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).

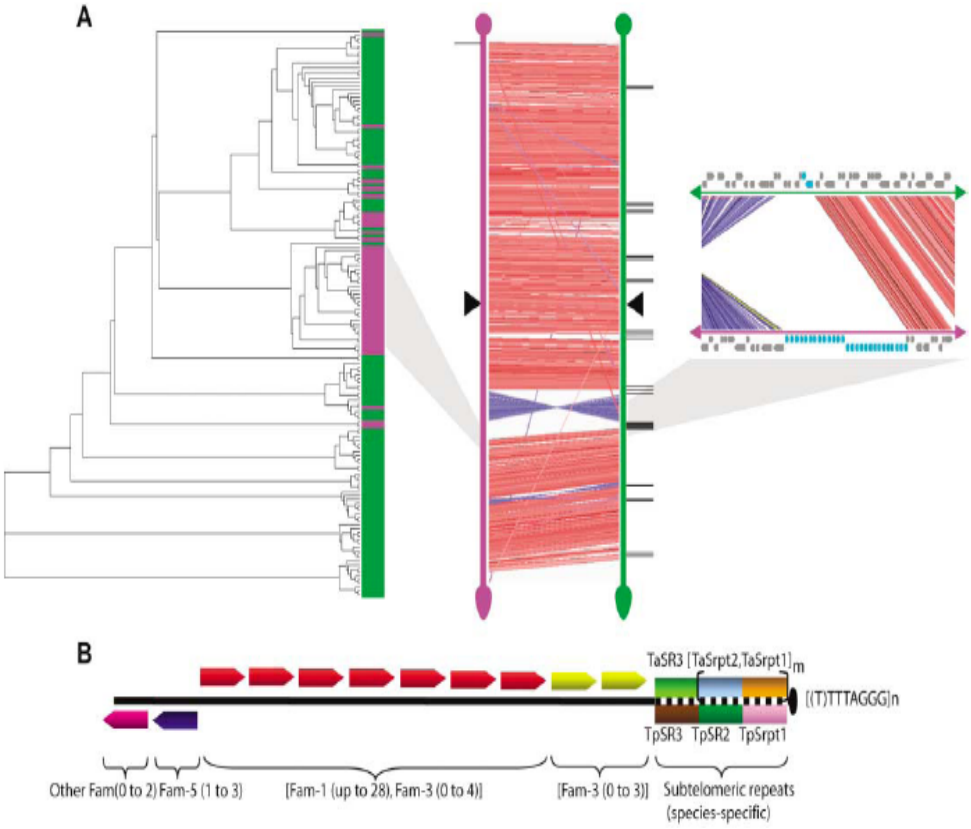

Figure 1.1: Large-scale synteny between *T. annulata* (TA) and *T. parva* (TP) chromosomes. (A) Synteny breaks of chromosome 3 of TA (green) and TP (purple) are located at Tpr genes. (Middle) Chromosome 3 of TA and chromosome 3 of TP are aligned. Connecting lines show maximal unique matches between the 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.

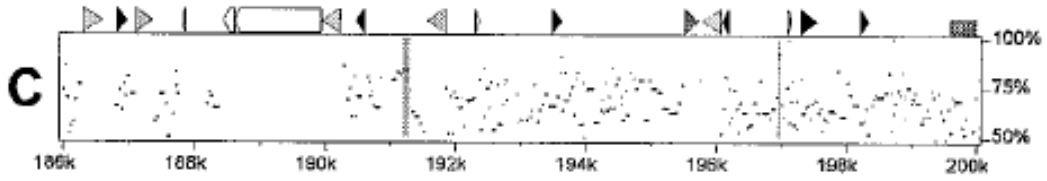

Figure 1.2: Percent identity plots (PIP) for region immediately upstream of CFTR/Cftr exon 1 (nucleotides 5,425–19,425). The vertical stripes are used to highlight the gap-free regions in a '28-kb interval encompassing CFTR/Cftr exon 1 that have a higher percent identity than other gap-free regions in 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.

