New tools for comparative genomics based on oligonucleotide compositional constraints and single nucleotide polymorphisms

by

Hamilton Ganesan

Submitted in partial fulfillment of the requirements for the degree Philosophiae Doctor
(Bioinformatics)
in the Faculty of Natural and Agricultural Sciences
Bioinformatics and Computational Biology Unit
Department of Biochemistry
University of Pretoria
Pretoria
June 2009

Declaration

I, Hamilton Ganesan, declare that this thesis/dissertation, which I hereby submit for the degree Philosophiae Doctor at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

SIGNATURE . . . . . DATE . . . . . .

Publications relevant to this thesis

The following manuscript has been published :

Ganesan, H.; Rakitianskaia, A. S.; Davenport, C. F.; Tümmler, B. & Reva, O. N. (2008) The SeqWord Genome Browser: an online tool for the identification and visualization of atypical regions of bacterial genomes through oligonucleotide usage. BMC *Bioinformatics* **9**, 333.

Acknowledgments

I would first and foremost like to thank the Lord Jesus Christ without whom, I would accomplish nothing. ALL my successes i owe to You.

My family (Mum, Dad, Dane and Alida) who always loved and supported me in all things, I don't have the words that can fully express my thanks.

My supervisor Oleg, Thanks for going the extra-mile and helping me see this work through to completion. Your ever open doors and welcome are truly appreciated. I am indebted to you.

To my co-supervisor Fourie, thanks for all your efforts, friendliness and ever helpful attitude. You have made my PhD an enjoyable endevour.

To all my past and present colleagues at the Pretoria bioinformatics unit (Ayton, Charles, Corne, Oliver, Pieter, Tjaart), you guys have been an awesome bunch and I feel privileged to have worked with you all.

Summary

Tuberculosis is one of the leading causes of mortality globally. Although this disease has been around for many generations, treatment and management of the disease remains a daunting challenge. *M. tb*, is one of the most famous tuberculosis causing organisms, however there are many other mycobacterial strains and species that are also responsible for human mortality, globally. Not all mycobacterial species, however, are disease causing. It is only a few strains such as *M. tb* H37Rv, *M. tb* CDC1551, *M. tb* F11 and *M. bovis* which are responsible for causing disease. The rest are relatively harmless. What are the genetic differences between these virulent and avirulent strains that dictate a strain's behavior? The answers to these and many other questions lie hidden within the genomes of these organisms. Due to the great advances in DNA sequencing techniques, it is now now possible to more quickly and cheaply, sequence whole bacterial genomes in a single experimental run (High-throughput sequencing). Comparative genomics is therefore extremely relevant and important to be able to handle the dubious amounts of genomic data being poured into our public databases. Several comparative genomics environments already exist on the web today, however the goal of this project is to produce a web-based, comparative genomics environment which not only incorporates basic comparative genomics functions but also, novel tools such as the Seqword Genome Browser (SWGB) and the Mycobacterial Comparison Project (MCP). Using these tools, some interesting comparative genomics findings regarding certain strains of Mycobacteria are made. We reveal several genomic islands within *M. avium* and *M. tb* H37Rv. It is shown that certain genes which are usually found to be conserved among other bacteria, tend to be rather divergent among the mycobacteria. 'Mutational hotspots' containing many DNA replication genes are observed to have higher mutation rates relative to the rest of the genome which perhaps accounts for the slow-growth rate of these bacteria. By looking at the genetic profile of PE-PGRS genes in mycobacteria it was shown that *M. tb* H37Rv and *M. tb* F11 were actually closer for several genes than when compared to strain H37Ra. The finding was unexpected as H37Ra is known to be derived from H37Rv. These findings are extremely important in the area of TB research as it is of extreme importance to be able to trace areas of greater or lower selection within mycobacteria. Automated sequence comparison such as this is also important for tracking drug resistance markers and other features within mycobacteria so that more focused research can be carried out. The built system was tested and validated with mycobacteria, however, the system is flexible and designed with the intent of inclusion of any prokaryotic organism. It is hoped that systems such as these, and other advances in sequence comparison technology in the future, will provide the understanding needed to better control and cure diseases in the future.

Contents

1 Introduction **1**
1.1 What is Comparative Genomics? . . . 1
1.2 Sequencing technologies and the need for comparative genomics . . . 2
1.3 Common Methods used in Comparative Genomics . . . 3
1.3.1 Sequence Alignment & BLAST . . . 3
1.3.2 Genome Alignment . . . 5
1.3.3 Synteny . . . 8
1.3.4 Gene-by-gene comparative genomics . . . 11
1.3.5 Single Nucleotide Polymorphism analyses in comparative genomics . . . 12
1.3.6 Phylogenetic Analyses . . . 14
1.3.7 Regulatory Motif Discovery . . . 17
1.4 A Novel Comparative Genomic Technique using Oligonucleotide usage pattern profiling . . . 18
1.4.1 Codon Usage Bias . . . 18
1.4.2 Oligonucleotide Usage Bias . . . 21
1.5 Conclusions . . . 23
1.6 Problem Statement . . . 24
1.7 Aims . . . 25

2 An integrated comparative genomics environment **26**
2.1 Introduction . . . 26
2.2 FunGIMS . . . 27
2.2.1 Overview of FunGIMS . . . 27
2.2.2 Model . . . 27
2.2.3 View . . . 28
2.2.4 Controller . . . 28
2.3 Examples of comparative genomics environments and what they have to offer . . . 29
2.4 Requirements . . . 30
2.4.1 User interface requirements . . . 30

2.4.2 Analysis Requirements . . . 30
2.4.3 Data structure requirements . . . 31
2.5 Design Principles . . . 32
2.5.1 User interface requirements . . . 32
2.5.2 Data structure requirements . . . 32
2.5.3 Software components and technologies employed . . . 33
2.6 Model-View-Controller Architecture and integration . . . 33
2.6.1 Model-View-Controller Pattern . . . 33
2.6.2 Integration of the various components under the M-V-C design pattern . . . 34
2.6.2.1 The Model Layer . . . 34
2.6.2.2 The View Layer . . . 35
2.6.2.3 The Controller Layer . . . 35
2.7 Technical implementation details . . . 37
2.7.1 Database implementation . . . 37
2.7.2 Graphical User Interface (GUI) . . . 39
2.7.3 The Controller . . . 41
2.8 Implementation of a general comparative genomics environment . . . 41
2.8.1 DNA sequence alignment . . . 42
2.8.2 Genome alignment with BlastZ . . . 43
2.8.3 Phylogeny analyses . . . 47
2.9 Conclusion . . . 48

3 The Seqword Genome Browser 50
3.1 Introduction . . . 50
3.2 Background . . . 50
3.3 Results . . . 53
3.4 Identification of divergent genomic islands . . . 60
3.5 Scientific Investigation – Application to mycobacteria . . . 63
3.6 Discussion . . . 69

4 The Mycobacterial Comparison Project 72
4.1 Introduction . . . 72
4.2 Tuberculosis . . . 73
4.3 The Mycobacterial genome . . . 74
4.4 Comparative genomics of Mycobacteria . . . 77
4.5 The Mycobacterial Comparison Project in context . . . 79
4.6 Data pre-processing . . . 81
4.6.1 Mycobacterial strain selection . . . 81
4.6.2 Annotation Data . . . 81
4.6.3 Gene-by-gene mutation data . . . 81

4.6.4 SNP Data . . . 82
4.6.5 Gene island data . . . 82
4.7 Database requirements . . . 82
4.8 Graphical User Interface requirements . . . 83
4.9 Workflow summary . . . 84
4.10 A comparative genomics investigation of key genomic loci in mycobacterial genomes and their role in virulence . . . 84
4.11 Discussion . . . 95

5 Concluding Discussion **97**

Abbreviations

BLAST : Basic Local Alignment Search Tool

CAI : Codon Adaptation Index

CF : Cystis Fibrosis

CFTR : Cystic Fibrosis Transmembrane Conductance Regulator

D : Distance

DNA : Deoxy Ribo Nucleic Acid

FuGE : Functional Genomics Experiment

FunGIMS : Functional Genomics Information Management System

GC : Guanine Cytosine

GCS : Guanine Cytosine Skew

GRV : Global Relative Variance

HTGE : Horizontally Transferred Genomic Elements

HTML : Hyper Text Mark-up Language

HTTP : Hyper Text Transfer Protocol

KB : Kilobase

MB : Megabase

MSP : Maximal Scoring Pair

MUMmer : Multiple Unique MatchER

MVC : Model-View-Controller

nsSNP : non-synonymous Single Nucleotide Polymorphism

ORM : Object Relational Mapper

OU : Oligonucleotide Usage

PIP : Percentage Identity Plot

PS : Pattern Skew

RNA : Ribo Nucleic Acid

RPC : Remote Procedure Call

rRNA : ribosomal Ribo Nucleic Acid

RSCU : relative synonymous codon usage (RSCU)

RV : Relative Variance

SNP : Single Nucleotide Polymorphism

SQL : Structured Query Language Abbreviations

sSNP : synonymous Single Nucleotide Polymorphism

XML : eXtensible Mark-up Language

List of Figures

1.1 Large-scale synteny between *T. annulata* (TA) and *T. parva* (TP) chromosomes 9
1.2 Percent identity plots (PIP) for region immediately upstream of CFTR/Cftr exon 1 (nucleotides 5,425–19,425) . . . 11
1.3 Polymorphisms and genomic organization of ACHE . . . 13
1.4 Of the 5 nsSNPs (namely ACHE:c.169G4A; ACHE:c.1031A4G and ACHE:c.1057-C4A) were even able to be mapped directly onto the protein structure (Hasin *et al.*, 2004). . . . 14
1.5 Maximum likelihood phylogenetic tree depicting the relationships between the *T. pallidum* subspecies . . . 16
1.6 Values of RSCU and w for codons in very highly expressed genes from E. coli and yeast (Sharp *et al.*, 1987). . . . 19
1.7 GC contents of 1,294 *E. coli* genes. Gray bars denote native genes and black bars denote genes that are supposedly acquired by horizontal transfer (Lawrence *et al.*, 1997). . . . 20
1.8 Plot of CAI vs χ^2 of codon usage for 1,189 E. coli genes. . . . 21
1.9 Graph depicting the total counts of biased words. . . . 22
1.10 10 most over-represented and under-represented heptanucleotides found in the datasets. Ranked by decreasing z values therefore, the most biased words are found at the top of the list (Rocha *et al.*, 1998). . . . 23

2.1 Main base classes within FuGE. Newly developed classes developed within FunGIMS inherit from these classes (Pizarro *et al.*, 2006). . . . 28
2.2 Screenshot of Sybil's synteny gradient . . . 29
2.3 Figure illustrating the MVC design pattern in the context of Turbogears. Numbers represent the order of events subsequent to a user making a server request from the browser. . . . 36
2.4 UML class diagram showing some of the major classes used in the database and the relationships between them. . . . 38

2.5 An example of a typical view that a user sees. Everything visualized on the page is essentially HTML generated by KID. The 'ALIGN' button is the users way of communicating with the controller and in-turn, the underlying data. Javascript is responsible for user-input validation. 40
2.6 All functionality within the software suite is accessible via either the main-menu dropdown at the top of the screen or through the sub-menu at the botton of the screen. 42
2.7 Sample page showing the ClustalW alignment results. 43
2.8 Main BlastZ submission page for the alignment of whole genomes. 44
2.9 A successful BlastZ job submission will direct users to this page. Here, users may check the progress of their jobs by clicking the 'CHECK PROGRESS' button. . . 45
2.10 Main result page of a BlastZ submission. 46
2.11 Graphical display of alignment results using the Laj applet. 47
2.12 Neighbor-joining tree result page. 48

3.1 General view of the web-based SWGB 55
3.2 Identification of divergent genomic regions on the 'Gene Map' view 56
3.3 The 'Diagram' view of SWGB 58
3.4 Identification of divergent genomic regions by plotting and highlighting 59
3.5 Filtering genomic regions by multiple parameters. Click the 'Filter' button to open a dialog as shown in the figure. Setting up border values of multiple OU statistical parameters allows more precise localization of regions of interest. . . . 61
3.6 Command-line interface of the OligoWords program. 63
3.7 RV, GRV and D gene diagram plot for *M. tb* H37Rv. 64
3.8 RV, GRV and D dot-plot generated for Mycobacterium avium K10. 64
3.9 SWGB view for genomic region 87000-892000 (highlighted). An arrow marks nramp (in red) on the border of the highlighted region. 65
3.10 RV, PS and GC gene diagram plot for *M. tb* H37Rv. 66
3.11 Global evolutionary changes in mycobacterial genomes as revealed by SWGB dot plots. Each dot corresponds to the calculated oligonucleotide usage pattern for an 8kb sliding window of step size 2kb. 68
3.12 SNP distribution in homologous loci of *M. tb* H37Ra and *M. tb* H37Rv 69

4.1 Experimental results where growth was monitored in BALB/c mice of strain INH34 74
4.2 Early comparison of *M. tb* and the vaccine strain M. bovis BCG based on IS6110 sites. 75
4.3 Circular map of M. tb H37Rv chromosome 76
4.4 Overview of the genomic organization in the corresponding regions proximal to the origin of replication in BCG Pasteur and *M. tb* H37Rv, revealed by BAC mapping, PCR and hybridization experiments (Brosch *et al.*, 2000). 78

4.5 Overview of the GenoMycDB user interface. Note the available options for searching and displaying (Catanho *et al.*, 2006). 80
4.6 Schema of the mycobacterial comparison project database. 83
4.7 dnaB gene details for *M. tb* H37Rv and its homologues. 86
4.8 dnaK gene details for *M. tb* H37Rv and its homologues. 87
4.9 mmpL4_1 gene details for *M. avium ssp paratuberculosis* K10 and its respective homologues. 88
4.10 gyrB gene details of *M. avium ssp paratuberculosis* K10 and its homologues. . . . 89
4.11 Genome atlas of *M. tb* H37Rv. Note the abundance of repeat regions especially in regions 3.9 – 4.0 MB (13). 92
4.12 Genome atlas of *M. avium* K10 (13). 93
4.13 A Region of *M. tb* CDC1551 that appears to lack annotation information and B the corresponding region in *M. tb* H37Rv. 94

List of Tables

1.1 Comparison of BLASTZ alignment results to other contemporary programs (Scwartz *et al.*, 2003) 7

3.1 Sliding window size and OU pattern types (oligomer lengths) selected for sequences of different length present in the SeqWord database. 53
3.2 Coordinates and annotations of the gene islands in the genome of *M. avium* K10. 65
3.3 Coordinates and annotations of the gene islands in the genome of *M. tb* H37Rv. 67

4.1 Table showing general order of events and options available to users when in the mycobacterial comparison project. 84
4.2 Coordinates and annotation of the genes islands in the genome of *M. avium* K10. 85
4.3 Summary of absence/presence of dna genes of *M. tb* H37Rv in *M. avium* K10. . 87
4.4 Annotations for the outlined genomic fragments for the *M. tb* H37Rv plot (Figure 3.10). 90
4.5 Summarised table showing cross-species comparison of loci 333437-3950263 of *M. tb* H37Rv. 91

Chapter 1

Introduction

1.1 What is Comparative Genomics?

Comparative genomics is a relatively new field in biological research where genome sequences or genomic fragments are used (directly or indirectly) to compare various organisms. This type of comparison allows scientists to study many aspects of an organisms biology including discovery of new genes and protein structure, evolution within and between species and many more (Cole, 1998; (1)). For example, understanding of our own genome has substantially increased after examining genetic feature counterparts in other organisms such as the mouse (Ureta-Vidal *et al.*, 2003). When performing comparative genomics studies, researchers compare many different features contained within genomes such as genes, introns, conserved regions, repeat regions, re-arrangements and single nucleotide polymorphisms (SNPs) to make inferences about the evolution, physiology, pathogenicity (Mulder *et al.*, 2008) and genetic structure (Badger & Olsen, 1999) of the organisms being studied. The fact that the genomes of all organisms are comprised of the same building blocks i.e. DNA, means that one could essentially compare the genomes of highly similar organisms (for population genomics) as well as phenotypically diverse organisms for example, mouse and human, anenomes and whales, grasses and trees etc. Comparative genomics is indeed a useful and insightful area of study producing many new biological insights and scientific breakthroughs. An overview of modern comparative genomics techniques will be presented in this chapter and further on in chapters 3 and 4, several innovative approaches developed in this project will also be presented.

1.2 Sequencing technologies and the need for comparative genomics

The first major breakthrough in DNA sequencing methodologies came about in the late 1970s with the introduction of the Sanger method which uses dideoxynucleotides in the sequencing process to sequence a few kilobases (KB) of DNA at a time (Sanger *et al.*, 1977). Since then, better, cheaper and much more efficient methods of sequencing have come about, so much so that, the sequencing and assembling of whole genomes, millions of base pairs in length, has become a reality (Moxon *et al.*, 2002; Franguel *et al.*, 1999). At the time of writing this work, several high-throughput sequencing technologies were available including ROCHE's GS FLX 'pyrosequencer', ILLUMINA's Genome Analyser 'sequence by synthesis' sequencer and APPLIED BIOSYSTEM's SOLiD 'sequencing by oligo ligation and detection' sequencer (Mardis, 2008). Each technology is extremely capable of sequencing single, if not, multiple genomes in a single run. Due to the ease with which biological DNA sequences and genomes can be produced, biological sequence databases have seen and continue to experience unprecedented exponential growth such as NCBI (2), GOLD (3), CAMERA (4) and a multitude of others. As a result, there is an ever increasing need to mine the wealth of information encoded within these sequences, and indeed, many great findings which have had medical, industrial and agricultural implications have been made possible. In an effort to highlight the significance of comparative genomics, a few major findings brought to light by comparative genomics will be presented. On a purely physical level, a simple comparison between a human and a mouse reveals no reasonable similarity between the two. However, Waterston *et al.*, (2002) in a highly international collaboration showed that there is more similarity between humans and mice than meets the eye. Waterston *et al.* compared the draft sequence of the mouse (Mus musculus) and human genome and made a few startling discoveries. The group found that over 90% of the human and mouse genomes can be grouped into corresponding regions that show conserved synteny, meaning that there are large portions of matching DNA segments in the mouse and human which exhibit the same genes and gene order (synteny). It was also found that over 40% of the human genome can be aligned to the mouse genome at the nucleotide level and even though transposable elements between mouse and humans have different activities, similar types of repeat regions have been found to accumulate in corresponding genomic regions in both species (Waterston *et al.*, 2000). These and many other findings presented illustrate the power of comparative genomics in our understanding of the relationships between these organisms. Availability of complete bacterial genomes significantly advanced our knowledge of bacterial virulence and general biology. Thus, in this work a comparative analysis was carried out between *M. tb* and other mycobacteria. Using the BLASTP and FASTA programs, it was found that among the 1439 genes present in both the above-mentioned species, 219 of these genes had no counterparts in any other organism (Marmiesse *et al.*, 2003). This interesting finding prompted further investigation which was carried out via macro-array experiments, to determine whether these 219 genes were indeed specific to the mycobacteria. It

was found that all but 9 of the 219 genes were present in all the mycobacterial species tested, which were *M. tb* H37Rv, *M. leprae*, *M. avium*, *M. marinum* and *M. smegmatis*. Some of the 'missing' genes, based on bioinformatics analyses, were found to code for proteins belonging to the ESAT-6 protein family. The ESAT-6 family of proteins was known to be highly immunogenic. This study highlighted the fact that, comparative genomics is an extremely useful tool in the discovery of 'core' genes within bacterial organisms. The discovery of these 'core' genes shared between *M. tb* H37Rv and *M. leprae* could prove vital in the identification and development of highly specific anti-mycobacterial drugs (Marmiesse *et al.*, 2003; Cole, 2002; Brosch *et al.*, 2001). In a similar comparative study, Garnier *et al.* (2003) also compared genome sequences of some mycobacteria namely *M. bovis*, *M. tb* and *M. leprae*. This group showed that *M. bovis* in fact, contained no unique genes within its genome related to the other mycobacteria. This thus led them to the conclusion that differential gene expression could be the reason for this pathogen's variation in host tropism (Garnier *et al.*, 2003). Comparative genomics again was useful in explaining even pathobiology of a pathogen. Modern sequencing technologies are highly publicized when it comes to the sequencing of bacterial genomes and vertebrate genomes however, sequencing technologies are not limited to nuclear DNA. Studies have been performed on mitochondrial genomes of angiosperms in order to augment current understanding in monocot and dicot lineages (Kubo & Newton, 2007). Even chloroplast (Chung *et al.*, 2006; Saski *et al.*, 2005), viral (Dolja *et al.*, 2006) and protozoan (Hall *et al.*, 2005) genomes were sequenced and proved very insightful subsequent to comparative genomics analysis.

1.3 Common Methods used in Comparative Genomics

Comparative genomics studies can be tackled in a variety of ways and by using various methods. Not only can pairs of genes or specific loci be used in comparison but nowadays, whole genomes can be used as a basis for species to species comparisons. As of late, bioinformaticists and software developers around the world have realized the significance and power of comparative genomics (Hartmans *et al.*, 2006) and have risen to the challenge by developing new tools and methods and improving old tools. This is a fortunate time for researchers as there is a vast collection of tools available on the internet which caters for most, if not all needs, of any one researcher. A few of the methods used in comparative genomics, and their associated tools will be dealt with next.

1.3.1 Sequence Alignment & BLAST

One of the most fundamental needs in any comparative genomics study is the ability to align various sequences to one another. Researchers may want to align for example, 2 genes, together because they want to search for homology between them and or discover differences and sim-

ilarities in order to draw meaningful conclusions. Also, a researcher may have sequenced an unknown gene and then wants to discover the function of this gene by doing a sequence similarity search against a database of known proteins with known functions. Hits to similar known sequences in the database would then help a researcher infer information about the sequence of interest. How is sequence alignment achieved? In what was perhaps one of the most famous bioinformatics research publications, Needleman and Wunsch (1970) publicized a dynamic programming algorithm approach to aligning sequences and also assessed the scores of these alignments by assigning scores to insertions, deletions and replacements in the alignment. This strategy proved extremely successful and has become a 'core' tool in the bioinformatics world with many improvements being made along the way (Waterman, 1984). This method however, is very computationally intensive and is only meant to align rather small sequences together. The question of aligning sequences to a database containing millions of entries is still a major problem. In a landmark publication released in 1990, Altschul *et al.* introduced the Basic Local Alignment Search Tool, better known as BLAST. This was a tool employing a novel approach to sequence alignments. The algorithm used for BLAST subscribed to a measure based on a set of well defined mutation scores and this measure was further optimized by an algorithm directly approximating results that would have been obtained by a dynamic programming algorithm (Altschul *et al.*, 1990). Although the algorithm used in BLAST is a lot less stringent than the dynamic programming approach, the major advantage is that it is orders of magnitude faster. Furthermore, the implementation of the algorithm is very versatile and can be applied to simple DNA and protein databases searches as well as gene identification and motif searches (Altschul *et al.*, 1990). The BLAST algorithm works intimately with the concept of the maximal segment pair measure or MSP measure. To understand the MSP measure one first needs to understand how BLAST scores sequence alignments. During an alignment of two DNA sequences, a score of +5 is awarded to an identical base match and -4 for mismatches (other scores may be used here). Note that BLAST generally uses the PAM-120 scoring matrix for scoring protein alignments). A sequence segment is a contiguous stretch of residues or nucleotides of any length and the similarity score for two aligned segments of the same length is the sum of the scores for each pair of aligned residues. Finally, an MSP is defined as the highest scoring pair of identical length segments chosen from two aligned sequences. Thus, given two sequences, a references and a query, BLAST attempts to create local alignments of the query onto the reference sequence, calculating MSP scores along the way. It is possible that several MSPs may be found, thus segments are defined to be locally maximal if the score of the alignment cannot be improved by either lengthening or shortening of the segments at that particular sequence location. BLAST aims to detect all these MSPs with scores above a given threshold. When scanning databases (often containing thousands and millions of sequences), it is not likely that many sequences will be found to be absolutely identical to a scientists query sequence. This necessitates the need for an MSP threshold, S. The MSP threshold will therefore incorporate sequence hits that are highly similar to the query, as well as those that are somewhat similar. Being able to detect these latter

sequences during database searches are important for the detection of sequences which may be biologically related to the query sequence at hand. BLAST is particularly rapid in its database searching because it minimizes the time spent on local alignments that have little chance of exceeding the threshold (S). This estimation is performed as follows. Firstly, allow a word pair to be a segment pair of fixed length w. BLASTs main strategy is to find only segment pairs that contain a word pair with a minimal score threshold of T. Scanning of a sequence allows one to quickly determine if it will contain a word that may align with the query sequencing yielding a score greater than or equal to T. Only these matches producing T satisfying scores are further extended to ascertain whether the containing segment pair may produce an alignment with a score greater than or equal to S. Using this T threshold efficiently allows a great speed-up of the algorithm, however, selection of a very small T score may yield many more undesirable hits and negatively influence algorithm performance (Altschul *et al.*, 1990). One of the most useful and informative scores when trying to assess one BLAST results is the e-value. The e-value is essentially the probability due to chance that there is another alignment with an S score greater than the given alignments S score or in other words, it is a measure of the reliability of the given S score (5). The e-value is calculated by the following equation :

$$E = Kmne^{-\lambda S}$$

Where K and m are the natural scales for the search space and scoring system respectively; m, the query sequence size; n, database size and S, the score. A typical good e-value is one that is less than 10^{-5}. There are several drawbacks with the e-value that however, must be noted. When query sequences are too short, e-values tend to be more conservative. Statistical integrity breaks down with the introduction of gaps in the alignment, therefore gap scores are used instead here. Furthermore, e-values may spring false positives due to some sequences exhibiting low-complexity regions. Therefore, ideally, one should run blast on longer rather than short sequences. One of the most popular BLAST servers is found at the NCBI (5). BLAST capability is indeed important to comparative genomics because when researchers deal with new or unknown sequences, BLAST will allow access to the wealth of information contained within biological databases in order to identify homologous sequences, annotate unknown sequences, search for close relatives and thus give clues to the phylogeny of an unknown sequence. Context of the database can even help attribute functions to a researchers sequence by identifying identical sequences of known functions within a database (Jones *et al.*, 2005; Reiter *et al.*, 2001) and thanks to constant improvements to BLAST such as Gapped BLAST and PSI-BLAST (Altschul *et al.*, 1997; Cameron, 2007) databases searches are now more sensitive and faster.

1.3.2 Genome Alignment

In cases where researchers are fortunate enough to have at their disposal completely sequenced and assembled genomes, alignment of these whole genomes will prove very informative in terms of discovering coding regions, regulatory signals and general mechanisms of genome evolution

(Chain *et al.*, 2003). Early sequence alignment work by Needleman and Wunsch (1990), etc. showed that sequence alignment is possible and can be sufficiently accurate, however, those early algorithms were meant for the alignment of proteins and genes spanning a few kilobases at most and those methods are inefficient when dealing with large, whole genome sequences. New methods are therefore required. Delcher *et al.* (1999) set out to do multiple genome comparisons on two similar *M. tb* strains and two less similar strains of mycoplasma. Using a program MUMmer, to perform the alignments on *Mycobacterium tuberculosis* strains H37Rv and CDC1551, Delcher's group were able to map each and every base from one genome onto the other thus identifying all SNPs between the species. There were quite a few differences identified between the two strains with most being single base changes. There were also a few dozen insertions found uniquely on each genome, some of which contained whole or partial genes (Delcher *et al.*, 1999). Other groups also endeavored to perform whole genome alignments on mycobacteria and similar comparative analyses of whole genome sequences allowed the discovery of genetic differences between various virulent mycobacterial strains and it was possible to in fact, associate these genetic features to clinical features exhibited by the pathogen (Fleischman *et al.*, 2002). The same method was also applied to studies involving other bacterial species (Guyon & Guenoche, 2008). MUMmer, which employs Multiple Unique Matching (MUMs) as the method of alignment was also later implemented to align up to 90 closely related species on a simple desktop PC (Treangen & Messeguer, 2006). MUMs represent maximum exact matching strings appearing only once in each of the sequences compared (Aluru, 2006). Due to these MUMs being unique words in each of the sequences, the false positive rate is significantly reduced, the drawback, however, is the inability to detect anchors between divergent genomes. There are also other methods that were used in the alignment of bacterial strains such as the anchor based whole genome comparison methods (Vishnoi *et al.*, 2007). Findings such as those found by Delcher *et al.* (1999) are extremely important as these genotypic changes inevitably impact phenotypically and bring about change in the disease process of these organisms. Making correlations between genotypic changes and pathogenesis is needless to say, helpful in understanding and combating the disease. MUMmer, although useful in comparing bacterial sequences, is not well suited to larger sequences such eukaryote genomes which span hundreds of millions to billions of nucleotides. Furthermore, vertebrate genomes that are structurally complex, pose new sets of challenges sometimes not encountered with smaller genomes (Couronne *et al.*, 2003). These and other challenges were acknowledged quite early on and it was noted that the choice of tools and approaches used in eukaryote genome scale comparisons would greatly affect alignment accuracy and thus the deductions that can be drawn from such alignments (Chain *et al.*, 2003) BLASTZ (Schwartz *et al.*, 2003) which is a program following the same strategy used by Gapped BLAST (Altschul *et al.*, 1997), was found to be effective in not only aligning bacterial sequences but also mammalian sequences of significantly large length. The algorithm was tested by aligning human and mouse sequences. BLASTZ was found to be more sensitive than contemporary programs such as PatternHunter and BLAT on several levels. BLASTZ successfully aligned over 96% of

human chromosome 20 onto mouse chromosome 2 which was pleasing given the known high level of homology between the two chromosomes. BLASTZ was also very sensitive in finding other homologous features between the species such as 3' UTR, 5' UTR and upstream regions (Table 1.1).

Table 1.1: Comparison of BLASTZ alignment results to other contemporary programs (Scwartz *et al.*, 2003)

Score	1 Mus	>1 Mus	1 Rev	>1 Rev
3000	0.36814	0.02340	0.00084	0.00080
4000	0.36859	0.02230	0.00040	0.00074
5000	0.36958	0.01975	0.00016	0.00059
6000	0.36992	0.01829	0.00013	0.00051
7000	0.36997	0.01697	0.00011	0.00043
8000	0.36966	0.01586	0.00010	0.00037
9000	0.36911	0.01490	0.00008	0.00033
10000	0.36831	0.01405	0.00007	0.00030

The columns have the following meanings: (1) score threshold for a gapped outer alignment (Step 2.2.2 of Fig. 1); (2) fraction of the genome covered by exactly one alignment; (3) fraction of the genome covered by more than one alignment; (4) fraction of the genome covered by exactly one alignment with reversed mouse; (5) fraction of the genome covered by more than one alignment with reversed mouse.

The algorithm is available at Pipmaker (http://bio.cse.psu.edu). Pipmaker is based on the BLASTZ algorithm but the ensuing alignment results may be viewed as Percentage Identity Plots (PIPs) which is a very informative graphical view that highlights conserved segments within alignments (Schwartz *et al.*, 2000) Genome rearrangements are one of the key evolutionary processes that shape genomes of subsequent generations. Recombination events are often responsible for large portions of genomes being shifted around and even duplicated. These types of changes make genome alignments all the more challenging. However, programs such as Mauve (Darling *et al.*, 2004), account for such changes and are even able to detect and align horizontally acquired elements. Their algorithm involves the following basic steps :

1. Identify local alignments (Multi-MUMs)
2. Calculation of phylogenetic tree guide using the multi-MUMs
3. Selection of a multi-MUM subset for use as anchors. The anchors are partitioned into local collinear blocks (LCBs).
4. Recursively perform anchoring to identify additional alignment anchors within and outside of LCBs

5. Perform progressive alignment of each LCB using the tree guide.

This strategy allows for the detection of large-scale evolutionary introduced rearrangements present in many genomes. Whole genome alignments as already seen, are useful in comparing genomes and acquiring large scale feature dichotomies between organisms such as large scale inversions, insertions and overall sequence identity. Another useful method in comparative genomics studies and somewhat related to genome alignment, is to analyze synteny observed between organisms. This will be dealt with next.

1.3.3 Synteny

Synteny which literally translates as "same thread" is basically a set of genes or features that share the same relative ordering between chromosomes of different species or between chromosomes within a specie (Pan *et al.*, 2005). A syntenic analysis between species would therefore help in the identification of homologous genes between organisms, aid in the understanding of evolution between organisms (See *et al.*, 2006) and within species chromosome evolution as well as highlight the presence of regulatory elements between genomes. A few studies highlighting the importance of synteny as a means of sequence analysis will now be presented. *Theilleria annulata* (TA) and *Theilleria parva* (TP) are intracellular eukaryotic, tick-borne hemoparasites that cause lymphoproliferative diseases in cattle such as tropical theileriosis (caused by TA) and East Coast Fever (ECF) caused by TP (Pain *et al.*, 2005) The two parasites share similarities in their life cycles involving intracellular stages in leucocytes and red blood cells, however, they are transmitted by different tick species and even transform different cell types when in the cattle's blood. The full genomes of these two protozoans were compared in order to understand the mechanisms for the differing tropisms and cell type transformation. Many new as well as established findings came to the front after their comparative analyses. As expected, the two species possessed tandem arrays of hypervariable genes families (which mapped adjacent to the telomeres) with an arrangement that was highly conserved (Figure 1.1).

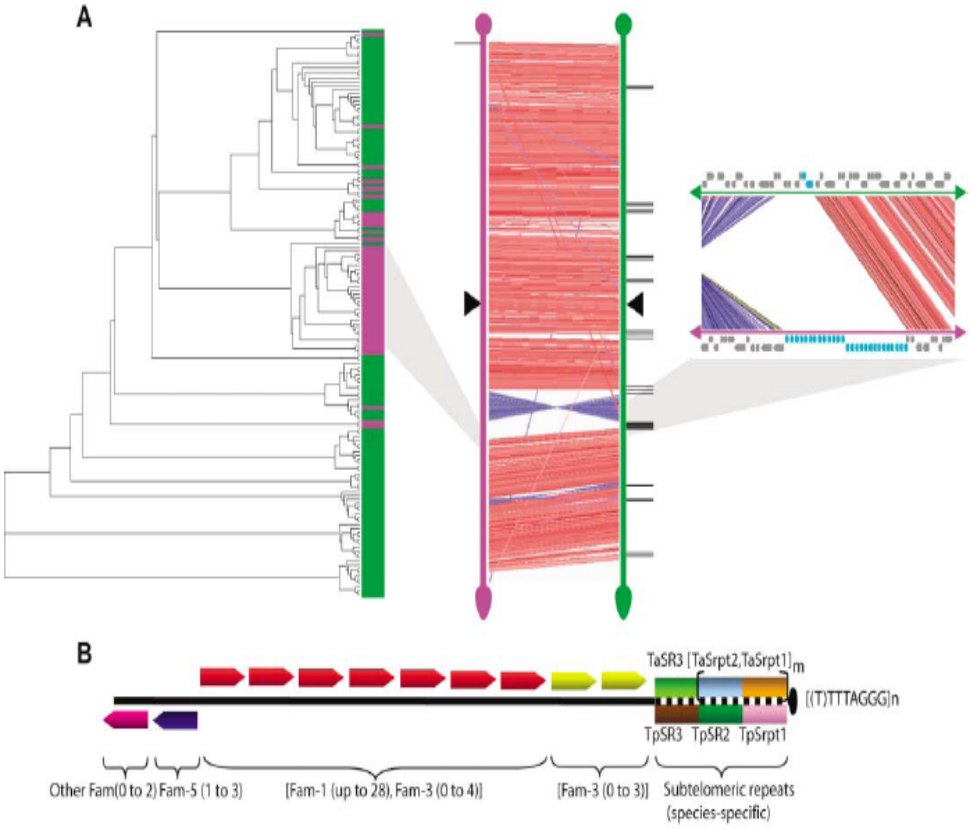


Figure 1.1: Large-scale synteny between *T. annulata* (TA) and *T. parva* (TP) chromosomes. (A) Synteny breaks of chromosome 3 of TA (green) and TP (purple) are located at Tpr genes. (Middle) Chromosome 3 of TA and chromosome 3 of TP are aligned. Connecting lines show maximal unique matches between the two chromosomes. Red lines, alignments in the same orientation; blue lines, alignments in opposing orientations; black triangles, putative centromeres; black lines, Tpr genes occurring outside the Tpr locus. The position of the Tpr locus of TP is aligned with the gray shaded area. (Left) The phylogenetic tree shows the clustering of the TP genes when compared with the TA genes. Branches ending in green boxes represent TA genes and purple boxes represent TP genes. All genes in the Tpr locus occur in the cluster which is aligned with the gray shaded area. (Right) A close-up of the insertion of the Tpr locus in TP (purple) with respect to TA (green), with Tpr and Tar genes (blue) and all other genes (gray). (B) Organization of a representative subtelomere (not to scale). The black line represents the coding part of the subtelomere, with the arrangement of gene families (arrowheads) shared between TA and TP. The arrowheads indicate the transcriptional orientation; the observed range in numbers of genes is in parentheses. The dotted black line represents the species-specific noncoding regions (upper, TA; lower, TP). Srpts, subtelomeric repeats; SR, subtelomeric regions (4) (Pain *et al.*, 2005).

