Classification of bacteria using oligonucleotide microarray: an *in silico* experiment

Are Klevan

Thesis for the degree of Siv.ing.
in Biotechnology.

Department of Molecular Biosciences
Department of biology
University of Oslo

2004

1 Acknowledgements

The work presented in this study was carried out at the Division of Molecular Bioscience, Department of biology, at the University of Oslo, from January 2002 to March 2004.

I want to thank my supervisor, Assistant Professor William Davies, for giving me the opportunity to do the work presented in this thesis, and for the support and expertise he provided along the way. I would also like to thank Simen Gaure and Andreas Botnen for collaboration and development of the programs used in this thesis, Dr. Kamran Shalchian-Tabrizi for his help and expertise in constructing phylogenetic trees, Assistant Professor Ole Christian Lingjærde for help with clustering of our data, and the crew at the computer department for drifting our servers.

I would of course like to thank my friends at 3525 for making my stay here unforgettable and my parents for their constant support and encouragement.

Finally a special thank to my fiancée and son for their unconditional love and support they always have shown me.

Oslo, March 2004

_____

Are Klevan

2 Summary

For evolutionary and medical reasons bacterial classification is an important field within microbiology. Before Carl Woese introduced the use of ribosomal RNA sequences for phylogenetic comparison, bacterial classification was based on different phenotypic methods. Today the primary center of attention is focused on making super trees (phylogenetic trees generated from multiple genes) and doing whole genome comparison. Still, problems resulting from non-orthogonal gene replacement and interference by lateral gene transfer make this matter far from trivial.

This study is based on the classification of bacteria using the distribution and frequency of selected 10-mer oligonucleotides in complete genome sequences. These frequencies will be detected by an oligonucleotide microarray and the occurring pattern will be compared to a reference in order to classify a particular organism. In this way it will be possible to compare many bacterial genomes with each other and organize them according to their pattern. Prior to this thesis a set of programs for extraction of informative oligonucleotides from genome sequence data, based on their entropy, have been developed. This study aims to evaluate this method using an *in silico* approach.

Different sub-sets of bacterial genome sequences were used to select sets of informative 10-mer oligonucleotides. In order the test this method a program simulating a microarray was written, such that a suitable output for further analysis was generated. 10-mer oligonucleotide frequencies from the genomes that are to be classified were computed and combined with a set of informative oligonucleotides, in the virtual microarray program. The output from this application was later used in construction of Dendrograms, using the microarray analysis program J-Express. These dendrograms were compared by visual inspection to phylogenetic reference trees made by conventional methods. The phylogenetic analysis was conducted on sequences encoding the 16S rRNA genes, the ATP synthase alpha chain, the prolyl-tRNA synthetase and the methionyl-tRNA synthetase. Our results indicate that the method obtains excellent resolution for discriminating bacteria at the species and strain levels, but not particularly good at the genus level.

Contents

3 Introduction

Bacterial classification has always been a major issue in microbiology. It allows us to see relationships between different microorganisms and to develop a more reasonable taxonomy. Classification is the part of taxonomy concerned with the grouping of bacterial species into taxa based on different characteristics. Classification can be divided into natural or artificial. Natural classification seeks to find evolutionary relatedness based on sequence similarities, while artificial systems are based upon expressed characters such as an organisms phenotype. Until the mid seventies no reasonable method to determine microbial relatedness and evolution were established, thus all bacterial classifications were artificial. In 1965 it was suggested that sequences from conserved macromolecules, such as rRNA, DNA or proteins could be used to reflect evolutionary relationship between organisms (Zuckerkandl and Pauling, 1965). More than ten years later the first phylogenetic trees made from 16S rRNA comparison were published. These trees provided important clues about relatedness, not only between prokaryotes, but to higher organisms as well (Woese and Fox, 1977). In the last few years an ever increasing number of genomes have been completely sequenced and whole genome comparison has been conducted between several different species. It is still is important to remember that genotype and phenotype are closely related and that they both should be accounted for in the field of classification.

The aim of this study is to establish a method, using oligonucleotides, for bacterial classification and to compare these results with already established methods. This introduction will begin by taking a glance at some conventional methods in bacterial classification followed by a broader discussion of more recent methods such as comparative genomics, phylogenetics, microarrays and clustering analysis.

3.1 Phenotypic classification of bacteria

The backbone of phenotypic classification is made up of different methods to determine morphology and biochemical properties, some of these methods are more than 100 years old and are still in use. Morphology, determined by light-microscopy, reveals characteristics such as size, shape and Gram-staining. To determine physiological and nutritional properties, a wide range of biochemical tests have been developed, which now are available in kits. In essence these kits are used to determine growth on particular substrates and/or to detect the production of particular metabolites under defined physiological conditions (Madigan *et al*., 2003).

The mechanisms of movement are also of interest, by flagella, by gliding, by gas vesicles or if the bacteria are non-motile. Further, tolerance to different antibiotics and the presence of specific surface antigens are widely used for identification in clinical diagnostic microbiology. Due to the diverse range of lipid compounds found in the bacterial cell membrane, methods for chemotaxonomic analysis of the outer and inner membrane have been developed. To a certain degree the cell wall is also suitable as a phylogenetic marker (Lengeler *et al*., 1999)

3.2 Genotypic classification of bacteria

3.2.1 GC ratios

The base composition of DNA, expressed in mol% G+C, varies with values ranging from 24 to 76 mol% G+C (Lengeler *et al*., 1999). The GC content can only be looked upon as an indication of relatedness, since closely related species should have approximately the same GC ratio, and distantly related species should have different GC content. Although the GC content is identical, the actual DNA sequence may be significantly different; as a result this method can only be used to exclude relatedness. GC contents is also being used as an indication of lateral gene transfer (LGT) since DNA acquired from distantly related species can have a significantly different GC ratio (Lawrence and Ochman, 1997).

3.2.2 DNA/DNA hybridization

If two organisms have a high sequence similarity, they probably also share highly similar genes and their DNA strands are likely to hybridize to one another in proportion to the similarities in their genes. DNA::DNA hybridization was the first comparative method to be used that gave specific values which could be used in a quantitative manner. As a result the method gives an indication of the degree of relatedness between two bacteria. Bacteria belonging to the same species are said to show a hybridization value above 60-70 %.

3.2.3 Fingerprinting techniques

The use of modern techniques to determine the degree of sequence conservation between bacterial genomes has lead to methods for detection of natural polymorphism. These techniques employs the usage of restriction enzymes, PCR or both, in order to distinguish between different organisms, based on their DNA sequence. Restriction enzymes are used in order to detect "restriction fragments length polymorphisms" (RFLP), which may be used as a tool in bacterial taxonomy. Originally, Southern hybridization (Southern, 1975) was used to type RFLPs, but today other techniques are more commonly used. One such method is the "amplified fragment length polymorphism" (AFLP) technique, which combines the usage of specific PCR amplification and treatment with restriction enzymes (Janssen *et al*., 1996). Another method for typing polymorphisms is the "random amplified polymorphic DNA" (RAPD) fingerprinting technique, which is a strictly PCR based method (Welsh and McClelland, 1990). These methods all have resolution at the strain level.

3.2.4 Ribotyping

Ribotyping is based on comparing the unique patterns generated when DNA from a particular organism is treated with restriction enzymes. The original method (Grimont and Grimont, 1986) is based on treating bacterial DNA with different restriction endonucleases, followed by separation using eletrophoresis. Fragments on the gel are transferred to a nylon filter and finally the DNA fragments carrying rRNA genes (rDNA) will be localized by hybridization with a labeled rRNA probe, analogous to Southern hybridization (Southern, 1975). The pattern obtained from the hybridized

fragments will then be compared between different organisms. In the new method the whole RNA operon is PCR amplified using specific fluorochrome labeled primers. The DNA product is treated with restriction enzymes, an finally separated by electrophoresis (Kostman *et al*., 1992). This reveals a pattern that is unique within a species, and can be compared to other patterns. The PCR based method is technically less demanding than the original one since there is no need for probing and hybridization.

3.2.5 Ribosomal RNA analyses

In the early 1970s Carl Woese introduced a method based on sequence analysis of the 16S ribosomal RNA molecule (Woese *et al*., 1975). He used the 16S sequences from different organisms to determine their phylogenetic relations, not only for prokaryotes, but for all living organisms. Today specific PCR amplification provides easy access to rRNA genes for sequencing. Since the 16S rRNA molecule has many regions that are highly conserved, a small set of PCR primers can be used to analyze a wide range of phylogeneticaly diverse organisms. Similar analysis has also been conducted on the 5S ribosomal RNA molecule, although it gives less information because of its limited size, and the 23S molecule which is approximately twice as large as the 16S molecule. However several findings suggest that the ribosomal operon has been subject to lateral gene transfer, which may give an incorrect evolutionary picture (Brochier *et al*., 2000).

3.2.6 Phylogenetic classification using conserved genes

In addition to rRNA, other conserved ubiquitous genes such as ATPase, DNA/RNA polymerase and elongation factors have been used in phylogenetic classification (Daubin *et al*., 2001; Gogarten *et al*., 1992). Due to the great diversity that exists among prokaryotes, finding genes common to all species is not a trivial manner. Never the less, it seems that conserved genes involved in translation, transcription, ATP synthesis/repair are present in nearly all species, but there are exceptions. This is an important field and hopefully it will give us a more complete phylogenetic classification. Influence by lateral gene transfer (LGT) and the introduction of new genes into an organism is a problem when constructing a reliable phylogenetic tree (Brown and Doolittle, 1997). If LGT cannot be limited to special categories of genes the basis for constructing a natural tree of life is eliminated, and that the tree of life may be

irresolvable (Doolittle, 1999; Martin, 1999). However findings suggest that informational genes are less frequently transferred than operational genes (Jain *et al*., 1999), nevertheless LGT has also been detected in some of these genes (Brochier *et al*., 2000). Informational genes are genes involved in transcription, translation, and related processes, while operational genes are more commonly referred to as housekeeping genes.

3.2.7 DNA arrays

Since the ultimate goal of this study is the construction a microarray for bacterial classification, a broad introduction will be given to microarrays. The basis for DNA arrays is hybridization between nucleic acids, as is the case with many other DNA based detection methods. On a single DNA microarray, thousands of single stranded cDNA molecules or oligonucleotides are attached to discrete regions on the same surface, measuring only a few square centimeters. Since this technology has the ability to detect tens of thousands of hybridizations in a single experiment, it is being referred to as a high through put method. It has proven to be extremely efficient, especially in gene expression experiments (Lockhart *et al*., 1996; Schena *et al*., 1995), detection of polymorphisms (Wang *et al*., 1998), and comparison of closely related species, e.g. when hybridizing *Bacillus cereus* to a *Bacillus anthracis* DNA microarray (Read *et al*., 2003).

3.2.7.1 Fabrication, hybridization and post analysis of DNA microarrays

A cDNA array is made by adding cDNA from any library of interest to the array (usually made by quartz), for prokaryotes and yeast this is usually done by amplifying genomic DNA with gene specific primers, while for eukaryotes EST positions are usually chosen (Duggan *et al*., 1999).

A different type of microarrays is the oligonucleotide array (Affymetrix GeneChip®), which is constructed in a fundamentally different manner. Instead of printing whole cDNA molecules to the matrix, the four different nucleotides are added by parallel addition using a light masking technique, see **Figure 1**. The oligonucleotides are usually between 25-70 bases long, depending on the type of array. The shorter they are more

stringent conditions are necessary to give a satisfactory hybridization. As a result, the GC ratio has to be approximately equal in all oligos, which in turn limits the number of possibilities.

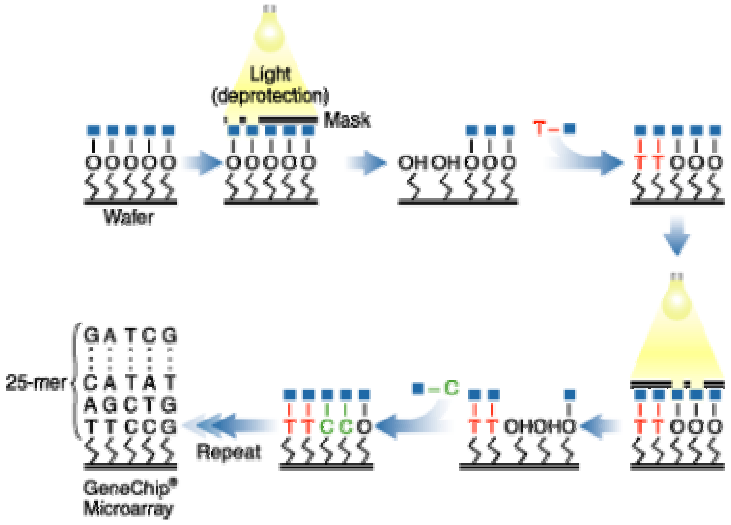


Figure 1: Affymetrix use a combination of photolithography and combinatorial chemistry to manufacture their GeneChip® Arrays (taken from the Affymetrix GeneChip® web site, http://www.affymetrix.com/technology/manufacturing/index.affx).

When conducting experiments with cDNA arrays, mRNA from the tissue of interest and the reference tissue, has to be extracted, purified and labeled before it is allowed to hybridize with DNA on the array. In this way gene expression between e.g. cancer cells and healthy cells, can be compared and quantitatively measured. Extraction and purification is a crucial step, as the quality of the mRNA has great influence on the final results. The labeling is usually done be using fluorescent dyes, where Cye3-dUTP (red) and Cye5-dUTP (green) are most commonly used. In some cases radioactive labeling is being employed, incorporating ^{33}P, ^{35}S or ^{3}H directly into the nucleotides. **Figure 2** shows a chart revealing the correlation between amount of starting material, total RNA and detection limit. It also shows that indirect and radioactive labeling has a much lower detection limit than direct labeling (Duggan *et al.*, 1999).

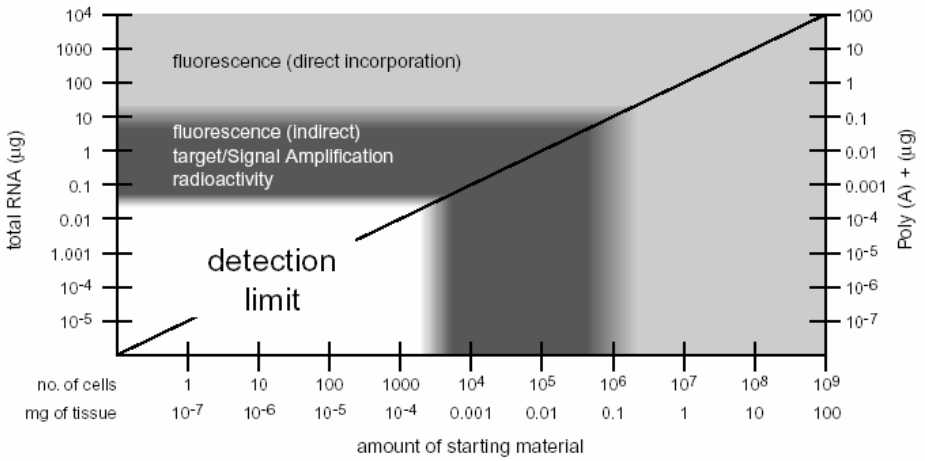


Figure 2: Chart showing the correlation between amount of starting material (for eukaryotic cells), total RNA and detection limits using different kinds of labeling methods. (taken from (Duggan *et al*., 1999))

When the material has been label, it is ready to hybridize with DNA on the array. This is a sensitive step and any physical contact with the array, such as dust or scratches, and/or too little or too much washing, will greatly affect the final result. The figure below summarizes the procedure for conducting a cDNA microarray experiment.

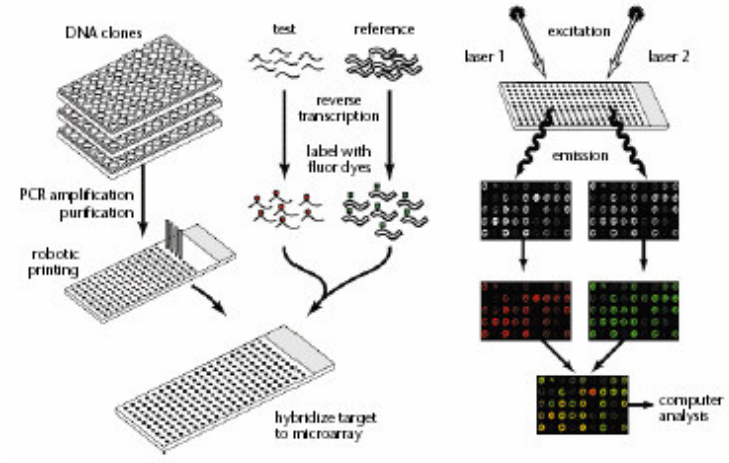


Figure 3: Chart summarizing the procedure for conducting a cDNA microarray experiment. Starting with the construction of an array by applying genes of interest followed by labeling and hybridization of the test and reference DNA. Finally the array is scanned and analyzed (Duggan *et al*., 1999).

The hybridized target molecules, on the microarray, are visualized by laser induced fluorescence, detected by a high resolution CCD camera, and a two channel image (red and green) is saved on a computer for further analysis (Gibson and Muse, 2002). Since

the two dyes always are incorporated a little differently, the data has to undergo a normalization process before further analysis. Finally the image is interpreted using sophisticated computational algorithms, such as hierarchical clustering (Alizadeh *et al*., 2000; Eisen *et al*., 1998; Sokal and Michener, 1958), κ-means clustering (Brazma and Vilo, 2000; MacQueen, 1967; Tavazoie *et al*., 1999) and self organizing maps (SOM) (Tamayo *et al*., 1999; Toronen *et al*., 1999). The most common technique is hierarchical clustering, and this is the only method to generate a dendrogram. Hierarchical clustering is fast and the process is relatively simple, starting by calculating a distance matrix between all genes. In the next step the distance matrix is traversed to find the two most similar genes or clusters, and placing them in a common cluster. In the last step the distances between the new cluster and all the other clusters or genes are calculated. The process is repeated until all objects are clustered. When calculating distances or similarities between two objects, there are a variety of different algorithms to choose from, etc. Euclidian, Manhattan or Pearson correlation. All these methods will generate a slightly different outcome (Quackenbush, 2001). As can be seen in **Figure 4** results from using Euclidian distance measures will give a completely different outcome compared to Pearson correlation. While Euclidian distance measures the distance between x and y, Pearson correlation calculates the angle θ between x and y, which is unaffected by parallel shifts in the data. When detecting co-expressed genes Pearson correlation is probably the most suitable method, while Euclidian distance measurements is better in comparing absolute gene products (Amaratunga and Cabrera, 2003).

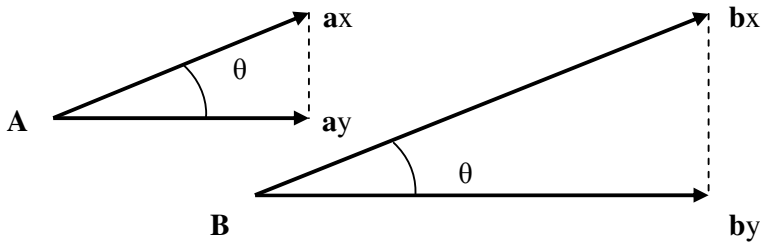


Figure 4: While the distance between x and y increases from figure A to B (measured as Euclidian distance), θ remains constant and is unaffected by a parallel shift in the data (corresponding to Pearson correlation).

When doing hierarchical clustering one of the following methods can be applied to cluster the information in the distance matrix; Single-linkage clustering, Complete-linkage clustering or Average-linkage clustering. The unweighted pair-group method average (UPGMA) is the most common average-linkage clustering method. As an alternative the weighted pair-group average method (WPGMA) might be a better choice if the cluster sizes are expected to be greatly uneven (Quackenbush, 2001). Since these methods all yield different results, biological knowledge concerning the input data will be of great value in choosing which method to use. Without any biological basis the average-linkage clustering method is usually the best choice.

