UNIVERSITY OF PRETORIA

Mathematical modeling of evolutionary changes of oligonucleotide frequency patterns of bacterial genomes for genome-scale phylogenetic inferences

PhD Dissertation

Xiaoyu Yu

9/30/2018

Submitted in partial fulfilment of the degree: PhD Bioinformatics, Bioinformatics and Computational Biology Unit, Department of Biochemistry School of Biological Sciences, Faculty Natural and Agricultural Science University of Pretoria

Table of Contents

Chapter 3. Creation and Comparison of OUP-based Tree to Common Phylogenetic Inferences

Chapter 4. Design and Implementation of the Program SeqWord Phylogenomics

Declaration

I, **Xiaoyu Yu** declare that the thesis, which I hereby submit for the degree **PhD Bioinformatics** 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: 30th September 2018

PLAGIARISM STATEMENT

UNIVERSITY OF PRETORIA
FACULTY OF NATURAL AND AGRICULTURAL SCIENCES
DEPARTMENT OF BIOCHEMISTRY
CENTRE FOR BIOINFORMATICS AND COMPUTATIONAL BIOLOGY

Full name: Xiaoyu Yu
Student number: 29020230
Title of the work: Mathematical modeling of evolutionary changes of oligonucleotide frequency patterns of bacterial genomes for genome-scale phylogenetic inferences

Declaration
1. I understand what plagiarism entails and I am aware of the University's policy in this regard.
2. I declare that this thesis is my original work. Where someone else's work was used (whether from a printed source, the internet or any other source) due acknowledgment was given and reference was made according to departmental requirements.
3. I did not make use of another student's previous work and submit it as my own.
4. I did not allow and will not allow anyone to copy my work with the intention of presenting it as his or her work.

Signature: ………………………….
Date: 30th September 2018

Abbreviation

HGT – Horizontal Gene Transfer

OTU - Operational Taxonomic Units

ANI – Average Nucleotide Identity

DDH - DNA-DNA hybridization

ML – Maximum Likelihood

MCMC – Markov Chain Monte Carlo

UCE - Ultra-Conserved Elements

MSA – Multiple Sequence Alignment

ILS – Incomplete Lineage Sorting

LRT – Likelihood Ratio Test

AIC – Akaike Information Criterion

BIC – Bayesian Information Criterion

OUP – Oligonucleotide Usage Pattern

PS – Pattern Skew

OUV – Oligonucleotide Usage Variance

TUD - Tetranucleotide Usage Deviation

CDS – Coding DNA Sequences

COG – Cluster of Orthologous Gene

WGS – Whole Genome Supermatrix

NJ – Neighbour Joining

GUI – Graphical User Interface

BSD – Branch Score Distance

HL – High Light

LL – Low Light

List of Figures

List of Tables

List of Supplementary Figures and Tables

Summary

Modern phylogenetic studies from the advancement of next generation sequencing can benefit from an analysis of complete genome sequences of various microorganisms. Evolutionary inferences based on genome scale analysis were believed to be more accurate than gene-based ones. However, the computational complexity of current phylogenomic procedures and lack of reliable annotation and alignment free evolutionary models keep microbiologists from wider use of these opportunities. For example, the super-matrix approach of phylogenomics requires identification of clusters of orthologous genes in compared genomes followed by alignment of numerous sequences to proceed with reconciliation of multiple trees inferred by traditional phylogenetic tools. In fact, the approach potentially multiplies the problems of gene annotation and sequence alignment, not mentioning the computational difficulties and laboriousness of the methods. For this research, we identified that the alignment and annotation-free method based on comparison of oligonucleotide usage patterns (OUP) calculated for genome-scale DNA sequences allowed fast inferring of phylogenetic trees. These were also congruent with the corresponding whole genome supermatrix trees in terms of tree topology and branch lengths. Validation and benchmarking tests for OUP phylogenomics were done based on comparisons to current literature and artificially created sequences with known phylogeny. It was demonstrated that the OUP diversification between taxa was driven by global adjustments of codon usage to fit fluctuating tRNA concentrations that were well aligned to the species evolution. A web-based program to perform OUP-based phylogenomics was released on http://swphylo.bi.up.ac.za/. Applicability of the tool was proven for different taxa from species to family levels. Distinguishing between closely related taxonomic units may be enforced by providing the program with alignments of marker protein sequences, e.g. *gyrA*.

Chapter 1) Literature Review

1.1) Introduction to Phylogenetics and Phylogenomics

In the mid-1800s, Darwin's theory of the origin of species gave birth to the field of evolution. Ernst Haeckel, a German zoologist also came up with a sketch which became a blueprint to what we know today as a phylogenetic tree. At that time, evolutionary relationships were built upon the similarities between specie morphology. It was assumed that sharing of common phenotypic traits might indicate a common ancestry of organisms represented in a tree by branches joined by an intermediate node. In Figure 1.1, the branches outlined in purple represent the species tree. As time progressed and the genetic basis of life was generally recognised, species comparison evolved into gene sequence comparisons, leading to gene trees. Gene trees (Figure 1.1 in blue, red and green) do not always agree with species trees owing to events such as horizontal gene transfer (HGT), gene duplication and an uneven rate of evolution of different genes (Figure 1 blue to green and red). This sometimes led to conflicting predictions of speciation events (Lin *et al*., 2011; Swenson and El-Mabrouk, 2012; Bezuidt *et al*., 2016).

Nowadays, phylogenetics is used in many aspects of biology. These fields include analysis of relationships between species (Takahashi *et al*., 2001; Zhaxybayeva *et al*., 2006), improvement of methods utilising annotation information such as paralogues (Finnerty *et al*., 2009; Berendzen *et al*., 2012; Chai *et al*., 2014), population evolution discovery (Francois and Mioland, 2007; Li and Durbin, 2011), pathogen and cancer studies in the field of medicine (Cawley and Talbot, 2006; Stecher *et al*., 2013) and detection of HGT through phylogenetic tree comparisons (Poptsova and Gogarten, 2007). Phylogenetics and its toolset are becoming ever more important, as applying it is a vital skill in supporting other type of studies, such as metagenomics (Filipski *et al*., 2015).

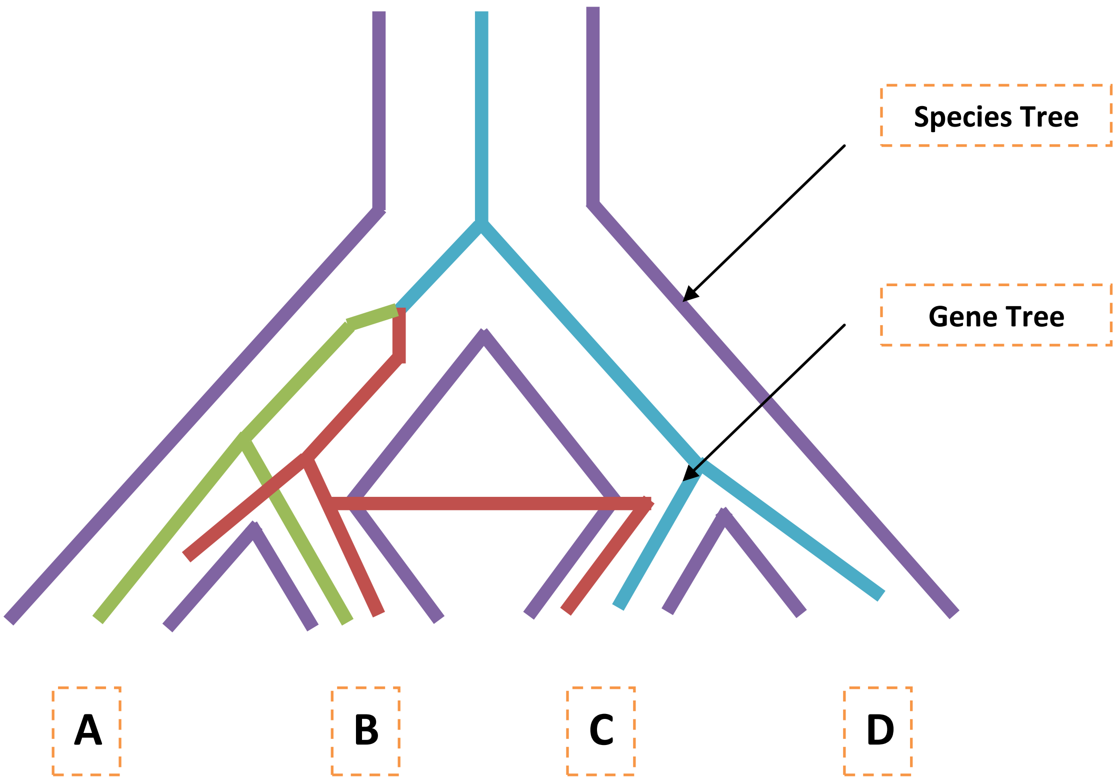


Fig. 1.1 An example of a species tree (purple) and a gene tree (green, blue and red). Within the gene tree, one can see gene duplication (blue splitting into red and green) and horizontal gene transfer events, which could create conflict between the two trees (red). From the above, because of horizontal gene transfer, the resulting gene tree could group species B and C more closely together. Similarly, with gene duplication, there could be a further separation of species B into B1 and B2, creating a more complex gene tree compared to the species tree.

The essential part of phylogenetic studies is based on three important aspects, namely the collection of proper phylogenetic markers, the use of the most appropriate evolutionary model, e.g. models of rates of substitutions in aligned DNA, and protein sequences and methods of implementation of evolutionary models compatible with the context of the study. Within these three domains, a vast variety of new innovative methods has been developed in the last decades. The field of phylogenetics itself has evolved to adapt to new technical advances in the current era. Despite continued progress in the design and development of new technologies and algorithms, supported by the advance in computational facilities, a number of underlining problems and limitations, some of which are not immediately obvious, still exist.

The major problem of phylogenetics is the choice of data to use for phylogenetic inferences. Currently, the main source of information on phylogenetic relations between organisms from the level of the tree of life to the level of subspecies is sequencing of DNA and protein samples. Large databases of sequences have been created; however, it is often noted in the literature that the sequenced data on various organisms are still not fully comprehensive and rather biased towards organisms of medical or economic value (Chan and Ragan, 2013). Hence, in the past, the tree of life itself was limited by the available sequence data. Non-sequenced organisms nevertheless constituted a significant domain within the tree of life. Filling this lack of data with new sequences may in future require reconsideration of phylogenetic relations between organisms. The current tree of life could look entirely different if these sequences are produced to fill missing branches within the phylogenetic tree (Puigbò *et al*., 2013). This is more evident in prokaryotes, where HGT and uneven evolutionary rates in different taxa and in different parts of genomes cause significant problems with the identification of universal phylogenetic marker genes. This limits application of gene tree approaches in inferring phylogenetic relations between bacterial taxa. The recent advances in genome sequencing techniques allow resolution of such problems by replacing single-gene phylogenetics with whole-genome phylogenomics.

With next generation sequencing (NGS) being the current focus, an immense amount of data is readily available at a low cost and more sequence data covering a wide variety of organisms are becoming a reality (Figure 1.2). Phylogenetics can take a new step forward to correct the tree of life and overcome the potential problems from the past. However, the implementation of techniques of whole-genome comparison is not trivial and simple because of many problems inherited from single-gene comparison, including orthology identification and proper alignment. There is also an overwhelming amount of genome-scale data calling for new, effective computational tools. Therefore, in the current NGS era, researchers are heading towards a new field of phylogenomics that benefits from the availability of large genome-scale sequences but requires development of new tools for inferring evolutionary relationships based on large sets of data.

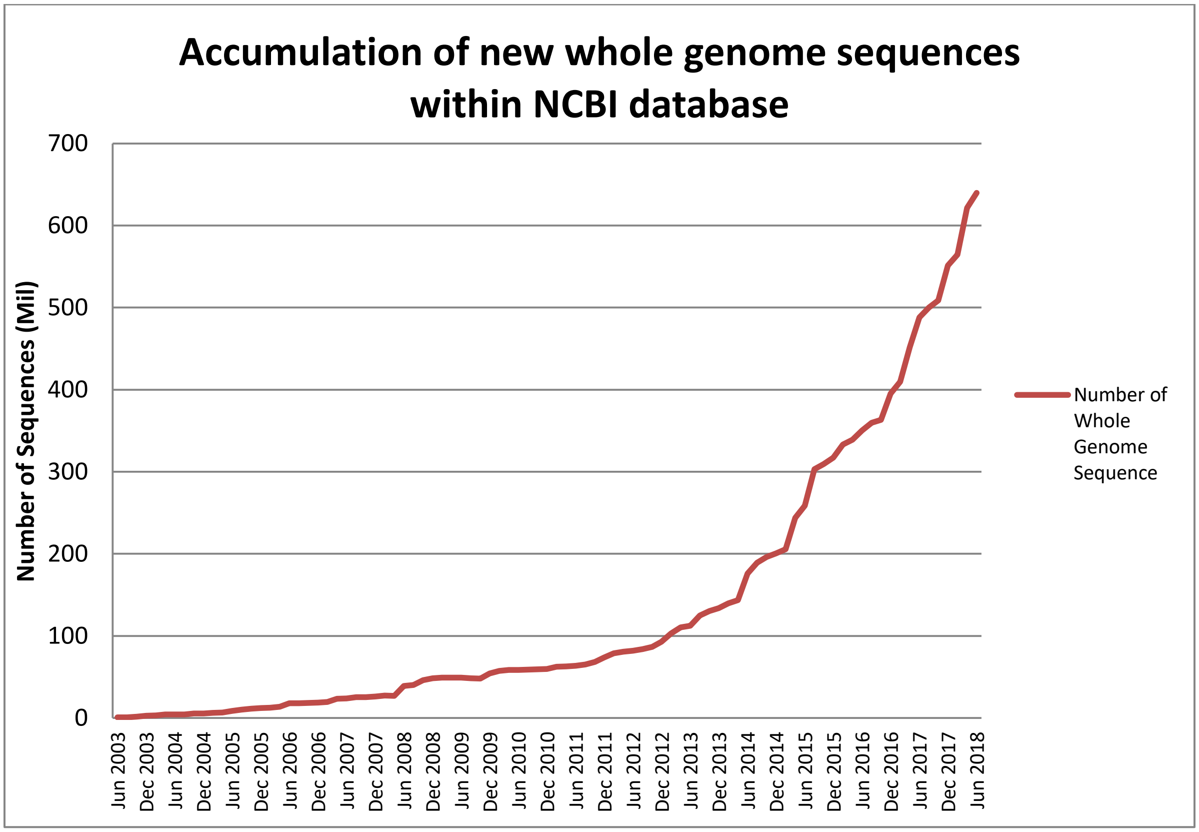


Fig. 1.2 Accumulation of new sequences in the NCBI database using whole genome shotgun method for assembling incomplete genomes or chromosomes dating from June 2003 to June 2018 (NCBI, 2018).

Based on the definition in the journal *Nature* (Nature, 2018), phylogenomics involves the reconstruction of evolutionary relationships by comparing sequences of whole genomes or sufficiently large portions of genomes. With the analysis of whole genome data, several advantages have become apparent over the traditional phylogenetic analysis. The most obvious advantage is operation with much longer genome-scale sequences compared to single genes. This brings about an expectation of better accuracy of phylogenetic reconstructions (Beiko, 2010). It looks plausible that a group of species can be distinguished better based on a number of orthologous genes being shared rather than by comparing only a single gene (Jeffroy *et al*., 2006). The multiple orthologous gene analysis can also infer functional and/or ecological factors, which singular genes do not take into consideration (Kumar *et al*., 2012; Chai *et al*., 2014). With a large array of data to be analysed, additional statistical support should be provided to reduce the impact of the stochastic noise on phylogenetic tree inferences (Kumar *et al*., 2012). Population genomics can also benefit from an analysis of variations in multiple genes displaying responses to the environment, which single-gene phylogenetic tools may not offer (Kvitek and Sherlock, 2013).

With multiple projects aimed at high throughput sequencing of bacterial genomes and populations still under way, i.e. the Genomic Encyclopaedia of Bacteria and Archaea project (Kyrpides *et al*., 2014), phylogenomics has become a new vital technique that researchers will need for mining phylogenetic relations in enormous sequence datasets. In the following sections of the literature review, a comparison of different phylogenetic and phylogenomic tools will be provided. Current problems of phylogenomics will also be discussed and potentially innovative approaches to the improvement of current approaches will be proposed.

1.2) Current Methods and Approaches to Phylogenetic Inferences

1.2.1) Gene-based Phylogenetics

A phylogenetics study includes many steps, of which the final aim is to create a phylogenetic tree consisting of branches denoting the relationship of common ancestry between each species in the study. Phylogenetic tree construction is divided into two main types of methods, consisting of character-based and distance-based approaches. For distance-based approaches, as the name states, the sequence comparison between two species is calculated as a number of weighted evolutionary events estimated by a certain criterion or algorithm. These distances, which denote the diversity between species, are then used in a tree construction algorithm such as neighbour joining (NJ) (Saitou and Nei, 1987) to resolve the final phylogeny. Character-based approaches, on the other hand, look at alignments of all sequences simultaneously and consider every single character difference along all possible (or plausible in the case of heuristics) tree topologies as a likelihood penalty with the aim to identify the most likely tree topology. Based on the different methods, the best tree is chosen upon a tree score for which each method has its own selection criteria. For example, maximum parsimony considers the smallest number of single-character substitutions between aligned sequences as the most likely tree path. Maximum likelihood (ML) considers the log likelihood score based on a chosen substitution model, and the Bayesian method the best posterior probability (Yang, 1996; Yang and Rannala, 1997).

When considering the use of distance-based methods, the distance calculation becomes of vital importance, as this directly influences the actual similarity measure between species in study. Pairwise sequence distances are calculated under different assumptions, using different models of either nucleotides or amino acid substitutions. The commonly used nucleotide substitution models are based on the Markov chain assumption. These models include JC69 (Jukes and Cantor, 1969), which assumes an equal substitution rate between any two nucleotides. The K80 (Kimura, 1980) model assumes different rates for transitions and transversions. HKY85 (Hasegawa *et al*., 1985) assumes non-equal base frequencies for different nucleotides. General time reversible (Tavaré, 1986) models assume an equal substitution rate for reversal substitution (r_{GC}=r_{CG}). Because of selective restraints, the gamma model, which measures the variation between different sites in terms of substitution rates, can also be added to existing models to improve distance measures by allowing different rates of evolution.

Protein substitution models are based on log-odds matrices. A likelihood of substitution of one amino acid for another is converted into a log-odds score. Such matrices include the point accepted mutation (PAM) matrix, looking at the differences observed in closely related proteins. The BLOck Substitution Matrix (BLOSUM) looks at blocks of conserved sequences in multiple alignments of functional importance. This matrix reduces the bias from divergent sequences over a long period of time (Dayhoff *et al*., 1978; Henikoff and Henikoff, 1992). The protein distance matrix is believed to work better than nucleotide substitution models, as base nucleotides become saturated much faster over time, which therefore reduces the amount of positive information, leading to long branch attraction problems. Amino acid sequences, on the other hand, are under codon restrictions with a high probability of synonymous mutation, therefore contain less noise due to saturated sites (Shapiro *et al*., 2006).

The most popular distance-based algorithms of phylogenetic inferences are NJ, minimum evolution and the least squared method (Bulmer, 1991). The least squared method can be applied to estimate mathematically optimal distances between operational taxonomic units (OTUs) and then uses the differences between initial and theoretical distances for statistical evaluation of branch lengths. These theoretical values are calculated such that the tree with the lowest sum between optimal branch lengths is considered the true one. This is very similar to the least regression fitting of a straight line where the parameters are estimated such that the least squared error is minimised to minimum to get the best fit. Minimum evolution, on the other hand, uses the sum of all branch lengths and the smallest tree that satisfies the data with the shortest branch length is considered the true one. This is considered under the minimum evolution criterion, which assumes that a shorter tree is more likely to be correct (Hartigan, 1973). The final NJ method is a cluster algorithm, which analyses the initial distance matrix in an attempt to construct an additive tree representing the distances between OTU by the length of branches with minimal loss of information. The algorithm starts in a star tree-like shape and pairs of OTU with the closest distance measure are joined together to create a new joint node. Then the distances between OTU are recalculated using the NJ equation (Figure 1.3). The distance matrix is then updated and the process repeats itself until all species are resolved (Saitou and Nei, 1987). Of the given methods here, NJ is the most popular because of the computational efficacy and repeatability.

The major advantage of distance-based methods is without any doubt the computational efficiency of these methods. Not all of the methods above in the procedure compare multiple tree topologies, in contrast to character methods, and hence they are suitable for processing of large datasets. The major downside of the method however is selecting a sensible substitution model for the dataset under study. Very divergent datasets in combination with nucleotide substitution models for a high level of saturation of nucleotide sequences will lead to false phylogenetic inferences with a low level of resolution between taxa, leading to long branch attraction problems (Bergsten, 2005). The long branch attraction problem is inherent to the matrix-based phylogenetic algorithms. It consists of shortening phylogenetic distances in a tree due to data saturation.

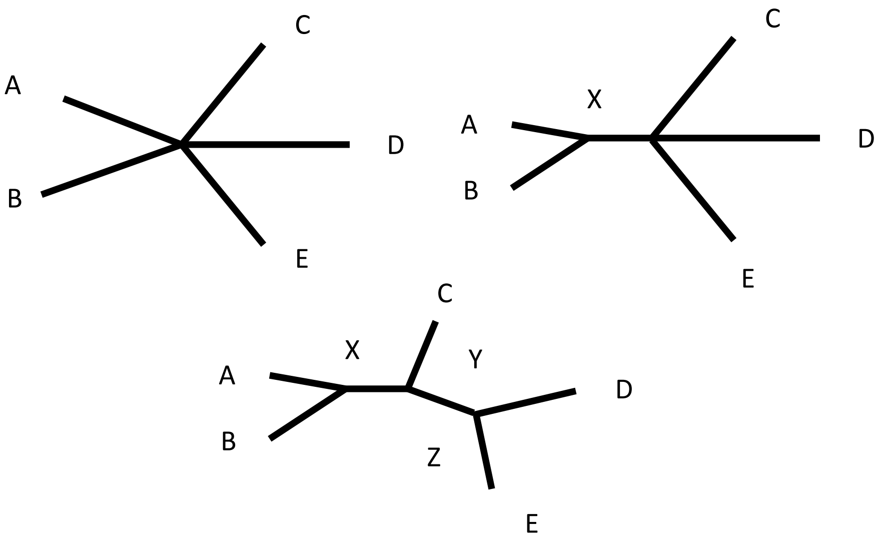


Fig. 1.3 Neighbour joining clustering algorithm. Species A and B are closest in similarity and hence clustered together first by node X (top left). The distance matrix is then updated accordingly by X, C, D and E. Species D and E are then the two closest species and clustered together by Z (top right). Finally, a fully resolved phylogenetic tree is created according to the NJ algorithm (bottom).

Character-based approaches are an alternative tool kit to do phylogenetic analysis that can cover some shortfalls of the distance-based approaches to infer a better phylogenetic tree in certain circumstances. The maximum parsimony method calculates the minimum number of character changes required at different sites under study to explain given tree topologies (Yang, 1996). The final tree score is determined by the sum of all character changes at all sites and the tree with the smallest sum difference being considered the most parsimonious one. When considering this method, some sites in the sequence can be ignored, such as conserved sites with little to no change. This method produces reasonable results, as its simplicity makes it straightforward. However this method is not computationally efficient and too simple It contains little to no biological assumption (such as nucleotide or amino acid substitution model), meaning that this model fails to work with diverse sequences. Maximum parsimony also suffers from the long branch attraction problem and often infers a large collection of trees of different topologies, but with the same scores, that can be confusing. Hence, this method is statistically less consistent compared to other character-based methods, such as ML and Bayesian statistics (Felsenstein, 1978).

ML is one of the more popular and consistent methods currently used in phylogenetics. It was first developed by Ronald Fisher in 1920 as a method to estimate unknown parameters in a model (Felsenstein, 1981; Yang, 1994). The assumption of the function states that the observations in the data can estimate the parameters explaining the given data. In phylogenetics, these parameters are the tree topology, branch lengths and substitution models. The ML algorithm analyses substitutions in sequence alignments to score and select the best tree topology with a maximum likelihood. This is achieved by an analysis of the nature of substitution and referring to the expected likelihood of the event stated in the selected evolutionary model, i.e. PAM or BLOSUM tables with additional assumptions of Gamma distribution, molecular clock and some others. This method is also consistent owing to its asymptotic nature and unbiasedness. It allows the use of different substitution models to estimate phylogenetic relationships. All these features make this method more advantageous than other phylogenetic inferencing methods. The only current limitation on this method is the computational intensity that limits it to analysis of small datasets only.

Bayesian inference is the last method mentioned in this section on character-based phylogenetic methods. The Bayesian method, also known as Bayesian statistics, is very popular in the field of statistical simulations (Andrew *et al*., 1995). Similar to the ML method, likelihoods of all tree topologies possible for given alignments are estimated. However, unlike ML, which assumes the parameters of a substitution model to be fixed and computable by numerical methods, the Bayesian method assumes the parameters to be variables of a statistical distribution. The advantage of this method over the popular ML method is that by using posterior distribution, one can estimate the accuracy of the estimated parameter without resorting to the computationally costly method of bootstrap resampling (Bollback, 2002).

In terms of phylogenetics, Bayesian inferences only became popular in the late 1990s (Yang and Rannala, 1997). At first, this approach was used to infer ultrametric trees under a molecular clock assumption. Later, the approach was supplemented with an algorithm of Markov Chain Monte Carlo (MCMC) that improved estimation of unrooted additive trees with different branch lengths. Further improvement on this method introduced a relaxed molecular clock algorithm that provided users with higher flexibility regarding control over the program workflow and outputs (Drummond and Rambaut, 2007). Just like ML methods, the Bayesian method is consistent, efficient and statistically powerful compared to other methods of phylogenetic inference. The Bayesian method is also easy to interpret by means of preference of the final best tree by referring to the posterior distribution. The major flaw however is that in certain cases the prior distribution is not known or well stated. It may put a burden upon users, as the prior distribution heavily influences the accuracy of the posterior distribution, which in turn determines which tree topology is the best one (Lemmon and Moriarty, 2004). Aside from the prior distribution, the substitution model that provides likelihoods is also highly sensitive with regard to the posterior distribution calculation. Overly simplified models tend to inflate the posterior probabilities, which results in wrong trees. Hence, it is highly important to use an additional toolkit to assess the correctness of the posterior distribution and/or analyse the effect of the model being used for the inference to ensure the appropriateness of the results (Zhaxybayeva and Gogarten, 2002).

In summary, a wide variety of techniques is available to do phylogenetic studies, with each method having its own pros and cons. Depending on the context of the study, one can consider the best toolset to use for the different types of datasets. In Table 1.1, the methods are summarised with their advantages and disadvantages, along with some popular programmes that implement these methods.

Table 1.1 Phylogenetic Algorithms and Toolkits

List of Methods	Description and Assumption	Type of Dataset	Toolkit
Distance Methods	• Uses pairwise alignments and substitution models to determine distance metric based on dissimilarities. • Substitution models include nucleotide or amino acid based algorithms. • Tree construction methods include neighbour joining, least squared or minimum evolution based on distances calculated during pairwise comparisons.	• Large dataset can be used as the algorithm is not computationally costly.	**PHYLIP*** A toolkit containing multiple programs for phylogenetic inference using distance, parsimony and ML methods.
Maximum Parsimony	• Assumes that evolution takes the form of the lowest number of substitutions as the true evolutionary path. • Easy to use and computationally efficient. • Fails to work for sites where multiple substitutions occur and because of its simplicity, no biological assumptions are taken into consideration.	• Datasets that do not span a long period of time. • Closely related organisms	**MEGA**** Molecular evolutionary genetic analysis. Program with strong graphical interface that does parsimony, ML and distance-based inferences.
Maximum Likelihood	• Mathematical function that estimates unknown constant parameters in terms of the dataset given. • Can incorporate complex substitution models to add true evolutionary history to phylogenetic inferences. • Powerful and consistent in estimating and testing model parameters needed for phylogenetic inferences. • Computationally expensive	• Size of the dataset cannot be too large owing to computational intensity • Any type of data	**HYPHY***** Hypothesis testing using phylogenies is a program that fits models of evolution using ML approaches. **PhyML****** Program that conducts fast searches for phylogenetic trees using ML methods.

Bayesian Inference	• Statistical method based on Bayesian statistics. • Estimates unknown parameter as a random variable of a statistical distribution. • Incorporates prior knowledge and observation data to create a posterior distribution leading to true phylogenetic tree • Can incorporate complex substitution models and posterior distribution is easy to interpret, making determining true tree simplistic. • Method limited by prior knowledge rarely being available. • Posterior distribution is highly dependent on prior and substitution model, making model selection difficult for users.	• Size of dataset cannot be too large owing to computational intensity of MCMC algorithm • Any type of data with preference where prior knowledge on data is available	**MrBayes******* Bayesian MCMC program for phylogenetic inference with all models of substitution available. **BEAST******** Similar program to MrBayes but for inferring rooted trees under molecular clock or relaxed molecular clock assumption models.

*(Tuimala, 2006), **(Tamura *et al*., 2013), *** (Kosakovsky Pond *et al*., 2005), **** (Guindon *et al*., 2010), ***** (Ronquist et al., 2012), ****** (Bouckaert *et al*., 2014)

Alongside selection between different phylogenetic algorithms, identification of proper phylogenetic marker genes is another task of great importance, which influences the correctness of phylogenetic inferences. The DNA sequence of 16S ribosomal ribonucleic acid (rRNA) has long been the most successful and most used phylogenetic marker since the 1970s, when it was introduced by Woese and Fox (1977). In this publication, the authors demonstrated successful application of 16S rRNA sequences for distinguishing between different prokaryotic domains. Sequences of 16S rRNA are characterised by a high degree of conservation and this is assumed to result from the importance of the 16S rRNA as a critical component of cell function (Clarridge, 2004). Very few gene sequences share this degree of conservation as well as being available in all bacterial genomes. Based on the conservative nature of 16S rRNA, the absolute rate of change in this sequence can be seen as an evolutionary distance of relatedness of organisms (Thorne *et al*., 1998). It has also been shown that phylogenetic trees produced by 16S rRNA data are congruent with the whole genome aligned trees (Bansal and Meyer, 2002). However, other authors reported conflicts between 16S rRNA inferences and species trees (Haggerty *et al*., 2009; Takahashi *et al*., 2009; Rajendhran and Gunasekaran, 2011; Prabha *et al*., 2014). In addition, because of the properties of 16S rRNA, this small subunit is used in clinical identification of bacteria and pathogens and has been successful in terms of being a genetic barcode for studies such as metagenomics (Richter *et al*., 2008; Fuks *et al*., 2017; Tran *et al*., 2017). However, unlike specie identification such as barcodes, there is not enough resolution power in this method to make reliable phylogenetic inferences for higher taxonomic levels (Janda and Abbott, 2007). The conservational nature of 16S rRNA underestimates the evolutionary rates of distantly related species and will appear more closely related for taxonomic groups (Prabha *et al*., 2014).

Other phylogenetic marker genes have also proven useful for phylogenetic inferences aside from 16S rRNA. The DNA gyrase (type II topoisomerase), subunit A (GyrA) protein sequence has also been considered a good phylogenetic marker gene for certain taxonomic groups (Huang, 1996; Menard *et al*., 2016). This gene shares many similarities with 16S rRNA, in which the motif is highly conserved owing to its function and this family of proteins is also prevalent in prokaryotic organisms. The variation in this gene has also proven to distinguish individual isolates and species and it is hence considered a feasible phylogenetic marker. Aside from traditional phylogenetic markers, other markers also exist in the form of ultra-conserved elements (UCE) (Faircloth *et al*., 2012). UCEs are regions of DNA that serve specific functions and hence are highly restricted to any change in composition. UCEs may be regulators, enhancers of gene expression and of other functional importance still being actively researched (Woolfe *et al*., 2005; Pennacchio *et al*., 2006). These sequences are easy to identify by their conservative nature and can easily align across divergent genomes within large datasets. These regions do not intersect with most types of paralogous genes and are not prone to insertions. It was shown that sequence similarity between UCE regions correlates with the evolutionary relatedness between organisms (McCormack *et al*., 2012). However, this method has its own share of problems of having highly non-neutral evolution which may infer false phylogenetic relationship for specific datasets. Using character-based methods, one can resolve phylogenetic relationships using UCE as a marker for comparison, providing an appropriate evolutionary nucleotide substitution model. A small flaw of these genetic markers however lies in the fact that there are no universal primers for amplification of UCE regions and there is a lack of any good resource to provide reference sequences of these regions.

Lastly, ribosomal proteins have proven to be the most likely phylogenetic markers able to outperform 16S rRNA in terms of phylogenetic inferencing (Martini *et al*., 2007). Ribosomal proteins in all bacteria have the same well-established functions that make it possible to avoid mixing up speciation and functional diversification processes when the sequences are compared (Hug *et al*., 2016). Ribosomal proteins are found co-located in narrow genomic regions, usually in the vicinity of clusters of genes for ribosomal RNA. Therefore, they can be retained in short reads or partial sequences generated from metagenomes. Read binning against gene sequences for ribosomal proteins is also characterised by higher accuracy compared to binning against 16S rRNA sequences that often leads to chimera sequence production. Chimeras and allele copy variations of 16S rRNA often decrease the accuracy of species identification in metagenomes. The final major advantage of application of ribosomal protein sequences over their predecessors is the ability to concatenate multiple ribosomal proteins to increase the statistical power over just a single marker gene for phylogenetic inferences (Jolley *et al*., 2012). The use of multiple concatenated ribosomal proteins could resolve conflicts between single-gene trees, which many phylogenetic markers have difficulty to deal with. The selection of multiple genes maximises the information used to find the most correct phylogenetic tree (Blair and Murphy, 2011). However, this method shares the same problems as phylogenomic methods, as more information gained from comparison of multiple concatenated ribosomal proteins leads to more conflicts between tree topologies (Jeffroy *et al*., 2006). As gene duplication and HGT are still prevalent within ribosomal proteins, the phylogenomic problem remains of concern for ribosomal proteins as a good phylogenetic (Yutin et al., 2012).

1.2.2) Approaches to Phylogenomics

As explained in the introductory section, phylogenomics was derived from phylogenetics to cover its shortfalls in handling large amounts of sequencing data in larger regions produced from NGS technologies (Chan and Ragan, 2013). Because of this fact, some phylogenomic approaches are very similar to phylogenetics, with some phylogenomics tools being upgraded versions of current phylogenetics toolsets. However, for the relevance of this section, we took a more in-depth look at other phylogenomic methods, which take different approaches. These include supermatrix and supertree methods, average nucleotide identity (ANI), genome BLAST sequence phylogeny, pangenomic analysis of clusters of orthologous genes, multi-locus sequence typing (MLST), alignment-free compositional algorithms and whole genome alignment. We will also discuss each method in detail, as well as its relevance in terms of current phylogenomic research and the tools that use these approaches. Finally, the problems of phylogenomics in the current context and the pros and cons of each method are evaluated.

1.2.2.1 Supermatrix and Supertree-based approaches

The first on the list of the best known and most commonly used alignment-based approaches to phylogenomics is the supermatrix and supertree method. The principal idea of using alignments of multiple clusters of orthologous genes (COG) instead of individual marker genes was that the comparison of multiple homologous genomic regions would resolve possible disagreements between the evolution scenarios of individual genes. This would allow reconstruction of more reliable phylogenetic relationships between organisms. The supermatrix and supertree approaches were exploited for integration of all coding sequences from genomes either by combining multiple alignments of homologous genes or encoded proteins into a supermatrix, or by finding consensus of multiple gene-based trees. A supertree is assembled using all taxa within every source tree, where shared taxa between source trees are connected in a heuristic approach (Bininda-Emonds, 2004). A consensus supertree consists of all source trees and resolving any conflicts between the individual gene trees. This method however does not evaluate how accurate the common shared phylogeny for each gene is. Toolsets for the supertree method include PhySIC (Scornavacca, 2009) and SuperFine (Swenson *et al*., 2012).

The supermatrix approach takes a different ordering to supertree by combining sequences of different COG together to create a supermatrix for tree construction (de Queiroz and Gatesy, 2007). Within a supermatrix alignment, missing data between sequences may be filled in by the addition of important accessory genes evident in some of the sampled organisms. An advantage of this approach over the supertree method is that the use of all characters in a super-alignment has a better resolution in estimating the final tree than a combination of only tree topology data (de Queiroz *et al*., 2003). In other words, phylogenetic signals in a supermatrix analysis are stronger because of the combined information in the supermatrix instead of combining individual trees from separate analyses. In terms of statistics, this type of analysis achieves higher statistical power by reducing noise within the study and therefore leads to more resolved phylogenetic trees. The supermatrix approach can also combine different types of data, allowing users to be flexible in adding different biological information beneficial to inferring the most correct phylogenetic tree. For example, in the study by Wheeler *et al.* (2001), a combined study of morphological and molecular

(18S and 28S rDNA) data was used in determining hexapod orders (Wheeler *et al*., 2001).

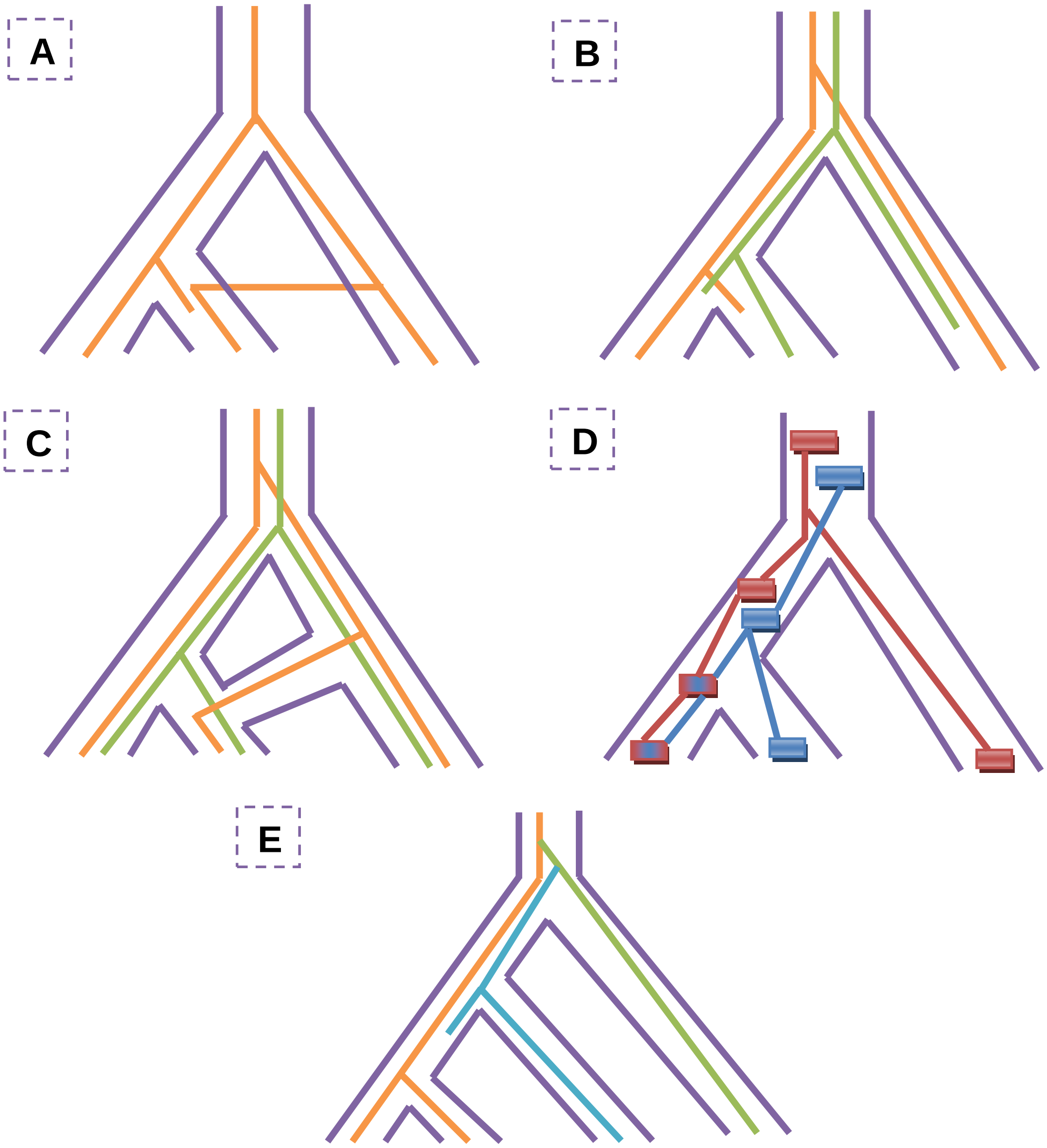


Fig. 1.4 Sources of gene tree discordance. A) Horizontal gene transfer: transferring of genetic material across lineages creating a relationship from ((A to (A(BC)). B) Gene duplication and loss: apparent distinction where gene trees are not congruent with specie trees. This leads to the extinct lineage appearing more distant, e.g. from ((AB)C) to (A(BC)). C) Hybridisation: A certain lineage B might descend from two lineages (AB) and (BC). D) Recombination results in different histories for different segments of DNA. Segments in red result in tree ((AB)C) and segments in blue result in tree (A(BC)). E) Incomplete lineage sorting (ILS): failure of two or more lineages in a population to coalesce, creating a possibility that at least one of the lineages coalesces with another closely related population, e.g. ((AB)(CD)) gene tree to ((AB)C)D) specie tree.