Proximal to telomeres in both species the genes (described as being related to the SfiI restriction enzyme fragment) designated, family 3) were found. This gene family is then followed by the appearance of Pro/Gln-rich proteins. The designated boundary between sub-telomeric gene

families and "house-keeping genes" is occupied by the adenosine 5'-triphosphate-binding cassette (ABC) transporter genes (designated, family 5). Members of these above-mentioned families 3 and 5 also occur internally within the genomes. This was an interesting find as internal clusters of these gene families act as reservoirs while their sub-telomeric counterparts actively exchange genetic material, which is a mechanism for expansion and antigen diversification (Pain *et al.*, 2005). The finger millet is an allotetraploid grass that belongs to the Chloridoideae subfamily. It is cultivated mainly in East Africa and Southern India where it makes a significant contribution to the countrys' food stockpiles due to its high nutritional quality and desirable storage quality. However, genetic improvement of this crop is lagging behind relative to its counterparts, for example, rice thus yield for the finger millet is far below optimal. In an attempt to understand the genetics of this organism in order to improve its agricultural yield and elucidate its evolutionary history, a comparative analysis was done against rice. Of the nine finger millet homologous groups identified, 6 corresponded to a single rice chromosome each, with the remaining three being orthologous to two rice chromosomes. In the remaining three cases, one rice chromosome was found inserted into another rice chromosome giving rise to the finger millet chromosomal configuration (Srinivasachary, 2007). All in all, there was quite a large degree of synteny observed between the rice and the finger millet with only 10% of markers employed not finding corresponding syntenic locations in either of the chromosomes. A host of other interesting synteny studies have been published such as Kubo's group that looked at angiosperm mitochondrial genome organization (Kubo & Newton, 2007), Lyon's group that established methods to perform plant genes and chromosomal comparisons (Lyons & Freeling, 2008) and many others (Saski *et al.*, 2005; Waterston *et al.*, 2002; Pain *et al.*, 2005). Indeed, studies such as these will continue to pervade the literature as more and more complete genomes become publicly available. However, before moving on, it must be noted that in studies such as synteny, visualization methods play very important roles in the success of a study. Many examples exist of such synteny visualization tools however for conciseness a few major tools will be mentioned. One of the most established synteny tools available is SynBrowse (Pan *et al.*, 2005). SynBrowse is a highly customizable, web-based synteny browser built on Gbrowse (Stein *et al.*, 2002). SynBrowse, which works off a relational database of pre-calculated data, allows users to view macro-, microsynteny, homologous regions, identify uncharacterized genes, regulatory elements and a host of other features. Synbrowse is freely available at (8). OrthoCluster (Zeng *et al.*, 2008) is also a powerful synteny browsing tool with built-in algorithms able to handle more advanced tasks in synteny such as gene strandedness, gene-inversions, gene duplications and the ability to allow several genome comparisons simultaneously. BlastAtlas (Hallin *et al.*, 2008), also built for viewing cross genome homology, can also handle metagenomic information. SyMAP (Soderlund *et al.*, 2006), also recently published, offers users an entirely new algorithmic approach to studying synteny by employing FPC-based physical maps.

1.3.4 Gene-by-gene comparative genomics

Thus far, the focus was mainly on large-scale features. Yet another approach within the scope of comparative genomics is to do simple gene-by-gene or region by region analyses among species. Gene by gene or region by region analyses is a very narrow and detailed analysis but it is useful in that sometimes researchers know exactly their gene or region of interest and thus want to only focus their energies on the differences of these genes or regions amongst the various species (Matolweni *et al.*, 2006). A few studies illustrating this point follow. Cystic Fibrosis (CF) is a fatal genetic linked (autosomal recessive) disorder. It causes the body to produce a thick, sticky mucus that clogs the lungs thus leading to infection. Furthermore, the mucus that is produced blocks the pancreas, precluding digestive enzymes from reaching the intestines which in turn affects food digestion. The mutated gene, directly linked to CF is the cystic fibrosis transmembrane conductance regulator (CFTR) gene (Li & Godzik, 1996). CFTR codes for a cyclic AMP-activated chloride channel crucial to salt and water transport in epithelial cells. In order to gain more insight into the spatial and temporal regulation of CFTR expression, homologous CFTR-containing regions in mouse and humans were sequenced and compared (Ellsworth *et al.*, 2000). After detailed comparative analyses, the group showed that human and mouse CFTR-containing segments are highly conserved. Furthermore, it was noticed that there were rather large conserved segments even within introns, revealed by percentage identity plots (Figure 1.3). Similarly, genetic segments containing the WNT2 (human) and Wnt2 (mouse) genes exhibit several conspicuously conserved sequences within introns flanking exon3.

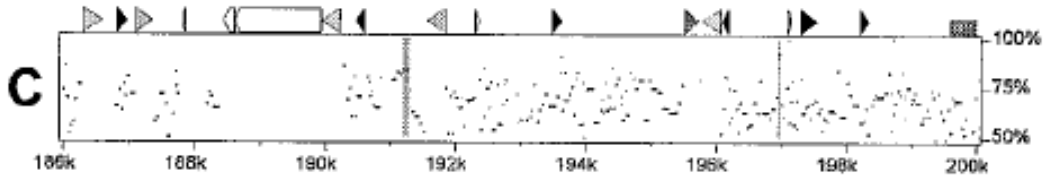


Figure 1.2: Percent identity plots (PIP) for region immediately upstream of CFTR/Cftr exon 1 (nucleotides 5,425–19,425). The vertical stripes are used to highlight the gap-free regions in a '28-kb interval encompassing CFTR/Cftr exon 1 that have a higher percent identity than other gap-free regions in that interval of the same or larger length. Features in the PIP: tall black rectangle, exon; white pointed box,L1-type repeat; dark gray pointed box, LTR repeat; black triangle, MIR-type repeat; light gray triangles, other SINE-type repeat; dark gray triangles, all other interspersed repeats; short white rectangle, CpG island where $0.6 <$ CpG/GpC $<= 0.75$; short dark gray rectangle, CpG island where CpG/GpC $>= 0.75$ (Ellsworth *et al.*, 2000).

These results strongly suggest that expression regulation of CFTR/cftr (Human/mouse) may indeed be orchestrated by the supposed non-coding regions, mechanisms by which this is done is still sadly, not well understood. In another study by Mallon *et al.* (2000), detailed comparative analysis of the Bpa/Str (X-linked disorder) gene regions were undertaken between mouse and human. Combining gene prediction tools and database searching, the group was able to find 11

genes in mouse and 13 in the human counterpart. Comparing the regions by pairwise alignments enabled the identification of a further four putative conserved genes. Prior non-sequence analyses of these regions led researchers to be believe that there were no substantial difference in these regions across humans and mice for instance. However, this sequence analyses has further shown that there is a considerable amount of rearrangement between the two species. Other features elucidated by this study was the unexpectedly high LINE and gene content but low SINE and G+C content which is unusual for regions such as these (Mallon *et al.*, 2000). Although the focus here was mainly on gene comparisons at the DNA level, protein level comparisons are also extremely informative (Muller *et al.*, 2005) but protein-protein comparison is an entirely different field and a detailed treatment is outside the scope of this work. Attention will now be turned to the concept of single nucleotide polymorphism, its definition and relation to comparative genomics.

1.3.5 Single Nucleotide Polymorphism analyses in comparative genomics

Single nucleotide polymorphisms or SNPs describe a type of genetic variation and has become a very popular approach in comparative genomics. For a nucleotide position to qualify as a SNP, there must exist at least two variants at that position and the least occurring variant must occur at a frequency greater than 1% (Ahmadian *et al.*, 2000). SNPs occur in the human genome for instance, at a frequency of 1 per 1000 base pairs and are thus a major source of genetic variation. Due to the invariable occurrence of SNPs within and between species, they can be used in the study of disease gene identification (White *et al.*, 2001), drug resistance (Nouvel *et al.*, 2006), phylogenetic analyses (Alland *et al.*, 2003), genotyping (Gutacker *et al.*, 2002; Filliol *et al.*, 2006), general evolution (Brosch *et al.*, 2001) and many more. Due to the high level of SNP detection activity around the world, there are initiatives to concertedly collect and collate all SNP data being generated into well organized and maintained databases (White *et al.*, 2001). This approach would hopefully streamline and speed up the rate at which SNP-related research is being performed. A few example SNP related studies will now be covered and some information regarding international SNP databases will follow. In excess of a hundred mycobacterial strains exist, a small proportion of which has been sequenced. Some are pathogenic and others not. Understanding the relationships between these virulent and non-virulent strains has clinical significance. Also, being able to trace the lineages of virulent to non-virulent phenotypes will afford a much needed understanding into the mechanisms of the organism's pathogenesis. After establishing a set of 148 synonymous SNPs (sSNPs) by comparing *M. tb* strains H37Rv and CDC1551, Gutacker and group used these sSNPs as a basis to genotype a 112-member 'core group' of mycobacterial isolates that represent the full diversity observed within the *M. tb* family. Gutacker *et al.*, based on their sSNP data managed to differentiate between the 112 isolates as well as categorize them into 8 major genotypically related clusters. Another impressive ability of the sSNP genotypic approach is the following: although the *M. tb* complex, comprising of

M. tb, *M. microti*, *M. bovis*, *M. africanum* and *M. canettii* are extremely closely related, sSNP genotyping was able to resolve the relationships between all these 5 species (Gutacker *et al.*, 2002) despite the failure of many other techniques in the past to do so. The presence of SNPs in genes and other genetic loci, depending on whether or not the SNPs are synonymous or not, have implications in disease processes. The following studies underscore this. Acetylcholinesterase (AChE), which catalyses the hydrolysis of acetylcholine (ACh) is responsible for the termination of impulse transmission at cholinergic synapses (Hasin *et al.*, 2004) as well as other biological functions. AChE is a highly conserved gene that shares over 88% identity with the mouse homolog. Due to SNP studies in the past, AChE has been implicated in Alzheimer's disease, Gulf War Syndrome and pesticide hypersensitivity. Due to the lack of SNP data for this gene, which was largely due to technology limitations and sample sizes, Hasin's group (2004) set out to increase the knowledge pool regarding this gene by analyzing the ACHE gene from 96 unrelated individuals representing 3 different ethnic groups. Their analyses revealed 13 SNPs in total (Figure 1.3), 10 of which were previously unknown. 5 of the 13 SNPs were nsSNPs thus resulting in downstream protein structural changes.

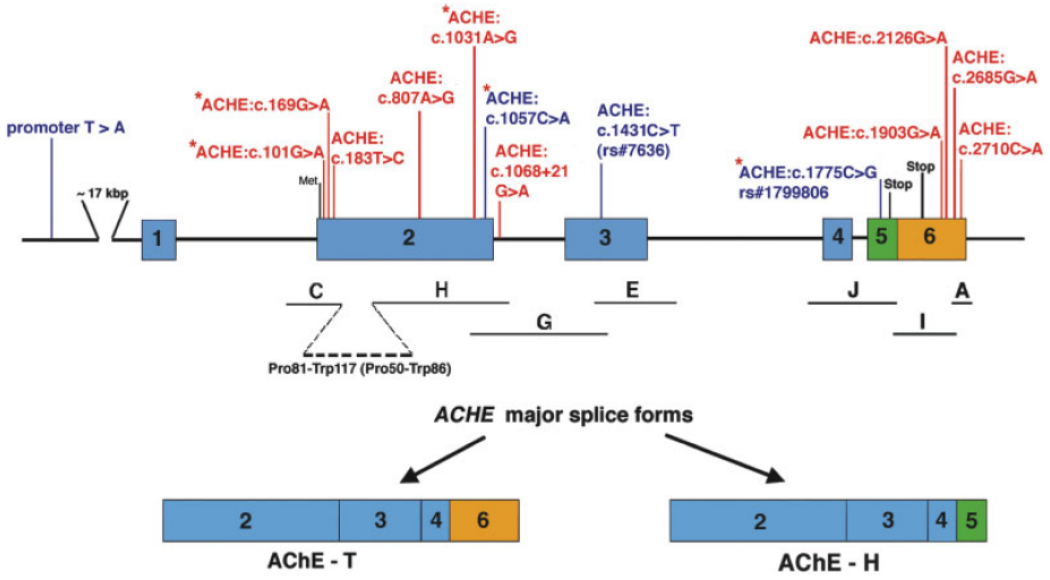


Figure 1.3: Polymorphisms and genomic organization of ACHE. Human AChE is encoded by a single ACHE gene composed of six exons which generates two major alternatively spliced forms that differ in quaternary structure and tissue distribution. Exons are depicted as boxes, and labeled from1 to 6. Previously reported SNPs are in blue, and novel SNPs in red. Nonsynonymous SNPs are marked with *. SNPs are identified based on their position in the cDNA sequence (Hasin *et al.*, 2004).

Hasin *et al.* (2004) successfully managed to highlight that SNPs in ACHE negatively affect the resulting protein structure and inevitable hamper the functioning of this protein. It should be noted here that, even a little change, such as a SNP, at the gene level has far reaching

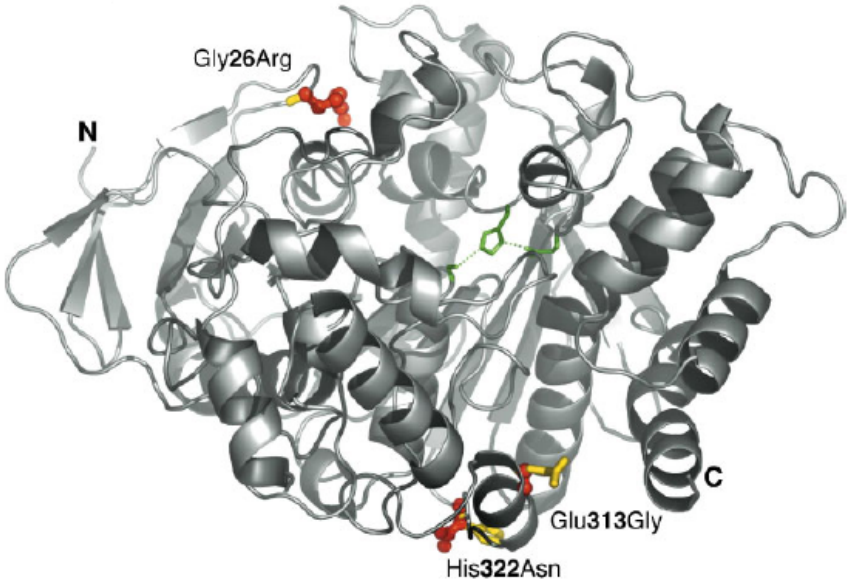


Figure 1.4: Of the 5 nsSNPs (namely ACHE:c.169G4A; ACHE:c.1031A4G and ACHE:c.1057-C4A) were even able to be mapped directly onto the protein structure (Hasin *et al.*, 2004).

consequences for an organism leading to disease states, high dysfunction and may even increase risk of acquiring a disease (Ozaki *et al.*, 2006). This study therefore, highlights the importance of SNP analyses and its importance on the clinical level. Needless to say, SNP analyses may also play an invaluable role in the agricultural and biotechnology industry. On a technical note, there are various ways to detect SNPs between sequences and many groups have dealt with this problem (Galves *et al.*, 2006; Tang *et al.*, 2006; Ahmadien *et al.*, 2000). The basic aim in any SNP analysis is to align sequences together and discover, single base changes at the same relative position in all the aligned sequences. This is mostly accomplished by sequence alignment (see earlier). There are several programs available on the commercial and open source market which offer users SNP discovery functionality, examples include MUMmer, CLCBio (16) and SNPdetector (Zangh *et al.*, 2005).

1.3.6 Phylogenetic Analyses

Phylogenetics is an approach with the objective of classifying entities such as biological organisms or molecules by virtue of the way they have evolved. Phylogenetic analysis could have the power to infer relationships between the various organisms within genera, elucidate genealogies of cultural groups, find common ancestors for certain related species, trace gene evolution within and between species and perhaps even trace human ancestry. The basis for grouping organisms or molecules into specific groups is dependent on similarities shared on a number of levels including biochemical, morphological, amino acid and DNA. Clusters of organisms or molecules

are presented in the form of phylogenetic or evolutionary trees. For the purpose of this work, focus will be on phylogenetic analyses based on DNA level comparisons between organisms or molecules. Syphilis, a sexually transmitted disease caused by the bacterium *Treponema pallidum sub-species pallidum*, was rumored to be brought to Europe by Christopher Columbus and his crew from the new world around the year 1495. In the twentieth century however, doubts began to arise as to the validity of the 'Columbian hypothesis' with some claiming that syphilis had already been present in Europe before but was merely indistinguishable (Harper *et al.*, 2008). To put an end to this debate, Harper *et al.* (2008) applied a phylogenetics approach to answering the question concerning the origins of the disease. Using 26 geographically disparate strains of the Treponema pathogen, 21 different genetic regions were examined in each of the strains. Phylogenetic trees incorporating all the variation present among the 26 strains were created by concatenating SNPs and indels into a single sequence in the same order they appeared in the genome (Figure 1.5). ClustalX was used to perform the alignment, Modeltest to choose the appropriate nucleotide substitution model and PAUP was used to subsequently construct maximum-likelihood and maximum-parsimony phylogenetic trees.

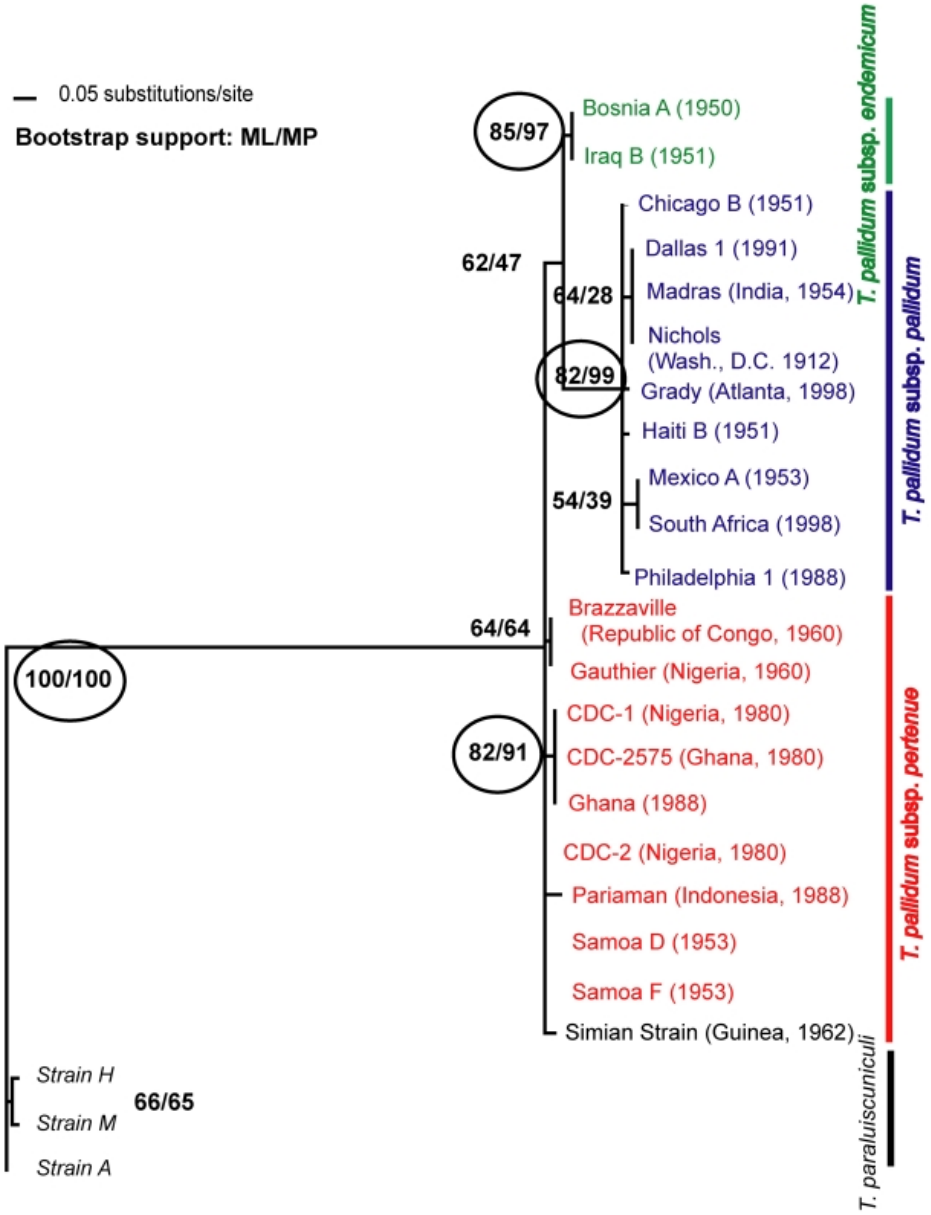


Figure 1.5: Maximum likelihood phylogenetic tree depicting the relationships between the *T. pallidum* subspecies. This tree is based on 20 polymorphic regions in the *T. pallidum* genome. Bootstrap support was estimated with 1,000 replicates in order to assess confidence at branching points and are shown within circles where values are high (.90%). Bootstrap support values for both maximum likelihood and maximum parsimony trees are shown, in that order (Harper *et al.*, 2008).

Based on the phylogenetic trees and geographical data, researchers were able to show that the 'Columbian hypothesis' was indeed plausible, however, treponemal diseases were very old

and travelled with humans along their migratory paths long before Columbus, though not as venereal diseases. Sub-species Pallidum strains (aka venereal causing pathogens) however, arose quite recently (Figure 1.5) as suggested by the phylogenetic tree and were introduced back into Europe as a venereal pathogen as *T. pallidum* most closely resembled disease causing strains from the south Americas than the non-venereal strains (Harper *et al.*, 2008). Due to the depth of the data, the researchers were also able to suggest quite a detailed model of *T. pallidum*'s evolution and dissemination throughout the world. The power of phylogenetics was also used to estimate the period when the Americas were first colonized by people and the haplogroups that were present at the time. They accomplished this using mitochondrial DNA (Achilli *et al.*, 2008). Phylogeny also has the power to determine to lineage of pathogenic bacteria from the non-pathogenic roots (Marmiesse *et al.*, 2004) and there are many other applications. Many programs exist to perform the wide range of phylogenetics tasks. Some of these programs include, Phylip (9), PAUP (10), MrBayes (Huelsenback & Ronquist, 2001), MEGA (Kumar *et al.*, 2008), CLCBio (11) and Clustalw (Thompson *et al.*, 1994). The choice of program largely depends on the desired outcomes of the user as every program though versatile, may not necessarily be able to perform all tasks required by a researcher.

1.3.7 Regulatory Motif Discovery

Regulatory motif (RM) analysis is yet another method by which genomes can be compared and analysed. Regulatory motifs are essentially, short DNA sequences involved in the control of gene expression. Regulatory motifs can dictate the conditions whereby genes are activated or in-activated. Transcription factors bind to specific promoter regions of target genes in a sequence-specific manner, but this binding may still happen even with slight sequence variation in the target site. Therefore, these binding sites, though specific, still exhibit slight sequence variation. Thus, a defined regulatory motif may contain slight sequence variation while still maintaining its specificity to transcription factors. Regulatory motif experimental determination is neither practical nor efficient for many biological systems (Conlon *et al.*, 2003) and they are also very difficult to detect directly by computational methods due to several reasons. They are often quite short, ranging from 6 to 15 nucleotides. They are also quite degenerate and occur at varying distances upstream of their target genes. Typically, when searching for RMs and when their patterns are expected to be found at a higher frequency relative to other sequence patterns of the same length, algorithms such as Expectation Maximization (EM) and Gibbs sampling may be used. Software such as MEME, AlignACE and BioProspector implement these algorithms (Kellis *et al.*, 2004). Discovery of regulatory motifs within and between genomes are interesting in that it allows one to examine the amount of gene regulatory mechanisms shared among organisms and detection of co-regulated genes. Also, conservation of RMs across different species can allow functional categorization of RMs and give insights into the physiology of organisms (Kamvysselis *et al.*, 2003). Thus far comparative genomics was performed directly on the sequence level. However, there are techniques that use metagenome information about a

genome and its genes to draw comparative genomics deductions. These methods will be disussed in the following sections.

1.4 A Novel Comparative Genomic Technique using Oligonucleotide usage pattern profiling

Sequence-centric comparative genomics analysis has gained much popularity over the years and proven its usefulness. However, performing comparative genomics on a more abstract level is also possible. For instance, one can make DNA comparisons on, not only the sequences themselves, but their physico-chemical properties, codon bias and the usage of oligo-nucleotide words, to name but a few approaches. This abstract level approach is more new to the field of comparative genomics, nevertheless producing many great insights involving topics such as bacterial mutation rates, DNA elemental transfer between bacteria, genome signature detection and a host of others. These will be dealt with next.

1.4.1 Codon Usage Bias

Due to the degeneracy of the genetic code, amino acids are encoded by several synonymous codons (Bulmer, 1998) and it has been demonstrated that these synonymous codons are not all used at the same frequencies. In an interesting study conducted by Sharp *et al.* (1987), it was shown that the use of codons is certainly non-random. In this study, a representative set of highly expressed genes from yeast and *E. coli* were chosen and their relative synonymous codon usage (RSCU) and w scores were compared. An RSCU score for a codon is basically the observed frequency of use of that particular codon's usage divided by the expected frequency of use under the assumption of 'equal usage of synonymous codons' for an amino acid. w on the other hand is the actual frequency of use for a codon compared to the frequency of use of the optimal codon for that amino acid.

The table in effect, illustrates how different codons are used in preference over others in specific genes. These RSCU values show that it is perhaps possible to use these values as indicators of highly expressed genes or as predictors of gene expression within an organism (Sharp & Li, 1987). Factors contributing to this preferred use of codons include translational selection, GC composition, RNA stability and others (Ermolaeva, 2001; Kiewitz, 2000). The codon adaptation index (CAI), a term introduced by Sharp *et al.* in the same study, was a numerical value representing the synonymous codon bias of a gene and was essentially a geometric mean of the RSCU values. Another very important reason for comparing codon usage biases across genomes is that it can give insight into the mutational processes and evolution of bacteria by tracking the transfer of genetic elements such as horizontal or laterally transferred elements between bacteria. At the time that genes or sub-genomic DNA are newly introduced into a bacterial cell, that 'new' DNA exhibits codon usage bias that is typical of its donor genome. However,

		E.coli RSCU	E.coli w	Yeast RSCU	Yeast w			E.coli RSCU	E.coli w	Yeast RSCU	Yeast w
Phe	UUU	0.456	0.296	0.203	0.113	Ser	UCU	2.571	1.000	3.359	1.000
	UUC	1.544	1.000	1.797	1.000		UCC	1.912	0.744	2.327	0.693
Leu	UUA	0.106	0.020	0.601	0.117		UCA	0.198	0.077	0.122	0.036
	UUG	0.106	0.020	5.141	1.000		UCG	0.044	0.017	0.017	0.005
Leu	CUU	0.225	0.042	0.029	0.006	Pro	CCU	0.231	0.070	0.179	0.047
	CUC	0.198	0.037	0.014	0.003		CCC	0.038	0.012	0.036	0.009
	CUA	0.040	0.007	0.200	0.039		CCA	0.442	0.135	3.776	1.000
	CUG	5.326	1.000	0.014	0.003		CCG	3.288	1.000	0.009	0.002
Ile	AUU	0.466	0.185	1.352	0.823	Thr	ACU	1.804	0.965	1.899	0.921
	AUC	2.525	1.000	1.643	1.000		ACC	1.870	1.000	2.063	1.000
	AUA	0.008	0.003	0.005	0.003		ACA	0.141	0.076	0.025	0.012
Met	AUG	1.000	1.000	1.000	1.000		ACG	0.185	0.099	0.013	0.006
Val	GUU	2.244	1.000	2.161	1.000	Ala	GCU	1.877	1.000	3.005	1.000
	GUC	0.148	0.066	1.796	0.831		GCC	0.228	0.122	0.948	0.316
	GUA	1.111	0.495	0.004	0.002		GCA	1.099	0.586	0.044	0.015
	GUG	0.496	0.221	0.039	0.018		GCG	0.796	0.424	0.004	0.001
Tyr	UAU	0.386	0.239	0.132	0.071	Cys	UGU	0.667	0.500	1.857	1.000
	UAC	1.614	1.000	1.868	1.000		UGC	1.333	1.000	0.143	0.077
ter	UAA	--	--	--	--	ter	UGA	--	--	--	--
ter	UAG	--	--	--	--	Trp	UGG	1.000	1.000	1.000	1.000
His	CAU	0.451	0.291	0.394	0.245	Arg	CGU	4.380	1.000	0.718	0.137
	CAC	1.549	1.000	1.606	1.000		CGC	1.561	0.356	0.008	0.002
Gln	CAA	0.220	0.124	1.987	1.000		CGA	0.017	0.004	0.008	0.002
	CAG	1.780	1.000	0.013	0.007		CGG	0.017	0.004	0.008	0.002
Asn	AAU	0.097	0.051	0.100	0.053	Ser	AGU	0.220	0.085	0.070	0.021
	AAC	1.903	1.000	1.900	1.000		AGC	1.055	0.410	0.105	0.031
Lys	AAA	1.596	1.000	0.237	0.135	Arg	AGA	0.017	0.004	5.241	1.000
	AAG	0.404	0.253	1.763	1.000		AGG	0.008	0.002	0.017	0.003
Asp	GAU	0.605	0.434	0.713	0.554	Gly	GGU	2.283	1.000	3.898	1.000
	GAC	1.395	1.000	1.287	1.000		GGC	1.652	0.724	0.077	0.020
Glu	GAA	1.589	1.000	1.968	1.000		GGA	0.022	0.010	0.009	0.002
	GAG	0.411	0.259	0.032	0.016		GGG	0.043	0.019	0.017	0.004

Figure 1.6: Values of RSCU and w for codons in very highly expressed genes from E. coli and yeast (Sharp *et al.*, 1987).

over time, the 'new' DNA changes to match the codon usage of its host as it under the same mutational constraints as the host (Ermolaeva, 2001). This process is known as 'amelioration'. In a landmark study in 1997, Lawrence and Ochman showed that more than 600 KB of DNA in Escherichia coli comprised of horizontally transferred elements and subsequently proposed a model for amelioration. In the study, a 1.43 MB contig for E. coli was constructed from various genbank sequences and the protein coding regions and open reading frames within the sequence was identified by existing annotations. To determine which sequences within E. coli appeared through horizontal transfer, gene features from a set of *E. coli* and *Salmonella enteric* were examined and a set of criteria for identifying horizontally transferred genes was developed in the following way. Codon-position-specific GC content was determined for each gene. On average, GC for the first, second and third codon positions were 59%, 43% and 56% respectively.

Atypical sequences were initially identified if their first and third codon position GC content were either 10% lower or 8% higher than their respective means. GC content for the second codon position could not be used as they are normally very similar across species to be able

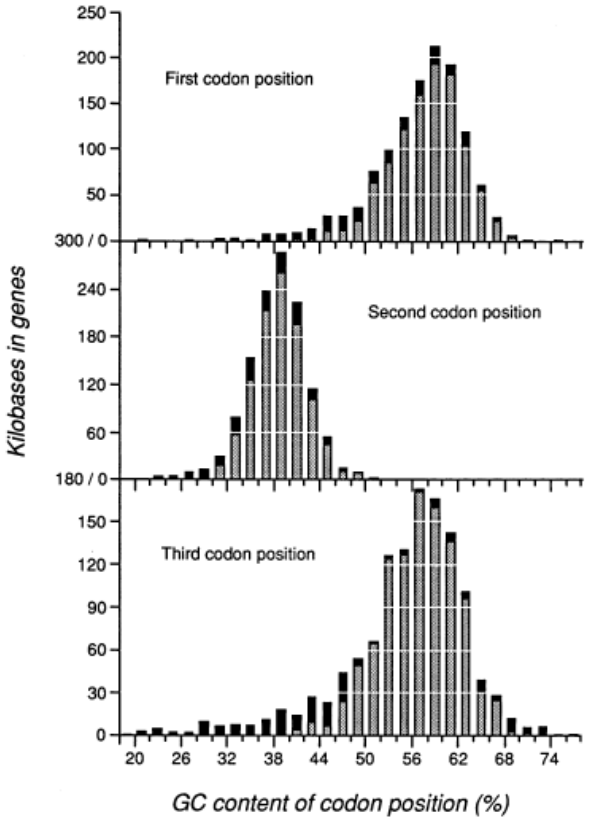


Figure 1.7: GC contents of 1,294 *E. coli* genes. Gray bars denote native genes and black bars denote genes that are supposedly acquired by horizontal transfer (Lawrence *et al.*, 1997).

to differentiate between different species. In addition to this GC content analyses, codon usage bias for the genes were also analyzed as GC content alone would not be sufficient to differentiate actual 'native genes' with atypical nucleotide composition (Koski *et al.*, 2001). To differentiate these 'native genes' from actual horizontally transferred genes, the CAI (in addition to χ^2) was used in order to determine if codon preferences were biased towards the codon sub-set employed by highly expressed genes. The logic followed that, if selection for preferred codons resulted in atypical GC content in a native E. coli gene, that gene would exhibit high χ^2 and CAI values. By employing this method, 200 protein coding regions were identified as being acquired by horizontal transfer. Also, about 29 genes were proposed to have been acquired by horizontal transfer due to their peculiar function and/or chromosomal location. Thus, a total of 229 genes were singled out as being present in this E. coli via horizontal transfer and have not yet 'ameliorated' to blend in with the E. coli codon usage landscape. Figure 1.7 illustrates that most of these genes exhibit atypical codon usage bias and atypical nucleotidic content.

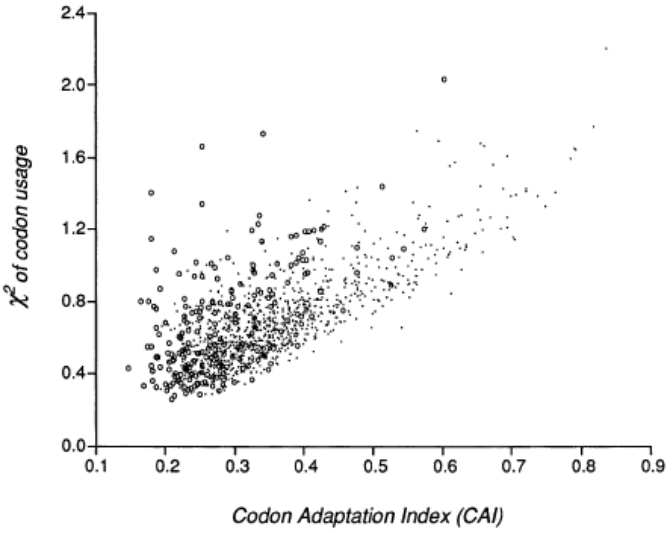


Figure 1.8: Plot of CAI vs χ^2 of codon usage for 1,189 E. coli genes. Points (n=1,024) represents native E. coli genes and open circles (n=165) represents genes inferred to be present in E. coli due to horizontal transfer (Lawrence *et al.*, 1997).

The 229 non-native genes present in this particular DNA stretch represent approximately 17% of the total and by extrapolation, the group proposed that about 618kb of protein-coding sequences within the E. coli K12 chromosome were introgressed. Similar studies have also been done on Mycobacteria, showing that it is possible to track mycobacterial evolution and the origin of its virulent genes (Becq *et al.*, 2007). Also interesting to note is that genetic exchange involving horizontally transferred elements may be influenced by organisms which in the first place, exhibit comparable codon usage statistics (Medrano-Soto *et al.*, 2004), thus environmental niches play a significant role in the amount of horizontally transferred elements found within bacteria. Codon usage has been shown to be an extremely insightful measure when comparing organisms. Related to the concept of codon usage bias is that of general oligonucleotide usage (OU). OU statistics have also been shown to be useful for several flavours of comparative genomics studies. An introduction to OU statistics and its varied uses and its application to this work will now be covered.

1.4.2 Oligonucleotide Usage Bias

The way in which bacteria use codons has already been shown to be non-random, in much the same way, oligonucleotide word usage among bacteria is also non-random with certain bacteria exhibiting an over- or under-representation of certain oligonucleotide words within their genomes. One of the earliest publications attempting to elucidate the oligonucleotide frequencies within genomes was in 1998 with Rocha *et al.* In this work, Bacillus subtilis strain 168 was chosen as a test specie. B. subtilis was chosen for analysis for various reasons. It is sufficiently long (~4.2Mbp)

containing over 4100 genes. It expresses and secretes heterologous proteins. Furthermore, it sporulates, making it a great candidate to study developmental processes. Several datasets were constructed using the *B. subtilis* sequence. These datasets include the Single-strand chromosome; Symmetrized chromosome; Leading and lagging strand; Genes, non-genes and prophage sets. These were constructed from biological criteria and for the purpose of being able to assign biological context to the bias use of words. A statistical method was devised to accurately count the word usage and this largely involved the use of maximal order Markov chains. Cross comparisons were performed on *Escherichia coli*, *Haemophilus influenzae* and *Methanococcus jannashii* complete genomes. Mono-, di and tri-nucleotide counts show a clear uneven distribution along the lengths of the sequence, irrespective of the dataset. There is also a clear trend when it comes to the usage of oligonucleotide words. Figure 1.8 depicts the total number of significantly over or under-represented words found in the single-strand chromosomes of the four species.

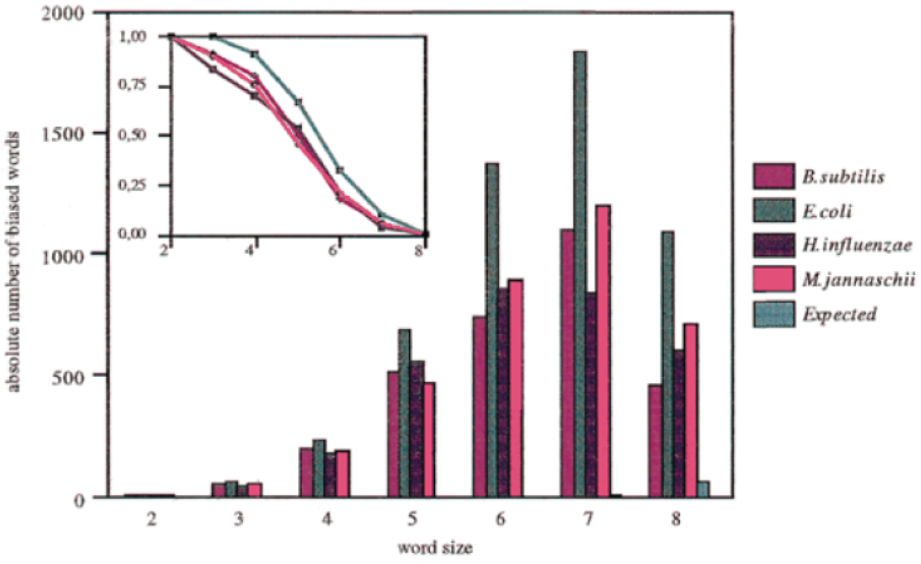


Figure 1.9: Graph depicting the total counts of biased words contained in the single-strand chromosome among the four species tested. Insert displays the relative number of biased words (i.e ratio of the number of biased words to the total number of possible words of that length) (Rocha *et al.*, 1998).