3.2.8 Genome sequencing and comparison

The ultimate bacteria genotype is the complete sequence of a whole genome. Since the first bacterial genome (*Hemophilus influenzae* (Fleischmann *et al*., 1995)) was sequenced in 1995, more than 155 complete genome sequences are now publicly available, where 144 are of prokaryotic origin (Entrez-Genome, February 2004), and many more are about to be completed. As this process becomes less labor intensive and less expensive, more and more complete genome sequences will become available for analysis. This may be looked upon as a new era in microbiology, allowing complete genotype::phenotype comparison to be made. It gives us the opportunity to study evolution, lateral gene transfer and the function of genes in a new perspective. Still the comparison of whole genome sequences is not straight forward and there are many complicating factors to overcome. It is difficult to compare genomes that are distantly related since the number of homolog sequences and conserved regions may be, very small, rearranged and scattered through out the genome. Thus, creating a good alignment is difficult but not impossible. Even thought the comparison of closely related species becomes difficult, mainly as a result of indels, inversions, tandem repeats, genome rearrangement and divergence in the third position of the codon. There is a lot of ongoing research seeking to find efficient methods for whole genome comparison. The BLAST program might be a useful tool in comparing genomes, although it is not designed to perform large scale genome alignments. Still, BLASTing whole genomes against each other and “three genome comparisons” might give crucial and valuable information about similar genes and relations (the Microbial Genome Database (MBGD) http://mbgd.genome.ad.jp). MUMmer (Delcher *et al*., 1999) is an application

meant for doing whole genome comparison, having the capability of rapid alignment between two genomes. The method is based on streaming the query sequence past a previously generated suffix tree, causing it to use less CPU time and memory. The output can be visualized as a plot and analyzed. Another program, called PROmer is the protein version of MUMmer, allowing comparison of large protein sequences (Delcher *et al*., 2002). Since protein sequences are much more conserved than nucleotide sequences, protein-based alignments are capable of detecting much older relationships than DNA alignments, making PROmer a natural choice if distantly related species are to be compared. An interesting fact that has emerged from genome analysis is the finding that the degree of horizontal gene transfer is surprisingly high (Eisen, 2000).

3.2.9 Phylogenetic analyses

Sequences that are to be compared phylogeneticaly must be of orthologous origin in order to reflect their true evolution, while paralogous have to be avoided. Orthologous sequences in two organisms are homologs that evolved from the same feature in their last common ancestor (Fitch, 1970). While paralogous are homologous sequences derived as a result of parallelism, usually by gene duplication. Prior to comparing sequences using phylogenetic methods, the sequences have to be aligned by multiple alignment program such as ClustalW (Higgins *et al*., 1996). Unless the sequences are too complex, having large indels and/or being of considerable different length, the program will compute an alignment close to ideal. Problems concerning the treatment of flanking positions and caps can be overcome by using a program such as Gblocks (Castresana, 2000), which removes weakly conserved regions, including gaps and flanking positions. A wide range of programs for phylogenetic analysis are available, PHYLIP (PHYLogeny Inference Package) (Felsenstein, 1993) and PAUP (Phylogenetic Analysis Using Parsimony) (Swofford, 1998) being the two most important ones. When measuring changes between sequences, nucleotide or protein, there are several methods available. Generally one of three methods is selected; maximum parsimony, distance methods or maximum likelihood. These methods both have their advantages and disadvantages, and the method chosen depends on the type of data that is to be analyzed and CPU time available.

Construction of phylogenetic trees using maximum likelihood (Felsenstein, 1981) is based on selecting trees that maximizes the probability of observing the data. For

sequences the data is the alignment of nucleotides or amino acids. These trees are calculated on the basis of the most suitable substitution model (see below). Since all possible topological trees that might fit the model have to be calculated, this method is extremely computer intensive and becomes virtually impossible if the data sets are large.

Maximum parsimony is based on the assumption that the most likely tree is the one that requires the fewest number of changes to explain the data (Swofford, 1993). Maximum parsimony is best suited to sequences that are quite similar, but if there are a large number of sequences to be analyzed the number of possible trees may become very large. The parsimony method is fairly computer intensive if the number of sequences and characters is large, but not as intensive as maximum likelihood.

Bayesian analysis is based on the idea of posterior probabilities, which is estimated probabilities based on a model that has learned something about the data (Huelsenbeck *et al*., 2001; Mau *et al*., 1999). As with maximum likelihood, the user has to postulate a model of evolution. This method searches for the best set of trees and generates a final consensus tree. Despite the fact that Bayesian analysis is relatively computer intensive it has the huge advantage of bypassing the time consuming bootstrapping algorithm.

Maximum likelihood, parsimony and Bayesian analysis uses tree-searching methods to find the tree that best meets certain criteria. When conducting an exhaustive search the user is guarantied to find the best tree, unfortunately it can be extremely computer intensive and in most cases impractical. The second best method is the branch-and-bound algorithm, but as with exhaustive searching it is also relatively slow (Hall, 2001). Usually a heuristic search has to be employed, a method often referred to as hill-climbing. Two extensively used methods within this category is branch swapping and stepwise addition, but there are many more. Using a heuristic method is always a trade off between the certainty of finding the best tree and CPU hours used. All these methods are character-based, meaning that they use the alignment directly without generating a distance matrix.

The distance methods are based on measuring the number of changes between pairs of sequences by generating a distance matrix. The sequences having the smallest number

of substitutions between them are placed as neighbors in the final tree. One of the big advantages using these methods is the fact that they are much faster than the other methods mentioned above. Common methods that relies on distances is the Neighbor-joining algorithm (Saitou and Nei, 1987), and the Fitch-Margoliash algorithm (Fitch and Margolia.E, 1987), employed in the programs FITCH and KITCH (Felsenstein, 1993). The Neighbor-joining method is very fast and suitable for sequences where the rates of evolution varies within the sequence (Jin and Nei, 1990). Another method is UPGMA which in fact is a clustering method. It assumes that all taxa are equally distant from the root, something that is not very likely; as a result UPGMA is rarely used in phylogenetic analysis. Neighbor joining, the Fitch-Margoliash method and UPGMA are algorithmic methods, meaning that they use an algorithm when doing tree construction, instead of tree-searching methods as mentioned above.

When corrections for multiple substitutions are made, maximum likelihood and distance methods have been shown to be more reliable than maximum parsimony (Mount, 2001). When branch lengths are varying the neighbor method has been shown to be more reliable than both standard and evolutionary parsimony (Jin and Nei, 1990).

It is impossible to mimic a true evolutionary process and statistical assumptions have to be made. Since transitions are more likely to occur than transversions some substitutions are more common than others. To cope with these problems, and the fact that there is a significant probability that a character has changed more than once, different kinds of substitution models have been made, (Jukes and Cantor, 1969), (Kimura, 1980), (Tajima and Nei, 1984), (Hasegawa *et al*., 1985), (Tamura and Nei, 1993). The model of choice is the one that has the greatest ability to predict the observed data and gives the highest likelihood score. The substitution rate might also vary within a sequence as a result of selection pressure. To compensate for this phenomenon a gamma distribution can be calculated, allowing variation in substitution rates.

To facilitate the process of choosing a model that best suits the data, a program such as Modeltest is helpful (Posada and Crandall, 1998), see 4.2.9 for further description. When a final tree has been computed it is always a good idea to generate other trees using different methods in order to verify support for the chosen model.

In order to test how well a particular data set fits a model or method, the final tree has to be bootstrapped. This is done by resampling the alignment, making pseudoalignments (usually 100 or 1000 times) by randomly reordering the columns in the multiple sequence alignment. A new tree is then made from each of the 100 or 1000 pseudoalignments, using the same settings as for the original tree. The original tree is then compared to one of the new trees, and for every clade that is present in both trees a score of 1 is given to that particular clade, if not a score of 0 is given. This process is repeated for each pseudoalignment. The final result is a bootstrapped tree, revealing the reliability of each clade. Clades having a score above 90% are pretty confident, while those having a value below 70% should be looked upon as less trustworthy. When using maximum likelihood methods bootstrapping can turn out to be extremely computer intensive and in many cases impractical. Fortunately MrBayes avoids these problems, instead of making pseudoalignments, it directly counts the fraction of times a clade occurs among the thousands of trees generated within the stable state.

3.3 Objectives of this study

Bacterial classification, natural or artificial, is a central field in microbiology and as a consequence many different methods have been developed.
I would like to point out that no method, despite new technology and whole genome comparison, is flawless. This problem also applies to phylogenetic trees, where orthologous genes are compared. Strictly speaking it is not possible to construct meaningful phylogenetic trees which are valid for all prokaryotes. By definition, such trees will only be valid for the molecules used in the tree construction.

In this study, a method for bacterial classification using oligonucleotides will be evaluated. This technique is based on the idea of selecting a set of informative oligonucleotides, to be placed on a DNA microarray, for the purpose of classifying bacteria based on their hybridization patterns. The empirical nature of this method circumvents problems created by horizontal gene transfer and non-orthogonal gene replacement. The greatest challenge lies in selecting a set of primers that, in a most efficient way, will be able to differentiate as many species and strains as possible and to evaluate the output made by the microarray.

Prior to this study a method was developed by W. Davies and S. Gaure for extraction and selection of informative oligonucleotides from genome data (using the programs Extseq, Gencnt and Selentprim written by Simen Gaure). Thus the aim of this thesis is to establish a method for testing different sets of oligonucleotides and their ability to classify bacterial species. In order to do this a method has to be established with the aim of testing and visualizing the generated data in a suitable manner for comparison with phylogenetic reference trees.

In summary our goals are:

1) To select different sets of complete genome sequence data suitable for extraction of informative primers, using previously written applications, and to generate a diverse range of primer sets for further analysis.

2) To develop a program or method, in collaboration with S. Gaure, to evaluate selected primer sets and their usefulness in the classification of bacterial species based on their 10-mer oligonucleotide frequencies on the primer set. A method also has to be developed in order to visualize this information in a suitable manner for comparison with phylogenetic reference trees.

3) Make a set of robust reference trees using established methods.

4) Compare the results from the oligonucleotide method to the reference trees and evaluate its value in bacterial classification.

4 Materials and methods

4.1 Computers and databases

The following computers were used in our research:

The Biotin EMBOSS server at the Biotechnology Center of Oslo, with a 1 GHz Pentium 3 and 2 GB RAM, running Linux 7.3 2.96-112.

The Macduff server at UiO, with a 1.8 GHz Pentium 4, 512 MB RAM, running Linux 8.0 3.2-7

The Darwin server at UiO, with 2 X 400MHz UltraSPARC-II, 1 GB RAM, running SunOs 5.6

A private laptop, with a 1.4 GHz Pentium M, 512 MB RAM, running windows XP.

A private desktop computer with a 2.26 GHz Pentium 4, 512 MB RAM, running windows XP.

An Apple Macintosh computer at UiO, with a 400 MHz G3, 128 MB RAM, running Mac OS 9.2

The following databases were used in this work:

http://www.kegg.com/kegg/kegg2.html
ftp://ftp.genome.ad.jp/pub/kegg/
http://www.ncbi.nlm.nih.gov/genomes/static/micr.html
ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/

4.2 Description of programs used in this study

Programs made specific for this work		
Program	**Author**	**Group**
Extseq	S. Gaure and W. Davies	USIT
Gencnt	S. Gaure and W. Davies	USIT
Selentprim	S. Gaure and W. Davies	USIT
Testprimers	S. Gaure and W. Davies	USIT
Testarray	S. Gaure and A. Klevan	USIT
Revperl	A. Botnen and A. Klevan	USIT
Extract	A. Botnen and A. Klevan	USIT
Consrun	A. Botnen and A. Klevan	USIT
Free or commercially available programs		
ClustalX 1.83	(Thompson, 1994)	
Gblocks 0.91b	(Castresana, 2000)	
PAUP	(Swofford, 1998)	
Modeltest 3.06	(Posada and Crandall, 1998)	
PHYLIP 3.5	(Felsenstein, 1993)	
TREE-PUZZLE 5.1	(Schmidt *et al*., 2002)	
MrBayes 3.0	(Huelsenbeck and Ronquist, 2001)	
J-Express 1.1	(Dysvik and Jonassen, 2001)	Molmine
ReadSeq	(Gilbert, 1999)	
Revseq	(Williams, 1999)	EMBOSS
Comseq 1.12	(Williams, 2000)	EMBOSS
Cons	(Carver, 2000)	EMBOSS

Table 1: List of programs used in this study.

4.2.1 Extseq

The program Extseq reads a list of files with the “fna” extension and makes them suitable for further processing. This involves collecting the selected genome files into one file and changing their names according to certain rules. The output file contains all inserted genomes in a concatenated file, their length and their new names. For each genome there is 8 bytes at the beginning of the genome containing the number of bases, followed by 8 bytes with the length of the genome, in bytes, and the genome name. The names are fetched from the “fna” files and abbreviated. The output is a binary file meant

to be an input file for Gencnt. The symbols a,c,g,t are replaced by a two bits code in the following manner:

a: 00
c: 01
g: 10
t: 11

The program can be executed like this:

$ Extseq inputfolder/.fna outputfolder/outputfile.seq*

Executing this command will read all files in the “inputfolder” with the extension .fna, write the output to the “outputfolder” and give the “outputfile” the extension .seq. Note that this program does not read from standard input, thus implementing < and/or > will not have any effect.

4.2.2 Gencnt

This program read files generated by Extseq. The output contains a list of all 10-mer oligonucleotides in a certain genome, or genome set, and their frequencies. The output file is made up by a primer sequence followed by its melting point in 4 bytes. Then comes a list as long as the total number of genomes, and for each genome there is one pare of 16-bits digits with the number of forward and reverse matches. This is repeated for each 10-mer oligonucleotide. The melting temperature for the 10-mers is calculated in the following way:

For each A & T the melting point is 2
For each G & C the melting point is 4
These numbers for all ten bases will then be summarized.

To avoid improper 10-mer oligonucleotides in the final array, palindromes and primers with bad energy are removed. In this context a palindrome is defined as a 10-mer where the three first bases can pair with the three last bases. To avoid primers with improper

energy only 10-mer oligonucleotides with a free energy in the last pentamer between -9 kcal/mol and -5 kcal/mol will be kept. The free energy is calculated according to the nearest neighbor method (Rychlik, 1995). A set with all possible 10-mer oligonucleotides will contain more than a million different primers (10^4 = 1.048.576), but after filtration this number is reduce to approx 700.000. Finally the extension "pri" will be added to the output. In addition Gencnt will output, to the terminal, the number of discarded 10-mer oligonucleotides in each genome, due to unspecific symbols such as Y, N, M, R, S, W, K, generated by sequencing errors.

The program can be executed in the following way:

$ Gencnt < outputfolder/outputfile.seq > outputfolder/outputfile.pri

This command will read the output from Extseq, "outputfile.seq" execute it in Gencnt and give the output file the extension "pri".

4.2.3 Selentprim

This program is used for primer selection, taking the output from Gencnt. The primers are selected according to melting point, GC content, minimum frequency and entropic distribution. Below is a list with arguments accepted by Selentprim:

-e	number between 0-1	minimum entropy
-E	number between 0-1	maximum entropy
-t		minimum melting temperature
-T		maximum melting temperature
-c	number between 0-10	minimum C bases
-C	number between 0-10	maximum C bases
-f	number greater than 0	minimum frequency for which a primer should occur in at least one genome

The entropy H of a discrete distribution p is given by:

$$H(p) = \sum_i p_i \log p_i$$

By measuring the uniformity of the primers in the Gencnt output file, Selentprim has the capability to extract primers on the basis of their ability to distinguish between different genomes. A primer that is present in all genomes at the same frequency is not very informative neither is a primer that hardly ever appears, thus the ideal primers lies some where in between. The goal is to select a set of primers, with a skewered distribution, being able to differentiate between different genomes in a most efficient way. The higher the entropy the more uniform is the distribution. So if $H = 1$ the distribution is uniform, if $H = 0$ the distribution is concentrated in a single point. The entropy is normalized to be a number between 0 and 1. This is done by dividing H with log (N), where N is the number of genomes. The output from Selentprim is a list of primers that fulfills the conditions made at the command line. The file has the same binary format as Extseq. Since GC ratio and melting temperature are in correlation to each other only the GC ratio has been used during this study.

The program can be executed in the following way:

$ Selentprim –e 0.3 –E 0.6 –c 4 –C 5 –f 2 < outputfolder/outputfile.pri > outputfolder/outputfile.dat

Executing this command will take the output from Gencnt and generate a file with primers containing from 4 to 5 C-bases, having an entropy between 0.4 and 0.6 and a minimum frequency of f = 2. The output file is given the extension “dat”. This program also outputs to screen how many available primers it has (usually 718.744, see above) and the number of primers extracted using the given settings. The goal is to select 4.000 primers suitable for classification purposes.

4.2.4 Testprimers

This program sorts the output from Selentprim for presentation purposes. Both the output from Gencnt and Selentprim are used as input. An ASCII file is made for each genome containing a list of primers, extracted by Selentprim, and their 3log frequency in that particular genome. The output from Testprimers can be used to make gnuplots or a list expressing distances between two genomes as a number between 0 and 1. This is done by calculating the Euclidian distances between two and two primer frequencies in

a pair of genomes. Finally the distances between all pairs in these two genomes will be summarized and normalized. This is done for every single genome, thus all pair of genomes are compared to each other and given a number reflecting their relatedness. The pair having the lowest number is probably the most similar.

The program can be executed in the following way:

$ testprimers –f outputfile.dat –p outputfile.pri

The file "outputfile.dat" is the output from Selentprim, holding the primer set, while the file "outputfile.pri" is the output from Gencnt holding the 10-mer oligonucleotide frequencies for genomes that are to be classified. Testprimers automatically generate several output files, one file for each genome in the Gencnt output file. As described above the output files from Testprimers can be used to create gnuplots or a 1:1 comparison of the genomes. This comparison is employed by writing the following command.

$ sort +1 –n dfile.dat

By executing this command a list with normalized Euclidian distances between all pair of genomes will be written to the screen. This list can be converted into a graph using J-Express.

4.2.5 Testarray

Testarray is a program that combines the output from GENCNT and SELENTPRIM (see flowchart on page 50) and produces a table in which there is a column for each bacteria/genome and rows reflecting the actual frequency of each primer in that particular species (see **Figure 5**). In this way we can generate a file containing a set of primers (e.g. made by five different Proteobacterial genomes) and test it against a completely different set of Proteobacterial genomes. The process mimics a true DNA microarray by virtually hybridizing genome DNA to the primers (selected by Selentprim) on the virtual array. The application can analyze several genomes at the same time, thus outputting a multiple experiment file for genome comparison, using J-

Express. Two different versions of Testarray have been made, Testarray and Testarray-v2, the later having a feature for dividing primer frequencies according to genome size. Without this kind of normalization the size of the different genomes will influence the clustering process. The output is an ASCII file that has to be edited in Excel before further processing in J-Express.

1		9	Es.col.CFT073	Es.col.K-12.MG1655	Es.col.O157:H7,	Es.col.O157:H7,-4	Sa.ent.Typ.(Sa.CT18,	Sa.ent.Ty
2	aaaacagctc		12	10	13	13	24	
3	aaaacagtgc		22	22	22	22	9	
4	aaaacggatc		11	12	17	17	13	
5	aaaacgtacg		10	6	7	7	35	
6	aaaacgtcgt		12	10	15	15	20	
7	aaaactctgc		17	20	16	16	9	
8	aaaagcaagc		13	12	16	16	23	
9	aaaagcactg		32	31	25	27	11	
10	aaaagctgac		21	15	30	32	25	
11	aaacaacgag		10	14	9	9	7	
12	aaacaccaca		12	15	16	16	7	
13	aaacaccatc		24	31	26	26	12	
14	aaacagatgg		17	12	23	24	23	

1		9	Es.col.CFT073	Es.col.K-12.MG1655	Es.col.O157:H7,	Es.col.O157:H7,-4	Sa.ent.Typ.(Sa.CT18,	Sa.ent.Ty
2	aaaacagctc		1.146914	1.077767	1.182151	1.175737	2.495302	2.504194
3	aaaacagtgc		2.102676	2.371088	2.000564	1.989710	0.935738	0.939073
4	aaaacggatc		1.051338	1.293321	1.545890	1.537503	1.351622	1.356438
5	aaaacgtacg		0.955762	0.646660	0.636543	0.633089	3.638982	3.651950
6	aaaacgtcgt		1.146914	1.077767	1.364021	1.356620	2.079418	2.086828
7	aaaactctgc		1.624795	2.155534	1.454955	1.447062	0.935738	1.043414
8	aaaagcaagc		1.242491	1.293321	1.454955	1.447062	2.391331	2.295511
9	aaaagcactg		3.058438	3.341078	2.273368	2.441916	1.143680	1.147756
10	aaaagctgac		2.007100	1.616651	2.728042	2.894123	2.599273	2.504194
11	aaacaacgag		0.955762	1.508874	0.818412	0.813972	0.727796	0.730390
12	aaacaccaca		1.146914	1.616651	1.454955	1.447062	0.727796	0.730390
13	aaacaccatc		2.293829	3.341078	2.364303	2.351475	1.247651	1.043414
14	aaacagatgg		1.624795	1.293321	2.091499	2.170592	2.391331	2.295511

Figure 5: The upper picture shows the output from Testarray prior to any normalization. The lower picture shows the output from Testarray-v2, where the primer frequencies have been divided with their associated genome size (Screenshot from Excel).