There are several problems with the supermatrix method. Because of the collection of data from multiple sources in a single supermatrix, the assumption of all characters having the same branching history or substitution model is not always valid. The number of COG may differ depending on the size and diversity of sampled genomes, which will eventually affect the topology of the resulting tree. One has to accept that supermatrix trees as well as supertrees are sample-dependent and may not be comparable to one another. This leads to the second problem, where conflicts in gene trees will be caused by missing data or HGT. Gene exchange, gene gain-and-loss events, and improper annotation of genomes can create homoplasy, leading to inference of erroneous trees. Several examples of these problems are shown in Figure 1.4. The first problem however has been addressed by the addition of model-based and parsimony-based methods for the correct application of a substitution model to infer true evolutionary history (Dickerman, 1998; Swofford *et al*., 2001).

Supermatrix and supertree approaches are examples of scaling up commonly used approaches of phylogenetics to the level of genome comparison. They are based on the same methods of sequence alignment followed by applying evolutionary models to explain substitutions of residues in the alignments. The applicability of gene-based substitution models to the whole genome is questionable. This is due to the fact that different genes have different substitution rates and applying a single substitution model for the whole genome may lead to incorrect phylogenetic inference. An even bigger problem derives from the fact that no manual checking of the results of orthology prediction and alignment correction is done when multiple genes are processed in an automatic or semi-automatic manner. This leads to multiplication of error in these error-prone procedures. Attempts have been made to design approaches to the evaluation of phylogenetic distances between genomes without doing multiple alignments.

1.2.2.2 Ortholog-based Approaches

One commonly used phylogenomic method is comparing the distribution of orthologous genes in genomes. The presence and absence of genes can determine the similarities of different taxonomic units. Techniques for identification of orthologous genes have been proposed by a reciprocal BLASTP alignment of translated complete DNA sequence (CDS), by complete genome alignment, or by combinatorial approaches (Sims *et al*., 2009). Efficient Database framework for comparative Genome Analyses using BLAST score Ratios (EDGAR) is a good platform that can identify orthologs using comparative analysis (Blom *et al*., 2009). This platform contains a large database containing orthologs from over 500 genomes across 75 genera in the National Centre for Biotechnology Information (NCBI) database. Orthologs in this case are defined under a strict criterion as genes with conserved function and diverged from a speciation event (Fitch, 1970). Hence, based on ortholog comparison, one can identify evolutionary events through speciation.

Usually, ortholog detection requires absolute bidirectional best BLAST hits. However, because of the variation in BLAST scores for different genera, the EDGAR platform calculates BLAST score ratio values instead for each genus to ensure the most correct ortholog detection. This ratio is achieved by plotting all resulting BLAST hit scores based on each genome assessed against all others in a histogram; one would typically see a bimodal distribution. This is observed as one group of genes with low similarity with unspecific hits and another with high similarity hits representing possible orthologs. To determine true orthologs from a group of potential ones, a sliding window approach is used to assess all bar widths containing similar blast hit scores with potential orthologs in the histogram to calculate a cutoff. The lowest scoring window, i.e. the group of alignments with the lowest BLAST scores, is set as the final cutoff. This cutoff is set as the BLAST score ratio for this group of alignments and potential orthologs are identified in this region. Comparison of concatenated orthologs between genomes containing core genes from each OTU is then used to calculate a distance score using alignments produced by the MUSCLE algorithm (Edgar, 2004). The distances are then used to create a phylogenetic tree using the NJ method with the PHYLIP package (Tuimala, 2006).

Edgar 2.0 is now available with enhanced functionality, including improved graphical interface, genome size statistics and other phylogenetic analysis features such as ANI. In general, orthology identification is not a trivial task because of many complications such as gene paralogy resulting from gene duplication and HGT events, which can lead to false phylogenetic inferences (Figure 1.4) (Boussau *et al*., 2008). Orthology prediction in diverse organisms may be problematic because of the accumulation of multiple mutations in homologous sequences preventing proper alignments. Another serious limitation of sequence-based methods applied to complete genomes is computation time, which is sensitive to the size of datasets. Heuristic approaches were used instead with a trade-off in terms of accuracy of resulting inferences (Woolley *et al*., 2008). The quality and reliability of alignments of multiple genomic loci are other issues of concern (Conte *et al*., 2008; Dwivedi and Gadagkar, 2009).

Long before sequencing techniques were introduced into practice, whole genome similarity comparison by chromosomal DNA-DNA hybridisation (DDH) was a popular method of measuring phylogenetic distances and species delineation (Wayne *et al*., 1987). Since the main source of differences in the level of hybridisation of genomic DNA is not sequence divergence but the presence of non-homologous regions. This method became obsolete and generally neglected when 16S rRNA sequencing was recognised as the gold standard of phylogenetics. However, later, with the advance of whole genome sequencing, several computational tools were developed to mimic genome-scale DNA-DNA hybridisation by analysing mismatches in homologous parts of genomes. This measure is known as the ANI, which shares strong correlation to experimental DDH values and has proven to be highly useful in determining bacterial specie definition (Arahal, 2014; Zhang *et al*., 2014). ANI is a relative measure of similarity between sequences based on comparison of widely distributed genes in addition to multiple lineage-specific genes. The selection of these genes must reach a certain cutoff in terms of a BLASTN match in order to reduce error arising from falsely inferenced homology due to a low level of similarity. The extraction of a phylogenetic signal from these genes (typically over 1000 genes considered) is genetically descriptive and robust owing to the selection of genes. Because of the quantity of data considered, ANI is a better measure of similarity than 16S rRNA and not prone to effects such as varied evolutionary rates and HGT of singles genes (Goris *et al*., 2007). This is due to the large number of genes being considered, whereby the effects of fast evolving genes are mitigated by slow evolution of others. As ANI is comparable to DDH, comparison of ANI can be seen as a golden standard in terms of specie definition and phylogenetic reconstruction. As more sequences are being produced through NGS platforms, the only shortfall of this method can be reduced through annotation of new sequences. Popular programs for ANI calculation are JSpecies (Richter and Rosselló-Móra, 2009) written in Java, and ANItool (Han *et al*., 2016), a web-based ANI implementation tool.

A similar approach aimed at replacing the tedious laboratory procedures of DDH for the delimitation of prokaryotic sequences is genome BLAST distance phylogeny (GBDP), which infers genome-to-genome distance between pairs of entirely or partially sequenced genomes (Meier-Kolthoff *et al*., 2013). The GBDP measure works similarly, compared to a combination of orthology and the alignment-based method. This distance is calculated based on a comparison of high-scoring segment pairs consisting of highly similar intergenomic regions identified through algorithms such as BLAST (Altschul *et al*., 1990). Information contained in these high-scoring segment pairs, such as the total number of identical base pairs, can be used to transform it into a single genome-to-genome distance value by the use of a specific distance formula (Henz *et al*., 2005). Phylogenetic trees can then be inferred from such distance matrices using distance comparison algorithms such as NJ (Tuimala, 2006). This method is comparable to other well-established methods such as ANI and the results have shown reasonable comparison to DDH values for species delineation. This method also contains statistical validations where results can be tested through bootstrapping and jackknifing methods. Although the method is significantly powerful in taxonomic classification, its application in phylogenomics is still questionable in comparison to other well-stated methods. However, with current toolsets this method contains, such as species delineation on par with DDH and ANI with statistical validation of bootstrapping, this method is promising in terms of phylogenomic inferencing.

Another interesting approach to infer phylogenetic relationships between species is to use species definition from pangenomic analysis in a recent publication (Moldovan and Gelfand, 2018). A pangenome is defined as a set of orthologous gene groups comprising all genes from a sample of genomes (Lapierre and Gogarten, 2009). Therefore, through identification of lineage-specific gene sets, we can differentiate between species and their relationships. In contrast to the methods considered above, this approach does not rely on DNA similarity levels and thus does not require alignment of sequences and/or analysis of homologous recombination between genomes. Instead, a gene frequency spectrum function, which defines the number of orthologous gene groups from exactly *k* genomes, was used to assess the homogeneity of the dataset. The shape of the distribution of this function over a set of *k* genomes can indicate if this set is homogenous or not. Smooth U-like distribution represents a homogenous dataset, while U-like distribution with internal peaks represents a non-homogenous dataset. In terms of evolution, when a homogenous dataset becomes non-homogenous, it is often associated with speciation where a single dataset is under independent directional selection and gene gain and loss, leading to two gene sets. The other possible scenario could be a single strain being under strong selection and gene gain leading to a single peak in the distribution. Under both scenarios, one can identify possible branching of non-homogenous sets into the formation of two monophyletic homogenous groups. Therefore, through the analysis of the distribution of spectrum function of the number of core genes in the pangenome of sets of strains, one can distinguish paraphyletic samples of strains from monophyletic ones. Therefore, it was shown in this paper that the spectrum function can potentially reflect the level of phylogenetic relatedness between organisms in the sample using pangenomic definition of species. Although the method is appealing in the sense that it does not require alignment between sequences to determine species definition, the authors did not come up with any new tools for phylogenetic inferences. However, this method does show potential in terms of phylogenomic inferencing using orthologous comparisons.

MLST (Maiden *et al*., 2013) can be seen as another form of orthology-based approach where this method convert a group of genes into a series of numbers that represents the sequence for easy comparison. MLST, which is mainly used for bacterial genomes, uses several housekeeping gene loci as identification measures for phylogenetic analysis (Katz *et al*., 2017). Each locus is assigned a specific unique allele number and multiple loci, specifically chosen, are combined and assigned an allelic profile also known as a sequence type. Each sequence under study is summarised in this way and searched through a database. Comparison of these conserved parts allows identification of differences between sequences. This method is efficient in converting large amounts of sequence data into a short sequence type without accommodating conflicting signals of HGT, which are common in bacteria. As the chosen loci are user-specific, signals such as rare point mutation, recombination and the time frame in which these mutations occur are not considered. This allows the user to control or reduce unwanted noise from these factors which may influence the resulting phylogenomic inference. Other major advantages of this method are that it allows easy reproducibility of the study, is expandable in the sense that these sequence types can be changed upon new information, portable as the database is freely exchanged between researchers and scalable, as this method is sufficiently fast in comparison. The only major drawback of this approach is the putative bias on gene selection, which may influence the resulting phylogenetic tree. ANI can cover this shortfall, as all orthologous protein genes shared between pairwise genome comparisons are considered. ANI is currently also used as a golden standard in terms of prokaryotic species definition (Richter and Rosselló-Móra, 2009).

1.2.2.3 Sequence Alignment Approaches (MAUVE)

Multiple Alignment of Conserved Genomic Sequence with Rearrangements (MAUVE) is an interesting program that creates phylogenetic trees through multiple alignments of large genomes as a byproduct (Darling et al., 2004a). MAUVE was designed to align long genome sequences quickly, using the anchored alignment approach by identifying and using multiple maximal unique matches of length *k* sub-sequences as anchors. These sub-sequences must be exact matches in at least two or more genome sequences in alignment. This strategy is repeated, with *k* being reduced to search for other smaller anchors in unmatched regions. To speed up this process, the seed and extend hash method is used, whereby subsets of genomes are aligned each time and additional genomes are added once matches have been identified (Darling et al., 2004b). Once the anchors have all been found, MAUVE filters out random spurious matches by determining local collinear blocks, which are homologous regions of sequences shared by two or more genomes that did not go through rearrangements. A phylogenetic guide tree is then created using the anchor information between pairs of genomes and converting this into a distance value for an NJ distance matrix. Specifically the ratio of shared base pairs between genomes over their average genome length will result in the similarity distance value. Finally, with the guided tree, MAUVE resolves the intervening regions between anchor sets through the progressive dynamic programming approach of CLUSTAL W (Thompson *et al*., 1994) as an optimal progression route to align. In the case under study, the guide tree created as a byproduct of MAUVE is the phylogenetic tree for the dataset of genomes that was aligned using MAUVE. Since this method was not strictly created for phylogenomic analysis, it is simplistic and efficient to use while taking into account large genome sequences with alignment. Nevertheless, because of the method using NJ and a relatively simple calculation scheme of similarity between sequences, this method lacks a sensible evolutionary model that distinguishes the variation in evolutionary rate of different organisms.

The above-mentioned methods of whole genome comparison using identification of homolog sequences in different genomes are based either on amino acid or DNA similarity. Homology prediction is an error-prone approach owing to different rates of substitutions in genomic loci, gene duplication and abundance of repeated elements, including generally used conserved protein domains, as well as for some other reasons. An attractive idea was to use homology and alignment-free parametric approaches to compare genomes by several common genome-signature properties. The program CVTree, an alignment-free method proposed by Qi *et al.,* estimates the phylogenetic relationship of sequences using comparison of frequencies of oligopeptides in complete proteomes (Qi *et al*., 2004b). The authors claim that the use of oligopeptides instead of oligonucleotides, which is other popular subject for K-mer statistics, decreases the range of possible K parameter estimations. In other words, because amino acids consist of three nucleotides in a codon, a varying K parameter range for nucleotide sequences is much larger compared to the amino acid range. Amino acid sequences are also believed to be more consistent in referring phylogenetic trees than nucleotide sequences owing to saturation of substitutions at specific codon positions (Jeffroy *et al*., 2006), leading to artefacts such as long branch attraction problems. However, the problems of short sequence alignments should not be transferred mechanically to genome-scale sequences and the alignment problems may not be attributed to the K-mer statistics.

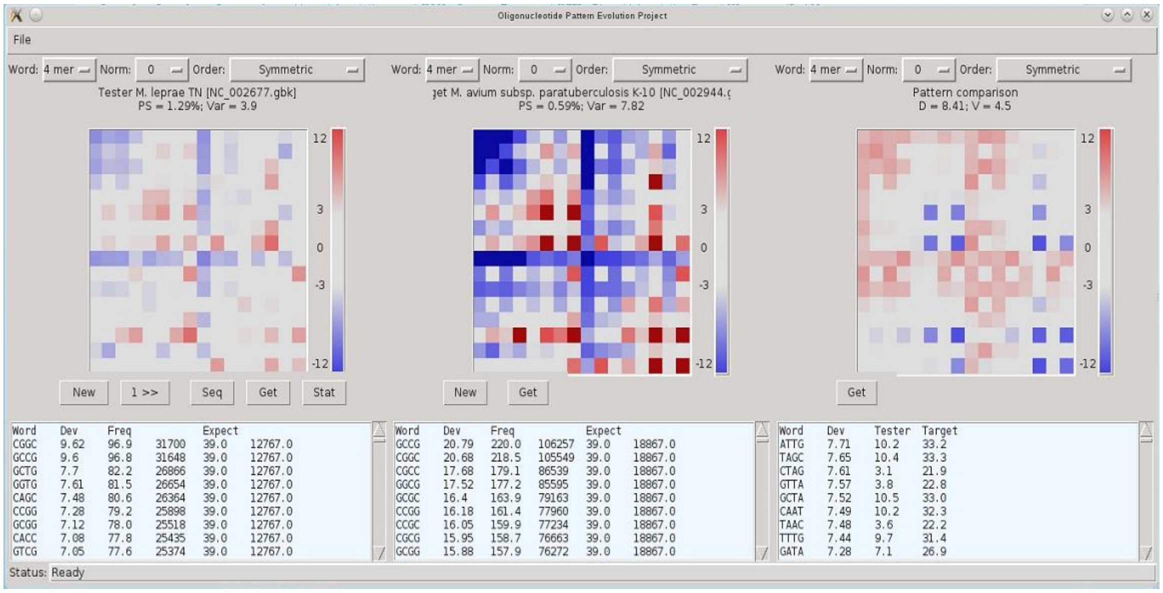
1.2.2.4 Alignment and Annotation Free Methods

Another innovative approach of CVTree was the reduction of the background noise resulting from an assumption of context-independent substitution of residues in protein sequences. This is done by considering amino acids as part of evolutionary stable K-mer oligopeptides shaped and governed by natural selection pressure, which is achieved by applying the Markov chain model (Brendel *et al*., 1986). The K-string oligopeptides with non-zero difference between the observed K-string frequency and estimated frequency calculated based on frequencies of K-2 substrings were used for the construction of genome-specific compositional vectors. These compositional vectors were the key elements of measuring the evolutionary distances between genomes. The similarity measure consists of calculating the correlation between two compositional vectors estimated for given genomes. Finally, the correlation values were converted into distance values ranging from 0 to 1 and the distance matrix was processed by the NJ method to plot the final phylogenetic tree. The first limitation of this method is that it translates sequences of protein-coding genes instead of whole genome information. The method assumes that the annotation is correct and that the annotated CDS can be translated directly to protein sequences, which can only be true for prokaryotes and archaea. A second problem is the rather oversimplified conversion of vector correlation values into evolutionary distances. Lack of evolutionary models does not allow further conversion of the correlation coefficients into biologically meaningful distances comparable with those estimated by other phylogenetic algorithms discussed above. As the proposed research focuses primarily on K-mer based approaches to phylogenomics, this topic will be discussed and presented in more detail in the following section.

1.3) Application of Nucleotide K-mers in Phylogenetics and Concept of Oligonucleotide Usage Pattern

Compositional methods that branch off distance-based methods are well known in terms of phylogenetic inferencing and are based on the similarities of nucleotide or amino acid frequencies between OTUs. Furthermore, genome-wide frequencies of K-mers, which are short oligonucleotide sequences of k nucleotides, have been proven to contain more biological information than just single nucleotide substitutions in aligned sequences. In other words, an analysis of changes on the level of individual nucleotides gives less information, which is also prone to a higher level of noise owing to the data saturation problem, than the analysis of frequencies and distribution of K-mer words (Bergsten, 2005). Frequencies of K-mers also reflect genome-specific preferences such as stereochemical properties of DNA substrings, including nucleotide stacking energy, DNA strand bendability, nucleotide position preferences and some others (Reva and Tummler, 2005). Therefore, by analysing specific K-mer metrics of sequences, one can estimate comparable genome signatures (Reva and Tummler, 2005; Bohlin *et al*., 2008). By comparing these signatures, one can derive a numerical measure based on how similar or different these sequences are (Berendzen *et al*., 2012). In many studies, it was found that certain K-mers have stronger signals than others do. For example, dinucleotide frequencies can produce genomic signatures resulting from selection pressure of dinucleotide stacking, DNA conformational tendencies, DNA replication and repair mechanisms (Karlin, 1998). Frequencies of tri-nucleotide and longer words may be affected by biased codon usage, which is specifically adjusted in every organism to ensure translational efficiency (Sharp and Li, 1987). It was also found that other specific K-mers (four- and eight-mers) can be used to differentiate bacterial and Archaea species (Bohlin *et al*., 2008). Therefore, by means of K-mer statistics, one can gather useful information regarding genome-specific DNA stereochemistry constraints and substitution preferences, which are applicable to evolutionary inferences.

K-mer statistics have been applied successfully, for example in identifying HGT inserts within genomes (Pierneef *et al*., 2015). Since K-mer can represent a genomic signature, K-mer patterns of foreign DNA inserts differ from the host genome signature and can therefore be identified as HGT events. However, the DNA of genomic islands may gain properties of the host organism owing to the amelioration process consisting in an unbalanced rate of nucleotide substitution favouring accumulation of the preferable oligonucleotides of the host organism (Lawrence and Ochman, 1997). Because of this, HGT events usually do not cause serious problems in comparison of K-mer patterns. An example of a program that uses the oligonucleotide usage metric is SeqWord Genomic Island Sniffer (SWGIS) (Ganesan *et al*., 2008). This program uses oligonucleotide usage patterns (OUP) representing genome signatures of various K-mer (two- to seven-mers) to identify HGT by estimating several statistical parameters of K-mer distribution. OUP is a matrix of frequencies of all K-mers. For example, a tetramer pattern comprises 256 possible permutations (Figure 1.5).



Oligonucleotide Pattern Evolution Project

Word: 4 mer Norm: 0 Order: Symmetric

Tester M. leprae TN [NC_002677.gbk]
PS = 1.29%; Var = 3.9

Word: 4 mer Norm: 0 Order: Symmetric

jet M. avium subsp. paratuberculosis K-10 [NC_002944.ç
PS = 0.59%; Var = 7.82

Word: 4 mer Norm: 0 Order: Symmetric

Pattern comparison
D = 8.41; V = 4.5

```
Word    Dev     Freq            Expect
CGGC    9.62    96.9    31700   39.0    12767.0
GCCG    9.6     96.8    31648   39.0    12767.0
GCTG    7.7     82.2    26866   39.0    12767.0
GGTG    7.61    81.5    26654   39.0    12767.0
CAGC    7.48    80.6    26364   39.0    12767.0
CCGG    7.28    79.2    25898   39.0    12767.0
GCGG    7.12    78.0    25518   39.0    12767.0
CACC    7.08    77.8    25435   39.0    12767.0
GTCG    7.05    77.6    25374   39.0    12767.0
```

```
Word    Dev     Freq            Expect
GCCG    20.79   220.0   106257  39.0    18867.0
CGGC    20.68   218.5   105549  39.0    18867.0
CGCC    17.68   179.1   86539   39.0    18867.0
GGCG    17.52   177.2   85595   39.0    18867.0
GCGC    16.4    163.9   79163   39.0    18867.0
CCGG    16.18   161.4   77960   39.0    18867.0
CCGC    16.05   159.9   77234   39.0    18867.0
CGCG    15.95   158.7   76663   39.0    18867.0
GCGG    15.88   157.9   76272   39.0    18867.0
```

```
Word    Dev     Tester  Target
ATTG    7.71    10.2    33.2
TAGC    7.65    10.4    33.3
CTAG    7.61    3.1     21.9
GTTA    7.57    3.8     22.8
GCTA    7.52    10.5    33.0
CAAT    7.49    10.2    32.3
TAAC    7.48    3.6     22.2
TTTG    7.44    9.7     31.4
GATA    7.28    7.1     26.9
```

Status: Ready

Fig. 1.5 A representation of a tetra-mer (right) oligonucleotide usage pattern for the comparison of *Mycobacterium avium* (middle) and *Mycobacterium leprae* (left). The blue colour represents over-representation of a certain word (higher observed frequency than expected frequency), while the red represents under-representation of a certain word (lower observed frequency than expected frequency). The table below the oligonucleotide usage pattern displays a list of words

ranked by the highest difference between observed and expected frequencies (deviation) to the lowest difference.

K-mer frequencies are tallied for each word and normalised by estimated expected frequency estimated by the Markov chain approach. Distribution of rare and abundant tetramers in two mycobacterial genomes is depicted in Figure 1.5 by blue and red cells, respectively. The right outermost panel in this figure demonstrates differences in tetramer frequencies between the query genome (*M. leprae*) and the reference genome (*M. avium*) depicted by coloured cells. The program also calculates several other statistical parameters of oligonucleotide usage, such as pattern skew (PS), oligonucleotide usage variance (VAR) and distance between patterns calculated for whole genomes (D) (Figure 1.5). All these parameters were defined and explained in the publication by Reva and Tummler (2005). The D value, which will be used as a phylogenetic similarity measure, will be explained in more detail in Chapter 2.

The distance between two OUP patterns is calculated as the absolute difference between the rank position of each four-mer (word) within the OUP of two sequences after ordering the words by their abundances. In other words, OUPs with dissimilar distribution of oligomers are characterised by a large distance and vice versa. The program automatically calculates four OUP from combinations of direct and reverse strands and takes the minimum value of counterpart pattern comparison as the distance. Therefore, depending on the distance value, HGT events can be identified since the genomic signature is unique to each organism. A large distance value between patterns shows that there is foreign genetic material (Reva and Tummler, 2005). For example, one would compare the OUP of a single region within the genome compared to the global OUP. PS is a particular case of the distance measure, which calculates the distance between direct and inverse strands of the same DNA. Since for bacterial genomes the PS value tends to be low, a high PS value could imply insertion of phage elements

(Reva, 2004). Lastly, OU variance calculates the variance of the deviation between two patterns. Since patterns are unique, a large difference in variance between patterns is another criterion for identifying HGT events. However, since the number of combinations of nucleotide words is constrained, uncontrolled mutation (insertion) can cause higher oligonucleotide variance (OUV) values, which can be another validation for false positives.

Aside from HGT identification, several authors proposed a method to use K-mer statistics for taxonomic binning of short genomic fragments generated from metagenomics datasets (Berendzen et al., 2012). Because of the nature of metagenomic datasets, short reads are hard to align and difficult to classify. Hence, short K-mer comparison works well and no assembly or annotation is needed compared to other methods that require this information (Delsuc *et al*., 2005). With the properties of K-mer statistics, they are not region-specific and are considered a genome signature (Reva and Tummler, 2004). Hence, for metagenomic datasets with a large amount of read data assembled in multiple unidentified contigs, K-mer statistics have the advantage of taxonomic binning of these contigs without a need for identification of marker genes in these short genomic fragments. MetaProb is such a program that uses K-mers to bin metagenomic reads based on probabilistic sequence signatures (Girotto et al., 2016). The first phase within this program utilizes the advantages of K-mers as sequence signatures to cluster reads together with q-mer overlaps between reads. Due to read sequences matching up to q-mers, q being a certain number of base nucleotides, the reads clustered is more likely to be from the same species. This technique is also widely used within de novo assembly. The second phase filters artefacts using a K-means algorithm identifying true clusters from false positives by analyzing if the clustered K-mers are probabilistic signatures of the binned species. This method is computational efficient with short run time and suitable for short metagenomic datasets. Other examples of programs include Kraken and Clark which also uses K-mers methods for metagenomic analysis and classification (Wood and Salzberg, 2014; Ounit et al., 2015). In another study, it has been estimated that theoretically, a reliable four-mer statistic can be estimated only for sequences longer than 5 kbp (Reva and Tummler, 2004). However, K-mer methods are still powerful in terms of binning contigs of metagenomics origin (Ondov et al., 2016), but uncertain to how feasible this method is with phylogenetic inferences.

Initially, linguistic approaches was created as a way to convert DNA sequences into words which can serve as a basis for revealing functional and evolutionary relatedness of sequences (Brendel et al., 1986). These methods were further developed into new tools which utilize K-mer counts between sequences as a measure of similarity largely used for sequence comparisons fundamental to molecular biology (Haubold et al., 2005). Analyzing abundance of K-mers within groups of genomes can also identify key words which might be functionally or structurally important within these groups (Davenport and Tümmler, 2010). A study by Castellini *et. al.* has also went one step further in classifying genomic information within a dictionary fashion whereby analyzing the dictionary index, one can analyze and visualize the genome under study (Castellini et al., 2012). In this way, one can efficiently do comparative studies using these dictionaries which highlight key factors such as genome structure and functional attributes. Linguistic methods have also been used for phylogenetic reconstruction based on comparison of subwords as a measurement of similarity (Comin and Verzotto, 2012). However the results from this method are only promising in the sense that it can identify major clusters well with low resolution for closely related sequences. The application of this program was also limited to assembled genomes and not usable for short reads from NGS platforms.

Fan *et al.* (2015) proposed and validated an alignment and assembly-free K-mer phylogenetic inferencing method for NGS read data (Fan *et al*., 2015). Through this research, K-mer statistics were proven to be a feasible approach to phylogenetic inferences. The authors proposed an evolutionary model of diversification of K-mer frequencies in genomes by applying a nucleotide substitution model based on a Poisson distribution. It was assumed that the substitution rate of a single nucleotide within a K-mer is relative to the number of K-mers shared between the two sequences under study; i.e. if the number of shared K-mers is high, the probability of a substitution in such K-mers is reduced. The accuracy of the resulting trees was validated by a benchmark test using bootstrapping analysis and by comparison to simulated sequences with known phylogeny. The method allowed bypassing the error-prone steps of identification of orthologs and sequence alignments. Moreover, the computational efficiency was another appealing aspect of the proposed method. There was however a downside to this method, as the trees inferred based on K-mer patterns with different K-values were not always congruent with one another.

Statistical approaches estimating phylogenetic distances between genomes by comparison of K-mer patterns were reviewed by Fan *et al.* (2015)(Fan et al., 2015). It was demonstrated that the calculated distances were congruent with those estimated by traditional phylogenetic methods, which was in agreement with previously published data by Takahashi *et al*. (2009). In other publications it was demonstrated that patterns of tetramers were optimal genome signatures in terms of noise-to-signal ratio (Reva and Tummler, 2004; Reva and Tummler, 2005; Reva and Tummler, 2008). Therefore, comparison of K-mer-based OUP between organisms can be viewed as a plausible phylogenetic distance. This approach also has the potential to identify possible outliers with abnormal genomic signatures that might be misplaced within a phylogenetic tree (Ganesan *et al*., 2008; Elhai *et al*., 2012). However, because of the lack of sensible evolutionary models of OUP diversification and tools to do such studies, the question still remains as to how appropriate the conversion of OUP dissimilarity values into phylogenetic distances is.

1.4) Tree comparison and Model Evaluation

With modern phylogenomics and an increasingly large number of datasets being processed, true phylogenetic tree inference is no longer a simplistic procedure and it is getting harder to distinguish with multiple specie nodes and countless choices of substitution models. Therefore, it is important to know the correct validation and evaluation techniques to ensure the most correct phylogenetic inference. When considering phylogenetic reconstructions, model choice is the vital first step and an incorrect choice can under/overestimate rates of evolutionary changes, leading to false trees or long branch attraction problems (Misof *et al*., 2014). When problems occur here, tree comparisons at a later stage are no longer important, as all trees are incorrect and wrong inferences can be made. Hence, it is evident that model selection is highly important and an objective measure is needed to determine the best model for the data under analysis. Before selecting the best model, one needs to first analyse the dataset and then choose several models that might fit it. To determine which evolutionary models work with the dataset, determination of the rate of changes in sequences in the dataset and the driving forces of these changes is important, as certain models might not work owing to under/over-parameterisation of these models. Under-parameterised models resolve phylogenetic relationships poorly owing to disproportioning evolutionary distances between sequences (long branch attraction). Over-parameterising does not always guarantee the production of better phylogenetic trees (Posada and Crandall, 2001) and overly complex models are not intuitive to use and understand. If time is not an issue, one can test all models under selection. Nevertheless, it is good practice to reduce the model space for a subset of substitution models as an important first step.

When the model space is determined, one needs to use different selection criteria to determine the best evolutionary model for a given dataset. The mostly widely used method is the likelihood ratio test (LRT). It can be interpreted as the measure of fit between model and data comparable to other models. Hence, by using this test, one can determine if either addition or reduction of parameters can optimise the model of choice. These approaches are developed for the best comparison of different nucleotide or amino acid substitution models.

The likelihood measure takes the form of the classical test statistic in the following equation:

$$\boldsymbol{\delta = 2(ln\, L_1 - ln\, L_0)} \qquad [1.1]$$

L_1 denotes the likelihood score in terms of observed data of the more complex model, while L_0 denotes the likelihood score of another model. The test statistic is evaluated under the chi-squared distribution with the degree of freedom being the difference in number of free parameters between the two models. Based on a cutoff on the test statistic, one can determine if a more complex model fits better than the other model. This approach has been incorporated in Modeltest (Posada and Crandall, 1998) in a hierarchical fashion, where an initial model is used to construct an initial tree and more complex models are tested compared to the previous one. The method itself determines the best model in terms of the dataset but with a clear weakness. The models tested against one another in the hierarchical LRT do not have a clear guide to the order of comparison (either added or reduced parameters). The order in which parameters are added or removed influences the model choice and often favours over-parameterised models.

Another commonly used selection criterion worth considering and a good method to use in comparing different types of models is the Akaike Information Criterion (AIC) (Akaike, 1974). The AIC measure is calculated as follows:

$$\boldsymbol{AIC_i = -2\, ln\, L_i + 2\, K_i} \qquad [1.2]$$

AIC is calculated as the log likelihood of the maximum likelihood value of the model *i* (joint ML estimate across all parameters) subtracted by the number of parameters within the model. AIC is derived from the Kullback-Leibler distance from information theory and can be explained by quantifying the information lost based on the estimated model approximating the true model (Kullback and Leibler, 1951). Therefore, parameter K in this case is a parameter for penalising over-parameterisation of the model, as the lower the AIC value, the less loss of information the model explains the dataset. Hasegawa (Hasegawa, 1990) first used this method in phylogenetics in 1990 and substitution models were selected using AIC. The difference between AIC values between models can indicate to the researcher that one model is better than other. Differences of two indicate strong support, between four and ten imply weak support and more than ten no support. Since AIC is calculated independently, if comparison of AIC using differences is nested, as in the hLRT method, AIC does not have the weakness of ordering. AIC has another form, AICc, which is used for datasets with smaller sample sizes. The equation takes into consideration the sample size n and adjusts the measure accordingly. In terms of phylogenetics, this n value could be the number of sites within the sequence.

$$\boldsymbol{AIC_{Ci} = -2\ ln\ L_i + 2\ K_i\ + \frac{2K_i(K_i+1)}{n-K_i-1}} \qquad [1.3]$$

The Bayesian selection criterion includes two measures, namely the Bayes factor and the Bayesian information criterion (BIC). BIC takes a similar form as AIC, the difference being that BIC penalises more than AIC for over-parameterisation. BIC takes the form shown below:

$$\boldsymbol{BIC_i = -\ 2\ ln\ L_i + 2K_i\ ln\ n} \qquad [1.4]$$

Derived by Schwarz (1978)(Schwarz, 1978), one would consider this selection criterion if the user is strict on over-parameterisation. Since the penalty function K is multiplied by the logarithm of sample size n, BIC resists the tendency to select complex models as n increases. The Bayes factor, on the other hand, takes the

Bayesian assumption and is calculated as the ratio between the probability of data based on model one over the probability of data based on model two (Kass and Raftery, 1995). The prior of each model to give the posterior odds to favour one model above another can multiply this ratio. However, if both priors are equal, the Bayes factor is equal to the LRT. Since the dependent probabilities are calculated for each parameter within the model, the Bayes factor accounts for uncertainty for each parameter estimated. To determine which model performs better according to the Bayes factor test, a score of more than 20 supports the first model, a score between 3 and 20 implies that model one is slightly favoured, while a score of less than 3 shows equivalence between the two models. Both methods can be used in a nested form for multiple comparisons between models for best model selection.

Based on the different selection criteria, one can assume the best possible substitution model for the dataset, but in terms of phylogenetic reconstruction, is this enough? Minin *et al.* (2003) created an approach that goes in the opposite direction to analyse model adequacy based on model performance in estimating the correct tree topology and branch lengths (Minin *et al*., 2003). Instead of analysing how well the model represents the data, this approach ranks the models based on the expected error in branch length estimates derived using BIC. For this method to work, one needs accurate results on both the tree topology and branch length, irrespective of the model used for the estimate, in order to assess the model quality accurately. This performance measure was feasible because of the inconsistency of results from ML under different tree topologies and hence if one assumes tree topology to be constant and correct, one can assess the model quality through the estimations of branch length. Since the method uses BIC weights, models with more parameters are penalised more than simpler models. This method will choose the simplest models estimating branch length with the least error and will be chosen as correct in terms of phylogenetic reconstructions (Abdo *et al*., 2005). Therefore one could also consider this method as selection criterion under specific circumstances (known tree topology), which might work better in testing for the best model or if certain models are adequate for tree construction.

The above covered the model selection phase, indicating how to choose a substitution model in terms of the dataset using different tests to minimise error in the final reconstruction phase. Other evaluation techniques include using artificial genome sequences to test the method of inference. By creating artificial sequences with known substitution models, one can test the accuracy of the inferencing method. If the chosen method produces the same resulting tree based on the artificial sequences, it is suitable for a given dataset alongside the chosen substitution model. This method might be trivial, as it does not allow testing if the created tree is the correct one. However, it allows testing and comparing different inferencing methods. There are currently many software tools available for the creation of different types of simulation datasets (Escalona *et al*., 2016). For example, SimBac creates artificial bacterial sequences of any size, which undergo recombination and substitution events (Brown *et al*., 2016). Researchers should be able first to identify the correct method before doing the phylogenomic analysis, as many methods are time-consuming and might not give the best results.

Lastly, final evaluation has to be done to ensure the correctness of the resulting tree. Different methods are available and can be applied. The most commonly used method is bootstrapping (Efron *et al*., 1996). Bootstrapping is a statistical method in which one can assign a measure of accuracy to a certain parameter by doing resampling with replacement (Efron and Tibshirani, 1993). This parameter in the case of phylogenetics is the correctness of the tree topology (Soltis and Soltis, 2003). With a bootstrapping test, using the whole pool of sequences as an initial dataset, sampling is performed to create a subset allowing random substitutions, monomeric insertions of residues or partitioning of aligned sequences. Then multiple phylogenetic trees are created from these random subsets. Hence, based on variations in tree topologies of sampled trees, one can assess the correctness of the resulting tree based on the number of repeated grouping of OTUs to the same clusters. Bootstrapping is a powerful method because of its simplicity and is applicable to almost all types of datasets, as there are various types of bootstrapping methods (Makarenkov *et al*., 2010; Fan *et al*., 2015). However, owing to its algorithm, this method is computationally intensive and may not be applicable to large datasets, such as supermatrix alignments of multiple proteins.

Another tree evaluation technique aside from bootstrapping is Bayesian posterior distribution. Because of the nature of the Bayesian method, the resulting tree is evaluated in the form of a posterior statistical distribution (Bouckaert *et al*., 2014). However, as previously stated (Chapter 1.2), this posterior distribution is heavily dependent on the prior distribution, which is sometimes hard to define. Since this is the case for many phylogenomic datasets, this method is not always possible to apply and hence not often used compared to bootstrapping. Other methods are also available to evaluate and improve final phylogenetic trees. Lin *et al.* (2010) propose a method based on Bayesian logistic regression to resolve polytomic branches (multiple branches from a single node) into dichotomic ones (binary branching from a single node). Polytomic branching results from poorly resolved phylogenetic trees, when selected sequences do not provide sufficient information to distinguish between related organisms. Hence, the algorithm resolving polytomic cases provides an additional level of mining of initial sequences to resolve this problem and increase the accuracy of the final tree (Lin *et al*., 2011). Some other examples also include resolving complex microbial phylogenetic networks using linear programming by comparing characteristic traits compositionally (Holloway and Beiko, 2010). The disk covering method analyses multiple related datasets, combining different trees to create the best tree (Bayzid *et al*., 2014). Lastly, a more accurate tree branch estimation method is proposed, which analyses the dataset by means of the weighted least square method to identify the best branch length combination for the resulting phylogenetic tree (Binet *et al*., 2016). Despite all these methods at researchers' disposal, the problem of incongruence between trees generated by different methods remains unsolved (Jeffroy *et al*., 2006). A clear benchmarking test or a golden standard among methods is still lacking for different types of datasets.

1.5) Aims of Current Project

To summarise, a common flaw shared by many of the above-mentioned methods is that they require proper annotation of genomes followed by aligning of predicted homologous loci of the genomes of interest. These steps are error-prone and may lead to false phylogenetic inferences (Ogden and Rosenberg, 2006). An attractive alternative is using linguistic approaches which were first introduced by Brendel *et. al.* (Brendel et al., 1986). These methods avoids sequence annotation, orthology prediction and alignment steps in phylogenomics (Reva and Tummler, 2004; Sims *et al*., 2009). To date, several studies have been proposed to use linguistic-based approaches for species identification and binning of metagenomic reads (Richter and Rosselló-Móra, 2009; Berendzen *et al*., 2012; Maiden *et al*., 2013; Tran and Chen, 2014; Blaimer *et al*., 2015; Filipski *et al*., 2015; Ondov *et al*., 2016). These alignment and annotation-free approaches were explored and tested in many areas of research, such as metagenomics (Wood and Salzberg, 2014; Ondov *et al*., 2016), evolutionary partitioning (Frandsen *et al*., 2015), branch length estimations (Binet *et al*., 2016) and phylogenetics (Yi and Jin, 2013; Tran and Chen, 2014), with promising yet sometimes controversial results. This problem is also evident through phylogenetic tree comparisons using different methods (alignment and annotation-free method vs parameterised methods), as there is no consensus between them. Evolutionary assumptions behind the models used for evolutionary inference relationships differ significantly (Koonin *et al*., 2011). This leads to another vital problem of phylogenomics, namely incongruency of trees constructed by various methods, even for the same dataset, that confuses one’s understanding of evolutionary relationships between organisms (Galtier and Daubin, 2008; Chan and Ragan, 2013). Although each method has its pros and cons specific to certain dataset or time scales (Bochkareva et al., 2018), the problem is still present to researchers if a consensus is needed in determining the most correct tree from several different programs. There are currently several studies and programmes have been proposed that specifically tackle the incongruence problem, phylogenomics still has a long way to go in many aspects. Many problems are still evident for current toolsets, such as computational efficiency, evolutionary model assumptions, tree construction algorithms and statistical consistency between methods in reducing errors in different phases of the study (de Oliveira Martins *et al*., 2008; Gori *et al*., 2016). However, there is still a bright future perspective for innovative methods such as annotation and alignment-free algorithms, promising tools and an open field in resolving current needs for phylogenomic studies.