It is clearly visible that the total number of detected biased words increase with word length up until 7 nucleotides. The reason cited for this finding was that there are three competing effects 1) total number of possible words increases with word length (as more smaller words constitute larger words); 2) relative number of biased word statistics are able to detect decreases with word length and 3) for a specific dataset, longer words usually play a minor role as strict signals, thus, counting exact words also tends to underestimate the importance of larger signals. Amongst many other interesting finds made by Rocha *et al.* (1998), another worth mentioning is that a

uniform series of A and T heptanucleotides are consistently the most under-represented words in all datasets (Figure 1.10)

Symmetrized	Single-strand	Genes	Intergenic	Leading
Under-represented				
AAAAAAA/TTTTTTT	TTTTTTT	AAAAAAA	TTTTTTT	AAAAAAA
CTTTTTA/TAAAAAG	AAAAAAA	TAAAAAG	AAAAAAA	TAAAAAG
ATTTTTC/GAAAAAT	CTTTTTA	GAAAAAT	TAAAAAG	TTTTTTT
CAAGCAA/TTGCTTG	TAAAAAG	TTGATGG	CACCTCC	GAAAAAT
CAACCGA/TCGGTTG	ATTTTTC	TCGGTTG	CTTTTTA	TTGCTTG
CGATGAA/TTCATCG	CAAGCAA	TAAATTG	GTTTTTA	TTGATGG
CAATGAA/TTCATTG	GAAAAAT	GGAAAAA	TTTAATC	CTTTTTA
GGAAAAA/TTTTTCC	CAACCGA	TTGCTTG	TACAATC	TCGGTTG
CAACAAA/TTTGTTG	CAACGAA	GTAAAAA	TTCCTTT	TAAGAAG
CTTCTTA/TAAGAAG	CAAGCGA	GCTTTTT	TTAAAAA	TTGATTG
Over-represented				
CTTTTCC/GGAAAAG	TTTTTTA	GGAAAAG	TTTTTTA	GGAAAAG
TAAAAAA/TTTTTTA	GGAAAAG	GTAAAAG	TAAAAAA	TAAAAAA
GTAAAAG/CTTTTAC	CTTTTCC	AAAAAAT	AAAAAAG	GTAAAAG
AAAAAAG/CTTTTTT	GTAAAAG	TAAAAAA	CTTTTTT	AAAAAAT
CAATGAC/GTCATTG	CAATGAC	TAAAAGA	GTTTTTT	GAAATCG
CAAGCTC/GAGCTTG	CTTTTAC	GGAATCG	AAAAAAC	CTTTTTT
CAAGCAC/GTGCTTG	TAAAAAA	ATAAATT	TTTTTTG	GTCATTG
AAATCAA/TTGATTT	GAGCTTG	AATTTGA	CAAAAAA	GTGCTTG
CATTTAC/GTAAATG	CTTTTTT	AAGAGCT	TTCCTTC	AAGAGCT
TAAGAAA/TTTCTTA	CTCCGCC	GCGGCAG	TAAAGAT	AATTTGA

Figure 1.10: 10 most over-represented and under-represented heptanucleotides found in the datasets. Ranked by decreasing z values therefore, the most biased words are found at the top of the list (Rocha *et al.*, 1998).

Further analysis was also done with palindromic sequence distribution, and word usage contrast between leading and lagging strands. What is clear is that there is a definite bias in the way nucleotides and oligonucleotides are used within and between genomes. The availability of whole genomes means that it is now possible to test for longer words which will possibly allow the detection of biological signals within genomes of various species. Although the signals produced by the biased use of oligonucleotides are not biologically understood, in the future, with further comparative analyses and experimental data, a better understanding will be possible. Comparative studies such as these may also shed light on the preferential use of genes between organisms or aid in the understanding of transfer of genetic elements between species. Indeed, the same is also possible on a protein level (Bastien *et al.*, 2004).

1.5 Conclusions

Due to the ever advancing sequencing techniques, biologists are now more than ever being faced with massive amounts of sequence data. Even whole genome sequencing has become commonplace so much so that many researchers have become inundated with the amount of genome

data that has to be processed. Sequence databases such as NCBI and GOLD are growing at an unprecedented rate due to this high sequence production rate. Also, the internet has seen a concomitant rise in the number of sequence databases becoming available with each database group offers their own flavor of sequence analyses and data. This new era in genomic sequencing will undoubtedly have to draw on phenomenal amounts of computing power to be able to handle the load of data analyses. Due to the rise in publication of whole genome sequences, many researchers are now fortunate in that they have access to this genome data and can now conduct direct whole genome comparative studies. Indeed this trend is evident by the great spectrum of comparative genomic software tools available on the web. Whole research teams are sometimes dedicated to the development of comparative genomics tools. The advantage of this is that there is absolutely no shortage of free, open-source tools covering a wide range of analyses types available on the web for researcher to download and use. The basic types of comparative genomics tasks that most researchers undertake include whole genome alignments, gene-gene comparisons of genomes, inter-genomic SNP studies and phylogenetic studies. The range of tools available to handle these tasks, however, is overwhelming. There are algorithms and tools being published on a regular basis outlining improvements to old techniques, algorithms and data analyses tools. Often, researchers only require a few basic tools to handle most of their data analyses needs and have little time to seek out and install all the new software that becomes available. It is a sensible idea to be able to centralize those comparative genomics tools and data in order to streamline comparative genomics studies.

1.6 Problem Statement

Tuberculosis is one of the most prevalent diseases in South Africa. Globally, it has claimed and continues to claim millions of lives annually. The causative agent. *Mycobacterium tuberculosis*, has been fully sequenced together with several other closely related species. Several strains within the mycobacterium family are responsible for causing illness in humans, some strains even cause illness in cattle and others in birds. The multitude of diseases caused by mycobacteria and variation in host range specificity is a phenomenon not well understood. What is the genetic basis for these strains pathogenesis and host range specificity? We have developed a web-based, comparative genomics environment that seeks to study the level of similarity and differences of these different mycobacterial strains based on the available whole genome sequences. The comparative genomics system showcased here offers novels tool in sequence comparison which employs the use of oligonucleotide usage statistics to compare genomes.This is showcased by the SeqWord Genome Browser (SWGB). Furthermore, differences between these strains are examined with our, also novel, mycobacterial comparison project (MCP) on a gene-by-gene and SNP level. A sequence-based comparison among these various strains may prove key in understanding the physiology of these organisms.

1.7 Aims

In order to create a general comparative genomics environment supplemented by inclusion of a novel analyses tool, the following aims were set :

1. Construction of a database schema (model) to handle the variety of data expected.
2. Construction of a general comparative genomics environment (built on the model) complete with a few basic analysis tools and data storage options. (Chapter 2)
3. Implementation and addition of a novel tool, the seqword genome browser (SWGB) into the system for identification of gene islands and other sequence features. (Chapter 3)
4. Implementation of the novel Mycobacterial Comparison Project (MCP) in order to perform a gene-by-gene and SNP analyses between a few key mycobacterial species. (Chapter 4)

At the end, the value of these tools for comparative analyses will be demonstrated and then it will be shown how the SWGB and MCP system complement each other in their analytical functions.

Chapter 2

Design of an integrated comparative genomics environment

2.1 Introduction

There are many types of studies and tools that researchers may want to use when performing comparative genomic studies. As seen in chapter 1, some methods intimately linked with comparative genomics include sequence alignment, BLAST, synteny studies, SNP analyses etc. Often, analysis servers are dedicated to hosting specific analysis tools. This requires a user to navigate to the specific server, upload data, submit data for processing and then finally download the results. This process is often time consuming. If a user, on the other hand, chooses to download and use a specific tool on their local computer, there is still the requirement of downloading and installation of the actual software. The skills and technical know-how required for these tasks are often outside the scope of the regular lab-based biological researcher. Furthermore, after attempting to use software, the results produced by the software may not be in a form compatible for input into other software, thus requiring further processing of the results which is not a trivial task. This justifies the need for an integrated system that not only offers easy access to all the main comparative genomics tools, but also a degree of compatibility and communication between the software, thus the need for an integrated comparative genomics environment. In this study, such a system was developed as a sub-module of the larger FunGIMS (Functional Genomics Information Management System) project. The comparative genomics environment was built as part of FunGIMS for several reasons. Firstly, it offered a pre-built and stable database schema with which to expand on. FunGIMS also already featured a high performance server that was suitable for the inclusion of a new module. Also, importantly, is that the main function of FunGIMS was to offer a highly integrated software environment catering for the main biological data-types which should ideally include comparative genomics. This chapter firstly deals with a brief overview of FunGIMS, then with comparative genomics environments in general and what

they have to offer, followed by the design of the new comparative genomics environment which was developed in this study.

2.2 FunGIMS

2.2.1 Overview of FunGIMS

As already mentioned, FunGIMS or the Functional Genomics Information Management System is a system aimed at providing a web-based environment where biological researchers are able to manage a variety of functional genomics data types (private and public) including micro-array data, protein and DNA sequences as well as small-molecule data. Sub-modules within the project, each specifically geared toward the different data types, allow for users to also perform general, commonly performed analyses tasks on the relevant data types. These sub-modules are built into FunGIMS in such a manner so as to facilitate 'cross-talk' between the various data-types and establish biologically relevant linkages between the various data-types. Core modules comprising FunGIMS are responsible for the security, database access, data storage and user management of the system. FunGIMS design was based on the Model-View-Controller design paradigm (discussed later) and henceforth, a brief description of FunGIMS will be given within this context.

2.2.2 Model

One of the main purposes for the development of FunGIMS was the need for integration of several biological data types. For this reason, a versatile data model had to be employed. Although, no one data model at the time of development, catered fully for the needs of this project, the FuGE (Functional Genomics Experiment) data model however was most suited to our needs as it was developed to facilitate convergence of data standards for high-throughput, comprehensive analyses in biology (Jones *et al.*, 2007). The FuGE data model, in essence provided a framework upon which custom designed sub-models could be built, specifically geared toward the various data types handled by the system. The FuGE object model, in its base classes Describable and Identifiable handle all aspects of the security and access to data. (Pizarro *et al.*, 2006)

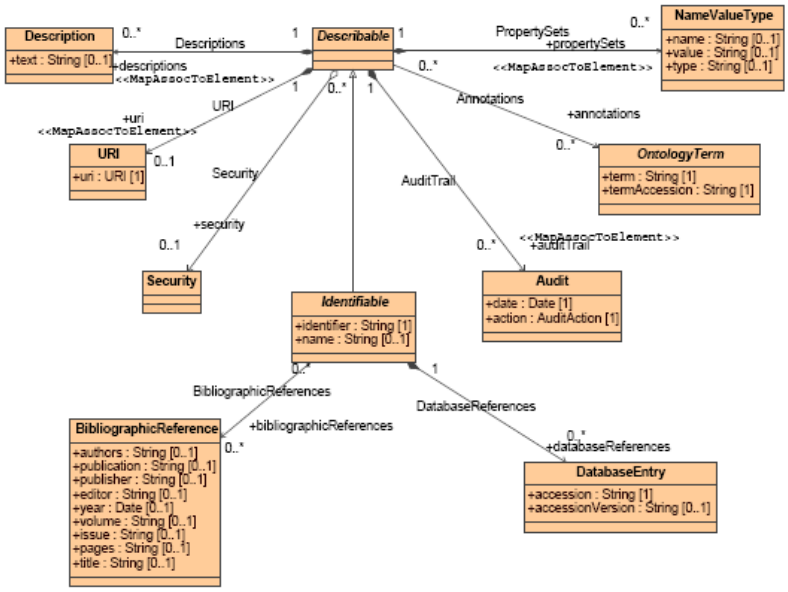


Figure 2.1: Main base classes within FuGE. Newly developed classes developed within FunGIMS inherit from these classes (Pizarro *et al.*, 2006).

By the process of polymorphic inheritance, all subsequent classes developed, inherited from the Identifiable and Describable base classes and thus inherited all their functionality. More of FuGE will be discussed later.

2.2.3 View

All views within FunGIMS were generated with the KID templating language and converted to HTML before being delivered to the client's browser. Due to FunGIMS being a collection of software modules, common themes and cascading style sheets (CSS) were employed by the various developers working on the various modules. This sharing of styles helped to maintain a common look and feel when within FunGIMS or any of the sub-modules. Style sharing was facilitated by each webpage inheriting from a single 'master.kid' file. Subsequent to inheritance, develops then added and customized their web pages to suit the data-type and context.

2.2.4 Controller

The FunGIMS controller is where all the functionality for the system is held. The root controller class contains all functions necessary to run the system. Functions declared within the controller are registered by the CherryPy server and request URLs (as well as parameters) from a clients browser are mapped to these functions. Turbogears makes extensive use of CherryPy which forms the controller layer of the FunGIMS system.

2.3 Examples of comparative genomics environments and what they have to offer

Several comparative genomics environments are available through the web and the variety of functionalities they offer are quite broad. One example of a contemporary comparative genomics environment is VISTA found at (http://genome.lbl.gov/vista/index.shtml). VISTA is a web-based comparative genomics environment offering a range of tools such as whole genome alignment wgVISTA), normal sequence comparison (mVISTA), and basic phylogenetic analyses (Phylo-VISTA). VISTA provides many great features but lacks the ability to store user data and user results generated by the software. Furthermore, the various tools runs from different servers therefore, there is no integration of the results with the different software components. On the whole, VISTA is a great comparative genomics environment offering most of the major tools that are required for comparative genomics studies. Sybil, another example of a comparative genomics environment is a web-based system allowing users to perform tasks such as genome browsing, genomic-region comparison, syntenic viewing (Figure 2.2) and a few others.

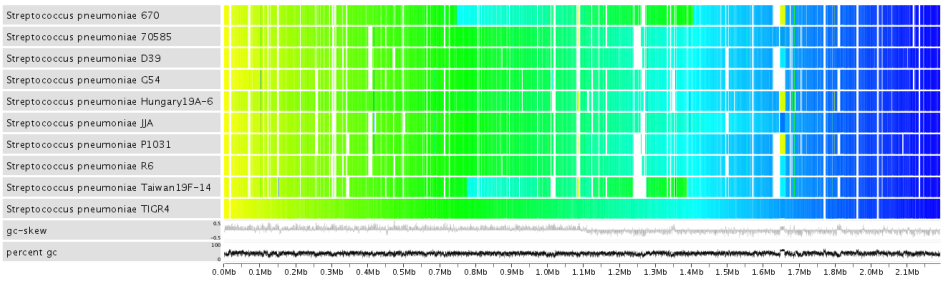


Figure 2.2: Screenshot of Sybil's synteny gradient

Sybil offers 2D graphical displays of back-end pre-calculated comparative datasets however, the scope of its analytical software offering is very limited and users must rely on data that is pre-calculated on their system. The BioMAX knowledge management environment (www.biomax.com) is yet another example of a comparative environment. BioMAX system serves to integrate data from various sources (i.e. databases such as KEGG, BIND, DIP etc) by dynamically building semantic networks between the various data-types. BioMAX then allows the user to perform advanced queries on the underlying data in a very simple way. Users of BioMAX may create, manage and visualize scientific models as an extendible network of interrelated concepts. Through APIs, other tools such as R and Bioconductor scripts may be integrated into the system. BioMAX is quite an advanced distributed software platform allowing users to manage data from various sources and subsequently analyse the data based on the semantic networks built by BioMAX. This is very useful when dealing with matured data sources containing well annotated data. However, when it comes to new sequence data, and the fact many more organisms are being sequenced, often unknown, means that this system is not well suited due to

the inadequacy of publicly available data.

2.4 Requirements

2.4.1 User interface requirements

At the heart of design of any software project is the requirements and skill level of the end user. The software in question must therefore, cater for all the requirements of the user as closely as possible and be suited to the users level of computing knowledge. In terms of requirements, this software needed to offer a unique range of tools dealing with comparative genomics. All tools needed to be integrated into one interface which would provide access to the various functions by the click of a button. The software had to allow for the tools to be used with uploaded data as well as data stored on the central system. Needless to say, the software had to therefore make provisions for the storage of user data on a central database. Storage of personal data on an external server also needs a security component to protect data privacy, thus the software needed a security component. Lastly, access to the software needed to be versatile and easy to maintain, therefore the entire software was created as a web based system affording users access from any computer with an internet connection. In terms of user skill level, the end users of this software include researchers and students alike. This group is mostly comprised of experimental based biologists who often do not have very advanced skills in terms of software and databse installation, software usage and data handling. This system had to therefore, abstract these components from their view and merely offer a very simplified, intuitive web-based interface to the full functionality of the back-end system. A simple login and password, set up by the administrator will allow any user to access the full functionality of the system, and access to their allowed sub-set of data. In summary, design of this software was closely guided by the abovementioned requirements of the users and as far as possible, usage of the system was intended to be as simple and intuitive.

2.4.2 Analysis Requirements

Prior to the actual addition of software into the system, careful consideration had to be given to to the types of analyses and scientific investigation typically undertaken by biological researchers in comparative genomics studies. Typically, researchers want to upload nucleotide and protein sequences and BLAST these against non-redundant databases in order to gauge taxonomy or identify their sequences. Alignment of two or more sequences (nucleotide or protein) in order to assess sequence similarity and differences is also a popular task. Downstream of this, users often want to examine single nucleotide polymorphisms between their sequences in order to make informed decisions regarding primer design, protein conformational changes and various other mutations. The construction of phylogenetic trees has become common place in assessing phylogenetic relationships amongst various sequences in question. Based on the researchers' need to perform these and other tasks, tools such as BLAST, MAFFT, ClustalW, Phyllip and BlastZ

were integrated into the system. Integration, a key theme of this work has implications on several layers. In terms of software integration, this merely describes the centralized accessibility to the various tools and inter-compatibility of output and inputs from and to the various software. On the database level, integration points to the fact that various sub-models (sub-schemas) form part of the larger database model functioning in a compatible and complimentary manner, although each sub-model is capable of functioning exclusively. On the project level, the comparative genomics module is also integrated into the FunGIMS project as it subscribes to the technologies, database structure and coding standards employed by FunGIMS. Strong emphasis was placed on integration on all these levels throughout development of the system.

2.4.3 Data structure requirements

In the design of this data model, an extension of the Functional Genomics (FuGE) object model (Pizarro *et al.*, 2006) was employed with major adaptations (see later). A complete stand-alone solution, incorporating all biological data-types is an impossible task due to the vast array of experimental techniques and data-types that are currently and prospectively available. In light of this, the FuGE development team instead attempted to model only those aspects shared amongst the various functional genomics experiments such as researcher contact information, sample preparation and protocols. This was achieved in two ways. Firstly FuGE makes available a general database structure which enables for the referencing of external data formats which captures the meta-data which in turn provides context to a specific scientific investigation. Secondly, and most importantly, these abovementioned reference points acts as a start point for extensions to the FuGE model which translates to the defining of sub-models specific to any functional genomics technology (Pizarro *et al.*, 2006). In this way, the FuGE object model allows for limitless extension while maintaining a high level of integration and thus provided the perfect base to build on. Ease of data integration, with the help of FuGE thus became a reality, however, the data model developed for this system had to also take into account several key factors. Volume is one of the major concerns especially in the context of biological data. Biological experiments produce data at enormous quantities. Furthermore, with the latest DNA sequencing technologies being implemented globally, a concomitant rise in database growth is also expected. Biological sequence databases can range in volume from thousands to billions and even trillions of entries, thus the design of the database should be appropriately scalable. With databases of such large proportions comes the question of speed. The design of the database should take into account that every entry in the database may need to be searched upon at some stage, therefore this is another crucial aspect to bear in mind during the design. Data integration is also quite an important feature that needs to feature in this data structure. Due to the great heterogeneity in sequence data, such as annotations, various sequence types, genomic sequences, protein sequencing, referencing material and so forth, great care needs to be taken to make sure that the overall data model can appropriately handle the various types of sequence and meta-data. Furthermore, new types of technologies and sequence classes are being produced all the time, therefore the

design of the system should be built with the prospect of extensibility to other future data types. Data security is extremely important in this day and age when academic publications depend on raw data being kept private. This system needs to be designed with great care being taken to protect visibility of individual as well as group data. The security classes controlling data visibility should make sure that only system users with the proper privileges may see certain data and the system data should also be totally protected from the individuals external to the system. Details regarding how these criteria were fulfilled in the design measure taken as well as the choice of technologies used at every stage will be covered in the design principles section next.

2.5 Design Principles

In the design of this software, several key principles were taken into account:

- The system must be web-based and intuitive to use
- The back-end data structures must be able to handle large amounts of data while not compromising on speed and agility in the integration of the various types.
- The software should take care of data security precluding wrongful visibility of data.
- Lastly, the software must offer the range of comparative genomics software tools providing a range of useful functionalities from a single interface.

2.5.1 User interface requirements

Principles used in the design of the interface for this software were largely based on the user requirements set out. Therefore, first and foremost, the web-based interface was designed so that usage would be very simple and efficient. This meant that there would not be the addition of too many icons and menus on the screen to inundate the user. Rather, fewer screen items were added and navigation to tools and user data was made possible in a maximum of two mouse clicks. Soft colors were chosen in the design of the interface so as to not be irritable to the eyes after many hours of usage. The inclusion of the software tools to this system was based on what were the major and most common tools used by researchers. Furthermore our system was to offer two unique tools in addition to the commonly available ones which were, the Seqword Genome Browser as well as the Mycobacterial comparison analyses suite. Details of the design and implementation follows.

2.5.2 Data structure requirements

Design of the data structures to handle the back-end data was largely influenced by the data requirements. Three main design aspects needed to be taken into account. Firstly, the data

structure needed to be highly scalable and should be able to perform optimally even after being populated with millions of entries. Secondly, the data structure needed to cater for several different data types while maintaining a high degree of integrity. Thirdly, database security was a crucial aspect and data privacy was therefore crucial in the design of the database schema. Lastly, provision needed to be made for the addition of new data types if the need did arise, thus, the database model was designed in a manner that was conducive to extension to other data types.

2.5.3 Software components and technologies employed

The software and technologies employed for this project are outlined below.

- A MySQL database
- The Turbogears web-development toolkit
- Several open-source biological analyses and graphical software tools (example BlastZ, Phylip, Laj and Clustal)
- Several programming languages (Python, BioPython, Javascript, HTML, KID)
- Integrated development environment (IDE)
- SQLAlchemy and SQLObject
- XML-RPC
- AJAX

Details regarding how each of these components was integrated will be discussed in the next section.

2.6 Model-View-Controller Architecture and integration

2.6.1 Model-View-Controller Pattern

The model-view-controller (MVC) pattern is a concept in software development that can be implemented as both a design pattern as well as an architectural pattern. The MVC concept basically aims to separate an application into tiers or layers viz the model, view and controller layers such each of the layers can be independently modified without adversely affecting the other layers. The model layer represents the actual data or content as well as the rules governing how the data is managed. The view is essentially the component that is involved with presentation of the data in the model such as the text, checkboxes, buttons etc. The controller is the business layer of the system and handles all the logic of the system as well as communication between the view and model. This is comprised of the actual programming code. Successful implementation

of the MVC pattern means that the user interface or the business logic layer can be modified independently while still maintaining system integrity. Using the MVC design pattern also allows several views of same underlying model. One of the simplest examples demonstrating MVC design pattern is found in a browser. Where the model is represented by the content (HTML), the view is represented by the CSS as it dictates how the data will be presented and finally the controller is represented by the controller as this controls communication between the content (model) and the css (view) and controls which data items will be displayed. This same MVC design pattern was employed for this project and details of its implementation will herewith be discussed.

2.6.2 Integration of the various components under the M-V-C design pattern

For this project, Turbogears was the python based web-development framework used. Turbogears is designed around the MVC paradigm and thus allows for separation of the various MVC components. For each of the various MVC components, the strategies, software and technologies used in the context of this project will now be discussed.

2.6.2.1 The Model Layer

The model layer represents the actual database, the data and the classes defining the various object types. In this layer the following technologies and software were used

- MySQL (Database server)
- SQLAlchemy (Object relational mapper)
- SQLObject (Object relational mapper)

MySQL, a robust, fast, reliable and highly scalable database server with a proven track record was used for data storage (6). MySQL comes in open-source database versions and also satisfied all the design criteria set-out above and was thus the database of choice. MySQL uses SQL (Structured Query Language) to communicate with other programs. MySQL also has its own set of augmented SQL commands which offer users more advanced and specific functionality. Communication with the database and its data was accomplished by using the object relational mappers (ORM) SQLObject and SQLAlchemy. SQLObject is an extremely popular and well established object relational manager that allows a user to interface with databases via objects. SQLObject treats tables as classes, rows as instances and table columns as class attributes. SQLObject also includes a python-object based query language that allows higher level usage of SQL thus providing users with more versatile database interfacing and a greater independence from actual SQL code (7). SQLAlchemy is another SQL toolkit and ORM for the python programming language. SQLAlchemy is quite a mature ORM offering very efficient and high-powered database access. This high-powered access is partly due to the fact that SQLAlchemy does not view database

tables as mere tables but rather as 'relational algebra engines'. SQLAlchemy allows mapping of classes against databases in several ways thus allowing many useful and powerful features such as cascading, complex selects, complex joins, sub-queries unions and many others (8). Both SQLObject and SQLAlchemy were used in this project as they both provided abstract access to the database in a python friendly manner while still offering all the power and functionality of SQL from within the Turbogears framework.

2.6.2.2 The View Layer

The view layer is essentially the user interface and it is what the user interacts with. This includes the web browser, buttons, etc. In this layer, the following technologies were employed

- Kid templating
- Javascript & Mochikit

All the above-mentioned items were used to create graphical user interfaces for the underlying data model. Users can point their web browser to the correct url and thus interact with the underlying data model via the controller. Kid is essentially a template generation engine that is used for XML compatible vocabularies written in the python programming language (9). Within the Turbogears framework, Kid is extremely useful for several reasons. It allows one to incorporate python into the actual templates. It allows inheritance, thus many pages can display from one template by merely including the appropriate XML tags. Kid also forces users to adhere to valid XML standards thus precluding syntax errors such as mismatched tags within the template. The kid templates produced within Turbogears then gets converted to XHTML and along with the data, is delivered to the client (i.e a browser) and displayed. Also involved with the view layer is Javascript and Mochikit. Javascript is a popular scripting language that is used in client-side web development. Javascript is useful in that it makes the interface more interactive. Also, due to the fact that it runs on the client side (i.e in the browser) it reacts very quickly to users inputs. It is used in this project for quick validation of user input on the web-page as well as for inter-communication with DOM components within the web-page. Mochikit, is essentially a lightweight javascript library that adds python-like features to javascript. Both javascript and mochikit are highly compatible with the Turbogears framework and both contributed substantially to the creation of a user friendly view.

2.6.2.3 The Controller Layer

The controller layer is the most computationally intensive and complex layer. This layer is typically responsible for receiving and responding to requests made by the user and also for communication and modification of the underlying data model. The main technologies used here are the CherryPy server and the python programming language. CherryPy is an object oriented web-application framework that uses the python programming language (10). CherryPy was

designed with the aim of facilitating rapid development of web based applications by wrapping the HTTP protocol. Due to CherryPy being very pythonic in nature, CherryPy may be used as a regular python module. Turbogears makes extensive use of CherryPy to map user request URLs and parameters to python functions. Python functions are written within Turbogears and handles data and user requests in specific ways. The results are then passed back to the user/client via CherryPy as a server response. All the analyses functionalities offered by the system are essentially python functions contained within the Turbogears framework and handled by the CherryPy module. Python is a dynamic, object-oriented programming language (11) and it can be used as a simple scripting language or for the development of high powered web-development frameworks such as Turbogears. Due to its ease of use and versatility, it was chosen for the development of this project. The various aspects of the MVC design pattern has been explained, the following figure aims to summarize the MVC design through interaction of the various technologies within Turbogears.

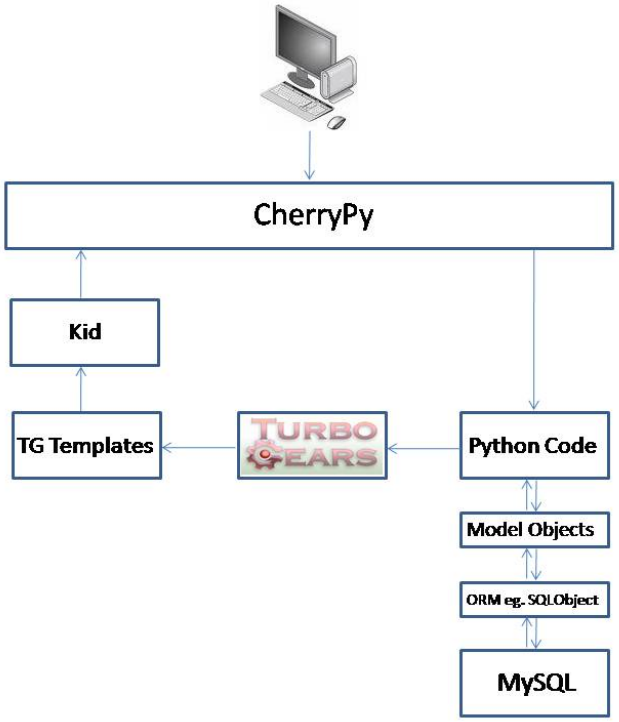


Figure 2.3: Figure illustrating the MVC design pattern in the context of Turbogears. Numbers represent the order of events subsequent to a user making a server request from the browser.

Using Turbogears and its MVC architecture, an integrated web-based comparative genomics analysis suite was developed. The analysis functions offered by the system are very varied but are all integrated into one web-based environment. The general system implementation will now be discussed.

2.7 Technical implementation details

This section will deal with the technical details of the project implementation in the context of the MVC design pattern that was previously explained. For every MVC layer, an explanation will be given regarding the software and technologies and how exactly they were used and integrated.

2.7.1 Database implementation

This essentially represents the model layer. MySQL was the database server used and all data in the system was stored in specific tables defined, created and populated using the SQLAlchemy ORM. The actual database structure was based on the FuGE model however, the structure was substantially adapted in order to deal with all the unique data types within the system. The basic class of the model was the Identifiable class and all datatypes inherit from this class (Figure 2.4).

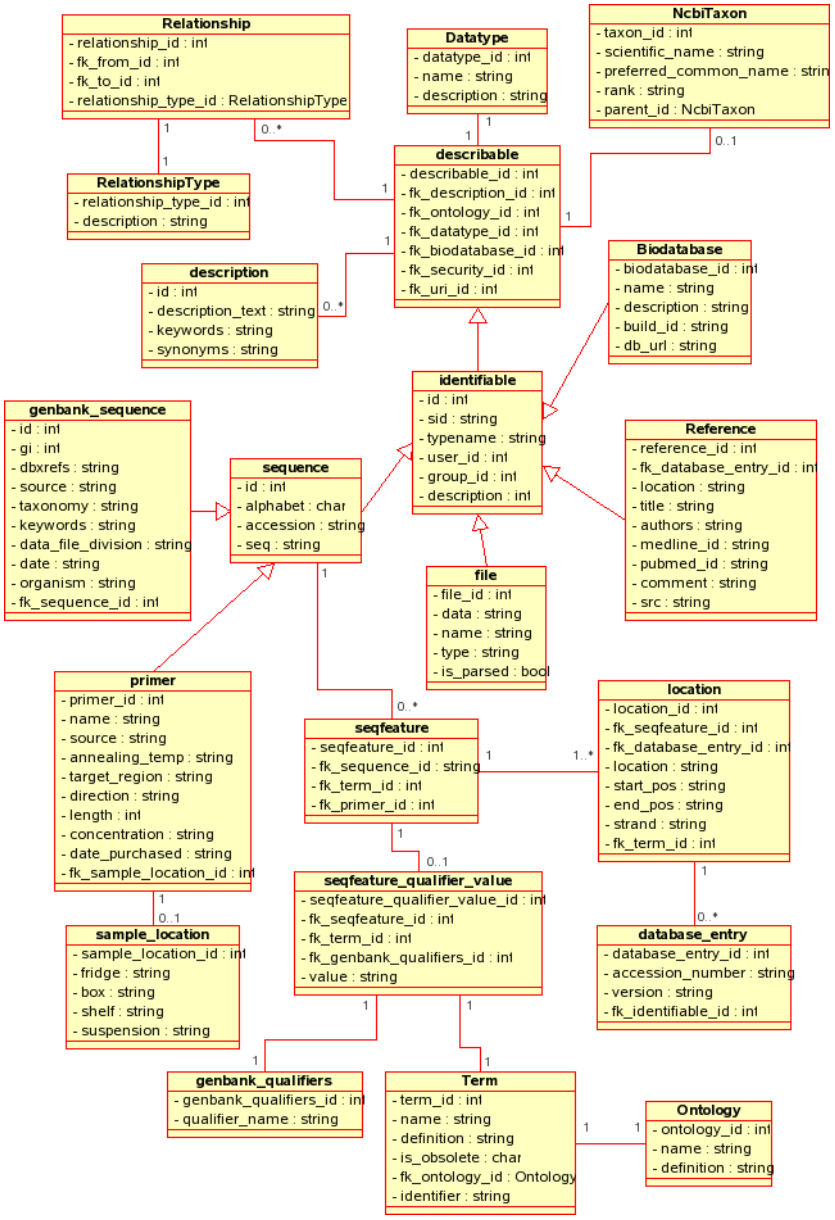


Figure 2.4: UML class diagram showing some of the major classes used in the database and the relationships between them.

Using eric3 as the Integrated Development Environment (IDE), all SQLAlchemy classes, tables and mappers were defined in the 'model.py' file under a Turbogears project. Turbogears then offers a command-line utility 'tg-admin sql create' which creates all the tables in MySQL and maps them together according to the definitions laid out in the model.py file. SQLAlchemy was used extensively because it also allows one to perform basic and complex MySQL commands from within the controller (more about this to follow). SQLAlchemy connects to the database

via drivers configured in the 'dev.cfg' file also located under the Turbogears project folder.

2.7.2 Graphical User Interface (GUI)

The GUI is essentially all the user sees and interacts with and thus represents the view layer. In Turbogears, views are generated using the KID templating engine which generates HTML, recognized by all browsers. Using Eric3 again, the KID templating language was used to script HTML and Javascript to define how the data and the HTML page elements would be displayed. KID eventually generates HTML which is delivered to the browser by the server (i.e CherryPy). Users may through the use of this HTML rendered by web browser such as Firefox, Konqueror or Mozilla interact with the underlying system or in other words, communicate with the controller (which in turn communicates with the model). This interaction includes simple visualization of their data but also manipulation of their data and other requests. Also, the view is responsible for displaying of data. A simple scenario is as follows. A user navigates to his data storage page and then wants to delete a specific entry. This is made possible by a form on the page. The user then clicks on the "delete entry" button thus invoking a request to the controller which then interprets the request and performs the action on the database with the help of the ORM (i.e SQLAlchemy). All requests are handled by the CherryPy server (see later). Also involved in the view is Javascript. Javascript in this project was used for various purposes such as for notification to the users, user-input validation and for inter-communication of the data within the page (Figure 2.5).

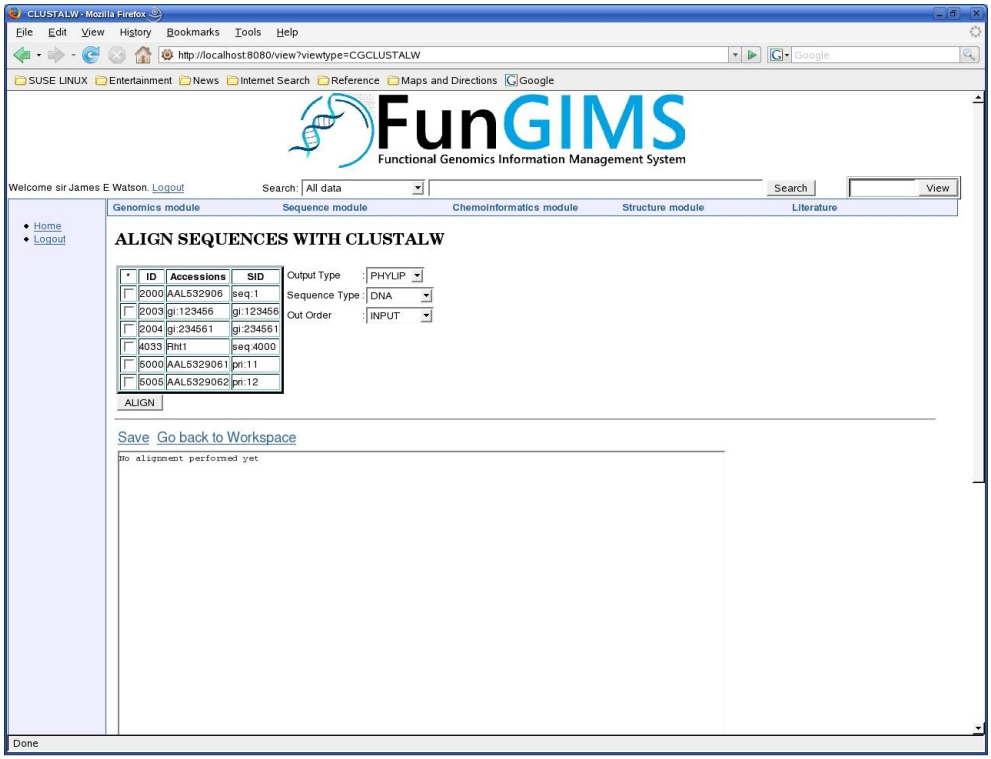


Figure 2.5: An example of a typical view that a user sees. Everything visualized on the page is essentially HTML generated by KID. The 'ALIGN' button is the users way of communicating with the controller and in-turn, the underlying data. Javascript is responsible for user-input validation.

The above figure (Figure 2.5) is an example of a typical view a user will be faced with. The 'ALIGN' button is provided so that the user may communicate with his data via the server (controller). Javascript, used for input validation will in this case, check that the user first selected several sequences before allowing the request to be sent to the server. Checking request submission in this way is advantageous in that it precludes invalid requests from being submitted and unnecessarily utilizing the server. In the project, the view was designed to be intuitive and user-friendly. These two attributes are important for success of software. The colors chosen were soft and easy on the eyes. Drop-down menus were added to facilitate easy navigation and cluttered screens which often overwhelms users, were avoided. As already mentioned previously, the view is a means for a user to communicate with the controller. The controller, which is by far the 'nerve center' of the project will now be dealt with.

2.7.3 The Controller

The controller layer is the most computationally intensive layer and it controls the model layer beneath it, the view layer above it, communicates with 3rd party software and is responsible for sometimes carrying out analyses functions. Within the Turbogears framework, the analyses functions and all instructions for the system, correspond to methods contained within the main class which is the 'Root Controller'. The Root controller and its methods are found in the 'controllers.py' file. When a user makes a request to the server via the view, CherryPy maps this URL and the parameters to a specific function contained within the root controller. The function within the root class is then executed either by CherryPy itself or instructions are handed to a third party software via an extended-mark-up-language – remote procedure call or XML-RPC. Remote procedure calls are essential in cases when third-party software or data sources are located on separate computers. The controller class lastly, receives results or data from the analyses software or databases respectively and decides on which view to invoke. In order to elucidate the functioning of the controller, a typical example of a request/response model will now be given continuing with the 'delete entry' example used above. From the view, the user will make a request to delete an entry by clicking a button. CherryPy then receives this request (along with the parameters such as the ID of the entry to be deleted) and maps it to the correct 'delete entry' method contained within the root controller. The method, with the aid of the ORM (e.g SQLAlchemy) then communicates directly with the database to delete the user specified entry. SQLAlchemy first compiles the command into a MySQL friendly format and issues this to the database. The ORM then receives communication from MySQL detailing whether the 'delete entry' command was successful or not and passes this back to CherryPy. CherryPy via the same method that was first invoked, then decides which KID template and data to display. CherryPy , now generates the HTML from the specifications within the KID template and passes this via the HTTP protocol back to the browser which is then visible to the user. Needless to say, the controller (represented by CherryPy) shoulders a lot of the communication and computational burden involved with interaction of the system. It is responsible to relay messages to and from the user to the server and for the relay of messages to and from the model as well as deciding which views are to be passed back to the browser (client). The above example outlines much of the technical details involved in this system. In the next section, however, a more high level, user centered view will be used to explain the implementation system in terms of the general CG tools offered by the system.