The program can be executed in the following way:

$ Testarray-v2 outputfolder/outputfile.dat outputfolder/genomes.pri > outputfolder/outputfile.ary

4.2.5.1 Revperl

Revperl is a script written to automatically run multiple files through the EMBOSS program Revseq (described in 4.2.15.1), which calculates the reverse, the compliment or the reverse compliment of the input sequence. When the input sequences have been

converted they are merged into the end of their original input file. In this way the forward and reverse stretch of DNA can be made available in one single stranded FASTA file.

4.2.5.2 Extract

Extract is a small Perl script that extracts gene sequences from a multiple FASTA file according to one or more specified search word, e.g. “16S” or “ribosomal”. The program searches through every file in the folder specified in the program code.

4.2.5.3 Consrun

Consrun is a script written to automatically run multiple files through the EMBOSS program Cons.

4.2.6 ClustalX

ClustalX is a window interface for the ClustalW multiple sequence alignment program. It provides an integrated package for performing multiple sequence alignments, profile alignments and result analysis.
The user can cut-and-paste sequences to change the order of the alignment, select a subset of sequences to be aligned and select a sub-range of the alignment to be realigned and inserted back into the original alignment. Alignment quality analysis can be performed and low-scoring segments or exceptional residues can be highlighted.
All input sequences must be in 1 file. 7 formats are automatically recognized: Clustal, Fasta, PHYLIP, GDE, NBRF/PIR, GCG/MSF, and Nexus. All non-alphabetic characters (spaces, digits, punctuation marks) are ignored except "-" which is used to indicate a GAP. Unless the sequences are too complex, having large indels and/or being of considerable different length, the program will compute an alignment close to ideal.
The program can be downloaded from ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX/.

4.2.7 Gblocks

Gblocks is a program for eliminating poorly aligned positions and divergent regions of an alignment of DNA or protein sequences. These positions may not be homologous or may have been subject to multiple substitutions and it is convenient to eliminate them

prior to phylogenetic analysis. Gblocks selects blocks in a similar way as it is usually done by hand, but following a reproducible set of conditions. The selected blocks must fulfill certain requirements with respect to the lack of large segments of contiguous nonconserved positions, lack of gap positions and high conservation of flanking positions, making the final alignment more suitable for phylogenetic analysis. Several parameters can be modified to make the selection of blocks more or less stringent. The program can be installed on virtually every system, or accessed on the Gblocks web server. The advantage of using this application is that it has been shown to give alignments that are virtually independent of the different options available in ClustalX (Daubin *et al*., 2002). The application can be accessed on http://woody.embl-heidelberg.de/phylo/.

4.2.8 PAUP

Phylogenetic Analysis Using Parsimony (PAUP) is a commercially available program for phylogenetic analysis. The package offers a number of options for conducting different types of phylogenetic analysis, such as parsimony, maximum likelihood and different distance methods. The input file has to be writen in the nexus file format. The output created by PAUP is a visualization of the phylogenetic relation between the organisms of interest visualized by a phylogenetic tree. Unfortunately PAUP doesn't have the ability to construct maximum likelihood protein trees. The program can be ordered at the PAUP home page, http://paup.csit.fsu.edu/index.html.

4.2.9 Modeltest

Modeltest is designed to compare different nested models of DNA substitution in a hierarchical hypothesis-testing framework. It compares 56 (in version 3.06) different likelihood models to find the one that bests suits the data set (Posada and Crandall, 1998). The program is meant to be used together with PAUP. A script, called "modelblock" is inserted at the end of the nexus file that is to be analyzed, PAUP is executed and the likelihood scores for 56 different models of evolution will be computed. The results will be written to a new file named model.scores. The file model.scores can be opened with the program modeltest and the most suitable model will be selected, including its parameter settings. A new block containing these data can

then be inserted into the nexus file, substituting the first block, “modelblock”. PAUP is then executed one more time, using the best model and settings for the data set. The application can be downloaded from the Modeltest home page at, http://inbio.byu.edu/Faculty/kac/crandall_lab/modeltest.htm.

4.2.10 The PHYLIP package

The phylogenetic interference package (PHYLIP) is a package of programs for construction of phylogenetic trees. Instead of being one program with many different functions the PHYLIP package is divided into many small programs having specific tasks, making it an extremely dynamic tool. All programs are menu based, but no window interface has been developed. The programs read files written in PHYLIP format. In this study five different PHYLIP programs has been used, BOOTSEQ, PROTDIST, FITCH, NEIGHBOR and CONSENSE. BOOTSEQ is a program for resampling datasets by the bootstrapping method, giving multiple datasets that can be used as input by most PHYLIP programs. PROTDIST is an application for computation of a distance measures for protein sequences using different substitution models. FITCH is a program to estimate phylogeny from distance matrix data using the "additive tree model". NEIGHBOR is an application for construction of phylogenetic trees by Neighbor joining or UPGMA, using a distance matrix as input. CONSENSE is a program used to compute consensus trees, using the majority-rule consensus tree method.

PHYLIP can be downloaded from the PHYLIP home page at http://evolution.genetics.washington.edu/phylip.html.

4.2.11 TREE-PUZZLE

TREE-PUZZLE is program suitable for maximum likelihood protein analysis. TREE-PUZZLE uses an algorithm called quartet puzzling, which is a maximum likelihood distance method, allowing analysis of large data sets. In addition the program can calculate a clock assumption, has a wide range of substitution models and provides gamma distribution. The program is relatively computer intensive which makes bootstrapping (by using SEQBOOT) virtually impossible when dealing with large data

sets. TREE-PUZZLE reads files written in the PHYLIP file format. The program can be downloaded from http://www.TREE-PUZZLE.de/

4.2.12 MrBayes

MrBayes is a program for phylogenetic studies based on Bayesian analysis (Mau *et al.*, 1999; Rannala and Yang, 1996) and the idea of posterior probabilities, which is estimated probabilities based on a model that has learned something about the data. Instead of seeking the best tree, as with maximum likelihood, MrBayes search for the best set of trees. From this set a consensus tree is calculated, thus bypassing the time consuming bootstrapping algorithm. Since MrBayes as default use four independent chains, the probability of being fixed on a local top is smaller than for other likelihood methods. For further information see (Huelsenbeck *et al.*, 2001).
The program can be downloaded from
http://morphbank.ebc.uu.se/mrbayes/download.php.

4.2.13 J-Express

J-Express is a software package for analysis and visualization of microarray data. The program gives access to multidimensional scaling and different clustering methods. J-Express has the ability to read output from TESTARRAY without any further conversion. Its efficiently allows interactive clustering of our genomes and construction of dendrograms. J-Express is a commercial program owned by MolMine A/S (http://www.molmine.com).

4.2.14 Readseq

Readseq is a sequence conversion program that can read, write and convert between any file written in one of the following formats:

Abstract syntax notation (ASN.1)
DNA strider
European Molecular Biology Laboratory (EMBL)
Fasta/Pearson
FITCH
Genbank
Genetics Computer Group (GCG)
Intelligenetics/Stanford
Multiple Sequence Format (MSF)
National Biomedical Research Foundation (NBRF)
Olsen
Nexus format
PHYLIP
Plain text
Pretty format for publication
Protein Information Resource (PIR or CODATA)
Zuker for RNA analysis

Table 2: Formats accepted by ReadSeq.

The program can be accessed at: http://searchlauncher.bcm.tmc.edu/seq-util/readseq.html

4.2.15 EMBOSS

The European Molecular Biology Open Software Suite (EMBOSS) is a package of academic sequence analysis software. The software automatically copes with data in a variety of formats and allows transparent retrieval of sequence data from the web. The EMBOSS package contains more than 100 different applications. EMBOSS can be accessed at the Norwegian EMBnet node (http://www.no.embnet.org). Below is a short explanation of different EMBOSS programs and scripts used in this study:

4.2.15.1 Revseq

Revseq takes a sequence and outputs its reverse complement. It can also output just the reversed sequence or the complement.

4.2.15.2 Compseq

Compseq counts the composition of dimer/trimer/etc words in the input sequence(s).

4.2.15.3 Cons

Cons calculates the consensus sequence from a multiple sequence alignment. To obtain the consensus a scoring matrix is used to calculate a score for each position in the alignment.

4.3 Phylogenetic classification

In this section methods for construction of the different reference trees will be explained. Since phylogenetic analysis based on different genes reveals some differences between the final trees, multiple trees should be generated using different genes. As a result, four different housekeeping genes have been selected in construction of the reference trees. These four genes are the 16S rRNA gene, the ATPase alpha chain gene, the Prolyl-tRNA synthetase gene and the Methionyl-tRNA synthetase gene.

4.3.1 Phylogenetic analysis of the 16S rRNA gene

The 16S rRNA gene is by far the most common sequence used in phylogenetic comparison. However when extracting these genes from the different bacterial genomes a problem appeared. Most bacteria have multiple copies of the rRNA operon, varying from one to more than eight. Aligning these genes revealed small but significant differences, thus making further analysis complicated. When doing analysis on the 16S rRNA genes from the different species in the EcoSalmoFlex set (see **Table 3**), some of the genes were intermingled between two or more species (see **Figure 9,** page 57).

EcoSalmoFlex
Escherichia coli CFT073
Escherichia coli K-12 MG1655
Escherichia coli O157 EDL933
Escherichia coli O157 Sakai
Salmonella typhi CT18
Salmonella typhi Ty2
Salmonella typhimurium
Shigella flexneri 2457T (serotype 2a)
Shigella flexneri 301 (serotype 2a)

Table 3: Set containing 9 different closely related enteric Bacteria used in phylogenetic analysis.

As a result of this phenomenon the phylogenetic comparison of 16S rRNA genes is divided in two parts. In the first part the aim is to compare every single gene (a total of 320 genes from 61 different organisms). In the second part only the consensus sequence from each organism will be subject to comparison. The flowchart below shows how this was carried out.

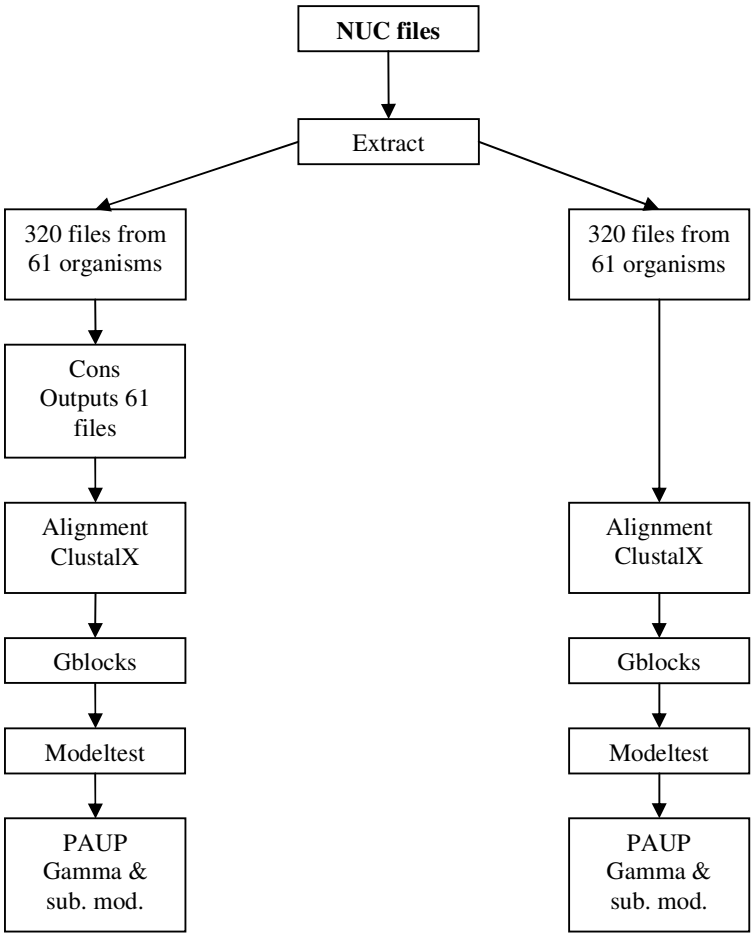


Figure 6: Flow chart showing the procedure for 16S analyses using Extract, Cons, ClustalX, Gblocks, Modeltest and PAUP. See below for explanation of the method.

Genome sequences were obtained for all bacteria listed in **Table 4** and used for gene extraction in order to generate the reference trees. The same bacteria were also used in evaluation of the dendrograms made later in this thesis. All bacteria have a unique abbreviation, analogous to the KEGG web site (www.kegg.com/kegg/kegg2.html), in order to make file handling more convenient. These abbreviations are used all through the study, but always together with a list similar to those below.

Gram-positive bacteria	
Bacillus anthracis	ban
Bacillus cereus	bce
Bacillus halodurans	bha
Bacillus subtilis	bsu
Bifidobacterium longum	blo
Clostridium acetobutylicum	cac
Clostridium perfringens	cpe
Clostridium tetani	ctc
Corynebacterium efficiens	cef
Corynebacterium glutamicum	cgl
Enterococcus faecalis	efa
Lactobacillus plantarum	lpl
Lactococcus lactis	lla
Listeria innocua	lin
Listeria monocytogenes	lmo
Mycobacterium bovis	mbo
Mycobacterium leprae	mle
Mycobacterium tuberculosis CDC1551	mtc
Mycobacterium tuberculosis H37Rv (lab strain)	mtu
Oceanobacillus iheyensis	oih
Staphylococcus aureus Mu50 (VRSA)	sav
Staphylococcus aureus MW2	sam
Staphylococcus aureus N315 (MRSA)	sau
Staphylococcus epidermidis	sep
Streptococcus agalactiae 2603	sag
Streptococcus agalactiae NEM316	san
Streptococcus mutans	smu
Streptococcus pneumoniae R6	spr
Streptococcus pneumoniae TIGR4	spn
Streptococcus pyogenes MGAS315 (serotype M3)	spg
Streptococcus pyogenes MGAS8232 (serotype M18)	spm
Streptococcus pyogenes SF370 (serotype M1)	spy
Streptococcus pyogenes SSI-1 (serotype M3)	sps
Streptomyces avermitilis	sma
Streptomyces coelicolor	sco
Thermoanaerobacter tengcongensis	tte

Proteobacteria	
Bordetella bronchiseptica	bbr/bre
Bordetella parapertussis	bpa
Bordetella pertussis	bpe
Brucella melitensis	bme
Brucella suis	bms
Campylobacter jejuni	cje
Caulobacter crescentus	ccr
Coxiella burnetii	cbu
Escherichia coli CFT073	ecc
Escherichia coli K-12 MG1655	eco
Escherichia coli O157 EDL933	ece
Escherichia coli O157 Sakai	ecs
Haemophilus ducreyi	hdu
Haemophilus influenzae	hin
Helicobacter hepaticus	hhe
Helicobacter pylori 26695	hpy
Helicobacter pylori J99	hpj
Mesorhizobium loti	mlo
Neisseria meningitidis MC58 (serogroup B)	nme
Neisseria meningitidis Z2491 (serogroup A)	nma
Nitrosomonas europaea	neu
Pasteurella multocida	pmu
Pseudomonas aeruginosa	pae
Pseudomonas putida	ppu
Pseudomonas syringae pv. tomato	pst
Ralstonia solanacearum	rso
Rickettsia conorii	rco
Rickettsia prowazekii	rpr
Salmonella typhi CT18	sty
Salmonella typhi Ty2	stt
Salmonella typhimurium	stm
Shewanella oneidensis	son
Shigella flexneri 301 (serotype 2a)	sfl
Sinorhizobium meliloti	sme
Vibrio cholerae	vch
Vibrio parahaemolyticus	vpa
Vibrio vulnificus	vvu
Xanthomonas axonopodis	xac
Xanthomonas campestris	xcc
Xylella fastidiosa 9a5c	xfa
Xylella fastidiosa Temecula1	xft
Yersinia pestis CO92	ype
Yersinia pestis KIM	ypk

Table 4: Lists showing sets of gram-positive bacteria and Proteobacteria used in this study. *(The bacteria *Bordetella bronchiseptica* has mistakenly been given a faulty abbreviation in some of the analysis).

NUC files, which are multiple nucleotide FASTA files, were downloaded from the KEGG database at ftp://ftp.genome.ad.jp/pub/kegg/. Each NUC file contains every annotated gene for a certain bacteria, as a result each completely sequenced bacteria has its own NUC file (all NUC files used in this study are included on the DVD). All files were uploaded to the Biotin server for further analysis.

Before running the program "Extract", a few script adjustments had to be made in order to locate the input/output directories. In addition, two different versions of Extract were constructed to obtain as many 16S rRNA genes a possible, respectively "extractA" and "extractB". Version A contains the search strings "16S", "RNA" and "RIBOSOMAL" and only genes annotated with these three words were extracted. Version B contains the string "16S_". Due to insufficient naming only 65 out of 80 bacteria had their 16S sequences extracted. It seemed to be virtually impossible to extract genes for the remaining bacteria, even by manual inspection.

"Extract" overwrites every file in the output folder after each execution. Since the program had to be run two times (version A, B), a temporary folder was made. Files in the "extract" output folder were moved to the temporary folder after the first computation to conserve the files. Every output file from "extract" was automatically given the extension ".16s" by the application. The programs were executed as shown below:

```
$ mkdir output16s
$ ./extractA.pl
$ mkdir output16sTmp
$ mv output16s/* output16sTmp/
$ ./extractB.pl
$ mv output16sTmp/* output16s/
$ rm output16s/b.melitensis.nuc.16s
```

Before "conruns.pl" was executed, a couple enhancements had to be made. This involved deleting the file "b.melitensis.nuc.16s", since in did not contain any 16S rRNA genes. Removing a faulty 16S rRNA gene in the 16s file for the bacteria *S.flexneri*, so that a consensus could be calculated. And finally in the 16S file for *S.coelicolor* a gene annotated as probable was removed since it appeared to disturb the final consensus

sequence. Here the experiment took two different directions, one in which the 16S rRNA consensus for each bacteria were obtained, and one in which all 320 16S rRNA genes were concatenated and aligned directly, using ClustalX. The concatenated file containing all 320 genes was named “ClustalX16Sallegener” and will be discussed at the ending of this chapter.

All of the 65 multiple FASTA files, containing one or more 16S rRNA gene, were then run through another script called consrun. By doing this a consensus sequence was made for each bacteria and given the extension “.cons”. Cons only outputs a file if the input contains two or more 16S rRNA sequences (genes), so for species that only has one 16S rRNA gene there will not be an output file. For all species that both had a “16s” file and a “cons” file the 16s. file was deleted, leaving only the consensus file. (See list with command lines below).

$ ./consruns.pl
Creates a consensus from multiple alignments
.....
Creates a consensus from multiple alignments
$ rm output16s/b.anthracis.nuc.16s
.
.
.
$ rm output16s/y.pestis.nuc.16s
$ rm output16s/y.pestis_kim.nuc.16s

When all unwanted files were removed the remaining files were concatenated and downloaded to a notebook for further analysis. (See command line)

$ cat output16s/ > ClustalXKomplettCons*

Sequences for the bacteria *H.pylori*, *S.coelicolor*, and *R.solanacearum* were removed because they contained more than 50% unknown characters. After removing unwanted sequences the remaining FASTA file contained 16S rRNA genes from 61 different bacteria.

These genes were aligned by ClustalX using default settings. The final alignment was uploaded to the Gblocks server for extraction of conserved regions (the final sequence alignments, both before and after block extraction, are included on the DVD). The resulting file was converted into the Nexus format, at the ReadSeq server, and some of the genes had to be renamed due to restrictions concerning this format.
Before executing the file in PAUP a "modelblock" was inserted at the end of the file, see 4.2.9 for further details. Finally the symbol "*" was replaced by "@" since the program mistakenly analyzed it as being a character. After execution the following parameters were obtained by Modeltest:

Base frequencies:
A = 0.2511
C = 0.2010
G = 0.2927
T = 0.2552

Substitution model:
R(a) [A-C] = 1.0000
R(b) [A-G] = 2.8866
R(c) [A-T] = 1.0000
R(d) [C-G] = 1.0000
R(e) [C-T] = 3.9617
R(f) [G-T] = 1.0000

Proportion of invariable sites (I) = 0.4753
Gamma distribution shape parameter (G) = 0.7353

The Tamura and Nei model was selected to be the most appropriate substitution model for this particular set of sequences (Tamura and Nei, 1993).