In the new era of NGS, phylogenomics is essential in analysing a large quantity of data with large sequences compared to phylogenetic analysis, as explained in the previous sections (Chapter 1.1). However, the shortfalls of these approaches are also apparent. For the purpose of this research, this project is aimed at three main problems addressed in the previous section (Chapter 1.2, 1.3 and 1.4). These three objectives will be covered in the next three chapters, each dedicated to a specific problem.

- As previously defined, alignment and annotation-free methods have several advantages, such as bypassing the erroneous steps associated with the alignment and annotation while being computationally efficient (Fan *et al*., 2015). Although K-mer statistics have proven to be a good phylogenetic measure (Reva and Tummler, 2004; Reva and Tummler, 2005; Bohlin *et al*., 2008; Takahashi *et al*., 2009), there is a need for a deeper analysis of possible conflicts between OUP-based trees and those trees inferred from common gene-based and genome-based approaches. In terms of evolutionary models, OUP lacks a sensible evolutionary model to back up its phylogenetic assumptions (Chan and Ragan, 2013). Hence, the first aim of the project was to assess the feasibility of OUP in terms of phylogenetic inferencing of different taxonomic groups of bacteria and to identify possible evolutionary forces shaping OUP. This objective also involves a comparison of OUP-based phylogenetic trees with currently used phylogenomic methods, using the tree topology and branch length criteria (Tuimala, 2006). These methods include whole genome alignment, comparison of individual orthologous genes, MAUVE, CVTree, phylogenetic marker *gyrA* gene and 16s rRNA (Qi et al., 2004a; Darling et al., 2004a). Phylogenetic trees were inferred from genome sequences of selected microorganisms and based on randomly generated sequences simulating evolutionary processes using the SimBac program.

- The second aim is to reconcile the discrepancy between different phylogenomic methods, specifically the *gyrA* phylogenetic marker and OUP-based method revealed in the case studies. The objective is to integrate the two different tree-inferencing approaches and infer a more accurate phylogenetic tree sourced from information gathered through both methods.
- Finally, to tackle the problem of lacking efficient tools to do such a study, the aim is to create an OUP-based phylogenomic inferencing tool available to the public. This will take the form of a downloadable program and an online web-based tool, which will allow users to implement the OUP-based phylogenomic approach.

Chapter 2) Analysis of Possible Evolutionary Forces Shaping Oligonucleotide Usage Patterns

2.1) Introduction and Theory Overview

In Chapter 1.3, it was explained that OUP comparison is a promising new approach in the field of phylogenomics owing to its ability to identify a genomic signature based on biological factors influencing the formation of OUP. Therefore, to assess if OUP is a viable phylogenomic method, analysis of these factors is important for understanding OUP evolution, as different signals can be subject to different evolutionary pressures (Trifonov, 1989). Hence, we analyses in depth three possible factors shaping genome-specific OUP and the way in which these relate to substitution rates of nucleotides within OUP. These three factors include mutations of nucleotides influenced by local patterns of nucleotide combinations, codon usage bias affecting the usage of OUP and lastly the simultaneous effect of both above-mentioned factors.

In a study by Baldi and Baisnee (2000), it was identified that the genetic code with different nucleotide models (di- and tri-mers) provided insight into the flexibility of DNA structure (Baldi and Baisnée, 2000). Five different nucleotide models representing stereo-chemical properties of oligonucleotides, including bendability (Brukner *et al*., 1995), positional preference (Satchwell *et al*., 1986), propeller twist angle (el Hassan and Calladine, 1996), protein deformability (Olson *et al*., 1998) and base stacking energy (Ornstein and Rein, 1978), were assessed. It was noted from the correlation between nucleotide models and the structural flexibility of DNA that the mutation of nucleotides is dependent on specific local patterns of nucleotide models (di- or tri-mers). This may be influenced by local patterns being associated specifically with genomic mechanisms of DNA replication and regulation, of which enzymes involved may sense differences in stereo-chemical properties.

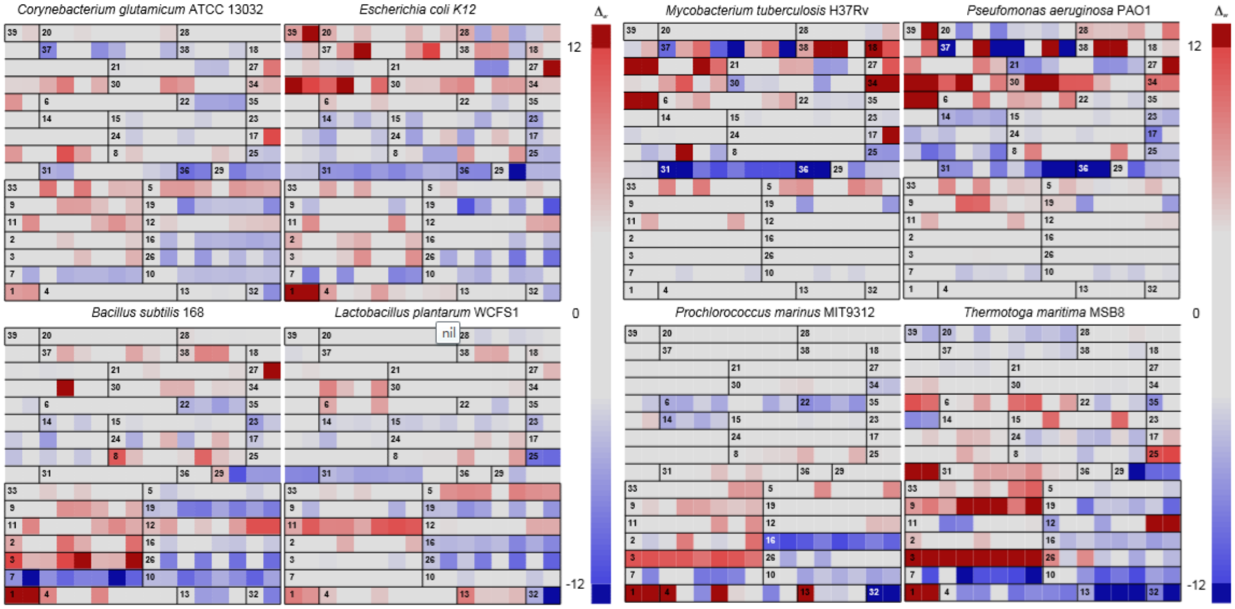


Fig. 2.1 Distribution of abundant and rare tetranucleotide words grouped according to structural features for eight bacterial species belonging to different taxonomic groups. Each cell represents one tetranucleotide. Frequencies of tetranucleotide were normalised by GC-content. They are ordered by base stacking energy from groups with higher energy at the top to lower energy at the bottom. Tetranucleotides belonging to the same group (there are 39 groups in total) have the same stacking energy and the same other stereo-chemical characteristics.

Reva and Tummler (2004) went further, using tetranucleotide patterns to assess the different conserved signatures for bacterial genomes by analysing oligonucleotide frequencies (Reva and Tummler, 2004). In this research, tetranucleotide frequencies were assessed according to the five stereo-chemical properties and it was found that the abundant and rare tetranucleotide words were distributed and grouped according to structural features (Figure 2.1). Two global features, PS and OUV, were further proposed as local and global compositional polymorphism measures calculated on the basis of the relative distribution of oligonucleotides. These measures also reflect taxon-specific properties such as DNA conformation, codon usage and specificity of DNA repair; restriction-modification systems provide the distinct ability to distinguish between bacterial sequences. Although these traits can be used in identifying distinct taxon-specific features through genome signatures, the evolutionary implications of OUP are still unknown.

Codon bias has been well studied by many researchers in the past decade owing to its importance in terms of translational efficacies and protein coding (Sharp and Li, 1987). Its influence in terms of oligonucleotide usage in E. *coli* has also been studied in research by Phillips (Phillips et al., 1987). It was shown that K-mers of up to hexamers were highly asymmetric, with highly abundant K-mers often associated with highly expressed codons. The opposite is also true, with some exceptions, where a few rarely used codons are associated with low abundant K-mers (e.g. palindromes). It was further shown in this paper that mononucleotides were heavily constrained in coding regions where the substitution rates of each codon position were context-dependent. This could result from the abundance of tRNA or the influence of translational machinery in the cytoplasm (Ikemura, 1981). Results also showed that dinucleotides do not randomly occur at each codon position and an unequal distribution of di-mer words is seen across different codons. For example, the presence of nucleotide G is higher in the first codon position but infrequent in the second compared to other nucleotides (Phillips *et al*., 1987). This could be related to the translational efficiency of each genome representing specific signatures of the sequence. Trinucleotides were found to be highly correlated to codons in coding regions of bacterial genomes where the density of genes is very high. Conversely, non-coding regions of the E. *coli* genome have been proven to correlate to 70% of the least abundant trinucleotides (Phillips *et al*., 1987). Although most of the K-mers up to hexa-mers were highly correlated to codon usage, frequencies of tetra- to hexa-mers were also regulated by some other factors, probably by DNA conformation constraints, as discussed above. Comparison of K-mer frequencies in coding and non-coding regions of prokaryotic genomes showed no significant difference, which again indicates an influence of some trans-genome acting forces (Reva and Tummler, 2004). Therefore, it was concluded that codon usage could not be a single evolutionary force driving the formation of OUP.

This view was shared by Fedorov *et. al.* (2002) in research on the regularities of context-dependent codon bias in eukaryotic genes (Fedorov *et al*., 2002). It was shown in this paper that the nucleotides surrounding codons could influence the choice of possible synonymous codons. Through the analysis of sequenced eukaryotic genomes, it was found that the immediate nucleotide denoted by the N_1 context-dependent nucleotide was statistically most significant. This was shown through the analysis of tetranucleotide frequencies, where the context-dependent nucleotide alongside the chosen synonymous codon represents half of the genome sequence in terms of composition. Because the proportion of eukaryotic coding regions is much smaller than the prokaryotic sequences, half of the genome sequence displaying such characteristics shows that the immediate context-dependent nucleotide applies selection pressure regulating codon selection. Conversely, the formation of tetranucleotides is non-random and affected by codon usage. Based on the above, the third hypothesis whereby the evolutionary diversification of oligonucleotide usage patterns can be driven by both codon usage and other factors such as context-dependent nucleotides, structural and functional constraints simultaneously is a distinct possibility. A common conclusion can be expressed through the discussion by Pride *et. al.* (2003), where both factors are taken into consideration and used as the evolutionary implications of the tetranucleotides-based phylogenetic method (Pride *et al*., 2003).

Pride *et. al.* (2003) analysed the evolutionary implications of microbial genomes using tetranucleotide frequency bias in 27 well-represented microbial genomes (Pride *et al*., 2003). Two different Markov methods were designed to determine the expected number of tetranucleotides by removing biases in mononucleotide frequencies and oligonucleotide components (Almagor, 1983; Rocha *et al*., 1998).

This was to ensure that the statistically meaningful tetranucleotide words were selected in the study and to eliminate possible biases from unequal base frequencies. An in-house program, Swaap, was used for the calculation of tetranucleotide usage deviation (TUD) between the observed and expected number of tetranucleotide words for various genomes (Pride and Blaser, 2002). The distribution of TUD patterns was shown to be well conserved for both intra- and interspecies comparisons. TUD between closely related species was well conserved irrespective of the GC content of genomes tested. This was also true for horizontally acquired genomic islands showing a substantial correlation to host genomes. A phylogenetic distance measure was proposed through this research, based on the differences between the TUD of two sequences.

Phylogenetic trees based on the TUD method were comparable to 16S rRNA trees. However, there were several differences between the resulting trees, which showed diverse evolutionary implications. The most important difference, which also ties with the concept of incongruence, was the unequal evolutionary rate after divergence from the common ancestor. Because of the conservative nature of 16s rRNA, OUP tends to evolve more rapidly based on the substantial difference in TUD between sequences (e.g. amelioration processes) (Lawrence and Ochman, 1997). Other possible explanations include functional constraints such as codon usage, which induce selective pressures speeding up nucleotide substitution rates shaping changes in TUD patterns (Pride and Blaser, 2002). Although there are several differences between the two methods, the OUP-based method undoubtedly also has several advantages over previous approaches. For nucleotide usage patterns reflecting structural and functional constraints, OUP analysis can provide alternative insights into selective pressure for microbial evolution. Nonetheless, the congruence between 16s rRNA trees and nucleotide usage pattern-based trees has shown the feasibility of this method in terms of phylogenetics and contains signals with evolutionary implications.

2.2) Relations Between OUP and Codon Usage in Bacteria Genomes

It was hypothesised that the driving forces of OUP diversification could be identified by an analysis of frequencies of context-dependent nucleotide substitution emissions in tetramers. The view of mutational signatures from context dependent mutation frequencies has also been shared by previous research (Meier et al., 2014). Emission was denoted as the likelihood of a given nucleotide in a sequence being substituted by one of the three other nucleotides if the states of the preceding and/or following nucleotide(s) are known. Several alternative hypotheses were also considered. One possibility was that the pattern of substitutions (emissions) may depend only on the state of the residue to be mutated but does not depend on the states of any neighbouring nucleotides. A second possibility was that the pattern of emissions may depend on the context of all surrounding nucleotides. Lastly, it was assumed that the pattern of emissions may depend on the context and also on the location of the mutated residue in the corresponding codon. In short, factors influencing the nucleotide substitution rate within tetra-mers can be related to the neighbouring nucleotides, or to the nucleotide position in the sequence (codon position) or to a combination thereof.

To evaluate this hypothesis, the analytical procedure described below was designed. Nucleotide sequences of homologous genes in different organisms were pairwise aligned and the number of substitutions was calculated. Then the subsets of substitutions taking place in a given context (position within the codon and/or the states of preceding or following nucleotides) were compared to the general emission pattern by using vector arithmetics. For example, a comparison of homologous sequences of *Corynebacterium jeikeium* K411 [NC_007164] and *C. kroppenstedtii* DSM 44385 [NC_012704] revealed that the mutated adenosine nucleosides (A) in the protein coding sequences of NC_007164 were substituted by C, G and T with frequencies of 0.5, 0.36 and 0.14, respectively. Taking a subset of these substitutions subject to a condition that the adenosine residues are located only in the second positions of codons (location factor) and are preceded by another A (context factor), the corresponding likelihoods values were 0.45, 0.5 and 0.05. Based on these two emission patterns, the vector distance was calculated as:

$$\sqrt{(\mathbf{0.5-0.45})^{\mathbf{2}}+(\mathbf{0.36-0.5})^{\mathbf{2}}+(\mathbf{0.14-0.05})^{\mathbf{2}}} = 0.18$$

Vector distances were calculated for all possible combinations of mutated (location factor) and context nucleotides (context factor) in a range of 10 residues upstream to 10 residues downstream from the mutated residue. This was an attempt to identify all possible forces influencing nucleotide substitutions within a 10-base nucleotide flanking region for tetra-mer OUP.

2.2.1) Selection of Bacterial Genomes for Case Studies

To perform this case study, several groups of microorganisms representing different phylogenetic branches and various taxonomic levels were selected. These included subspecies of *Prochlorococcus marinus*; representatives of genera *Bacillus*, *Corynebacterium*, *Lactobacillus*, *Mycobacterium* and *Pseudomonas*; and representatives of different orders of Gamma-Proteobacteria and archaeal group *Thermotogaceae*. The following rationales were considered for choosing these groups of bacteria for the case studies.

Bacillus

The *Bacillus* genus belonging to the Firmicutes phylum is a Gram-positive rod-shaped bacterium and is well characterised in literature in terms of both phylogeny and its uses in medicine, biotechnology and as a model organism. *B. subtilis* is one of the best studied and understood bacteria in both molecular and cellular biology. Its superb genetic amenability and relatively large size have provided powerful tools to investigate a bacterium in all possible aspects (Graumann, 2012). This specie is also widely used in the food industry for high pressure processing for food such as milk, cheese and beef (Soni et al., 2016). *Bacillus* species such as a strain of *B. halodurans* reduce the alkalinity of cement industry waste through their alkaliphilic properties (Kunal *et al*., 2016). Certain strains of *B. licheniformis* have also been found to be associated with diesel fuel degradation and plant growth promotion (Stevens *et al*., 2017). In the field of medicine, *B. cereus* is pathogenic and associated with high toxicity that causes food poisoning (Miller *et al*., 2018). The genus *Bacillus* is very diverse and well characterised in terms of its phylogeny by Xu and Cote (2003), using 16S-23S internal transcribed spacer regions (Xu and Cote, 2003). The 11 different *Bacillus* species and strains used in this case study are shown in Table 2.1.

Table 2.1: Genomes of *Bacillus* used in this study

Genome sequence NCBI ID	Strains
NC_000964	Bacillus subtilis subsp. subtilis str. 168 chromosome, complete genome.
NC_002570	Bacillus halodurans C-125 chromosome, complete genome.
NC_003909	Bacillus cereus ATCC 10987, complete genome.
NC_006270	Bacillus licheniformis ATCC 14580 chromosome, complete genome.
NC_006582	Bacillus clausii KSM-K16, complete genome.
NC_009674	Bacillus cytotoxicus NVH 391-98 chromosome, complete genome.
NC_009848	Bacillus pumilus SAFR-032 chromosome, complete genome.
NC_013791	Bacillus pseudofirmus OF4 chromosome, complete genome.

NC_014103	Bacillus megaterium DSM 319 chromosome, complete genome.
NC_014829	Bacillus cellulosilyticus DSM 2522 chromosome, complete genome.
NC_015634	Bacillus coagulans 2-6 chromosome, complete genome.

Lactobacillus

Lactobacillus, similar to the *Bacillus* genus, is also a rod-shaped bacterium belonging to the Gram-positive Firmicutes phylum. These bacteria form a major part of the lactic acid bacteria group, which converts sugar to lactic acid (Makarova *et al*., 2006). This group of species is mostly associated with the microbiota within many body sites, such as the digestive and genital system. In women, the *Lactobacillus* count determines if the vaginal microbiota is healthy or not and is associated with a common disease in premenopausal women called bacterial vaginosis. This disease is characterised by a depletion of *lactobacilli* population and the presence of Gram-negative anaerobes, or in some cases Gram-positive cocci, and aerobic pathogens (Cribby *et al*., 2008). In the digestive system, *Lactobacillus* species are used as a probiotic to treat diarrhoea and other more serious digestive problems such as irritable bowel syndrome and infection by the ulcer-causing bacterium *Helicobacter pylori* (Martin *et al*., 2013; Ruggiero, 2014). For food production, bacterocin can be isolated from *lactobacillus* strains, which serve as food biopreservatives and are used for fermentation in foods such as yoghurt, cheese and beer. This group is also well characterised and its phylogeny has been thoroughly studied and identified by Zheng *et. al.* (2015). Table 2.2 shows the different strains and species of *Lactobacillus* used in this study.

Table 2.2: Genomes of *Lactobacillus* used in this study

Genome sequence NCBI ID	Strains
NC_005362	Lactobacillus johnsonii NCC 533, complete genome.
NC_006814	Lactobacillus acidophilus NCFM chromosome, complete genome.
NC_007576	Lactobacillus sakei subsp. sakei 23K chromosome, complete genome.
NC_007929	Lactobacillus salivarius UCC118 chromosome, complete genome.
NC_008054	Lactobacillus delbrueckii subsp. bulgaricus ATCC 11842 chromosome, complete genome.
NC_008497	Lactobacillus brevis ATCC 367, complete genome.
NC_008526	Lactobacillus casei ATCC 334 chromosome, complete genome.
NC_008530	Lactobacillus gasseri ATCC 33323 chromosome, complete genome.
NC_009004	Lactococcus lactis subsp. cremoris MG1363 chromosome, complete genome.
NC_009513	Lactobacillus reuteri DSM 20016 chromosome, complete genome.
NC_010080	Lactobacillus helveticus DPC 4571, complete genome.
NC_010610	Lactobacillus fermentum IFO 3956, complete genome.
NC_013198	Lactobacillus rhamnosus GG chromosome, complete genome.
NC_014106	Lactobacillus crispatus ST1, complete genome.
NC_014554	Lactobacillus plantarum subsp. plantarum ST-III chromosome, complete genome.
NC_014724	Lactobacillus amylovorus GRL 1112 chromosome, complete genome.
NC_015214	Lactobacillus acidophilus 30SC chromosome, complete genome.
NC_015428	Lactobacillus buchneri NRRL B-30929 chromosome, complete genome.
NC_015602	Lactobacillus kefiranofaciens ZW3 chromosome, complete genome.
NC_015930	Lactococcus garvieae ATCC 49156, complete genome.
NC_015975	Lactobacillus ruminis ATCC 27782 chromosome, complete genome.
NC_015978	Lactobacillus sanfranciscensis TMW 1.1304 chromosome, complete genome.

Corynebacteria

Corynebacteria belong to the phylum Actinobacteria and are Gram-positive rod-shaped bacteria that in some phases of life become club-shaped, leading to the name. *Corynebacterium* is a diverse group with a wide range of ecological niches such as soil, vegetables, sewage, skin and microbiota of animals and humans (Burkovski, 2008). The two most notable areas where this group of species is found are the industrial and medical fields. Diphtheria is a human infection caused by the species *C. diphtheria*, *C. ulcerans* and *C. pseudotuberculosis (Both et al., 2015)*. The ideal intrinsic attributes of *C. glutamicum* as a biocatalyst play an important role in the production of amino acids and other commodity chemicals (Vertes *et al*., 2012). The phylogeny of *Corynebacteria* is well documented in a study by Pascual *et. al.* (1995) using comparison of 16S rRNA gene sequences (Pascual *et al*., 1995). Table 2.3 shows the different strains and species of *Corynebacteria* used in this study.

Table 2.3: Genomes of *Corynebacteria* used in this study

Genome sequence NCBI ID	Strains
NC_003450	Corynebacterium glutamicum ATCC 13032, complete genome.
NC_004369	Corynebacterium efficiens YS-314 chromosome, complete genome.
NC_007164	Corynebacterium jeikeium K411 chromosome, complete genome.
NC_010545	Corynebacterium urealyticum DSM 7109 chromosome, complete genome.
NC_012704	Corynebacterium kroppenstedtii DSM 44385 chromosome, complete genome.
NC_015673	Corynebacterium resistens DSM 45100 chromosome, complete genome.
NC_015859	Corynebacterium variabile DSM 44702 chromosome, complete genome.
NC_016781	Corynebacterium pseudotuberculosis 3/99-5 chromosome, complete genome.
NC_016785	Corynebacterium diphtheriae CDCE 8392 chromosome, complete genome.

NC_017317	Corynebacterium ulcerans 809 chromosome, complete genome.
NC_020302	Corynebacterium halotolerans YIM 70093 = DSM 44683 chromosome, complete genome.
NC_021663	Corynebacterium terpenotabidum Y-11, complete genome.
NC_021915	Corynebacterium maris DSM 45190, complete genome.

Mycobacteria

Mycobacteria belong to the Actinobacteria phylum. Although they do not have wholly Gram-positive characteristics, they are characterised as Gram-positive bacteria owing to the lack of an outer membrane. The determination of bacterium *M. tuberculosis* as Gram-positive or Gram-negative is also controversial (Fu and Fu-Liu, 2002). The most notable species in this group are *M. tuberculosis* and *M. leprae,* causing tuberculosis and leprosy respectively. The large amount of study devoted to this group is due to the high impact of these deadly pathogens, which cause over 1.8 million deaths a year (Gordon and Parish, 2018). Because of the high antibiotic resistance by different strains of *M. tuberculosis*, the evolution and traits of this pathogen have been well researched and understood (Comas *et al*., 2013). The phylogeny as a whole for the group *Mycobacteria* has been analysed and well documented, based on comparison of 16S rRNA or on concatenated housekeeping genes. Both these methods support the separation of rapidly growing and slow-growing species in this genus (Tortoli *et al*., 2017). Table 2.4 shows the different strains and species of *Mycobacterium* used in this study.

Table 2.4: Genomes of *Mycobacterium* used in this study

Genome sequence NCBI ID	Strains
NC_000962	Mycobacterium tuberculosis H37Rv complete genome.
NC_002677	Mycobacterium leprae TN chromosome, complete genome.
NC_002944	Mycobacterium avium subsp. paratuberculosis K-10, complete genome.

NC_008146	Mycobacterium sp. MCS chromosome, complete genome.
NC_008595	Mycobacterium avium 104 chromosome, complete genome.
NC_008596	Mycobacterium smegmatis str. MC2 155 chromosome, complete genome.
NC_008611	Mycobacterium ulcerans Agy99 chromosome, complete genome.
NC_008705	Mycobacterium sp. KMS chromosome, complete genome.
NC_008726	Mycobacterium vanbaalenii PYR-1 chromosome, complete genome.
NC_009077	Mycobacterium sp. JLS chromosome, complete genome.
NC_009338	Mycobacterium gilvum PYR-GCK chromosome, complete genome.
NC_010397	Mycobacterium abscessus chromosome, complete sequence.
NC_010612	Mycobacterium marinum M chromosome, complete genome.
NC_015576	Mycobacterium sp. JDM601 chromosome, complete genome.
NC_016947	Mycobacterium intracellulare MOTT-02 chromosome, complete genome.
NC_017904	Mycobacterium sp. MOTT36Y chromosome, complete genome.

Pseudomonas

Pseudomonas belongs to the Proteobacteria phylum. It is a Gram-negative bacterium with a great deal of metabolic diversity colonised in a wide range of niches (Silby *et al*., 2011). *P. aeruginosa* specifically is an excellent focus for scientific research and has proven important in clinical studies as a model organism in biofilm formation (Mann and Wozniak, 2012). This specie is an opportunistic pathogen with low antibiotic susceptibility adding value to clinical studies (Lister *et al*., 2009). *P. syringae* is another well studied and classified specie owing to its effects on plants as a plant pathogen (Marcelletti and Scortichini, 2014). Some other *Pseudomonas* species are used as bioremediation agents, such as *P. putida*, which contains two operons that specify a pathway for the degradation of aromatic hydrocarbons (Marques and Ramos, 1993), and the KC strain of *P. stutzeri,* which has also been found to have the ability to degrade carbon tetrachloride (Sepulveda-Torres *et al*., 1999). Because of its wide applications and being the largest Gram-negative genus, the phylogeny of *Pseudomonas* has been well studied and identified using multilocus sequence analysis approaches (Gomila *et al*., 2015). Table 2.5 shows the different strains and species of *Pseudomonas* used in this study.

Table 2.5: Genomes of *Pseudomonas* used in this study

Genome sequence NCBI ID	Strains
NC_002516	Pseudomonas aeruginosa PAO1 chromosome, complete genome.
NC_002947	Pseudomonas putida KT2440 chromosome, complete genome.
NC_004129	Pseudomonas protegens Pf-5 chromosome, complete genome.
NC_007005	Pseudomonas syringae pv. syringae B728a chromosome, complete genome.
NC_007492	Pseudomonas fluorescens Pf0-1 chromosome, complete genome.
NC_008027	Pseudomonas entomophila L48 chromosome, complete genome.
NC_015379	Pseudomonas brassicacearum subsp. brassicacearum NFM421 chromosome, complete genome.
NC_015410	Pseudomonas mendocina NK-01 chromosome, complete genome.
NC_015556	Pseudomonas fulva 12-X chromosome, complete genome.
NC_015740	Pseudomonas stutzeri ATCC 17588 = LMG 11199 chromosome, complete genome.
NC_017986	Pseudomonas putida ND6 chromosome, complete genome.
NC_020829	Pseudomonas denitrificans ATCC 13867, complete genome.

Gamma-Proteobacteria

A higher taxonomic level was chosen grouping orders of *Enterobacteriales*, *Alteromonadales*, *Xanthomonadales* and *Pseudomoadales* as a test case for OUP based method. This group of species includes many familiar pathogens such as *Salmonella* and *Escherichia coli* and plant pathogens such as *Xylella fastidiosa*. Throughout a long history, *E. coli* has been used as a model organism in microbiology owing to its ease of manipulation and laboratory culture and its important application in biological engineering and industrial microbiology (Lee, 1996). *E. coli* was one of the first organisms to be sequenced and named in the complete genome sequence *E. coli* K-12 (Blattner *et al*., 1997). *E. coli* has industrial importance as the host organism for the enhancement of biochemical production in metabolic engineering (Chen *et al*., 2013). In terms of clinical importance, *S. enterica* is responsible for a variety of diseases transmitted through food, including gastroenteritis and typhoid fever (Blanc-Potard *et al*., 1999). *X. fastidiosa* is responsible for many plant diseases all over the world, such as citrus variegated chlorosis and Pierce's disease, which affects grapevines, citrus, coffee and almonds and has a great agricultural impact on yields (Richard A. Redak et al., 2004). These orders are also well classified in terms of their phylogeny, with many studies using various techniques such as comparison of housekeeping genes and 16S rRNA (Williams et al., 2010). Table 2.6 shows the different strains and species of Gamma-Proteobacteria used in this study.

Table 2.6: Genomes of Gamma-Proteobacteria used in this study

Genome sequence NCBI ID	Strains
NC_002488	Xylella fastidiosa 9a5c chromosome, complete genome.
NC_003919	Xanthomonas axonopodis pv. citri str. 306 chromosome, complete genome.
NC_004347	Shewanella oneidensis MR-1 chromosome, complete genome.
NC_004556	Xylella fastidiosa Temecula1 chromosome,

	complete genome.
NC_004631	Salmonella enterica subsp. enterica serovar Typhi str. Ty2 chromosome, complete genome.
NC_007954	Shewanella denitrificans OS217, complete genome.
NC_009085	Acinetobacter baumannii ATCC 17978 chromosome, complete genome.
NC_009832	Serratia proteamaculans 568 chromosome, complete genome.
NC_010468	Escherichia coli ATCC 8739 chromosome, complete genome.
NC_010506	Shewanella woodyi ATCC 51908 chromosome, complete genome.
NC_013971	Erwinia amylovora ATCC 49946 chromosome, complete genome.

Prochlorococcus marinus

Prochlorococcus marinus are small Gram-negative bacteria belonging to the Cyanobacteria phylum. These species are among the major primary producers in the ocean, responsible for a large percentage of the photosynthetic production of oxygen (Flombaum *et al*., 2013). *Prochlorococcus* is the smallest known photosynthetic organism and probably the most abundant of this type on earth owing to its size. These organisms are mainly found from the surface of the ocean to a depth of 200 m in the 40°S to 40°N latitudinal band and can adapt to a nutrient-deprived environment (Partensky *et al*., 1999). The environmental constraints linked to the evolution of these organisms and their ecological importance have also proven interesting and therefore the phylogeny differentiating different ecotypes has been well studied (Prabha *et al*., 2014). Table 2.7 shows the different strains and species of *P. marinus* used in this study.

Table 2.7: Genomes of *P. marinus* used in this study

Genome sequence NCBI ID	Strains
NC_005042	Prochlorococcus marinus subsp. marinus str. CCMP1375 chromosome, complete genome.

NC_005072	Prochlorococcus marinus subsp. pastoris str. CCMP1986 chromosome, complete genome.
NC_007335	Prochlorococcus marinus str. NATL2A chromosome, complete genome.
NC_007577	Prochlorococcus marinus str. MIT 9312, complete genome.
NC_008816	Prochlorococcus marinus str. AS9601, complete genome.
NC_008817	Prochlorococcus marinus str. MIT 9515, complete genome.
NC_008819	Prochlorococcus marinus str. NATL1A, complete genome.
NC_009091	Prochlorococcus marinus str. MIT 9301, complete genome.
NC_009840	Prochlorococcus marinus str. MIT 9215 chromosome, complete genome.
NC_009976	Prochlorococcus marinus str. MIT 9211, complete genome.

Thermotogaceae

Thermotogaceae belongs to the Thermotogae phylum and is a Gram-negative hyperthermophilic bacterium whose cell is wrapped in a unique sheath-like outer membrane, called a "toga". The genus Thermotoga consists of some of the most thermophilic bacteria known, with optimum growth temperatures of up to 80°C (Huber *et al*., 1986). Because of this trait, these organisms are viewed as model systems for studying adaptation and microbial evolution at high temperatures (Mongodin *et al*., 2005). These properties also have biotechnological applications such as catalysing a variety of high-temperature reactions and degrading simple and complex carbohydrates (Conners *et al*., 2006). The diversity of genomic region within Thermotoga species is also of interest, with high horizontal gene transfer rates from both archaeal and bacterial species (Nelson *et al*., 1999). The physiological difference in gene content between ecotypes for the adaptation to different environments has driven understanding of the evolution of this genus (Nesbo *et al*., 2006; Bhandari and Gupta, 2014). Table 2.8 shows the different strains and species of Thermotogaceae used in this study.

Table 2.8: Genomes of Thermotogaceae used in this study

Genome sequence NCBI ID	Strains
NC_000853	Thermotoga maritima MSB8 chromosome, complete genome.
NC_009486	Thermotoga petrophila RKU-1 chromosome, complete genome.
NC_009828	Thermotoga lettingae TMO chromosome, complete genome.
NC_010483	Thermotoga sp. RQ2 chromosome, complete genome.
NC_011978	Thermotoga neapolitana DSM 4359 chromosome, complete genome.
NC_013642	Thermotoga naphthophila RKU-10, complete genome.
NC_014926	Thermovibrio ammonificans HB-1 chromosome, complete genome.
NC_015707	Thermotoga thermarum DSM 5069 chromosome, complete genome.
NC_016148	Thermovirga lienii DSM 17291 chromosome, complete genome.

2.2.2) Analysis of Emission Patterns Calculated for Different Groups of Microorganisms

Diagrams in Figure 2.2 represent distributions of average values (AVR) ± 2.5× standard deviations (STD) calculated for the first, second and third codon positions of mutated residues. This interval was chosen based on a 95% confidence interval of a normal distribution. An assumption was that the higher vector distances with a smaller STD range should be an indication of stronger specificity of the emission pattern. In other words, this will imply stronger selective pressure on nucleotide substitutions. Inspection of the diagrams in Figure 2.2 showed that the emission pattern constraints were predominantly codon-specific in all taxonomic groups. Thus, the emission patterns of residues at the first codon position were influenced by the states of the second and third residues in the same codon. Similarly, for the second and third codon positions, their substitution patterns were influenced by the states of other nucleotides in the same codon. This can be seen from the change in fluctuation of likelihood of substitution rate change from other nucleotides flanking the base nucleotides. On the contrary, the emission patterns were generally not influenced at all by the states of neighbour residues from the other codons. These signals were recognisable in all the taxonomic groups of Eubacteria. However, the differences between the emission patterns calculated for several groups were statistically unreliable owing to strong background noise (see Gamma-Proteobacteria, *Lactobacillus* and *Thermatoga* in Figure 2.2). These background noises could be contributed by other forces driving the rate of substitution.

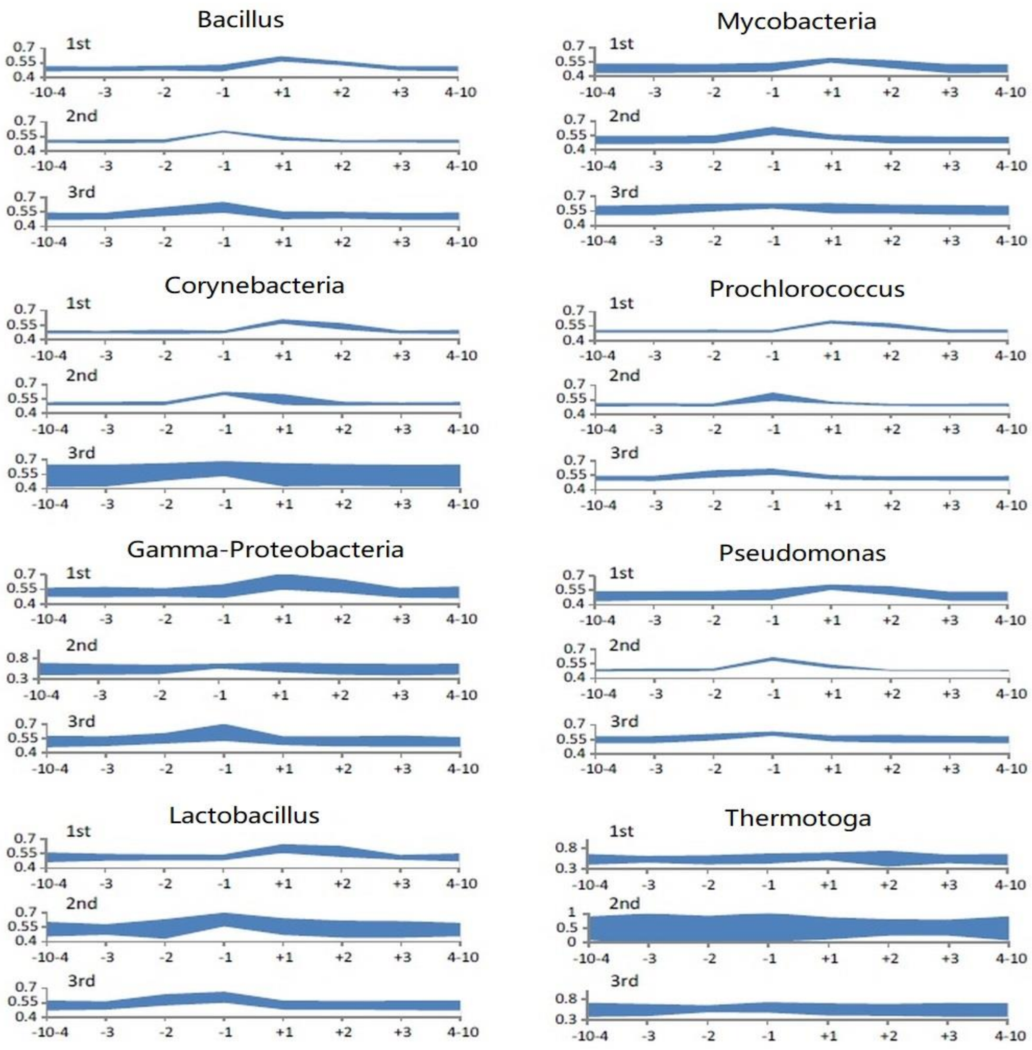


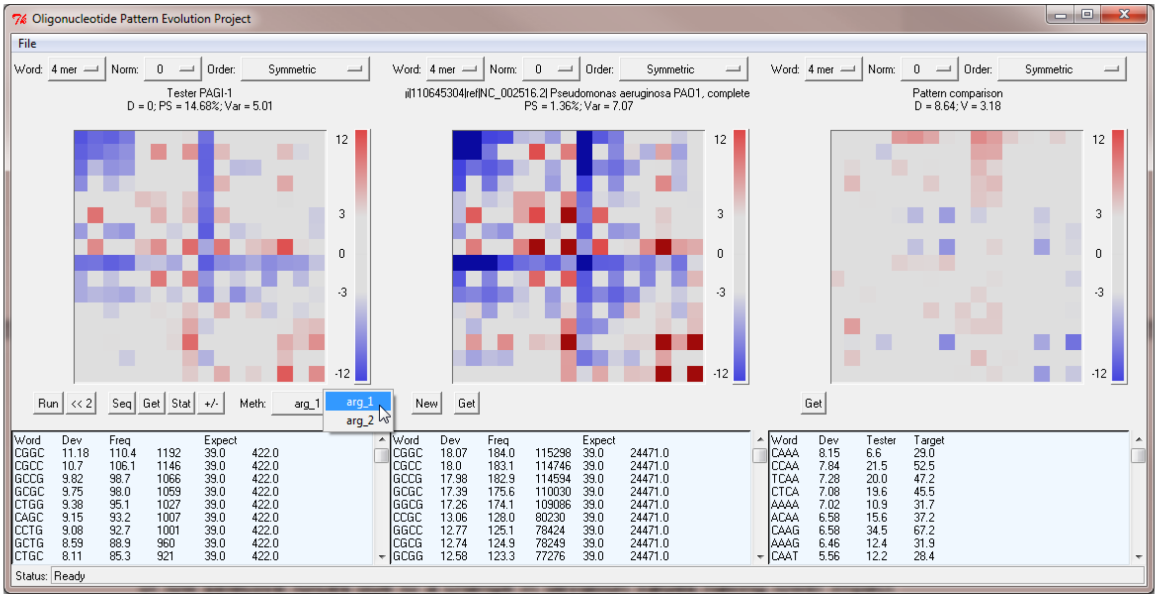
Fig. 2.2 Emission patterns of the codon-specific residues influenced by the states of context residues. The diagrams of the emission pattern deviations were organised by the first, second and third codon positions. X axes depict the positions of the controlled context residues relative to

the mutated residues. Data for the preceding and posterior 10-to-4 residues were summarised in the two outermost categories. Y axes show the vector distances between the global emission pattern and the patterns calculated for each category. Bandwidth depicts the values AVR ± 2.5×STD.