2.8 Implementation of a general comparative genomics environment

Although there are a multitude of analyses tools already available on the web, it is very useful and efficient to have a centralized server environment where users may have access to various tools.

This was the idea behind the development of this project. Addition of all available CG tools into one analysis suite is unrealistic, therefore a few key analyses tools were chosen for inclusion into the project. The sub-set of analysis functions chosen would be sufficient in meeting the basic needs of CG research, however, novel tools were also added (details of these will be dealt with later). The basic set of analysis tools included within the project are :

- DNA sequence alignment (MAFFT and Clustal)
- Genome Alignment (BlastZ)
- Phylogeny analysis (Phylip)

In order to upload the users data into system so that it conforms to the standards of the model. The data was first passed through the relevant analyses tools. The output from these would then be parsed by various python scripts which then subsequently inserted the formatted data into the relevant tables.

2.8.1 DNA sequence alignment

The alignment of nucleic acid sequences is one of the most basic and popular bioinformatics analyses being performed and was thus included in the system. Once logged into the system, a user may choose to perform sequence alignments by selecting the appropriate drop-down menus.

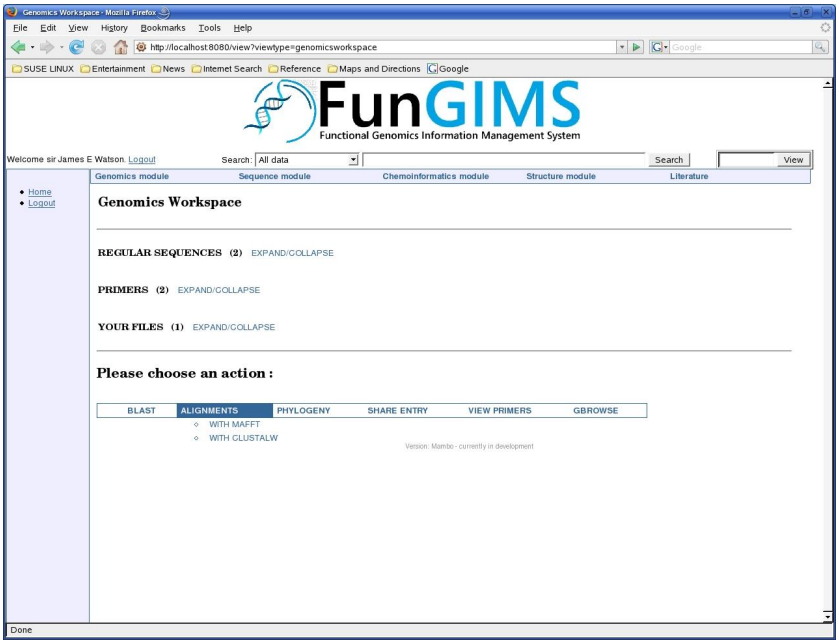


Figure 2.6: All functionality within the software suite is accessible via either the main-menu dropdown at the top of the screen or through the sub-menu at the botton of the screen.

A user may either choose to perform MAFFT alignments or ClustalW alignments. Once on the subsequent page, all the sequence data the current user has privileges to is displayed on the top of the screen (Figure 2.7). The user can then select a set of sequences that he/she wishes to align and also select certain available output format options.

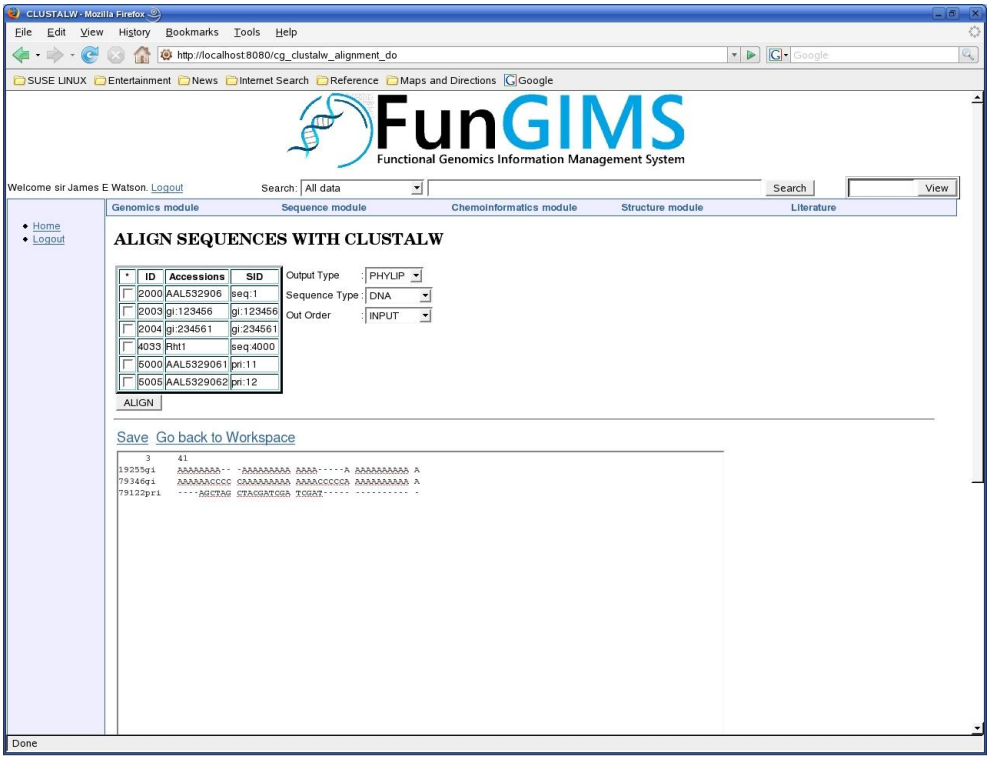


Figure 2.7: Sample page showing the ClustalW alignment results.

Note that, when users submit their requests, in page error checking is handled by Javascript and post-request errors are handled by the controller.

2.8.2 Genome alignment with BlastZ

Genome alignment using BlastZ may be performed by a simple three step procedure. From the main job submission page, users may upload two genome sequences at a time for alignment using BlastZ and then choose their comparison type (Figure 2.8).

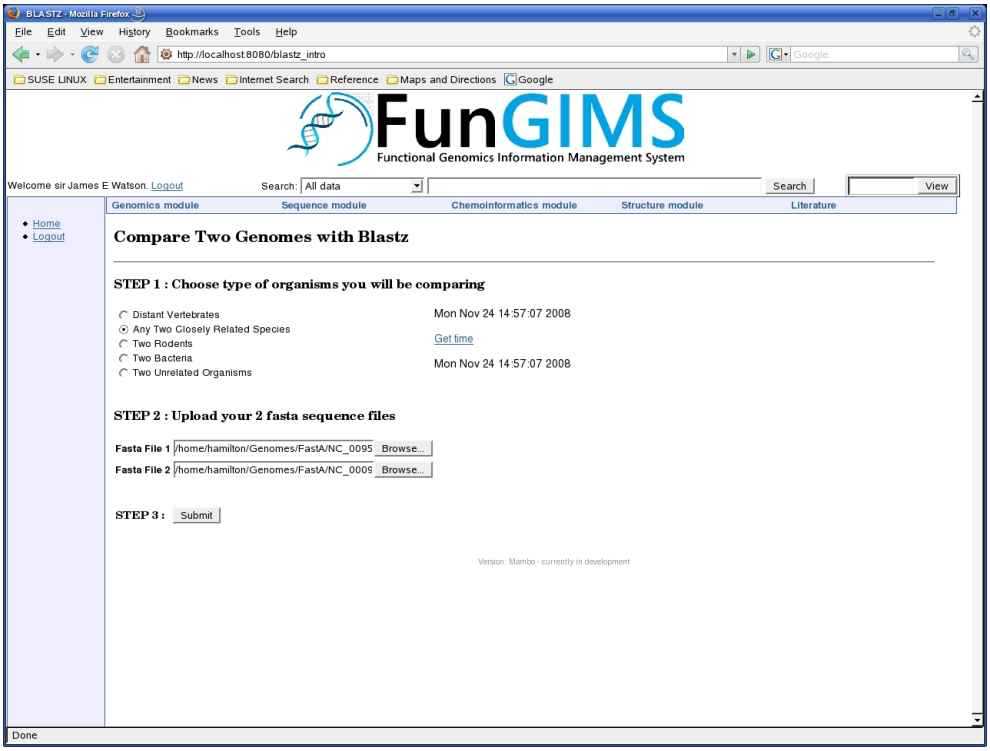


Figure 2.8: Main BlastZ submission page for the alignment of whole genomes.

Owing to the fact that genome sequence alignments are often time consuming, pages subsequent to job submission allow the user to check the status of the job at anytime (Figure 2.9). If the submitted job is unsuccessful due to problems such as inappropriate file formats and such, an error will be thrown and the job discarded. Users are then re-directed back to the main submission page (Figure 2.8).

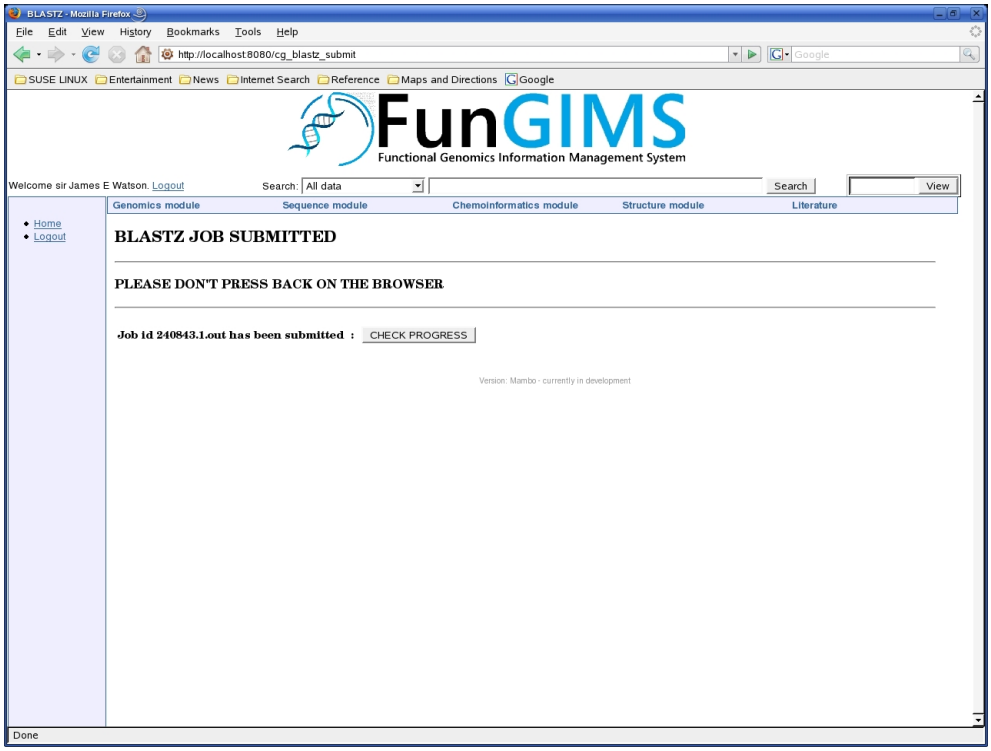


Figure 2.9: A successful BlastZ job submission will direct users to this page. Here, users may check the progress of their jobs by clicking the 'CHECK PROGRESS' button.

If the job, however, is successful the user is then presented with the result page where they are presented with a quick summary of the alignment results and the options to view their results in two ways. The main BlastZ result page (Figure 2.10) allows users to view the alignment result file either as plain text or graphically via the Laj applet (Figure 2.11). In terms of plain text results, clicking on the 'View original outfile' item will invoke a text box containing the raw text of the alignment file which a user may peruse. This file the user may download by clicking on the 'Download results' item. In terms of graphic results display, the Laj applet merely allows a user to view an interactive dot plot of their alignment results. The applet is easily viewed by clicking on the 'View results in applet' option. Laj also allows users to zoom into the actual alignment where they can explore sequence similarity on the nucleotide level.

Figure 2.10: Main result page of a BlastZ submission.

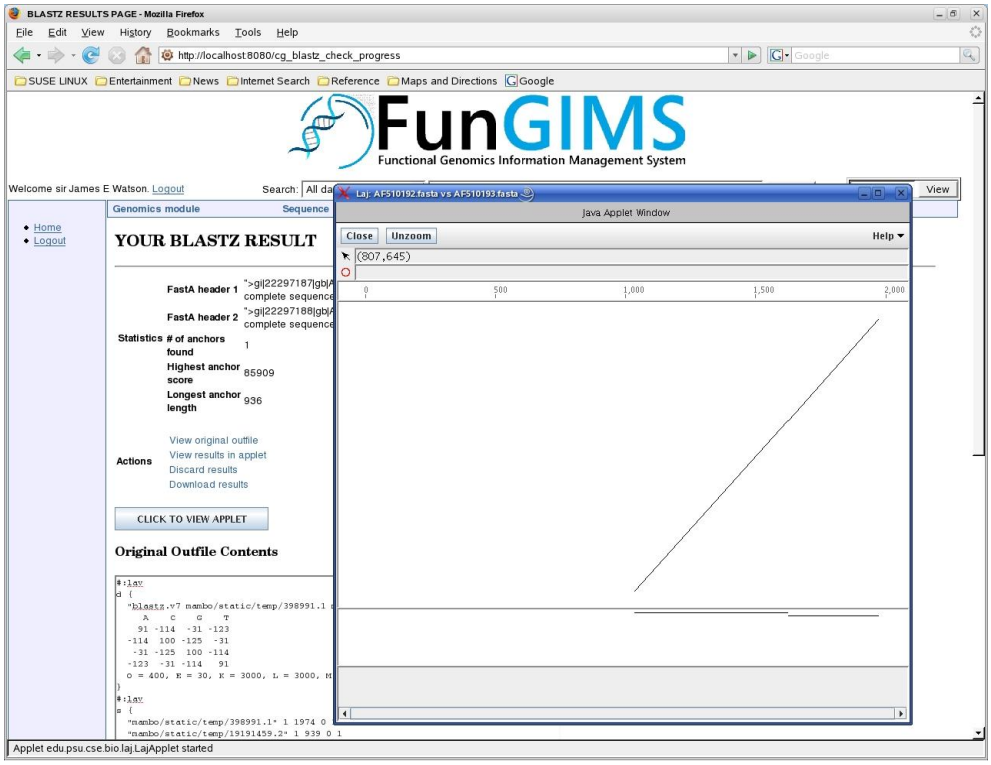


Figure 2.11: Graphical display of alignment results using the Laj applet.

Once a user is satisfied with his results, he or she may choose to download the results onto their local machine. If a user then exits the main result page, the result files are automatically deleted to save space on the server. Next, a brief overview of the phylogeny module will be presented.

2.8.3 Phylogeny analyses

The comparative genomics module also enables quick and easy neighbor-joining tree generation based on user uploaded sequences. On entering the main 'Genomics neighbor-joining tree' page, the user is presented with a list of genomic sequences that he or she has access to (Figure 2.12).

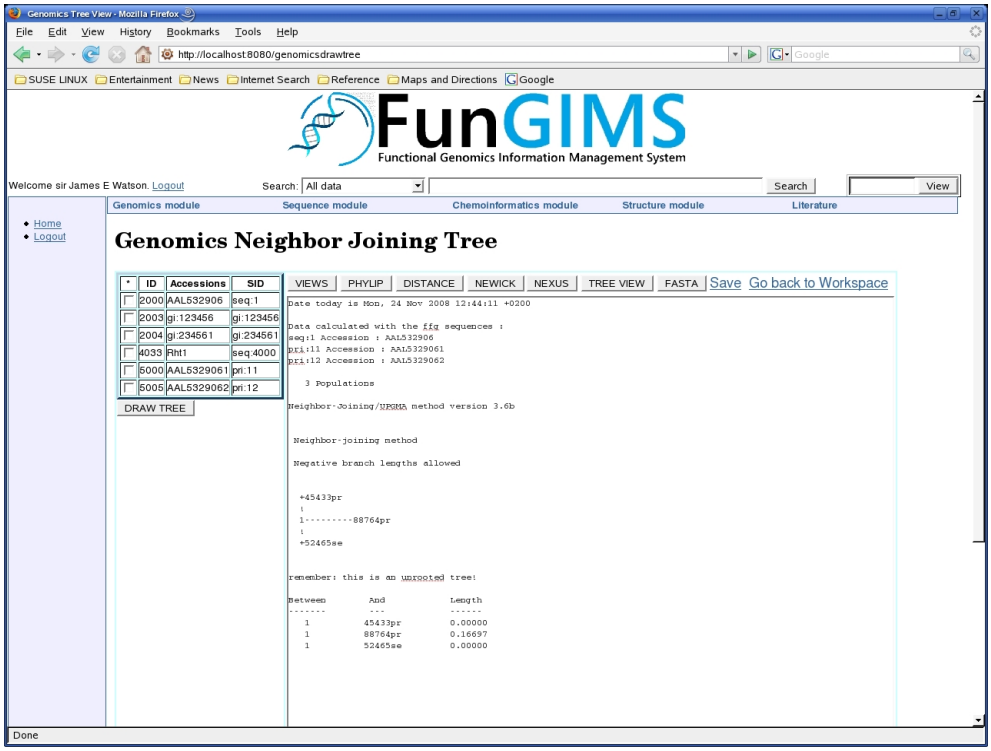


Figure 2.12: Neighbor-joining tree result page.

The user selects two or more sequences which they wish to align and then clicks 'DRAW TREE'. The server then responds by retrieving the actual sequences, performs clustalw alignments, calculates the distance matrix and lastly plots the tree. The text form on the tree is then displayed in a text box. When performing the above step, all default values are assumed as the output is meant to give users an overview of tree topology. Bootstrapping, though not a function in the current version of the software, may be added in subsequent releases. A user also has the option of viewing the tree file in one of several formats including nexus, newick or phylip format for example by simply clicking on one of the buttons above the text box (Figure 2.12). The program used to generate the matrix as well as produce the tree is phylip. The user also has the option of the saving their results onto their local P.C for the purposes of carrying out further analyses or viewing with other programs. This sub-module of the CG module serves to demonstrate the ease at which data and software can be centralized to perform basic analyses tasks.

2.9 Conclusion

Usability, convenience and relevance are crucial aspects to consider in software design. For this project, specifics needs by biologists which is namely, easy access to web tools and data

security were identified and guided the design of this piece of software which harmoniously incorporated all the crucial design aspects. FunGIMS, the general management system caters for a wide audience, offering everything from protein sequence analyses functions to chemi-informatics functions. FunGIMS, is a good example of how several types of software and data types can be integrated into one system while still maintaining user friendliness, convenience and relevance. Based on the successes of FunGIMS and using its development framework the comparative genomics sub-module was developed which offers users access to many useful tools pertinent to comparative genomics such as genome alignment and visualization, phylogenetic functions and others. All software and their results produced are seamlessly integrated into this single environment thus allowing researchers to more rapidly assimilate their data. This system serves as an example of how software and data of various types can be integrated 'under one roof' allowing users access to their data under various contexts (sub-modules) while still in a secure environment. The web-access to the system adds a further dimension of convenience permitting users to not only perform their research tasks from any location but also at any time. Some of the shortcomings of the system include performance, coding standards and documentation. Due to the large of amount of underlying relationships that exist between the datatypes, performance, in terms of speed, is sometimes compromised. During development of this system, time was crucial hence, coding standards did not always conform to best practices. This may hamper extensions to the system. Also, not not much emphasis was placed on proper documentation throughout the project and again, this could become problematic when extensions to the system need to be designed. It is hoped that these, and other shortcomings will be addressed as further development of the project continues. Further development of the system will see the addition of new functionalities. In terms of new functionalities, perhaps a high-throughput analyses pipeline may prove a useful addition as well as a full micro-array analyses module.

Chapter 3

The Seqword Genome Browser

A large part of the work in this chapter was published as:

Ganesan H., Rakitianskaia AS., Davenport CF., Tümmler B., Reva ON. (2008) The SeqWord Genome Browser: an online tool for the identification and visualization of atypical regions of bacterial genomes through oligonucleotide usage. BMC *Bioinformatics* **7**, p333.

3.1 Introduction

The greater FunGIM system includes the Seqword Genome Browser (SWGB) which is a subproject of the Seqword research team. The SWGB aims to visualize whole bacterial genomes on the level of oligonucleotide usage (OU) statistics thus providing a novel method of genome visualization. The specifics of OU statistics and the SWGB will be discussed.

3.2 Background

The study of genome OU signatures has a long history dating back to early publications by Karlin *et al.*, who focused mainly on dinucleotide compositional biases and their evolutionary implications (Karlin & Burge, 1995; Karlin, 1995; Pride *et al.*, 2003). Statistical approaches of OU comparison were further advanced by Deschavanne *et al.* (1999) who applied chaos game algorithms and by Pride *et al.* (2003) who extended the analysis to tetranucleotides using Markov Chain Model simulations. Later, a number of practical tools for phylogenetic comparison of bacterial genomes (Deschavanne *et al.*, 1999; Coenye & Vandamme, 2004; van Passel *et al.*, 2006), identification of horizontally transferred genomic islands (Mrazek & Karlin, 1999; ; Becq *et al.*, 2007; Dufraigne *et al.*, 2005; Nakamura *et al.*, 2004; Pride & Blaser, 2002) and assignment of unknown genomic sequences (Abe *et al.*, 2003; Teelin *et al.*, 2004) based on OU statistics became

publicly available. These approaches exploited the notion that genomic OU composition was less variable within genomes rather than between them, regardless of which genomic regions had been taken into consideration (Jernigan & Baran, 2002). A general belief was that if a significant compositional difference was discovered in genomic fragments relative to the core genome, these loci most likely can be assigned to horizontally transferred genetic elements (transposons, prophages or integrated plasmids). This approach was criticized by several researchers (Koski *et al.*, 2001; Wang, 2001) who pointed out that codon bias and base composition are poor indicators of horizontal gene transfer. Therefore, there is a need for more informative parameters which also take into account higher order DNA variation. An overview of the current OU statistical methods based on di-, tetra- and hexanucleotides has been published recently. The conclusion of the review was that all methods were context dependent and, though being efficient and powerful, none of them were superior in all applications (Bohlin, 2008). Thus, the major motivation in this work was to develop more flexible and informative algorithms seamlessly integrating di- to heptanucleotides OU analysis for reliable identification of divergent genomic regions.

Recently the concept of OU patterns was introduced into the literature (Reva & Tummler, 2004). Each OU pattern is characterized by a number of OU statistical parameters namely, local pattern deviation (D), pattern skew (PS), relative variance (RV) and others (see Methods section). Novelties of the developed algorithms relative to other existing methods include the following: i) distances between patterns of different word length (from di- through to heptanucleotides) calculated for the same sequences are comparable; i.e. one may use longer word patterns to perform a large scale analysis and then switch to shorter word patterns for a more detailed view; ii) OU patterns calculated for sequences of different lengths are comparable provided that the length of the sequence is longer than the corresponding thresholds (specified in the Methods section); iii) alterations of OU patterns may be analyzed by different non-redundant parameters (D, PS and RV with different schemes of normalization by frequencies of shorter constituent words). Superimposition of these OU characteristics allows better discrimination of divergent genomic regions relative to other contemporary approaches (Reva & Tummler, 2005). This is described by:

$$\Delta_{[\xi 1 \ldots \xi N]} = (C_{[\xi 1 \ldots \xi N]|obs} - C_{[\xi 1 \ldots \xi N]|e})/C_{[\xi 1 \ldots \xi N]|0}$$

where ξn is any nucleotide A, T, G or C in the N-long word; $C_{[\xi 1 \ldots \xi N]|obs}$ is the observed count of the word $[\xi 1 \ldots \xi N]$; $C_{[\xi 1 \ldots \xi N]|e}$ is the expected count and $C_{[\xi 1 \ldots \xi N]|o}$ is a standard count estimated from the assumption of an equal distribution of words in the sequence: ($C_{[\xi 1 \ldots \xi N]|o} = L_{seq} \times 4^{-N}$).

Expected counts of words $C_{[\xi 1 \ldots \xi N]|e}$ were calculated in accordance with the applied normalization scheme. Thus, $C_{[\xi 1 \ldots \xi N]|e} = C_{[\xi 1 \ldots \xi N]|0}$ if OU is not normalized, or $C_{[\xi 1 \ldots \xi N]|e} = C_{[\xi 1 \ldots \xi N]|n}$ if OU is normalized by empirical frequencies of all shorter words of the length n. The expected count of a word $C_{[\xi 1 \ldots \xi N]|e}$ of length N in a L_{seq} long sequence normalized by frequencies of n-mers (n <N) was calculated as follows:

$$C_{[\xi_1..\xi_{i_w}]n} = L_{seq} \times F_{[\xi_1..\xi_n]} \times \prod_{i=2}^{N-n+1} \left(\frac{F_{[\xi_i..\xi_{i+n-1}]\xi_{i+n}}}{\sum_{\xi}^{A,T,G,C} F_{[\xi_i..\xi_{i+n}]\xi}} \right)$$

where the $F_{[\xi 1...\xi N]}$ values are the observed frequencies of the particular word of length n in the sequence andξ is any nucleotide A, T, G or C. For example, expected count of a word ATGC in a sequence of L_{seq} nucleotides normalized by frequencies of trinucleotides is:

$$C_{ATGC} = L_{seq} \times F_{ATG} \times \frac{F_{TGC}}{F_{TGA}+F_{TGT}+F_{TGG}+F_{TGC}}$$

Two approaches of normalization have been exploited where the F values were calculated for the complete sequence of a chromosome, plasmid, etc (generalized normalization) or for a given sliding window (local normalization). The normalization by equation 2 allows identification of words, frequencies of which cannot be predicted exactly by frequencies of shorter constituent words.

The distance D between two patterns was calculated as the sum of absolute distances between ranks of identical words (w, in a total 4^N different words) after ordering of words by $\Delta_{[\xi 1...\xi N]}$ values (see equation 1) in patterns i and j as follows:

$$D(\%) = 100 \times \frac{\sum_{w}^{4^N} \left| rank_{w,i} - rank_{w,j} \right| - D_{min}}{D_{max} - D_{min}}$$

Application of ranks instead of relative oligonucleotide frequency statistics made the comparison of OU patterns less biased to the sequence length provided that the sequences are longer than the limits of 0.3, 1.2, 5, 18.5, 74 and 295 kbp for di-, tri-, tetra-, penta-, hexa- and heptanucleotides, respectively (Reva & Tummler, 2004)

PS is a particular case of D where patterns i and j were calculated for the same DNA but for direct and reversed strands, respectively.$D_{max} = 4^N \times (4^N - 1)/2$ and $D_{min} = 0$ when calculating a D or, in a case of PS calculation, $D_{min} = 4^N$ if N is an odd number or $D_{min} = 4^N - 2^N$ if N is an even number due to presence of palindromic words (Reva & Tummler, 2004) Normalization of D-values by Dmax ensures that the distances between two sequences are comparable regardless of the word length of OU patterns.

Relative variance of an OU pattern was calculated by the following equation:

$$RV = \frac{\sum_{w}^{4^N} \Delta_w^2}{\left(4^N - 1\right)\sigma_0^2}$$

where N is word length; Δ_w^2 is the square of a word w count deviation (see equation 1); and σ_0^2 is the expected variance of the word distribution in a randomly generated sequence that depends on the sequence length and the word length:

$$\sigma_0^2 = 0.14 + \frac{4^N}{L_{seq}}$$

where L_{seq} is sequence length, and N is word length. Normalization of OU pattern variance by σ_0 makes the variances comparable regardless of the word length of OU patterns and the sequence length. The regression equation was tested on 300 randomly generated sequences with an equiprobable occurrence of all 4 nucleotides by the DataFit 7.1.44 software. The SWGB is coded in Java to be used as an applet in a Web-browser either on the Internet or locally (the programs OligoWords in Python and SeqWord_Viewer, which respectively calculate and visualize the OU patterns for DNA sequences, are available for download from the SWGB website). SWGB should run on any platform with a Java 1.5.x runtime environment or newer. The pre-calculated data-sets are saved in a MySQL Server 5.0 database. The size of the sliding window and the OU pattern type were applied according to the sequence length (Table 3.1) At the time of writing, the SeqWord database contained OU patterns pre-calculated for the sequences of 682 bacterial chromosomes belonging to 637 different organisms (strains and species), 412 plasmids, 100 bacteriophages and 39 other viruses, which were downloaded from the NCBI (14).

Table 3.1: Sliding window size and OU pattern types (oligomer lengths) selected for sequences of different length present in the SeqWord database.

Sequence length	Sliding window	Step	OU pattern type
> 2 Mbp	8 kbp	2 kbp	4 mer
from 1 mbp to 2 Mbp	5 kbp	0.5 kbp	4 mer
from 0.5 mbp to 1 Mbp	3 kbp	0.3 kbp	3 mer
< 0.5 Mbp	1.5 kbp	0.15 kbp	3 mer

3.3 Results

User familiarity with the abbreviations of the various OU statistical parameters is important. Different types of OU patterns were abbreviated as type_Nmer. Types might be "n0" for non-normalized, or "n1" for normalized by mononucleotide frequencies. For example, the non-normalized tetranucleotide usage pattern is denoted as n0_4mer; tetranucleotide usage pattern normalized by mononucleotide content is n1_4mer etc. The genomes in the SWGB database were

analyzed by the following statistical parameters: D – distance between two patterns of the same type (in this work we used distances (D) between local patterns calculated for overlapping genome fragments and the global genome patterns calculated for the complete sequence – the local pattern deviation); PS – pattern skew, distance between the two patterns of the direct and reverse strands of the same DNA sequence; RV and GRV – oligonucleotide usage variances normalized locally and globally, respectively, and reduced to the OU variance expected for a randomly generated sequence (see Background section); GC-content (GC) and GC-skew (GCS) in DNA fragments. The SeqWord Genome Browser (SWGB) applet is available via the Internet through mirror sites (University of Pretoria, South Africa [http://www.bi.up.ac.za/SeqWord/mhhapplet.php]; Hannover Medical School, Germany [http://genomics1.mh-hannover.de/seqword/genomebrowser/mhhapplet.php]; Penn State University, USA [http://seqword.bx.psu.edu/mhhapplet.php]) and is mouse and menu driven. The Web-based applet is used to visualize DNA compositional variations in bacterial and viral genomes stored in the SeqWord database. Every genome in the database is represented by a set of statistical OU parameters (D, PS, GV, GRV, GC and GCS) calculated for genomic fragments, which were selected by a sliding window (sliding window length and step were set according to the total length of the sequence as demonstrated in Table 3.1). While in 70 to 99% of genomic fragments the OU compositional bias is similar to the complete genome OU pattern, some regions with atypical OU composition, however, are always present. Superimposition of different OU parameters allows discrimination of divergent genomic regions, as was published previously (Reva & Tummler, 2005). Briefly: rRNA operons are characterized by extremely high PS and low RV; giant genes with multiple repeated elements have high or moderate PS and high RV; horizontally transferred genetic elements are characterized by increased divergence between RV and GRV accompanied by high D; and genes for ribosomal proteins show a moderate increase of D, PS and RV above genomic averages. Having analyzed 1243 sequences of different microorganisms including viruses and plasmids in the SeqWord database, it was confirmed that the approaches developed and tested previously (Reva & Tummler, 2008) (mainly on Pseudomonas putida KT2440 chromosomal DNA) are appropriate and useful for analysis of genomic sequences of other microorganisms and viruses. In an open applet window, the user has the ability to choose from an ever growing list of available sequences (Figure 3.1) The user also has the option of restricting the list to display only bacterial chromosomes, plasmids, phages, viruses or all sequences by selecting the corresponding filter button. Users have to select a genome in the list and click the 'Display in the Applet' button to retrieve the pre-calculated data. All OU parameters calculated for a given genome may be exported to a local text file by using the 'Export' function from the applet's 'File' menu. Later, instead of again having to connect to the database, users may open and view their local files (previously exported from the applet or calculated by the OligoWords program, see below) via the 'Open' function in the 'File' menu.

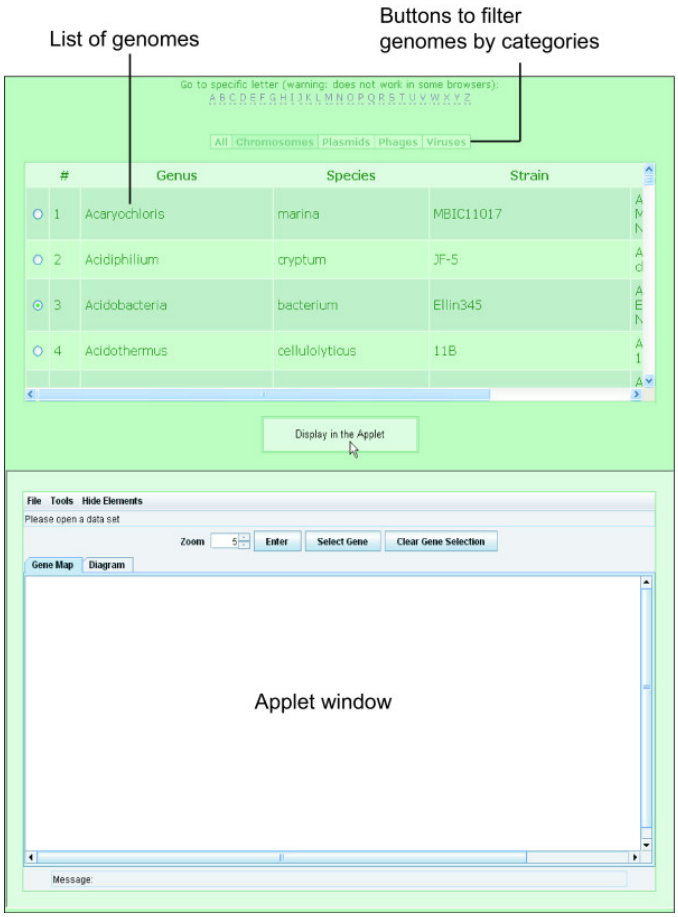


Figure 3.1: General view of the web-based SWGB with a list of genomes present in the database and an enclosed Java applet for data visualization. To show OU statistical parameters for a selected genome, click the 'Display in the Applet' button. Click a filter button to order genomes by the corresponding category and use the interactive letters at the top to scroll the list to a sequence of interest.

The SWGB is basically comprised of two views, denoted by the 'Gene Map' and 'Diagram' tabs. The applet is instrumental for visualization of natural variation in DNA sequences by the interactive diagrams on the 'Gene Map' and 'Diagram' tabs. Users may save the current diagram in JPG format by using the 'Save picture' function in the 'File' menu. The 'Gene Map' tab offers a simple view of an entire genome at a glance and gives users access to a number of important pre-calculated OU statistics superimposed on the gene map (Figure 3.2) Displays for each of the statistical parameters can be toggled on/off by checking items in the 'Hide Elements' menu. By merely mousing over any region on the plot, a message displaying detailed information for the pointed curve will be shown in the 'Message' bar. Clicking a gene on the map displays a dialog with the annotation details (Figure 3.2).

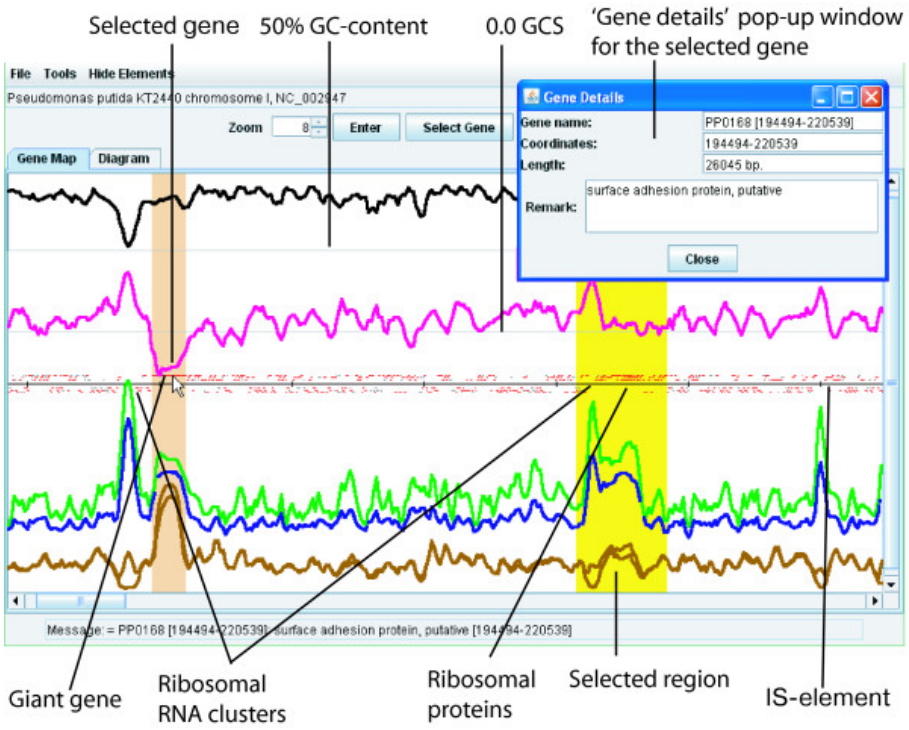


Figure 3.2: Identification of divergent genomic regions on the 'Gene Map' view. Superimposition of different OU parameters such as GC (black line), GCS (pink), PS (green), D (blue), GRV (upper brown line) and RV (lower brown line) allows discrimination of divergent genomic regions. In this example a part of the chromosome of Pseudomonas putida KT2440 (127–774 kbp) is displayed in the applet window. A genomic fragment was highlighted using the function 'Select region' and a giant gene, PP0168, was selected by 'Select gene'. A pop-up window 'Gene Details' was opened by double-clicking the gene on the map. Genes are indicated by red and grey (for hypotheticals) bars. The black horizontal line separates genes by their direction of translation.