When running the analysis, using the Tamura and Nei model, maximum likelihood would probably have been the method of choice, but it proved to be extremely computer intensive.
So instead a heuristic distance tree was computed using maximum likelihood distance measures, with parameters given by "modeltest". PAUP was set to generate random-sequence starting trees (10 replicates) by stepwise addition, using the tree-bisection-

reconnection (TBR) branch-swapping algorithm. Finally the tree was bootstrapped using 100 replicates. The final tree was saved and edited in Treeview.

The file containing all 320 16S rRNA genes was treated in the exact same manner as the 16S rRNA consensus file, using ClustalX, Gblocks, ReadSeq, Modeltest and PAUP (the final sequence alignments, before and after block extraction, are included on the DVD). After execution in PAUP the following parameters were obtained by Modeltest:

Base frequencies:
A = 0.2500
C = 0.2500
G = 0.2500
T = 0.2500

Substitution model:
R(a) [A-C] = 1.0000
R(b) [A-G] = 3.0400
R(c) [A-T] = 1.0000
R(d) [C-G] = 1.0000
R(e) [C-T] = 3.6190
R(f) [G-T] = 1.0000

Proportion of invariable sites (I) = 0.0851
Gamma distribution shape parameter (G) = 0.4741

The Tamura and Nei model, using equal base frequencies, was selected to be the most appropriate substitution model for this particular set of sequences (Tamura and Nei, 1993).
Due to a much larger set of sequences a less time consuming algorithm had to be employed. Thus a Neighbor-joining tree using maximum likelihood distance measures with parameters given by “Modeltest” was computed and bootstrapped with 100 replicates.

4.3.2 Phylogenetic analysis of the ATP synthase alpha chain gene

The gene encoding the ATP synthase α chain has proven to be suitable in phylogenetic studies (Gogarten *et al*., 1992). All genes where obtained from the KEGG database at ftp://ftp.genome.ad.jp/pub/kegg/ by downloading a PEP file for each bacteria. A PEP file is a multiple FASTA file containing every annotated protein in a particular bacterium genome (all PEP files used in this study are included on the DVD). All 79 files were uploaded to the Biotin server. The program "Extract" had to be modified in order to locate the ATPase genes and the output/input folders. The search strings were set to "ATP", "ALPHA" and "[EC:3.6.3.14]" (the enzyme number), in addition the output extension was changed to ".ATPase". The program was given the name "extractATPaseA.pl", and executed in the following way:

$ mkdir ATPase
$ ./extractATPaseA.pl
$ cat ATPase/ > ClustalXATPase*

The ClustalXATPase file was downloaded to a notebook and edited with a text editor. For both *Streptococcus pyogenes SF370* and *Streptococcus pyogenes MGAS315*, a sequence encoding a Na^+ driven ATPase was removed. Further on, two duplicated genes described as "similar to ATP synthase alpha chain" found in *Listeria innocua* and *Listeria monocytogenes* were deleted. Finally, one of two sequences was removed from the bacterium *Lactococcus lactis* since it turned out to be encoding the ATPase beta chain gene.

After deleting these five sequences the ATPase alpha chain sequences for *Haemophilus ducreyi* and *Helicobacter hepaticus* had to be manually inserted into the FASTA file due to lack of annotation. *Clostridium tetani* was not included in this alignment because it only has a Na^+ driven ATPase.

A total of 78 ATPase genes from 78 different bacteria were saved and later aligned in ClustalX. Genes described as putative were given a "*" and those described as probable with "**". Only default settings were used in ClustalX. Blocks were obtained at the Gblocks server and the file format converted to Nexus, using ReadSeq (the final sequence alignments, before and after block extraction, are included on the DVD).

The ATPase alignment was analyzed using three different phylogenetic programs, TREE-PUZZLE, MrBayes and the PHYLIP package (see flowchart below). Although

TREE-PUZZLE probably is the most suitable program for doing protein analysis, it turned out to be too time consuming to do a bootstrapping using this method. Instead a bootstrapped tree was computed by the PHYLIP package, using the Neighbor-joining method. In addition a consensus tree was computed by MrBayes. This method is summarized in the flowchart below.

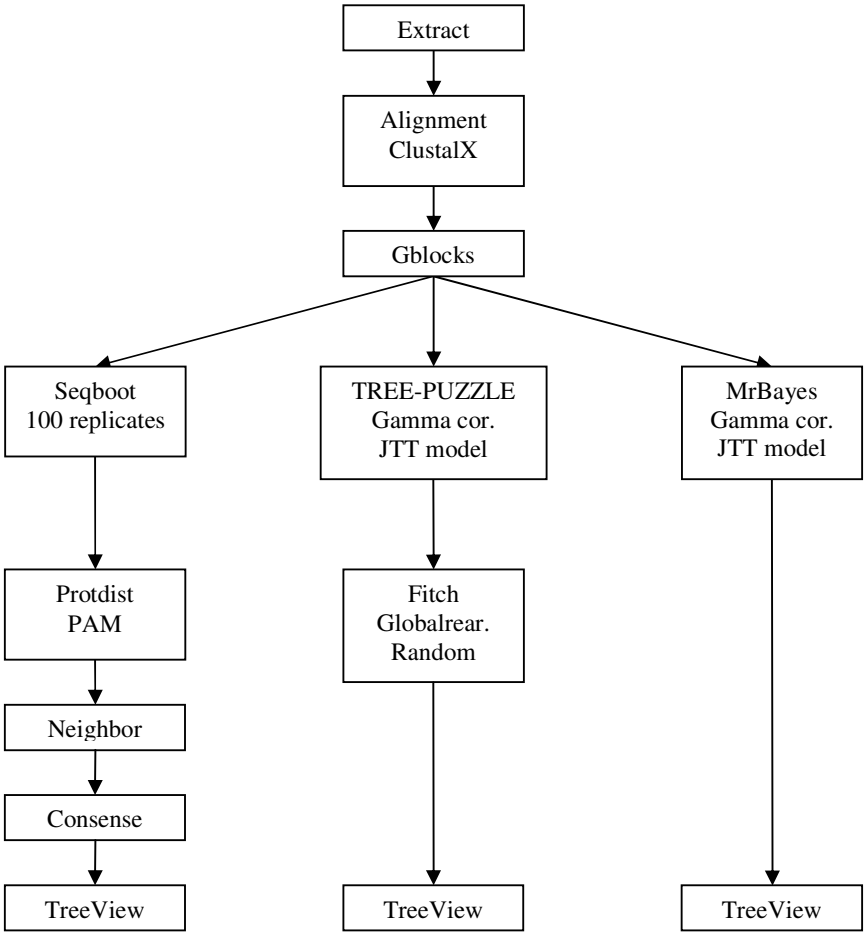


Figure 7: Flow chart showing the phylogenetic classification of protein sequences, and how this analyzes were divided into three different techniques subsequent to block extraction using Gblocks. Finally, trees from all three methods were constructed in TreeView.

The file was uploaded to the Macduff server and analyzed using MrBayes. The program was set to use the JTT substitution model (Jones *et al.*, 1992), gamma correction (using invgamma), and to for 1.000.000 generations. To reduce the risk of generating trees fixed on local tops it is necessary to run the analysis two times, each execution taking more than 70 hours. After the first run the command "sump" was given, revealing a plot

showing the probability of observing the data versus the generation. In this way it is possible to determine what the burn-in value for the analysis should be, thus discarding trees generated before a steady state has been reached. In this case the burn-in value was set to 500, discarding the first 5% of the generated trees. In the next step the command "sumt burnin = 500" was given, generating a consensus-tree from the remaining 95% of the trees. Finally the file generated in this last step can be used as input for Treeview. Each clade contains a probability, a number between zero and one, to determine the reliability of the clades, resembling the bootstrapping algorithm. MrBayes was executed on more time in order to verify the consistency of the tree; the two trees were compared and found to be nearly identical.

The output from Gblocks was also converted into the PHYLIP format and executed in TREE-PUZZLE on the Biotin server. The program was set to use the JTT substitution model and gamma distribution with four gamma rate categories. The output file from TREE-PUZZLE can be used directly as input in FITCH. Here the option, global rearrangement was activated and the species input order was randomized 10 times to make the final tree more reliable.

In a third approach programs from the PHYLIP package were used. In the first step the output from Gblocks was converted into the PHYLIP format and executed in SEQBOOT, on the Macduff server, giving 100 resampled versions of the original alignment. In the second step the output from SEQBOOT was analyzed in PROTDIST to generate 100 distance matrixes, one for each set, using the PAM substitution model. In the next step the files generated by PROTDIST were used as input in NEIGHBOR, thereby calculating 100 phylogenetic trees, the species input order was randomized to increase the reliability of the final trees. Finally the program CONSENSE was executed, generating a bootstrapped output.

4.3.3 Phylogenetic analysis of the Prolyl-tRNA synthetase gene

A new version of “Extract” was made, “extractPro.pl”, and the same PEP files as described above were analyzed. This version of “extract” made use of the search strings “PRO”, “TRNA” and “[EC:6.1.1.15]” in order to extract the prolyl-tRNA synthetase sequences. The location of the output directory was written into the application and named “Prolyl”. The application was executed in the following way:

$ mkdir Prolyl
$ ./extractPro.pl
$ cat Prolyl/ > ClustalXPro*

The concatenated file ClustalXPro was downloaded to a notebook an edited. All 78 bacteria had their genes extracted, except from the bacterium *Haemophilus ducreyi* which had its prolyl-tRNA gene inserted manually into the FASTA file. Both *Bacillus anthracis* and *Bacillus cereus* had two versions of the tRNA synthetase gene, however these sequences were included in the final alignment. The final input file for ClustalX contained 80 genes from 78 different bacteria. Genes described as putative were given a “*” and those described as probable with “**”. ClustalX was executed with standard settings and the output was saved as a FASTA file. When obtaining blocks at the Gblocks server the options ”Allow smaller final blocks” and “Allow gap positions within the final blocks” had to be employed in order to get reasonable sized blocks (the final sequence alignments, both before and after block extraction, are included on the DVD). The file was converted into the PHYLIP format and uploaded to the Macduff server. The file was analyzed using the program TREE-PUZZLE and FITCH, and the PHYLIP method (using SEQBOOT, PROTDIST, NEIGHBOR and CONSENSE). All programs used the same settings as when conducting the ATPase analysis. In addition the Gblocks output was converted into the Nexus format and used as input for MrBayes. The burnin-value was set to 1.000 instead of 500, the rest of the settings were identical to those used when comparing the ATPase genes. Trees from the first and second execution turned out to be almost identical.

4.3.4 Phylogenetic analysis of the Methionyl-tRNA synthetase gene

A fifth version of “Extract” was made, containing the search strings "MET", "TRNA", and "[EC:6.1.1.10]". The application was named “extractMet.pl. Sequences from all bacteria were obtained, except from *Haemophilus ducreyi*, which had its sequence inserted manually. The application was executed in the following way:

$ mkdir Methionyl
$ ./extractMet.pl
$ cat Methionyl / > ClustalXMet*

When editing the file a “putative” gene for the bacteria *Bacillus anthracis* was removed, and for *Clostridium perfringens* a “probable” gene had to be deleted. For *Bacillus cereus* two tRNA synthase genes were extracted and both were included. Also *Ralstonia solanacearum* had two genes encoding this enzyme, but both were annotated as probable, however they were included in the final file. A total of 80 genes were aligned with ClustalX. The rest of the analyses were done in the same manner as with the other genes (the final sequence alignments, both before and after block extraction, are included on the DVD).

4.4 Classification using 10-mer oligonucleotides

In this section a method for classification of bacteria using the genome frequencies of 10-mer oligonucleotides will be demonstrated. The procedure is divided into four different branches, as shown in the flowchart in **Figure 8**. First suitable genome sets for primer selection have to be assembled, 10-mer frequencies determined and finally informative primers extracted, as illustrated by the upper left branch in the flowchart. This will be done by using the programs Extseq, Gencnt and Selentprim, see program description for further details.

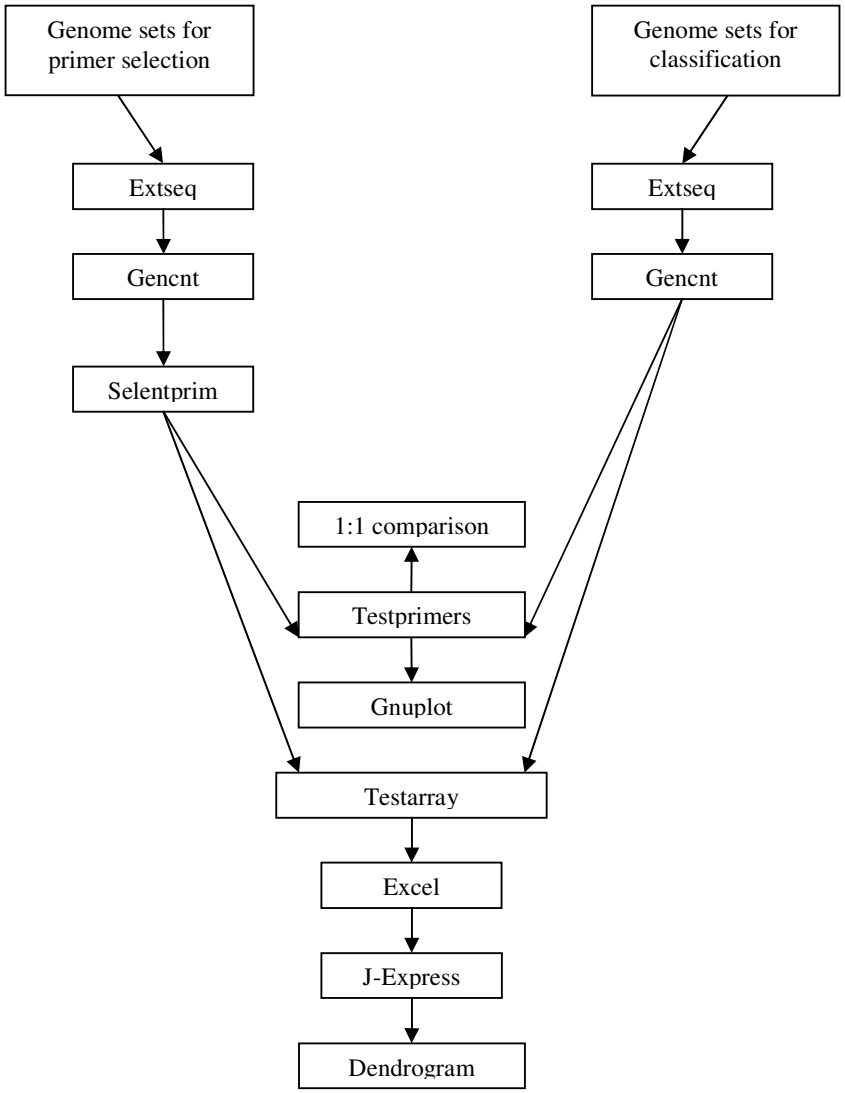


Figure 8: Flowchart showing the procedure for selection of primer sets and classification of bacterial genomes. Testarray takes two input files, the output from Selentprim and Gencnt, and outputs the result into a third file. As an alternative a gnuplot between the two files or a 1:1 comparison between all genomes, might be constructed, with the program Testprimers.

The next step, indicated by the upper right branch in the flowchart above, utilize computation of 10-mer oligonucleotide frequencies in genomes to be classified, thus ending with the output from Gencnt. The output from these two branches (the output from Selentprim and Gencnt) is combined in Testarray and further analyzed using Excel and J-Express, as shown in the lower part of the flowchart. An alternative these two files might be executed in Testprimers enabling the construction of gnuplots or 1:1 comparison. The programs Extseq, Gencnt, Selentprim, Testprimers, Gnuplot and the 1:1 comparison were all executed on the Darwin server. Revperl was executed on the Biotin server, and analysis using J-Express and Excel was performed on a laptop.

4.4.1 Selection of organisms and evaluation of genome sets

This method relies on the selection of 4.000 primers that will be able to discriminate between different bacterial species and strains. The selection of genomes, from which these primers will be extracted, is important. As the number of genomes in the extraction set increase, the number of possible 10-mer oligonucleotides becomes higher. In theory a genome a little larger than 1 million base pairs is sufficient to include all possible 10-mer oligonucleotides, if each 10-mer only occurs once. Still, since many 10-mers occur more than once, as can be seen by making a 10-mer oligonucleotide frequency plot (with the EMBOSS program Compseq), the size of a genome containing all possible 10-mers probably has to be significantly larger than one Mb. (Mb refers to Megabase pair of DNA, while MB is the abbreviation for Megabytes of data. When using FASTA files one MB of data is approximately equal to one Mb of DNA).
A three dimensional plot showing numbers of primers versus megabytes of DNA versus species combinations was constructed to see how many genomes or Megabytes (MB) of DNA data that is needed to include all possible 10-mer oligonucleotides. Three different groups of bacteria were selected, a gram-positive group, a group containing Proteobacteria and a mixed species group. Within these three groups sets containing 10, 20, 40 and 80 MB of DNA were made. The 20 MB set contains all bacteria in the 10 MB set, the 40 MB set contains all bacteria in the 20 MB and so on. All sets can be accessed on the included DVD, in the “Primers_vs_MB_vs_Species” folder. These sets were uploaded to the Darwin server and executed using Extseq, Gencnt and finally Selentprim. In this way all primers with unacceptable energies and palindromes were filtered out, and the number of total available 10-mer oligonucleotides was given by Selentprim. See **Figure 14** (page 67).

Another plot was constructed to reveal the correlation between MB of DNA data and the entropy interval needed to give 4000 primers (+/- 10 primers). The experiment was conducted by generating different sized genome sets ranging from 10-180 MB (the sets can be accessed on the included DVD in the "MB_of_Genomes_vs_Entropy" folder). These sets were analyzed using Extseq, Gencnt and Selentprim was executed with GC ratio set to 4-5 and f = 2, resembling typical settings later used in the study. The lower entropy was set to 0.0 while the upper entropy (Y-axis) and the genome sets were changed (X-axis) in order to extract 4000 primers. See **Figure 15** (page 68).

Based on the analysis made above, five sets were chosen to be used in the final primer extraction. In order to select a satisfactory number of informative primers all genome sequences involved in this selection has to be organized into sufficiently large sets. Since the size of the genome sets had a significant effect on the primer selection, four different sets were constructed, being approximately 40MB and 80MB (+/- 0.1 MB). Because one MB (Megabytes) of data is roughly one Mb (Megabase) of DNA, the sets have been assembled according to MB of data (these sets are included on the DVD in the "Genome_sets_for_primer_extraction" folder). The fifth set, named EcoSalmoFlex, contains 9 closely related species and size criteria were not applied (see **Table 3** on page 38). Species in the two gram-positive and Proteobacteria sets were selected in a way that best represents the diversity of the group (see **Table 5** and **Table 6**). To reveal the correlation between entropy, the minimum frequency "f" and the number of extracted primers in these final genome sets, four three-dimensional plots were made. Since all four plots reveals the same tendency only diagrams for the Proteobacteria are included in this thesis (**Figure 16** and **Figure 17**, page 69).

40 MB Gram-positive	
Bacillus anthracis	ban
Bacillus halodurans	bha
Bifidobacterium longum	blo
Clostridium tetani	ctc
Corynebacterium glutamicum	cgl
Enterococcus faecalis	efa
Lactococcus lactis	lla
Listeria innocua	lin
Mycobacterium tuberculosis H37Rv (lab strain)	mtu
Oceanobacillus iheyensis	oih
Staphylococcus aureus N315 (MRSA)	sau
Streptococcus mutans	smu
Streptococcus pneumoniae R6	spr

40 MB Proteobacteria	
Bordetella pertussis	bpe
Campylobacter jejuni	cje
Coxiella burnetii	cbu
Escherichia coli CFT073	ecc
Helicobacter pylori J99	hpj
Neisseria meningitidis MC58 (serogroup B)	nme
Nitrosomonas europaea	neu
Pseudomonas putida	ppu
Pseudomonas syringae pv. tomato	pst
Rickettsia conorii	rco
Vibrio parahaemolyticus	vpa
Xylella fastidiosa Temecula1	xft

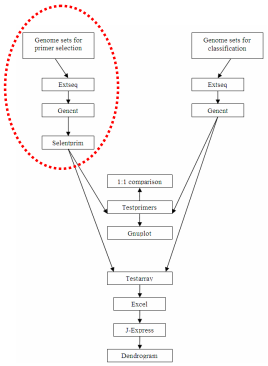
Table 5: Genome sets containing a total of 40 MB (Megabytes) DNA. The set to the left contains 13 gram-positive bacterial genomes, while the set to the right contains 12 genomes from Proteobacteria.