2.3) Program Modelling of Context and Codon Dependent Genome Evolution

Selective pressure on nucleotide substitution based on composition has been well studied and documented (Bulmer, 1991). In the case of foreign inserts of genomic islands, amelioration is a key aspect of directional selective pressure where the base DNA composition of the transferred genetic sequence undergoes nucleotide substitutions over time and reflects similarly in DNA composition to the recipient genome (Lawrence and Ochman, 1997). In a previous study by Yu (2014), analysis of change in the substitution rate with regard to OUP distance between different genomic islands and host was done (Yu, 2014). In the case study, a combination of four genomic islands from known hosts (tester) in five target sequences (target) creating 20 scenarios was analysed to identify the relationship between OUP distances and selective pressure from different genome compositions.

Selective pressure on a single nucleotide from oligonucleotide usage constraints (deviation) was calculated based on surrounding K-mers containing that specific nucleotide from di- to tetra-mer frequencies. The deviation for each oligonucleotide word was measured as the total observed K-mer word count decreased by the expected count. This measure emphasises the over- and under-represented K-mer words, which are important genome signatures in the sequence under study. Taking for example the sequence “GTGGGTCGTGTA” with T as the base nucleotide under selective pressure, the surrounding K-mers up to tetra-mers for this nucleotide include GGGT, GGT, GT, GGTC, GTC, TC, GTCG, TCG, TCGT. The likelihood of substitution to any other nucleotide under selective pressure was then calculated as the difference in all changes to surrounding K-mer words. For the above example, a change from T to A, the K-mer words taken into consideration are GGGA, GGA, GA, GGAC, GAC, AC, GACG, ACG and ACGT. In this case, the total likelihood measure for nucleotide A is the sum of differences of each word deviation between the tester and target normalised by a weighting scheme. This weighting scheme ensures that the impact of all word deviations on the base nucleotide are equal where di-mer deviations are halved, tri-mers are divided by three and so on. Other possible nucleotide substitutions were also calculated in this way, forming a row vector of normalised total deviation value for each nucleotide substitution [Dev_A, Dev_C, Dev_G, Dev_T]. This row vector was then converted to a substitution probability based on a logistic model where high deviation values have a higher probability of substituting to that specific nucleotide and vice versa. A 0.33 capacity value was set to restrict the total probability of substitution to be less than one. If the total probability of substitution to all other nucleotides did not equal to one, then the remaining probability was equal to the chance that the base nucleotide did not undergo substitution. The row vector for each nucleotide substitution probability was then accumulated up to a sum of one, with each cell within the vector representing a region of probability of substitution. For example, for vector [Dev_A, Dev_C, Dev_G, Dev_T], an accumulated vector was [0.02,0.24,0.75,1], representing the probability of substitution of A in region (0,0.02], C in (0.02,0.24] and so on. Nucleotide substitutions were generated depending on a random number generated between zero and one. This algorithm was implemented as a program written on Python 2.5 with a graphical user interface building upon the Python Tkinter and Pmw modules. The program interface is shown in Figure 2.3.



```
Word: 4 mer   Norm: 0   Order: Symmetric
Tester PAGI-1
D = 0; PS = 14.68%; Var = 5.01

Word: 4 mer   Norm: 0   Order: Symmetric
i|110645304|ref|NC_002516.2| Pseudomonas aeruginosa PAO1, complete
PS = 1.36%; Var = 7.07

Word: 4 mer   Norm: 0   Order: Symmetric
Pattern comparison
D = 8.64; V = 3.18

Meth: arg_1
```

```
Word   Dev     Freq            Expect
CGGC   11.18   110.4   1192    39.0    422.0
CGCC   10.7    106.1   1146    39.0    422.0
GCCG   9.82    98.7    1066    39.0    422.0
GCGC   9.75    98.0    1059    39.0    422.0
CTGG   9.38    95.1    1027    39.0    422.0
CAGC   9.15    93.2    1007    39.0    422.0
CCTG   9.08    92.7    1001    39.0    422.0
GCTG   8.59    88.9    960     39.0    422.0
CTGC   8.11    85.3    921     39.0    422.0
```

```
Word   Dev     Freq            Expect
CGGC   18.07   184.0   115298  39.0    24471.0
CGCC   18.0    183.1   114746  39.0    24471.0
GCCG   17.98   182.9   114594  39.0    24471.0
GCGC   17.39   175.6   110030  39.0    24471.0
GGCG   17.26   174.1   109086  39.0    24471.0
CCGC   13.06   128.0   80230   39.0    24471.0
GGCC   12.77   125.1   78424   39.0    24471.0
CGCG   12.74   124.9   78249   39.0    24471.0
GCGG   12.58   123.3   77276   39.0    24471.0
```

```
Word   Dev     Tester  Target
CAAA   8.15    6.6     29.0
CCAA   7.84    21.5    52.5
TCAA   7.28    20.0    47.2
CTCA   7.08    19.6    45.5
AAAA   7.02    10.9    31.7
ACAA   6.58    15.6    37.2
CAAG   6.58    34.5    67.2
AAAG   6.46    12.4    31.9
CAAT   5.56    12.2    28.4
```

```
Status: Ready
```

Fig. 2.3 Interface of the in-house program for simulation of OUP evolutionary changes. Left panel represents frequencies of oligonucleotides (from 2- to 7-mers) in the tester sequence, which will be modified by random mutations generated using an algorithm designed to match the OUP pattern of the target sequence shown in the centre panel. The right panel shows differences between tester and target OUP patterns at the current state. Different evolutionary algorithms were programmed as Python scripts and made available for trials by selecting them from the drop-box menu, as shown in the figure.

The first algorithm simulating the OUP evolution was designed to test the influence of possible selective pressure caused by surrounding K-mers and by the initial OUP distance between the tester and target sequences. The conversion of the substitution probability was highly influenced by surrounding K-mers. In other words, if the logistic function has a flatter gradient, then the conversion of deviation values to the probability of substitution is higher and vice versa. In biological terms, a flat gradient is a result of weak forces from neighbouring nucleotides on substitution probabilities (Figure 2.4). The converse is also true for sharp gradient curves showing high selective forces by the surrounding nucleotides. Hence, the impact of surrounding OUP in this model has a direct influence on the evolution of the tester sequence towards the target sequence.

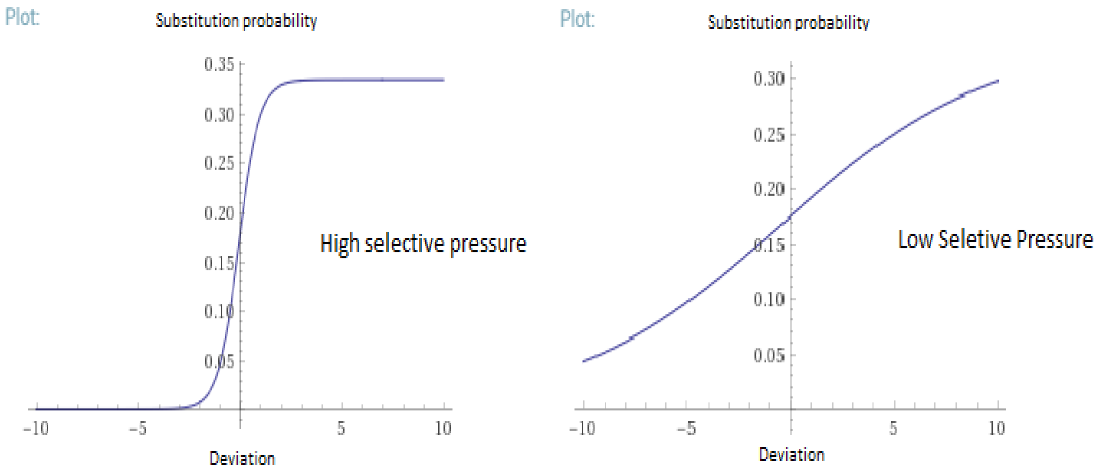


Fig. 2.4 Conversion of total deviation from surrounding K-mers to likelihoods of substitution of base nucleotide displaying selective pressure of OUP on nucleotide substitution. A sharp curve (left) shows higher selective pressure by surrounding K-mers whereby a smaller deviation value can be converted to a high substitution rate to another nucleotide. Vice versa, for the flatter curve (right), low selective pressure can be seen from a more lenient conversion of deviation to substitution rate.

Another influential factor was the relative difference in OUP between tester and target sequences. A large distance between OUP leads to a high initial substitution rate and vice versa. This model was designed to mimic the amelioration process where a tester sequence undergoes higher directional selective pressure caused by the difference in the oligonucleotide composition compared to the target sequence. This was later shown through a high correlation from a linear model fitting between the values of tester, target and tester target difference in terms of OUP distance against the substitution rate of the logistic model. The 20 combinations consist of four tester sequences, *Bacillus subtilis* 168, *Escherichia coli* CFT073, *Streptomyces coelicolor* A3, *Pseudomonas aeruginosa* pathogenicity island (PAGI) 1, and five target sequences, *Bacillus subtilis* 168, *Escherichia coli* K12 substr MG1655, *Streptomyces griseus* NBRC 13350, *Xylella fastidiosa* 9a5c and *Pseudomonas aeruginosa* PA01. This same OUP-based substitution model was tested on four PAGI from the same host pKLC102 in estimating the time of insertion of these genomic islands (Yu, 2014). Although the model was successful in estimating a relative time of insertion for each genomic island in comparison to the others, it failed to determine an absolute time frame in years. The model has also shown discrepancies in time estimations based on different K-mers, which was especially visible for genomic islands that had undergone evolution for a longer period of time. A possible hypothesis from this is that the substitution rate is not entirely based on the selection pressure of context residues and cannot be comprehended by only this OUP evolution model. The alternative hypothesis could be that the evolution of OUP is driven by other non-compositional factors such as the codon bias adaptation as well.

Context-dependent nucleotides and codon usage are other probable evolution implications for OUP, which can be tested by simulating substitutions of sequences based on target codon usage. An assumption of this model is that assigning different likelihoods for mutations will possibly merge both forces from codon usage and OUP of a tester sequence with those of the target sequence. Success with such a substitution model will give credit to the hypothesis that the OUP evolution is driven by the codon usage adjustment in bacterial genomes. For example, if one knows sequence A is the ancestral state evolving into sequence B, then the OUP distances between A and B should decrease alongside an increase in the number of substitutions from A under the selection pressure of B’s codon usage. The simulation algorithm was designed to first calculate the observed codon frequencies in the tester and target genomes as the base for calculating substitution likelihood probabilities in the tester sequence, allowing greater likelihood for codons showing greater difference in their frequencies (see an example of a codon frequency matrix in Table 2.9). The substitution likelihood probability for each base nucleotide position was derived as the codon frequency containing the base nucleotide over the sum of all other codons containing other variations of the base nucleotide of that codon position within other codons. Hence, the selection force on the base nucleotide takes into consideration the other codon positions as context residues. For example, for amino acid leucine, possible codons are CTT, CTG, TTA, TTG, CTA and CTC. If one looks at possible substitution at third codon position with nucleotide T, the likelihood to have a substitution of nucleotide A based on the frequency vector [CTT,304], [CTG,244], [CTC,105], [CTA,62] will be calculated as:

$$\mathrm{A} = \frac{62}{304 + 244 + 105 + 62} = 0.08$$

Similarly, for other nucleotides [C,0.15], [G,0.34] and [T,0.43] are also calculated in this way. These likelihood values are then accumulated to a sum of one and in each region, e.g. in A 0 to 0.08, C 0.08 to 0.23 etc, a random number generated between 0 and 1 will determine if there will be a nucleotide substitution or not based on the region in which this number falls. This is done for each nucleotide and a simulated sequence of a certain number of nucleotide substitutions is then created. A series of simulated sequences of various numbers of nucleotide substitutions can then be used as a probable evolutionary path between the two sequences based on the context and codon usage forces. An example is shown in Figure 2.5, where the comparison of OUP distances and nucleotide substitution numbers were simulated on a fragment of the genome of *Mycobacterium* sp. MCS was used as a tester sequence, which was allowed to evolve towards five different mycobacterial genomes used as target sequences. Genomic fragments of 50 kb comprising exclusively coding sequences taken from the respective genomes were used in this example.

Table 2.9 Example of a codon usage frequency matrix

Amino Acid	Codon	Frequency	Amino Acid	Codon	Frequency
Alanine (A)	GCA	260	Asparagine (N)	AAT	264
	GCT	239		AAC	202
	GCG	227	Proline (P)	CCG	156
	GCC	130		CCT	143
Cysteine (C)	TGT	45		CCA	81
	TGC	44		CCC	34
Aspartic Acid (D)	GAT	438	Glutamine (Q)	CAA	263
	GAC	239		CAG	228
Glutamic Acid (E)	GAA	699	Arginine (R)	AGA	159
	GAG	317		CGT	139
Phenyl-alanine (F)	TTT	293		CGC	105
	TTC	155		CGG	72
Glycine (G)	GGA	269		CGA	48
	GGC	222		AGG	45
	GGT	184	Serine (S)	TCA	161
	GGG	112		TCT	147
Histidine (H)	CAT	185		AGC	138
	CAC	83		TCC	77
Isoleucine (I)	ATT	469		AGT	75
	ATC	296		TCG	57
	ATA	116	Threonine (T)	ACA	276
Lysine (K)	AAA	635		ACG	177
	AAG	271		ACT	115
Leucine (L)	CTT	304		ACC	82
	CTG	244	Valine (V)	GTT	273
	TTA	242		GTA	203
	TTG	164		GTG	199
	CTC	105		GTC	187
	CTA	62	Tryptophan (W)	TGG	70
Methionine (M)	ATG	280	Tyrosine (Y)	TAT	251
				TAC	121

Figure 2.5 demonstrates a moderately high correlation between the decrease in OUP distances and the number of nucleotide substitutions based on r-squared values calculated for the target mycobacteria species (*M. avium* 104, *M. marinum M*, *Mycobacterium* sp. MOTT36Y). From this observation one can conclude that

adaptation of the codon usage of the organism to fluctuations of tRNA concentrations can explain to some extent the evolutionary diversification of OUP during speciation. As this adaptation has a global effect on the complete genome, the resulting OUP can also be considered as an overall genomic signature reflecting the evolutionary process.

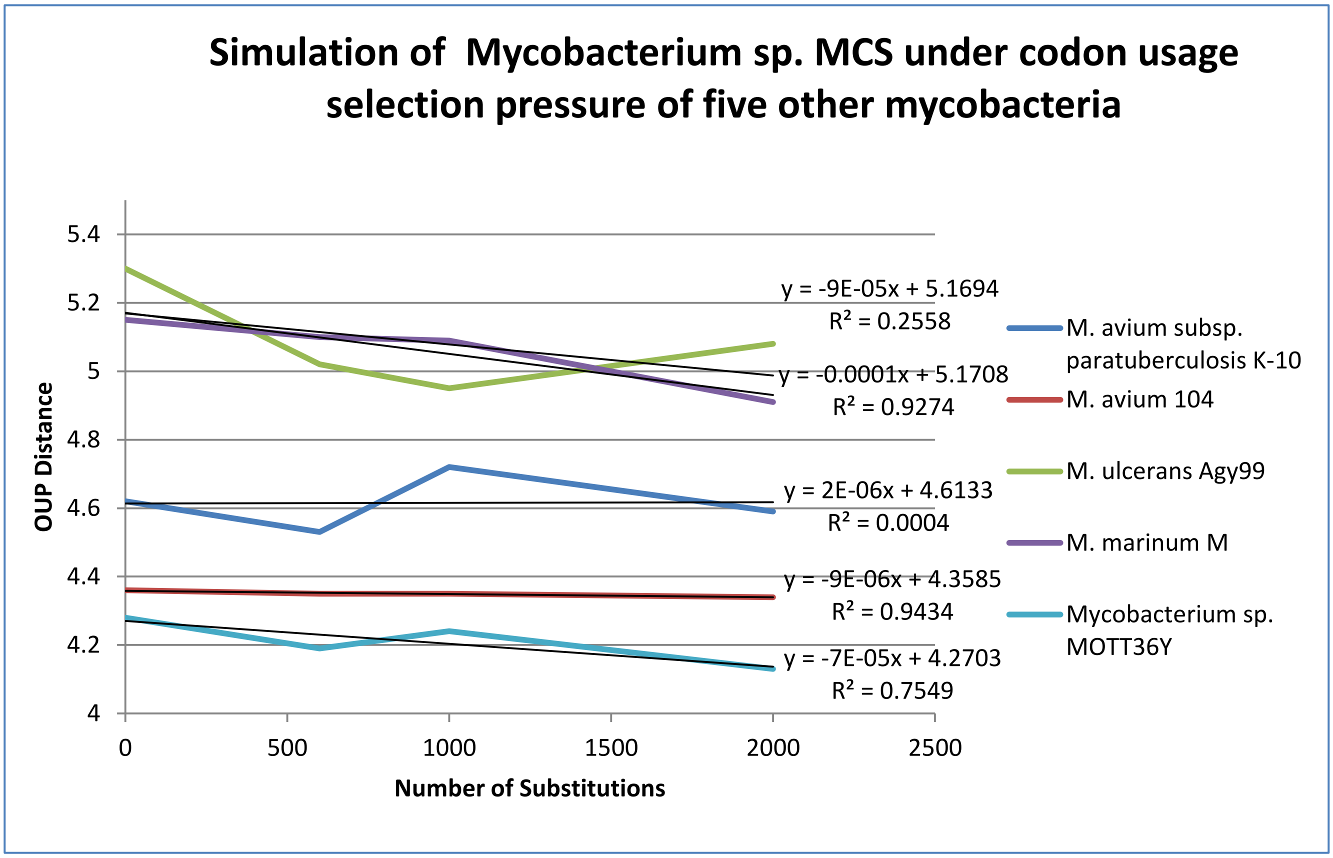


Fig. 2.5 Plotting of OUP distances against numbers of nucleotide substitutions in the tester sequence of *Mycobacterium* sp. MCS evolving towards codon usage of the target sequences of *M. avium* subsp*. paratuberculosis* K-10, *M. avium* 104, *M.* ulcerans Agy99, *M. marinum* M and *Mycobacterium* sp. MOTT36Y.

At the current stage, this simulation algorithm still needs several improvements to be used for modelling the speciation and evolution of bacterial genomic OUP. Firstly, since codon usage was calculated from coding regions only, non-coding regions bearing similar OUP parameters that limit the modelling of the whole genome OUP diversification were not considered. This limitation will be resolved by combining the codon driven and DNA conformation driven models considered above. A hypothesis is that the OUP diversification is initiated by the need for codon usage adaptation. This is followed by an adjustment of the DNA replication/reparation machineries, which recognise genome-specific DNA conformation parameters specifically for the most abundant oligonucleotides. Secondly, this model considers only synonymous substitutions in codons. Since non-synonymous mutations are prohibited, the sequence undergoing substitutions will never reflect the composition of the target sequence. This is evident in Figure 2.5 where OUP distance is decreasing at a low rate, reflecting small changes in composition with a large number of substitutions. To improve the model by allowing non-synonymous mutations, amino acid substitutions can be set on a specific ruleset in combination with ML estimating the most probable substitution pathway, e.g. BLOSUM, a likelihood score incorporating biological constraints in amino acid substitutions (Henikoff and Henikoff, 1992). Lastly, the current tester-to-target model may not be useful in the sense that OUP diversification is most probably not driven by any target but occurs as a random process similar to the Brownian motion. A target-driven model may significantly underestimate the number of substitutions required for OUP diversification. Another shortfall of this method, which ties to the previous constraint problem, is that the random substitutions are generated by computer randomisers, which in fact generate pseudo-random numbers. This limitations leads to losing vital biological information and is biased towards the way nucleotide substitutions are calculated.

The two simulated OUP evolutionary models discussed above, although imperfect, have demonstrated that OUP distances are whole genome signatures, which may have evolutionary implications. In the first model, the estimation of the time of insertion of a mobile genetic element was brought forward using a simulated amelioration process based on the comparison of OUP distances between tester and target sequences. In the second model, the simulated evolution of one sequence to another aimed at the acquisition of a target codon usage showed the correlation between numbers of substitutions and OUP distances. Although both methods are currently not practical in terms of estimating true biological distances between sequences because of their limitations, the theoretical aspect of OUP distance reflecting evolutionary implication was demonstrated. Both simulated models showed correlations between selection forces from context nucleotides and/or codon usage to OUP distances, which was also evident in literature (Fedorov *et al*., 2002; Pride *et al*., 2003). With such correlation, it can be concluded that OUP distances can have an application in phylogenomic inferences.

2.4) Discussion

Therefore, in this project, we use and validate a matrix-based OUP approach for phylogenomic inferencing because of its simplicity and freedom from any evolutionary hypothesis. It can be concluded from our case studies that OUP evolution in bacteria is mostly driven by codon selection and codon-dependent context nucleotides. This has also been reflected in literature displaying evolutionary implications with possible applications in phylogenetics (Pride *et al*., 2003). Driving forces could be deduced by biased codon usage reflecting unequal concentrations of tRNA molecules in the cytoplasm of bacterial cells (Shah and Gilchrist, 2010). Indeed, the abundance of different tRNAs depends on the number of allelic copies of the corresponding genes and unequal gene expression efficacy from different loci (Elf *et al*., 2003). Fluctuations of tRNA concentrations in bacterial species can engender a steady rate of directed mutations adjusting the codon usage, thus influencing the global OUP pattern. This hypothesis is consistent with previous publications (Marquez *et al*., 2005; Bofkin and Goldman, 2007).

It was shown that exclusively the neighbour residues affiliated with the same codon had influenced the patterns of nucleotide substitution emissions (Figure 2.1). Therefore, OUP being driven based on codon bias and adaptation is a valid phylogenetic signature that is feasible in explaining the phylogenetic relationship between sequences. However, codon adaptation as a single driving force cannot explain the fact that non-coding intragenic regions of bacterial genomes conform to the same OUP characteristic of the whole genome. This was also evident in the result where context-dependent and codon usage models could not fully explain the evolutionary diversification of OUP, as was discussed above in this chapter. In the paper by Reva and Tummler (2004), it was shown that coding and non-coding regions of bacterial genomes share the same abundant oligonucleotides characterised by similar stereo-chemical properties. It was hypothesised that the bacterial DNA reparation system could allow more mutations in the DNA fragments with alternative OUP by recognising an alternative conformation of these DNA loci. However, this driving force of OUP diversification is probably weaker than codon usage adaptation and requires a longer period of evolution. This assumption is supported by the fact that the horizontally transferred genomic islands comprising important protein coding genes rapidly gain the host-specific OUP (Sueoka, 1988; Lawrence and Ochman, 1997). Insertions of prophages comprising non-coding sequences and selfish genes may be identified by their specific OUP even in several related bacterial species. This implies that these were inherited from one common ancestor a long time ago without losing the OUP specificity of these loci (Pierneef *et al*., 2015). This might also be taxon-specific, as certain taxonomic groups were shown to have large noise regions, as seen in Figure 2.2.

Comparison to literature and the case studies demonstrated here showed that OUP contains reliable phylogenetic signatures for phylogenomic inferencing. The advantage of this method is that OUP can easily be calculated for non-annotated DNA sequences of complete bacterial chromosomes and/or large genomic fragments. Theoretically it was predicted that even a 5 kbp DNA sequence could be sufficient for a statistically reliable OUP estimation (Reva and Tummler, 2004). However, to avoid the influence of horizontally transferred genomic islands and other genomic loci with alternative OUP, it is recommended that the genomic fragments subjected for phylogenomic inferences should be 50 kbp or longer. In the next chapter, this assumption will be checked experimentally. With this conclusion, we can relate OUP to phylogenetic inferences based on its evolutionary implications and taxonomic binning abilities. In the next chapter, we take a look at the implementation of the OUP-based phylogenetic approach and how well it compares to other traditional phylogenomic methods currently in use, such as marker gene sequence comparison and whole genome alignments.

Chapter 3) Creation and Comparison of OUP-based Tree to Common Phylogenetic Inferences

3.1) Methods Used for OUP Calculation and Comparison

The concept of OUP has been defined in previous chapters (Chapter 1.3, Chapter 2.1) and will be used as a core metric for OUP-based phylogenomic comparisons. OUP metrics will be calculated according to the methods stated by Reva and Tummler (2004). Briefly: K-mers (tetranucleotide in this work) were ordered by descending frequencies of occurrence in the genome and then ranked as seen in the example in Figure 1.5. The patterns of oligonucleotides of paired sequences were compared by using equation 1,

$$D_{ij} = 100 \times \frac{\sum_{w}^{4^k} \left| rank_{w,i} - rank_{w,j} \right|}{4^k \times (4^k - 1)/2} \qquad [\mathbf{1}]$$

where D_{ij} is the distance between the patterns *i* and *j*; *k* is the length of the K-mer, i.e *k* = 4 in this current work; and $rank_{w,i/j}$ are integer rank numbers of the word (K-mer) *w* in the patterns *i* and *j*. This D metric represents the similarity measure between the two sequence patterns.

Another oligonucleotide usage statistical parameter, termed oligonucleotide usage variance (OUV), was used in this study to identify possible outlier genomes. It is defined and calculated by the use of equation 2:

$$\mathrm{OUV} = \frac{\sum_{w}^{4^k} \Delta_w^2}{(4^k - 1) \times \sqrt{0.02 + \frac{4^k}{L_{seq}}}} \qquad [\mathbf{2}]$$

where L_{seq} is the sequence length and Δ_w was calculated as the difference between the observed and expected frequencies of a word *w*. The expected frequency was calculated under an assumption of an equal distribution of words in the sequence. Random sequences with a high rate of mutations were characterised with lower OUV values (Reva and Tummler, 2004), i.e. the deviation is low because the sequence is characterised by the expected frequency of words across the genome implying a random sequence with no signature. This metric has been shown to reflect the stringency of selection of specific oligonucleotides in a genome. In this research, this metric was used to identify outliers in samples of genomes, of which phylogenetic relations may be falsely predicted. These genomes will be characterised by extreme difference (outside the 2.5 x STD range) in OUV compared to other sequences in the dataset under study.

OUP calculation and comparisons were implemented by a graphical user interface (GUI) program, MetaLingvo 1.0, written in Python 2.5, which is available for download from the project website (www.bi.up.ac.za/SeqWord/metalingvo/index.html). A command line version of the program named LingvoCom 1.0 is available at www.bi.up.ac.za/SeqWord/lingvocom/index.html. These websites provide users with detailed guidelines on the usage of these programs. The programs analyse input genome-scale DNA sequences and return PHYLIP format distance tables. The distance table will consist of OUP comparisons between genomes, which will then be processed by the PHYLIP package program *neighbour* for the construction of phylogenetic trees (Tuimala, 2006).

3.2) Selection of Taxonomic Groups for Case Study

Various groups of microorganisms were selected for this study to represent different bacterial provenances by taxonomically well-characterised species. Complete genome sequences of different taxonomic groups of microorganisms were obtained from GenBank (Chapter 2.2.1). In total, 11 species of the genus *Bacillus*; 13 species of the genus *Corynebacteria*; 11 species from different orders of Gamma-Proteobacteria; 22 species of the genus *Lactobacillus*; 16 species of the genus *Mycobacterium*; 12 species of the genus *Pseudomonas*; and nine archaeal species of genera *Thermotoga*/*Thermovibrio* were chosen (see Tables 2.1-2.8 in the previous chapter). These groups represent a vast variation of phylum such as Firmicutes, high GC Gram-positive bacteria of the genera *Corynebacterium* and *Mycobacterium*, several orders representing Gamma-Proteobacteria such as *Enterobacteriales*, *Alteromonadales*, *Xanthomonadales* and *Pseudomoadales*. Moreover, a group of 12 strains of *Prochlorococcus marinus* was used to study OUP evolution among closely related microorganisms belonging to the same species. Discrepancies between gene-based and genome-based phylogenetic trees were reported for the latter group of microorganisms in a previous publication (Prabha *et al*., 2014). An attempt was made in this work to resolve this discrepancy by using OUP approaches.

Identification of COG in each taxonomic group was performed by an in-house python pipeline running reciprocal local BLASTP alignments of all protein-coding genes of a genome against protein-coding genes of all other genomes in the same taxonomic group. This pipeline can also be referred to the EDGARs platform of the ortholog identification method as stated in Chapter 1.2.2 (Blom *et al*., 2009). Pairs of genes showing a reciprocal sequence similarity with a cutoff of e-values ≤ 0.0001 were considered orthologous. Whole genome data were extracted from Genbank files and were used for MAUVE tree inferences. Similarly, 16S rRNA and gyrase A gene sequence data were also extracted from the complete genome sequences of microorganisms obtained from the Genbank database. Sequences for each species used within the case study for the analysis using CVTree was taken directly from the CVTree database consisting of annotated proteome data on the CVTree website http://tlife.fudan.edu.cn/cvtree/cvtree3/ (Qi *et al*., 2004a).

3.3) Methods Used for the Construction of Phylogenetic Trees by Alignment-based and Alignment-free Approaches

All COGs were aligned using the MUSCLE algorithm (Edgar, 2004). Alignment ambiguities were removed by the program Gblocks (Castresana, 2000). Evolutionary distances between proteins were estimated based on the Jones-Taylor-Thornton (JTT) substitution model implemented in the program *protdist* (Jones *et al*., 1992). For alignments of 16S rRNA sequences, the Felsenstein F84 substitution model was implemented in the program *dnadist* of the PHYLIP package (Felsenstein, 1985; Tuimala, 2006). Phylogenetic inferences were performed based on the JTT/F84 distance tables by the NJ algorithm using the program *neighbor* of the PHYLIP package. The NJ algorithm was chosen for this study because of two important aspects. Firstly, NJ is computationally efficient, with high performance allowing analysis of big datasets. Secondly, this algorithm is universal, which allows easy comparison between different methods of phylogenetic inferences. The NJ algorithm can furthermore produce phylogenetic trees from any set of distances calculated using other methods such as MAUVE, whole genome supermatrix (WGS), OUP or 16S rRNA comparisons. Lastly, this method is simplistic and free from any evolutionary pre-assumptions, which are needed by other methods such as ML and minimal parsimony (MP) algorithms.

NJ trees were inferred for every COG as well as for alignments of 16S rRNA. WGS trees were inferred based on concatenated alignments of all COG translated into protein sequences (excluding 16S rRNA). Phylogenomic trees based on whole genome sequence alignment were inferred by the program

MAUVE 10 . Lastly, two types of alignment-free trees were calculated by using whole genome sequence data. OUP comparison was performed by using the program LingvoCom 1.0 (http://www.bi.up.ac.za/SeqWord/lingvocom/index.html) with NJ trees constructed using the program *neighbor* from the PHYLIP package. CVTree was an online program based on the comparison of whole genome proteomes by means of calculation of genome-scale oligo-protein k-string vectors to estimate phylogenomic distances between microorganisms (Xu and Hao, 2009).

Topologies of phylogenetic trees were compared by using the symmetric and branch score distance (BSD) algorithms implemented in the program *treedist* of the PHYLIP package (Kuhner and Felsenstein, 1994). The symmetric algorithm compares the topologies of trees only, while the BSD algorithm also accounts for the branch length differences (Tuimala, 2006).

3.4) Evaluation of the OUP Based Algorithm by Comparison of Resulting Trees

3.4.1) Comparison of OUP Inferences to Other Genome-based and Gene-based Phylogenetic Trees

Phylogenetic trees based on alignments of individual COG and on alignment-free methods including OUP comparison were compared to both the WGS and *gyrA*-based trees using the PHYLIP *treedist* algorithm to identify the level of congruence between their tree topologies (Tuimala, 2006). The symmetric algorithm of *treedist* calculates the distance between tree topologies by counting the number of rearrangements between clades defined in different tree topologies. Relocation of one end-node element between clades in compared trees will give a distance of two. Distributions of symmetric distances calculated for gene trees (MAUVE and 16S rRNA) and alignment-free trees (OUP and

CVTree) compared to the WGS and *gyrA*-based trees used as references are shown in Figure 3.1.

Remarkably, in almost all taxonomic groups, the OUP tree topologies were identical and/or very similar to those of WGS trees. However, the topologies of the gene trees of individual COGs were generally dissimilar to those of the WGS trees and to each other (Figure 3.1 and Figure 3.2). For example, the trees based on GyrA protein alignments, which are generally recognised as phylogenetic markers (Huang, 1996), shared topological similarities with only a few gene-based trees and were dissimilar to the WGS and OUP trees (Figure 3.1). For a better visualisation of relationships between the phylogenetic trees created using the various methods including 16S rRNA alignments, WGS, OUP, MAUVE, GyrA protein sequence alignments and CVTree, dendrograms based on tree topology distances are shown in Figure 3.2.

These dendrograms were inferred using the NJ algorithm based on matrices of symmetric distances calculated between phylogenetic trees using the program *treedist* (Tuimala, 2006). The OUP trees were usually the most congruent with the respective WGS trees except for the groups of Lactobacilli and Prochlorococcus. The trees based on alignments of marker genes/proteins were often grouped together with the CVTree cladograms, while the grouping of the MAUVE trees was rather controversial. GyrA and 16S rRNA trees were generally dissimilar to the WGS trees, although these genes were considered a universal phylogenetic marker for a long time (Janda and Abbott, 2007; Shapiro *et al*., 2016).

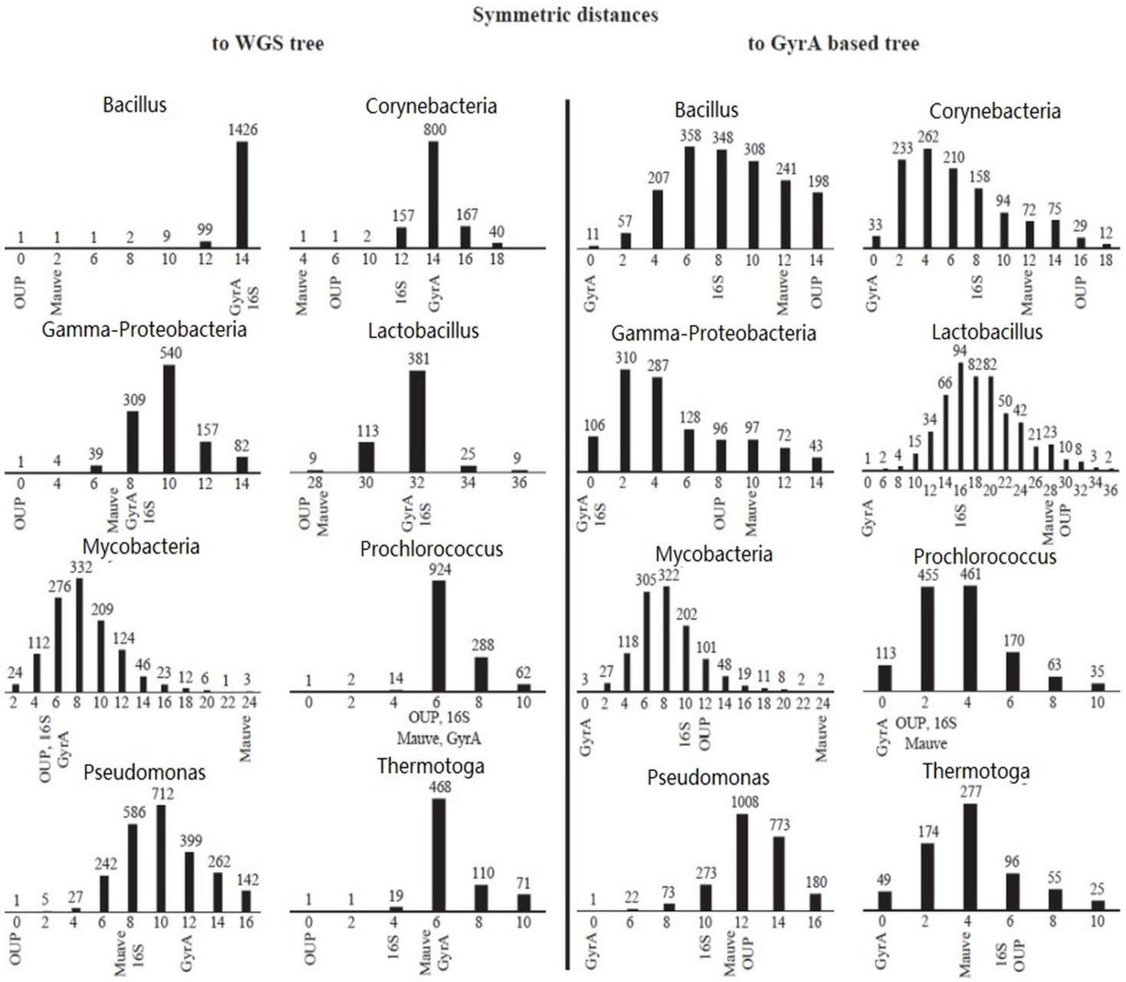
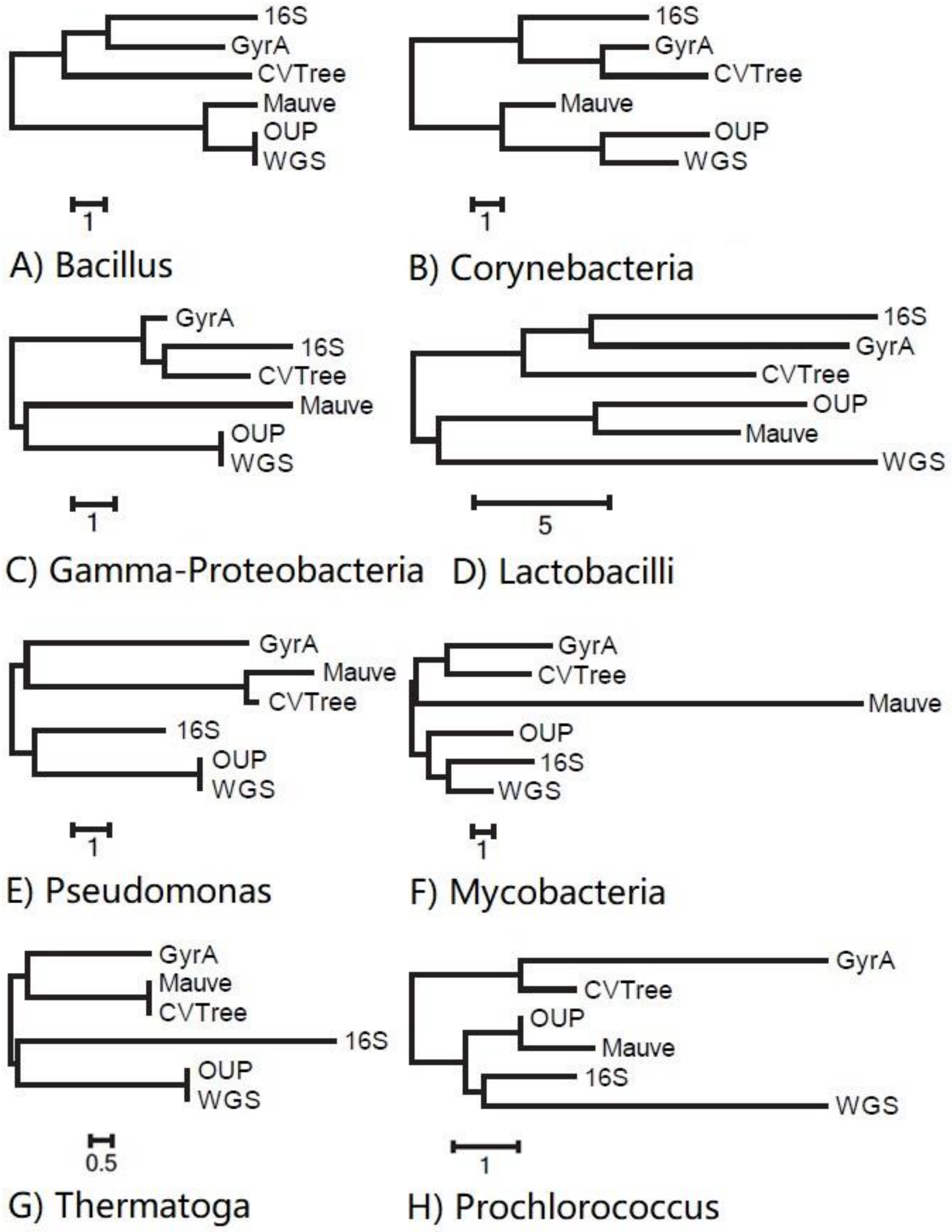


Fig. 3.1 Distribution of symmetric distances between pangenome trees and reference trees: WGS (left) and GyrA (right). Trees are grouped by their symmetrical distance categories. The numbers of trees of each category are shown above the corresponding columns. Distance categories containing OUP, MAUVE, GyrA and 16S rRNA-based trees are marked accordingly.