The 'Zoom' function is straight-forward and allows users to control the amount of data viewed in the plot area. Clicking the 'Enter' button after setting the desired zoom value will then redraw the map. A 'Zoom into region' function under the 'Tools' drop-down menu allows users to zoom into exact genomic regions by merely entering their desired co-ordinates into the pop-up dialog box. The 'Tools' → 'Select region' menu item allows highlighting of selected regions without zooming. Use the option 'Clear ...' in the 'Tools' menu to undo zooming or highlighting. To locate a genomic region by gene, click the button 'Select Gene'. In the pop-up dialog box one may order the gene list by gene names, functionality or coordinates, then select a gene in the list and click 'OK'. When a gene annotation is not available, the values of the locus coordinates are used as a gene name. The applet window will be scrolled to the selected gene highlighted on the map (see Figure 3.2).

The 'Diagram' tab allows flexible filtering of the underlying data based on the criteria chosen by users. Although the underlying data is pre-calculated, the user may, by simply changing selected parameters, generate very different images which give different insights into the natural genomic variation. To start with, the 'Diagram' view offers a bar chart or a dot-plot presentation of the pre-calculated data. To view a bar chart of the distribution statistics for a given OU parameter, select the desired parameters from the X or Y-axis drop-downs and click 'Enter'. The number of bars displayed can be adjusted using the '# Bars' selector.

On the dot-plot diagram, each genomic fragment (selected by the sliding window) is represented by a dot with X and Y coordinates that correspond to values of OU parameters chosen from X and Y drop-down lists, respectively. The Z axis parameter may be set as well. In this case, the dots are colored by values of OU parameters selected for the Z axis, and the color range is displayed on the vertical color bar on the left of the plot area (Figure 3.3).

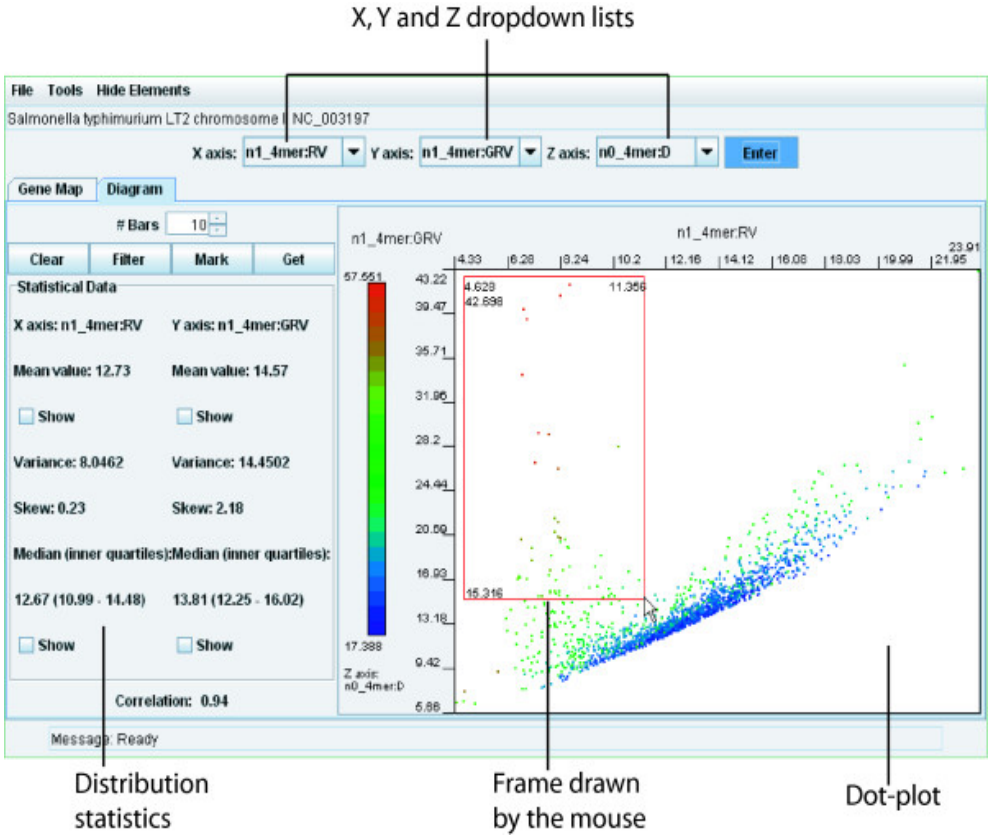


Figure 3.3: The 'Diagram' view. To draw a diagram, first select corresponding OU parameters using the dropdown lists and click the 'Enter' button. In this example n1_4mer:RV, n1_4mer:GRV and n0_4mer:D were selected for the X, Y and Z axes, respectively. Every dot on the dot-plot corresponds to a genomic fragment selected by the sliding window. Dots are spread and colored in accordance with their values of the selected statistical OU parameters. Information for each dot may be found by one of the following methods: i) information for a dot under the mouse pointed by the mouse is shown in the 'Message' bar; ii) double clicking a dot returns us to the 'Gene map' tab with the corresponding genomic fragment highlighted; iii) framing the dots and clicking the 'Get' button opens a new applet window with the information about all selected regions. In this example the genomic regions of Salmonella typhimurium LT2 (NC_003197) that correspond to horizontally transferred genetic elements were selected (see discussion in the text).

Having set up the dot-plot, users will be able to identify divergent genomic regions (see next section). To retrieve annotations of genomic fragments corresponding to a group of dots, frame the dots of interest by clicking and dragging over the desired area. A selector frame then appears around the dots (Figure 3.3). Clicking the 'Get' button displays the selected genomic fragments with their coordinates and gene annotations. Furthermore, identification and isolation of specific genomic regions may be improved significantly by filtering dots by OU parameters. The simplest way of filtering is by the third (Z axis) parameter. One may select an area on the color bar to exclude all dots from the plot lying outside of the selected color range (see an example in help

files on-line). The hidden dots will not be selected by the 'Get' button. A more sophisticated way to filter genomic regions is provided by the 'Filter' button. An example will be discussed below. The 'Mark' button enables genomic fragments to be selected by their coordinates and highlighted on the dot-plot. Click the 'Mark' button to open a dialog and enter coordinates of one or multiple fragments (Figure 3.4). Co-ordinates of each fragment must be added to the list by clicking the 'Add' button. Close the dialog by clicking 'OK'. The corresponding dots on the dot-plot will be highlighted as shown in Figure.3.4.€

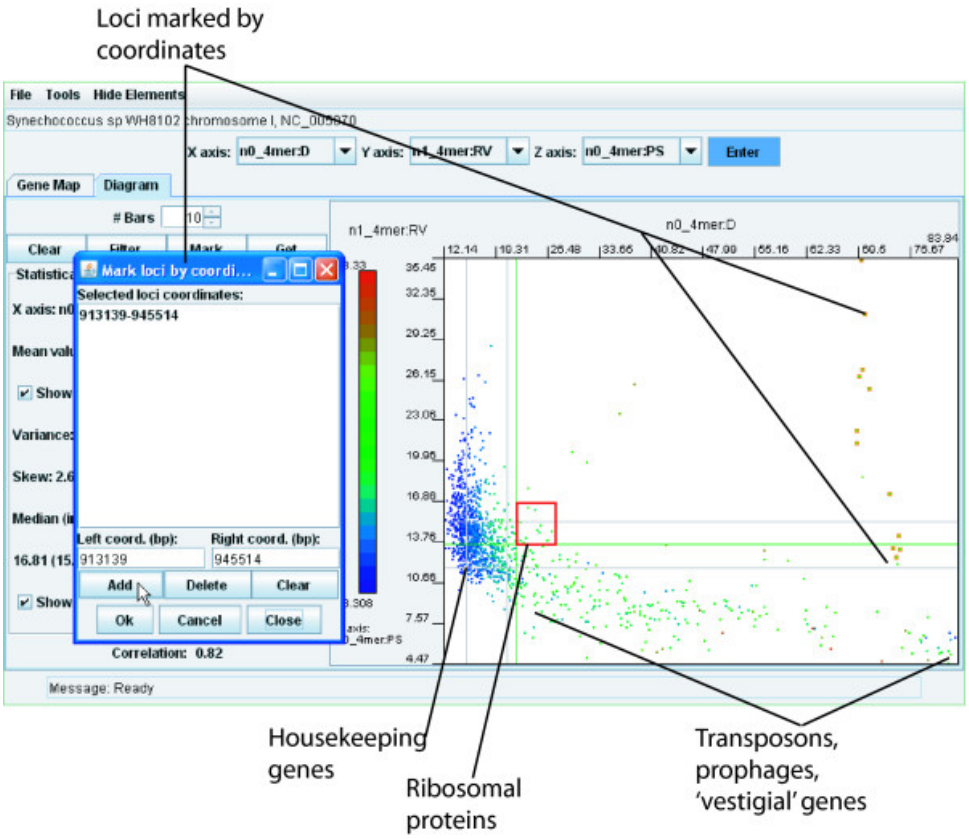


Figure 3.4: Identification of divergent genomic regions by plotting and highlighting. In this example the genome of Synechococcus sp. WH8102 was analyzed. The parameters n0_4mer:D, n1_4mer:RV and n0_4mer:PS were selected for the X, Y and Z axes, respectively. The genomic regions covering the giant gene for the surface protein SwmB (Reva & Tummler, 2008) were highlighted by entering the coordinates of this gene into the 'Mark loci by coordinates' dialog. The genomic regions enriched with i) housekeeping genes; ii) genes for ribosomal proteins; iii) vestigial genetic elements (comprising pseudogenes, transposons, prophages and IS-elements) are indicated.

3.4 Identification of divergent genomic islands

Several routines have been developed to identify the horizontally transferred genomic islands, genes for ribosomal RNA and proteins, non-functional pseudogenes and genes of other functional categories. All these routines are described in detail with illustrations in supplementary web-pages (use the 'Help' link in the applet window). The approach to identify inserts of foreign genomic elements by OU statistical parameters have been described recently (Reva & Tummler, 2005). While several algorithms allow identification of horizontally transferred genomic islands (Mrazek & Karlin, 1999; Azad & Lawrence, 2005; Becq *et al.*, 2008; Dufraigne *et al.*, 2005; Nakamura *et al.*, 2004; Pride & Blaser, 2002), the multiple oligomer parameters used in the SWGB even allows tentative attribution of genomic fragments (and, given the right scale, genes or gene clusters) to different functional classes using only a FASTA sequence as input. However, the emphasis of the SWGB is not primarily its annotation capability, but its ability to display the natural internal variability of genome sequences. *Pseudomonas putida* KT2440 was used, which is a known mosaic genome with 105 genomic islands above 4000 bp in length (Weinel *et al.*, 2004) as an example. Many of these features can be visualized at a glance using the SWGB without any in depth analysis (see Figure 3.2). On the 'Diagram' view the parameters n1_4mer:RV, n1_4mer:GRV and n0_4mer:D were selected for the X, Y and Z axes, respectively, as shown previously (see Figure 3.3). Plotting local relative oligomer variance (RV) against global relative variance (GRV) basically shows the effect of normalization by global mononucleotide content. The core genome is then represented on the dot plot as the positive linear correlation line where RV $\approx$ GRV (Figure 3.3). In other words, these fragments exhibit such compositional closeness to the core genome that normalizing by local mononucleotide content does not have a different effect compared to normalizing by global content. These genomic fragments also exhibit a low distance from the genomic average; and are therefore colored blue. Scattered dots lying peripheral to the expected strong linear correlation do not belong to the core genome and also have a higher distance from the genomic average and are hence colored green. Using the filter settings recommended in Figure 3.5, twenty one fragments were found to be genomic islands (note that while border values of OU parameters are not the same for different genomes, the grading notches of the sliders represent relative values that allows identification of homologous regions in many different genomes). For a number of reasons, many more islands were found in a similar analysis by Weinel *et al.*, (2004) Firstly, the sliding window size of 8 kbp means many of the 4 kbp features from their analysis were not identified automatically. Furthermore, they were looking for all compositionally atypical regions, whereas here, restriction was made to horizontally transferred regions.

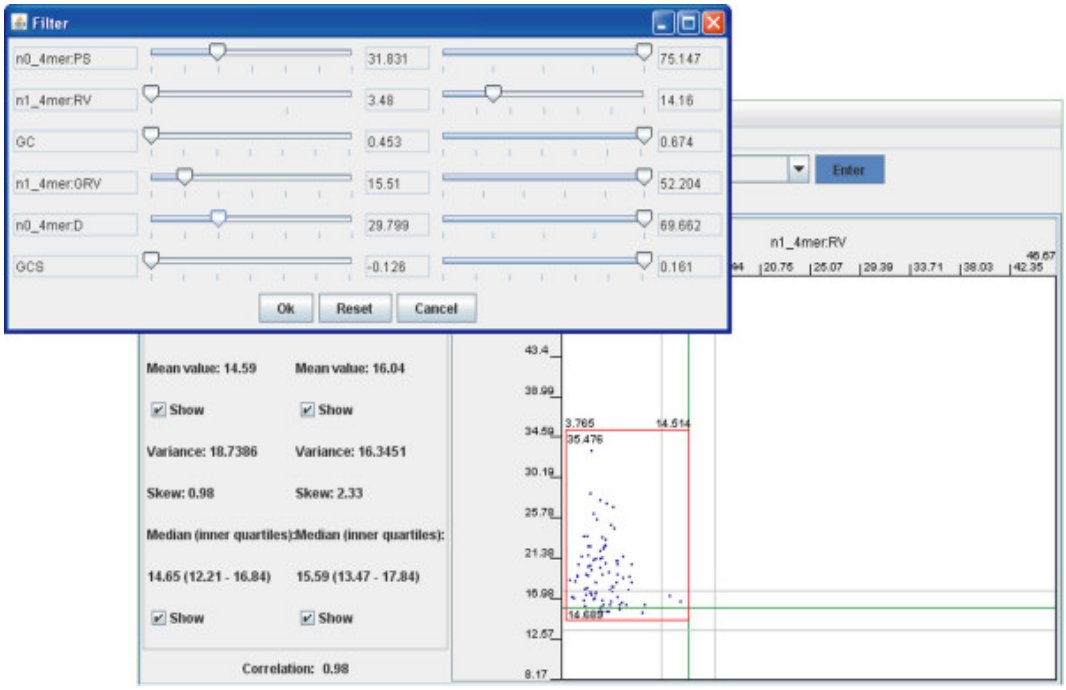


Figure 3.5: Filtering genomic regions by multiple parameters. Click the 'Filter' button to open a dialog as shown in the figure. Setting up border values of multiple OU statistical parameters allows more precise localization of regions of interest.

A known 40 kbp bacteriophage insertion [2586000–2626000] is, surprisingly, not among the genomic fragments selected in the SWGB using this filter. Although the prophage is still perceptible on the 'Gene Map' view (see a figure in the supplementary help web-pages), the OU parameters of the region do not differ markedly enough from the core sequence to be isolated automatically as a horizontally transferred region.

As the SWGB uses parameters that are based on comparison of local fragments to the global genomic average, strains with abundant insertions of homogenous DNA can confound this form of analysis. One example is the Methanosarcina acetivorans C2A genome which is composed of an estimated 25% of putatively horizontally acquired DNA, one of the highest amounts discovered to date (Dufraigne *et al.*, 2005). As a result of these insertions, the genomic signature has been strongly influenced, resulting in a large amount of scatter and a poorly defined core genome on the plots. On the other hand, this type of analysis allows estimation of genome stability in a simple, multi parameter view (see the *Vibrio cholerae* N16961-O1-eltor example in the online help files). To conclude, filtering provides a convenient way to automatically isolate divergent genomic regions of interest. However, some regions may erroneously remain undetected due to possible amelioration of older inserts (Lawrence & Ochman, 1997) or a higher level of noise in unstable genomes. However, many problematic genomic fragments can in some cases be easily attributed to functional gene categories using the SWGB 'Diagram' window (see Figure 3.2).

Methodologies for discovering long modular genes have already been discussed in a previous publication (Reva & Tummler, 2008). Briefly, long genes display a particular tetranucleotide usage and can be discovered by plotting n0_4mer:D (X axis) versus n1_4mer:RV (Y axis). The positively linear correlated outlier fragments (towards the top right of the image) are often fragments of long genes with their characteristic repeats. An example using the gene encoding the 1.12 megadalton cell surface protein of *Synechococcus sp.* WH8102 (McCarren & Brahamsha, 2007) marked on the dot-plot is shown in Figure. 3.4. Ribosomal RNA operons (but not genes for ribosomal proteins) are characterized by extremely high pattern skew and a large distance from the core genome (Figure 3.2). Thus, there is a tendency to find many genomic fragments containing rRNA genes colored dark brown to red in the bottom right section of the 'Diagram' tab. The annotation for rRNA operons is not present in the database; therefore, these are seen in the 'Gene Map' tab as un-annotated areas with high pattern skew (Figure 3.2). Ribosomal proteins tend to be increasingly present at a slightly greater than average RV and above average D (see Figure 3.2), which is in agreement with observations that highly expressed genes for ribosomal proteins have a highly specific codon usage compared to housekeeping genes of the organism (Puigbo *et al.*, 2008).The majority of genomic fragments form a cluster characterized by average and higher than average RV, stable OU patterns (low D) and low PS. These tend to be the core, or bulk genes and genomic regions with their typical tetranucleotide usage. Some other core sequence fragments spread from this area toward lower RV and less specific OU patterns (higher D and PS) – these are all characteristics of an unstable or randomly generated sequence (Reva & Tummler, 2004). These regions were found to be enriched with many hypothetical genes, prophages and transposons (the data is not shown but is easily verified with any genome using the 'Get' button. Consider, for example, this area in the pseudogene rich *Mycobacterium leprae* TN or *Methanosarcina acetivorans* C2A genomes (Dufraigne *et al.*, 2005; Klockgether *et al.*, 2006) and the relatively homogenous *Alcanivorax borkumensis* SK2 genome (Reva *et al.*, 2008). These regions were thus categorized as rich in 'vestigial' genes in contrast to the core genome regions rich in housekeeping genes (Figure 3.4).

It must be stressed that with an average length of genes being around 1 kbp and overlapping sliding windows of 8 kbp, one cannot expect precise separation of housekeeping and vestigial genes by the method described above. However, when analyzing an unknown DNA sequence prior to annotation, it may be helpful to identify genomic regions enriched with a higher proportion of these so called housekeeping genes and other regions rich in vestigial genes. These tentative results should be verified with other complementary algorithms such as BLAST, gene finding and annotation techniques.

The most important feature of the supplemented software available from the SWGB web-server for download is the ability to quickly and easily analyze a novel sequence on a local computer. The command-line Python program OligoWords is first used to analyze FASTA or GenBank formatted sequences. The program is available for download in several packages as precompiled executable files and as Python source code. The command-line interface of the

OligoWords program is shown in Figure 3.6. Parameters such as oligomer length and window size can all be set depending on the sequence length and desired resolution (see Table 3.2 for suggestions). Since the SWGB is implemented as a Java applet, it can be run within a web browser locally. The HTML-embedded applet is available for download from the same FTP site (15) (select SeqWord_Viewer.zip). The output file from OligoWords is read into the SWGB via the 'Open' function of the 'File' menu, and the complete functionality of the online system is then available. For example, a new sequence can be analyzed for ribosomal gene clusters, putative horizontally transferred elements or other regions of atypical DNA structure prior to the lengthy annotation step. A complete description of how to run the SWGB and OligoWords locally is presented in the online help files.

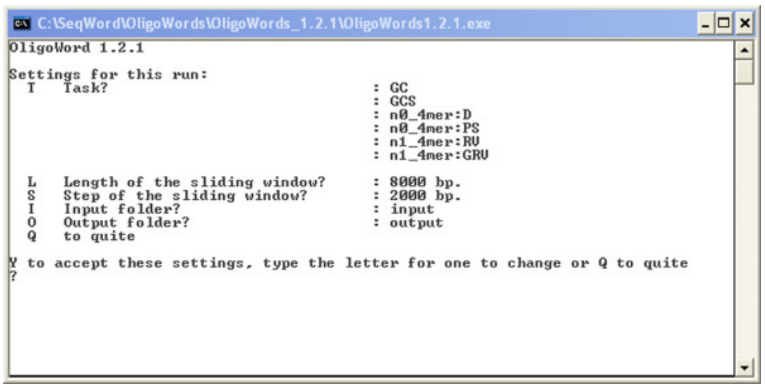


Figure 3.6: Command-line interface of the OligoWords program. To change the setting for the current run, type the option's letter and enter a new value as prompted. Users may change: T) the set of statistical OU parameters to be calculated for every local pattern; L) length of the sliding window; S) step of the sliding window; I) the name of the input folder that contains FASTA and/or GenBank files with source DNA sequences; and O) the name of the output folder where the result files will be stored.

3.5 Scientific Investigation – Application to mycobacteria

In this section, the results of a comparative study of mycobacterial genomes by using SWGB routines described above will be shown and described. The aim of this study was to identify a few of the main sub-genomic components contributing toward genetic variation seen among mycobacteria. With the help of literature, we also aim to gauge these identified components contribution toward virulence of the organism.

Firstly, few mycobacterial genomes were analysed for the presence of horizontally transferred genetic elements (HTGE). RV (X-axis), GRV (Y-axis) and D (Z-axis) gene diagram plots were generated for *M. tb* H37Rv (Figure 3.7). The expected area on the plot where fragments of horizontally transferred genetic elements were expected to appear was empty (see Figure 3.7 below). Similar plots lacking traces of HTGE were obtained for all other *M. tb* and *M. bovis*

genomes.

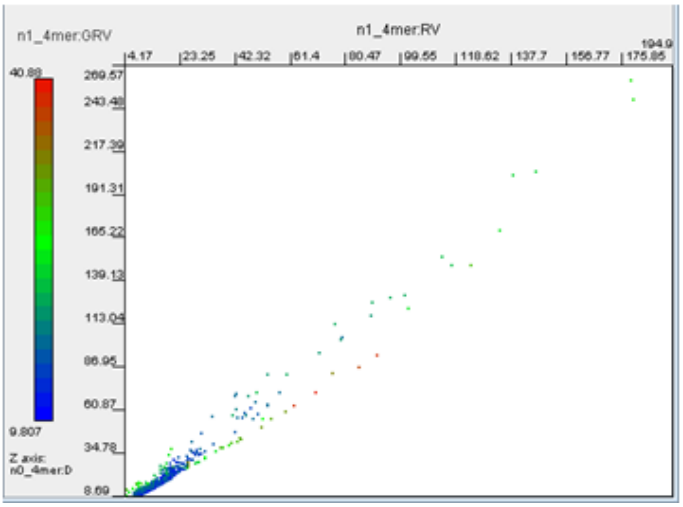


Figure 3.7: RV, GRV and D gene diagram plot for *M. tb* H37Rv.

On the contrary, the SWGB plot for *Mycobacterium avium* K10 (NC_002944) revealed many genomic regions of putatively lateral origin (Figure 3.8). The coordinates and annotation of these identified regions are shown in Table 3.2.

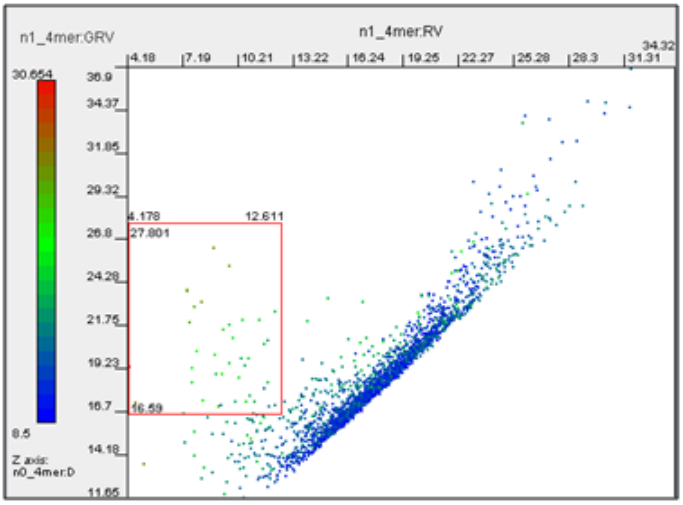


Figure 3.8: RV, GRV and D dot-plot generated for Mycobacterium avium K10.

The area of the distribution of horizontally transferred genomic elements is outlined in the figure above (Figure 3.8) and the coordinates of the outlined genes are shown in the table below.

Table 3.2: Coordinates and annotations of the gene islands in the genome of *M. avium* K10.

Left	Right	n1_4mer:RV	n1_4mer:GRV	n0_4mer:D	Annotations
78000	86000	11.6606	18.0853	13.7828	dnaB; mmpL4_1 and 4 genes for hypothetical proteins
870000	892000	8.6089	22.8127	22.0832	nramp and 20 hypothetical proteins
1290000	1304000	8.7803	20.6197	21.1895	lipL and a hypothetical gene

The RV, GRV, D plot above highlights several areas of possible horizontally transferred origin. The first genomic island (as outlined in the table above) is in the region 78000–86000 bp. In this region dnaB, mmpL4 and 4 genes of hypothetical proteins were found. Based on the RV, GRV distribution for this region, the 4 hypothetical proteins indeed exhibit atypical oligonucleotide usage relative to the rest of the genome. dnaB however, is found very proximal to this genomic island but is not necessarily horizontally transferred. This is the same situation with the next genomic island (870000 – 892000 bp) and nramp, where nramp lies proximal to the terminal regions of the genomic islands. If OU parameters are examined for this genomic region the following is seen.

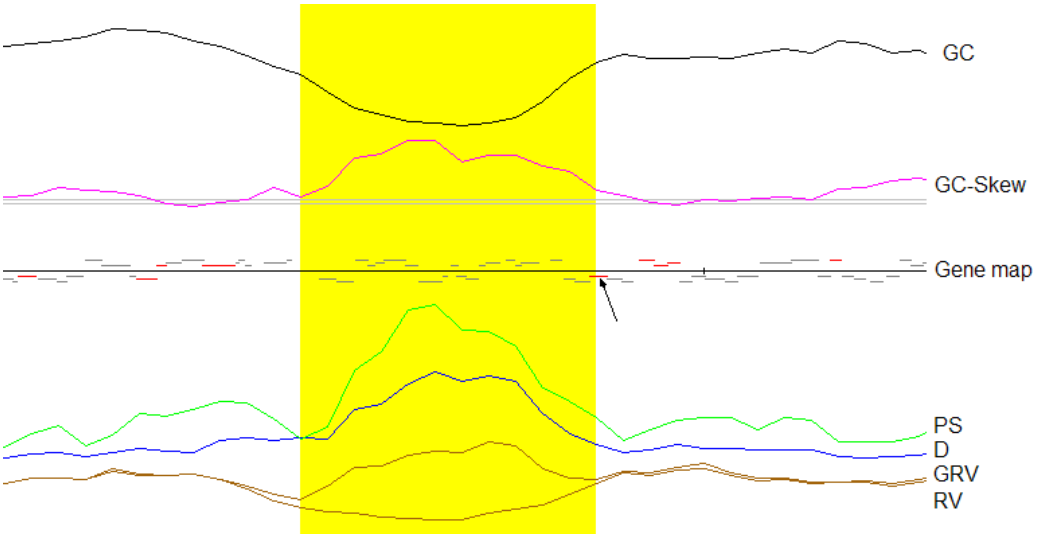


Figure 3.9: SWGB view for genomic region 87000-892000 (highlighted). An arrow marks nramp (in red) on the border of the highlighted region.

The above view of the genomic region reveals typical attributes for a horizontally transferred region. RV (the bottom most line) represents oligonucleotide usage normalized by mononucleotide content for the local pattern. GRV (above RV) represents oligonucleotide usage normalized by mononucleotide content for the whole genome. When regions (such as this) contain

atypical oligonucleotide usage patterns, GRV and RV diverge from each other as evident in the above figure. Nramp lies on the border of the genomic island is its horizontal origin should be checked separately.

Using different plot parameters (RV, PS and GC) on *M. tb* H37Rv revealed some regions of atypical and quite unusual OU.

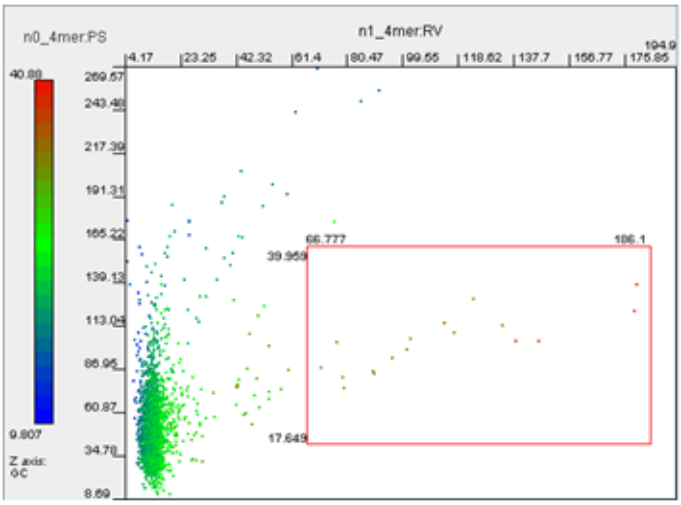


Figure 3.10: RV, PS and GC gene diagram plot for *M. tb* H37Rv.

Note that the core genomic elements form a dense cloud of dots with two jets of outliers directed rightward and upward relative to the core genome sequence. RV increase in local OU patterns shows an increased compositional bias in oligonucleotide frequencies that may correlate with the codon usage bias in coding sequences. A simultaneous increase of PS and RV parameters usually imply multiple tandem repeats in these genomic regions (Reva and Tummler 2008)

Annotations for the outlined genomic fragments for the *M. tb* H37Rv plot above (Figure 3.10) is shown in Table 3.3 below.

Table 3.3: Coordinates and annotations of the gene islands in the genome of *M. tb* H37Rv.

LEFT	RIGHT	n1_4mer:RV	n0_4mer:PS	GC	ANNOTATION
332000	342000	75.5	25.7	73%	PE-PGRS and PPE family proteins
1630000	1638000	77.1	29.1	75%	PE-PGRS and PPE family proteins
3734000	3746000	120.5	31.5	73%	PE-PGRS and PPE family proteins
3924000	3954000	121.6	29.1	76%	PE-PGRS and PPE FAMILY PROTEINS; acyl-CoA synthase; acyl-CoA dehydrogenase; acyl-CoA lygase FADD18; enoyl-CoA hydratase;thiamine-pyrophosphate requiring enzyme and many hypotheticals

This fragment selection from the RV, PS and GC SWGB dot-plot screen (Figure 3.10) allows identification of PE-PGRS and PPE family proteins which in *M. tb* are indirectly associated with virulence (Zheng *et al.*, 2008).

These PE-PGRS genes acquired highly peculiar characteristics in the *M. tb*/*M. bovis* genomes as compared to *M. avium* which contains homologous genes that can not be distinguished from the core genome sequences by the method described above. The genomes of *M. ulcerans* and *M. marinum* on the other hand, show intermediate states of evolution from the pattern of *M. avium* toward that of *M. tb* (Figure 3.11)

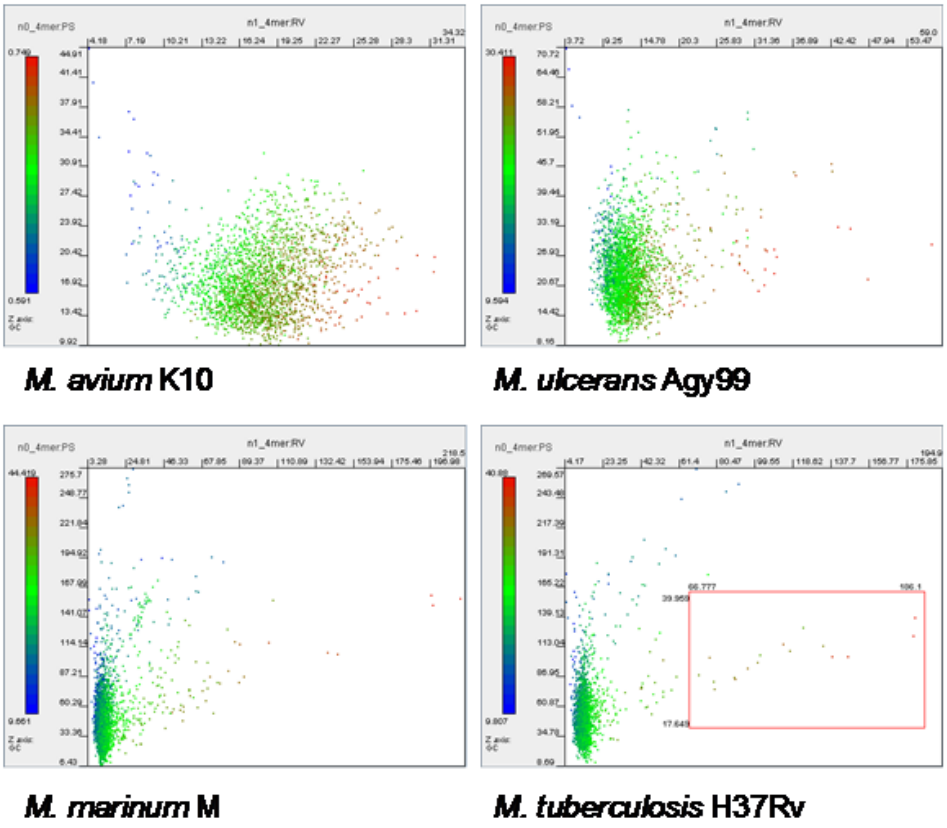


Figure 3.11: Global evolutionary changes in mycobacterial genomes as revealed by SWGB dot plots. Each dot corresponds to the calculated oligonucleotide usage pattern for an 8kb sliding window of step size 2kb.

Taking into account that the differences between 16S rRNA sequences of these genomes are less than 3%, the rate of evolutionary changes of the genes of PE_PGRS and PPE families is fascinating and yet poorly-studied (Harmsen *et al.*, 2003).

The rate of mutations in these genomic loci are several fold higher than the average rate of mutations per genome. In Figure 3.12 below, single nucleotide polymorphisms (SNPs) between the two closely related *M. tb* strains H37Rv and H37Ra is shown. Note the frequency of SNPs accompanied with genome rearrangements in a hypervariable locus toward the right of the figure.

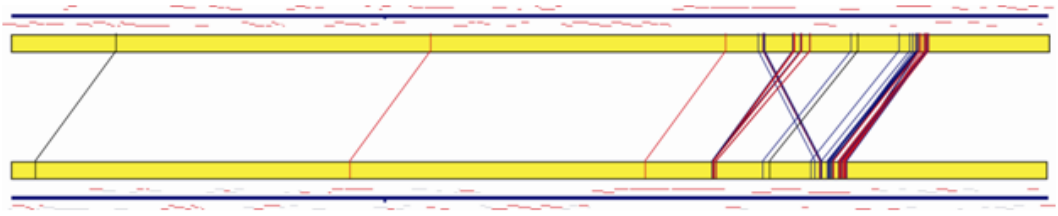


Figure 3.12: SNP distribution in homologous loci of *M. tb* H37Ra [3,800,000-4,000,000] (above) and *M. tb* H37Rv [3,850,000-4,000,000] (below) genomes. Transitions, transversions and deletions are depicted by blue, red and black connecting lines, respectively. Genes are depicted by red and grey (hypothetical) bars above (forward) and below (reverse) the blue lines according to the direction of their transcription.

Figure 3.11 above emphasizes the differences of these organisms by displaying the genomic fragments that have experienced evolutionary changes from M. avium lineage to that of *M. tb*. For *M. avium* the variability of the calculated parameters RV is found to be in the range 5-35 while PS found in the range 10-45. These ranges are considered homogenous as they are consistent among many other bacterial strains including *E. coli* and *B. subtilis*. On the contrary, some loci in *M. tb* exhibits extreme divergence for these calculated parameters. For example, having a RV up to 200 and PS up to 270. This signifies local mutational hotspots which have not been well studied thus far. In the following chapter the mycobacterial comparison project (MCP) will be used to further elucidate these regions of interest.

3.6 Discussion

In this chapter, the SWGB and its functionality was used to reveal differences among mycobacterial genomes that would have otherwise gone un-noticed. Based on sequence studies thus far, it is known that mycobacterial genomes are largely similar on the nucleotide level (Fleischmann *et al.*, 2002). Using conserved genes such as 16s RNA, dnaB and gyrB genes, it is possible to sometimes differentiate between various strains but not to a high level. Using our novel algorithmic approaches to the genomic analyses and novel display techniques (Ganesan *et al.*, 2008), however, we were able to not only identify regions of gross differences between some mycobacterial strains, but also identity specific genes and genomic islands that account for much of the sequence differences exhibited in these organisms. Several interesting genomic islands containing important genes were identified.

Several gene islands which had most likely been acquired by the lateral transfer but had been significantly ameliorated towards the OU pattern of the core genome sequence were identified in the *M. tb* H37Rv genome by an in-house program Gene Island Sniffer (12). Many virulence associated proteins were found in these former gene islands. The *M. tb* genome contains 13 genes encoding RND (resistance, nodulation & cell division) proteins designated, MmpL (Mycobacterial membrane protein Large). RND proteins are a family of multi-drug resistance pumps that

function to recognize and mediate the transport of a wide variety of cationic, anionic or neutral compounds such as various drugs, fatty acids, bile salts etc. (Domenech *et al.*, 2005). Although MmpL proteins play a role in drug resistance in certain bacteria, it was found not to be the case within mycobacteria however, MmpL4 mutants showed a decreased level of virulence in a low-dose aerosol murine model of infection. The study supports the concept that MmpL-mediated lipid secretion affects both the pathogens ability to survive intracellularly as well as host-pathogen dialogue which determines the ultimate outcome of infection (Domenech *et al.*, 2005; Rodriguez *et al.*, 2002). lipL was one of the other elements identified. lipL is a gene belonging to the hormone-sensitive lipase family and is responsible for fatty acid metabolism during an adverse nutrient climate (Deb *et al.*, 2006). This activity may account for the organisms utilization of stored triacylglycerols during dormancy and its subsequent reactivation.

Also identified were genomic islands in *M. tb* H37Rv rich in PE-PGRS genes, perhaps representing a gene cluster of some sort. Studies in the past have shown that virulent mycobacterial genes sometimes do appear clustered together in specific loci. Camacho *et al.*, (1999) identified a 50kb chromosomal region in *M. tb* which contained several virulence genes. The group created a library of signature-tagged transposon mutants of *M. tb* and then screened for those which had their ability to replicate within lung of mice negatively affected. The insertions for those mutants which had their virulence inhibited were then mapped onto the *M. tb* genome. Apart from the identification of the 'pathogenicity island' the group also noticed that most of the mutated loci seemed to be involved in lipid metabolism and transport across the membrane (Camacho *et al.*, 1999). Similarly, Danelishvili *et al.*, (2007) also conducted studies with some transposon mutant mycobacteria this time with the intention of identifying *M. avium* genes and host cell pathways involved in their uptake by macrophages. In the clones with impaired macrophage uptake, they revealed that 4 of the six genes examined, all lie within the same region of the chromosome. Analysis of this chromosome region revealed a pathogenicity island of 58% GC content (compared to 69% for the genome) inserted between two tRNA sequences. This region was also found to be unique to *M. avium* and absent in *M. tb* and *M. tb paratuberculosis*. Gene islands indeed play a role in the life-cycle and virulence of mycobacteria thus it is imperative that these regions can be accurately and efficiently identified.