80 MB Gram-positive	
Bacillus anthracis	ban
Bacillus halodurans	bha
Bacillus subtilis	bsu
Bifidobacterium longum	blo
Clostridium acetobutylicum	cac
Clostridium tetani	ctc
Corynebacterium glutamicum	cgl
Enterococcus faecalis	efa
Lactobacillus plantarum	lpl
Lactococcus lactis	lla
Listeria innocua	lin
Listeria monocytogenes	lmo
Mycobacterium tuberculosis H37Rv (lab strain)	mtu
Oceanobacillus iheyensis	oih
Staphylococcus aureus N315 (MRSA)	sau
Staphylococcus epidermidis	sep
Streptococcus agalactiae 2603	sag
Streptococcus mutans	smu
Streptococcus pneumoniae R6	spr
Streptococcus pyogenes MGAS8232 (serotype)	spm
Streptomyces avermitilis	sma
Streptomyces coelicolor	sco
Thermoanaerobacter tengcongensis	tte

80 MB Proteobacteria	
Bordetella pertussis	bpe
Campylobacter jejuni	cje
Caulobacter crescentus	ccr
Coxiella burnetii	cbu
Escherichia coli CFT073	ecc
Escherichia coli O157 EDL933	ece
Helicobacter pylori J99	hpj
Mesorhizobium loti	mlo
Neisseria meningitidis MC58 (serogroup B)	nme
Nitrosomonas europaea	neu
Pseudomonas aeruginosa	pae
Pseudomonas putida	ppu
Pseudomonas syringae pv. tomato	pst
Ralstonia solanacearum	rso
Rickettsia conorii	rco
Shewanella oneidensis	son
Shigella flexneri 301 (serotype 2a)	sfl
Vibrio parahaemolyticus	vpa
Xanthomonas axonopodis	xac
Xylella fastidiosa Temecula1	xft

Table 6: Genome sets containing a total of 80 MB (Megabytes) DNA. The set to the left contains 23 gram-positive bacterial genomes, while the set to the right contains 20 genomes from Proteobacteria.

4.4.2 Construction of different primer sets

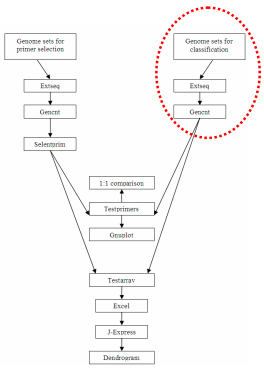


The figure to the left shows a miniaturized picture of the flowchart in **Figure 8**. The branch described in this section is indicated by the dotted circle.

The sets assembled above were then used with Extseq, Gencnt and informative primers extracted using Selentprim. Determining the optimal minimum frequency (the f-value in Selentprim) is difficult. As a result, multiple primer sets were constructed for each genome set; keeping the entropy interval within 0.3 and 0.7, the GC ratio at 4-5 and varying the f-value in order to extract 4000 primers (see **Table 10**, page 70). Finally these primer sets were saved, in total 16, and further analyzed as described below.

An important fact that has to be accounted for is the relatively small number of 10-mer oligonucleotides that actually will be available after filtration. As mentioned above (see 4.2.2) only a little more than 700.000 primers will remain after removing those with unfavorable energies and palindromes. Since the differences in GC ratio has to be as small as possible, to avoid incomplete hybridization, only a fraction of the 700.000 10-mers can be chosen, dramatically reducing the number of primers. As a consequence some parameters might have to be set to less optimal than preferable, to extract a sufficient number of primers.

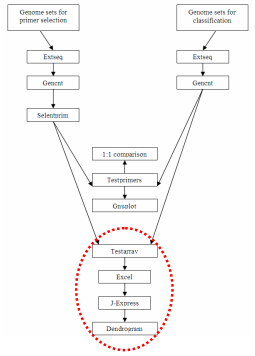
4.4.3 Computation of 10-mer frequencies in genomes to be classified



The figure to the left shows a miniaturized picture of the flowchart in **Figure 8**. The branch described in this section is indicated by the dotted circle.

To test the selected primer sets, two large sets of genomes, one containing Gram-positive bacteria, and one with Proteobacteria (see **Table 4** on page 40) were assembled. For these genome sets the reverse compliment sequences were computed from the original FASTA input files, using revperl.pl (for further details on revperl.pl see 4.2.5.1), to make both DNA strands available. By making the genome sequences double-stranded the number of 10-mer oligonucleotides available for hybridization will be the same as it would be *in vitro*, when conducting a real microarray experiment. The new sets containing the double-stranded genomes were then executed in Extseq and Gencnt to calculate 10-mer oligonucleotide frequencies for every genome. The outputs from Gencnt were later used as one of two input files in Testarray.

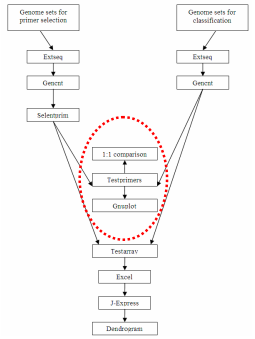
4.4.4 Analyzing output from Selentprim and Gencnt in Testarray



The figure to the left shows a miniaturized picture of the flowchart in **Figure 8**. The branch described in this section is indicated by the dotted circle.

Using the program Testarray, files containing 10-mer oligonucleotide frequencies from the genomes that are to be classified (the output from Gencnt) and the primer sets made by Selentprim are executed together to mimic the hybridization occurring when using an actual microarray (as shown in the flowchart). Only Testarray-v2, which normalizes the primer frequencies according to genome size, was used in these final analyses. The output from Testarray was later edited in Excel and executed in J-Express in order to cluster the array data. In order to find the algorithm that best corresponds with our data the sets were analyzed using the clustering algorithms UPGMA, WPGMA, Single-linkage or Complete-linkage, and a variety of correlations or distance measurements. Based on comparing the different dendrograms generated from these algorithms, Pearson correlation and UPGMA proved to be most suitable in comparing species at the strain level. WPGMA and Canberra distance measures or Pearson correlation seems to be most appropriate when making a global tree for all Proteobacteria or Gram-positive bacteria. These decisions were made by comparing the dendrograms to the phylogenetic trees by visual inspection. The data was also analyzed using κ–means clustering and SOM.

4.4.5 Making gnuplots and doing 1:1 comparison



The figure to the left shows a miniaturized picture of the flowchart in **Figure 8**. The branch described in this section is indicated by the dotted circle.

As an alternative to Testarray and J-Express the output from Selentprim and Gencnt can be executed in Testprimers, in order to make a gnuplot or doing a 1:1 comparison. Many gnuplots have been made during this study, in addition to the 1:1 comparison, but only one of each are included in the results.

5 Results and discussion

Due to the large amount of data produced during this study, and the need for direct comparison between the dendrograms and the reference trees, it is more convenient to have results and the discussion in the same chapter. This chapter starts with a presentation of the reference trees, followed by a discussion concerning their quality. Then, results produced by the oligonucleotide classification method are shown, compared to the reference trees and finally discussed.

5.1 Results and discussion of the phylogenetic reference trees

In this section results from the phylogenetic classification are represented.
Figure 9 shows the intermingling of 16S rRNA genes between closely related species in the EcoSalmoFlex (see **Table 3**, page 38), and explains the need for calculating the 16S rRNA consensus sequences for the different organisms. A tree made from these consensus sequences and a tree holding all 320 sequences, from 61 bacteria, were constructed.

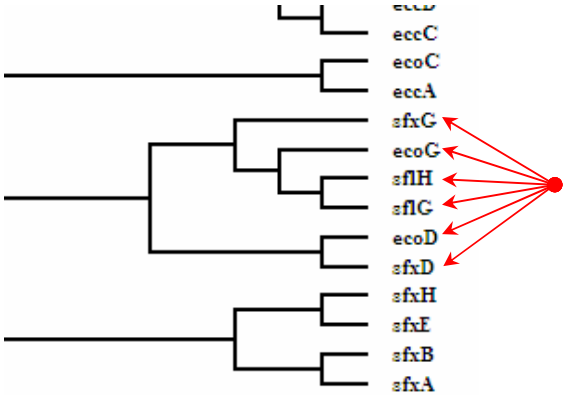


Figure 9: Part of phylogenetic tree showing 16S rRNA genes from *E.coli* and *S.flexneri* intermingling. (Screenshot from Treeview).

In the following section the phylogenetic trees are presented. Due to the need for easy file handling, all species in these trees are named by a three letter abbreviation. To facilitate interpretation lists with abbreviations and the corresponding bacterial names are given on the opposite page of the phylogenetic tree. Similar lists are also included when the dendrograms are presented in section 5.2.1.

The smallest 16S rRNA tree containing 61 taxas (see **Figure 10**) was calculated using a distance with parameters estimated by Modeltest, the set was randomly generated 10 times and bootstrapped with 100 replicates. For information on specific settings and methods see 4.2.9.

Gram-positive bacteria	
ban	*Bacillus anthracis*
bce	*Bacillus cereus*
bha	*Bacillus halodurans*
blo	*Bifidobacterium longum*
bsu	*Bacillus subtilis*
cac	*Clostridium acetobutylicum*
cef	*Corynebacterium efficiens*
cgl	*Corynebacterium glutamicum*
cpe	*Clostridium perfringens*
ctc	*Clostridium tetani*
efa	*Enterococcus faecalis*
lin	*Listeria innocua*
lla	*Lactococcus lactis*
lmo	*Listeria monocytogenes*
lpl	*Lactobacillus plantarum*
mbo	*Mycobacterium bovis*
mle	*Mycobacterium leprae*
mtc	*Mycobacterium tuberculosis CDC1551*
mtu	*Mycobacterium tuberculosis H37Rv (lab strain)*
oih	*Oceanobacillus iheyensis*
sag	*Streptococcus agalactiae 2603*
sam	*Staphylococcus aureus MW2*
san	*Streptococcus agalactiae NEM316*
sau	*Staphylococcus aureus N315 (MRSA)*
sav	*Staphylococcus aureus Mu50 (VRSA)*
sco	*Streptomyces coelicolor*
sep	*Staphylococcus epidermidis*
sma	*Streptomyces avermitilis*
smu	*Streptococcus mutans*
spg	*Streptococcus pyogenes MGAS315 (serotype M3)*
spm	*Streptococcus pyogenes MGAS8232 (serotype M18)*
spn	*Streptococcus pneumoniae TIGR4*
spr	*Streptococcus pneumoniae R6*
sps	*Streptococcus pyogenes SSI-1 (serotype M3)*
spy	*Streptococcus pyogenes SF370 (serotype M1)*
tte	*Thermoanaerobacter tengcongensis*

Proteobacteria	
bbr/bre	*Bordetella bronchiseptica*
bme	*Brucella melitensis*
bms	*Brucella suis*
bpa	*Bordetella parapertussis*
bpe	*Bordetella pertussis*
cbu	*Coxiella burnetii*
ccr	*Caulobacter crescentus*
cje	*Campylobacter jejuni*
ecc	*Escherichia coli CFT073*
ece	*Escherichia coli O157 EDL933*
eco	*Escherichia coli K-12 MG1655*
ecs	*Escherichia coli O157 Sakai*
hdu	*Haemophilus ducreyi*
hhe	*Helicobacter hepaticus*
hin	*Haemophilus influenzae*
hpj	*Helicobacter pylori J99*
hpy	*Helicobacter pylori 26695*
mlo	*Mesorhizobium loti*
neu	*Nitrosomonas europaea*
nma	*Neisseria meningitidis Z2491 (serogroup A)*
nme	*Neisseria meningitidis MC58 (serogroup B)*
pae	*Pseudomonas aeruginosa*
pmu	*Pasteurella multocida*
ppu	*Pseudomonas putida*
pst	*Pseudomonas syringae pv. tomato*
rco	*Rickettsia conorii*
rpr	*Rickettsia prowazekii*
rso	*Ralstonia solanacearum*
sfl	*Shigella flexneri 301 (serotype 2a)*
sme	*Sinorhizobium meliloti*
son	*Shewanella oneidensis*
stm	*Salmonella typhimurium*
stt	*Salmonella typhi Ty2*
sty	*Salmonella typhi CT18*
vch	*Vibrio cholerae*
vpa	*Vibrio parahaemolyticus*
vvu	*Vibrio vulnificus*
xac	*Xanthomonas axonopodis*
xcc	*Xanthomonas campestris*
xfa	*Xylella fastidiosa 9a5c*
xft	*Xylella fastidiosa Temecula1*
ype	*Yersinia pestis CO92*
ypk	*Yersinia pestis KIM*

Table 7: Lists showing sets of gram-positive bacteria and Proteobacteria used in this study. Due to faulty annotation in the NUC files, not all species in these lists are included in the 16S rRNA tree. *(The bacteria *Bordetella bronchiseptica* has mistakenly been given a faulty abbreviation in some of the analysis).

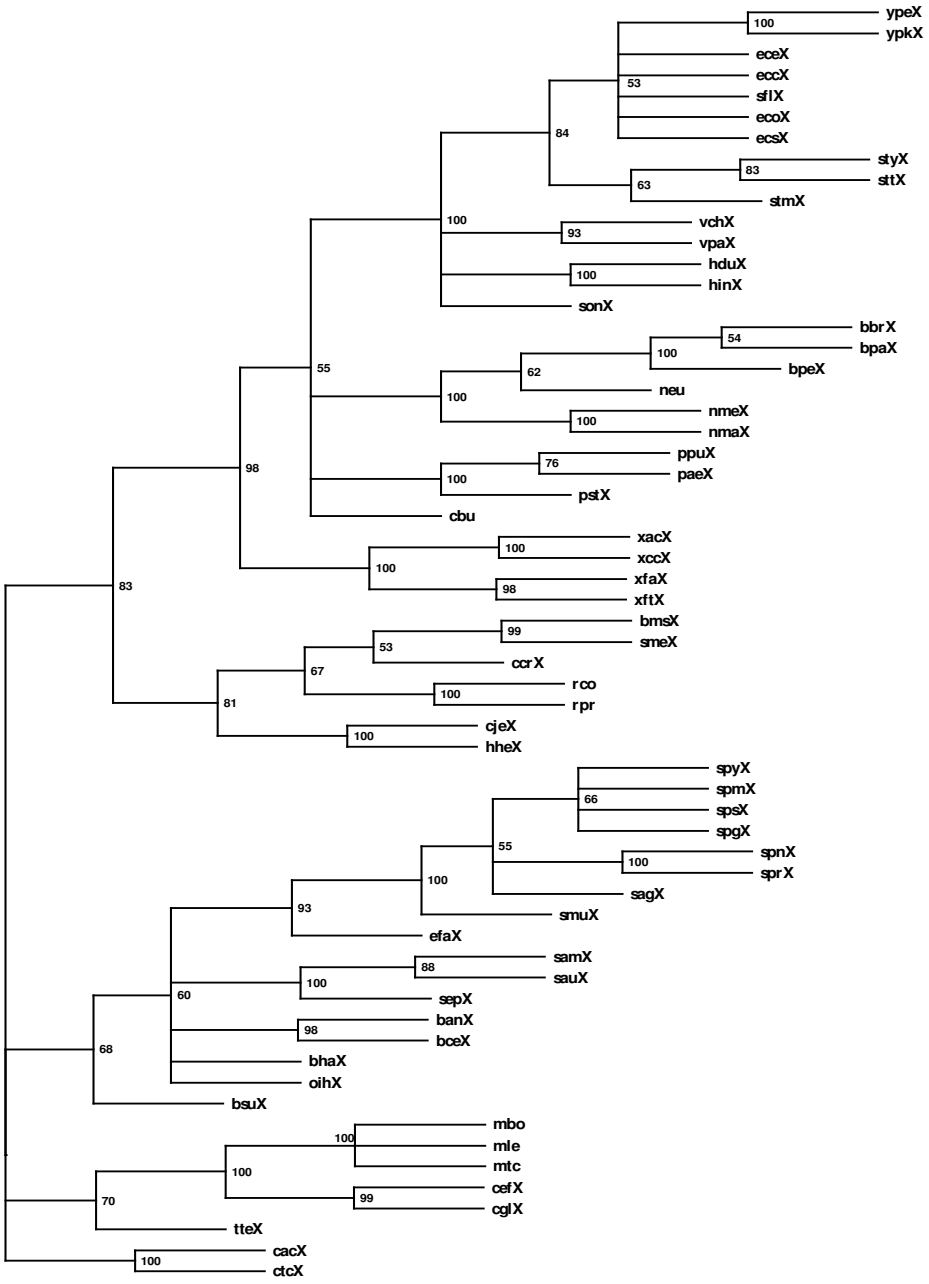


Figure 10: 16S rRNA consensus phylogenetic tree constructed using ClustalX, Gblocks, Modeltest and PAUP, with bootstrapping values. X indicates genes annotated as putative, and XX when annotated as probable.

The larger tree containing all 320 16S genes also had its parameters generated by Modeltest (see 4.2.9 for settings), and its phylogeny determined by the Neighbor Joining method, finally it was bootstrapped. Due to the total size of all 320 branches the tree is impractically large and should be viewed directly from its file on the included DVD.

The first ATP synthase alpha chain tree presented here was constructed using MrBayes. The data was executed two times (generating two trees) in order to order to verify the consistency of the tree. The execution of MrBayes took more than 140 hours (2 X 70 hours), on a 2.26 GHz Pentium 4.

Gram-positive bacteria	
ban	*Bacillus anthracis*
bce	*Bacillus cereus*
bha	*Bacillus halodurans*
blo	*Bifidobacterium longum*
bsu	*Bacillus subtilis*
cac	*Clostridium acetobutylicum*
cef	*Corynebacterium efficiens*
cgl	*Corynebacterium glutamicum*
cpe	*Clostridium perfringens*
ctc	*Clostridium tetani*
efa	*Enterococcus faecalis*
lin	*Listeria innocua*
lla	*Lactococcus lactis*
lmo	*Listeria monocytogenes*
lpl	*Lactobacillus plantarum*
mbo	*Mycobacterium bovis*
mle	*Mycobacterium leprae*
mtc	*Mycobacterium tuberculosis CDC1551*
mtu	*Mycobacterium tuberculosis H37Rv (lab strain)*
oih	*Oceanobacillus iheyensis*
sag	*Streptococcus agalactiae 2603*
sam	*Staphylococcus aureus MW2*
san	*Streptococcus agalactiae NEM316*
sau	*Staphylococcus aureus N315 (MRSA)*
sav	*Staphylococcus aureus Mu50 (VRSA)*
sco	*Streptomyces coelicolor*
sep	*Staphylococcus epidermidis*
sma	*Streptomyces avermitilis*
smu	*Streptococcus mutans*
spg	*Streptococcus pyogenes MGAS315 (serotype M3)*
spm	*Streptococcus pyogenes MGAS8232 (serotype M18)*
spn	*Streptococcus pneumoniae TIGR4*
spr	*Streptococcus pneumoniae R6*
sps	*Streptococcus pyogenes SSI-1 (serotype M3)*
spy	*Streptococcus pyogenes SF370 (serotype M1)*
tte	*Thermoanaerobacter tengcongensis*

Proteobacteria	
bbr/bre	*Bordetella bronchiseptica*
bme	*Brucella melitensis*
bms	*Brucella suis*
bpa	*Bordetella parapertussis*
bpe	*Bordetella pertussis*
cbu	*Coxiella burnetii*
ccr	*Caulobacter crescentus*
cje	*Campylobacter jejuni*
ecc	*Escherichia coli CFT073*
ece	*Escherichia coli O157 EDL933*
eco	*Escherichia coli K-12 MG1655*
ecs	*Escherichia coli O157 Sakai*
hdu	*Haemophilus ducreyi*
hhe	*Helicobacter hepaticus*
hin	*Haemophilus influenzae*
hpj	*Helicobacter pylori J99*
hpy	*Helicobacter pylori 26695*
mlo	*Mesorhizobium loti*
neu	*Nitrosomonas europaea*
nma	*Neisseria meningitidis Z2491 (serogroup A)*
nme	*Neisseria meningitidis MC58 (serogroup B)*
pae	*Pseudomonas aeruginosa*
pmu	*Pasteurella multocida*
ppu	*Pseudomonas putida*
pst	*Pseudomonas syringae pv. tomato*
rco	*Rickettsia conorii*
rpr	*Rickettsia prowazekii*
rso	*Ralstonia solanacearum*
sfl	*Shigella flexneri 301 (serotype 2a)*
sme	*Sinorhizobium meliloti*
son	*Shewanella oneidensis*
stm	*Salmonella typhimurium*
stt	*Salmonella typhi Ty2*
sty	*Salmonella typhi CT18*
vch	*Vibrio cholerae*
vpa	*Vibrio parahaemolyticus*
vvu	*Vibrio vulnificus*
xac	*Xanthomonas axonopodis*
xcc	*Xanthomonas campestris*
xfa	*Xylella fastidiosa 9a5c*
xft	*Xylella fastidiosa Temecula1*
ype	*Yersinia pestis CO92*
ypk	*Yersinia pestis KIM*

Table 8: Lists showing sets of gram-positive bacteria and Proteobacteria used in this study. Due to faulty annotation in the PRO files not all species in these lists are included in the ATPase tree. *(The bacteria *Bordetella bronchiseptica* has mistakenly been given a faulty abbreviation in some of the analysis).