A) Bacillus

B) Corynebacteria

C) Gamma-Proteobacteria

D) Lactobacilli

E) Pseudomonas

F) Mycobacteria

G) Thermatoga

H) Prochlorococcus

Fig. 3.2 Topological similarity between the trees calculated for the selected taxonomic groups by different algorithms: GyrA protein distances, 16S rRNA distances (depicted as 16S), OUP distances, whole genome sequence alignment distances (WGS), MAUVE and CVTree. Dendrograms were designed by an NJ algorithm based on the matrices of distances between the trees calculated by the *treedist* symmetric approach.

In addition to tree topology, other important metrics of comparison of phylogenetic inferences are the lengths of branches, which reflect the amount of dissimilarity between taxa. The BSD algorithm accounts for the branch length differences between phylogenetic trees calculated by different methods. All the phylogenetic methods mentioned above were ordered ascendingly by BSD values. Rank orders of the trees based on the alignments of 16S rRNA, GyrA proteins and MAUVE, and based on the alignment-free OUP approach, which were ordered by their similarity to the corresponding WGS trees, are shown in Table 3.1. A higher rank (low value with least branch length difference) represents the highest similarity to the reference WGS tree. The program CVTree was excluded from this comparison, as it produces distances known not to be comparable to the distance values produced by other methods. Again, it became evident that branch lengths calculated by the OUP approach in many cases were congruent with those produced by the WGS algorithm (Table 3.1).

Table 3.1 Ranks of congruence of several gene-based and alignment-free-based trees with the WGS reference tree calculated for different groups of microorganisms

Taxonomic group	Number of genes in pangenome	Ranks of congruence with the WGS trees			
		16S rRNA	gyrA	OUP	Mauve
Bacillus	1820	1810	1551	1	489
Corynebacteria	1182	1165	1053	1182	33
Gamma-Proteobacteria	1144	1115	935	1	773
Lactobacillus	540	533	420	2	26
Mycobacteria	1168	775	1115	1009	1168
Prochlorococcus	1311	1285	305	720	802
Pseudomonas	2418	2248	1775	2416	7
Thermotoga	683	682	469	1	564

From the above two metric comparisons, several important conclusions can be made from the results in Figure 3.1, Figure 3.2 and Table 3.1. First, the trees inferred from the alignments of the traditional phylogenetic marker sequences, the 16S rRNA nucleotide and GyrA protein sequences, were least similar to the WGS trees by topology and by lengths of branches in most of the taxonomic groups in this case study. On the contrary, both tree comparison algorithms showed that the OUP trees were most similar to WGS trees calculated for Bacillus, Gamma-Proteobacteria and Thermotoga, and the best but one for Lactobacillus. However, this was not always the case. OUP trees inferred for Corynebacteria, Mycobacteria and Pseudomonas were similar to the corresponding WGS tree by the topology but dissimilar by branch length. The results were ambiguous for subspecies of Prochlorococcus and were investigated further in another case study. The trees constructed by MAUVE, based on whole genome alignments, also shared similarity with the corresponding WGS trees. However, the OUP approach usually outperformed the MAUVE trees in this regard (Figure 3.1, Figure 3.2 and Table 3.1). It may be concluded that OUP comparison is a promising approach for phylogenomics, as this procedure produces trees congruent to WGS trees but is more efficient in terms of computational power and run time compared to the latter. This is also true for higher taxonomic levels as for the group Gamma-Proteobacteria performed well for both symmetrical and branch score distance. To visualise differences and similarities of tree topologies, several examples of WGS, OUP and GyrA-based trees inferred for the taxonomic groups of Bacillus and Corynebacteria are shown in Figure 3.3.

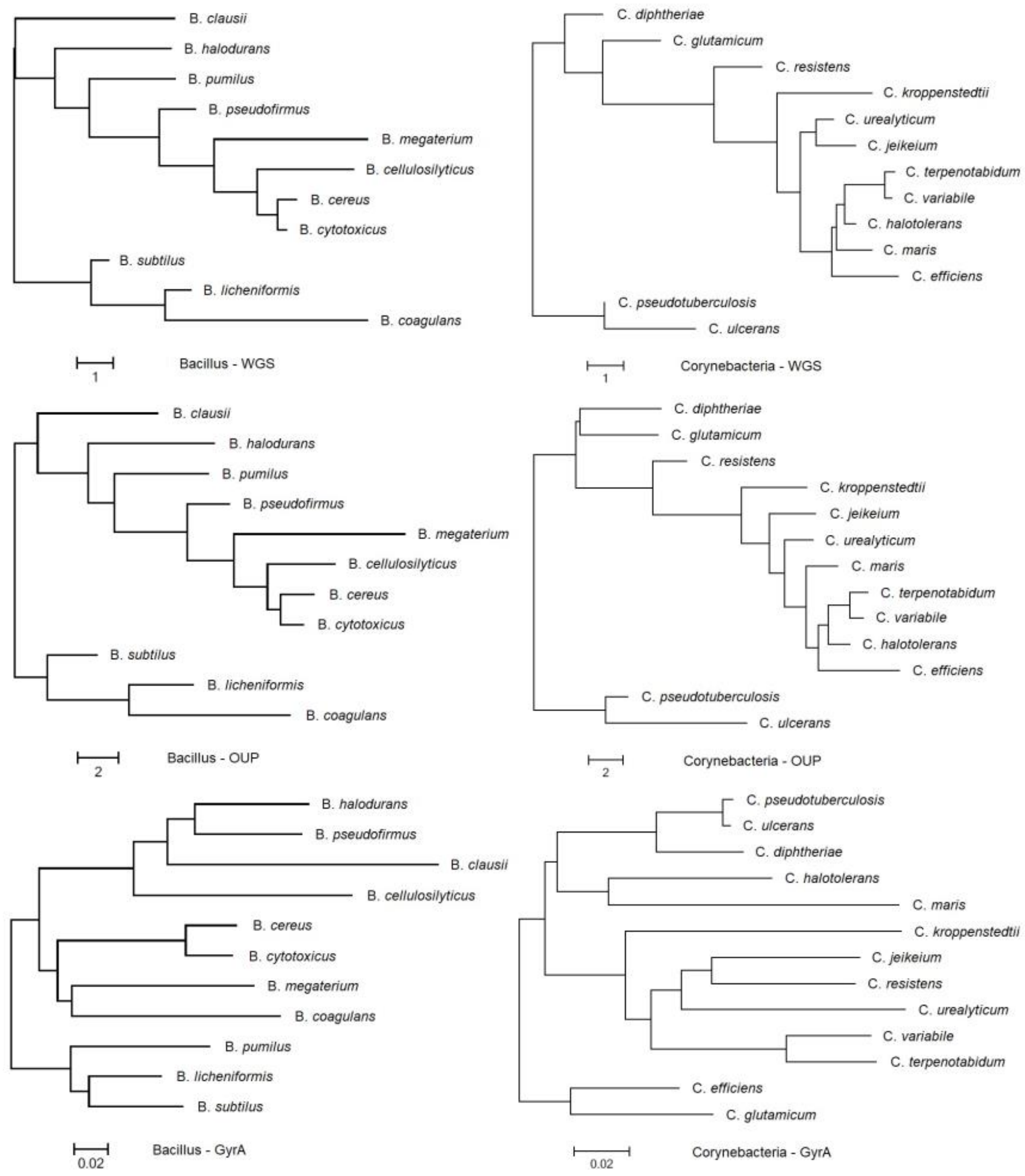


Fig. 3.3 WGS-, OUP- and GyrA-based trees inferred for the taxonomic groups Bacillus (left) and Corynebacterium (right). From both taxonomic groups, one can clearly see from the resulting tree that OUP trees are almost identical to WGS trees, while the GyrA-based trees differ from both OUP- and WGS-based trees.

From the above comparison of results, it can be seen that based on different phylogenomic methods, incongruence of results is evident for the same genomic datasets. The advance in sequencing technologies has created a new paradigm of evolutionary reconstructions based on complete genome sequence data (Richter and Rosselló-Móra, 2009; Blaimer *et al*., 2015). Several case studies were designed for this research to assess the reliability of different phylogenetic and phylogenomic approaches. However, it has to be admitted that the lack of experimentally proven models of species evolution does not allow the performance of any formal statistical validation or benchmarking of available phylogenetic approaches. An indirect indication giving extra credits towards the genome-based approaches is that in four out of eight inferences shown in Figure 3.2, the WGS and OUP trees shared identical topologies with the highest similarity in branch lengths. All other trees were algorithm-specific except for one case of congruence between the Mauve and CVTree trees calculated for the Thermatoga group.

3.4.2) Resolving Phylogenetic Relations between Prochlorococcus Strains by OUP Approach

OUP-based trees were often found to be congruent with those of the corresponding WGS trees, with some topological differences being observed (Figure 3.1). These topological differences were shown in the dataset between OUP and WGS trees calculated for the group Prochlorococcus. We hypothesised that these misalignments between OUP and WGS trees may result from errors in the multistep procedure of WGS inferences, while the OUP approach is more straightforward in comparison. These multistep procedures include genome annotation, identification of orthologous genes, multiple sequence alignment and concatenation. To validate this hypothesis, the OUP tree calculated for the group Prochlorococcus was compared to published phylogenetic trees calculated for the same organisms by the whole genome alignment algorithm, comparison of the 16S rRNA gene (Prabha *et al*., 2014) and pangenomic definition analysis (Moldovan and Gelfand, 2018).

The authors of the first publication claimed that the superstring whole genome comparison fitted the known phylogeny between ecotypes of this species better than that based on 16S rRNA sequence comparison (Figure 3.4). These results were congruent with the above analysis of topological differences between different methods for the taxonomic group Prochlorococcus (Figure 3.1). Relations between the high light- (HL) adapted and low light- (LL) adapted ecomorphs of this species were considered, which were well segregated by the WGS approach but were mixed when compared by 16S rRNA. The results from the second publication using pangenomic definition of prokaryotes also shared this delineation between two ecotypes for this species. However, for each ecotype, more in depth subgroups was created for better definition of each cluster of strains.

The OUP tree designed in this study was consistent with the delineation of the HL and LL ecotypes of *P. marinus* and it distinguished properly between the ecotypes. The resulting OUP comparison fits very well to the tree topology produced by the WGS method (Figure 3.4) as well as to the pangenomic definition method, especially, the subgrouping of the low light ecotype (Figure 3.5). The OUP-based method made it possible to distinguish correctly between divergent sequences (MIT9303 and MIT9313 strains) through the use of the OUV metric that indicated these genomes as outliers in the given group (Equation 2). These outliers were marked and noted by asterisks in the resulting tree. Outlier detection is explained in more detail in the next chapter. Although 16S rRNA did not perform well in distinguishing between the two ecomorph-adapted strains, this method did perform well in the delineation of the tightly clustered HL strains. Although OUP performed better in general, this method did not perform well for this group (HL-clustered strain). This shows that the properties of phylogenetic marker genes such as 16S rRNA have the advantage of well-established phylogenetic relationships of closely related organisms. One can also conclude that OUP-based phylogenetics is more suitable for distantly related organisms and that the diversification of OUP is more constant over time, which is suitable for phylogenetic studies, i.e. the overall good inferencing of the phylogenetic relationship of the OUP-based tree compared to WGS methods, while the 16S rRNA method only performed well for closely related strains.

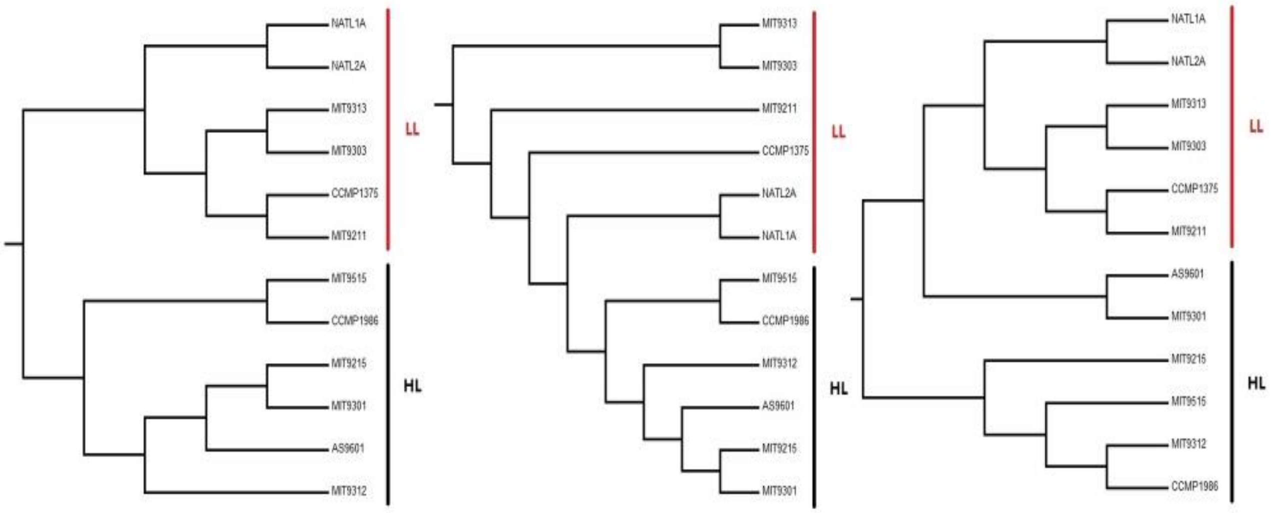


Fig. 3.4 Comparison between phylogenetic trees using whole genome gene overlapping (left), 16S rRNA (centre) and OUP (right) inferencing methods on the *Prochlorococcus marinus* subspecies dataset. The first two inferences were produced by the study by Prabha *et al.* (2014). The OUP method distinguish the phylogenetic relationship of this dataset more clearly than 16S rRNA with regard to the different light-adapted strains (LL: Low light, HL: High light) compared to the whole genome gene overlapping method.

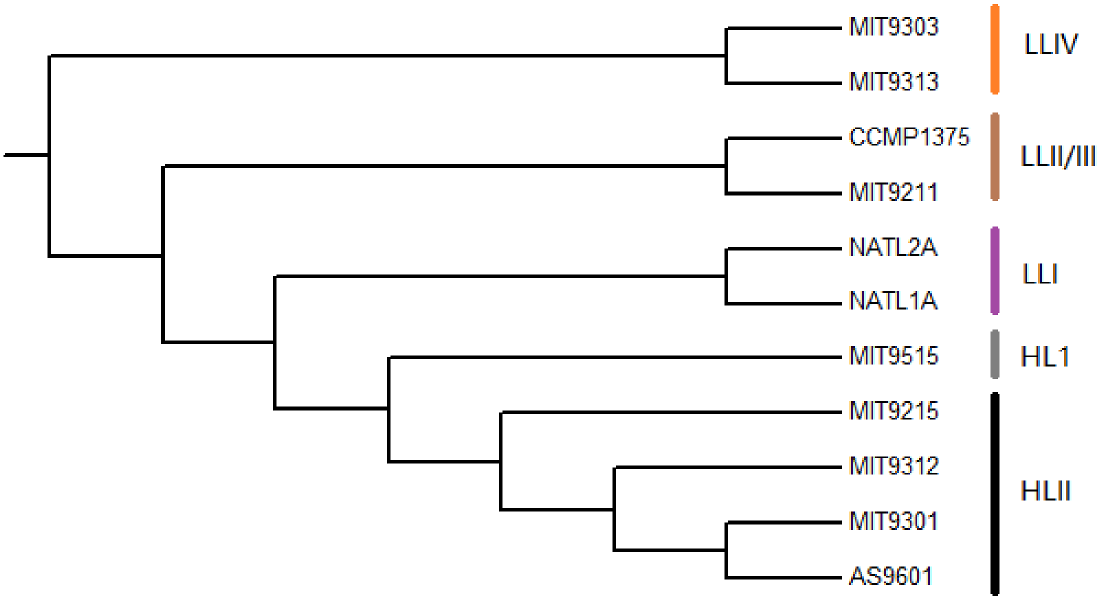
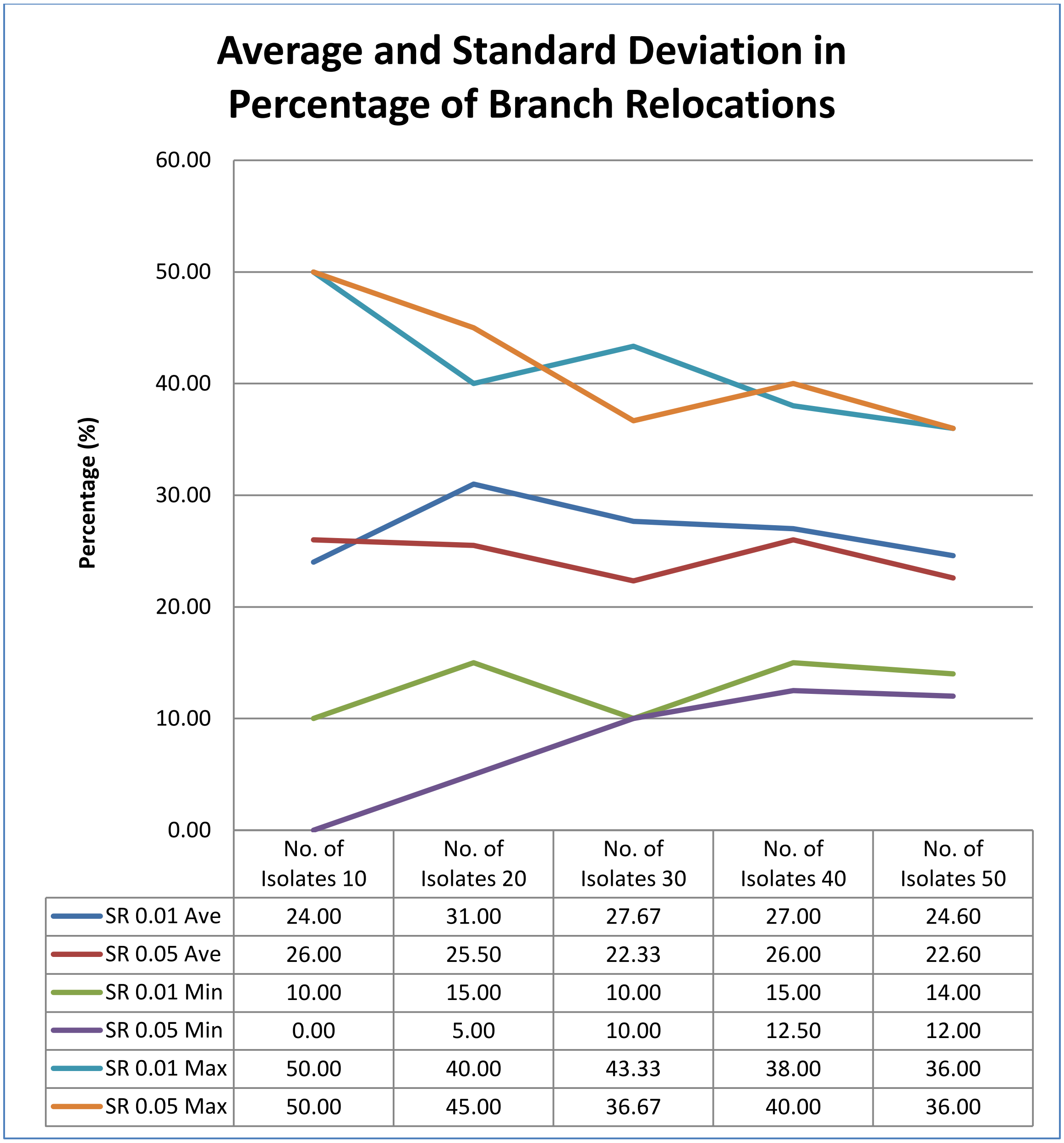


Fig. 3.5 Phylogenomic tree constructed using pangenomic definition method on the dataset Prochlorococcus. Eleven strains that was used within this research was extracted from the resulting tree inferenced by the study done by Moldovan and Gelfand (Moldovan and Gelfand, 2018). The strain CCMP1986 was not available in the resulting tree. Five subgroups branching off from two ecotypes are clearly shown in five colours.

3.4.3) Testing of the OUP Approach on Artificial Sequences Simulating Speciation Events

To ascertain the accuracy and consistency of OUP-based methods proposed in this research, we used the program SimBac to simulate sets of artificial DNA sequences of 1 Mbp in length as a case study (Brown *et al*., 2016). These sequences were generated based on a pre-assigned phylogenetic tree, therefore comparison with known phylogeny allows a measure of accuracy. In total 10 sets of sequences were generated with sample sizes of 10, 20, 30, 40 and 50 sequences, as well as different substitution rates, 0.01 and 0.05, of which every set was repeated 10 times. Generation of multiple datasets was done to assess the consistency of the program under a different number of sequences and different substitution rates. The OUP algorithm was used to construct phylogenetic trees based on the generated sets of sequences. The OUP trees were then compared to the reference trees produced by the SimBac program by using the *treedist* symmetric algorithm. The symmetrical distance was divided by two to calculate the number of branch relocations between compared trees. Average (ave), minimum (min) and maximum (max) percentages of taxonomic units producing topological mismatches were normalised by the sample sizes as displayed in Figure 3.6. The normalisation is calculated based on the total number of branch relocations over the total number of samples in a single simulation run (shown as isolates number in Figure 3.6) Normalisation allows easier comparison between dataset sizes in terms of accuracy and consistency.

The average number of relocations of taxonomic units was around 27% (standard deviation = 9.6%) for the substitution rate 0.01 and it was equal to 24.5% (standard deviation = 11%) for the substitution rate 0.05. The average and maximum percentages of relocations for both substitution rates decrease with an increase in the sample size. The decrease in the margin between maximum and minimum values as sample size increases shows an increasing consistency and decreasing variance of false branch relocations. The dataset with a substitution rate of 0.05 displayed better accuracy in terms of lower average percentages of false branch relocations at 95% confidence interval (P-value of 0.1307 rejecting null hypothesis of equal mean between two groups supporting lower average on 0.05 substitution rate dataset).



	No. of Isolates 10	No. of Isolates 20	No. of Isolates 30	No. of Isolates 40	No. of Isolates 50
SR 0.01 Ave	24.00	31.00	27.67	27.00	24.60
SR 0.05 Ave	26.00	25.50	22.33	26.00	22.60
SR 0.01 Min	10.00	15.00	10.00	15.00	14.00
SR 0.05 Min	0.00	5.00	10.00	12.50	12.00
SR 0.01 Max	50.00	40.00	43.33	38.00	36.00
SR 0.05 Max	50.00	45.00	36.67	40.00	36.00

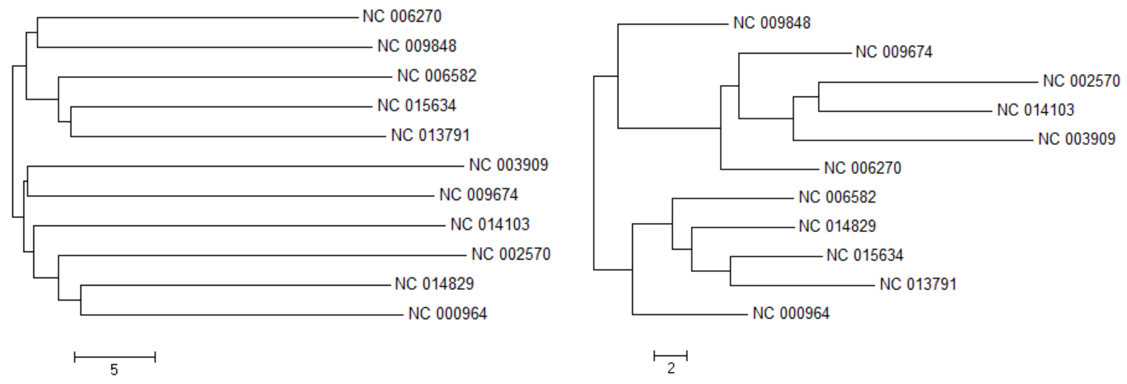
Fig. 3.6 The average, minimum and maximum percentage of branch relocations based on the comparison between the OUP tree and the reference tree in different simulation datasets with different sets of initial program parameters. The graph shows that an increase in the number of simulated sequences (labelled as isolates) leads to an increase in the consistency of OUP inferences reflected by a decreasing margin between minimum and maximum values (lower variance). the accuracy of OUP inferences in this instance was shown to increase corresponding to an increase in substitution rates (SR).

These results showed that for certain combinations of parameters when creating simulation datasets, the OUP method can improve its accuracy while not losing its consistency in producing true phylogenetic relationships. Overall, relatively high levels of topological mismatches are present in small datasets (50% wrongly placed artificial sequences). This may result from the randomness of the created artificial sequences that are not compatible with OUP. OUP compares genome signatures with high emphasis on codon usage and adaptation (Chapter 2.2), while artificial sequences are built on a known substitution model and recombination for their phylogeny (Brown *et al*., 2016). Therefore, this may be an influential factor in determining the accuracy of the OUP-based algorithm with artificial sequences. However, with larger datasets with more sequences, one can observe that OUP-based algorithms are more consistent and accurate with less variation in the accuracy of the infer trees. Unfortunately, it was not possible to use these datasets to produce trees with other methods to compare their performance with that of the OUP algorithm, for example the 16S rRNA and CVTree, as these methods require annotation information.

3.4.4) Bootstrapping Test for the Consistency of OUP Approach Based on the Variation in Lengths

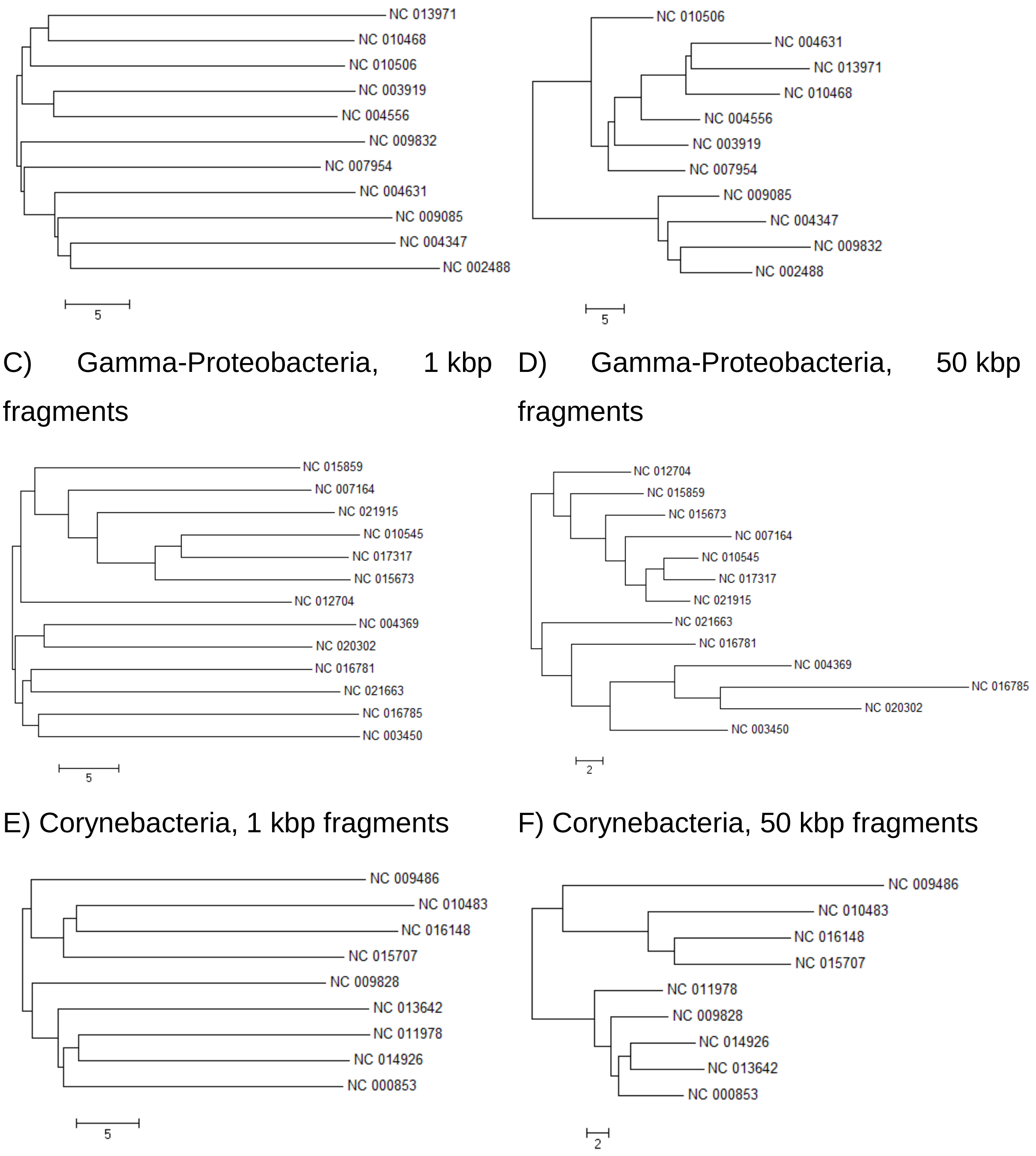
The bootstrapping test was designed to test the consistency of the OUP-based method as well as for establishing a cutoff determining how long the sequences need to be to obtain a consistent phylogenetic inference based on OUP. It is theoretically possible to calculate OUP based on a sequence length of 5 kbp. However, it was hypothesised that 50 kbp was a better length to infer phylogenetic relationships by reducing possible horizontal gene transfer bias within sequences (Chapter 2.4). Two bootstrap approaches were used. Firstly, 100 replicated genomic fragments of various lengths (1 kb, 5 kb, 10 kb and 50 kb) were selected in a random fashion from genome sequences of every group of microorganisms (Tables 2.1-2.8). OUP trees were inferred for every replicate and then consensus trees were created using the majority rule algorithm implemented in the program *consense* of the PHYLIP package (Tuimala, 2006). In addition, the sequencing error tolerance of the approach was tested by allowing random permutations in randomly selected genomic fragments. Genomic fragments of different lengths were tested as in the previous experiment. In total, 100 replicates of initial DNA sequences were generated by the permutation algorithm of the program *seqboot* of the PHYLIP package. Then, as in the previous procedure, OUP-based phylogenetic trees were calculated for every replicate and analysed by the program *consense*.

The OUP-based phylogenetic method showed significant robustness in terms of clustering of genomic fragments even when short sequences of 1 kbp were used. In all experiments, unambiguous tree topologies were created for all groups of microorganisms with bootstrap numbers of 100 assigned to each split in all trees. Shortening of genomic fragments influenced the ability of the program to estimate proper branch lengths. When OUP were calculated based on short 1 kbp sequences, the resulting trees were star-like, which indicates insufficient information to resolve differences between taxa. Longer genomic sequences allowed better cladding of the organisms. Examples of phylogenetic trees were inferred for the groups Bacillus, Gamma-Proteobacteria, Corynebacteria and Thermotoga to represent different phyla of microorganisms (Figure 3.7).



A) Bacillus, 1 kbp fragments

B) Bacillus, 50 kbp fragments



C) Gamma-Proteobacteria, 1 kbp fragments

D) Gamma-Proteobacteria, 50 kbp fragments

E) Corynebacteria, 1 kbp fragments

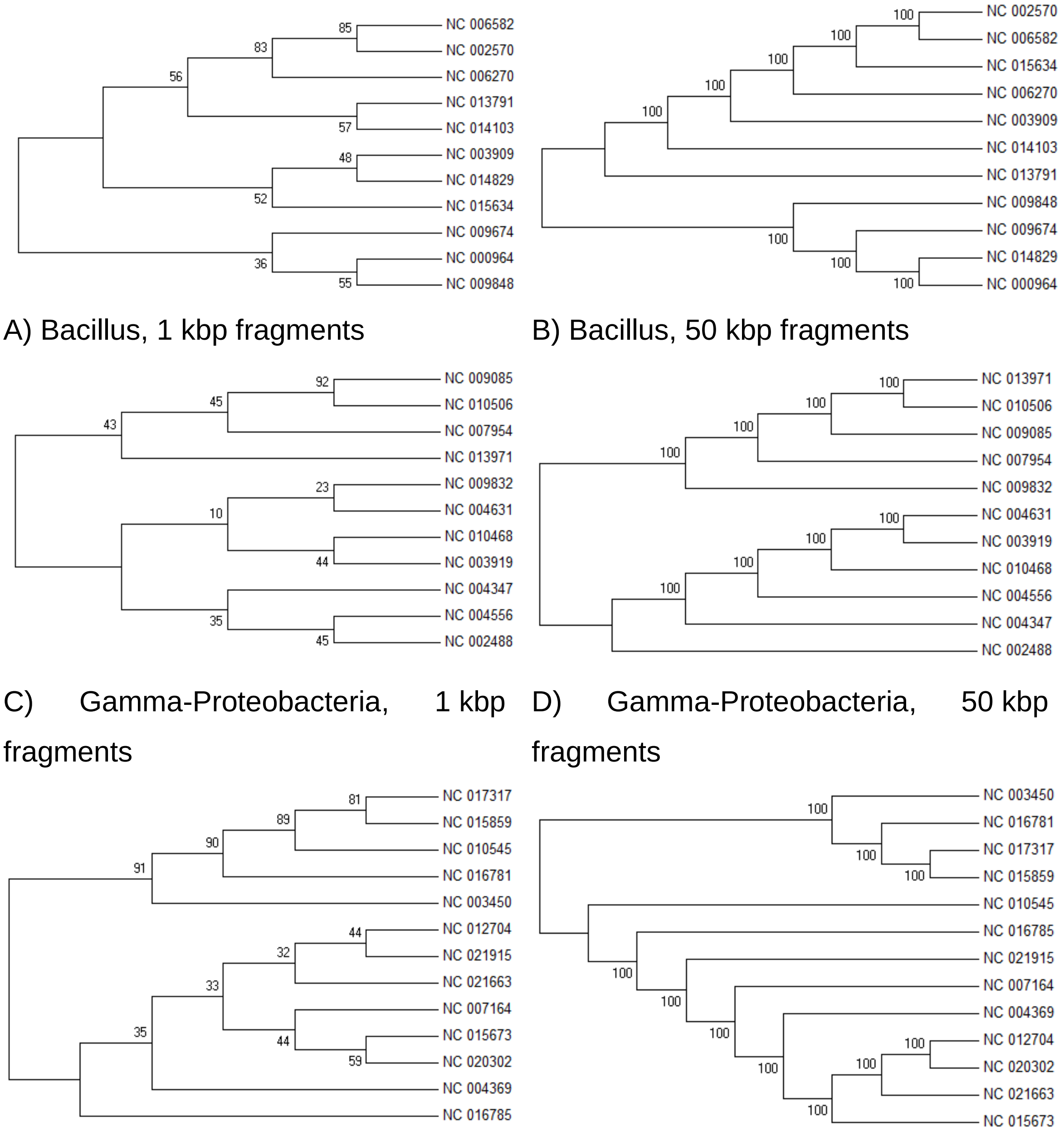
F) Corynebacteria, 50 kbp fragments

G) Thermotoga, 1 kbp fragments

H) Thermotoga, 50 kbp fragments

Fig. 3.7 Examples of NJ trees generated from randomly selected genomic fragments of different length for various taxonomic groups.

The regions selected and extracted from each dataset based on various lengths are randomly selected and then bootstrapped using a Python script (data not shown). The bootstrapping results of four out of eight taxonomic groups, *Bacillus*, Gamma-Proteobacteria, *Corynebacteria* and *Thermotoga* dataset for sequence length 1 kbp and 50 kbp, are shown in Figure 3.8. Summarised histograms of average bootstrapping values for all resulting bootstrap runs for all taxonomic groups with various sequence lengths are shown in Figure 3.9.



A) Bacillus, 1 kbp fragments

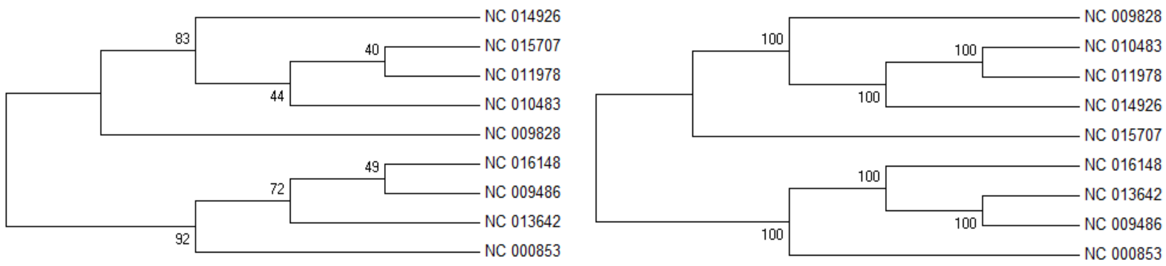
B) Bacillus, 50 kbp fragments

C) Gamma-Proteobacteria, 1 kbp fragments

D) Gamma-Proteobacteria, 50 kbp fragments

E) Corynebacteria, 1 kbp fragments

F) Corynebacteria, 50 kbp fragments



G) Thermotoga, 1 kbp fragments H) Thermotoga, 50 kbp fragments

Fig. 3.8 Examples of NJ trees generated from randomly selected and permutated genomic fragments of different length for various taxonomic groups.

It can be seen in Figure 3.9 that the Mycobacteria were more sensitive with regard to sequence length of genomic fragments compared to other taxonomic groups. Higher order groups of Gamma-Proteobacteria were not affected as much by shorter sequence length compared to closely related groups. However, when the length of fragments was 50 kbp, phylogenetic OUP-based inference was highly consistent with a bootstrapping value of over 90% in all taxonomic groups. To conclude, phylogenetic inferences based on OUP calculated for randomly generated genomic fragments showed strong robustness even when the fragments were rather short and random mutations were allowed. However, reliable results in terms of clustering of taxonomic units (tree topology) and accuracy of branch length estimation may be achieved when the length of genomic fragments is equal to or above 50 kbp.

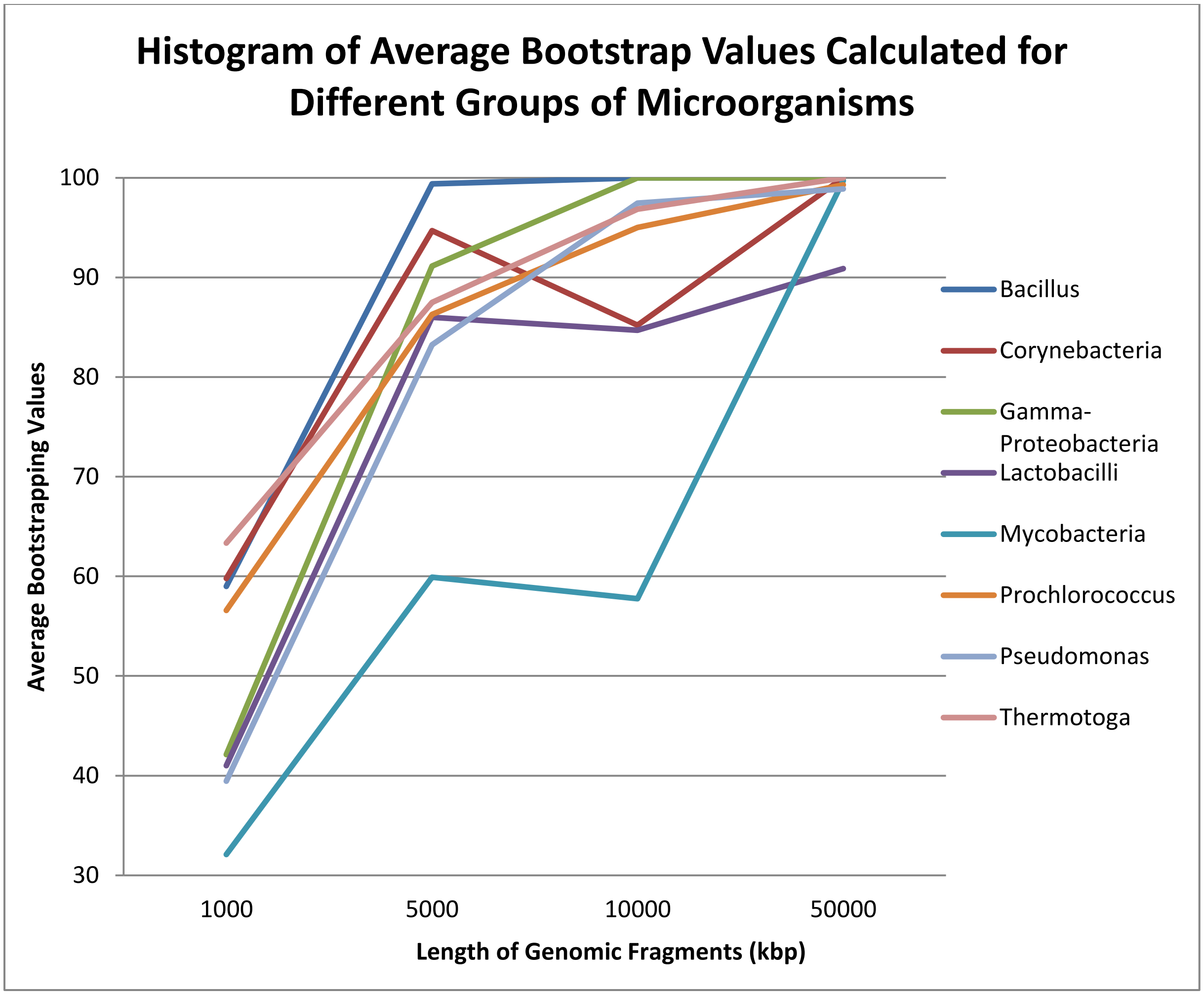


Fig. 3.9 Histograms of average bootstrap values calculated for different groups of microorganisms using 100 replicates of randomly selected genomic fragments of different length.