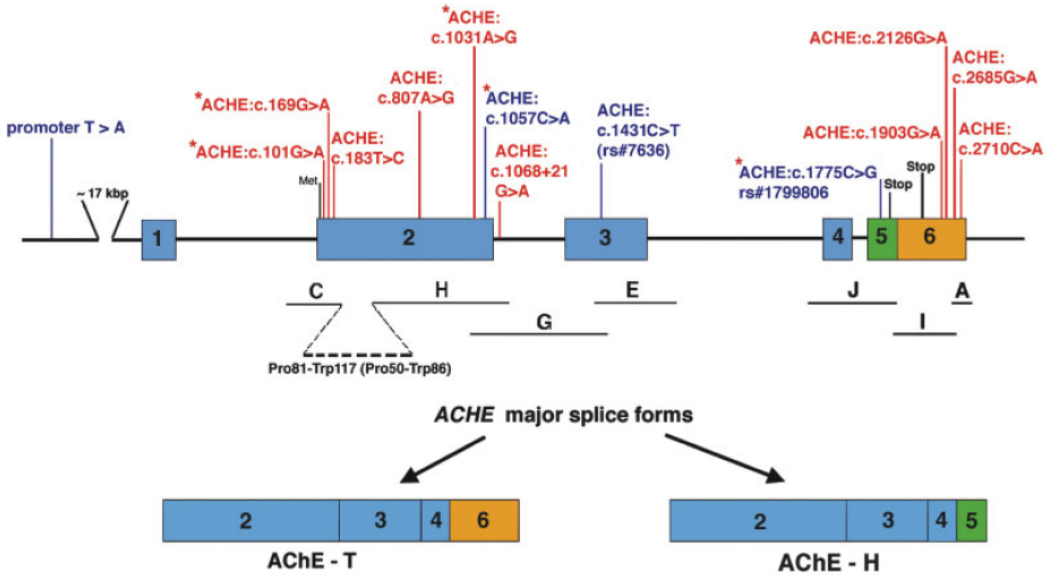

Figure 1.3: Polymorphisms and genomic organization of ACHE. Human AChE is encoded by a single ACHE gene composed of six exons which generates two major alternatively spliced forms that differ in quaternary structure and tissue distribution. Exons are depicted as boxes, and labeled from1 to 6. Previously 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

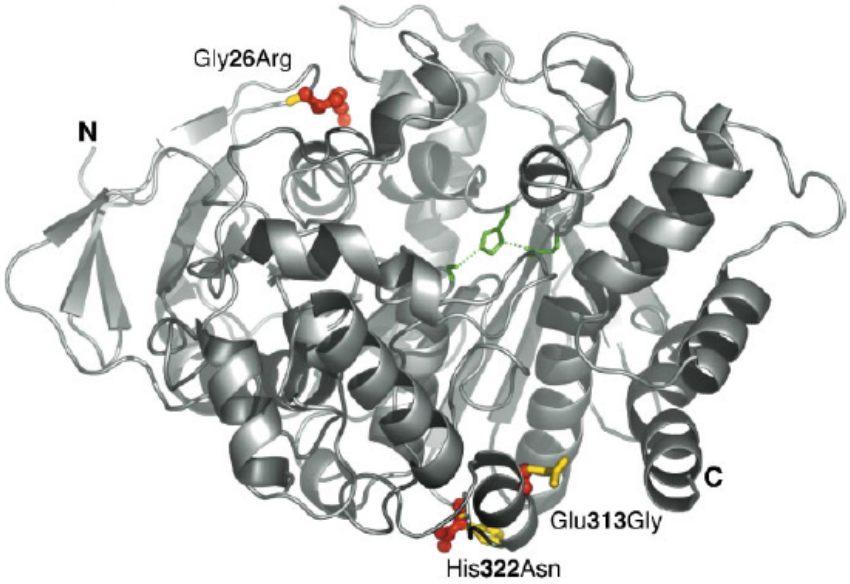

Figure 1.4: Of the 5 nsSNPs (namely ACHE:c.169G4A; ACHE:c.1031A4G and ACHE:c.1057-C4A) were even able to be mapped directly onto the protein structure (Hasin *et al.*, 2004).

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.