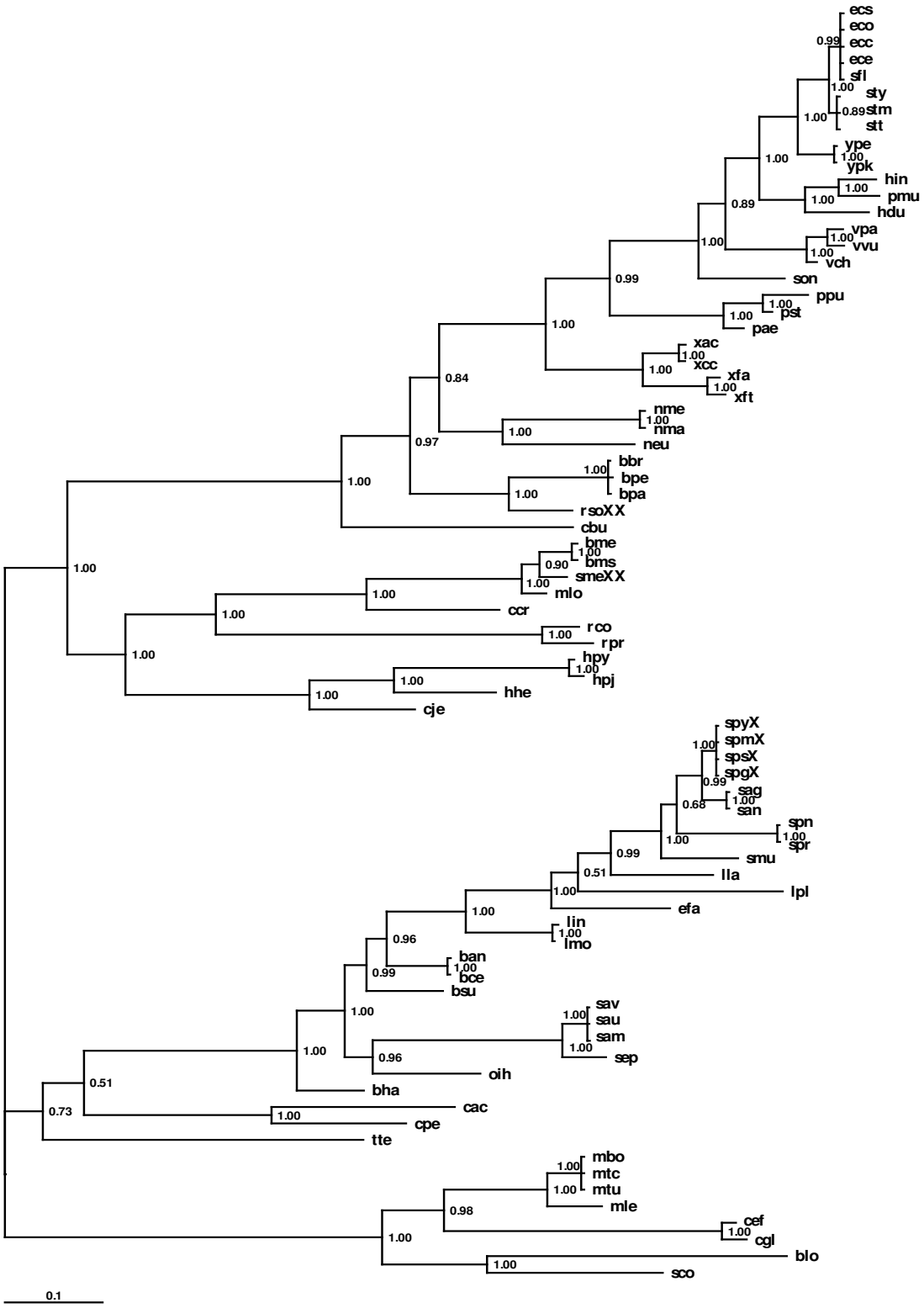


Figure 11: ATPase alpha chain tree from first execution with MrBayes. The clade reliabilities from first and second executions proves to be consistent (within +/- 0.05), with exception of the "lpl" clade thas has a difference between 0.51 and 0.99. The scale bar to the left displays number of substitutions per site. X indicates genes annotated as putative, and XX when annotated as probable.

This tree is also made using the ATP synthase alpha chain gene, calculated and constructed by TREE-PUZZLE and FITCH. Unfortunately this tree is not bootstrapped, thus no information is given regarding the probability of the clades.

Gram-positive bacteria	
ban	*Bacillus anthracis*
bce	*Bacillus cereus*
bha	*Bacillus halodurans*
blo	*Bifidobacterium longum*
bsu	*Bacillus subtilis*
cac	*Clostridium acetobutylicum*
cef	*Corynebacterium efficiens*
cgl	*Corynebacterium glutamicum*
cpe	*Clostridium perfringens*
ctc	*Clostridium tetani*
efa	*Enterococcus faecalis*
lin	*Listeria innocua*
lla	*Lactococcus lactis*
lmo	*Listeria monocytogenes*
lpl	*Lactobacillus plantarum*
mbo	*Mycobacterium bovis*
mle	*Mycobacterium leprae*
mtc	*Mycobacterium tuberculosis CDC1551*
mtu	*Mycobacterium tuberculosis H37Rv (lab strain)*
oih	*Oceanobacillus iheyensis*
sag	*Streptococcus agalactiae 2603*
sam	*Staphylococcus aureus MW2*
san	*Streptococcus agalactiae NEM316*
sau	*Staphylococcus aureus N315 (MRSA)*
sav	*Staphylococcus aureus Mu50 (VRSA)*
sco	*Streptomyces coelicolor*
sep	*Staphylococcus epidermidis*
sma	*Streptomyces avermitilis*
smu	*Streptococcus mutans*
spg	*Streptococcus pyogenes MGAS315 (serotype M3)*
spm	*Streptococcus pyogenes MGAS8232 (serotype M18)*
spn	*Streptococcus pneumoniae TIGR4*
spr	*Streptococcus pneumoniae R6*
sps	*Streptococcus pyogenes SSI-1 (serotype M3)*
spy	*Streptococcus pyogenes SF370 (serotype M1)*
tte	*Thermoanaerobacter tengcongensis*

Proteobacteria	
bbr/bre	*Bordetella bronchiseptica*
bme	*Brucella melitensis*
bms	*Brucella suis*
bpa	*Bordetella parapertussis*
bpe	*Bordetella pertussis*
cbu	*Coxiella burnetii*
ccr	*Caulobacter crescentus*
cje	*Campylobacter jejuni*
ecc	*Escherichia coli CFT073*
ece	*Escherichia coli O157 EDL933*
eco	*Escherichia coli K-12 MG1655*
ecs	*Escherichia coli O157 Sakai*
hdu	*Haemophilus ducreyi*
hhe	*Helicobacter hepaticus*
hin	*Haemophilus influenzae*
hpj	*Helicobacter pylori J99*
hpy	*Helicobacter pylori 26695*
mlo	*Mesorhizobium loti*
neu	*Nitrosomonas europaea*
nma	*Neisseria meningitidis Z2491 (serogroup A)*
nme	*Neisseria meningitidis MC58 (serogroup B)*
pae	*Pseudomonas aeruginosa*
pmu	*Pasteurella multocida*
ppu	*Pseudomonas putida*
pst	*Pseudomonas syringae pv. tomato*
rco	*Rickettsia conorii*
rpr	*Rickettsia prowazekii*
rso	*Ralstonia solanacearum*
sfl	*Shigella flexneri 301 (serotype 2a)*
sme	*Sinorhizobium meliloti*
son	*Shewanella oneidensis*
stm	*Salmonella typhimurium*
stt	*Salmonella typhi Ty2*
sty	*Salmonella typhi CT18*
vch	*Vibrio cholerae*
vpa	*Vibrio parahaemolyticus*
vvu	*Vibrio vulnificus*
xac	*Xanthomonas axonopodis*
xcc	*Xanthomonas campestris*
xfa	*Xylella fastidiosa 9a5c*
xft	*Xylella fastidiosa Temecula1*
ype	*Yersinia pestis CO92*
ypk	*Yersinia pestis KIM*

Table 9: Lists showing sets of gram-positive bacteria and Proteobacteria used in this study. Due to faulty annotation in the PRO files not all species in these lists are included in the ATPase tree. *(The bacteria *Bordetella bronchiseptica* has mistakenly been given a faulty abbreviation in some of the analysis).

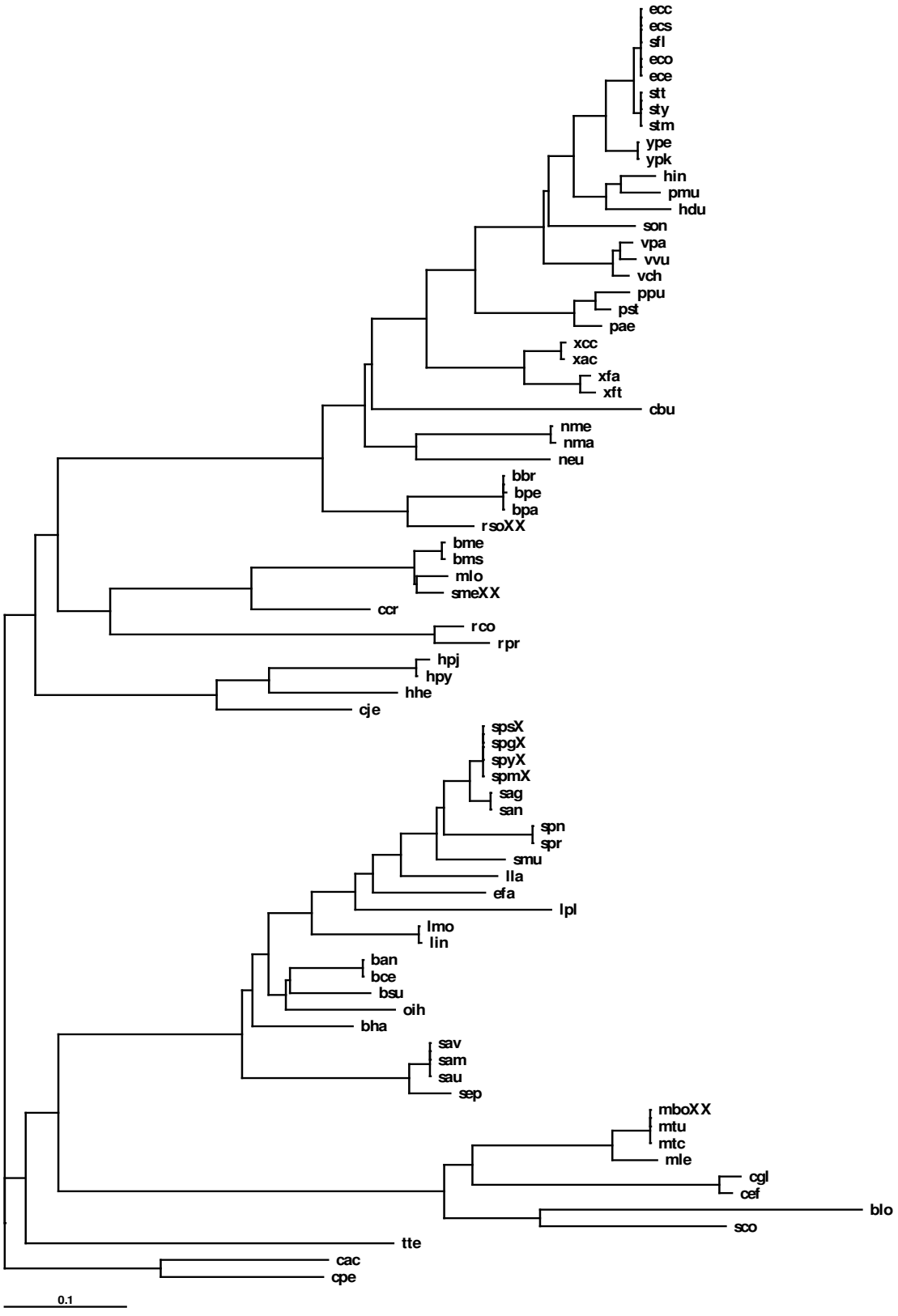


Figure 12: ATPase alpha chain tree from TREE-PUZZLE and FITCH, due to computational limitations the trees has not been bootstrapped. The scale bar to the left displays number of substitutions per site. X indicates genes annotated as putative, and XX when annotated as probable.

Due to computational limitations the different methods, employed for phylogenetic analysis, were not ideal. Although some of the generated trees have a high probability of being optimal, no guarantees can be given regarding their reliability. The two factors having greatest influence on the final trees is probably the selection/extraction of genes from the NUC and PRO files, and the numerous options available in each phylogenetic program. Every gene used in this analysis was extracted from NUC or PRO files, downloaded from the KEGG database, using Extract (see 4.2.5.1). This program extracts any FASTA sequence, from a multiple FASTA file, using gene annotation as the only searching criteria. If any of these genes have been given an incorrect annotation the final results will be affected. In addition, genes annotated as probable or putative were included in the selection. Some of these genes turned out to be false, thus affecting the results (see phylogenetic trees for the Prolyl-tRNA synthetase and the Methionyl-tRNA synthetase gene on the included DVD). These genes should have been removed and the dataset reanalyzed, but due to computational limitations this was impractical. Only the 16S rRNA and ATP synthase trees are included in the results (the remaining phylogenetic trees are available on the included DVD). These trees are consistent with those that were excluded, thus being of confirmative value. When picking a method for phylogenetic analysis there is always a balance between choosing the most suitable method and CPU hours available. Below is a discussion on the four different phylogenetic methods employed in constructing the reference trees, one method for DNA sequences and three for amino acid analysis;

The 16S rDNA consensus was calculated using a distance method with maximum likelihood measures and 10 random starting trees, which is a relatively robust method. Still the strength in this tree lies in the parameters suggested by Modeltest, ensuring the most suitable model of evolution. Modeltest was also employed in constructing the 16S rRNA tree containing all 320 genes. Due to the large number of sequences involved in this alignment a less favorable method was used in constructing a phylogenetic tree, the Neighbor-joining method. Still it is comparable to the consensus tree in **Figure 10**, which to a certain degree ensures its quality. When looking at the intermingling of 16S rRNA, as seen in **Figure 9** (page 57), it looks like the majority of the multiple 16S rRNA genes, within a single species, are placed together and probably do not affect the final result.

Trees made by MrBayes, calculated using maximum likelihood and containing clade probabilities, are probably the most certain ones. Any uncertainties should therefore lie in the selected substitution model and/or in the gamma distribution.

Having automatic parameter estimation and the ability to compute pairwise maximum likelihood distances, trees made by TREE-PUZZLE are likely to be reliable. FITCH was later employed in tree construction, taking the distance matrix from TREE-PUZZLE, using global rearrangement and randomized input order. Unfortunately, bootstrapping these results was impractical, thus there is no way to judge the clade reliabilities. Still, trees made by TREE-PUZZLE resemble trees made by other methods.

Trees made using different programs in the PHYLIP package are certainly the least certain ones. Still it is a good sign that they resemble trees made with other methods, and thus strengthen the overall results. Ideally FITCH should have been used in tree construction, but due to some unknown computational error causing problems during the bootstrapping, NEIGHBOR was used.

Therefore, regarding the reference trees, a conclusion can be drawn that the quality is sufficiently high to be used in evaluating trees made by the oligonucleotide method.

There is also a lot of on going research to make phylogenetic trees based on multiple genes, commonly referred to as supertrees (Brown *et al*., 2001; Daubin *et al*., 2002). The maximum likelihood tree generated by Daubin *et al*., from a core of 118 genes, is included in this study for comparison, see **Figure 13**. The supertree represented here contains 11 gram-positive and 12 Proteobacterial species. Some of the species included in the supertree are not included in the reference trees and vice versa, thus reducing the possibilities for comparison. Still, many of the deep branches are included in both the supertree and the reference trees, and there seems to be an almost perfect correspondence between the different phylograms.

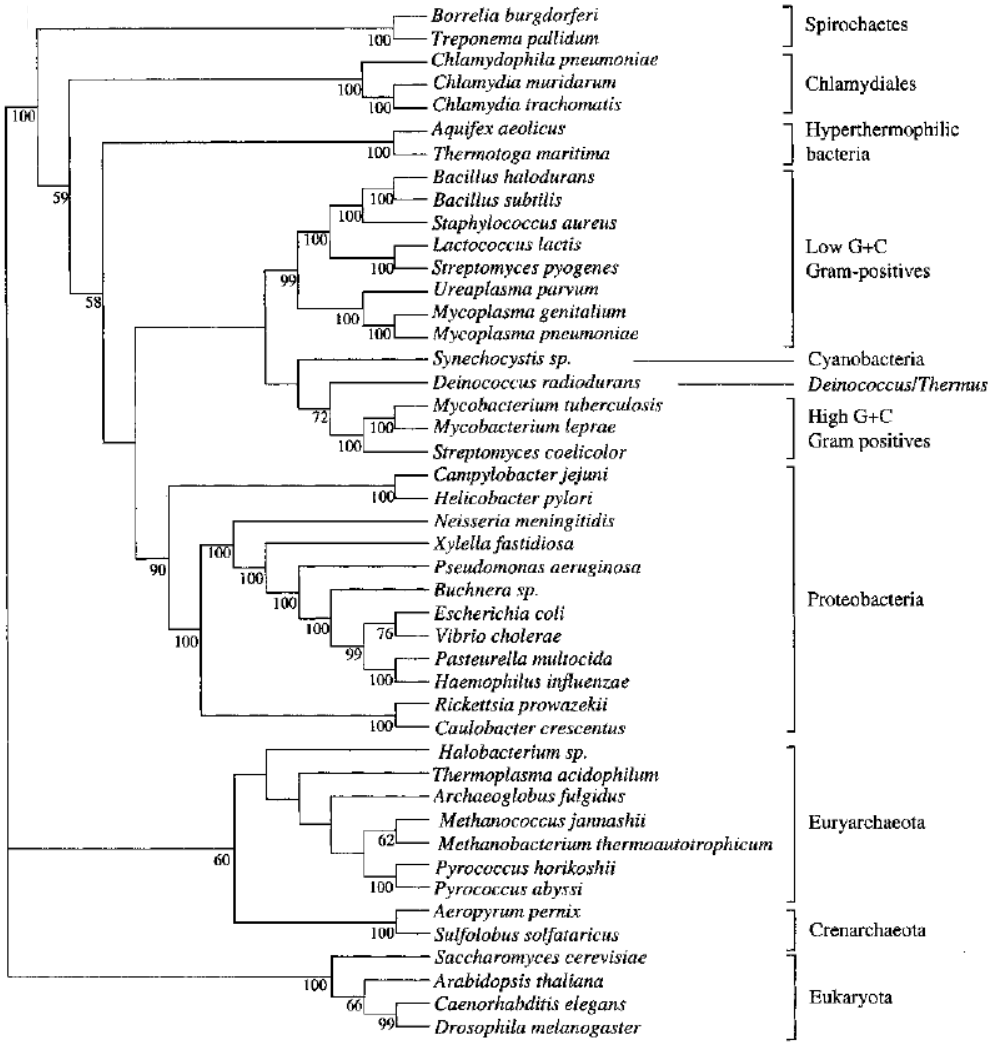


Figure 13: Maximum likelihood supertree based on 118 genes. Taken from (Daubin *et al*., 2002).

