. 1972. Matching sequences under deletion/insertion constraints. *Proc. Natl. Acad. Sci. U S A*, **69(1)**, 4–6.

Sankoff, D., & Kruskal, J. 1983. *Time warps, string edits and macromolecules: The theory and practice of sequence comparison*. Center for the Study of Language and Inf; Reissue ed edition (December 1, 1999).

SanMiguel, P., & Bennetzen, J. L. 1998. Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotranposons. *Annals of Botany*, **82**, 37–44.

Savageau, M. A. 1986. Proteins of escherichia coli come in sizes that are multiples of 14 kda: domain concepts and evolutionary implications. *Proc. Nat.l Acad. Sci. U S A*, **83**, 1198–1202.

Schlesinger, Y., Straussman, R., Keshet, I., Farkash, S., Hecht, M., Zimmerman, J., Eden, E., Yakhini, Z., Ben-Shushan, E., Reubinoff, B. E. Bergman, Y., Simon, I., & Cedar, H. 2007. Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. *Nat. Genet.*, **39**, 232–236.

Schneider, E., Blundell, M., & Kennell, D. 1978. Translation and mRNA decay. *Molecular Genetics and Genomics*, **160(2)**, 121–129.

Sellars, P. H. 1974. On the theory and computation of evolutionary distances. *SIAM J. Appl. Math.*, **26**, 787–793.

Sellars, P. H. 2000. Pattern recognition in genetic sequences. *Proc. Natl. Acad. Sci. U S A*, **76**, 3041.

Siddiqui, A. S., & Barton, G. J. 1995. Continuous and discontinuous domains - an algorithm for the automatic generation of reliable protein domain definitions. *Protein Sci.*, **4**, 872–884.

Silverstein, C., Brin, S., Motwani, R., & Ullman, J. 1998. Scalable techniques for

mining causal structures. *Pages 594–605 of: In Proc. 1998 Int. Conf. Very Large Data Bases (VLDB98).*

Smith, G. P. 1976. Evolution of repeated dna sequences by unequal crossover. *Science*, **191**, 528–535.

Smith, T. F., & Waterman, M. S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Smith, T. F., Waterman, M. S., & Fitch, W. M. 1981. Comparative biosequence metrics. *J. Mol. Evol.*, **18**, 38–46.

Sneath, P. H. A., & Sokal, R. R. 1973. *Numerical taxonomy.* Freeman, San Fransisco.

Spilianakis, C. G., Lalioti, M. D., Town, T., Lee, G. R., & Flavell, R. A. 2005. Interchromosomal associations between alternatively expressed loci. *Nature*, **435**, 637–645.

Srikant, R., Vu, Q., & Agrawal, R. 1997. Mining association rules with item constraints. *Pages 67–73 of: In Proc. 1997 Int. Conf. Knowledge Discovery and Data Mining (KDD97).*

Steinke, Dirk, Salzburger, Walter, & Meyer, Axel. 2006. Novel relationships among ten fish model species revealed based on a phylogenomic analysis using ESTs. *J. Mol. Evol.*, **62**, 772–784.

Surani, M. A., *et al.* . 2007. Genetic and epigenetic regulators of pluripotency. *Cell*, **128**, 747–762.

Szallasi, Z. 2001. To kill two birds with one stone: a general concept in gene regulation? *TRENDS Pharmacolog. Sci.*, **22**, 110.

Tabachnick, B. G., & Fidell, L. S. 1996. *Using multivariate statistics (3rd ed.).* New York: Harper Collins.

Tabata, S., Kaneko, T., Nakamura, Y., Kotani, H., Kato, T., Asamizu, E., Miyajima, N., Sasamoto, S., Kimura, T., Hosouchi, T., *et al.* . 2000. Sequence

and analysis of chromosome 5 of the plant *Arabidopsis thaliana*. *Nature*, **408**, 823–826.

Tamarin, R. H. 1999. *Principles of genetics (sixth ed.)*. McGraw-Hill.

The Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **480**, 796–815.

The Gene Ontology Consortium. 2000. Gene ontology: Tool for the unification of biology. *Nature Genetics*, **25**, 25–29.

Theologis, A., Ecker, J. R., Palm, C. J., Federspiel, N. A., Kaul, S., White, O., Alonso, J., Altafi, H., Araujo, R., Bowman, C. L., *et al.* . 2000. Sequence and analysis of chromosome 1 of the plant *Arabidopsis thaliana*. *Nature*, **408**, 816–820.

Thompson, J. D., Higgins, D. G., & Gibson, T. J. 1994. CLUSTALW:improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Tukey, J. W. 1977. *Exploratory data analysis*. Addison-Wesley, London.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., *et al.* . 2001. The sequence of the human genome. *Science*, **291**, 1304–1351.