In terms of the genomic islands (loci with atypical OU) found within *M. tb* H37Rv (Figure 3.10), it is seen that the PE-PGRS, PPE gene family are a dominant feature. What are these genes and what role do they play? Approximately 8% of the potential coding capacity of *M. tb* H37Rv was found to be accounted for by two unrelated gene families encoding the PE and PPE proteins (Banu *et al.*, 2002). The PE/PPE names are derived from the motifs Pro-Glu/Pro-Pro-Glu which in most cases are found near the N-terminus of these glycine and alanine rich proteins. The PE and PPE family comprises of about 100 and 68 members, respectively. The largest class of the PE family, having 67 members in *M. tb* H37Rv is referred to as the PE-PGRS sub-family. These proteins consist of the PE domain followed by C-terminal extension with multiple tandem repetitions of Gly-Gly-Ala or Gly-Gly-Asn encoded by the PGRS (polymorphic GC-rich repeti-

tive sequence) motif. PE-PGRS proteins may contain up to 1900 amino acids (based on predictive models), up to 50% of these can be glycine (Banu *et al.*, 2002). Implicit in the name, PGRS genes are GC rich and may be a major source of polymorphism, this then lead to the question of whether PE-PGRS proteins of *M. tb* variable surface antigens? Banu and group tested this hypothesis by raising antibodies in mice against 5 PE-PGRS proteins. These antibodies detected single proteins when the original plasmid constructs (used for immunization) were expressed in epithelial and reticulocyte extracts, thus confirming the proteins antigenicity. Furthermore, the antibodies cross reacted with several PE-PGRS proteins suggesting that different proteins share common epitopes. The group then went on to perform sub-cellular fractionation studies and immunoelectron microscopy which localized many PE-PGRS proteins in the cell wall and membrane of *M. tb*. Their findings further suggested that PE-PGRS proteins play a role in antigenic variability on the cell surface (Banu *et al.*, 2002; Okkels *et al.*, 2003). Similarly, in a comparative study of gene products of key metabolic pathways among 5 mycobacterial genomes (*M. tb*, *M. leprae*, *M. avium*, *M. bovis* and *M. avium ssp. Paratuberculosis* K10), it was shown that the major differences between these species is accounted for by gene products constituting the cell wall and gene families encoding the PE/PPE/PGRS proteins. What is interesting is that *M. avium ssp. Paratuberculosis* lacks PE-PGRS genes. This gene set is the very likely set of genes responsible for the survival of *M. tb* in macrophages, which then leads to the idea that *M. tb* and *M. avium ssp. Paratuberculosis* exhibit differences in the survival mechanisms of these species within macrophages (Marri *et al.*, 2006). There is also evidence that PE-PGRS may be important for bacterial survival during early stages of infection indirectly through alternative sigma factor (SigD) (Raman *et al.*, 2004) as well as provide resilience for a changing host micro-environment through differential expression (Voskuil *et al.*, 2007)

It has been shown how easily gross differences between mycobacterial genomes can be detected using our SWGB applet. Several gene islands have been identified and along with genes that account for the differences between the various mycobacterial species. In the following chapter, a closer look will be taken at these genes and gene families using the mycobacterial comparison project (MCP) system in an attempt to provide deeper understanding of these differences and the role they play in the evolution and virulence of mycobacteria.

Chapter 4

The Mycobacterial Comparison Project

4.1 Introduction

The possibility and usefulness of a central framework to incorporate data analyses and data storage facilities has been demonstrated. This was shown by example, using general bioinformatics analyses such as general sequence alignments, phylogenetic analyses and large scale genome alignment. What shall now be attempted is further building on this framework and applying it in the comparison of Mycobacterial genomes. The 'Mycobacterial Comparison Project' (MCP) has been implemented upon the built framework and is potentially useful for gene-by-gene comparison of the mycobacteria (though the analyses techniques employed here may be extended to any organism). Mycobacteria research in South Africa is an especially relevant area of research as tuberculosis is one of the leading infectious diseases and also responsible for millions of deaths globally (Fu & Fu-Lui, 2007). In South Africa, widespread emergence of multi and extreme drug resistant strains of mycobacteria further exacerbates the problem. Research into the biology of these micro-organisms is therefore very relevant and necessary if there is to be any hope of combating the deadly disease. In light of this, the Mycobacterial Comparison project was initiated in order to elucidate some of the biology of these organisms by means of comparative genomics. Due to the availability of many sequenced Mycobacterium genomes, a comparative genomics study using our FunGIMS-based framework as a backbone became possible. One aim of the whole project was to collect and process data from several key mycobacterial species and structure this data so that meaningful biological conclusions could be drawn. Details regarding the implementation and scientific aims will be discussed, first however, a brief overview of tuberculosis will be supplied, followed by an introduction to the Mycobacterial genome and the current state of comparative Mycobacterial genomics research.

4.2 Tuberculosis

Mycobacterium tuberculosis (*M. tb*) is the etiologic agent of tuberculosis which accounts for more deaths each year than any other disease caused by a single pathogen. Ninety-five percent of these cases are found in developing countries due to inadequacies in the healthcare resources and patient follow-ups (Nouvel *et al.*, 2006). Tuberculosis is predominantly spread by aerosol transmission whereby droplet nuclei often containing several bacilli (particle size < 5 microns) gain access to alveoli of the lungs. It is here that the bacilli are engulfed by alveolar macrophages, which is a cell line equipped with multiple microbiocidal mechanisms including phagolysosome fusion. In order for the bacteria to establish infection, it must first survive this phagolysosome fusion and make its way to the lymphatics or bloodstream (McDonough *et al.*, 1993). It is this pathogen's ability to withstand the macrophages defense mechanisms and survive within the phagosomal compartment of the macrophage that makes it so deadly. Another reason contributing to *M. tb*'s high rate of successful disease establishment is its ability to develop resistance to drugs. It is this drug resistance that especially exacerbates disease control and management (Fu *et al.*, 2007) and poses a significant threat to the global control of the disease.

The relationship of mycobacterial genetics on drug resistant phenotypes and downstream clinical effects has now become a topic of great public interest and much work in this area has recently been conducted yielding very significant findings. One such finding was made by Pym *et al.* (2002) where a study was done on the KatG gene's 315 serine to threonine (S315T) mutation in *M. tb*. Clinically significant isoniazid (INH) resistance is most often linked to the S315T KatG phenotype. KatG in native form codes for a catalase-peroxidase enzyme which converts INH into its bioactive form. KatG is also known to be a virulence factor accounting for heavy attenuation of strains lacking KatG in a variety of animal models. In reality, it was observed that INH-resistant strains are transmitted even in the context of MDR phenotypes likely to be associated with other 'fitness-reducing' mutations. This phenomenon goes against the standard that resistance-conferring mutations significantly reduce bacterial fitness. This hypothesis was tested by constructing a panel of isogenic strains of *M. tb* with different katG alleles and characterizing them in an animal model of tuberculosis.

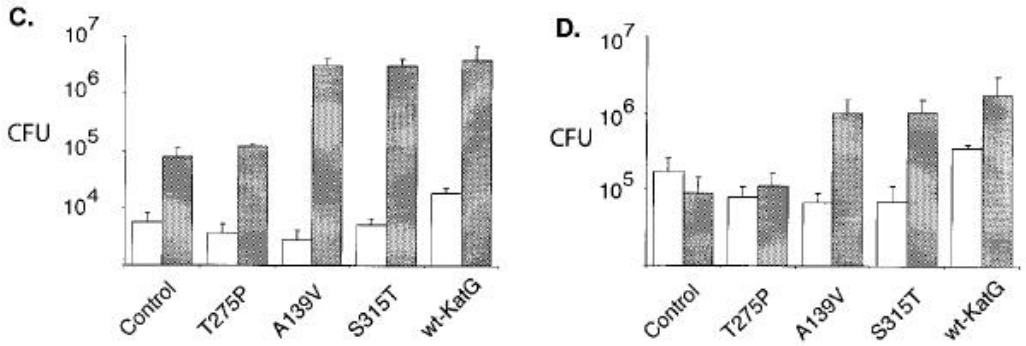


Figure 4.1: Experimental results where growth was monitored in BALB/c mice of strain INH34 complemented with the various katG allele plasmids. White bars correspond to the CFU after day one of infection and gray bars, after 40 days of infection in the lung (C) and spleen (D). Each time point represents the mean result of three or four mice and error bars correspond to standard deviation (Pym *et al.*, 2002).

One interesting point to note in this figure is that KatG-proficient strains were more prolific in their growth in the spleen and lung and the S315T recombinants grew much more vigorously than their respective negative controls.
These results conclusively demonstrated that antibiotic-resistance-conferring mutations do not necessarily carry a high fitness cost or reduction in virulence. The same concept was also ratified in a similar study showing, by knockout studies, that *M. tb*'s Nramp orthologues (mntH) are not vital determinants of virulence (Domenech *et al.*, 2002). This is a potentially dangerous situation as about a third of the world's population are allegedly latently infected with *M. tb* and a large proportion of these individuals carry MDR-TB. Due to it being shown that resistance-conferring-mutations need not negatively affect virulence, future re-activation of latent MDR-TB is a major concern (Pym *et al.*, 2002).

4.3 The Mycobacterial genome

Before moving on to comparative mycobacterial genomics, a short description of the mycobacterial genome will now be undertaken. A general understanding of the mycobacterial genome will form a foundation on which comparative genomics can be built. Using techniques such as pulsed field gel electrophoresis, fingerprinting and hybridization, scientists earlier on were able to establish the genome structure of two mycobacteria (Figure 4.2).

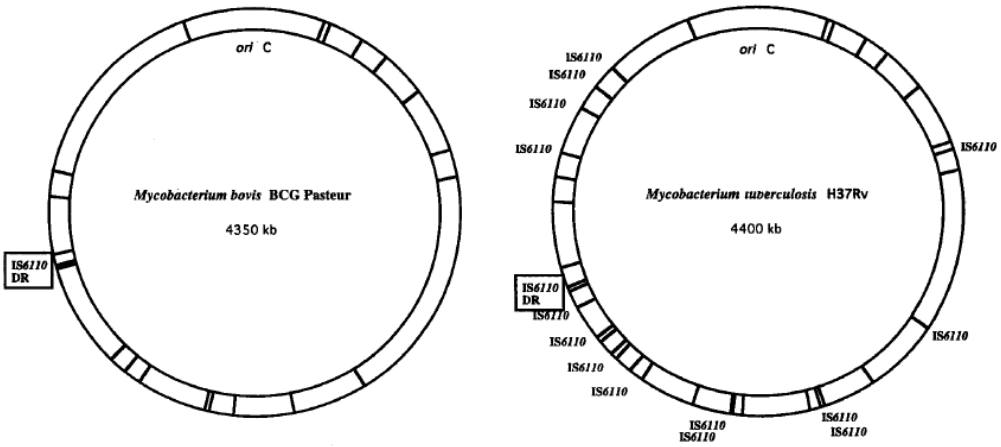


Figure 4.2: Early comparison of *M. tb* and the vaccine strain M. bovis BCG based on IS6110 sites.

Even at such an early stage, scientists were able to identify 16 copies of the insertion sequence IS6110, 8 copies of IS1081 and 26 copies of PGRS. The distribution of these insertion sequences were found to be non-random with a strong concentration of these inserts in a region of about 40% of the *M. tb* chromosome near the putative terminus of replication (terC). With the advancements of many scientific techniques, many more insights were to follow. The first *M. tb* strain to be isolated was H37Rv in 1905 (Camus *et al.*, 2002) taken from sputa from a 19 year-old male suffering from pulmonary tuberculosis (Kato-Maeda *et al.*, 2001). Even through years of passage, the bacterium has remained virulent in the animal model. In 1998, Cole *et al.* were the first to publish a completely sequenced and annotated genome of H37Rv highlighting many critical facts about the genome and opening doors for many other downstream comparative studies (see later). A few of the group's observations will be noted here.

By using large insert BACs, cosmids and random small-insert clones from whole-genome shotgun libraries, the group systematically pieced together the genome. The result was a 4,411,529 composite sequence with a 65.6% G+C (GC) content. The genome was found to be rich in repetitive DNA, particularly in insertion sequences (IS), new multi-gene families and duplicated house-keeping genes. The GC content was found to be relatively constant throughout the genome. This uniformity of the GC content is probably due to the lack of atypical base composition pathogenicity islands. There were however, regions that exhibited higher than average GC content, but these corresponded to sequences belonging to the large gene family which include the polymorphic GC-rich sequences (PGRSs).

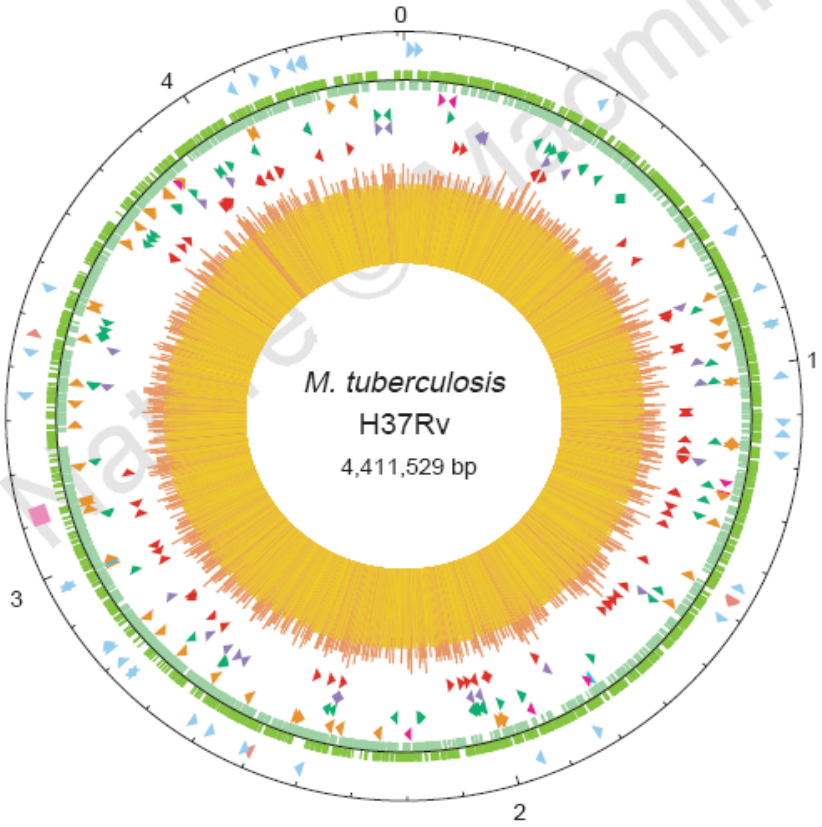


Figure 4.3: Circular map of M. tb H37Rv chromosome. The outer circle shows the scale in megabases. 0 represents the origin of replication. Moving inwards, the next ring denotes the position of the stable RNA genes (tRNA-blue; other are in pink); the second ring shows the coding sequence by strand (clockwise-darkgreen; anti-clockwise-lightgreen); the third ring denotes repetitive DNA (IS elements-orange; 13E12 REP family-dark pink; prophage-blue). The fourth ring denotes positions of the PPE family members (green). The fifth ring shows the PE family members (purple, excluding PGRS) and the sixth ring shows the positions of the PGRS sequences (dark red). The histogram (centre) represents GC content ($<$ 65% GC-yellow; $>$ 65%-red) (Cole et al., 1998).

In terms of RNA, 50 genes encoding functional RNA molecules were detected. The following were the three molecular species produced by the unique ribosomal RNA operon, the 10Sa RNA (involved in degradation of proteins encoded by abnormal mRNA), the RNA component of RNase P and 45 transfer RNA. The rrn operon is situated unusually as it is located about 1,500 kilobases (kb) from the putative oriC. This is significant as most eubacteria have one or more rrn operons proximal to oriC in order to exploit the gene-dosage effect obtained during replication. This placement may account for the slow-growth of *M. tb*.

In terms of insertion sequences and prophages, access to the full genomic sequence of H37Rv led to the detection of a further, pre-dominantly undescribed, 32 different IS elements in addition

to those presented by Philipp *et al*, (1998). The newly discovered IS elements mainly belong to the IS3 and IS256 families, though 6 constitute a brand new group. Most insertion sequences found in this genome appear to have been inserted into non-coding or intergenic regions and often near tRNA genes. Many have also clustered thus suggesting the possibility of actual insertional hot-spots which may prevent gene activation. At least 2 prophages were detected in the genome explaining *M. tb*'s persistent low-level lysis in culture. Prophages phiRv1 and phiRv2, both approximately 10kb in length and some of their gene products show marked similarity to those encoded by certain bacteriophages from Streptomyces and saprophytic mycobacteria.

In terms of protein coding genes, 3,924 ORFs were identified thus accounting for about 91% of the genomes potential coding capacity. Consistent with the high GC content, GTG initiation codons (35%) are used more frequently than in *Bacillus subtilus* (9%) and *E. coli* (14%), although ATG (61%) is the most common translational start. The even distribution in gene polarity seen in *M. tb* (as opposed to fast growing bacteria such as *E. coli*) may account for the slow-growth and the infrequency of replication cycles. Recently, many more fully sequenced Mycobacteria genomes have entered the public domain such as *M. tb* strains H37Rv, CDC1551, H37Ra and F11 as well as *M. bovis* AF2122/97 and *M. bovis* BCG str. Pasteur 1173P2. These are available through via the NCBI server and other public databases (Vishnoi *et al*, 2008). A list of fully sequenced genomes may be found at the Genome News Network website (17).

4.4 Comparative genomics of Mycobacteria

Thanks to the availability of *M. tb* and other mycobacterial whole genome sequences, comparative genomic studies are now flourishing helping us gain ground in the elucidation of tuberculosis pathogenesis amongst other things. An overview of Mycobacterial comparative genomics (ranging from the last decade to the present) will be treated here thus giving a good idea of the progression and current state of the field.

Comparative studies of Mycobacteria is by no means a new field. Scientists from very early on realized the potential of these comparisons and resorted to techniques such as PFGE, genetic fingerprinting, hybridization and restriction maps to compare mycobacteria (Philipp *et al.*, 1998; Philipp *et al.*, 1996; Cole *et al.*, 1998). These techniques were extremely effective and made many great discoveries possible. One of the first questions that scientists tried to understand was the mechanism responsible for the attenuation of *Mycobacterium bovis* in becoming the reliable BCG vaccine strain. In 2000 Brosch *et al.* set out to characterize the vaccine strain of M. bovis BCG Pasteur 1173P2. Due to the inability of DNA hybridization techniques to detect insertion sequences and translocations, a complimentary method based on PFGE of HinDIII restriction fragments from selected *M. tb* H37Rv and *M. bovis* BCG Pasteur (vaccine strain) BACs was used. Using this technique, two major rearrangements were identified in BCG. These were shown to correspond to two tandem duplications, DU1 (29 668 bp) and DU2 (36 161 bp). DU1 was the result of a single duplication event but DU2 is supposed to have arisen from a 100

kb genomic segment which subsequently incurred a 64 kb internal deletion. BCG strains that still contain DU1 and DU2 were found to be diploid for at least 58 genes and contain two oriC copies (Figure 4.4). Some evidence at the time of the study suggested that DU2 was still undergoing expansion as two copies were detected in a few sub-populations of BCG Pasteur cells. Although the exact impact of the duplications on the pathogenicity of BCG is unclear, mere knowledge of these regions would aid in the quality control of BCG vaccines (Brosch *et al.*, 2000).

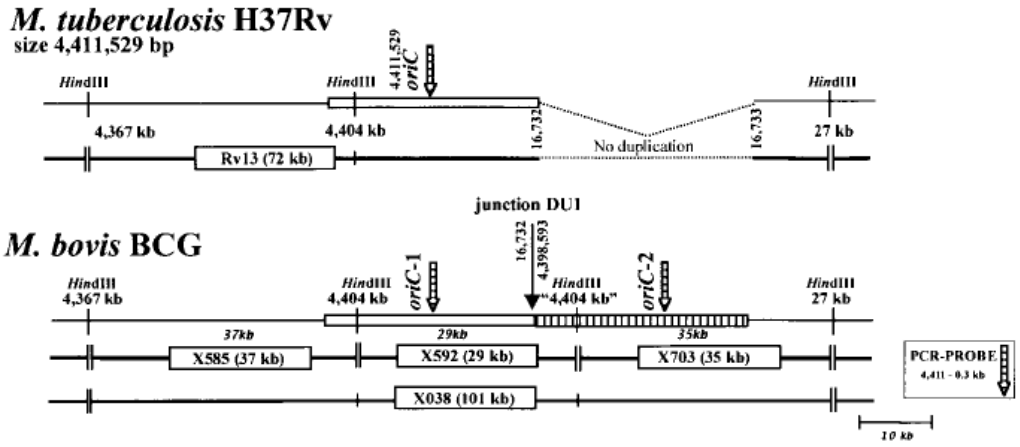


Figure 4.4: Overview of the genomic organization in the corresponding regions proximal to the origin of replication in BCG Pasteur and *M. tb* H37Rv, revealed by BAC mapping, PCR and hybridization experiments (Brosch *et al.*, 2000).

Comparison of mycobacterial genomes is also important for purposes of genotyping. Already covered in chapter one was the use of SNPs in order to categorize or group mycobacteria with their clonal relatives (Gutacker *et al.*, 2002). Fleischmann *et al.* (2002) in the same year also produced similar results. After sequencing the complete genome of the *M. tb* clinical strain, the group was able to perform comparisons with *M. tb* lab strain, H37Rv and showed that the two strains exhibited several differences on the SNP and larger-sequence level such as genomic re-arrangements. These changes were more apparent at specific loci, thus suggesting mutational hotspots and selective pressure. A more detailed look at these differences could elucidate their role in pathogenesis and host immunity (Fleischman *et al.*, 2002).

Drug target discovery is undeniably, a crucial aspect of medicine and biological sciences with far reaching implications. Comparative genomics has shown to be a powerful tool in this field as well (Marmiesse *et al.*, 2003). One study presented by Cole (2002) clearly illustrated this point. Gene duplication events are an unavoidable part of mycobacterial evolution leading to functional redundancies in biochemical pathways. It is sometimes difficult to predict with certainty, which of the duplicated genes impact on which function. In *M. tb*, there were five proteins that showed strong similarity (based on database searches) to various lipoamide dehydrogenase components of the crucial pyruvate dehydrogenase complex. The proteins in question were Rv0462, Rv0794c,

Rv2855c, Rv2713 and Rv3303c. During early analyses of the *M. tb* genome it was shown that proteins Rv3303c and Rv0794c (termed lpdA and lpdB respectively) showed the highest similarity to lipoamide dehydrogenase. Subsequent biochemical studies of the gene products however, showed that it was Rv0462 that actually encoded authentic lipoamide dehydrogenase. Comparative genomics would have helped come to this conclusion much faster as only one of the five *M. tb* proteins had a functional orthologue with *M. leprae* and that was Rv0462. The remaining four genes were present in pseudogene form in M. leprae (Cole *et al.*, 2002).

Scientists are often interested in tracing the lineage of bacteria (and organisms in general) for various reasons such as observing the effects of micro-evolution, tracing pathogenesis (Kato-Maeda *et al.*, 2001), examining what micro-evolutionary effects have on some protein families downstream in terms of drug resistance (Gagneux *et al.*, 2006) and also to determine their origins (Tsolaki *et al.*, 2005). Using comparative genomics, Brosch *et al.*, (2002) shattered the long held belief that *M. tb* evolved from *M. bovis*. By looking at a specific set of 20 variable regions among a 100 mycobacterial strains as well as specific deletion events (TbD1) the groups evidence pointed to the fact that *M. tb* existed before *M. bovis* and may have been a human pathogen far longer than previously believed. Using similar analyses Pym *et al.*, (2002) also managed to show mechanisms for *Mycobacterium bovis* BCG and *Mycobacterium microti* attenuation into vaccine strains. Comparative analyses (amongst other techniques such as gene knock-out) revealed that a specific deletion event of RD1 has a major role to play in the efficacy of the above-mentioned vaccine strains. In addition, further comparative analyses carried out by Garnier *et al.* (2003) revealed that *M. bovis* contains no unique genes when compared to other members of the *M. tb* complex. This implies that it is the differential gene expression that dictates host tropism and virulence of the various strains.

Comparative mycobacterial genomics has indeed proven to be a useful scientific tool over the years and with the availability of more sequence data, the scientific world is poised for greater insights into this pathogen. To proceed, we aim to exploit new mycobacterial sequence data by the implementation of a mycobacterial comparison project. This is dealt with next.

4.5 The Mycobacterial Comparison Project in context

Owing to the widespread effect of the disease and growing international concern, many other projects involving the comparisons of genomes of Mycobacteria have also been undertaken. Some of these projects include MycDB (Bergh & Cole, 1994), GenoMycDB (Cantanho *et al.*, 2006) and MycoperonDB (Rangan *et al.*, 2006) to name a few. MycDB is one of the earliest examples highlighting the need for an integrated platform to combine sequence data for mycobacteria. The MycDB project was essentially a database management system which combined mycobacterial specific data such as antigen lists, reagent details from CDC/WHO antibody bank, a few thousand gene sequences, reference data and physical maps (Bergh & Cole, 1994). The system was a stand-alone database management software which basically allows users to view data as well as

show relationships between the various data types in the database by complex query processes. Although useful at the time, it is now obsolete, as access to completely sequenced and finished genomes with substantial annotation and post genomic analyses information is widely available.

GenoMycDB (Cantanho *et al.*, 2006) is another more contemporary, mature and comprehensive web-based database example of a mycobacterial comparative database. At the core structure of this database lies the pairwise sequence alignments (and parameters) of all predicted proteins for six mycobacterial strains including *M. tb* (strains H37Rv and CDC1551), M. bovis AF2122/97, *M. avium subsp. paratuberculosis* K10, *M. leprae* TN, and *M. smegmatis* MC2 155. For each protein, the database also stores information such as sub-cellular localization, assigned cluster of orthologous groups (COGs), corresponding gene features and more. Users are also allowed to carry out complex queries in order to produce tables containing groups of potential homologous proteins. Furthermore, queries may be restricted by localization, DNA strand of corresponding gene and/or protein annotation.

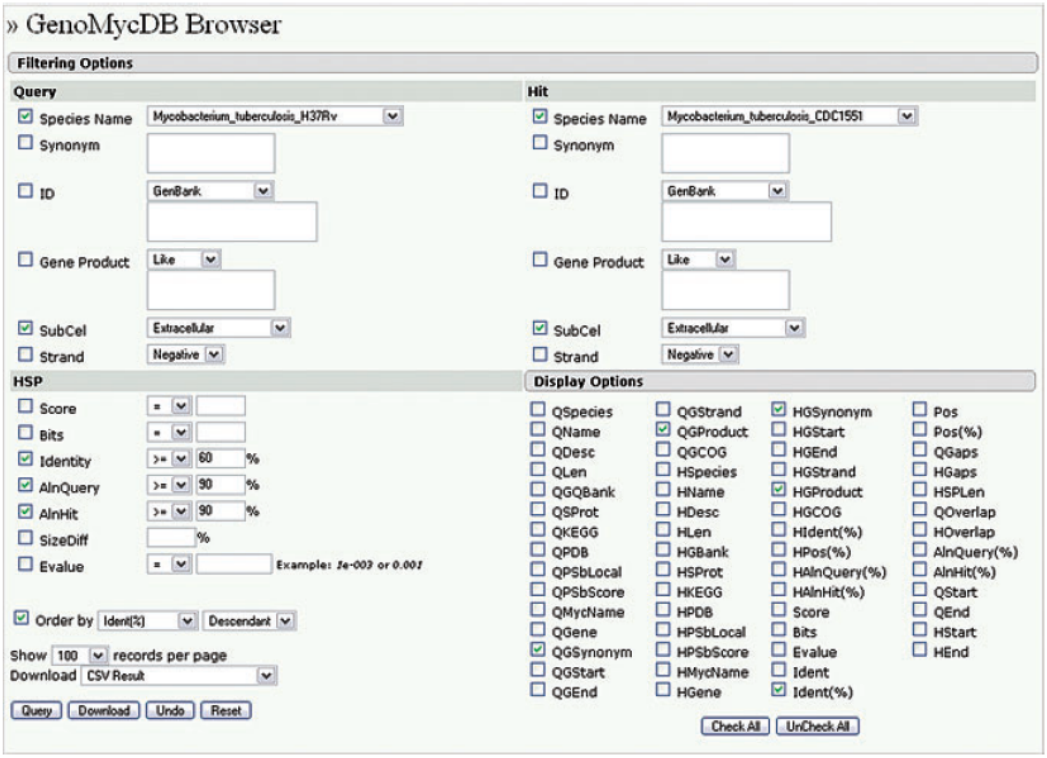


Figure 4.5: Overview of the GenoMycDB user interface. Note the available options for searching and displaying (Catanho *et al.*, 2006).

GenoMycDB is extremely useful allowing users to functionally classify their mycobacterial proteins of interest as well as elucidate the genome structure of the contained mycobacteria. The mycobacterial comparison project does not contain the same information as GenoMycDB but

rather seeks to create a richer source of post-analyses mycobacterial comparison data not only on the protein, but on the DNA level as well.

Yet another example is the MycoperonDB (Rangan *et al.*, 2006). This is a database of computationally predicted transcriptional units and operons from five different strains of mycobacteria. The database also combines literature information for experimentally validated mycobacterial operons with the aim of validating computationally predicted operons. All the above mentioned data is contained with a relational database with a user-friendly web-interface freely available at http://www.cdfd.org.in/mycoperondb/index.html.

The mycobacterial comparison project seeks to create a new type of comparative environment where on-the-fly comparisons can be done and meaningful conclusions can be drawn. This will be achieved by integrating a few related datatypes all of which are crucial in understanding the inter-relatedness of the various mycobacteria. The mycobacterial species chosen for analyses will now be discussed as well as how the various datatypes were generated.

4.6 Data pre-processing

4.6.1 Mycobacterial strain selection

There are quite a large number of mycobacterial strains that have been sequenced as well as fully assembled and annotated. These sequences are found in Genbank. However, initially only a few strains of mycobacteria were chosen so as to be representative of Mycobacteria as a genus. Both virulent and avirulent strains, spanning a moderate range of hosts, were chosen with which to conduct analyses. The strains chosen were *M. tb* H37Rv, *Mycobacterium tuberculosis* H37Ra, *Mycobacterium tuberculosis* F11, *Mycobacterium tuberculosis* CDC1551, *Mycobacterium ulcerans* Agy99, *Mycobacterium bovis subsp. bovis*, *Mycobacterium bovis* BCG, *Mycobacterium leprae* TN and *Mycobacterium avium subsp. paratuberculosis*.

4.6.2 Annotation Data

Annotation data for all mycobacterial strains were extracted from their full Genbank records. The files were parsed using Biopython and the annotations were subsequently stored in their respective tables in the database. Python scripts were used to extract the annotations as well as to populate the database (See figure below).

4.6.3 Gene-by-gene mutation data

After the gene data (i.e. annotations) for each species was committed to the database, the actual gene sequences for each gene was then searched with BLAST against all genes for the next species. In this way, the gene sequences for all the mycobacterial species were 'aligned' by BLAST. This analysis was a crude way of establishing whether a gene of one species was present in the next. If the gene was present in another species, CAI values were calculated for each

gene and more importantly, mutations between the sequences were recorded at each of the three codon positions. This data was stored in the 'gene_mutations' table.

4.6.4 SNP Data

For each of the nine species, SNP data was generated in the following way. Using M. tb F11 as the reference strain, each of the other strains of mycobacteria were aligned to the reference strain using the 'nucmer' tool (part of the MUMmer suite (Kurtz *et al.*, 2004)). Thereafter, 'show-snps' (also part of the MUMmer suite) was used to extract single-nucleotide polymorphisms (SNPs) from the alignment data produced by nucmer. SNP data was stored in species specific tables.

4.6.5 Gene island data

Using an in-house algorithm, developed by co-workers (see Seqword Genome Browser chapter), whole genome sequences of each of the mycobacterial strains were scanned for the presence of genomic islands which imply horizontally transferred elements. The genomic island co-ordinates for each of the strains were stored in the 'gi' (gene islands) table.

4.7 Database requirements

The database schema needed to be fast, stable and cater for the various data-types mentioned above. Separate tables were created for each of the data-types and foreign keys were subsequently used to associate the tables (Figure 4.6). There is a high degree of inter-connectivity between the sub-schema and within the sub-schema thus creating a highly flexible and highly semantic relationship of all the data. This sub-schema designed for the mycobacterial project had to also be compatible with the main Fungims project schema and this was successfully accomplished.

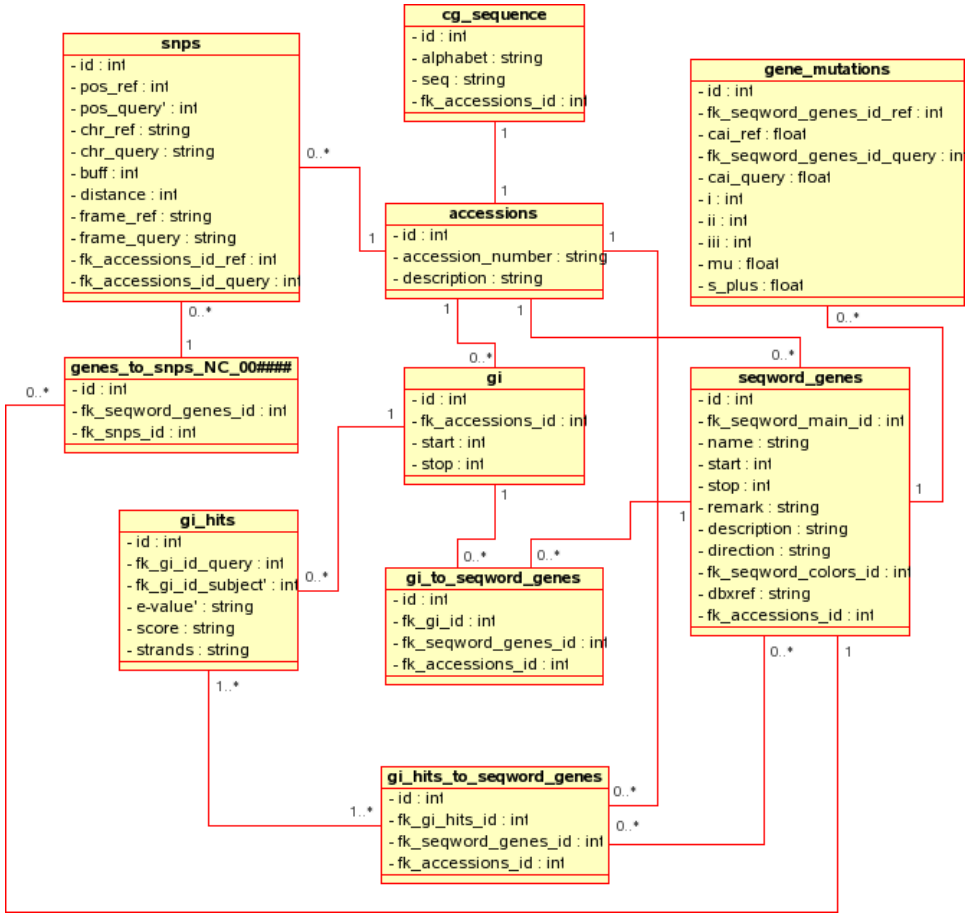


Figure 4.6: Schema of the mycobacterial comparison project database.

4.8 Graphical User Interface requirements

The design of the graphical-user interface (GUI) had to first and foremost, relay the significance of the inter-relationships of the various data-types in a manner that had the most impact. This was accomplished by the use of contrasting colors and an effective layout. Also, due to the many data-types and information that had to be displayed, a simple layout pattern had to be used to display all the data on the page while avoiding clutter and over usage of screen real estate. This was accomplished by simple tables and side-by-side display of the data. Furthermore, user-friendliness was a top priority and no more than two or three buttons had to be clicked for users to dynamically generate data. Also, the screen layout was extremely simple and un-ambiguous leaving users with little to be confused about when using mycobacterial comparison project.

4.9 Workflow summary

The mycobacterial comparison project essentially allows for the integration and visualization of several key data-types significant in mycobacteria. In addition, simple tree drawing functionality is also available. The basic underlying functions include annotation retrieval, database filtering of overlapping data, sequence alignment of genes and lastly, neighbor-joining tree generation. The following table depicts the general flow contained within the mycobacterial comparison project (Table 4.1)

Table 4.1: Table showing general order of events and options available to users when in the mycobacterial comparison project.

#	Web page/User choice	Tasks performed by server
1	Mycobacterial Comparison Project introductory Page	- Retrieval and display of mycobacterial genome metadata including genome length; accession numbers and number of gene islands found
2	Choice of strain to explore	- Retrieval and display of annotation information for chosen strain
3	Choice of gene of interest	- Retrieval of gene information of user specified gene - Retrieval of homologues (gene-by-gene data) - Retrieval of SNPs corresponding to specified gene - Autogeneration of protein alignment of homologues - Formatting and display of all above data
5	Creation of phylogenetic trees	- Redisplay of all homologues on new web page
6	Choice of DNA or Protein alignment based neighbour-joining tree	- Retrieval of all sequences (sequence conversion if necessary) - Remote-procedure call for clustalw alignment - Collection of result files and error checking - Creation of distance matrix (phylip) - Neighbour-joining tree generation (phylip) - Conversion of tree file to other formats - Display of tree and homologues

4.10 A comparative genomics investigation of key genomic loci in mycobacterial genomes and their role in virulence

In this section an investigation built off the findings of the SWGB chapter is carried out. The aim is to further identify significant genetic differences between specific mycobacterial strains such as *M. tb* H37Rv, *M. tb* H37Ra and *M. tb* F11. Starting off with the genomic islands identified by the SWGB, this system was used to analyze these specific areas in more detail and show how the genes contained within these loci are not conserved as would be expected for these essential genes. Genes that have important functions are often found conserved within a group of bacteria. For example, DNA replication genes are core to bacterial genomes and are thus expected to be conserved. An investigation into these DNA polymerase genes was also under-

taken for a few mycobacterial species in order to inspect their level of conservation among the mycobacteria. Other genes of interest, identified by their atypical nucleotide usage (in chapter 3) were the PE-PGRS genes. A profile of these genes in particular was built up for the mycobacteria in order to guage their relationship with the organism's virulence status. Profiles for this gene family was created by, selecting specific genes (one at a time) and then retrieving the homologue data generated by the MCP. The data for each gene was then tabularized (see below).