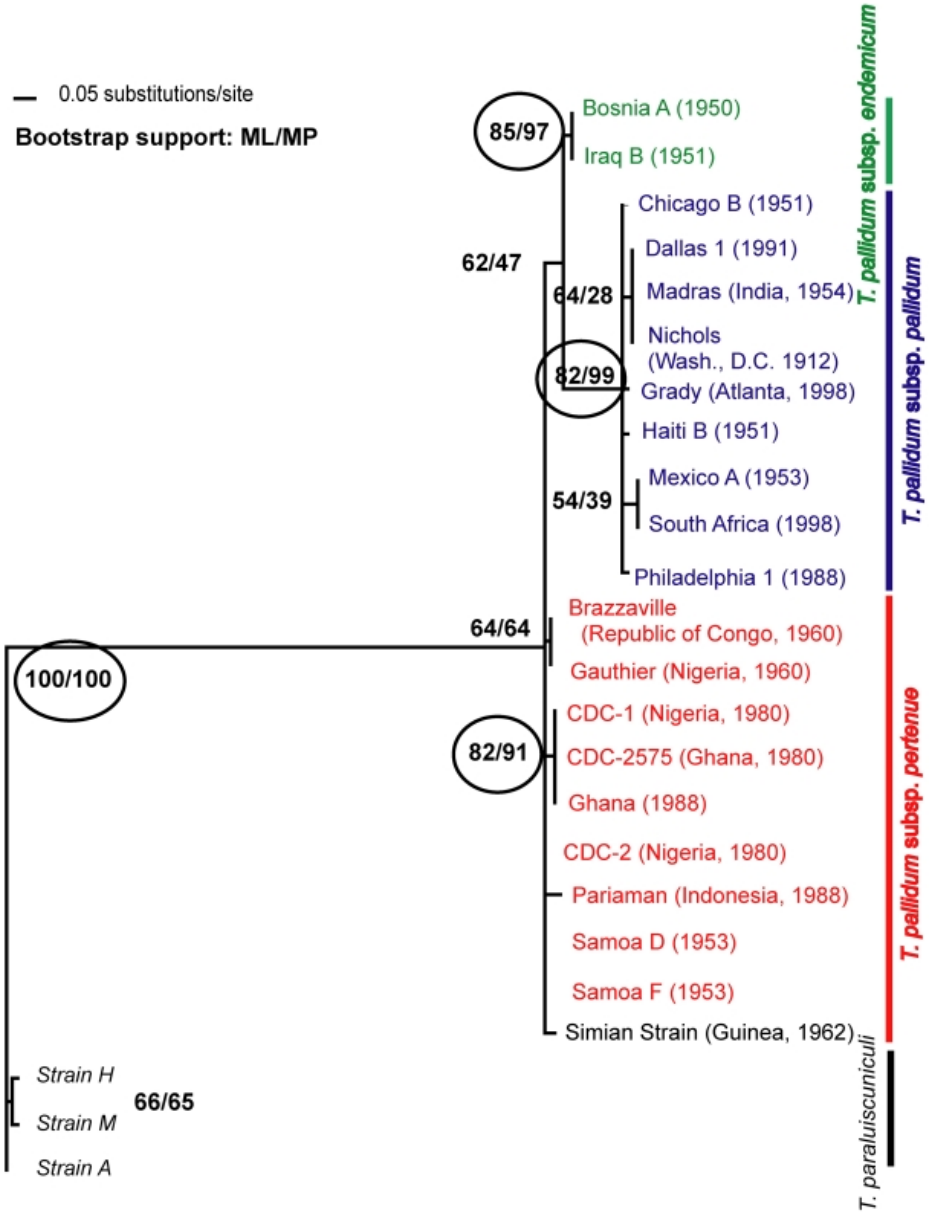

Figure 1.5: Maximum likelihood phylogenetic tree depicting the relationships between the *T. pallidum* subspecies. This tree is based on 20 polymorphic regions in the *T. pallidum* genome. Bootstrap support was estimated with 1,000 replicates in order to assess confidence at branching points and are 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