Verstrepen, K. J., Jansen, A., Lewitter, F., & Fink, G. F. 2005. Intragenic tandem repeats generate functional variability. *Nature Genetics*, **37(9)**, 986–990.

Waddington, C. H. 1942. The epigenotype. *Endeavour*, **1**, 18–20.

Warren, D. H. D. 1983. *An abstract Prolog instruction set*. Tech. rept. Technical report 309. SRI.

Warren, P.B., & ten Wolde, P. R. 2004a. Enhancement of the stability of genetic switches by overlapping upstream regulatory domains. *Phys. Rev. Lett.*, **92(12)**, 1281.

Warren, P.B., & ten Wolde, P. R. 2004b. Statistical analysis of the spatial distribution of operons in the transcriptional regulation network of *escherichia coli*. *J. Mol. Biol*, **342(5)**, 1379–1390.

Wetlaufer, D. B. 1973. Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc. Natl. Acad. Sci. U S A*, **70**, 697–701.

Wheeler, D. L., Chappey, C., Lash, A. E., Leipe, D. D., Madden, T. L., Schuler, G. D., Tatusova, T. A., & Rapp, B. A. 2000. Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **28(1)**, 10–14.

Wilbur, W. J., & Lipman, D. J. 1983. Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl. Acad. Sci. USA*, **80**, 726–730.

Willy, P. J., & Kobayashi, R. 2000. A basal transcription factor that activates or represses transcription. *Science*, **290**, 982.

Woese, C., & Fox, G. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. U S A*, **74(11)**, 5088–90.

Woese, C., Magrum, L., & Fox, G. 1978. Archaebacteria. *J Mol Evol*, **11(3)**, 245–51.

Woese, C., Kandler, O., & Wheelis, M. 1990. Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. *Proc. Natl. Acad. Sci. U S A*, **87(12)**, 4576–9.

Xiong, J. 2008. *Essential bioinformatics*. Cambridge University Press, New York.

Yule, G. U., & Kendall, M. G. 1946. *An introduction to the theory of statistics*. Charles Griffin & Co. Ltd.

Zhang, S., Herbert, G. G., Wang, J. T. L, H., Piel W., & Stockwell, D. R. B. 2006. PhyloMiner: A Tool for Evolutionary Data Analysis. *Pages 129–132 of: In Proceedings of the 18th International Conference on Scientific and Statistical Database Management (SSDBM'06)*.

Index

16s ribosomal RNA, 9

Alpha helix, 20
Apriori algorithm, 39
Arabidopsis thaliana, 51
Auxin binding protein, 176

Beta sheet, 20
Binomial coefficient, 79
Binomial distribution, 80
Biological process, 16
BLAST, 31
Bonferroni correction, 82

Cellular component, 16
Chromatin, 12
Chromosomal interaction, 13
Class (protein structure), 22
ClustalW, 30
Common fold, 22
Cytochrome c, 176
Cytochrome oxidase, 176

Dynamic programming algorithm, 26

E-values, 31
Elementary operations, 26
Epigenetics, 109

Factorial, 75
Family, 21

G-protein complex, 15
Gene location, 6
Gene ontology (GO), 16
Genes, 6
Genome, 6
Greenwood statistic, 83

Housekeeping genes, 15

Kurtosis, 71

Levenshtein distance, 28
Linnaean taxonomy, 45
Logic programming, 36

Markov chains, 34
Mean filtering, 73
Mobile genetic elements, 7
Model organisms, 51
Molecular function, 16
Molecular phylogeny, 45
Moments, 70
MRE11 complex, 13
Multi domain proteins, 12
Multinomial distribution, 81

NCBI taxonomy, 49

New Hampshire format, 50
Newick format, 50
Non-synonymous substitutions, 177

Operons, 10
Organellar DNA, 8

Phylogenetic trees, 46
Poisson approximation, 81
Poisson distribution, 77
Polecat problem, 171
Probability, 74
Probability mass function, 80
Profile based alignment, 33
Prolog, 36
Protein complex, 13
Protein domains, 18
Protein evolution, 176
Proteins, 18
Proteolysis, 11

Saccharomyces cerevisiae, 64
SCOP, 21
Secondary structure, 20
Sequence alignment, 26
Single locus mutation, 7
Skew, 71
Standard deviation, 71
Standard error, 72
Stirling's approximation, 76
Superfamily class, 22
Superfamily library, 24
Synonymous substitutions, 177

Tandem duplication, 7
Tertiary structure, 21
Transposons, 8
Tree of Life, 49, 179
TreeBASE, 48

Viral DNA, 8

WARMR, 43
Warren Abstract Machine, 37
Whole genome duplication, 8