In the previous chapter, the SWGB identifed genomic islands within the various mycobacterial strains and using its comparative view, several key genomic islands were identified. As an example, attention will be turned to the gross differences identified between *M. tb* H37Rv and *M. avium ssp Paratuberculosis*. With the aid of gene diagram dot-plots (RV, GRV and D) in chapter 3, several genomic islands were identified in *M. avium ssp Paratuberculosis* that were clearly absent in *M. tb* H37Rv. A closer look within these genomic islands revealed several genes and gene families. Similarly, using different dot-plot criteria, genomic islands unique to *M. tb* H37Rv were revealed.

Several genomic islands within *M. avium ssp Paratuberculosis* (K10) were identified. The table beneath shows the coordinates and annotations for the identified genomic islands.

Table 4.2: Coordinates and annotation of the genes islands in the genome of *M. avium* K10.

Left	Right	n1_4mer:RV	n1_4mer:GRV	n0_4mer:D	Annotations
78000	86000	11.6606	18.0853	13.7828	dnaB; mmpL4_1 and 4 genes for hypothetical proteins
870000	892000	8.6089	22.8127	22.0832	nramp and 20 hypothetical proteins
1290000	1304000	8.7803	20.6197	21.1895	lipL and a hypothetical gene

Using the mycobacterial comparison project, a closer look is taken into each of these genomic island sub-elements to see how they relate to other mycobacterial strains. dnaB was examined first. Although dnaB may not necessarily be of 'horizontal' origin it was investigated to see whether this very important gene contained homologues and to what extent the homologues differed among the various strains.

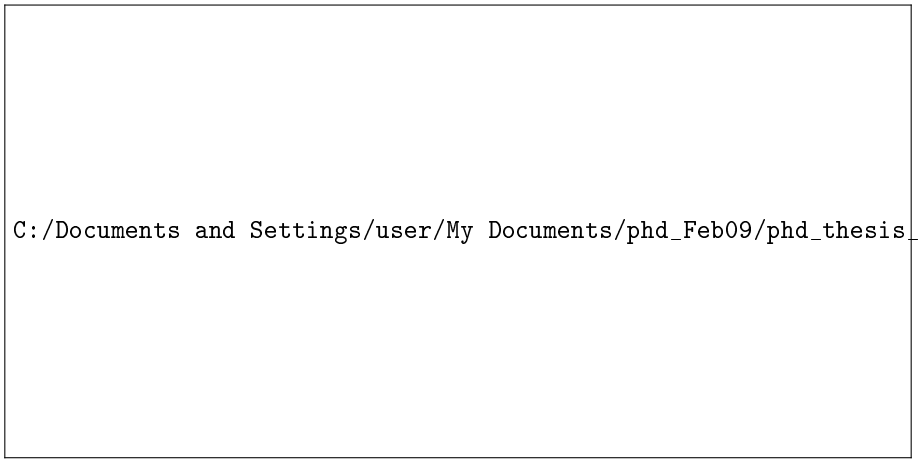


Figure 4.7: dnaB gene details for *M. tb* H37Rv and its homologues.

The gene dnaB was found present in all the database species present. The MCP system essentially extracts the annotation and its related sequence information for the gene of interest from the species of interest and nucleotide BLASTs it against the annotations and sequences for the other strains. Based on this analyses steps, the dnaB gene of *M. tb* H37Rv has been identified in (i.e aligned to) all other species in the database. However, on closer examination, it was found that the dnaB *M. avium* K10 homologue is the most different when compared to the dnaB of *M. tb* H37Rv, containing over 640 single nucleotide differences which ultimately lead to a downstream protein alteration. When aligning, the MCP also takes into account the nucleotide mismatches within the alignment and counts them. The MCP also performs on the fly translation of the genes to inspect whether the mismatches were synonymous (coded in green) or non-synonymous (coded in red) relative to the reference species protein. Although at the gene level dnaB is relatively unchanged among *M. tb* CDC1551, *M. bovis* and *M. tb* F11, the protein sequence was still affected. Note that this protein remains fully conserved between the *M. tb* strains H37Ra and H37Rv. *M. Leprae* exhibited 291 nucleotide differences relative to *M. tb* H37Rv which lead to an altered protein. This many differences in the nucleotide sequence inevitably means great differences in the amino-acid sequence and subsequent protein. The possibility that the dnaB of *M. Leprae* functions slightly different to that of *M. tb* H37Rv is great.

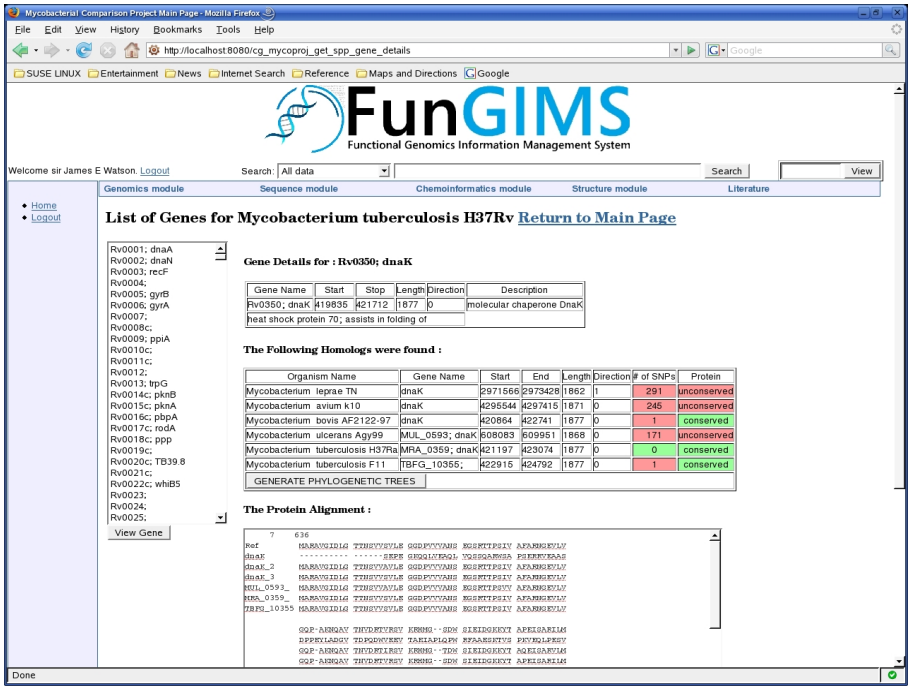


Figure 4.8: dnaK gene details for *M. tb* H37Rv and its homologues.

Several other dna genes were looked at using the MCP. Figure 4.8 above shows again how different *M. avium's* dnaK homologue is relative to *M. tb* H37Rv. The table below highlights the dna genes of *M. tb* H37Rv and whether there are homologues in *M. avium* K10.

Table 4.3: Summary of absence/presence of dna genes of *M. tb* H37Rv in *M. avium* K10.

Gene in *M. tb* H37Rv	Present in *M. avium* K10
dnaB	Yes
dnaA	No
dnaN	No
dnaK	Yes
dnaJ1	Yes
dnaE1	Yes
dnaJ2	Yes
dnaE2	Yes
dnaQ	Yes
dnaZX	Yes

The table above clearly shows that most dna genes of H37Rv are also present in *M. avium* K10. This may imply that the mechanisms for dna replication between these species are highly similar, however, the MCP also revealed that none of these dna genes are actually conserved in *M. avium*. In all cases, where there is an *M. avium* homologue, there are over 100 nucleotide differences between the *M. tb* H37Rv copy and its *M. avium* counterpart. This could further

imply that although all the protein components may be shared among species, the mechanisms by which they are used may differ.

In terms of the other genes identified within *M. avium*'s gene-island, mmpL4_1 was examined.

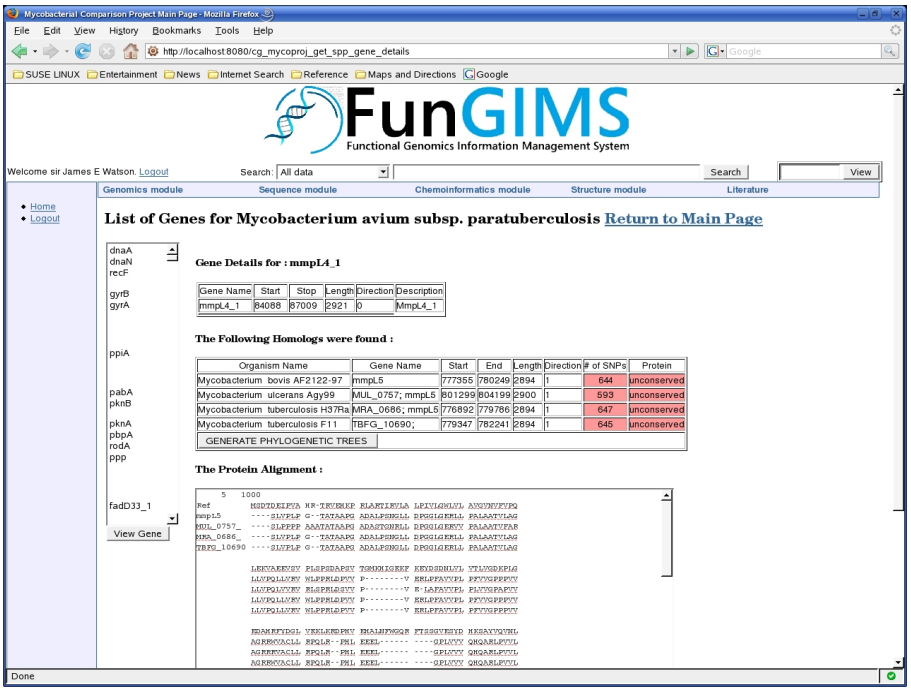


Figure 4.9: mmpL4_1 gene details for *M. avium ssp paratuberculosis* K10 and its respective homologues.

Supportive of the SWGB gene-plot findings, it is seen here that this gene, is totally absent from *M. tb* H37Rv. Furthermore, the gene does not seem to exhibit a high degree of conservation among the various mycobacterial strains as shown by the high number of SNPs in the homologues. Whether this gene is functional or not in the species tested is not known, however, if they are functional, the mechanisms are likely to be quite variable.

The gyr gene is another gene of interest commonly used to differenciate between different mycobacterial strains (Chimara *et al.*, 2004). It was decided that the MCP system would be used to check if this gene could be used to differenciate between these two strains.

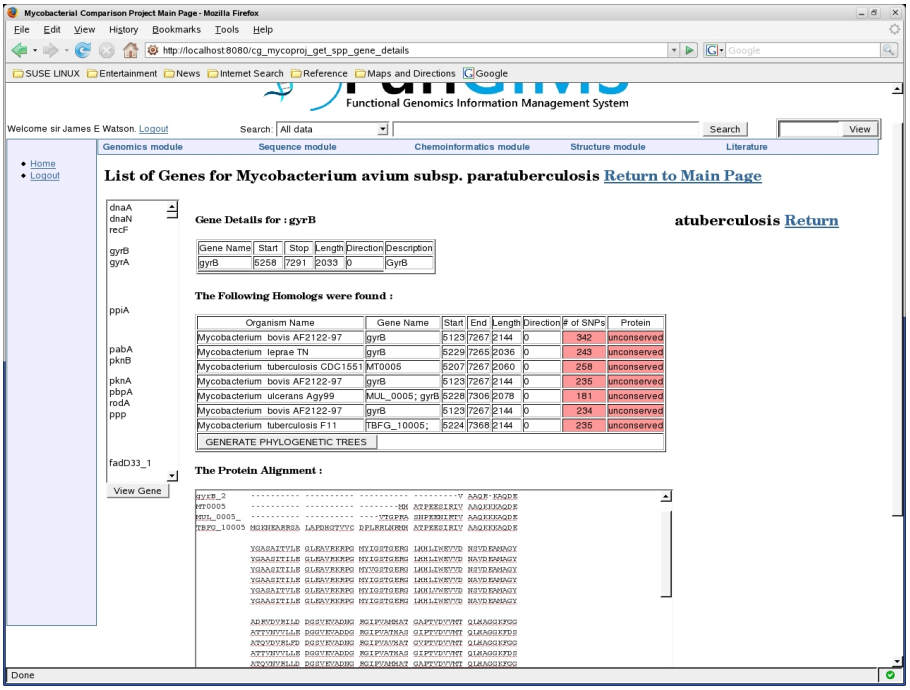


Figure 4.10: gyrB gene details of *M. avium ssp paratuberculosis* K10 and its homologues.

It is interesting to note that no homologues for gyrB are detected within *M. tb* H37Ra or strain H37Rv. The gene however is known to be present in *M. tb* H37Rv, but is so vastly different on the nucleotide level to gyrB in *M. avium* that the MCP fails to detect it. Due to the way in which the MCP searches for homologues, gyrB could have gone undetected because the *M. tb* genomes (H37Rv, H37Ra etc) were not properly annotated or incompletely annotated. gyrB holds an essential function by encoding the B sub-unit of DNA gyrase and its presence is crucial to the organism. The fact that this gene exhibits such a high level of mutation among the mycobacteria is significant and not well understood. The variability for this gene though high, is controlled.

In terms of the gene islands detected for *M. tb* H37Rv, most of the features within these islands are comprised of genes belonging to the PE-PGRS/PPE family of proteins (Table 4.4)

Table 4.4: Annotations for the outlined genomic fragments for the *M. tb* H37Rv plot (Figure 3.10).

LEFT	RIGHT	n1_4mer:RV	n0_4mer:PS	GC	ANNOTATION
332000	342000	75.5	25.7	0.73	PE-PGRS and PPE family proteins
1630000	1638000	77.1	29.1	0.75	PE-PGRS and PPE family proteins
3734000	3746000	120.5	31.5	0.73	PE-PGRS and PPE family proteins
3924000	3954000	121.6	29.1	0.76	PE-PGRS and PPE FAMILY PROTEINS; acyl-CoA synthase; acyl-CoA dehydrogenase; acyl-CoA lygase FADD18; enoyl-CoA hydratase; thiamine-pyrophosphate requiring enzyme and many hypotheticals

Using the MCP, another comprehensive table was constructed containing PE-PGRS/PE/PPE comparisons (for loci 333437-3950263 of *M. tb* H37Rv) among all the database species present. The table was constructed in the following way. Starting with *M. tb* H37Rv, every annotation appearing between region 333437 and 3950263 was individually analysed using the MCP and the results stored in the table. Results that were recorded included, does the annotation appear in other species (i.e is there a homologue) or not; location of the homologue if found; is the homologue conserved or not; if not, how many SNPs appear in the homologue.

Table 4.5: Summarised table showing cross-species comparison of loci 333437-3950263 of *M. tb* H37Rv. (Note the SNP column indicates - total number of SNPs in homologue / number of SNPs per 100 bp).

H37Rv		H37Ra		F11	
Genes		NC_009525		NC_009565	
Code	Name/Position	Coord	SNP	Coord	
Rv0287c	PE_PGRS3 [333437-336310]	334799-337672	1 \ 0	336752-339472	193 \ 6
Rv1450c	PE_PGRS27 [1630638-1634627]	1631948-1636144	98 \ 2	1635000-1639223	109 \ 2
Rv3343c	PPE54 [3729364-3736935]	3740205-3746048	981 \ 14	3741691-3749289	209 \ 2
Rv3347c	PPE55 [3743711-3753184]	3752824-3762261	36 \ 0	3756176-3763594	2058 \ 24
Rv3350c	PPE56 [3755952-3767102]	3765065-3776089	126 \ 1	3768418-3779781	66 \ 0
Rv3507	PE_PGRS53 [3926569-3930714]	3935246-3939391	1 \ 0	3938922-3943133	36 \ 0
Rv3508	PE_PGRS54 [3931005-3936710]	3939682-3944319	1124 \ 21	3959787-3963047	968 \ 21
Rv3511	PE_PGRS55 [3939617-3941761]	3949108-3951261	7 \ 0	3953290-3958956	6 \ 0
Rv3512	PE_PGRS56 [3941724-3944963]	3949108-3951261	1285 \ 47	3953290-3958956	1246 \ 27
Rv3514	PE_PGRS57 [3945794-3950263]	**3939682-3944319**	**626 \ 13**	**3959787-3963047**	**727 \ 18**
H37Rv		CDC 1551		Bovis	
Genes		NC_009755		NC_002945	
Code	Name/Position	Coord	SNP	Coord	SNP
Rv0278c	PE_PGRS3 [333437-336310]	333551-336268	157 \ 5	334697-337330	360 \ 13
Rv0280	PPE3 [339364-340974]	339309-341036	30 \ 1	340366-341976	1 \ 0
Rv1450c	PE_PGRS27 [1630638-1634627]	No homologs		1632588-1634975	1287 \ 40
Rv1452c	PE_PGRS28 [1636004-1638229]	1636119-1638335	32 \ 1	1632588-1634975	112 \ 4
Rv3343c	PPE54 [3729364-3736935]	366212-375772	2018 \ 23	3686614-3690630	3708 \ 64
Rv3345c	PE_PGRS50 [3738158-3742774]	No homologs		3694689-3696308	3039 \ 97
Rv3347c	PPE55 [3743711-3753184]	424940-434767	2353 \ 24	3700427-3706717	3205 \ 40
Rv3350c	PPE56 [3755952-3767102]	424940-434767	2608 \ 24	3719324-3720628	9848 \ 158
Rv3507	PE_PGRS53 [3926569-3930714]	No homologs		3872192-3876274	215 \ 5
Rv3508	PE_PGRS54 [3931005-3936710]	No homologs		3890501-3893479	1154 \ 26
Rv3511	PE_PGRS55 [3939617-3941761]	No homologs		3883854-3889670	7 \ 0
Rv3512	PE_PGRS56 [3941724-3944963]	No homologs		3883854-3889670	1151 \ 25
Rv3514	PE_PGRS57 [3945794-3950263]	No homologs		**3890501-3893479**	**859 \ 23**
H37Rv		Ulcerans		Avium K10	
Genes		NC_008611		NC_002944	
Code	Name / Position	Coord	SNP	Coords	SNP
Rv0278c	PE_PGRS3 [333437-336310]	575111-576157	974 \ 49	4622313-4622930	889 \ 50
Rv0279c	PE_PGRS4 [336560-339073]	575111-576157	942 \ 52	No homologs	
Rv0280	PPE3 [339364-340974]	1289688-1291244	386 \ 24	3880292-3881887	300 \ 18
Rv1450c	PE_PGRS27 [1630638-1634627]	4845511-4848744	960 \ 26	No homologs	
Rv1452c	PE_PGRS28 [1636004-1638229]	4845511-4848744	562 \ 20	No homologs	
Rv3343c	PPE54 [3729364-3736935]	95273-96187	2229 \ 52	4399276-4399920	2257 \ 54
Rv3345c	PE_PGRS50 [3738158-3742774]	1221305-1223809	1057 \ 29	No homologs	
Rv3347c	PPE55 [3743711-3753184]	829924-831396	2867 \ 52	2895832-2896983	3706 \ 69
Rv3350c	PPE56 [3755952-3767102]	829924-831396	2804 \ 44	1673194-1674579	3534 \ 56
Rv3507	PE_PGRS53 [3926569-3930714]	402249-403910	1808 \ 62	No homologs	
Rv3508	PE_PGRS54 [3931005-3936710]	4456652-4457851	1764 \ 51	No homologs	
Rv3511	PE_PGRS55 [3939617-3941761]	4913637-4915358	540 \ 27	No homologs	
Rv3512	PE_PGRS56 [3941724-3944963]	4859588-4862377	935 \ 31	No homologs	
Rv3514	PE_PGRS57 [3945794-3950263]	4456652-4457851	1225 \ 43	No homologs	

The table essentially represents the gene island identified by the SWGB and all the sub-genetic elements found within this island. It is clear that this region is dominated by PE/PE-PGRS/PPE proteins. Several important findings are highlighted in the above table. Firstly the loci marked in blue for *M. tb* F11 are quite interesting as it appears to be genetically closer to *M. tb* H37Rv than strain H37Ra. This was not expected as the strain H37Ra was generated from strain H37Rv in a series of passages in petri dishes under controlled laboratory conditions. This finding will be discussed below.

It is also interesting to note that many of the genes above are not found within *M. avium* K10. Examination of the genome atlas of the *M. tb* H37Rv genome shows that loci 333437 to 3950263 are rich in repeats (direct and inverted). As already mentioned, this is characteristic is typical for this family of proteins.

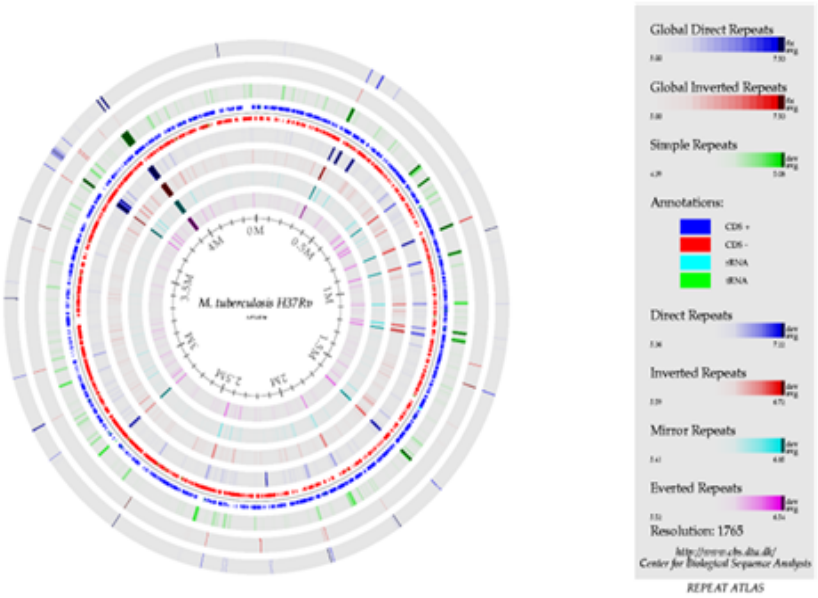


Figure 4.11: Genome atlas of *M. tb* H37Rv. Note the abundance of repeat regions especially in regions 3.9 – 4.0 MB (13).

Looking at a similar atlas for *M. avium* K10 reveals that these genes or gene area, unlike strain H37Rv, is actually core to the *M. avium* K10 genome (Figure 4.12) and does not contain an unusually high amount of repeats.

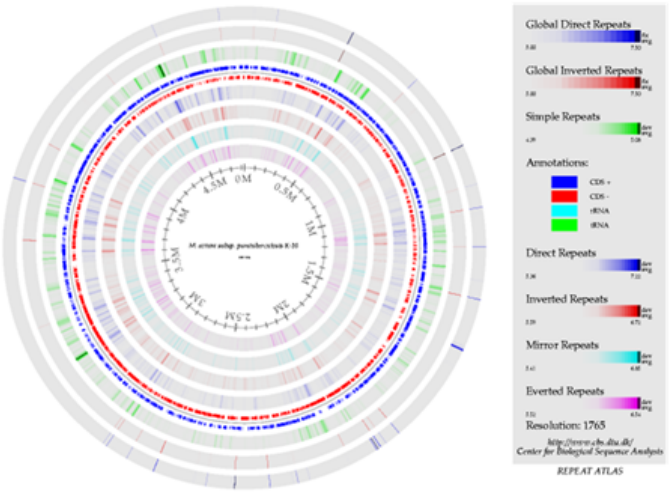
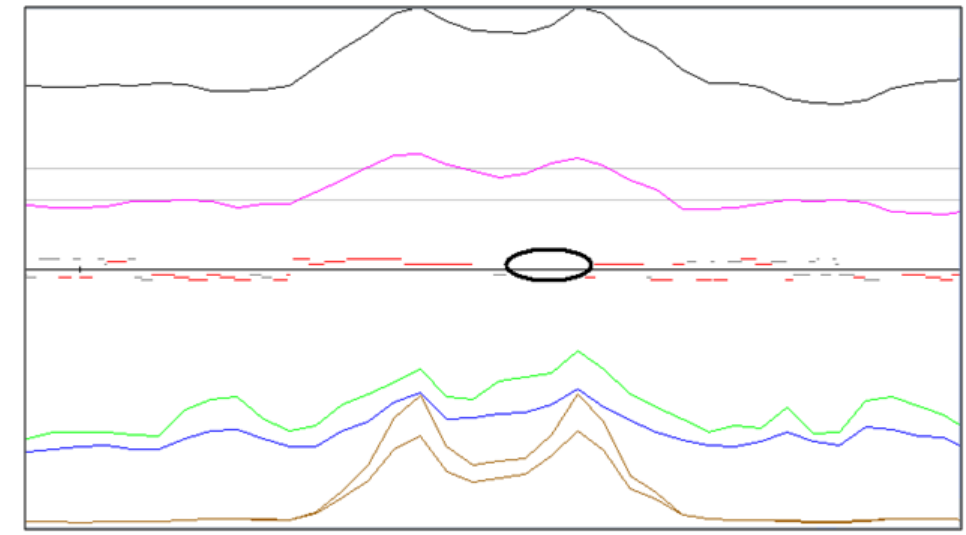
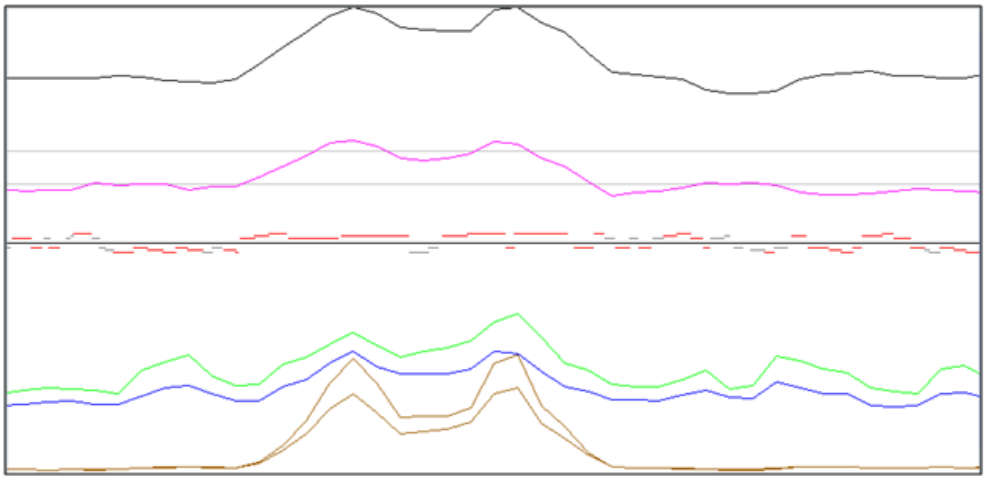


Figure 4.12: Genome atlas of *M. avium* K10 (13).

Another point of note is for that of annotation Rv3514, especially in relation to strain CDC1551. The lack of homologues is highly unusual. Going back to a SWGB view for this location in CDC1551 it can be seen that this discrepancy may have resulted from incompleted annotation information for the genome *M. tb* CDC1551 (Figure 4.13).



(A)



(B)

Figure 4.13: A Region of *M. tb* CDC1551 that appears to lack annotation information and B the corresponding region in *M. tb* H37Rv.

This illustrates that there still much vital information missing from our databases. In time however, with the increase in the amount of curated data, sequence information will definitely grow in richness.

4.11 Discussion

In this work the MCP system was used to compare rates of mutations in different genes of mycobacteria. One interesting finding was that the genes gyrA, gyrB, dnaA and dnaB, which are key players in the bacterial replication system and thus show high level of conservation among bacteria, were shown to exhibit a high degree of heterogeneity among the tested mycobacterial strains. Due to this heterogeneity, it can be suggested that the replication system among these mycobacterial strains are not conserved. There is a possibility that the genes in question may have evolved faster and these gene changes could have, in conjunction with several other events, lead up to the emergence of the slow-growing mycobacteria. However, to verify this, a genome-wide scale investigation would have to be undertaken for all genes amongst all the strains in question to acscertain with confidence if the gyr and dna genes are more susceptible to mutation relative to the other genes.

mmpL4_1 was also one of genes identified within *M. avium*'s K10 GI, and the MCP showed that this gene was entirely absent from H37Rv as suggested by the SWGB dot-plot.

The SWGB was able to identify crude regions of differences between each strain and the MCP is complementary in that it allows us to zoom into these areas of interest and allows us to see exactly how these genes differ even on the SNP level.

In terms of *M. tb* H37Rv, the GIs detected for this strain was found roughly between the loci 3.3-3.96 MB and was predominantly comprised of PE/PE-PGRS/PPE gene families. Using the MCP, a table was constructed summarizing all the genes within this GI in relation to other strains within the database collection. Several noteworthy observations will now be discussed.

M. tb strain H37Rv is the virulent precursor to strain H37Ra and thus should be the closest genetic relative to H37Ra. However, it is seen that in terms of the PE-PGRS genes in the above-mentioned loci, for several genes, *M. tb* H37Rv is actually closer to *M. tb* F11. Strain H37Ra is a laboratory strain while F11 is a clinical isolate. Therefore, the evolutionary pressure that H37Ra is under is quite different to what F11 had experienced over time. The time frames are not clear but what is clear is that these genes code for proteins that exhibit antigenic variation and thus contribute significantly to B-cell responses in TB patients (Tundup *et al.*, 2006). A global gene study needs to be unteraken to acsertion if these genes undergo greater mutation activity ('mutational hotspots') relative to the other genes within a mycobacterial genome. What is known is that strains H37Rv and F11 are both virulent whereas H37Ra is not. The contribution of these genes and the mutations they underwent, toward the virulence of *M. tb* (H37Rv and F11) and acquired avirulence of *M. tb* H37Ra is an area still not well understood.

The MCP system, showed the PE/PE-PGRS/PPE arrangement is indeed, non-random. There

could be several reasons for the clustering of these gene families. van Pittius *et al.* showed that some members of the PE/PPE families are associated with the ESAT-6 (esx) genes cluster. The ESAT-6 gene clusters have been shown to be involved immunopathogenic activity. By employing techniques such as phylogenetics, DNA hybridization and comparative genomics the group showed the PE/PPE gene families expansion to be linked with the duplications occurring within the ESAT-6 gene clusters. They also propose that the duplication and distribition of the ESAT-6 gene clusters over time could explain the lineage of the slow-growing mycobacteria (van Pittius *et al.*, 2006).

The MCP was shown to be quite useful in highlighting useful features of mycobacterial genomes and seeing how various genes relate to other strains and shows the level of conservation between elements among strains. Another feature of the MCP, though not shown, is its ability to quickly construct phylogenetic trees. This is useful in allowing users to view phylogenetic distances of their genes/features of interest relative to the various mycobacterial species. The system could easily be adapted to include other mycobacterial species (as they become sequenced) as well as any other prokaryotic group with simple pre-calculations and database inclusions.

Research into tuberculosis is very much a growing concern and due to the bacteria's slow growth, sequence based research and diagnostic methods are advantageous. With the greater access researchers have to high-throughput sequencing techniques, means, these sequence analyses methods will find a niche. The ability to centralize and rapidly compare mycobacterium sequences is becoming an indispensible tool to researchers and diagnostic services alike.

Chapter 5

Concluding Discussion

The development of a comparative genomics environment is a task that has been tackled by many research groups in the past and indeed has proven to be quite successful for scientific purposes. In this study however, the aim was to create an integrated comparative genomics environment that combined together not only basic comparative genomics tools but novel tools as well. In the end, the usefulness of the various tools and approaches is shown by making some interesting discoveries about mycobacterial genomes and the behaviour of key mycobacterial sub-genetic elements amongst a few representative species. This project formed part of the greater Functional Genomics Information Management System (FunGIMS). Before dealing with the specifics of the findings made, a summary of the steps taken in development from conception to design and implementation will now be discussed.

The concept of a web-based, integrated comparative genomics environment in this context meant the combining of various tools and algorithmic approaches, in a highly accessible way, in order to facilitate scientific investigation of genomic sequence data. Combining various software tools under one environment while achieving integration relied on one very important factor. There had to be a highly versatile, robust and flexible data structure (MODEL) which would be able to handle the inputs and outputs from the various tools and data types within the environment. This was the first step toward realization of the project bar technology choice.

Before embarking on detailed design for the project, decisions had to be made regarding which software would be employed to construct the system. The project policy held for development was in favor of the open-source community, thus all software used for this project was borrowed from the open source community. Turbogears, a rapid web-development framework was used to develop the core of the system and the interface. CherryPy was the server of choice where the modules were deployed. SQLAlchemy, an object relational mapper was used in conjunction with Turbogears for the development of the database schema. Likewise, all scientific tools example

BlastZ, phylogenetic packages and sequence alignment programs were all sourced from the open-source community.

The model for this system was built upon the already established Functional Genomics Experiment (FuGE) model (Pizarro *et al.*, 2005). FuGE was designed with the prospect for extension to include other biological data types under a single model (Jones *et al.*, 2007) and thus laid the platform upon which a comprehensive customized model was built.

The basic analysis tools incorporated into the system included BlastZ for genome alignment, MAFFT and ClustalW for regular sub-genomic sized sequence alignment (protein and DNA), BLAST and Phylip. Using the Turbogears web-development framework, a common web-interface allowing access to the various tools was designed. The web-interface also allowed communication with the back-end database thus giving users access to their stored data and simultaneous usage of the software tools.

Together with the comparative genomics project, the novel offering of this system was the SeqWord Genome Browser (SWGB) and the Mycobacterial Comparison Project (MCP). These were included as sub-modules of the larger system. The SWGB provided users with a novel way of visualizing genomes based on its novel algorithmic approach to sequence analyses. The MCP system on the other hand, allowed users to perform gene-by-gene analyses on several mycobacterial genes simultaneously for various mycobacterial species, this, combined with SNP and annotation data allowed users a detailed view of specific parts of the genome and allowed rapid cross-species comparison.

Tuberculosis is a very relevant area of scientific interest, both locally and abroad. Due to the devastating impact this disease has had on the human population, mycobacteria was chosen as a test organism with which to validate the performance and value of the system. Using the unique offerings of both the SWGB and the MCP, a scientific investigation was carried out to explore a few of the major genetic differences between several key strains of mycobacteria.

By using the SWGB and adjusting its various plot parameters several genomic islands (i.e regions of atypical oligonucleotide usage) were identified in *M. avium* K10 and *M. tb* H37Rv. These gene islands oligonucleotide usage patterns vary so greatly from the rest of the genome that they are deemed to be of horizontal origin or may represent mutational hotspots. Gene islands may cover large areas of a genome and thus contain may sub-genomic elements that are of interest. The SWGB further enabled zooming into these genomic islands and retrieve annotations for the sub-genomic elements contained within the genomic islands.

In the case of *M. avium* K10, the sub-genomic elements contained within its gene islands were all

found to be linked with virulence or drug metabolism. For instance MmpL was one of the genes found. It is known that MmpL-mediated lipid secretion affects both the pathogens ability to survive intracellularly as well as host-pathogen dialogue which determines the ultimate outcome of infection (Domenech *et al.*, 2005). lipL was one of the other elements identified. lipL belongs to the hormone-sensitive lipase family and is responsible for fatty acid metabolism during an adverse nutrient climate (Deb *et al.*, 2006). This activity may account for *M. avium* K10's ability to utilize stored triacylglycerols during dormancy and its subsequent reactivation. These findings may be key to the understanding of the variation in pathogenesis and host interaction exhibited by *M. avium* and *M. tb* H37Rv.

In terms of genomic islands identified for *M. tb* H37Rv. It was found that PE-PGRS and PPE family of proteins were the dominant feature comprising the islands. PGRS genes are GC rich and have been shown to be a source of antigenic variation. This set of genes plays a role in the bacterium's survival within macrophages. This again points to the differences in the survival mechanisms between these species. Furthermore, these PE-PGRS genes were shown to cluster together. This phenomenon has been observed by other researchers and it has been shown that the duplication dynamics of the PE/PPE gene families is strongly linked to the duplication of the ESAT-6 gene clusters within a genome (van Pittius *et al.*, 2006). Although we have seen (amongst our tested strains), that the PE/PPE gene families do exhibit mutation activity, we are yet to show that it occurs at a rate greater than that observed for the rest of the other gene families within the genome. In a related study however, an examination of the PPE38 genomic region was performed. PPE38 belongs to the PPE gene family of *M .tb* (and other related mycobacteria). By looking at the PPE38 region in a cohort of clinical *M. tb* isolates, it was demonstrated that this region was in fact hypervariable due to insertion events involving IS6110 elements, IS6110-associated recombination, homologous recombination and gene conversion events between its PPE38 and PPE71 constiutent homologues (McEvoy *et al.*, 2009). Rapid molecular evolution was therefore established for the region. This then lends support to the idea that PPE/PE related genomic loci within our tested strains may also be under similar 'increased' rates of mutational activity.

Also revealed by the SWGB was that the genomes of *M. ulcerans* and *M. marinum* show intermediate states of evolution (based on RV-PS plots) from the pattern of *M. avium* toward that of *M. tb* H37Rv. This was interesting in that it almost summarizes the mutational events occurring as these species evolve.

The SWGB afforded us a unique view of various mycobacterial genomes and highlighted areas of 'non-normal' oligonucleotide usage patterns. Within these gene islands, specific genes were then identified. The mycobacterial comparison project was then used to look deeper into these genes in order to gauge the roles they may play within the mycobacteria.

Continuing from the SWGB findings, the MCP system was used to make more interesting findings. Amongst the elements found within the genomic islands of *M. avium* K10, were *dna*A/B genes. These genes have important roles to play in the bacterial replication system and are thus expected to have high levels of conservation. The MCP system revealed, however, that the genes showed a high-degree of non-conservation among the tested mycobacteria. This may suggest that the replication system of mycobacteria has evolved faster. Whether the changes within this gene set is one of the contributers to the eventual appearance of slow-growing mycobacterial phenotypes remains to be proved.

The SWGB highlighted a high PE-PGRS/PPE gene content within the gene islands of *M. tb* H37Rv. Using the MCP system, a PE-PGRS profile was created for all the mycobacterial genomes within the system. The results achieved were very interesting. Mycobacterial strain H37Ra is essentially a laboratory derivative of *M. tb* H37Rv. As a result, it was expected that the genetic profile of these two strains would be closest to each other. However, the MCP system revealed that *M. tb* H37Rv was actually closer to *M. tb* F11(a virulent strain) in its PE-PGRS profile than it was to H37Ra. This further underscored the roles these genes play in virulence of mycobacteria. It is accepted that the evolutionary pressures experienced by H37Ra is quite different to that of *M. tb* F11. The time frames over which these changes took place are also not clear but what is clear from the literature is that some gene families within the PPE family do give insight into the lineage of various species within the mycobacterial genus (McEvoy *et al.*, 2009).