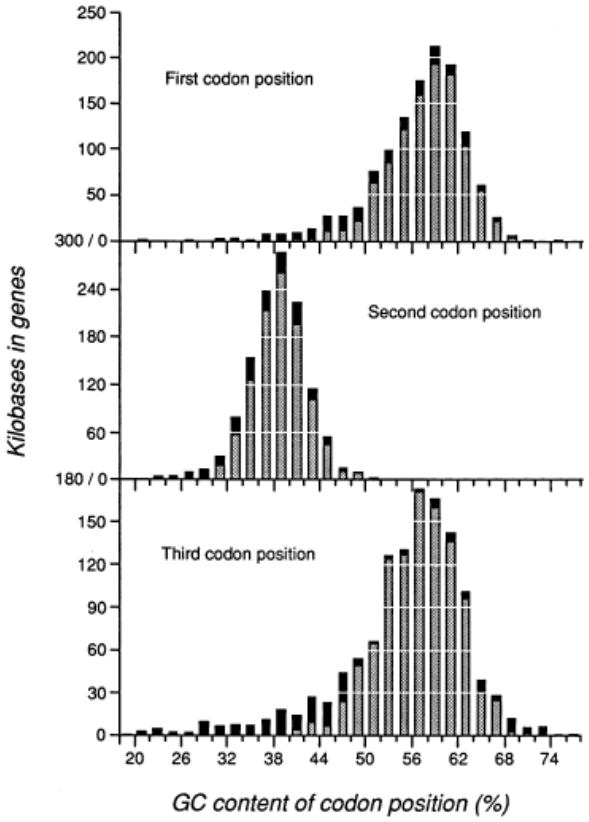

Figure 1.7: GC contents of 1,294 *E. coli* genes. Gray bars denote native genes and black bars denote genes that are supposedly acquired by horizontal transfer (Lawrence *et al.*, 1997).

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.

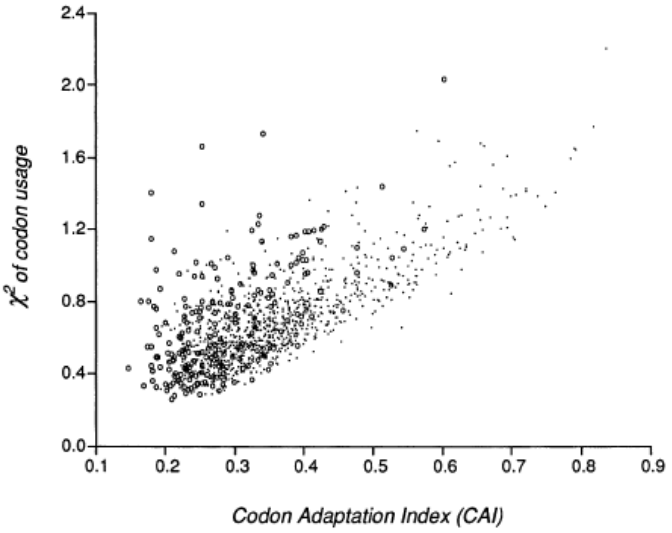

Figure 1.8: Plot of CAI vs χ^2 of codon usage for 1,189 E. coli genes. Points (n=1,024) represents native E. coli genes and open circles (n=165) represents genes inferred to be present in E. coli due to horizontal transfer (Lawrence *et al.*, 1997).

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.

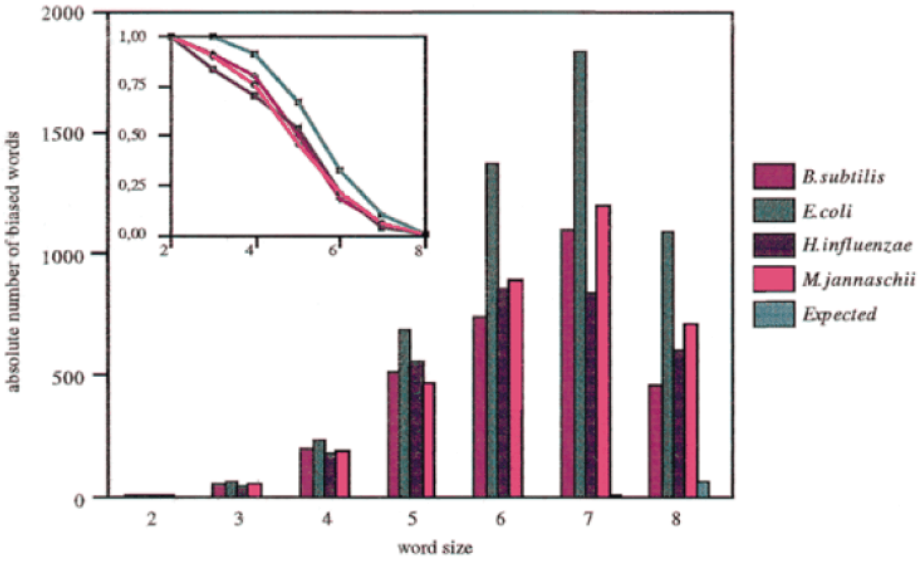

Figure 1.9: Graph depicting the total counts of biased words contained in the single-strand chromosome among the four species tested. Insert displays the relative number of biased words (i.e ratio of the number of biased words to the total number of possible words of that length) (Rocha *et al.*, 1998).

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.