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