The ability to identify the similarities and differences between mycobacterial strains is of the utmost importance to understanding the genetic causes of diseases and the differences in disease processes and tropisms exhibited by the various mycobacterial strains. In this project, several novel approaches to identify key areas of interest between various mycobacteria were used. In identifying these areas scientists can now be more focused in their research and solidify the links between clinical phenotypes and genomic loci. TB researchers will find it very useful to be able to now pin-point and analyse specific genomic areas and immediately see how comparative loci in other species compare. This functionality is also helpful in the ability to rapidly detect drug resistant/susceptible phenotypes. In this day and age as high-throughput sequencing techniques becomes more cost effective and faster, comparative sequenced based methods will become a most valuable research and discovery tool.

Bibliography

1. www.genome.gov
2. http://www.ncbi.nlm.nih.gov/
3. www.genomesonline.org/
4. http://camera.calit2.net/index.php/
5. http://www.ncbi.nlm.nih.gov/BLAST
6. www.mysql.com
7. http://www.sqlobject.org/
8. www.sqlalchemy.org
9. www.kid-templating.org
10. www.cherrypy.org
11. www.python.org
12. http://www.bi.up.ac.za/SeqWord/sniffer/
13. www.cbs.dtu.dk/services/GenomeAtlas/
14. http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi
15. SWGB FTP site [ftp://milliways.bi.up.ac.za/SeqWord/GenomeBrowser/]
16. www.clcbio.com
17. http://www.genomenewsnetwork.org/resources/sequenced_genomes/genome_guide_p2.shtml

Abe, T., Kanaya, S., Kinouchi, M., Ichiba, Y., Kozuki, T. & Ikemura, T. (2003) Informatics for unveiling hidden genome signatures. *Genome Res* **13**(4), 693-702.

Achilli, A., Perego, U. A., Bravi, C. M., Coble, M. D., Kong, Q.-P., Woodward, S. R., Salas, A., Torroni, A. & Bandelt, H.-J. (2008) The phylogeny of the four pan-American MtDNA haplogroups: implications for evolutionary and disease studies. *PLoS ONE* **3**(3), e1764.

Ahmadian, A., Gharizadeh, B., Gustafsson, A. C., Sterky, F., Nyrén, P., Uhlén, M. & Lundeberg, J. (2000) Single-nucleotide polymorphism analysis by pyrosequencing. *Anal Biochem* **280**(1), 103-110.

Alland, D., Whittam, T. S., Murray, M. B., Cave, M. D., Hazbon, M. H., Dix, K., Kokoris, M., Duesterhoeft, A., Eisen, J. A., Fraser, C. M. & Fleischmann, R. D. (2003) Modeling bacterial evolution with comparative-genome-based marker systems: application to *Mycobacterium tuberculosis* evolution and pathogenesis. *J Bacteriol* **185**(11), 3392-3399.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) Basic local alignment search tool. *J Mol Biol* **215**(3), 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 Res* **25**(17), 3389-3402.

Aluru, S. (2006) Handbook of Computational Molecular Biology, Chapman & Hall/CRC Computer & Information Science Series.

Azad, R. K. & Lawrence, J. G. (2007) Detecting laterally transferred genes: use of entropic clustering methods and genome position. *Nucleic Acids Res* **35**(14), 4629-4639.

Azad, R. K. & Lawrence, J. G. (2005) Use of artificial genomes in assessing methods for atypical gene detection. *PLoS Comput Biol* **1**(6), e56.

Badger, J. H. & Olsen, G. J. (1999) CRITICA: coding region identification tool invoking comparative analysis. *Mol Biol Evol* **16**(4), 512-524.

Banu, S., Honoré, N., Saint-Joanis, B., Philpott, D., Prévost, M.-C. & Cole, S. T. (2002) Are the PE-PGRS proteins of *Mycobacterium tuberculosis* variable surface antigens?, *Mol Microbiol***44**(1), 9-19.

Bastien, O., Lespinats, S., Roy, S., Métayer, K., Fertil, B., Codani, J.-J. & Maréchal, E. (2004) Analysis of the compositional biases in *Plasmodium falciparum* genome and proteome using *Arabidopsis thaliana* as a reference. *Gene* **336**(2), 163-173.

Becq, J., Gutierrez, M. C., Rosas-Magallanes, V., Rauzier, J., Gicquel, B., Neyrolles, O. & Deschavanne, P. (2007) Contribution of horizontally acquired genomic islands to the evolution of the tubercle bacilli. *Mol Biol Evol* **24**(8), 1861-1871.

Bergh, S. & Cole, S. T. (1994) MycDB: an integrated mycobacterial database. *Mol Microbiol* **12**(4), 517-534.

Bohlin, J., Skjerve, E. & Ussery, D. W. (2008) Investigations of oligonucleotide usage variance within and between prokaryotes. *PLoS Comput Biol* **4**(4), e1000057.

Brosch, R., Philipp, W. J., Stavropoulos, E., Colston, M. J., Cole, S. T. & Gordon, S. V. (1999) Genomic analysis reveals variation between *Mycobacterium tuberculosis* H37Rv and the attenuated *M. tuberculosis* H37Ra strain. *Infect Immun* **67**(11), 5768-5774.

Brosch, R., Pym, A. S., Gordon, S. V. & Cole, S. T. (2001) The evolution of mycobacterial pathogenicity: clues from comparative genomics. *Trends Microbiol* **9**(9), 452-458.

Brosch, S., Häge, A. & Johannsen, H. S. (2002) Prognostic indicators for stuttering: the value of computer-based speech analysis. *Brain Lang* **82**(1), 75-86.

Bulmer, M. (1998) Galtons law of ancestral heredity. *Heredity* **81**(5), 579-585.

Camacho, L. R., Ensergueix, D., Perez, E., Gicquel, B. & Guilhot, C. (1999) Identification of a virulence gene cluster of *Mycobacterium tuberculosis* by signature-tagged transposon mutagenesis. *Mol Microbiol* **34**(2), 257-267.

Cameron, M., Bernstein, Y. & Williams, H. E. (2007) Clustered sequence representation for fast homology search. *J Comput Biol* **14**(5), 594-614.

Cameron, M. & Williams, H. E. (2007) Comparing compressed sequences for faster nucleotide BLAST searches. IEEE/ACM *Trans Comput Biol Bioinform* **4**(3), 349-364.

Camus, J.-C., Pryor, M. J., Médigue, C. & Cole, S. T. (2002) Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. Microbiology **148**(10), 2967-2973.

Catanho, M., Mascarenhas, D., Degrave, W. & de Miranda, A. B. (2006) GenoMycDB: a database for comparative analysis of mycobacterial genes and genomes. *Genet Mol Res* **5**(1), 115-126.

Chain, P., Kurtz, S., Ohlebusch, E. & Slezak, T. (2003), An applications-focused review of comparative genomics tools: capabilities, limitations and future challenges. *Brief Bioinform* **4**(2), 105-123.

Chimara, E., Ferrazoli, L. & Leão, S. C. (2004) *Mycobacterium tuberculosis* complex differentiation using gyrB-restriction fragment length polymorphism analysis. *Mem Inst Oswaldo Cruz* **99**(7), 745-748.

Chimara, E., Giampaglia, C. M. S., Martins, M. C., da Silva Telles, M. A., Ueki, S. Y. M. & Ferrazoli, L. (2004) Molecular characterization of *Mycobacterium kansasii* isolates in the State of São aulo between 1995-1998. *Mem Inst Oswaldo Cruz* **99**(7), 739-743.

Chung, H.-J., Jung, J. D., Park, H.-W., Kim, J.-H., Cha, H. W., Min, S. R., Jeong, W.-J. & Liu, J. R. (2006) The complete chloroplast genome sequences of *Solanum tuberosum* and comparative analysis with Solanaceae species identified the presence of a 241-bp deletion in cultivated potato chloroplast DNA sequence. *Plant Cell Rep* **25**(12), 1369-1379.

Coenye, T. & Vandamme, P. (2004) Bacterial whole-genome sequences: minimal information and strain availability. *Microbiology* **150**(7), 2017-2018.

Coenye, T. & Vandamme, P. (2004) A genomic perspective on the relationship between the Aquificales and the epsilon-Proteobacteria. *Syst Appl Microbiol* **27**(3), 313-322.

Cole, S. T. (2002) Comparative mycobacterial genomics as a tool for drug target and antigen discovery. *Eur Respir J Suppl* **36**, 78s-86s.

Cole, S. T. (1998) Comparative mycobacterial genomics. *Curr Opin Microbiol* **1**(5), 567-571.

Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S., Barry, C. E., Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L., Oliver, K., Osborne, J., Quail, M. A., Rajandream, M. A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, J. E., Taylor, K., Whitehead, S. & Barrell, B. G. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**(6685), 537-544.

Conlon, E. M., Liu, X. S., Lieb, J. D. & Liu, J. S. (2003), Integrating regulatory motif discovery and genome-wide expression analysis. *Proc Natl Acad Sci U S A* **100**(6), 3339-3344.

Couronne, O., Poliakov, A., Bray, N., Ishkhanov, T., Ryaboy, D., Rubin, E., Pachter, L. & Dubchak, I. (2003), Strategies and tools for whole-genome alignments. *Genome Res* **13**(1), 73-80.

Danelishvili, L., Cirillo, S. L. G., Cirillo, J. D. & Bermudez, L. E. (2007) Virulent mycobacteria and the many aspects of macrophage uptake. *Future Microbiol* **2**(5), 461-464.

Darling, A. C. E., Mau, B., Blattner, F. R. & Perna, N. T. (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* **14**(7), 1394-1403.

Deb, C., Daniel, J., Sirakova, T. D., Abomoelak, B., Dubey, V. S. & Kolattukudy, P. E. (2006) A novel lipase belonging to the hormone-sensitive lipase family induced under starvation to utilize stored triacylglycerol in *Mycobacterium tuberculosis*. *J Biol Chem* **281**(7), 3866-3875.

Delcher, A. L., Kasif, S., Fleischmann, R. D., Peterson, J., White, O. & Salzberg, S. L. (1999) Alignment of whole genomes. *Nucleic Acids Res* **27**(11), 2369-2376.

Deschavanne, P. J., Giron, A., Vilain, J., Fagot, G. & Fertil, B. (1999) Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol Biol Evol* **16**(10), 1391-1399.

Dolja, V. V., Kreuze, J. F. & Valkonen, J. P. T. (2006) Comparative and functional genomics of closteroviruses. *Virus Res* **117**(1), 38-51.

Domenech, P., Pym, A. S., Cellier, M., Barry, C. E. & Cole, S. T. (2002) Inactivation of the *Mycobacterium tuberculosis* Nramp orthologue (mntH) does not affect virulence in a mouse model of tuberculosis. *FEMS Microbiol Lett* **207**(1), 81-86.

Domenech, P., Reed, M. B. & Barry, C. E. (2005) Contribution of the *Mycobacterium tuberculosis* MmpL protein family to virulence and drug resistance. *Infect Immun* **73**(6), 3492-3501.

Dufraigne, C., Fertil, B., Lespinats, S., Giron, A. & Deschavanne, P. (2005) Detection and characterization of horizontal transfers in prokaryotes using genomic signature. *Nucleic Acids Res* **33**(1), e6.

Ellsworth, R. E., Jamison, D. C., Touchman, J. W., Chissoe, S. L., Maduro, V. V. B., Bouffard, G. G., Dietrich, N. L., Beckstrom-Sternberg, S. M., Iyer, L. M., Weintraub, L. A., Cotton, M., Courtney, L., Edwards, J., Maupin, R., Ozersky, P., Rohlfing, T., Wohldmann, P., Miner, T., Kemp, K., Kramer, J., Korf, I., Pepin, K., Antonacci-Fulton, L., Fulton, R. S., Minx, P., Hillier, L. W., Wilson, R. K., Waterston, R. H., Miller, W. & Green, E. D. (2000) Comparative genomic sequence analysis of the human and mouse cystic fibrosis transmembrane conductance regulator genes. *Proc Natl Acad Sci U S A* **97**(3), 1172-1177.

Ermolaeva, M. D. (2001) Synonymous codon usage in bacteria. *Curr Issues Mol Biol* **3**(4), 91-97.

Filliol, I., Motiwala, A. S., Cavatore, M., Qi, W., Hazbón, M. H., del Valle, M. B., Fyfe, J., García-García, L., Rastogi, N., Sola, C., Zozio, T., Guerrero, M. I., León, C. I., Crabtree, J., Angiuoli, S., Eisenach, K. D., Durmaz, R., Joloba, M. L., Rendón, A., Sifuentes-Osornio, J., de León, A. P., Cave, M. D., Fleischmann, R., Whittam, T. S. & Alland, D. (2006) Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set. *J Bacteriol* **188**(2), 759-772.

Fleischmann, R. D., Alland, D., Eisen, J. A., Carpenter, L., White, O., Peterson, J., DeBoy, R., Dodson, R., Gwinn, M., Haft, D., Hickey, E., Kolonay, J. F., Nelson, W. C., Umayam, L. A., Ermolaeva, M., Salzberg, S. L., Delcher, A., Utterback, T., Weidman, J., Khouri, H., Gill, J., Mikula, A., Bishai, W., Jr, W. R. J., Venter, J. C. & Fraser, C. M. (2002) Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. *J Bacteriol* **184**(19), 5479-5490.

Frangeul, L., Nelson, K. E., Buchrieser, C., Danchin, A., Glaser, P. & Kunst, F. (1999) Cloning and assembly strategies in microbial genome projects. *Microbiology* **145**(10), 2625-2634.

Fu, L. M. & Fu-Liu, C. S. (2007) The gene expression data of *Mycobacterium tuberculosis* based on Affymetrix gene chips provide insight into regulatory and hypothetical genes. *BMC Microbiol* **7**, 37.

Fu, L. M. & Fu-Liu, C. S. (2002) Genome comparison of *Mycobacterium tuberculosis* and other bacteria. *OMICS* **6**(2), 199-206.

Fu, L. M. & Shinnick, T. M. (2007) Genome-Wide Analysis of Intergenic Regions of *Mycobacterium tuberculosis* H37Rv Using Affymetrix GeneChips. *EURASIP J Bioinform Syst Biol,* **23054**.

Gagneux, S., Burgos, M. V., DeRiemer, K., Encisco, A., Muñoz, S., Hopewell, P. C., Small, P. M. & Pym, A. S. (2006) Impact of bacterial genetics on the transmission of isoniazid-resistant *Mycobacterium tuberculosis*. *PLoS Pathog* **2**(6), e61.

Galves, M., Quitzau, J. A. A. & Dias, Z. (2006) New strategy to detect single nucleotide polymorphisms. *Genet Mol Res* **5**(1), 143-153.

Ganesan, H., Rakitianskaia, A. S., Davenport, C. F., Tümmler, B. & Reva, O. N. (2008) The SeqWord Genome Browser: an online tool for the identification and visualization of atypical regions of bacterial genomes through oligonucleotide usage. *BMC Bioinformatics* **9**, 333.

Garnier, T., Eiglmeier, K., Camus, J.-C., Medina, N., Mansoor, H., Pryor, M., Duthoy, S., Grondin, S., Lacroix, C., Monsempe, C., Simon, S., Harris, B., Atkin, R., Doggett, J., Mayes, R., Keating, L., Wheeler, P. R., Parkhill, J., Barrell, B. G., Cole, S. T., Gordon, S. V. & Hewinson, R. G. (2003), The complete genome sequence of *Mycobacterium bovis*. *Proc Natl Acad Sci U S A* **100**(13), 7877-7882.

Gutacker, M. M., Smoot, J. C., Migliaccio, C. A. L., Ricklefs, S. M., Hua, S., Cousins, D. V., Graviss, E. A., Shashkina, E., Kreiswirth, B. N. & Musser, J. M. (2002) Genome-wide analysis of synonymous single nucleotide polymorphisms in *Mycobacterium tuberculosis* complex organisms: resolution of genetic relationships among closely related microbial strains. *Genetics* **162**(4), 1533-1543.

Guyon, F. & Guénochea, A. (2008) Comparing bacterial genomes from linear orders of patterns, *Discrete Applied Mathematics* **156**, 1251-1262.

Hall, N., Karras, M., Raine, J. D., Carlton, J. M., Kooij, T. W. A., Berriman, M., Florens, L., Janssen, C. S., Pain, A., Christophides, G. K., James, K., Rutherford, K., Harris, B., Harris, D., Churcher, C., Quail, M. A., Ormond, D., Doggett, J., Trueman, H. E., Mendoza, J., Bidwell, S. L., Rajandream, M.-A., Carucci, D. J., Yates, J. R., Kafatos, F. C., Janse, C. J., Barrell, B., Turner, C. M. R., Waters, A. P. & Sinden, R. E. (2005) A comprehensive survey of the Plasmodium life cycle by genomic, transcriptomic, and proteomic analyses. *Science* **307**(5706), 82-86.

Hallin, P. F., Binnewies, T. T. & Ussery, D. W. (2008) The genome BLASTatlas-a GeneWiz extension for visualization of whole-genome homology. *Mol Biosyst* **4**(5), 363-371.

Harper, K. N., Liu, H., Ocampo, P. S., Steiner, B. M., Martin, A., Levert, K., Wang, D., Sutton, M. & Armelagos, G. J. (2008) The sequence of the acidic repeat protein (arp) gene differentiates venereal from nonvenereal *Treponema pallidum subspecies*, and the gene has evolved under strong

positive selection in the subspecies that causes syphilis. *FEMS Immunol Med Microbiol* **53**(3), 322-332.

Harper, K. N., Ocampo, P. S., Steiner, B. M., George, R. W., Silverman, M. S., Bolotin, S., Pillay, A., Saunders, N. J. & Armelagos, G. J. (2008) On the origin of the treponematoses: a phylogenetic approach. *PLoS Negl Trop Dis* **2**(1), e148.

Hartmans, S., Bont, J. A. M. D. & Stackerbrandt, E. (2006) The Genus Mycobacterium—Nonmedical, Prokaryotes **3**, 889-918.

Hasin, Y., Avidan, N., Bercovich, D., Korczyn, A., Silman, I., Beckmann, J. S. & Sussman, J. L. (2004) A paradigm for single nucleotide polymorphism analysis: the case of the acetylcholinesterase gene. *Hum Mutat* **24**(5), 408-416.

Huelsenbeck, J. P. & Ronquist, F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**(8), 754-755.

Jernigan, R. W. & Baran, R. H. (2002) Pervasive properties of the genomic signature. *BMC Genomics* **3**(1), 23.

Jones, A. R., Miller, M., Aebersold, R., Apweiler, R., Ball, C. A., Brazma, A., Degreef, J., Hardy, N., Hermjakob, H., Hubbard, S. J., Hussey, P., Igra, M., Jenkins, H., Julian, R. K., Laursen, K., Oliver, S. G., Paton, N. W., Sansone, S.-A., Sarkans, U., Stoeckert, C. J., Taylor, C. F., Whetzel, P. L., White, J. A., Spellman, P. & Pizarro, A. (2007) The Functional Genomics Experiment model (FuGE): an extensible framework for standards in functional genomics. *Nat Biotechnol* **25**(10), 1127-1133.

Jones, A. R., Pizarro, A., Spellman, P., Miller, M. & Group, F. E. W. (2006) FuGE: Functional Genomics Experiment Object Model. *OMICS* **10**(2), 179-184.

Jones, C. E., Baumann, U. & Brown, A. L. (2005) Automated methods of predicting the function of biological sequences using GO and BLAST. *BMC Bioinformatics* **6**, 272.

Kamvysselis, M., Patterson, N., Birren, B., Berger, B. & Lander, E. (2003), Whole-genome Comparative Annotation and Regulatory Motif Discovery in Multiple Yeast Species, *Recomb*, 10-13.

Karlin, S. (1998) Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr Opin Microbiol* **1**(5), 598-610.

Karlin, S. & Burge, C. (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet* **11**(7), 283-290.

Karlin, S., Campbell, A. M. & Mrázek, J. (1998) Comparative DNA analysis across diverse genomes. *Annu Rev Genet* **32**, 185-225.

Kato-Maeda, M., Bifani, P. J., Kreiswirth, B. N. & Small, P. M. (2001) The nature and consequence of genetic variability within *Mycobacterium tuberculosis*. *J Clin Invest* **107**(5), 533-537.

Kato-Maeda, M., Rhee, J. T., Gingeras, T. R., Salamon, H., Drenkow, J., Smittipat, N. & Small, P. M. (2001) Comparing genomes within the species *Mycobacterium tuberculosis*. *Genome Res* **11**(4), 547-554.

Kellis, M., Patterson, N., Birren, B., Berger, B. & Lander, E. S. (2004) Methods in comparative genomics: genome correspondence, gene identification and regulatory motif discovery. *J Comput Biol* **11**(2-3), 319-355.

Kiewitz, C., Larbig, K., Klockgether, J., Weinel, C. & Tümmler, B. (2000) Monitoring genome evolution ex vivo: reversible chromosomal integration of a 106 kb plasmid at two tRNA(Lys) gene loci in sequential *Pseudomonas aeruginosa* airway isolates. *Microbiology* **146**(10), 2365-2373.

Kiewitz, C. & Tümmler, B. (2000) Sequence diversity of *Pseudomonas aeruginosa*: impact on population structure and genome evolution. *J Bacteriol* **182**(11), 3125-3135.

Klockgether, J., Reva, O., Larbig, K. & Tümmler, B. (2004) Sequence analysis of the mobile genome island pKLC102 of *Pseudomonas aeruginosa C*. *J Bacteriol* **186**(2), 518-534.

Koski, L. B., Morton, R. A. & Golding, G. B. (2001) Codon bias and base composition are poor indicators of horizontally transferred genes. *Mol Biol Evol* **18**(3), 404-412.

Kubo, T. & Newton, K. J. (2008) Angiosperm mitochondrial genomes and mutations. *Mitochondrion* **8**(1), 5-14.

Kumar, S., Nei, M., Dudley, J. & Tamura, K. (2008) MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. *Brief Bioinform* **9**(4) 299-306.

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C. & Salzberg, S. L. (2004) Versatile and open software for comparing large genomes. *Genome Biol* **5**(2), R12

Lawrence, J. G., Hendrix, R. W. & Casjens, S. (2001) Where are the pseudogenes in bacterial genomes?, *Trends Microbiol* **9**(11), 535-540.

Lawrence, J. G. & Ochman, H. (1997) Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol* **44**(4), 383-397.

Li, W. & Godzik, A. (2006) VISSA: a program to visualize structural features from structure sequence alignment. *Bioinformatics* **22**(7), 887-888.

Lyons, E. & Freeling, M. (2008) How to usefully compare homologous plant genes and chromosomes as DNA sequences. *Plant J* **53**(4), 661-673.

Mallon, A. M., Platzer, M., Bate, R., Gloeckner, G., Botcherby, M. R., Nordsiek, G., Strivens, M. A., Kioschis, P., Dangel, A., Cunningham, D., Straw, R. N., Weston, P., Gilbert, M., Fernando, S., Goodall, K., Hunter, G., Greystrong, J. S., Clarke, D., Kimberley, C., Goerdes, M., Blechschmidt, K., Rump, A., Hinzmann, B., Mundy, C. R., Miller, W., Poustka, A., Herman, G. E., Rhodes, M., Denny, P., Rosenthal, A. & Brown, S. D. (2000) Comparative genome sequence analysis of the Bpa/Str region in mouse and Man. *Genome Res* **10**(6), 758-775.

Marmiesse, M., Brodin, P., Buchrieser, C., Gutierrez, C., Simoes, N., Vincent, V., Glaser, P., Cole, S. T. & Brosch, R. (2004) Macro-array and bioinformatic analyses reveal mycobacterial core genes, variation in the ESAT-6 gene family and new phylogenetic markers for the *Mycobacterium tuberculosis* complex. *Microbiology* **150**(2), 483-496.

Marri, P. R., Bannantine, J. P. & Golding, G. B. (2006) Comparative genomics of metabolic pathways in Mycobacterium species: gene duplication, gene decay and lateral gene transfer. *FEMS Microbiol Rev* **30**(6), 906-925.

McCarren, J. & Brahamsha, B. (2005) Transposon mutagenesis in a marine synechococcus strain: isolation of swimming motility mutants. *J Bacteriol* **187**(13), 4457-4462.

McDonough, K. A., Kress, Y. & Bloom, B. R. (1993), The interaction of *Mycobacterium tuberculosis* with macrophages: a study of phagolysosome fusion. *Infect Agents Dis* **2**(4), 232-235.

McDonough, K. A., Kress, Y. & Bloom, B. R. (1993), Pathogenesis of tuberculosis: interaction of *Mycobacterium tuberculosis* with macrophages. *Infect Immun* **61**(7), 2763-2773.

McEvoy, C. R. E., van Helden, P. D., Warren, R. M. & van Pittius, N. C. G. (2009) Evidence

for a rapid rate of molecular evolution at the hypervariable and immunogenic Mycobacterium tuberculosis PPE38 gene region. *BMC Evolutionary Biology* **9**, 237-248.

Medrano-Soto, A., Moreno-Hagelsieb, G., Vinuesa, P., Christen, J. A. & Collado-Vides, J. (2004) Successful lateral transfer requires codon usage compatibility between foreign genes and recipient genomes. *Mol Biol Evol* **21**(10), 1884-1894.

Moxon, E. R., Hood, D. W., Saunders, N. J., Schweda, E. K. H. & Richards, J. C. (2002) Functional genomics of pathogenic bacteria. *Philos Trans R Soc Lond B Biol Sci* **357**(1417), 109-116.

Mrázek, J. & Karlin, S. (1999) Detecting alien genes in bacterial genomes. *Ann N Y Acad Sci* **870**, 314-329.

Mulder, N., Rabiu, H., Jamieson, G. & Vuppu, V. (2007) Comparative analysis of microbial genomes to study unique and expanded gene families in *Mycobacterium tuberculosis*. *Infect Genet Evol*.

Muller, J., Oma, Y., Vallar, L., Friederich, E., Poch, O. & Winsor, B. (2005) Sequence and comparative genomic analysis of actin-related proteins. *Mol Biol Cell* **16**(12), 5736-5748.

Nakamura, Y., Itoh, T., Matsuda, H. & Gojobori, T. (2004) Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat Genet* **36**(7), 760-766.

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**(3), 443-453.

Nouvel, L. X., Kassa-Kelembho, E., Vultos, T. D., Zandanga, G., Rauzier, J., Lafoz, C., Martin, C., Blazquez, J., Talarmin, A. & Gicquel, B. (2006) Multidrug-resistant *Mycobacterium tuberculosis*, Bangui, Central African Republic. *Emerg Infect Dis* **12**(9), 1454-1456.

Okkels, L. M., Brock, I., Follmann, F., Agger, E. M., Arend, S. M., Ottenhoff, T. H. M., Oftung, F., Rosenkrands, I. & Andersen, P. (2003), PPE protein (Rv3873) from DNA segment RD1 of *Mycobacterium tuberculosis*: strong recognition of both specific T-cell epitopes and epitopes conserved within the PPE family. *Infect Immun* **71**(11), 6116-6123.

Ozaki, K., Sato, H., Iida, A., Mizuno, H., Nakamura, T., Miyamoto, Y., Takahashi, A., Tsunoda, T., Ikegawa, S., Kamatani, N., Hori, M., Nakamura, Y. & Tanaka, T. (2006) A functional SNP in PSMA6 confers risk of myocardial infarction in the Japanese population. *Nat Genet* **38**(8), 921-925.

Pan, X., Stein, L. & Brendel, V. (2005) SynBrowse: a synteny browser for comparative sequence analysis. *Bioinformatics* **21**(17), 3461-3468.

Philipp, W. J., Gordon, S., Telenti, A. & Cole, S. T. (1998) Pulsed field gel electrophoresis for mycobacteria. *Methods Mol Biol* **101**, 51-63.

Philipp, W. J., Nair, S., Guglielmi, G., Lagranderie, M., Gicquel, B. & Cole, S. T. (1996) Physical mapping of *Mycobacterium bovis* BCG pasteur reveals differences from the genome map of *Mycobacterium tuberculosis* H37Rv and from *M. bovis*. *Microbiology* **142**(11), 3135-3145.

Philipp, W. J., Schwartz, D. C., Telenti, A. & Cole, S. T. (1998) Mycobacterial genome structure. *Electrophoresis* **19**(4), 573-576.

van Pittius, N. C. G., Sampson, L. S., Lee, H., Kim Y., van Helden, P. D. & Warren, R. M. (2006) Evolution and expansion of the Mycobacterium tuberculosis PE and PPE multigene families and their association with the duplication of the ESAT-6 (esx) gene cluster regions. *BMC Evolutionary Biology* **6**, 95-126.

Pizarro, A., Jones, A., Spellman, P., Miller, M., Whetzel, P. & working group, F. (2006) Extensible Framework for Standards in Functional Genomics.

Pride, D. T. & Blaser, M. J. (2002) Concerted evolution between duplicated genetic elements in *Helicobacter pylori*. *J Mol Biol* **316**(3), 629-642.

Pride, D. T., Meinersmann, R. J., Wassenaar, T. M. & Blaser, M. J. (2003), Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res* **13**(2), 145-158.

Puigbò, P., Bravo, I. G. & Garcia-Vallve, S. (2008) CAIcal: a combined set of tools to assess codon usage adaptation. *Biol Direct* **3**, 38.

Puigbò, P., Bravo, I. G. & Garcia-Vallvé, S. (2008) E-CAI: a novel server to estimate an expected value of Codon Adaptation Index (eCAI). *BMC Bioinformatics* **9**, 65.

Pym, A. S., Brodin, P., Brosch, R., Huerre, M. & Cole, S. T. (2002) Loss of RD1 contributed to the attenuation of the live tuberculosis vaccines *Mycobacterium bovis* BCG and *Mycobacterium microti*. *Mol Microbiol* **46**(3), 709-717.

Pym, A. S., Saint-Joanis, B. & Cole, S. T. (2002) Effect of katG mutations on the virulence

of *Mycobacterium tuberculosis* and the implication for transmission in humans. *Infect Immun* **70**(9), 4955-4960.

Raman, S., Hazra, R., Dascher, C. C. & Husson, R. N. (2004) Transcription regulation by the *Mycobacterium tuberculosis* alternative sigma factor SigD and its role in virulence. *J Bacteriol* **186**(19), 6605-6616.

Raman, S., Puyang, X., Cheng, T.-Y., Young, D. C., Moody, D. B. & Husson, R. N. (2006) *Mycobacterium tuberculosis* SigM positively regulates Esx secreted protein and nonribosomal peptide synthetase genes and down regulates virulence-associated surface lipid synthesis. *J Bacteriol* **188**(24), 8460-8468.

Ranjan, S., Gundu, R. K. & Ranjan, A. (2006) MycoperonDB: a database of computationally identified operons and transcriptional units in Mycobacteria. *BMC Bioinformatics* **7** Suppl 5, S9.

Reiter, L. T., Potocki, L., Chien, S., Gribskov, M. & Bier, E. (2001) A systematic analysis of human disease-associated gene sequences in *Drosophila melanogaster*. *Genome Res* **11**(6), 1114-1125.

Reva, O. & Tümmler, B. (2008) Think big-giant genes in bacteria. *Environ Microbiol* **10**(3), 768-777.

Reva, O. N., Hallin, P. F., Willenbrock, H., Sicheritz-Ponten, T., Tümmler, B. & Ussery, D. W. (2008) Global features of the *Alcanivorax borkumensis* SK2 genome. *Environ Microbiol* **10**(3), 614-625.

Reva, O. N. & Tümmler, B. (2005) Differentiation of regions with atypical oligonucleotide composition in bacterial genomes. *BMC Bioinformatics* **6**, 251.

Reva, O. N. & Tümmler, B. (2004) Global features of sequences of bacterial chromosomes, plasmids and phages revealed by analysis of oligonucleotide usage patterns. *BMC Bioinformatics* **5**, 90.

Reva, O. N., Weinel, C., Weinel, M., Böhm, K., Stjepandic, D., Hoheisel, J. D. & Tümmler, B. (2006) Functional genomics of stress response in *Pseudomonas putida* KT2440. *J Bacteriol* **188**(11), 4079-4092.

Rocha, E. P., Viari, A. & Danchin, A. (1998) Oligonucleotide bias in *Bacillus subtilis*: general trends and taxonomic comparisons. *Nucleic Acids Res* **26**(12), 2971-2980.

Rodríguez, J. C., Jennings, P. A. & Melacini, G. (2002) Effect of chemical exchange on radiation damping in aqueous solutions of the osmolyte glycine. *J Am Chem Soc* **124**(22), 6240-6241.

Sanger, F., Nicklen, S. & Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**(12), 5463-5467.

Saski, C., Lee, S.-B., Daniell, H., Wood, T. C., Tomkins, J., Kim, H.-G. & Jansen, R. K. (2005) Complete chloroplast genome sequence of *Gycine max* and comparative analyses with other legume genomes. *Plant Mol Biol* **59**(2), 309-322.

Schwartz, S., Kent, W. J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R. C., Haussler, D. & Miller, W. (2003), Human-mouse alignments with BLASTZ. *Genome Res* **13**(1), 103-107.

Schwartz, S., Zhang, Z., Frazer, K. A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R. & Miller, W. (2000) PipMaker-a web server for aligning two genomic DNA sequences. *Genome Res* **10**(4), 577-586.

See, D. R., Brooks, S., Nelson, J. C., Brown-Guedira, G., Friebe, B. & Gill, B. S. (2006) Gene evolution at the ends of wheat chromosomes. *Proc Natl Acad Sci U S A* **103**(11), 4162-4167.

Sharp, P. M. & Li, W. H. (1987) The codon Adaptation Index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15**(3), 1281-1295.

Soderlund, C., Nelson, W., Shoemaker, A. & Paterson, A. (2006) SyMAP: A system for discovering and viewing syntenic regions of FPC maps. *Genome Res* **16**(9), 1159-1168.

Srinivasachary; Dida, M. M., Gale, M. D. & Devos, K. M. (2007) Comparative analyses reveal high levels of conserved colinearity between the finger millet and rice genomes. *Theor Appl Genet* **115**(4), 489-499.

Stein, L. D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J. E., Harris, T. W., Arva, A. & Lewis, S. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res* **12**(10), 1599-1610.

Teeling, H., Meyerdierks, A., Bauer, M., Amann, R. & Glöckner, F. O. (2004) Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ Microbiol* **6**(9), 938-947.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994) Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Comput Appl Biosci* **10**(1), 19-29.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**(22), 4673-4680.

Treangen, T. J. & Messeguer, X. (2006) M-GCAT: interactively and efficiently constructing large-scale multiple genome comparison frameworks in closely related species. *BMC Bioinformatics* **7**, 433.

Tsolaki, A. G., Gagneux, S., Pym, A. S., de la Salmoniere, Y.-O. L. G., Kreiswirth, B. N., Soolingen, D. V. & Small, P. M. (2005) Genomic deletions classify the Beijing/W strains as a distinct genetic lineage of *Mycobacterium tuberculosis*. *J Clin Microbiol* **43**(7), 3185-3191.

Ureta-Vidal, A., Ettwiller, L. & Birney, E. (2003), Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat Rev Genet* **4**(4), 251-262.

Vishnoi, A., Roy, R. & Bhattacharya, A. (2007) Comparative analysis of bacterial genomes: identification of divergent regions in mycobacterial strains using an anchor-based approach. *Nucleic Acids Res* **35**(11), 3654-3667.

Vishnoi, A., Srivastava, A., Roy, R. & Bhattacharya, A. (2008) MGDD: Mycobacterium tuberculosis genome divergence database. *BMC Genomics* **9**, 373-377.

Voskuil, M. I. (2004) *Mycobacterium tuberculosis* gene expression during environmental conditions associated with latency. *Tuberculosis (Edinb)* **84(3-4)**, 138-143.

Wang, B. (2001) Limitations of compositional approach to identifying horizontally transferred genes. *J Mol Evol* **53**(3), 244-250.

Waterhouse, R. M., Kriventseva, E. V., Meister, S., Xi, Z., Alvarez, K. S., Bartholomay, L. C., Barillas-Mury, C., Bian, G., Blandin, S., Christensen, B. M., Dong, Y., Jiang, H., Kanost, M. R., Koutsos, A. C., Levashina, E. A., Li, J., Ligoxygakis, P., Maccallum, R. M., Mayhew, G. F., Mendes, A., Michel, K., Osta, M. A., Paskewitz, S., Shin, S. W., Vlachou, D., Wang, L., Wei, W., Zheng, L., Zou, Z., Severson, D. W., Raikhel, A. S., Kafatos, F. C., Dimopoulos, G., Zdobnov, E. M. & Christophides, G. K. (2007) Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. *Science* **316**(5832), 1738-1743.

Waterman, M. S. (1984) Efficient sequence alignment algorithms. *J Theor Biol* **108**(3), 333-337.

Waterman, M. S., Arratia, R. & Galas, D. J. (1984) Pattern recognition in several sequences: consensus and alignment. *Bull Math Biol* **46**(4), 515-527.

Waterston, R. H., Lander, E. S. & Sulston, J. E. (2002) On the sequencing of the human genome. *Proc Natl Acad Sci U S A* **99**(6), 3712-3716.

Weinel, C., Nelson, K. E. & Tümmler, B. (2002) Global features of the *Pseudomonas putida* KT2440 genome sequence. *Environ Microbiol* **4**(12), 809-818.

White, P. S., Kwok, P.-Y., Oefner, P. & Brookes, A. J. (2001) 3rd International Meeting on Single Nucleotide Polymorphism and Complex Genome Analyses: SNPs: Some Notable Progress, *European Journal of Human Genetics* **9**, 316-318.

Zhang, H., Duan, X., Yuan, Z., Li, W., Zhou, G., Zhou, Q., Bing, L., Min, F., Li, X. & Xie, Y. (2006) Chromosomal aberrations induced by (12)C6+ ions and 60Co gamma-rays in mouse immature oocytes. *Mutat Res* **595**(1-2), 37-41.

Zheng, H., Lu, L., Wang, B., Pu, S., Zhang, X., Zhu, G., Shi, W., Zhang, L., Wang, H., Wang, S., Zhao, G. & Zhang, Y. (2008) Genetic Basis of Virulence Attenuation Revealed by Comparative Genomic Analysis of *Mycobacterium tuberculosis* Strain H37Ra versus H37Rv, *PLoS One* **3**, 2375.