3.5) Reconciliation of Tree Topologies by Logistic Functions

The results from the previous sections have highlighted the discrepancies and incongruences between the resulting inferences based on different methods. This was especially true for the case study of the group *Prochlorococcus marinus* strains where the three methods used all gave different outputs when two clear strain ectomorphs were present. It is then of interest to look at possible methods of integration and reconciliation where combining the advantage of different methods could possibly increase resolution of phylogenomic inferences, especially on low taxonomic levels, e.g. phylogenetic marker genes work better for closely related organisms while the OUP approach yields better results for distantly related organisms.

To obtain insight into intrinsic relationships between the OUP- and protein GyrA-based trees, the distance metrics of both methods were analysed by a pairwise comparison of distance matrices. Non-linear dependence between corresponding distance values was observed that explains the incongruence between OUP- and GyrA-based trees. If there were linear dependence with a linear direct transformation model (one OUP distance converted to one corresponding GyrA distance), the resulting phylogenetic tree would be congruent, i.e. with a direct transformation, the GyrA distance between sequences would be the same or a scalar multiple of the OUP-based distances. An example of the pairwise distance plot for the taxonomic group Mycobacteria is shown in Figure 3.10. The set of OUP and GyrA distances for Mycobacteria may be found in supplementary Table 1.

In Figure 3.10, the two non-linear curves fitted onto the pairwise distance plot could be explained as follows: At the beginning of speciation (plot close to the origin), a higher rate of substitutions in protein sequences may be expected owing to positive selection. Accumulation of mutations then comes to a saturation point when the purifying selection takes over, allowing only sporadic neutral mutations in non-conserved regions (Shapiro *et al*., 2016). This change in the rate of accumulation of new mutations, which give change to phylogenetic distance over time, could lead to the non-linear dependence (curve) between the two methods. This concept is in agreement with the hypothesis of gene fixation in ecological niches proposed by Shapiro *et al.* as the Stable Ecotype Model (Shapiro and Polz, 2014).

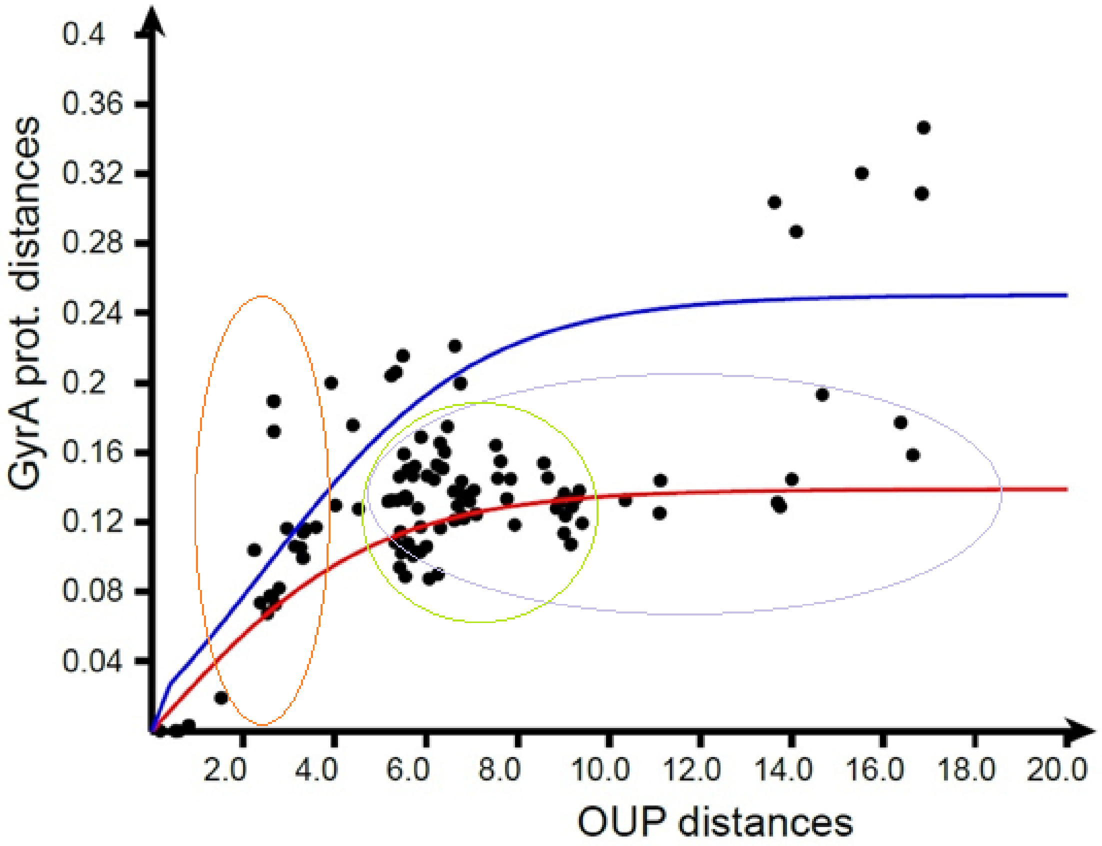


Fig. 3.10 Plotting of OUP distances (axis X) against GyrA sequence distances (axis Y) calculated for pairs of organisms of the taxonomic group Mycobacteria. Each pair of organisms on the plots is depicted by a dot. Distribution of dots is fitted to two logistic curves, which could potentially explain the incongruence between inferred trees. The orange region shows that the phylogenetic marker gene GyrA distinguishes closely related species better. The purple region shows that the OUP-based method distinguishes more diverse species better. The green region shows a large cluster of distance comparisons between sequences, which is difficult for a single method to determine.

Various different equations were fitted with the MatLab version R2015a (2015) in order to determine the type of non-linear dependence between the two distance methods for all taxonomic groups. The four types of equations that fitted well overall took the form of power function, exponential function, polynomial function and logistic function based on the R squared fitting criterion. The ranking based on the function performance is shown in Table 3.2. Overall, the polynomial function was on average the best fit to the data. However, in terms of parameters and function type, this model was least intuitive in terms of phylogenetic reconstructions (There is no convergence value and the distance can be infinite with the origin value being non-zero). The other three types of functions showed similar rankings, but the logistic function was chosen for the best fit to the evolutionary hypothesis of counter-play between the positive selection of beneficial mutation and the purifying counter-selection acting together against a background of a constant neutral mutation rate. This can also be seen in Figure 3.10, where the logistic functions are represented by the red and blue curves. This logistic dependence with its s-shaped curve is a good representation of the rate of accumulation of mutations in household proteins over the rate of changes in OUP patterns. For this equation [3], it can be assumed that the OUP diversification was constant over time, while the rate at which mutations are accumulated in a population may vary significantly, depending on the stage of the speciation process. The shape of the logistic curve also has certain implications for true biological assumptions, such as a common ancestor being the origin, gradient scalar being a mutation rate factor and capacity being the limit to speciation or change. If the capacity is reached and surpassed, a new fitting is done such that each curve fitting can be seen as a new cluster. These different clusters can relate to probable speciation leaps associated with changed habitats or the lifestyle strategy of a microorganism, which may also happen in reaction to massive HGT events.

Table 3.2 Rank of eight taxonomic group data fitting of different mathematical functions

Taxonomic Group	Polynomial Function	Exponential Function	Power Function	Logistics Function
Bacillus	1	2	4	3
Gamma-Proteobacteria	1	3	4	2
Mycobacteria	4	2	3	1
Pseudomonas	1	2	4	3
Lactobacillus	1	3	2	4
Prochlorococcus	2	4	1	3
Thermatoga	1	3	2	4
Corynebacteria	1	2	4	3
Average Rank	1.5	2.625	3	2.875

When in-depth analysis is done, the diversification of OUP and GyrA protein distances results from the difference in the rate of mutation accumulation at different stages of speciation. In the long run, it can be seen from the plot in Figure 3.10 that the diversification of OUP (x-axis) may better reflect the longer time span of evolutionary events. This is represented by a flat spread of distances estimated by GyrA comparison. Conversely, substitutions in GyrA sequences can reflect the early stages of speciation better (Figure 3.10). The fact that the genetic markers, such as GyrA, can distinguish closely related organisms better was also discussed in Chapter 1.2.1. In the present case, it may be explained as an acquisition of a few mutations, which were habitat-beneficial, possibly triggering further diversification of the population by an exponentially increasing rate of accumulation of secondary and compensatory amino acid substitutions in household proteins. However, as the number of possible beneficial mutations improving the adaptability of microorganisms at specific conditions is limited, the force of the purifying selection will very soon surpass the positive selection. This results in only neutral mutations, which will occur at a much lower rate. This study showed that the interplay of the purifying and positive selections can be simulated by an s-like logistic curve. It may furthermore be concluded from this that the common phylogenetic approaches based on the sequence comparison may result in an improper estimation of phylogenetic distances due to substantial differences in the substitution rates during the early stages of the evolutionary process. On the contrary, the rate of global OUP changes is constant in time. However, closely related organisms may be indistinguishable by their OUP. Hence, it is of interest to identify possible methods to integrate the advantages of both methods into one algorithm to distinguish phylogenetic relationships better. Therefore, further studies were aimed at integration and reconciliation of these two approaches of phylogenetics based on marker gene comparison and the OUP-based method to achieve better resolution between taxonomic clades.

As previously fitted, the inverse logistic function in the form of equation [3] showed the best fit to the distribution of the distance values calculated for GyrA protein alignments and OUP (Figure 3.10). The logistic equation is characterised by two parameters *K* and *g* and is well known for its uses in the field of population studies. Parameter *g* is often associated with the growth rate and *K* represents the capacity of the population. In terms of pairwise comparison of distance matrices, parameter *g* correlates to the rate of amino acid substitutions in the selected marker protein normalised by changes in OUP. Parameter *K* defines the boundary line of sequence substitutions limited by functional constraints of coding proteins and purifying selection.

$$\boldsymbol{OUP = f(GyrA) = \frac{-ln\left(\frac{2K}{gyrA + K} - 1\right)}{g}} \qquad [\mathbf{3}]$$

With the inverse logistic equation [3], one can convert GyrA distances into OUP distance or vice versa, which allows integration of OUP and GyrA evolutionary distances, as shown in equation [4]:

$$\boldsymbol{D} = \frac{\boldsymbol{OUP} + \boldsymbol{n} \times \boldsymbol{f}(\boldsymbol{gyrA})}{(\boldsymbol{n}+\mathbf{1})} \qquad [\mathbf{4}]$$

If *n* = 1, equation 4 returns an average value of the actual OUP distance and the estimated OUP distance calculated by the inverse logistic function in terms of GyrA protein comparison. As the value of *n* increases, greater weight is given to the protein distance values. This scalar value allows for better resolution between closely related organisms because of the increasing inference power of marker gene comparison. Fine-tuning of phylogenetic inferences by selecting an appropriate *n* value will be discussed in more detail in the next chapter.

The level of improvement of phylogenetic inferences by reconciliation of GyrA- and OUP-based trees was checked in a case study on the Prochlorococcus group of closely related organisms. Incongruence of phylogenetic trees developed for this group of microorganisms by different methods associated with poor alignment of the taxonomic grouping of these organisms with their biological peculiarities was also reported in other publications (Prabha *et al*., 2014). The authors of this publication claimed that the whole genome phylogenetic tree was much better fitted to known phylogenetic relations between ecotypes of this species, i.e. the tightly clustered HL-adapted and divergent LL-adapted strains, than the 16S rRNA based tree. In the present study, an integrated OUP+GyrA tree was very well aligned with the reported whole genome tree of *P. marinus* and it distinguished properly between the HL and LL ecotypes (Figure 3.11).

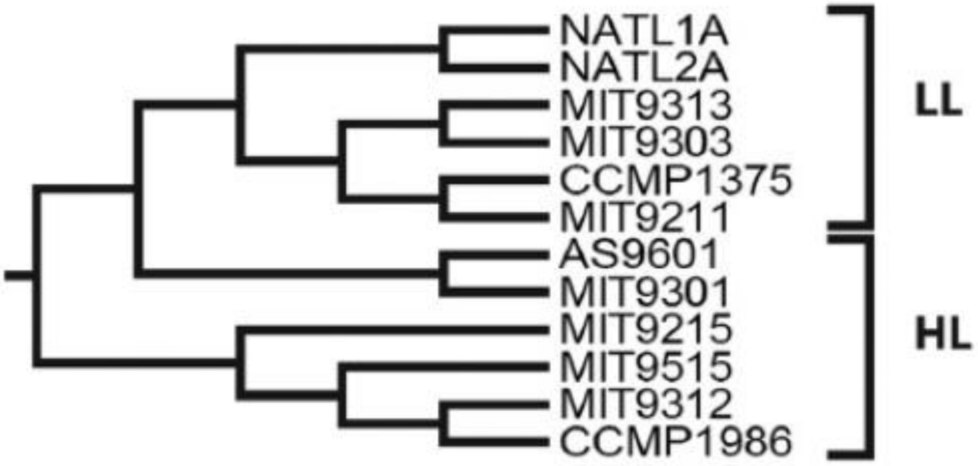


Fig. 3.11. An integrated OUP+GyrA phylogenetic tree constructed for the *Prochlorococcus marinus* subspecies data set. The inferred tree clearly distinguishes between different light-adapted ecotypes of this species (LL; HL) reported elsewhere (Prabha *et al*., 2014).

3.6) Discussion

Phylogenetic trees based on OUP comparison were generally more congruent with the corresponding WGS trees when compared with other methods (Figure 3.1, Figure 3.2 and Table 2). In cases when the congruence was ambiguous, a likely factor might be the numerous error-prone steps of the WGS phylogenomics, such as the genome annotation, orthology prediction and sequence alignment (Chan and Ragan, 2013). These types of systematic errors were also discussed in detail in the literature review in Chapter 1.2.2. Another factor one needs to note is the large branch length difference for some taxonomic groups, which might be ambiguous owing to possible branches being scalar multiples of each other, e.g. if the WGS tree has branch lengths 1, 3, 2, 4 and the OUP tree has branch lengths 2, 6, 4, 8 for four taxonomic units respectively, the branch length difference is 10 units. However, in reality, the tree is identical in terms of proportionality. Moreover, the discriminative power of the selected COG may vary in different groups of organisms assuming different prevalence of HGT events (Boto, 2010). This in turn influences the quality of WGS trees, which is evident in Figure 3.6 where the symmetrical distance of different COGs varies greatly compared to the WGS and *gyrA* gene reference tree. The case study of *Prochlorococcus marinus* subspecies also showed evidence of this. The tightly clustered HL-adapted strains were inferred more correctly with 16S rRNA gene data compared to divergent LL-adapted strains. HGTs will in many cases also not be a problem for OUP, as the DNA of genomic islands gain OUP features of the host chromosome in the amelioration process (Lawrence and Ochman, 1997).

Aside from the problem of HGT, comparison of approaches in general displayed the essential problem of phylogenetics, which consists of incongruence of trees produced by different approaches. OUP and WGS showed most congruence in their inferences calculated for different taxonomic groups. Therefore, OUP-based phylogenomic methods can be seen as a feasible method of choice for phylogenomic analysis, which performed on par with the WGS method. This was further evident in the comparison against literature for the Prochlorococcus dataset and artificially created sequences. As seen in Figure 3.2, overall incongruence between methods was evident, especially for OUP and *gyrA* distance-based trees, which showed the largest tree topology distances (Figure 3.6).

Integration of OUP inferences with those created by comparison of GyrA protein sequences was hypothesised to cover the shortfall of the OUP method of having less discriminating power when closely related organisms are studied. This limitation was evident in the case study of the Prochlorococcus group. In Figure 3.10, one can observe that OUP has a more gradual change compared to GyrA protein distance, having several jumps at the initial stages. In order to integrate the two methods, mathematical modelling, which identifies relationships between the two matrices of distances, was analysed to emphasise the different weighting at different stages of speciation for each taxonomic unit.

Another interesting result when integrating the two methods was that often multiple logistic curves fit the plotting of OUP and GyrA distance pairs better, which may indicate a discontinuity of speciation events. It means that a graduate delineation between species evolved from one common ancestor may be interrupted by a rapid leap to a new ectomorph when bacterial organisms change their habitat of the lifestyle strategy. Such outbreaks of new forms of life in the bacterial world may be associated with a massive HGT event. For example, environmental free-living bacteria acquire a set of genes making them potential pathogens, symbionts or survivors in harsh environments. To improve researchers' knowledge, there are currently no algorithms or computational tools to distinguish between graduate and chopped speciation events. The possibility to model these events was studied by allowing fitting of multiple logistic curves to distribution plots of OUP-GyrA distances followed by clustering of organisms, which follow different curves. An example is shown in Figure 3.12, in which the above-mentioned group of HL- and LL-adapted strains of *Prochlorococcus marinus* was used. First, the program identified that the best fitting to the distance plot is achieved by two logistic curves (Figure 3.12A). Then the program grouped the organisms into clusters by summarising estimated phylogenetic distances between organisms and sticking distance pairs to one of two logistic curves (Figure 3.12B). Clustering of organisms was well aligned with their belonging to two ectomorphs. LL-adapted strains but MIT9211 were grouped on the left panel of the graph termed Zone 1 (followers of the first logistic curve). LL-adapted strains except for MIT9515 and CCM1986 were grouped in Zone 2. The results are promising, but it should be noted that this algorithm was designed for future studies and its evaluation was not an aim of the current project.

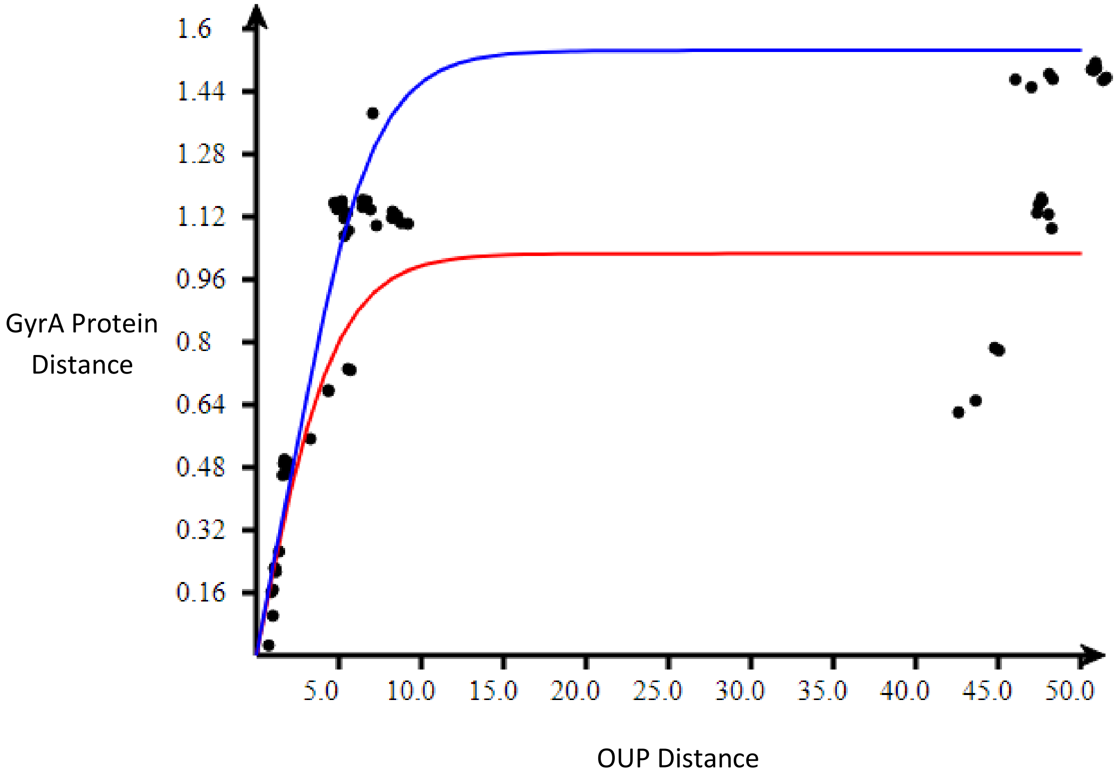


Fig. 3.12A Logistic curve fitting of *Prochlorococcus marinus* strains based on the comparison of OUP and GyrA distances.

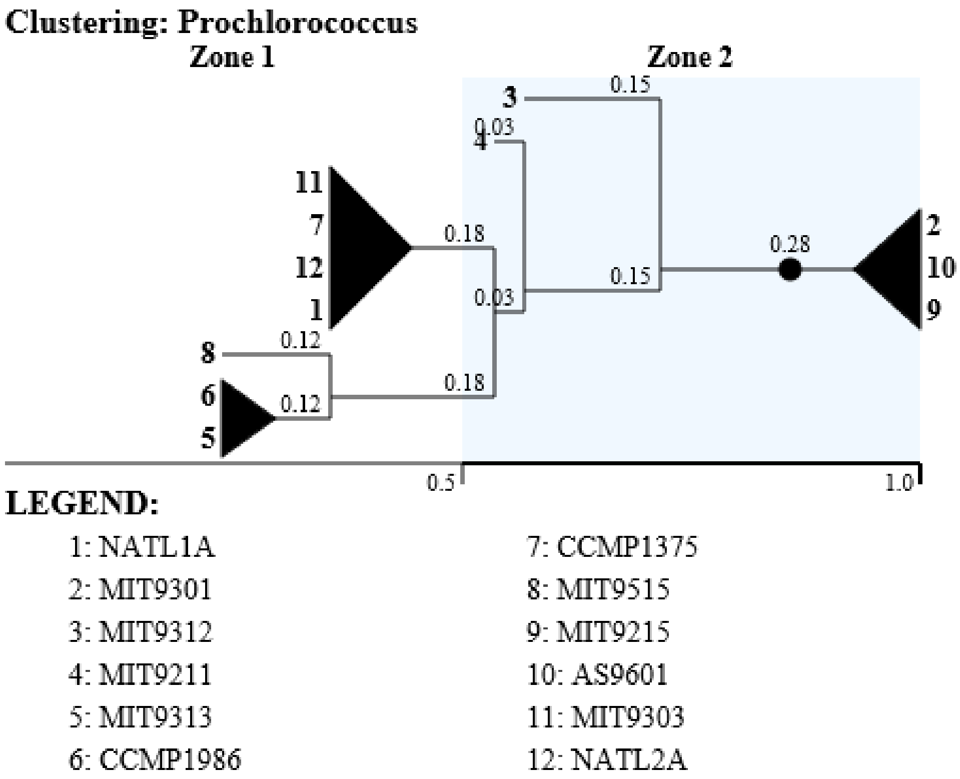
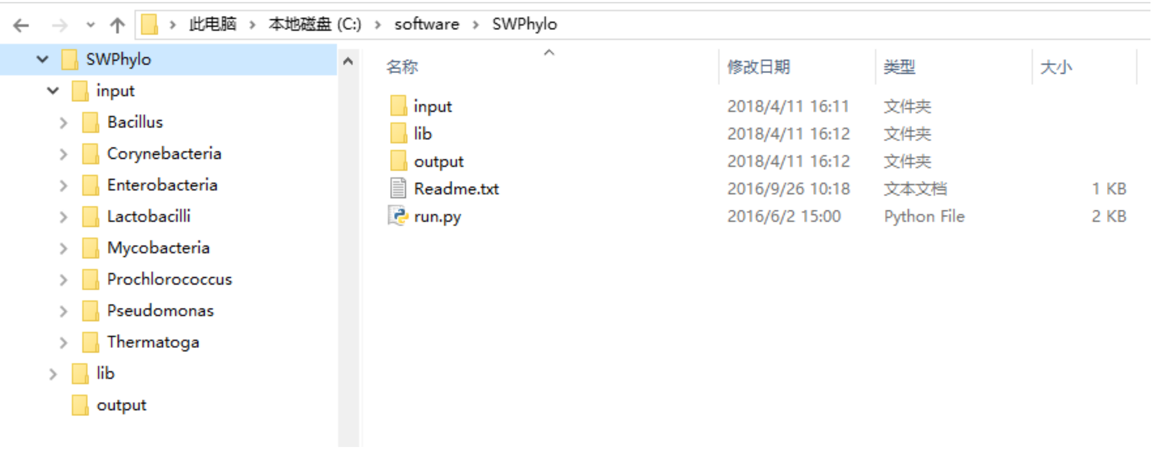


Fig. 3.12B Clustering of *Prochlorococcus marinus* strains based on the logistic model integrating OUP and GyrA distances. The different zones indicated different clusters of organisms following two different logistic curves. Taking the ecotype as an example, the clusters did not group well; the HL and LL strains as strain MIT9211 were grouped with the HL-adapted strain while CCMP1986 and strain MIT9515 were grouped with the LL-adapted strains.

Chapter 4) Design and Implementation of the Program SeqWord Phylogenomics

4.1) SWPhylo Algorithm of OUP-based Phylogenetic Inferences

SeqWord Phylogenomics (SWPhylo) was designed with two implementations, a command line and web-based GUI program, to infer OUP-based phylogenomics. The command line program also forms the basis for the web-based counterpart and is written in Python 3.4. Submodules of Scipy 0.16.1 (Oliphant, 2007), Numpy 1.10.3 (Oliphant, 2006) and Sympy 0.7.2 (Meurer *et al*., 2016) versions onwards are required for the program to run. The command line program takes a folder name as input that contains either a single FASTA file containing multiple sequences or individual Genbank files for phylogenomic inferencing (e.g. “Bacillus” in Figure 4.1). Another functionality of this program allows the input of an alignment file in FASTA format consisting of a phylogenetic marker gene for better resolution of the phylogenetic relationship for the input sequences. This file must also be in the input folder alongside the other input files for the phylogenomic analysis, of which each alignment in the phylogenetic marker gene file must correspond to the sequences in the input folder.



名称	修改日期	类型	大小
input	2018/4/11 16:11	文件夹	
lib	2018/4/11 16:12	文件夹	
output	2018/4/11 16:12	文件夹	
Readme.txt	2016/9/26 10:18	文本文档	1 KB
run.py	2016/6/2 15:00	Python File	2 KB

Fig. 4.1 SWPhylo command line program and folder structure.

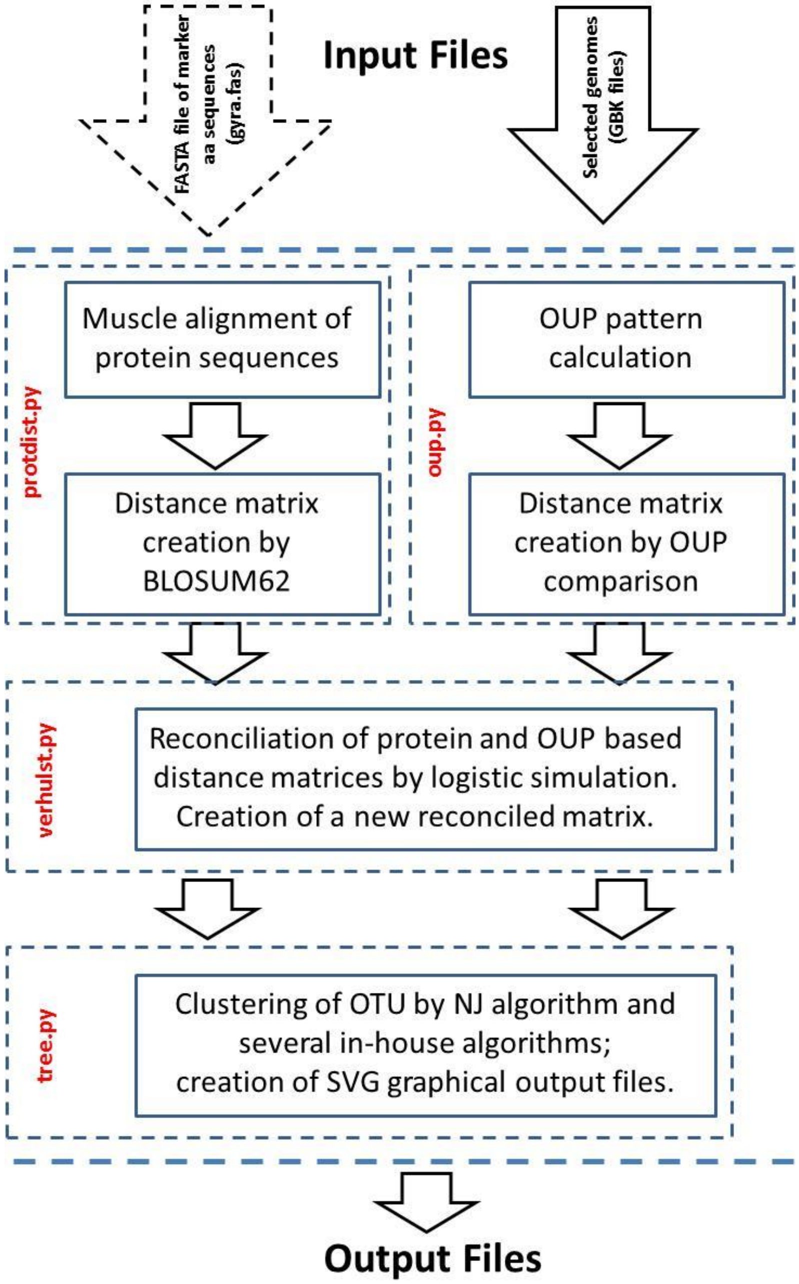
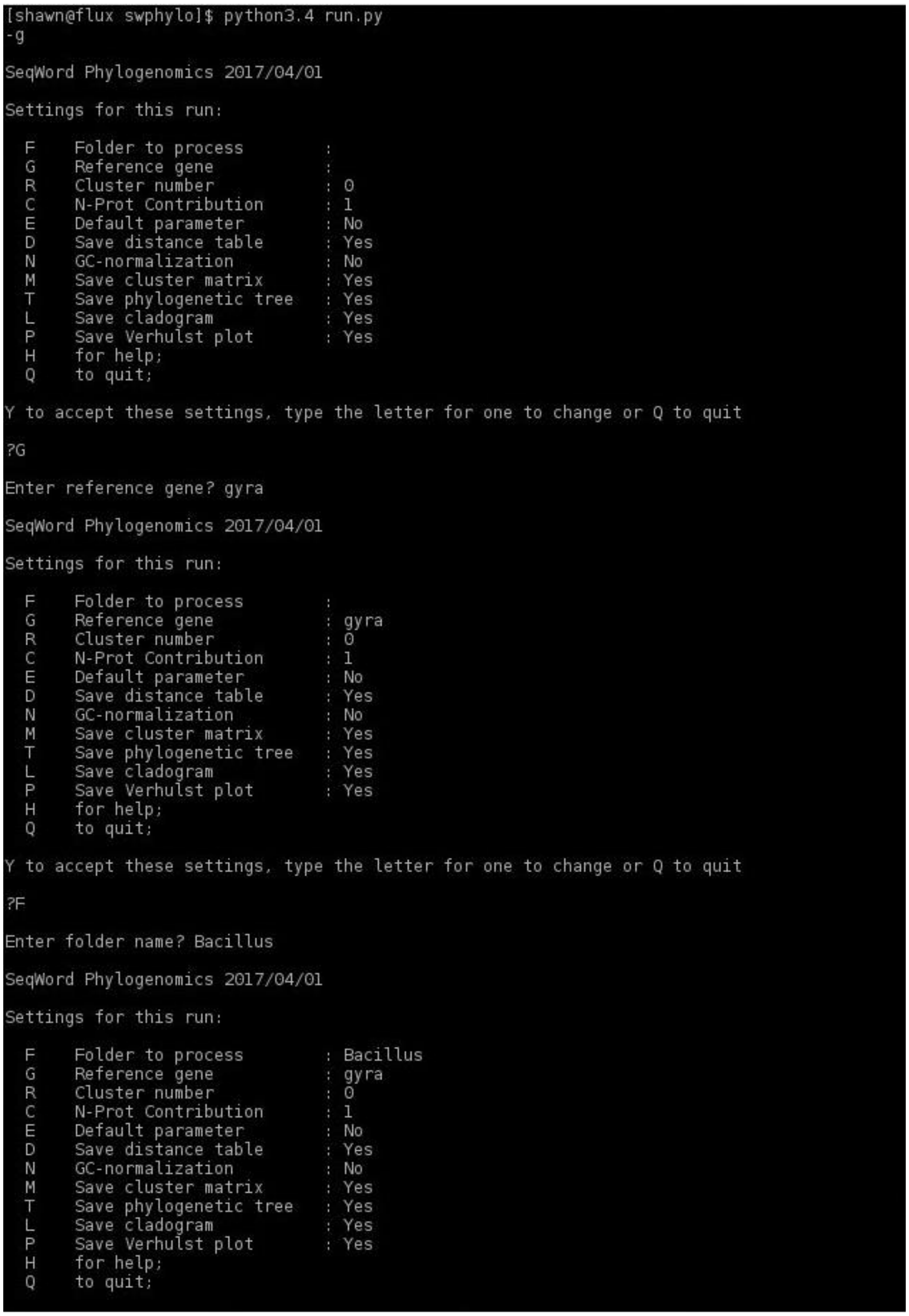

FASTA file of marker aa sequences (gyra.fas)
Input Files
Selected genomes (GBK files)
protdist.py
Muscle alignment of protein sequences
Distance matrix creation by BLOSUM62
oup.py
OUP pattern calculation
Distance matrix creation by OUP comparison
verhulst.py
Reconciliation of protein and OUP based distance matrices by logistic simulation. Creation of a new reconciled matrix.
tree.py
Clustering of OTU by NJ algorithm and several in-house algorithms; creation of SVG graphical output files.
Output Files

Fig. 4.2 Flow diagram explaining the SWPhylo program inferencing procedure of OUP-based phylogenomics.

The program has two workflows depending on the input files used for the startup module *run.py* (Figure 4.2). When the option of using only selected genomes (OTUs) in the input file (Genbank or FASTA) is selected, either using the command line or through changing options, the program will calculate OUP distances between these sequences using the module *oup.py* containing equation [1]. This results in an OUP distance matrix, which is then used for the clustering of OTUs through the *tree.py* module, creating the final phylogenetic tree using the NJ algorithm. Since only OTU sequences are provided, no integration is needed between protein and OUP distance. Therefore this procedure, which is carried out by *verhulst.py*, is skipped in this workflow. The final *tree.py* module creates three outputs, one in graphical .svg format for the display of the phylogenetic tree and the other two .txt files containing tree topology and OUP distance between the OTUs for further in-depth analysis using other programs such as MEGA and PHYLIP (Tamura *et al*., 2013; Tuimala, 2006). All output files are found in the output folder, as seen in the file structure of SWPhylo in Figure 4.1.

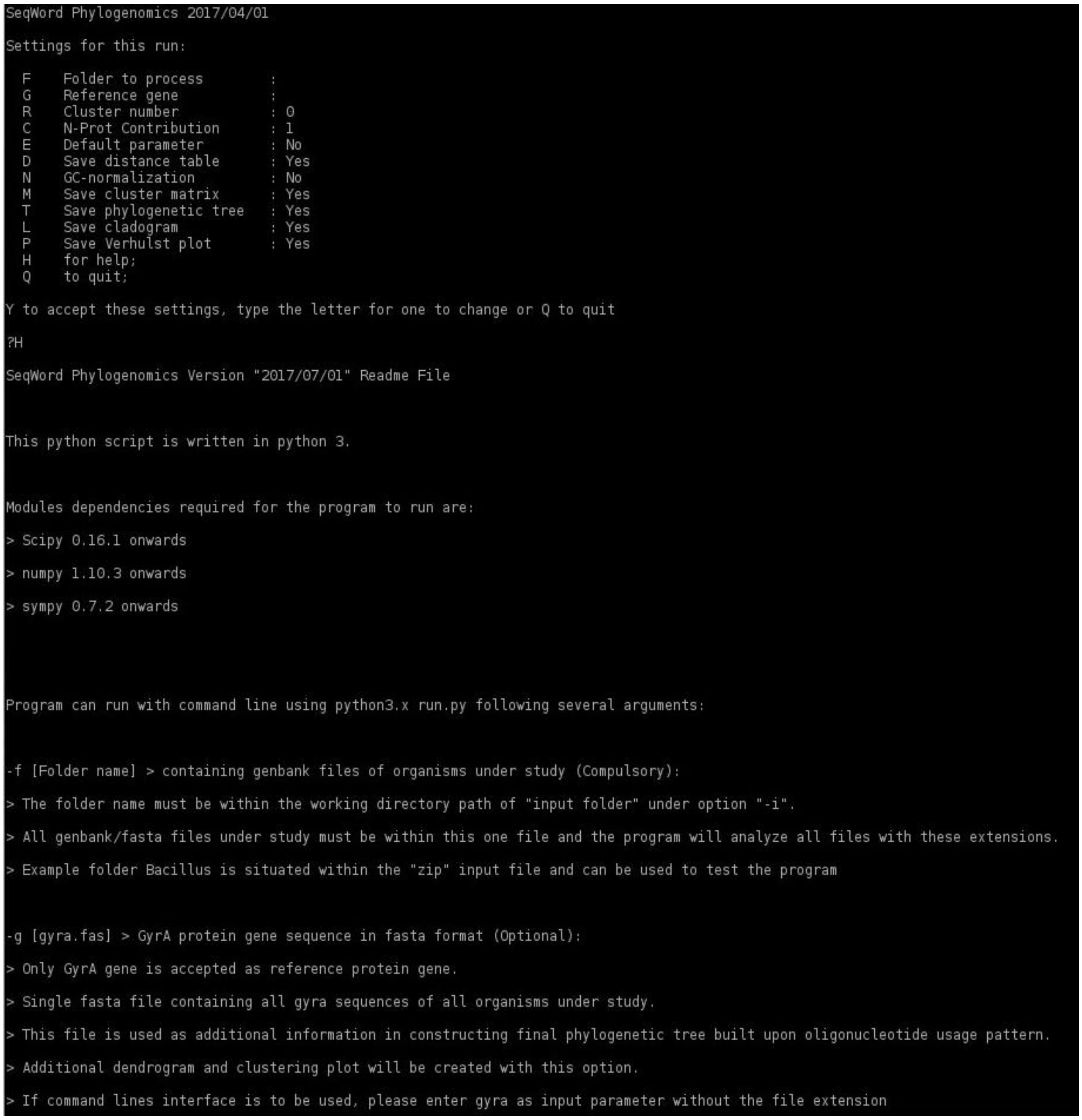
The other workflow is chosen when the option of the phylogenetic marker gene is set and the alignment file of this marker gene for the corresponding OTUs is available in the input folder. A distance matrix comparing the similarities of the marker gene between OTUs is calculated using the module *protdist.py*. This distance is calculated by first aligning this marker gene sequence between each OTU with MUSCLE (Edgar, 2004) and then using substitution matrix BLOSUM 62 (Henikoff and Henikoff, 1992) for the conversion of the similarity distance measure. The integration of OUP and protein distance is done through the module *verhulst.py* using equations 3 and 4, resulting in an integrated distance matrix for the final inference of the phylogenetic tree. This module also consists of the *lmfit.py* module (Newville *et al*., 2014) for the fitting of the Verhulst equation between the OUP and protein distance matrix. Two additional .svg figures are created in this module, including Verhulst fitting of the OUP and protein distance matrices and the clustering of OTUs, depending on the Verhulst fitting. The integrated distance matrix will also be produced by this module in .txt format, which will be used in the *tree.py* module for the final construction of the phylogenetic tree. For this workflow, the output folder will contain six files with three .svg figures and three .txt files. The figure consists of a Verhulst fitting plot, clustering of OTUs plot and a phylogenetic tree, while the three .txt files consist of a clustering matrix for the fitting of the Verhulst model, an integrated distance table and a tree topology text file.