5.2 Results and discussion of the oligonucleotide classification

Before any primer set could be generated a number of analyses had to be conducted in order to assemble suitable genome sets (of appropriate size and diversity) for primer extraction. The results presented here give the foundation for selecting the first four genome sets for primer extraction. First a chart was constructed to reveal the correlation between genome set size and the number of total available primers, without using any selective parameters.

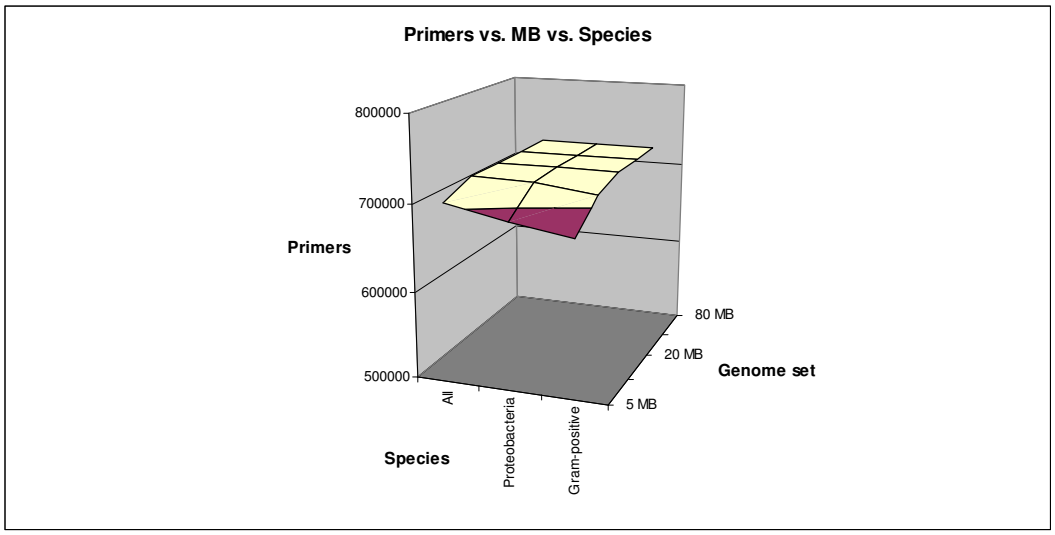


Figure 14: Chart showing number of total available primers vs. size of genome set vs. species set.

The plot shows that a relatively small number of bacterial genomes is sufficient to extract all possible 10-mer oligonucleotides. After filtration there is a maximum of 718.744 available primers. A little more than 10 MB of genome data appears to be sufficient to generate an adequate number of primers. 10 MB of genome data corresponds to three medium sized single stranded bacterial genomes. There also appears to be a correlation between the diversity of the species in the three different sets and the number of primers obtained.

The chart below show the relationship between the size of the genome set and the entropy interval needed to extract approximately 4000 primers. The diagram was constructed with the purpose of finding suitable intervals for primer extraction.

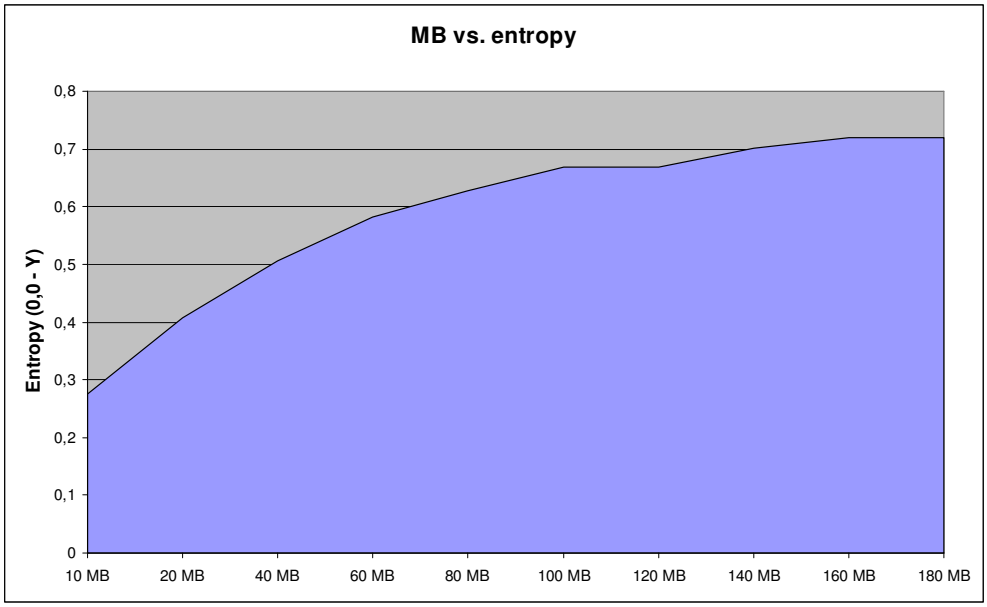


Figure 15: Chart revealing the correlation between size of entropy window and size of the genome set needed to extract approximately 4000 primers.

This chart shows that, if the genome sets are small it is sufficient with a relatively narrow entropy interval in order to extract a satisfactory number a primers. As the genome sets increase in size the entropy window has to be expanded to retain the same number of 10-mer oligonucleotides. This corresponds to the fact that it is easier to find primers with a scattered frequency distribution in a small set of genomes because there are more possibilities for variation. As the size of the genome sets increases towards 100 Mb or more, the entropy window seems to stabilize around 0.70-0.75. It is important to remember that when the genome sets becomes large the frequencies of the different primers becomes more uniformly distributed, resulting in a lower number of primers having a skewered and informative distribution. As a result of this investigation it was decided that the entropy used to generate the final primer sets should be kept within an interval of 0.3 and 0.7.

Four charts revealing the correlation between entropy, minimum frequency and the number of extracted primers were made in order to evaluate primer distribution in the different genome sets. Two sets for both the gram-positive and the Proteobacteria were made, using the 40 and 80 Mb sets. These four sets are the same sets used in the final step to extract primers. Only the sets for Proteobacteria are shown here (see **Figure 16** and **Figure 17**).

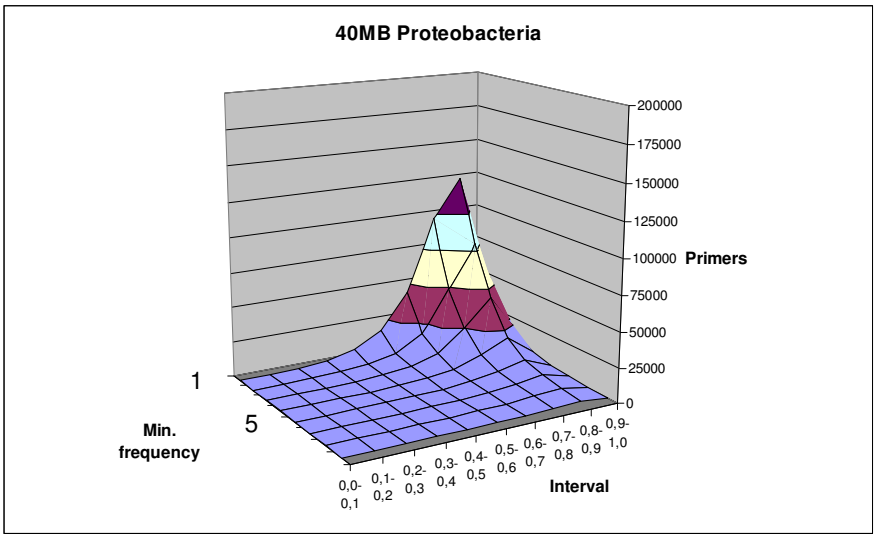


Figure 16: Chart showing minimum frequency vs. entropy interval vs. available primers, calculated by Selentprim using GC ratio 4-5 and input file generated from the "40 Mb Proteobacteria" set.

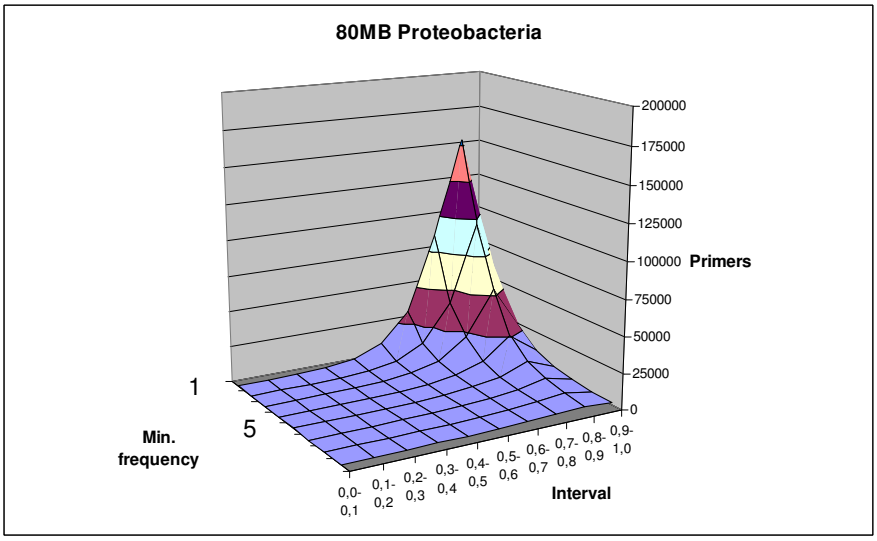


Figure 17: Chart showing minimum frequency vs. entropy interval vs. available primers, calculated by Selentprim using GC ratio 4-5 and input file generated from the "80 Mb Proteobacteria" set.

The entropy interval and minimum frequency are relatively limited since most primers are concentrated in one corner (see **Figure 16** and **Figure 17**). It is also interesting to see that there are more available primers with entropy below 0.7 among both the Proteobacterial sets than within the gram-positive sets. This could be due to the fact that the species within the Proteobacterial sets are more divergent, thus revealing greater differences concerning their primer frequencies. Since neither very high nor very low entropy will be suitable in selecting primers, it can be assumed that the entropy interval should lie somewhere between 0.3 and 0.7.

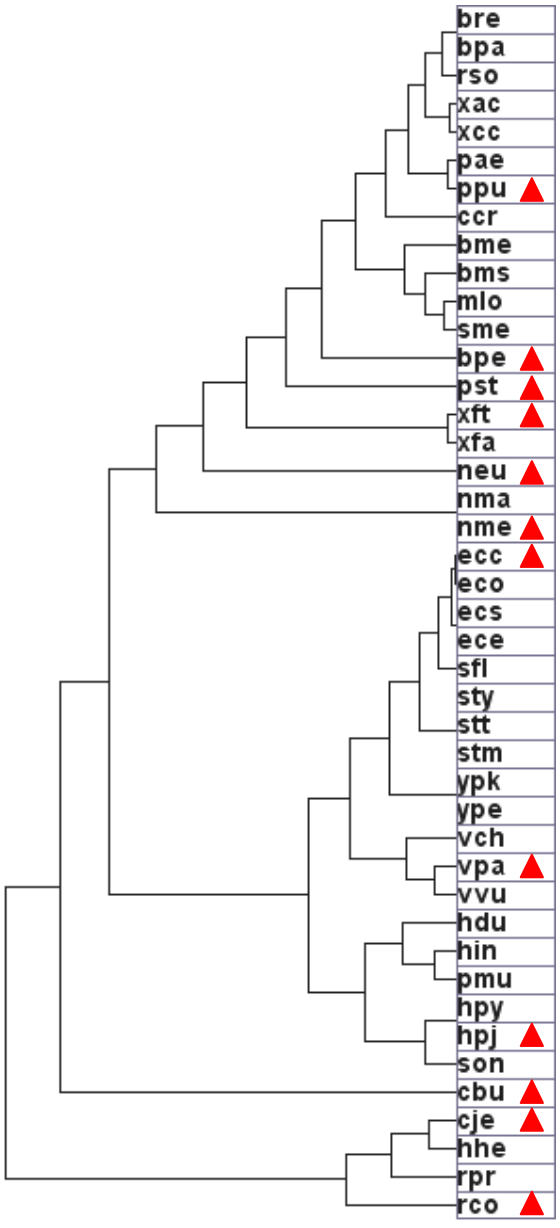
Table 10 shows the final primer sets generated from the four different genome sets and the parameters employed in order to reach approximately 4000 primers. These sets were later used as input in Testarray together with a file containing frequency data for genomes that were to be classified. The output from Testarray was edited in Excel and further analyzed by J-Express in order to generate dendrograms. These primers sets were also used in Testprimers to make gnuplots and 1:1 comparison. Results from these analysis will be presented and discussed in the next section.

Set	Discarded	-e	-E	f	G/C	Primers	File name
40Mb Proteobacteria	20	0,3	0,427	1	4-5	3999	030427-c45-f1.dat
40Mb Proteobacteria	20	0,3	0,5038	2	4-5	3999	0305038-c45-f2.dat
40Mb Proteobacteria	20	0,3	0,5488	3	4-5	3998	0305488-c45-f3.dat
40Mb Proteobacteria	20	0,3	0,6065	4	4-5	3996	0306065-c45-f4.dat
40Mb Proteobacteria	20	0,3	0,6518	5	4-5	4003	0306518-c45-f5.dat
80Mb Proteobacteria	126882	0,3	0,4764	1	4-5	4004	0304764-c45-f1.dat
80Mb Proteobacteria	126882	0,3	0,5417	2	4-5	4001	0305417-c45-f2.dat
80Mb Proteobacteria	126882	0,3	0,5785	3	4-5	4000	0305785-c45-f3.dat
80Mb Proteobacteria	126882	0,3	0,6266	4	4-5	4003	0306266-c45-f4.dat
80Mb Proteobacteria	126882	0,3	0,6673	5	4-5	4003	0306673-c45-f5.dat
40Mb Gram-positive	10	0,3	0,5596	1	4-5	3999	0305596-c45-f1.dat
40Mb Gram-positive	10	0,3	0,637	2	4-5	4004	030637-c45-f2.dat
40Mb Gram-positive	10	0,3	0,6967	3	4-5	3999	0306967-c45-f3.dat
80Mb Gram-positive	10	0,3	0,6299	1	4-5	4003	0306299-c45-f1.dat
80Mb Gram-positive	10	0,3	0,6859	2	4-5	4001	0306859-c45-f2.dat
80Mb Gram-positive	10	0,3	0,7303	3	4-5	4001	0307303-c45-f3.dat

Table 10: Table showing the different primer sets and their settings when executed in Selentprim. "-e" is the minimum entropy, while "–E" refers to the maximum entropy. "f" is minimum frequency, "G/C" the number of C-bases and "Primers" refers to the total number of extracted primers in each set. Finally, in the last column, the file name of the primer set. "Discarded" refers to the number of discarded primers, by the program Gencnt, due to sequencing errors.

5.2.1 Discussion and results on comparison of distantly related species

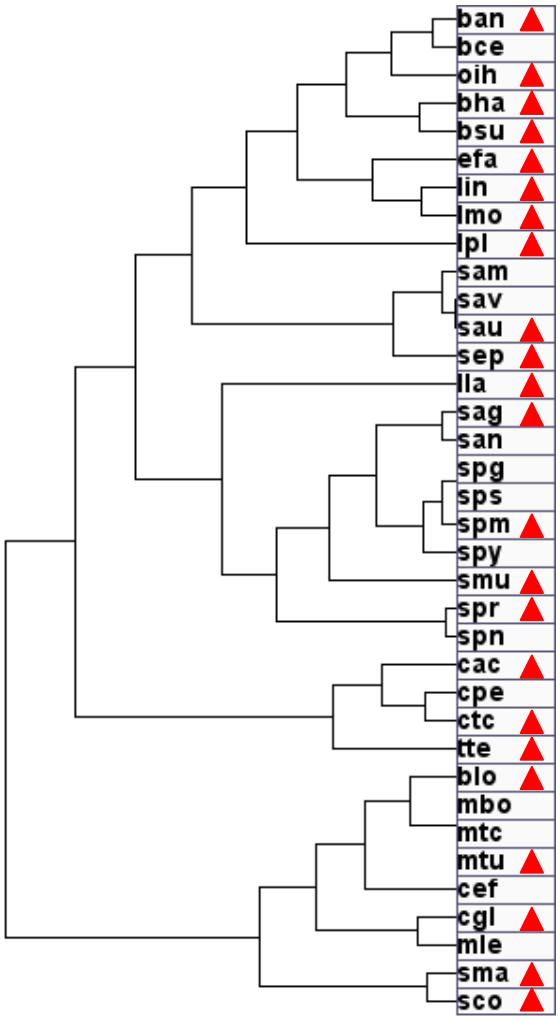
The output from Testarray, made by combining primer sets and files with 10-mer oligonucleotide frequencies for the species to be classified, were analyzed in J-Express. Different primer sets (see **Table 10**) were subject to different distance measures and clustering algorithms in order to generate a final dendrogram (shown below) that best resembles the phylogenetic reference trees.



bbr/bre	*Bordetella bronchiseptica*
bme	*Brucella melitensis*
bms	*Brucella suis*
bpa	*Bordetella parapertussis*
bpe	*Bordetella pertussis*
cbu	*Coxiella burnetii*
ccr	*Caulobacter crescentus*
cje	*Campylobacter jejuni*
ecc	*Escherichia coli CFT073*
ece	*Escherichia coli O157 EDL933*
eco	*Escherichia coli K-12 MG1655*
ecs	*Escherichia coli O157 Sakai*
hdu	*Haemophilus ducreyi*
hhe	*Helicobacter hepaticus*
hin	*Haemophilus influenzae*
hpj	*Helicobacter pylori J99*
hpy	*Helicobacter pylori 26695*
mlo	*Mesorhizobium loti*
neu	*Nitrosomonas europaea*
nma	*Neisseria meningitidis Z2491*
nme	*Neisseria meningitidis MC58*
pae	*Pseudomonas aeruginosa*
pmu	*Pasteurella multocida*
ppu	*Pseudomonas putida*
pst	*Pseudomonas syringae pv. tomato*
rco	*Rickettsia conorii*
rpr	*Rickettsia prowazekii*
rso	*Ralstonia solanacearum*
sfl	*Shigella flexneri 301*
sme	*Sinorhizobium meliloti*
son	*Shewanella oneidensis*
stm	*Salmonella typhimurium*
stt	*Salmonella typhi Ty2*
sty	*Salmonella typhi CT18*
vch	*Vibrio cholerae*
vpa	*Vibrio parahaemolyticus*
vvu	*Vibrio vulnificus*
xac	*Xanthomonas axonopodis*
xcc	*Xanthomonas campestris*
xfa	*Xylella fastidiosa 9a5c*
xft	*Xylella fastidiosa Temecula1*
ype	*Yersinia pestis CO92*
ypk	*Yersinia pestis KIM*

Figure 18: Proteobacteria tree clustered with J-Express using WPGMA and Pearson Correlation. Primer set generated from the "Proteobacteria 40 Mb" set, using the following settings in Selentprim; e0.3 E0.5038 c4 C5 f2. Species marked with a red triangle are included in the primer selection set.

A set of dendrograms were constructed for the gram-positive species, analogous to the Proteobacterial tree. The gram-positive dendrogram that best resembles the phylogenetic reference trees is shown below.



ban	*Bacillus anthracis*
bce	*Bacillus cereus*
bha	*Bacillus halodurans*
blo	*Bifidobacterium longum*
bsu	*Bacillus subtilis*
cac	*Clostridium acetobutylicum*
cef	*Corynebacterium efficiens*
cgl	*Corynebacterium glutamicum*
cpe	*Clostridium perfringens*
ctc	*Clostridium tetani*
efa	*Enterococcus faecalis*
lin	*Listeria innocua*
lla	*Lactococcus lactis*
lmo	*Listeria monocytogenes*
lpl	*Lactobacillus plantarum*
mbo	*Mycobacterium bovis*
mle	*Mycobacterium leprae*
mtc	*Mycobacterium tuberculosis CDC1551*
mtu	*Mycobacterium tuberculosis H37Rv*
oih	*Oceanobacillus iheyensis*
sag	*Streptococcus agalactiae 2603*
sam	*Staphylococcus aureus MW2*
san	*Streptococcus agalactiae NEM316*
sau	*Staphylococcus aureus N315 (MRSA)*
sav	*Staphylococcus aureus Mu50 (VRSA)*
sco	*Streptomyces coelicolor*
sep	*Staphylococcus epidermidis*
sma	*Streptomyces avermitilis*
smu	*Streptococcus mutans*
spg	*Streptococcus pyogenes MGAS315*
spm	*Streptococcus pyogenes MGAS8232*
spn	*Streptococcus pneumoniae TIGR4*
spr	*Streptococcus pneumoniae R6*
sps	*Streptococcus pyogenes SSI-1*
spy	*Streptococcus pyogenes SF370*
tte	*Thermoanaerobacter tengcongensis*

Figure 19: Gram-positive bacteria tree clustered with J-Express using WPGMA and the Canberra algorithm. Primer set generated from the "Gram-positive 80 Mb" set, using the following settings in Selentprim; e0.3 E0.7303 c4 C5 f3. Species marked with a red triangle are included in the primer selection set.