When the OTU sequence files are copied into the input file, as stated in the workflows, the program SWPhylo can be run using two types of command line commands. First, the user can use the *run.py* module only, which will prompt a menu of which the parameters can be set manually to run the program (Figure 4.3). By inserting the corresponding keywords as stated on the interface shown, one can input the necessary information for the program to run according to the needs of the user. An important note for this program is that the input file shown as the “folder to process” option denoted by the keyword “F” is vital for the program to run. This option will let the user choose the folder in which the OTU sequences are found, as shown in the example in Figure 4.3. The folder name was changed from “ ” to “Bacillus” by using the keyword “F”. For the second workflow of integrated method, the keyword “G” can be used for the input of the marker gene sequence data denoted by “gyra” in the example in Figure 4.3. Other options, such as “D”, “M”, “T”, “L” and “P”, allow users to select the outputs created by the program by using either “yes” or “no”. Keyword “N” allows the users to normalise the calculation of OUP distances to account for the difference in GC content of each OTU. This option was not used throughout the study, with unknown influences on resulting inferences. Options “R”, “C” and “E” will only influence the resulting phylogenetic inference when protein sequence data are used. Option “R” limits the number of logistic curves fitted onto the two distance matrices. Option “C” changes the degree of influence the protein distance has on the final integrated distance measure. Option “E” will allow the user to use a pre-calculated Verhulst model estimated by using all the bacterial sequences in this study as a baseline for the integration of protein and OUP-based distances for final phylogenetic inference. Finally, for more information, option “H” contains a help file, which explains each option in the program and how the program can be run through the command line (Figure 4.4).



```
[shawn@flux swphylo]$ python3.4 run.py
-g

SeqWord Phylogenomics 2017/04/01

Settings for this run:

  F    Folder to process        :
  G    Reference gene           :
  R    Cluster number           : 0
  C    N-Prot Contribution      : 1
  E    Default parameter        : No
  D    Save distance table      : Yes
  N    GC-normalization         : No
  M    Save cluster matrix      : Yes
  T    Save phylogenetic tree   : Yes
  L    Save cladogram           : Yes
  P    Save Verhulst plot       : Yes
  H    for help;
  Q    to quit;

Y to accept these settings, type the letter for one to change or Q to quit

?G

Enter reference gene? gyra

SeqWord Phylogenomics 2017/04/01

Settings for this run:

  F    Folder to process        :
  G    Reference gene           : gyra
  R    Cluster number           : 0
  C    N-Prot Contribution      : 1
  E    Default parameter        : No
  D    Save distance table      : Yes
  N    GC-normalization         : No
  M    Save cluster matrix      : Yes
  T    Save phylogenetic tree   : Yes
  L    Save cladogram           : Yes
  P    Save Verhulst plot       : Yes
  H    for help;
  Q    to quit;

Y to accept these settings, type the letter for one to change or Q to quit

?F

Enter folder name? Bacillus

SeqWord Phylogenomics 2017/04/01

Settings for this run:

  F    Folder to process        : Bacillus
  G    Reference gene           : gyra
  R    Cluster number           : 0
  C    N-Prot Contribution      : 1
  E    Default parameter        : No
  D    Save distance table      : Yes
  N    GC-normalization         : No
  M    Save cluster matrix      : Yes
  T    Save phylogenetic tree   : Yes
  L    Save cladogram           : Yes
  P    Save Verhulst plot       : Yes
  H    for help;
  Q    to quit;
```

Fig. 4.3 Command line interface for SWPhylo command line program. Parameters can be changed through keywords, as shown in this figure.



```
SeqWord Phylogenomics 2017/04/01

Settings for this run:

  F    Folder to process        :
  G    Reference gene           :
  R    Cluster number           : 0
  C    N-Prot Contribution      : 1
  E    Default parameter        : No
  D    Save distance table      : Yes
  N    GC-normalization         : No
  M    Save cluster matrix      : Yes
  T    Save phylogenetic tree   : Yes
  L    Save cladogram           : Yes
  P    Save Verhulst plot       : Yes
  H    for help;
  Q    to quit;

Y to accept these settings, type the letter for one to change or Q to quit

?H

SeqWord Phylogenomics Version "2017/07/01" Readme File


This python script is written in python 3.


Modules dependencies required for the program to run are:

> Scipy 0.16.1 onwards

> numpy 1.10.3 onwards

> sympy 0.7.2 onwards



Program can run with command line using python3.x run.py following several arguments:


-f [Folder name] > containing genbank files of organisms under study (Compulsory):

> The folder name must be within the working directory path of "input folder" under option "-i".

> All genbank/fasta files under study must be within this one file and the program will analyze all files with these extensions.

> Example folder Bacillus is situated within the "zip" input file and can be used to test the program


-g [gyra.fas] > GyrA protein gene sequence in fasta format (Optional):

> Only GyrA gene is accepted as reference protein gene.

> Single fasta file containing all gyra sequences of all organisms under study.

> This file is used as additional information in constructing final phylogenetic tree built upon oligonucleotide usage pattern.

> Additional dendrogram and clustering plot will be created with this option.

> If command lines interface is to be used, please enter gyra as input parameter without the file extension
```

Fig. 4.4 Help file associated with the SWPhylo program for command line and interface usage. This file can also be seen through text editors under readme.txt.

For more advanced users, the program allows a single command line to run the program. In the command line, each option required by the user must be set with a central slash “-” alongside the corresponding keyword in small letters and the information the user needs to input into the program. For the same input as the example in Figure 4.3, a single command line takes the form of:

```
>  python3.4 run.py –f Bacillus –g gyra
```

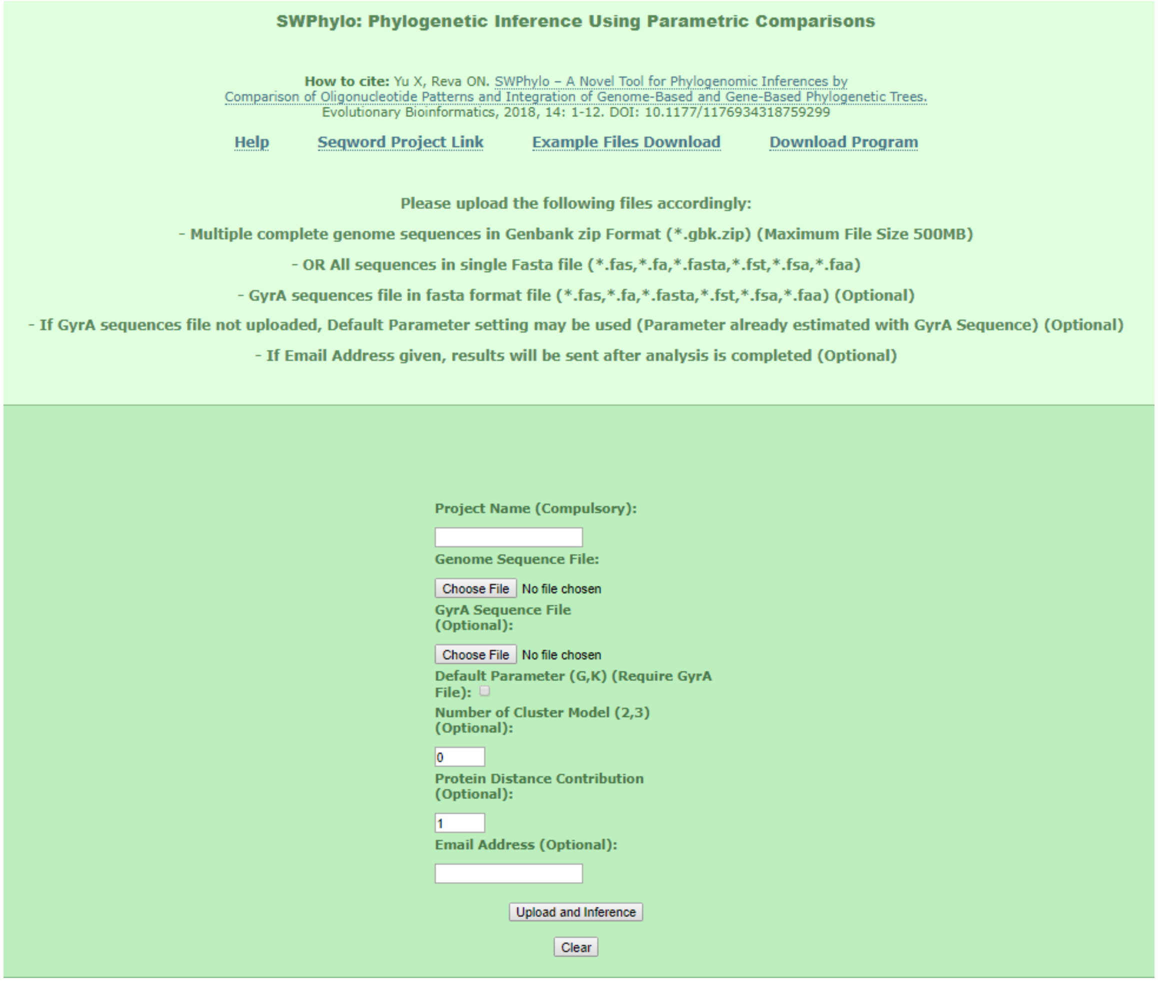
This command line will run the program SWPhylo with the input folder Bacillus and the GyrA protein sequence data corresponding to the OTU sequences found in the same folder. All other options can be changed using the same rule as stated above (e.g. –c 2 to change the protein contribution factor from the default value 0 to a user-set value 2). A *readme.txt* file is available in the SWPhylo program for more information on running the program using the command line, what the various options are and what they are associated with (Figure 4.1).

In the last section of this chapter, before concluding this research, we will look in more detail at the web-based SWPhylo program and its graphical interface created for easier representation and usage of the program for researchers to implement OUP-based phylogenomics. He will also go into detail as to how the program is used and familiarise users with the various parameter inputs to achieve the most accurate result in terms of different datasets.

4.2) Design of the Web-based Software Tool SWPhylo

The web-based SWPhylo implementation is a Python program integrated into a web-based user interface shown in Figure 4.5. SWPhylo is written in Python 3.4 and accessible at http://swphylo.bi.up.ac.za/, set up by a PHP framework. The program allows the submission of complete genome sequences or large genomic fragments either in FASTA or GenBank formats. The single FASTA file in comparison to GenBank files may contain multiple genome sequences stored for analysis. If the genomic sequences are represented by individual GenBank files, they must be compressed into a single archive file (.zip) before uploading. The program will need a project name in order to run, as this project name will be displayed on top of all output figures shown in Figure 4.6, Figure 4.7 and Figure 4.8. Therefore, the project name should be impactful and relevant to a user-submitted dataset for display in output figures. The project name will not allow any non-letter and number characters, which will be removed because of the file naming criteria on the server.

Optionally, the program allows submission of an additional FASTA file with an alignment of GyrA protein sequences in FASTA format. The number of sequences in this file must equal the number of submitted genomes and they must be given the same identifiers. This option allows better resolution of the phylogenetic tree based on the integration of the *gyrA* phylogenetic marker gene and OUP-based methods. Users may explore the functionality of the program by using example files available from the web page under “Example Files Download”, which will prompt a zip file for download. This example zip file contains a group of *Bacillus* species alongside a GyrA protein sequence file in FASTA format. Users may also use this file as a reference to validate their input file(s) to determine whether they are valid for use on the web page. If a protein alignment file is provided (i.e. gyrA.fas), the program will combine the input datasets in the resulting tree by using equation [4] based on the integration of OUP- and GyrA-based distances. Alternatively, the program will infer a phylogenetic tree solely through the OUP comparison (Figure 4.8).



SWPhylo: Phylogenetic Inference Using Parametric Comparisons

How to cite: Yu X, Reva ON. SWPhylo – A Novel Tool for Phylogenomic Inferences by Comparison of Oligonucleotide Patterns and Integration of Genome-Based and Gene-Based Phylogenetic Trees. Evolutionary Bioinformatics, 2018, 14: 1-12. DOI: 10.1177/1176934318759299

Please upload the following files accordingly:

- Multiple complete genome sequences in Genbank zip Format (*.gbk.zip) (Maximum File Size 500MB)

- OR All sequences in single Fasta file (*.fas,*.fa,*.fasta,*.fst,*.fsa,*.faa)

- GyrA sequences file in fasta format file (*.fas,*.fa,*.fasta,*.fst,*.fsa,*.faa) (Optional)

- If GyrA sequences file not uploaded, Default Parameter setting may be used (Parameter already estimated with GyrA Sequence) (Optional)

- If Email Address given, results will be sent after analysis is completed (Optional)

Project Name (Compulsory):

Genome Sequence File: No file chosen

GyrA Sequence File (Optional): No file chosen

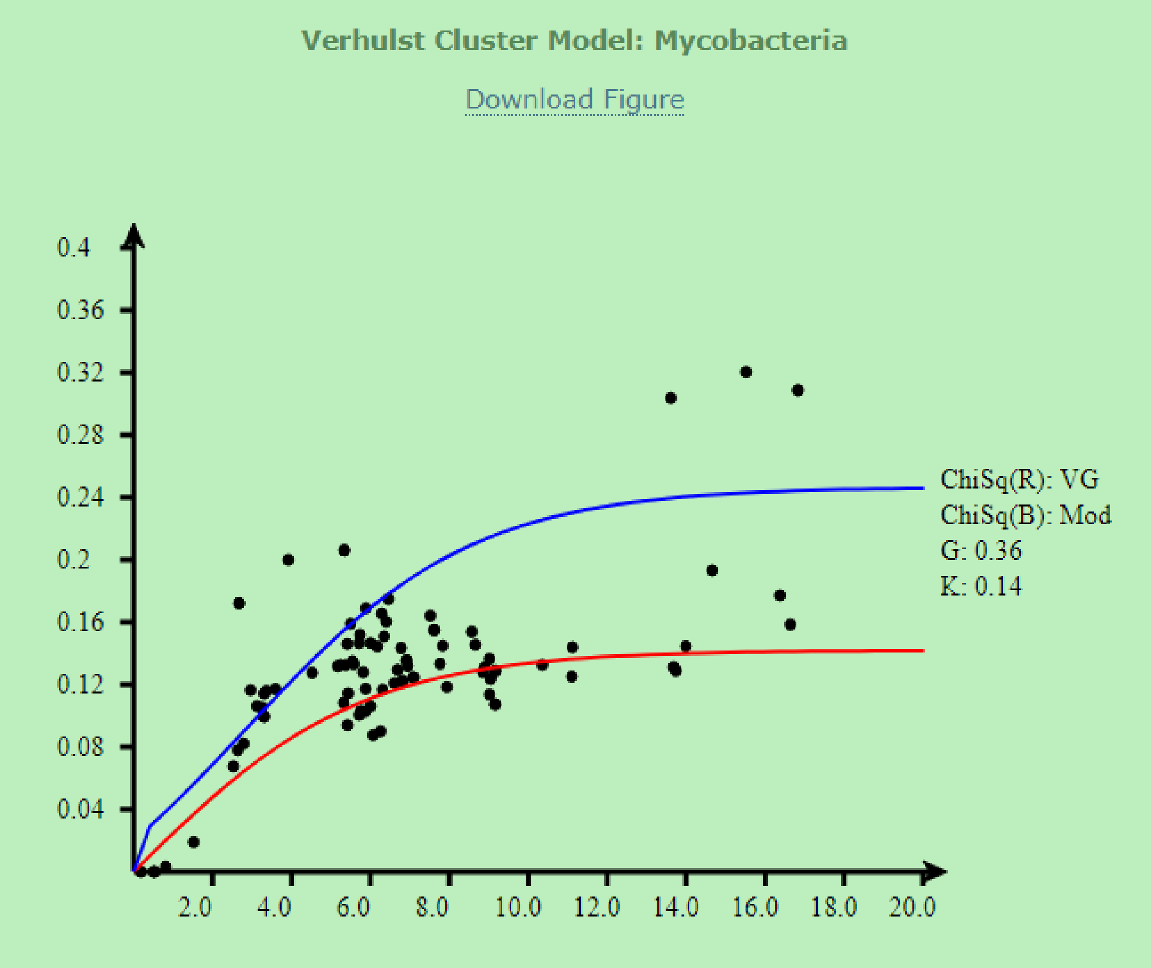
Default Parameter (G,K) (Require GyrA File): □

Number of Cluster Model (2,3) (Optional): 0

Protein Distance Contribution (Optional): 1

Email Address (Optional):

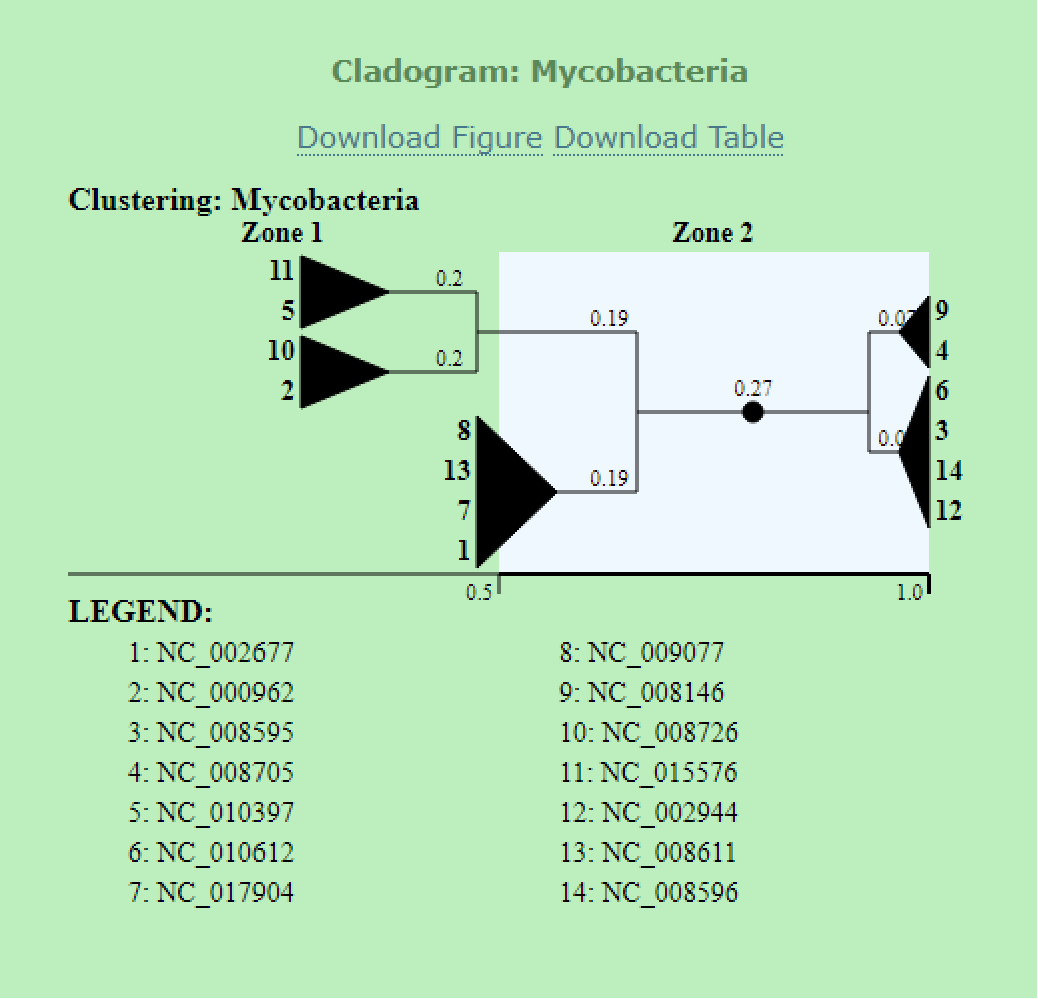
Fig. 4.5 SWPhylo web-based user interface.



Verhulst Cluster Model: Mycobacteria

Download Figure

Fig. 4.6 Logistic clustering output from SWPhylo web interface of taxonomic group Mycobacteria. On the right of the graph, parameter values g and K are shown, as well as how well the logistic curve was fitted according to chi-squared fitting criteria.



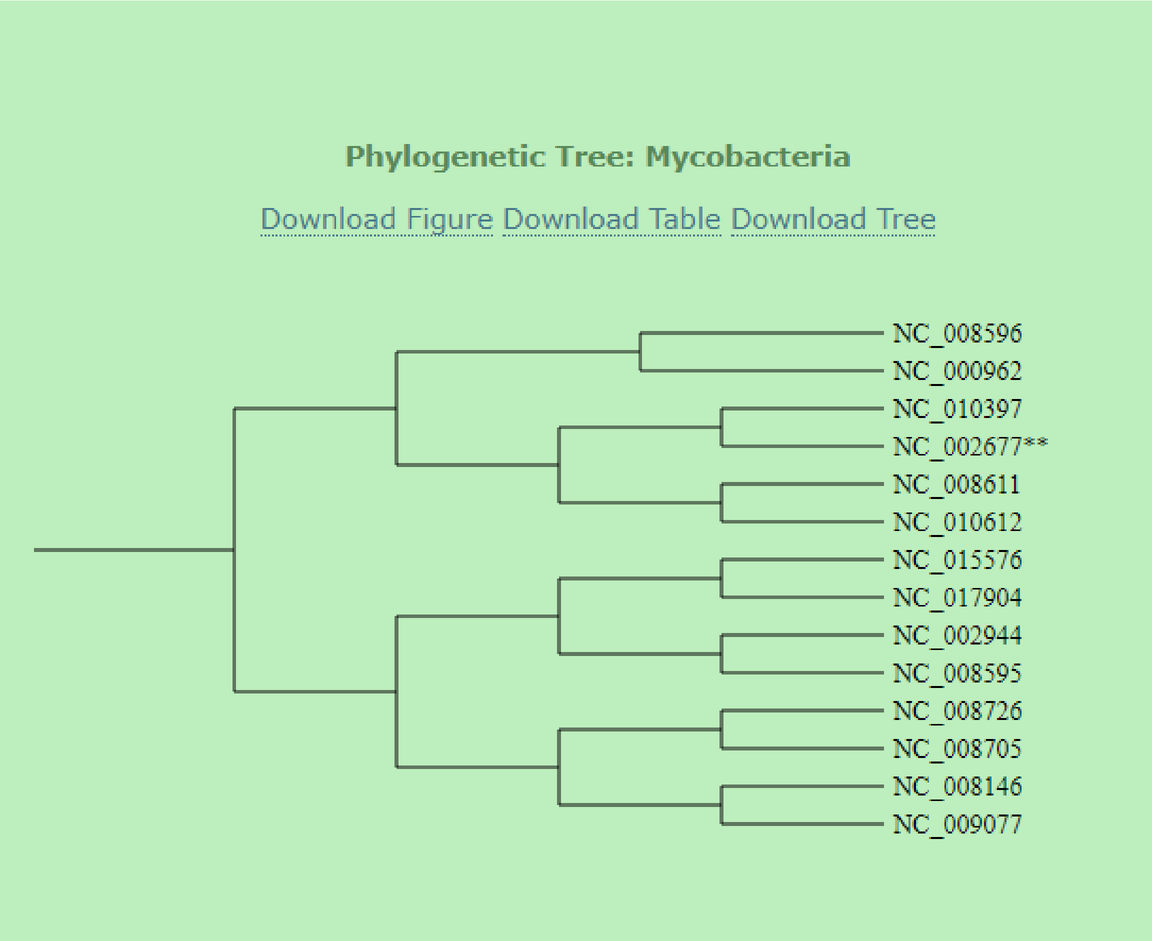
Cladogram: Mycobacteria

Clustering: Mycobacteria

LEGEND:

1: NC_002677
2: NC_000962
3: NC_008595
4: NC_008705
5: NC_010397
6: NC_010612
7: NC_017904
8: NC_009077
9: NC_008146
10: NC_008726
11: NC_015576
12: NC_002944
13: NC_008611
14: NC_008596

Fig. 4.7 Cladogram clustering organisms according to their respective logistic curves. Each logistic curve represents one zone and the number of pairs of distances (dots) organisms have within one curve will determine their position in the zone, e.g. organism 11 has a greater proportion of dots in zone 1 compared to organism 13.



Phylogenetic Tree: Mycobacteria

Fig. 4.8 Phylogenetic tree resulting from integrated method of gyrA and OUP distance. The asterisk on organism NC_002677 shows that in terms of Oligonucleotide usage variance, this organism might be an outlier of this dataset and may not be represented correctly in terms of its phylogeny

Several additional functional parameters were added to the SWPhylo website to give better resolution of phylogenetic relationships in the datasets. If a protein sequence alignment file is provided, an important parameter is the contribution of protein sequence distances to the estimation of phylogenetic distances between OTUs (Figure 4.8). This parameter is used to control the weighting of GyrA distances compared to OUP distances in accordance with equation [4] shown in the previous chapter. By default, this value is 1, which creates equal weighting between GyrA and OUP distances. This may be changed to either 2 or 3 to increase the contribution of marker sequence differences between strains over oligonucleotide composition differences in the process of phylogenetic inference. We recommend using value 2 when closely related species are compared, while value 3 is used to distinguish between subspecies of the same species. The rationale for this is that closely related organisms differ in marker gene sequences but share the same OUP. This concept was discussed in detail in the previous chapter. This function can only take effect if an aligned sequence of marker proteins (e.g. GyrA) has been uploaded in FASTA format together with a ZIP file of genome sequences.

Another parameter associated with processing of uploaded alignments of phylogenetic marker genes is the checkbox forcing the use of default parameters of reconciliation of sequence-based and OUP-based phylogenetic inferences. When this option is checked, the program uses the default values of the coefficients *g* and *K* (*g* = 0.0775; *K* = 1.3379, see equation [3]) estimated for a joined set of all the taxonomic groups used in this study (Chapter 2.2.1). The use of the default *g* and *K* parameters ensures independence of inference results on the submitted sample composition. When the checkbox is unchecked, the program recalculates these parameters for the given dataset to reflect the group-specific rates of evolutionary changes in OUP and marker gene sequences. This allows the parameters used for phylogenetic reference of the submitted dataset to be data-specific and dependent on the sample content. Recalculation of the parameters may improve the accuracy of a phylogenetic inference by a proper reconciliation of sequence and OUP signals. However, phylogenetic trees calculated in this way for different sets of genomes may not be comparable. One other reason to force recalculation of the default parameters is discussed below.

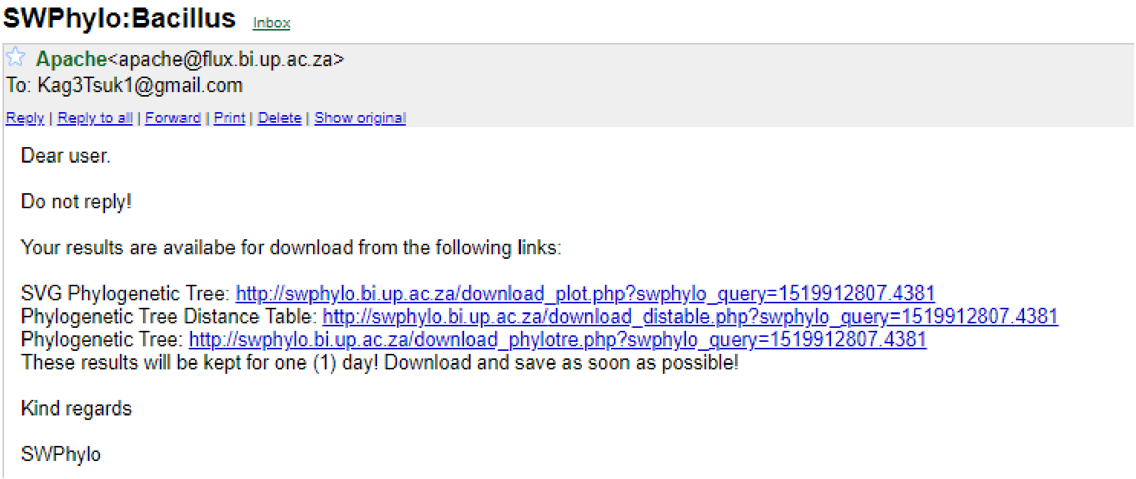
In the current project, we did not explore the possibility of using other housekeeping proteins instead of GyrA. Potentially there should be no problem in using other popular genetic markers such as ribosomal proteins recommended in multiple publications as universal phylogenetic markers (Yutin et al., 2012; Hug et al., 2016). If alternative protein sequences are submitted as an input, the checkbox of the default *g* and *K* parameters should be unchecked, as these default parameters were calculated specifically for reconciliation of OUP with GyrA protein distances. The logistic curve fits well to the distribution of GyrA protein distance and OUP, but may not be appropriate for other genetic markers. This is why the parameters of the logistic curve integrating OUP differences with phylogenetic distances between other marker proteins should be recalculated.

When a GyrA protein sequence is provided, the program returns two additional figures: the logistic curve diagram (Figure 4.6) and a cluster plot (Figure 4.7). As was discussed in the previous chapter, the fitting of OUP to protein distances (i.e. GyrA) often achieves the best result when several logistic curves with individual *g* and *K* parameters are applied instead of one. The hypothesis adopted in this thesis, which needs further experimental approval, is that micro-evolutionary speciation is not a graduated process but constitutes a series of evolutionary leaps associated with changes in lifestyle strategies or habitat specificities in bacterial populations. These evolutionary leaps are reflected by a series of logistic curves in Figure 4.6, with step-wise *K* parameters. In Figure 4.7, organisms are first grouped into clusters by their OUP/GyrA similarities, and then they are plotted into zones so that the organisms of each zone fit to the same logistic curve. The distance between zones is the number of evolutionary leaps. It should be noted that this analysis is applicable only for comparison of closely related organisms, as multiple leaps in different evolutionary branches lead to a random distribution of OUP-GyrA distance pairs. These logistic curves were fitted by the program using the Python module *lmfit* with the best fit chosen by AIC values, i.e. an optimal number of logistic curves fitted onto the dataset with the lowest AIC value are chosen (Chapter 1.4). The chi-squared goodness of fit test was also converted to a keyword for easier interpretation, indicating to the users if the fit was good or not, as seen in Figure 4.6 (Newville *et al*., 2014). If all data points are within the confidence interval of 90%, it is considered a very good fit (VG); between 75% and 90% is good; between 50% and 75% is moderate (mod); between 25% and 50% is bad and below 25% is very bad (VB).

The program performs clustering of the taxonomic operational units (i.e. genomes in this study) around different logistic curves. Depending on the proportion of pairs of distances between genomes belonging to different logistic curves, the positions within zones will differ, i.e. if there are 10 genomes under study, for each genome there should be nine dots representing pairs of distances to other genomes. For instance, if genome A produces five pair dots in logistic curve I and four dots in logistic curve II, genome A is assigned to the zone associated with curve I. However, the level of sharing of genome-specific pair dots between different logistic curves is reflected by cluster positions in particular zones. Terminal clusters constitute genomes strongly associated with specific logistic curves (Figure 4.7).

Clustering may reflect either different evolutionary rates in tree branches or evolutionary leaps towards occupation of new niches and/or habitats during speciation. These leaps may be associated with an abrupt burst of positively selected mutations in housekeeping genes. The number of clusters by default is determined by the program automatically based on AIC values, but may be set by the user in the parameter field ‘Number of clusters’. For more detail on each parameter field, users can consult the user guide on http://swphylo.bi.up.ac.za/ under the help tab.

The last output figure of the program SWPhylo (with or without GyrA protein alignment file provided) is a simple cladogram representing only the phylogenetic tree topology (Figure 4.8). Users may download the actual distance table in the standard PHYLIP format to analyse the phylogenetic relationships by using more sophisticated tools, such as the programs *neighbour*, *fitch* and *kitch* from the PHYLIP package, MEGA6 (Tamura *et al*., 2013) or SplitsTree4 (Huson and Bryant, 2006). The user can also choose to submit an email address, which allows the program to send hyperlinks to the result files when calculations are done. The email contains several links for the users to access and download all possible results from these websites (Figure 4.9). For both email and online results, for each output, raw data in the form of distance tables (.txt) and phylogenetic trees in two formats (.svg) and (.txt) can be downloaded. These results will be kept on the server for 24 hours before being deleted. SWPhylo is also downloadable and can be used through command line alongside Python 3.4 with detailed guidelines stated in the previous section.


Inbox
Reply | Reply to all | Forward | Print | Delete | Show original

SWPhylo:Bacillus

Apache<apache@flux.bi.up.ac.za>
To: Kag3Tsuk1@gmail.com

Dear user.

Do not reply!

Your results are availabe for download from the following links:

SVG Phylogenetic Tree: http://swphylo.bi.up.ac.za/download_plot.php?swphylo_query=1519912807.4381
Phylogenetic Tree Distance Table: http://swphylo.bi.up.ac.za/download_distable.php?swphylo_query=1519912807.4381
Phylogenetic Tree: http://swphylo.bi.up.ac.za/download_phylotre.php?swphylo_query=1519912807.4381
These results will be kept for one (1) day! Download and save as soon as possible!

Kind regards

SWPhylo

Fig. 4.9 Example of an email sent from SWPhylo website containing results based on the input data from users.

It was concluded from this research that the diversification of genomic OUP is more distinguishable with regard to a constant time factor compared to the rates of substitution in individual genes. This makes OUP comparison a promising approach to estimate the relative time of evolution of organisms. However, it should be noted that there may be exclusions from this assumption. In the paper by Reva and Tummler (2004), several bacterial genomes were noted in which the global OUP experienced a drastic demolition for unknown reasons. One of these organisms was *Xylella fastidiosa* 9a5c. The OUP characteristics of this genome were unprecedented for bacteria chromosomal with strand asymmetry and low OUV, implicating a mutator phenotype. The reason for these dramatic processes was assumed to be associated with the acquisition of a large genomic island of *Pseudomonas* origin comprising several active phage integrases. This has a degenerative effect on the whole chromosome (Klockgether *et al*., 2007). Interestingly, protein sequences of *X. fastidiosa* 9a5c remained very similar to those of *X. fastidiosa* Temecula1 in contrast to their differing OUP. The separation of *X. fastidiosa* 9a5c from *X. fastidiosa* Temecula1 in an OUP-based phylogenetic tree will therefore be an overestimation. Another example of a problematic organism is *Mycobacterium leprae*. The OUV of this genome is significantly lower than the other *Mycobacteria,* which implies a higher rate of mutations or weaker conservation of OUP. A relaxed codon bias could be beneficial to this pathogen, causing long-lasting chronic infection to slow down the growth rate. On the OUP-based tree, this bacterium seems more distant to the tuberculosis cluster than may be estimated by protein sequence comparison (Figure 4.8 marked by double asterisk).

To warn users that the phylogeny of a specific bacterium may not have been identified correctly, the program uses deviations in OUV values (see equation [2]). One asterisk displayed on an output phylogenetic tree marks the organisms with genomic OUV 2.5xSTD larger than the average OUV of the dataset. Two asterisks depict genomes characterised by OUV 2.5xSTD lower than the average. Sequences marked with either number of asterisks are shown be outliers and such sequences might be misplaced within the phylogenetic tree.

Chapter 5) Conclusions

The aim of this project was to investigate the evolutionary implications of OUP and in turn the feasibility of applying OUP as a phylogenomic inferencing measure to identify the relationships between OTUs. Through literature, it was identified that codon bias and adaptation, as well as context nucleotide selection forces, drove OUP formation. This finding was in congruence with the results, which showed high correlation between both above-mentioned factors and OUP in the simulation models in this study. Based on these driving forces, which act as fundamental aspects of species evolution, OUP has the potential to be used for phylogenomic inference between OTUs. To analyse the feasibility of this approach, a distance matrix-based OUP approach was chosen for phylogenomic inferencing owing to its simplicity and freedom from any evolutionary hypothesis.

The OUP-based algorithm was proven to be a feasible phylogenomic comparison metric based on multiple case studies on various groups of microorganisms selected to represent different bacterial provenances by taxonomically well-characterised species. OUP-based trees were most congruent with WGS trees for the majority of the bacterial groups used in this study, both in terms of topology and branch length. OUP-based methods were comparable to other well-known methods such as 16S rRNA phylogenetic comparisons, the supermatrix WGS approach, MAUVE whole genome alignment and CVTree amino acid k-word comparison. The method also creates consistent inferences with known phylogeny according to simulated datasets with varying combinations of different parameters creating different evolutionary scenarios. Through the case study with the two ecotypes of *Prochlorococcus,* OUP was able to outperform the traditional method of phylogenetics based on alignment of 16S rRNA sequences in terms of distinguishing between sub-species and ectomorphs of closely related organisms. OUP performed better for divergent sequences, with mixed results in terms of closely related organisms, of which phylogenetic markers such as *gyrA* and 16S rRNA might infer better results. Bootstrapping results using different sequence lengths also showed the high robustness of this method, through which sequences as short as 50 kbp were able to make highly reliable phylogenetic inferences.

Integration of methods using the marker gene gyrA and OUP-based method was created in order to resolve discrepancies and incongruence between phylogenomic methods and infer more accurate phylogenetic relationships between OTUs. The case study for resolving the phylogenetic relationship between different *Prochlorococcus* ecotypes has shown an improvement using an integrated method of both OUP and GyrA protein distances. The method was able to distinguish two logistic curves clearly, indicating two ecotypes in which only three of the 12 strains was misplaced, compared to literature. This method was also an improvement on the stand-alone OUP approach through which the addition of GyrA protein distance was able to resolve conflicts in closely related organisms, where OUP is inadequate.

The GyrA protein used in this analysis may potentially be replaced by other genetic markers to suit a specific set of organisms better. For example, the case study with the taxonomic group *Prochlorococcus*, 16S rRNA ribosomal proteins, might be more suitable to distinguish the closely related highlighted ecotypes more clearly. This work focuses on reconciliation of evolutionary distances calculated by comparison of OUP and GyrA sequences. Other phylogenetic markers were not considered and might be interesting to study in future. It has to be noted that in the context of this study, the logistic model only works well with the integrated method of OUP and GyrA. Furthermore, the logistic model cannot guarantee the accuracy of all taxonomic groups, as seen by the number of branch relocations between the true phylogenetic tree and OUP in terms of simulated sequences. Lastly, the distance-based approach has its own limitations and should optimally be supplemented in future with a likelihood model of OUP evolution.

OUP-based phylogenomics was built and implemented into both a web-based program and a command-line-based program named SWPhylo. SWPhylo allows researchers to infer phylogenetic relationships between OTUs using either OUP or integrated GyrA- and OUP-based approaches. In Figure 5.1, using Google Analytics, the geographical data of researchers using this program for their phylogenomic research are shown. SWPhylo is computationally efficient, not reliant on annotation and alignment information and can be used on large whole genome or partial genome datasets. This tool is robust and unique, which are its core advantages as a phylogenomic toolset. The program can be found and downloaded from the SWPhylo website at http://swphylo.bi.up.ac.za.

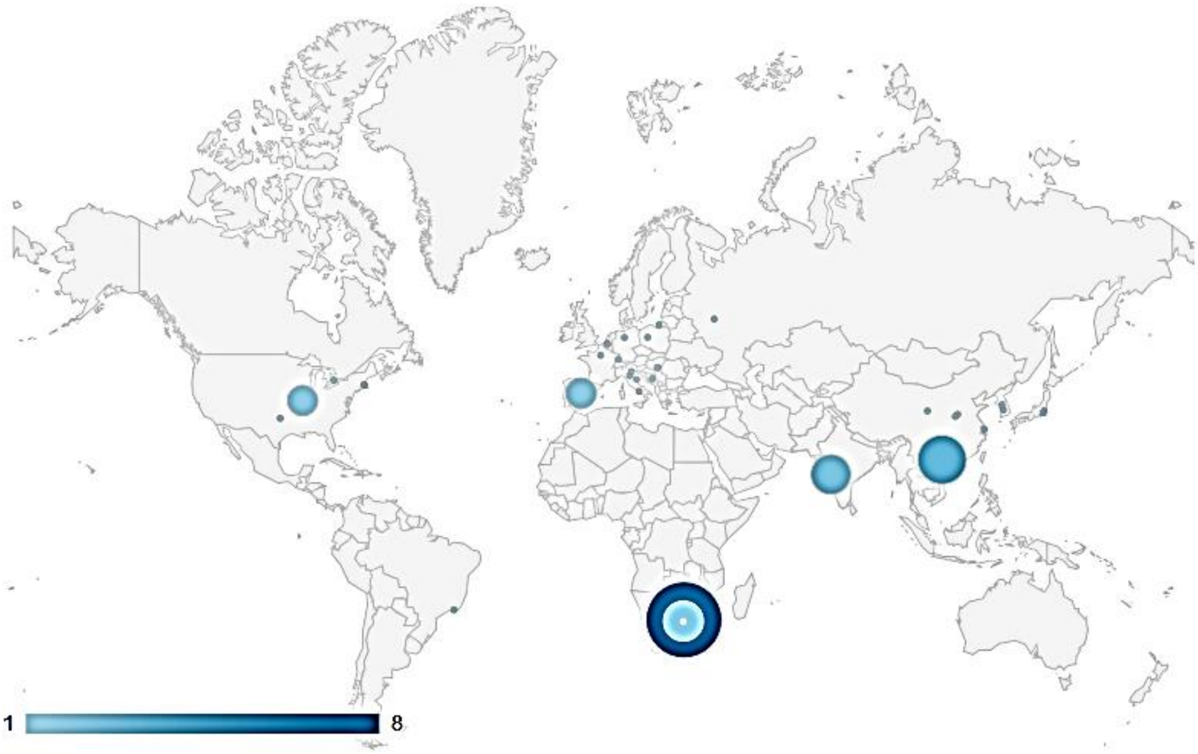


Fig. 5.1 Geographical view of users using the website program SWPhylo, taken from Google Analytic from the period March 2018 to June 2018 (Google.com, 2018).

.

Acknowledgement

I would like to thank all the members at the Bioinformatics and Computational Biology Unit at the University of Pretoria for all their support, guidance and assistance, in particular my supervisor Dr. O. Reva and Dr R. Pierneef. Finally, I would like to thank National Research Foundation and University of Pretoria for their funding for this project.

References

(2015) MATLAB. 2015a ed. Natick, Massachusetts: The MathWorks Inc.

Abdo Z, Minin VN, Joyce P, et al. (2005) Accounting for uncertainty in the tree topology Has little effect on the decision-theoretic approach to model selection in phylogeny estimation. *Mol Biol Evol* 22.

Akaike H. (1974) A New Look at the Statistical Model Identification. *IEEE Trans Automat Control* 19.

Almagor H. (1983) A Markov analysis of DNA sequences. *J Theor Biol* 104: 633-645.

Altschul S, Gish W, Miller W, et al. (1990) Basic Local Alignment Search Tool. *Jour Mol Biol* 215.

Andrew G, Carlin JB, Stern HS, et al. (1995) Bayesian data analysis. *Chapman and Hall*.

Arahal DR. (2014) Chapter 6 - Whole-Genome Analyses: Average Nucleotide Identity. In: Goodfellow M, Sutcliffe I and Chun J (eds) *Methods in Microbiology.* Academic Press, 103-122.

Baldi P and Baisnée P-F. (2000) Sequence analysis by additive scales: DNA structure for sequences and repeats of all lengths. *Bioinformatics* 16: 865-889.

Bansal AK and Meyer TE. (2002) Evolutionary analysis by whole-genome comparisons. *J Bacteriol* 184: 2260-2272.

Bayzid MS, Hunt T and Warnow T. (2014) Disk covering methods improve phylogenomic analyses. *BMC Genomics* 15: S7.

Beiko RG. (2010) Gene sharing and genome evolution: networks in trees and trees in networks. *Biol Philos* 25.

Berendzen J, Bruno WJ, Cohn JD, et al. (2012) Rapid phylogenetic and functional classification of short genomic fragments with signature peptides. *BMC Research Notes* 5: 1-21.

Bergsten J. (2005) A review of long-branch attraction. *Cladistics* 21.

Bezuidt OK, Pierneef R, Gomri AM, et al. (2016) The Geobacillus Pan-Genome: Implications for the Evolution of the Genus. *Frontiers in Microbiology* 7: 723.

Bhandari V and Gupta RS. (2014) Molecular signatures for the phylum (class) Thermotogae and a proposal for its division into three orders (Thermotogales, Kosmotogales ord. nov. and Petrotogales ord. nov.) containing four families (Thermotogaceae, Fervidobacteriaceae fam. nov., Kosmotogaceae fam. nov. and Petrotogaceae fam. nov.) and a new genus Pseudothermotoga gen. nov. with five new combinations. *Antonie Van Leeuwenhoek* 105: 143-168.

Binet M, Gascuel O, Scornavacca C, et al. (2016) Fast and accurate branch lengths estimation for phylogenomic trees. *BMC Bioinformatics* 17: 23.

Bininda-Emonds ORP. (2004) Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life.

Blaimer BB, Brady SG, Schultz TR, et al. (2015) Phylogenomic methods outperform traditional multi-locus approaches in resolving deep evolutionary history: a case study of formicine ants. *BMC Evolutionary Biology* 15: 1-14.

Blair C and Murphy RW. (2011) Recent trends in molecular phylogenetic analysis: where to next? *J Hered* 102.

Blanc-Potard A-B, Solomon F, Kayser J, et al. (1999) The SPI-3 Pathogenicity Island of Salmonella enterica. *Journal of Bacteriology* 181: 998-1004.

Blattner F, Plunkett G, Bloch C, et al. (1997) The complete genome sequence of Escherichia coli K-12. *Science* 277: 1453 - 1474.

Blom J, Albaum SP, Doppmeier D, et al. (2009) EDGAR: A software framework for the comparative analysis of prokaryotic genomes. *BMC Bioinformatics* 10: 154-154.

Bochkareva OO, Dranenko NO, Ocheredko ES, et al. (2018) Genome rearrangements and phylogeny reconstruction in Yersinia pestis. *PeerJ* 6: e4545.

Bofkin L and Goldman N. (2007) Variation in evolutionary processes at different codon positions. *Mol Biol Evol* 24.

Bohlin J, Skjerve E and Ussery D. (2008) Reliability and applications of statistical methods based on oligonucleotide frequencies in bacterial and archaeal genomes. *BMC Genomics* 9: 104.

Bollback JP. (2002) Bayesian model adequacy and choice in phylogenetics. *Mol Biol Evol* 19.

Both L, Collins S, de Zoysa A, et al. (2015) Molecular and Epidemiological Review of Toxigenic Diphtheria Infections in England between 2007 and 2013. *Journal of Clinical Microbiology* 53: 567-572.

Boto L. (2010) Horizontal gene transfer in evolution: facts and challenges. *Proc R Soc B Biol Sci* 277.

Bouckaert R, Heled J, Kühnert D, et al. (2014) BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* 10.

Boussau B, Guéguen L and Gouy M. (2008) Accounting for horizontal gene transfers explains conflicting hypotheses regarding the position of aquificales in the phylogeny of Bacteria. *BMC Evolutionary Biology* 8: 1-18.

Brendel V, Beckmann JS and Trifonov EN. (1986) Linguistics of nucleotide sequences: morphology and comparison of vocabularies. *Journal of Biomolecular Structure and Dynamics* 4: 11-21.

Brown T, Didelot X, Wilson DJ, et al. (2016) SimBac: simulation of whole bacterial genomes with homologous recombination. *Microbial Genomics* 2.

Brukner I, Sánchez R, Suck D, et al. (1995) Trinucleotide Models for DNA Bending Propensity: Comparison of Models Based on DNaseI Digestion and Nucleosome Packaging Data. *Journal of Biomolecular Structure and Dynamics* 13: 309-317.

Bulmer M. (1991) Use of the method of generalized least squares in reconstructing phylogenies from sequence data. *Mol Biol Evol* 8.

Burkovski A. (2008) *Corynebacteria: Genomics and Molecular Biology*: Caister Academic Press.

Castellini A, Franco G and Manca V. (2012) A dictionary based informational genome analysis. *BMC Genomics* 13: 485-485.

Castresana J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 17.

Cawley GC and Talbot NL. (2006) Gene selection in cancer classification using sparse logistic regression with Bayesian regularization. *Bioinformatics* 22.

Chai J, Kora G, Ahn T-H, et al. (2014) Functional phylogenomics analysis of bacteria and archaea using consistent genome annotation with UniFam. *BMC Evolutionary Biology* 14: 1-13.

Chan CX and Ragan MA. (2013) Next-generation phylogenomics. *Biology Direct* 8: 1-6.

Chen X, Zhou L, Tian K, et al. (2013) Metabolic engineering of Escherichia coli: A sustainable industrial platform for bio-based chemical production. *Biotechnology Advances* 31: 1200-1223.

Clarridge JE. (2004) Impact of 16S rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases. *Clinical Microbiology Reviews* 17: 840-862.

Comas I, Coscolla M, Luo T, et al. (2013) Out-of-Africa migration and Neolithic co-expansion of Mycobacterium tuberculosis with modern humans. *Nature genetics* 45: 1176-1182.

Comin M and Verzotto D. (2012) Alignment-free phylogeny of whole genomes using underlying subwords. *Algorithms for Molecular Biology* 7: 34.

Conners SB, Mongodin EF, Johnson MR, et al. (2006) Microbial biochemistry, physiology, and biotechnology of hyperthermophilic Thermotoga species. *FEMS Microbiol Rev* 30: 872-905.

Conte MG, Gaillard S, Droc G, et al. (2008) Phylogenomics of plant genomes: a methodology for genome-wide searches for orthologs in plants. *BMC Genomics* 9: 1-16.

Cribby S, Taylor M and Reid G. (2008) Vaginal Microbiota and the Use of Probiotics. *Interdisciplinary Perspectives on Infectious Diseases* 2008: 256490.

Darling ACE, Mau B, Blattner FR, et al. (2004a) Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. *Genome Research* 14: 1394-1403.

Darling AE, Mau B, Blattner FR, et al. (2004b) GRIL: genome rearrangement and inversion locator. *Bioinformatics* 20: 122-124.

Davenport CF and Tümmler B. (2010) Abundant Oligonucleotides Common to Most Bacteria. *PLOS ONE* 5: e9841.

Dayhoff M, Schwartz R and Orcutt B. (1978) model of evolutionary change in protein. *Atlas Prot Seq Struct* 5.

de Oliveira Martins L, Leal É and Kishino H. (2008) Phylogenetic Detection of Recombination with a Bayesian Prior on the Distance between Trees. *PLOS ONE* 3: e2651.

de Queiroz A and Gatesy J. (2007) The supermatrix approach to systematics. *Trends Ecol Evol* 22.

de Queiroz A, J Donoghue M and Doaoghue Mj Kim J. (2003) *Separate Versus Combined Analysis of Phylogenetic Evidence*.

Delsuc F, Brinkmann H and Philippe H. (2005) Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet* 6.

Dickerman AW. (1998) Generalizing Phylogenetic Parsimony from the Tree to the Forest. *Syst Biol* 47: 414-426.

Drummond AJ and Rambaut A. (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* 7.

Dwivedi B and Gadagkar S. (2009) Phylogenetic inference under varying proportions of indel-induced alignment gaps. *BMC Evol. Biol.* 9: 1471-2148.

Edgar R. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5: 113.

Efron B, Halloran E and Holmes S. (1996) Bootstrap confidence levels for phylogenetic trees. *Proc. Natl Acad. Sci. USA* 93: 7085-7090.

Efron B and Tibshirani R. (1993) An Introduction to the Bootstrap.

el Hassan MA and Calladine CR. (1996) Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA. *J Mol Biol* 259: 95-103.

Elf J, Nilsson D, Tenson T, et al. (2003) Selective Charging of tRNA Isoacceptors Explains Patterns of Codon Usage. *Science* 30.

Elhai J, Liu H and Taton A. (2012) Detection of horizontal transfer of individual genes by anomalous oligomer frequencies. *BMC Genomics* 13: 245.

Escalona M, Rocha S and Posada D. (2016) A comparison of tools for the simulation of genomic next-generation sequencing data. *Nature reviews. Genetics* 17: 459-469.

Faircloth BC, McCormack JE, Crawford NG, et al. (2012) Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst Biol* 61.

Fan H, Ives AR, Surget-Groba Y, et al. (2015) An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data. *BMC Genomics* 16.

Fedorov A, Saxonov S and Gilbert W. (2002) Regularities of context-dependent codon bias in eukaryotic genes. *Nucleic Acids Res* 30: 1192-1197.

Felsenstein J. (1978) Cases in which parsimony and compatibility methods will be positively misleading. *Syst. Zool.* 27: 401-410.

Felsenstein J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* 17.

Felsenstein J. (1985) Phylogenies and the comparative method. *Am Nat.*

Filipski A, Tamura K, Billing-Ross P, et al. (2015) Phylogenetic placement of metagenomic reads using the minimum evolution principle. *BMC Genomics* 16: 1-9.

Finnerty JR, Mazza ME and Jezewski PA. (2009) Domain duplication, divergence, and loss events in vertebrate Msx paralogs reveal phylogenomically informed disease markers. *BMC Evol Biol* 9.

Fitch WM. (1970) Distinguishing homologous from analogous proteins. *Syst Zool* 19: 99-113.

Flombaum P, Gallegos JL, Gordillo RA, et al. (2013) Present and future global distributions of the marine Cyanobacteria Prochlorococcus and Synechococcus. *Proceedings of the National Academy of Sciences of the United States of America* 110: 9824-9829.

Francois O and Mioland C. (2007) Gaussian approximations for phylogenetic branch length statistics under stochastic models of biodiversity. *Math Biosci* 209.

Frandsen PB, Calcott B, Mayer C, et al. (2015) Automatic selection of partitioning schemes for phylogenetic analyses using iterative k-means clustering of site rates. *BMC Evolutionary Biology* 15: 13.

Fu LM and Fu-Liu CS. (2002) Is Mycobacterium tuberculosis a closer relative to Gram-positive or Gram-negative bacterial pathogens? *Tuberculosis (Edinb)* 82: 85-90.

Fuks G, Elgart M, Amir A, et al. (2017) Combining 16S rRNA gene variable regions enables high-resolution microbial community profiling. *bioRxiv*.

Galtier N and Daubin V. (2008) Dealing with incongruence in phylogenomic analyses. *Philosophical Transactions of the Royal Society B: Biological Sciences* 363: 4023-4029.

Ganesan H, Rakitianskaia A, Davenport C, et al. (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.

Girotto S, Pizzi C and Comin M. (2016) MetaProb: accurate metagenomic reads binning based on probabilistic sequence signatures. *Bioinformatics* 32: i567-i575.

Gomila M, Peña A, Mulet M, et al. (2015) Phylogenomics and systematics in Pseudomonas. *Frontiers in Microbiology* 6: 214.

Google.com. (2018) *Features – Google Analytics*. Available at: http://www.google.com/analytics/features/

Gordon SV and Parish T. (2018) Microbe Profile: Mycobacterium tuberculosis: Humanity's deadly microbial foe. *Microbiology* 164: 437-439.

Gori K, Suchan T, Alvarez N, et al. (2016) Clustering Genes of Common Evolutionary History. *Molecular Biology and Evolution* 33: 1590-1605.

Goris J, Konstantinidis KT, Klappenbach JA, et al. (2007) DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *International journal of systematic and evolutionary microbiology* 57: 81-91.

Graumann P. (2012) *Bacillus: Cellular and Molecular Biology (2nd ed.),* University of Freiburg, Germany: Caister Academic Press.

Guindon S, Dufayard JF, Lefort V, et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 59.

Haggerty LS, Martin FJ, Fitzpatrick DA, et al. (2009) Gene and genome trees conflict at many levels. *Philos Trans R Soc Lond B Biol Sci* 364: 2209-2219.

Han N, Qiang Y and Zhang W. (2016) ANItools web: a web tool for fast genome comparison within multiple bacterial strains. *Database: The Journal of Biological Databases and Curation* 2016: baw084.

Hartigan JA. (1973) Minimum evolution fits to a given tree. *Biometrics* 29: 53-65.

Hasegawa M. (1990) Phylogeny and molecular evolution in primates. *The Japanese Journal of Genetics* 65: 243-266.

Hasegawa M, Kishino H and Yano T. (1985) Dating of the human ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* 22: 160 - 174.

Haubold B, Pierstorff N, Möller F, et al. (2005) Genome comparison without alignment using shortest unique substrings. *BMC Bioinformatics* 6: 123.

Henikoff S and Henikoff JG. (1992) Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* 89.

Henz SR, Huson DH, Auch AF, et al. (2005) Whole-genome prokaryotic phylogeny. *Bioinformatics* 21.

Holloway C and Beiko RG. (2010) Assembling networks of microbial genomes using linear programming. *BMC Evolutionary Biology* 10: 360.

Huang WM. (1996) BACTERIAL DIVERSITY BASED ON TYPE II DNA TOPOISOMERASE GENES. *Annual Review of Genetics* 30: 79-107.

Huber R, Langworthy TA, König H, et al. (1986) Thermotoga maritima sp. nov. represents a new genus of unique extremely thermophilic eubacteria growing up to 90°C. *Arch Microbiol* 144: 324-333.

Hug LA, Baker BJ, Anantharaman K, et al. (2016) A new view of the tree of life. 1: 16048.

Huson D and Bryant D. (2006) Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* 23: 254 - 267.

Ikemura T. (1981) Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. *J Mol Biol* 146.

Janda JM and Abbott SL. (2007) 16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls. *Journal of Clinical Microbiology* 45: 2761-2764.

Jeffroy O, Brinkmann H, Delsuc F, et al. (2006) Phylogenomics: the beginning of incongruence? *Trends Genet* 22.

Jolley KA, Bliss CM, Bennett JS, et al. (2012) Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology* 158: 1005-1015.

Jones DT, Taylor WR and Thornton JM. (1992) The rapid generation of mutation data matrices from protein sequences. *CABIOS* 8.

Jukes T and Cantor C. (1969) Evolution of protein molecules. In: Munro H (ed) *Mamm Protein Metab. Academy Press.*

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

Kass RE and Raftery AE. (1995) Bayes factors. *J Am Stat Assoc* 90.

Katz LS, Griswold T, Williams-Newkirk AJ, et al. (2017) A Comparative Analysis of the Lyve-SET Phylogenomics Pipeline for Genomic Epidemiology of Foodborne Pathogens. *Frontiers in Microbiology* 8: 375.

Kimura M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* 16: 111 - 120.

Klockgether J, Würdemann D, Reva O, et al. (2007) Diversity of the Abundant pKLC102/PAGI-2 Family of Genomic Islands in Pseudomonas aeruginosa. *Journal of Bacteriology* 189: 2443-2459.

Koonin EV, Puigbo P and Wolf YI. (2011) Comparison of phylogenetic trees and search for a central trend in the "forest of life". *J Comput Biol* 18.

Kosakovsky Pond SL, Frost SDW and Muse SV. (2005) HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21: 676-679.

Kuhner MK and Felsenstein J. (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol Biol Evol* 11.

Kullback S and Leibler RA. (1951) On information and sufficiency. *Annals Math Stat* 22.

Kumar S, Filipski AJ, Battistuzzi FU, et al. (2012) Statistics and truth in phylogenomics. *Mol Biol Evol* 29.

Kunal, Rajor A and Siddique R. (2016) Bacterial treatment of alkaline cement kiln dust using Bacillus halodurans strain KG1. *Brazilian Journal of Microbiology* 47: 1-9.

Kvitek DJ and Sherlock G. (2013) Whole genome, whole population sequencing reveals that loss of signaling networks is the major adaptive strategy in a constant environment. *PLoS Genet* 9: E10038972.

Kyrpides NC, Hugenholtz P, Eisen JA, et al. (2014) Genomic Encyclopedia of Bacteria and Archaea: Sequencing a Myriad of Type Strains. *PLOS Biology* 12: e1001920.

Lapierre P and Gogarten JP. (2009) Estimating the size of the bacterial pan-genome. *Trends Genet* 25: 107-110.

Lawrence JG and Ochman H. (1997) Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol* 44.

Lee SY. (1996) High cell-density culture of Escherichia coli. *Trends Biotechnol* 14: 98-105.

Lemmon AR and Moriarty EC. (2004) The importance of proper model assumption in Bayesian phylogenetics. *Syst Biol* 53.

Li H and Durbin R. (2011) Inference of human population history from individual whole-genome sequences. *Nature* 475: 493-496.

Lin GN, Zhang C and Xu D. (2011) Polytomy identification in microbial phylogenetic reconstruction. *BMC Systems Biology* 5: 1-11.

Lister PD, Wolter DJ and Hanson ND. (2009) Antibacterial-Resistant Pseudomonas aeruginosa: Clinical Impact and Complex Regulation of Chromosomally Encoded Resistance Mechanisms. *Clinical Microbiology Reviews* 22: 582-610.

Maiden MCJ, Jansen van Rensburg MJ, Bray JE, et al. (2013) MLST revisited: the gene-by-gene approach to bacterial genomics. *Nature reviews. Microbiology* 11: 728-736.

Makarenkov V, Boc A, Xie J, et al. (2010) Weighted bootstrapping: a correction method for assessing the robustness of phylogenetic trees. *BMC Evolutionary Biology* 10: 250-250.

Makarova K, Slesarev A, Wolf Y, et al. (2006) Comparative genomics of the lactic acid bacteria. *Proceedings of the National Academy of Sciences* 103: 15611-15616.

Mann EE and Wozniak DJ. (2012) Pseudomonas biofilm matrix composition and niche biology. *FEMS Microbiology Reviews* 36: 893-916.

Marcelletti S and Scortichini M. (2014) Definition of Plant-Pathogenic Pseudomonas Genomospecies of the Pseudomonas syringae Complex Through Multiple Comparative Approaches. *Phytopathology* 104: 1274-1282.

Marques S and Ramos JL. (1993) Transcriptional control of the Pseudomonas putida TOL plasmid catabolic pathways. *Mol Microbiol* 9: 923-929.

Marquez R, Smit S and Knight R. (2005) Do universal codon-usage patterns minimize the effects of mutation and translation error? *Genome Biol* 6: R91.

Martin R, Miquel S, Ulmer J, et al. (2013) Role of commensal and probiotic bacteria in human health: a focus on inflammatory bowel disease. *Microb Cell Fact* 12: 71.

Martini M, Lee IM, Bottner KD, et al. (2007) Ribosomal protein gene-based phylogeny for finer differentiation and classification of phytoplasmas. *Int J Syst Evol Microbiol* 57: 2037-2051.

McCormack JE, Faircloth BC, Crawford NG, et al. (2012) Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. *Genome Res* 22.

Meier-Kolthoff JP, Auch AF, Klenk H-P, et al. (2013) Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* 14: 60.

Meier B, Cooke SL, Weiss J, et al. (2014) C. elegans whole-genome sequencing reveals mutational signatures related to carcinogens and DNA repair deficiency. *Genome Research* 24: 1624-1636.

Menard A, Buissonniere A, Prouzet-Mauleon V, et al. (2016) The GyrA encoded gene: A pertinent marker for the phylogenetic revision of Helicobacter genus. *Syst Appl Microbiol* 39: 77-87.

Meurer A, Smith CP, Paprocki M, et al. (2016) SymPy: Symbolic computing in Python. *PeerJ Preprints* 4: e2083v2083.

Miller RA, Jian J, Beno SM, et al. (2018) Intraclade Variability in Toxin Production and Cytotoxicity of Bacillus cereus Group Type Strains and Dairy-Associated Isolates. *Applied and Environmental Microbiology* 84: e02479-02417.

Minin V, Abdo Z, Joyce P, et al. (2003) Performance-based selection of likelihood models for phylogeny estimation. *Syst Biol* 52.

Misof B, Meusemann K, von Reumont BM, et al. (2014) A priori assessment of data quality in molecular phylogenetics. *Algorithms for Molecular Biology* 9: 1-8.

Moldovan MA and Gelfand MS. (2018) Pangenomic Definition of Prokaryotic Species and the Phylogenetic Structure of Prochlorococcus spp. *Frontiers in Microbiology* 9.

Mongodin E, Hance I, Deboy R, et al. (2005) Gene transfer and genome plasticity in Thermotoga maritima, a model hyperthermophilic species. *J Bacteriol* 187: 4935 - 4944.

Nature. (2018) *Phylogenomics*. Available at: https://www.nature.com/subjects/phylogenomics.

NCBI. (2018) *GenBank and WGS Statistics.* Available at: https://www.ncbi.nlm.nih.gov/genbank/statistics/.

Nelson K, Clayton R, Gill S, et al. (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. *Nature* 399: 323 - 329.

Nesbo CL, Dlutek M and Doolittle WF. (2006) Recombination in Thermotoga: implications for species concepts and biogeography. *Genetics* 172: 759-769.

Newville M, Stensitzki T, Allen DB, et al. (2014) LMFIT: Non-Linear Least-Square Minimization and Curve-Fitting for Python.

Ogden TH and Rosenberg M. (2006) Multiple sequence alignment accuracy and phylogenetic inference. *Syst Biol* 55.

Oliphant T. (2006) *Guide to NumPy*.

Oliphant T. (2007) *Python for Scientific Computing*.

Olson WK, Gorin AA, Lu XJ, et al. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc Natl Acad Sci U S A* 95: 11163-11168.

Ondov BD, Treangen TJ, Melsted P, et al. (2016) Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology* 17: 132.

Ornstein RL and Rein R. (1978) An optimized potential function for the calculation of nucleic acid interaction energies I. base stacking. *Biopolymers* 17: 2341-2360.

Ounit R, Wanamaker S, Close TJ, et al. (2015) Clark: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics* 16.

Partensky F, Hess WR and Vaulot D. (1999) Prochlorococcus, a Marine Photosynthetic Prokaryote of Global Significance. *Microbiology and Molecular Biology Reviews* 63: 106-127.

Pascual C, Lawson PA, Farrow JA, et al. (1995) Phylogenetic analysis of the genus Corynebacterium based on 16S rRNA gene sequences. *Int J Syst Bacteriol* 45: 724-728.

Pennacchio LA, Ahituv N, Moses AM, et al. (2006) In vivo enhancer analysis of human conserved non-coding sequences. *Nature* 444: 499-502.

Phillips GJ, Arnold J and Ivarie R. (1987) The effect of codon usage on the oligonucleotide composition of the E.coli genome and identification of over-and underepresented sequences by Markow chain analysis. *Nucleic Acids Research* 15: 2627-2638.

Pierneef R, Cronje L, Bezuidt O, et al. (2015) Pre_GI: a global map of ontological links between horizontally transferred genomic islands in bacterial and archaeal genomes. *Database* 2015: bav058-bav058.

Poptsova MS and Gogarten JP. (2007) The power of phylogenetic approaches to detect horizontally transferred genes. *BMC Evolutionary Biology* 7: 45.

Posada D and Crandall K. (1998) MODELTEST: testing the model of DNA substitution. *Bioinformatics* 14.

Posada D and Crandall KA. (2001) Selecting the best-fit model of nucleotide substitution. *Syst Biol* 50.

Prabha R, Singh DP, Gupta SK, et al. (2014) Whole genome phylogeny of Prochlorococcus marinus group of cyanobacteria: genome alignment and overlapping gene approach. *Interdisciplinary sciences, computational life sciences* 6: 149.

Pride D and Blaser M. (2002) Identification of horizontally acquired elements in Helicobacter pylori and other prokaryotes using oligonucleotide difference analysis. *Genome Let* 1: 2 - 15.

Pride D, Meinersmann R, Wassenaar T, et al. (2003) Evolutionary implications of microbial genome tetanucleotide frequency biases. *Genome Res* 13: 145 - 155.

Puigbò P, Wolf YI and Koonin EV. (2013) Seeing the Tree of Life behind the phylogenetic forest. *BMC Biology* 11: 1-3.

Qi J, Luo H and Hao B. (2004a) CVTree: a phylogenetic tree reconstruction tool based on whole genomes. *Nucleic Acids Research* 32: 45 - 47.

Qi J, Wang B and Hao B-I. (2004b) Whole Proteome Prokaryote Phylogeny Without Sequence Alignment: A K-String Composition Approach. *Journal of Molecular Evolution* 58: 1-11.

Rajendhran J and Gunasekaran P. (2011) Microbial phylogeny and diversity: small subunit ribosomal RNA sequence analysis and beyond. *Microbiol Res* 166: 99-110.

Reva O and Tummler 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 and Tummler B. (2005) Differentiation of regions with atypical oligonucleotide composition in bacterial genomes. *BMC Bioinformatics* 6: 251.

Reva O and Tummler B. (2008) Oligonucleotide usage signatures of the Pseudomonas putida KT2440 genome. *Pseudomonas: Genomics and Molecular Biology* Chapter 3: 43 - 64.

Richard A. Redak, Alexander H. Purcell, João R.S. Lopes, et al. (2004) THE BIOLOGY OF XYLEM FLUID–FEEDING INSECT VECTORS OF XYLELLA FASTIDIOSA AND THEIR RELATION TO DISEASE EPIDEMIOLOGY. *Annual Review of Entomology* 49: 243-270.

Richter DC, Ott F, Auch AF, et al. (2008) MetaSim: a sequencing simulator for genomics and metagenomics. *PLOS ONE* 3.

Richter M and Rosselló-Móra R. (2009) Shifting the genomic gold standard for the prokaryotic species definition. *Proceedings of the National Academy of Sciences* 106: 19126-19131.

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

Ronquist F, Teslenko M, Mark P, et al. (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol* 61.

Ruggiero P. (2014) Use of probiotics in the fight against Helicobacter pylori. *World Journal of Gastrointestinal Pathophysiology* 5: 384-391.

Saitou N and Nei M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4: 406-425.

Satchwell SC, Drew HR and Travers AA. (1986) Sequence periodicities in chicken nucleosome core DNA. *J Mol Biol* 191: 659-675.

Schwarz G. (1978) Estimating the dimension of a model. *Ann Stat* 6.

Scornavacca C. (2009) *Supertree methods for phylogenomics*.

Sepulveda-Torres LC, Rajendran N, Dybas MJ, et al. (1999) Generation and initial characterization of Pseudomonas stutzeri KC mutants with impaired ability to degrade carbon tetrachloride. *Arch Microbiol* 171: 424-429.

Shah P and Gilchrist MA. (2010) Effect of Correlated tRNA Abundances on Translation Errors and Evolution of Codon Usage Bias. *PLOS Genetics* 6: e1001128.

Shapiro B, Rambaut A and Drummond AJ. (2006) Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Mol Biol Evol* 23.

Shapiro BJ, Leducq J-B and Mallet J. (2016) What Is Speciation? *PLOS Genetics* 12: e1005860.

Shapiro BJ and Polz MF. (2014) Ordering microbial diversity into ecologically and genetically cohesive units. *Trends in microbiology* 22: 235-247.

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

Silby MW, Winstanley C, Godfrey SAC, et al. (2011) Pseudomonas genomes: diverse and adaptable. *FEMS Microbiology Reviews* 35: 652-680.

Sims GE, Jun S-R, Wu GA, et al. (2009) Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proceedings of the National Academy of Sciences* 106: 2677-2682.

Soltis PS and Soltis DE. (2003) Applying the bootstrap in phylogeny reconstruction. *Statistical Science*: 256-267.

Soni A, Oey I, Silcock P, et al. (2016) *Bacillus Spores in the Food Industry: A Review on Resistance and Response to Novel Inactivation Technologies*.

Stecher B, Maier L and Hardt W-D. (2013) 'Blooming' in the gut: how dysbiosis might contribute to pathogen evolution. *Nat Rev Micro* 11: 277-284.

Stevens V, Thijs S, McAmmond B, et al. (2017) Draft Genome Sequence of Bacillus licheniformis VSD4, a Diesel Fuel–Degrading and Plant Growth–Promoting Phyllospheric Bacterium. *Genome Announcements* 5: e00027-00017.

Sueoka N. (1988) Directional mutation pressure and neutral molecular evolution. *Proc Natl Acad Sci U S A* 85: 2653 - 2657.

Swenson KM and El-Mabrouk N. (2012) Gene trees and species trees: irreconcilable differences. *BMC Bioinformatics* 13: S15.

Swenson MS, Suri R, Linder CR, et al. (2012) SuperFine: fast and accurate supertree estimation. *Syst Biol* 61.

Swofford DL, Waddell PJ, Huelsenbeck JP, et al. (2001) Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Syst Biol* 50: 525-539.

Takahashi K, Terai Y, Nishida M, et al. (2001) Phylogenetic relationships and ancient incomplete lineage sorting among cichlid fishes in Lake Tanganyika as revealed by analysis of the insertion of retroposons. *Mol Biol Evol* 18.

Takahashi M, Kryukov K and Saitou N. (2009) Estimation of bacterial species phylogeny through oligonucleotide frequency distances. *Genomics* 93: 525-533.

Tamura K, Stecher G, Peterson D, et al. (2013) MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Molecular Biology and Evolution* 30: 2725-2729.

Tavaré S. (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences* 17.

Thompson JD, Higgins DG and Gibson TJ. (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.

Thorne JL, Kishino H and Painter IS. (1998) Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution* 15.

Tortoli E, Fedrizzi T, Meehan CJ, et al. (2017) The new phylogeny of the genus Mycobacterium: The old and the news. *Infection, Genetics and Evolution* 56: 19-25.

Tran NH and Chen X. (2014) Comparison of next-generation sequencing samples using compression-based distances and its application to phylogenetic reconstruction. *BMC Research Notes* 7: 1-13.

Tran Q, Pham D-T and Phan V. (2017) Using 16S rRNA gene as marker to detect unknown bacteria in microbial communities. *BMC Bioinformatics* 18: 499.

Trifonov EN. (1989) The multiple codes of nucleotide sequences. *Bull Math Biol* 51: 417-432.

Tuimala J. (2006) *A primer to phylogenetic analysis using the PHYLIP package*.

Vertes AA, Inui M and Yukawa H. (2012) Postgenomic approaches to using corynebacteria as biocatalysts. *Annu Rev Microbiol* 66: 521-550.

Wayne LG, Brenner DJ, Colwell RR, et al. (1987) Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics. *International journal of systematic and evolutionary microbiology* 37: 463-464.

Wheeler WC, Whiting M, Wheeler QD, et al. (2001) The phylogeny of the extant hexapod orders. *Cladistics* 17: 113-169.

Williams KP, Gillespie JJ, Sobral BWS, et al. (2010) Phylogeny of Gammaproteobacteria. *Journal of Bacteriology* 192: 2305-2314.

Wood DE and Salzberg SL. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 15.

Woolfe A, Goodson M, Goode DK, et al. (2005) Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol* 3: e7.

Woolley SM, Posada D and Crandall KA. (2008) A comparison of phylogenetic network methods using computer simulation. *PLOS ONE* 3.

Xu D and Cote JC. (2003) Phylogenetic relationships between Bacillus species and related genera inferred from comparison of 3' end 16S rDNA and 5' end 16S-23S ITS nucleotide sequences. *Int J Syst Evol Microbiol* 53: 695-704.

Xu Z and Hao B. (2009) CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes. *Nucleic Acids Res* 37.

Yang Z. (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites. *J Mol Evol* 39.

Yang Z. (1996) Phylogenetic analysis using parsimony and likelihood methods. *J. Mol. Evol.* 42: 294-307.

Yang Z and Rannala B. (1997) Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo Method. *Mol. Biol. Evol.* 14: 717-724.

Yi H and Jin L. (2013) Co-phylog: an assembly-free phylogenomic approach for closely related organisms. *Nucleic Acid Res* 41.

Yu X. (2014) Mathematical simulations and Verhulst modeling of compositional changes in DNA sequences of acquired genomic islands due to bacterial genome amelioration. *Department of Biochemistry.* Pretoria: University of Pretoria South Africa.

Yutin N, Puigbò P, Koonin EV, et al. (2012) Phylogenomics of prokaryotic ribosomal proteins. *PLOS ONE* 7.

Zhang W, Du P, Zheng H, et al. (2014) Whole-genome sequence comparison as a method for improving bacterial species definition. *J Gen Appl Microbiol* 60: 75-78.

Zhaxybayeva O, Gogarten J, Charlebois R, et al. (2006) Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. *Genome Res* 16: 1099 - 1108.

Zhaxybayeva O and Gogarten JP. (2002) Bootstrap, Bayesian probability and maximum likelihood mapping: Exploring new tools for comparative genome analyses. *BMC Genomics* 3.

Appendix

Tables

Supplementary Table 1. Set of OUP and gyrA distances

Mycobacteria			
Genome pairs		**Distances**	
First	**Second**	**OUP**	**gyrA**
NC_008595	NC_002944	0.665806533	0.001201
NC_008595	NC_008726	3.209024198	0.077785
NC_008595	NC_016947	1.568659733	0.041694
NC_008595	NC_017904	1.610190176	0.041694
NC_008595	NC_009077	4.224719084	0.079335
NC_008595	NC_008705	4.260986489	0.079335
NC_008595	NC_008146	4.276542805	0.079335
NC_008595	NC_008596	4.243947825	0.064998
NC_008595	NC_009338	4.136348027	0.084246
NC_008595	NC_010397	4.057807912	0.081753
NC_008595	NC_008611	3.847747461	0.056726
NC_008595	NC_000962	3.927238535	0.04687
NC_008595	NC_010612	4.053708938	0.049214
NC_008595	NC_015576	3.416828794	0.077503
NC_008595	NC_002677	10.09020518	0.08197
NC_002944	NC_008726	3.297347049	0.079055
NC_002944	NC_016947	1.716884641	0.042929
NC_002944	NC_017904	1.75577573	0.042929
NC_002944	NC_009077	4.292725269	0.080609
NC_002944	NC_008705	4.324908552	0.080609
NC_002944	NC_008146	4.338737834	0.080609
NC_002944	NC_008596	4.284397722	0.066261
NC_002944	NC_009338	4.22724658	0.085521
NC_002944	NC_010397	4.120706809	0.083027
NC_002944	NC_008611	3.8885674	0.057971
NC_002944	NC_000962	3.926570194	0.048112
NC_002944	NC_010612	3.943819334	0.050454
NC_002944	NC_015576	3.59922179	0.078769
NC_002944	NC_002677	10.04630089	0.083246
NC_008726	NC_016947	3.513567106	0.074113

NC_008726	NC_017904	3.57569983	0.074113
NC_008726	NC_009077	1.97004778	0.05225
NC_008726	NC_008705	1.983632828	0.05225
NC_008726	NC_008146	2.023478975	0.05225
NC_008726	NC_008596	2.621277122	0.062333
NC_008726	NC_009338	1.754458606	0.021915
NC_008726	NC_010397	4.057681939	0.084772
NC_008726	NC_008611	5.555247862	0.076564
NC_008726	NC_000962	5.574774412	0.082058
NC_008726	NC_010612	5.590732208	0.068961
NC_008726	NC_015576	4.024805447	0.079561
NC_008726	NC_002677	11.18099224	0.112247
NC_016947	NC_017904	0.479499149	0.00001
NC_016947	NC_009077	4.386617875	0.080797
NC_016947	NC_008705	4.396083059	0.080797
NC_016947	NC_008146	4.427956218	0.080797
NC_016947	NC_008596	4.218106143	0.086047
NC_016947	NC_009338	4.402520516	0.075567
NC_016947	NC_010397	3.8198891	0.085914
NC_016947	NC_008611	3.92310232	0.059485
NC_016947	NC_000962	4.11627384	0.057062
NC_016947	NC_010612	4.061290385	0.053146
NC_016947	NC_015576	4.590223735	0.07646
NC_016947	NC_002677	10.40694208	0.085569
NC_017904	NC_009077	4.434030815	0.080797
NC_017904	NC_008705	4.433977475	0.080797
NC_017904	NC_008146	4.467382262	0.080797
NC_017904	NC_008596	4.29573671	0.086047
NC_017904	NC_009338	4.399899499	0.075567
NC_017904	NC_010397	3.746122263	0.085914
NC_017904	NC_008611	3.801846989	0.059485
NC_017904	NC_000962	4.067843166	0.057062
NC_017904	NC_010612	4.019193377	0.053146
NC_017904	NC_015576	4.748297665	0.07646
NC_017904	NC_002677	10.3441859	0.085569
NC_009077	NC_008705	0.446339277	0.00001
NC_009077	NC_008146	0.387467709	0.00001
NC_009077	NC_008596	2.397088383	0.052081
NC_009077	NC_009338	1.897713983	0.05457
NC_009077	NC_010397	4.815238669	0.100899
NC_009077	NC_008611	6.53942068	0.071724

NC_009077	NC_000962	6.475020314	0.082697
NC_009077	NC_010612	6.605050622	0.062715
NC_009077	NC_015576	5.350194553	0.07989
NC_009077	NC_002677	11.58945075	0.103843
NC_008705	NC_008146	0.151288412	0.00001
NC_008705	NC_008596	2.494766862	0.052081
NC_008705	NC_009338	1.831377315	0.05457
NC_008705	NC_010397	4.861453283	0.100899
NC_008705	NC_008611	6.583725672	0.071724
NC_008705	NC_000962	6.550882574	0.082697
NC_008705	NC_010612	6.683500145	0.062715
NC_008705	NC_015576	5.490029183	0.07989
NC_008705	NC_002677	11.65761593	0.103843
NC_008146	NC_008596	2.516713541	0.052081
NC_008146	NC_009338	1.869812975	0.05457
NC_008146	NC_010397	4.861608648	0.100899
NC_008146	NC_008611	6.601109452	0.071724
NC_008146	NC_000962	6.568364892	0.082697
NC_008146	NC_010612	6.670364245	0.062715
NC_008146	NC_015576	5.508268482	0.07989
NC_008146	NC_002677	11.68435512	0.103843
NC_008596	NC_009338	2.878013027	0.072288
NC_008596	NC_010397	4.308207712	0.099147
NC_008596	NC_008611	6.068043485	0.072757
NC_008596	NC_000962	5.967844949	0.08833
NC_008596	NC_010612	6.092065823	0.063768
NC_008596	NC_015576	5.763618677	0.087466
NC_008596	NC_002677	10.55479737	0.113936
NC_009338	NC_010397	4.961909354	0.08875
NC_009338	NC_008611	6.66991587	0.083031
NC_009338	NC_000962	6.652277231	0.084694
NC_009338	NC_010612	6.729642334	0.075402
NC_009338	NC_015576	5.79401751	0.073232
NC_009338	NC_002677	11.9627897	0.117513
NC_010397	NC_008611	4.053789866	0.094044
NC_010397	NC_000962	4.238257848	0.098204
NC_010397	NC_010612	4.142067482	0.086253
NC_010397	NC_015576	6.043287938	0.095772
NC_010397	NC_002677	9.609943454	0.125304
NC_008611	NC_000962	2.004876867	0.064578
NC_008611	NC_010612	0.984987099	0.010841

NC_008611	NC_015576	5.745379377	0.080236
NC_008611	NC_002677	8.182696848	0.094579
NC_000962	NC_010612	1.873921402	0.059467
NC_000962	NC_015576	6.140564202	0.081868
NC_000962	NC_002677	7.552029005	0.084253
NC_010612	NC_015576	6.055447471	0.072543
NC_010612	NC_002677	8.425247904	0.085447
NC_015576	NC_002677	9.900110695	0.108385