Only two of the 16 generated dendrograms are shown in the section above, one for the Proteobacterial species and one for the gram-positive species (see **Figure 18** and **Figure 19**), the remaining trees are included on the DVD. Since none if these 16 trees were identical and the differences turned out to be inconsistent, a selection was made to best represent the final dendrograms. There are virtually an infinite number of combinations in selecting 4000 primers and it is not possible to evaluate every possibility. The size of the genome sets, as well as their composition, most certainly has some effect on the final result. Still, it is difficult to point out any rules regarding how the parameters in Selentprim should be employed in order to extract a high-quality primer set. Neither can any conclusion be drawn regarding the size of the extraction set (40 or 80 MB).
Comparing trees made in J-Express with the reference trees reveals some differences regarding the Proteobacterial trees. Inconsistency is found when looking at the four species *Xanthomonas axonopodis* (xac), *Xanthomonas campestris* (xcc), *Xylella fastidiosa 9a5c* (xfa) and *Xylella fastidiosa Temecula1* (xft) which always appear on a common branch, with high bootstrapping values, in the reference trees. Comparing these results to those obtained in **Figure 18**, where the *Xanthomonas axonopodis/ Xanthomonas campestris* pair is placed far away from the *Xylella fastidiosa 9a5c/ Xylella fastidiosa Temecula1* pair, raises some questions about the oligonucleotide method and its ability to classify bacteria. This phenomenon is also observed with some of the other species that normally appear on the same branch.
One explanation to this abnormal classification would be to blame the clustering method itself, due to a general limitation in the hierarchical clustering algorithm. If a bad assumption is made early in the process it can not be corrected, thus affecting the final result (Quackenbush, 2001). When clustering objects the algorithm seeks to find the two species that are most closely related, placing them in a common cluster and repeat this procedure until all objects are clustered. During this procedure the four species mentioned above might have been placed in different cluster even thought they are related, and drawn further and further apart in the subsequent clustering process (see **Figure 20**). If the first cluster to be made had been different, resulting in a different starting point, the rest of the clustering would probably have been a little different and these four organisms possibly would have been clustered together.

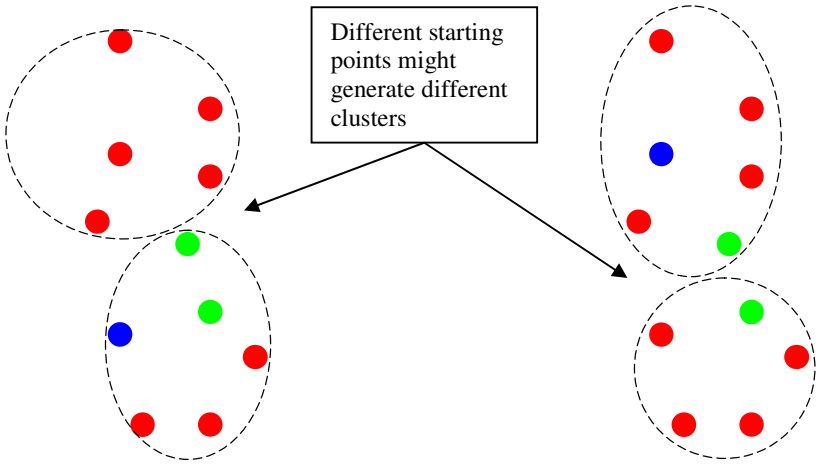


Figure 20: Illustration demonstrating how different starting points (blue dots) may result in different clustering of the objects involved (red dots) causing related objects (green dots) to be placed in separate clusters.

As an alternative to hierarchical clustering κ-means clustering and SOM were conducted. Although they both require knowledge regarding the number of clusters that best represents the available data. Unfortunately these methods will not generate any phylogenetic tree, only groups or networks. Still, both κ-means clustering and SOM place the *Xanthomonas axonopodis* (xac)/*Xanthomonas campestris* (xcc) pair and the *Xylella fastidiosa 9a5c* (xfa)/*Xylella fastidiosa Temecula1* (xft) pair in different groups and the grouping strongly resembles those generated by hierarchical clustering.

As pointed out in the introduction, it might be difficult to compare and cluster complex profiles if the differences are too large. In **Figure 21** the three most similar species were obtained by using Euclidian distance measures when *Bacillus halodurans C-125* (bha) was used as a starting point, *Bacillus subtilis 168* (bsu) was the second most similar species while *Lactobacillus plantarum* (lpl) was the third. When using *Bacillus subtilis 168* (bsu) as starting point the most similar species should either be *Bacillus halodurans C-125* (bha) or *Lactobacillus plantarum* (lpl), but this is not the case, see **Figure 22**. Instead *Listeria monocytogenes* (lmo) and *Bacillus cereus* (bce) were calculated to give the most similar profiles.

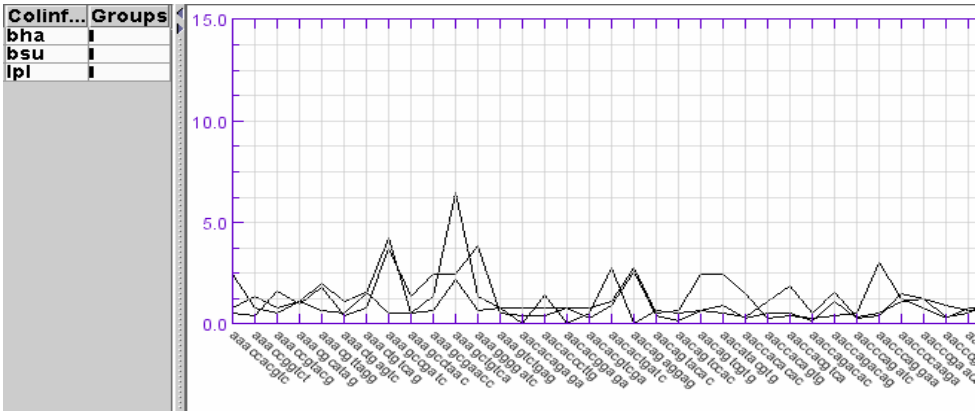


Figure 21: Figure showing the three most similar profiles when *Bacillus halodurans C-125* (bah) is used as starting point with Euclidian distance measures. The primer set was generated from the "Gram-positive 40 Mb" set, using the following settings in Selentprim; e0.3 E0.6967 c4 C5 f3 (Screen shot from J-Express).

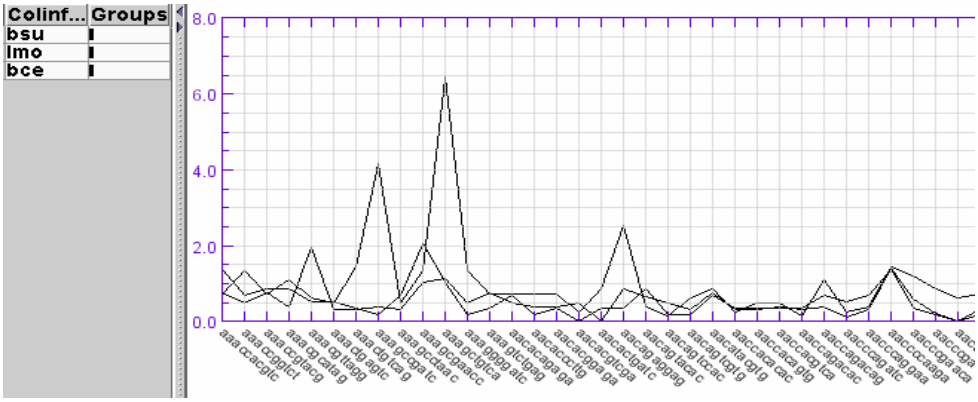


Figure 22: Figure showing the three most similar profiles when *Bacillus subtilis 168* (bsu) is used as starting point with Euclidian distance measures. The primer set was generated from the "Gram-positive 40 Mb" set, using the following settings in Selentprim; e0.3 E0.6967 c4 C5 f3 (Screen shot from J-Express).

One explanation to this phenomenon might be that the patterns are too complex to make a reasonable comparison, thus providing more than one possible solution. This problem also takes place when using different distance measures or correlations, even though the species selected to be the most similar may vary. Another explanation could be that high peaks in the primer frequencies, most probably as a result of repeated sequences in some genomes, strongly affects the algorithms for distance measures, thus having an effect on the final clustering (see **Figure 23**).

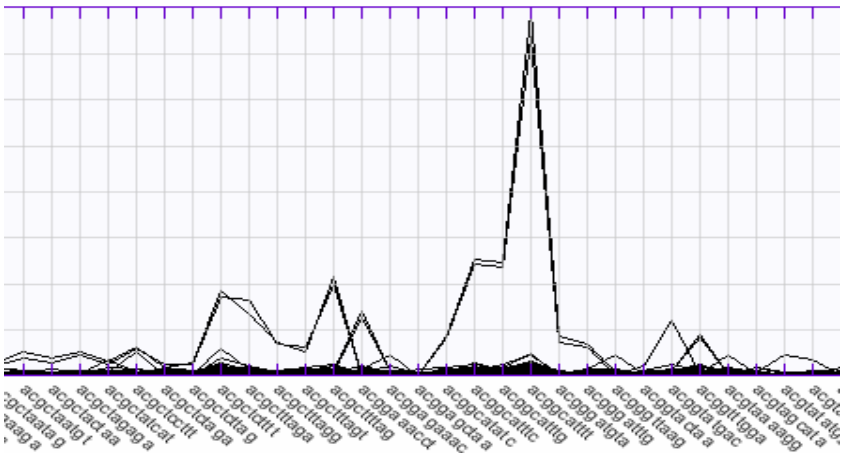


Figure 23: Plot showing primer frequencies for different Proteobacterial species, made using the output from Testarray. The high peak seen in the figure counts almost 100 primers with the sequence ACGGCATTTT. Several peaks like this one are distributed throughout the file, probably having a significant effect when calculating distances between species.

The MUMmer plot below clearly reveals a tighter relationship between *Bacillus subtilis 168* and *Bacillus halodurans C-125*, than between *Bacillus subtilis 168* and *Listeria monocytogenes EGD-e.*

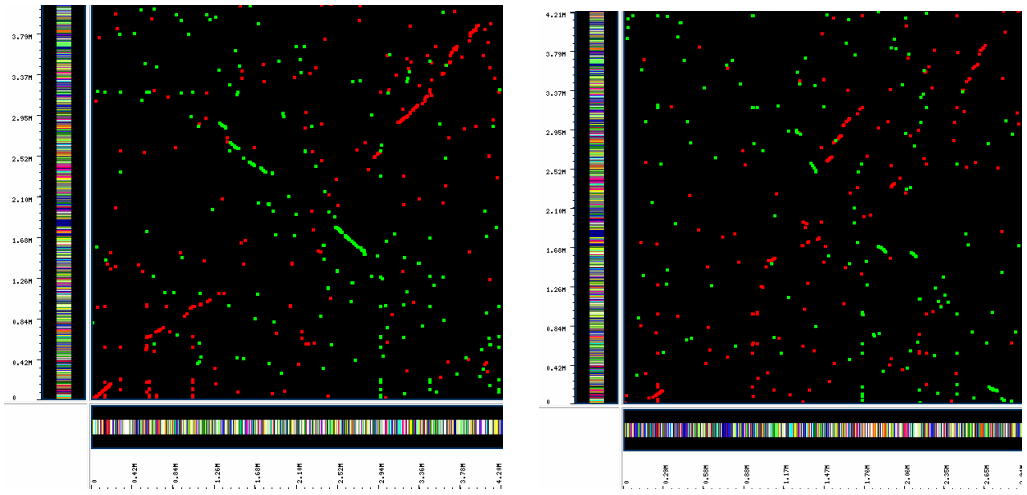


Figure 24: The MUMmer plot to the left shows *Bacillus subtilis 168* (Y-axis) vs. *Bacillus halodurans C-125* (X-axis). The plot to the right shows *Bacillus subtilis 168* (Y-axis) vs. *Listeria monocytogenes EGD-e* (X-axis). Both plots use 20 bp as minimum alignment length.

The rest of the dendrogram for the Proteobacterial species appears to be relatively consistent for the remaining species, compared to the phylogenetic reference trees. Taxa having high bootstrapping values in the reference trees also seem to be the most consistent in the dendrogram. Still there are exceptions, as mentioned above (the

Xanthomonas and *Xylella* species). No significant abnormalities can be found when evaluating the most favorable dendrogram for the gram-positive bacteria. The inconsistency between the gram-positive dendrogram and the phylogenetic reference trees appears to be no larger than the internal variations between the different phylogenetic trees. The gram-positive bacteria used in this thesis are less diverse than the set containing the Proteobacterial species, and this could explain the better clustering results observed for the former group. Despite some irregular clustering in the Proteobacterial dendrogram, this method appears to be suitable in classifying distantly related organisms. Nevertheless the results might have been even better if some of the problems mentioned above, concerning high peaks and clustering algorithms, were treated in a reasonable manner. This could be done by a broader evaluation of different clustering algorithms and distance measures, in addition to developing a method to reduce the influence of high peaks on the final dendrogram.

5.2.2 Discussion on comparison of closely related species and strains

In this part of the study the aim is to test how well the oligonucleotide method distinguishes between strains from the same species and species that are closely related, such as *Escherichia coli* and *Shigella flexneri*. Hopefully the generated profiles are similar enough not to be mixed with other species and still having a sufficient number of differences so that they can be resolved. The figures below shows four different array plots, made with J-Express, making it possible to create graphs of each profile in relation to another.

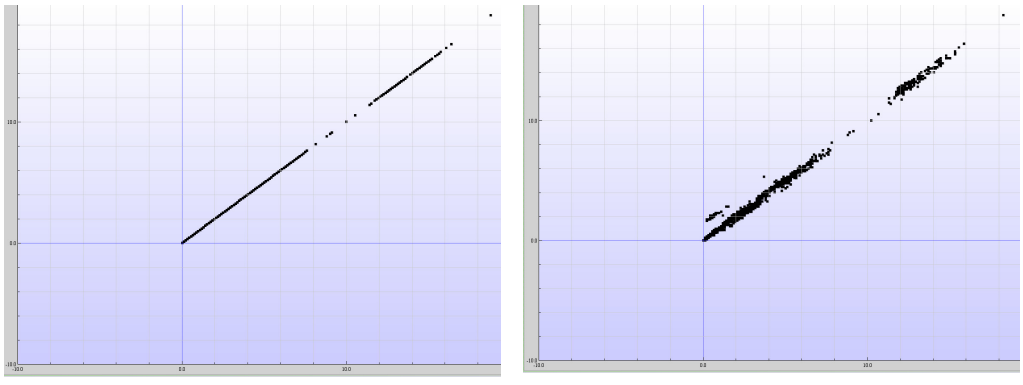


Figure 25: The plot to the left shows *Shigella flexneri* 301 (serotype 2a) against it self, while the plot to the right shows *Shigella flexneri* 301 (serotype 2a) vs. *Shigella flexneri* 2457T (serotype 2a). (Screen shot from J-Express). Using the medium EcoSalmoFlex set with the following settings, e0.0 E0.9 c4 C5 f2

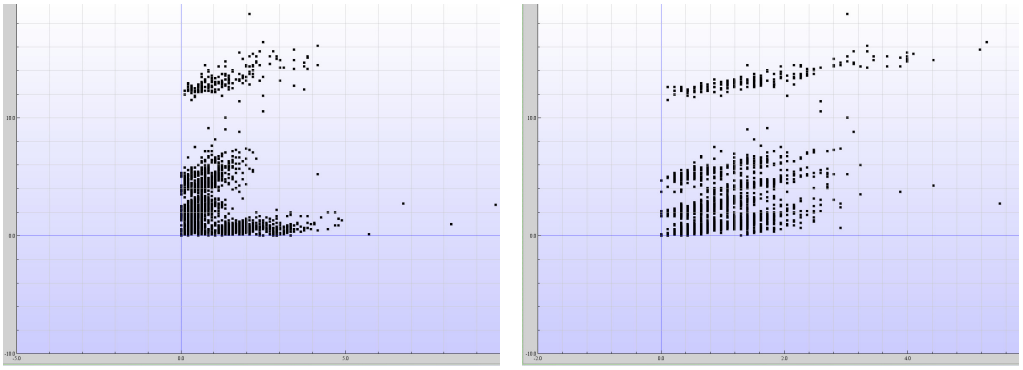


Figure 26: The plot to the left shows *Shigella flexneri* 301 (serotype 2a) vs. *Salmonella typhi* CT18, while the plot to the right shows *Shigella flexneri* 301 (serotype 2a) vs. *Escherichia coli* K-12 MG1655. (Screen shot from J-Express).

Looking at these profiles it is interesting to see that even between these closely related species the differences are significant, but far from random. These analyses were conducted on an array generated from all nine species in the EcoSalmoFlex set. Since some of these 9 species shows nearly 100% homology by DNA::DNA hybridization (70-100% between *E.coli* and *S.flexneri* and 50% between *E.coli* and *S.typhi* (Madigan *et al.*, 2003)) it is difficult to find primers having a skewered distribution, which gives a set containing a little more than 2400 primers.

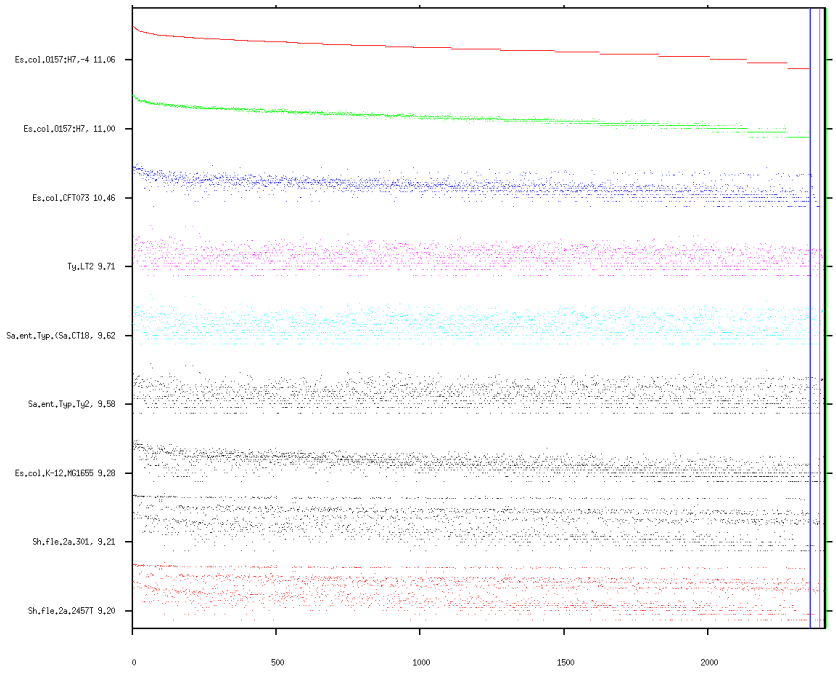


Figure 27: Gnuplot of the EcoSalmoFlex, showing primer distribution in the different genomes. In the largest genome primers are sorted according to their frequency.
The primer set was generated using e 0.0 E 0.9 c4 C5 f2.

The Gnuplot above, constructed using the output from Testprimers, reveals the distribution of primers across the different genomes. This kind of plot has two advantages; first of all it allows us to see if the primers are evenly distributed between and within each genome, which is true in this case. Secondly; it allows us, at least to a certain degree, to compare patterns from different species or strains by visual inspection. When running this set in Testarray, and analyzing the results in J-Express, the dendrogram shown in **Figure 28** was generated. In addition, a second primer set was generated, using only two species, and executed in Testarray and J-Express, see **Figure 29**.

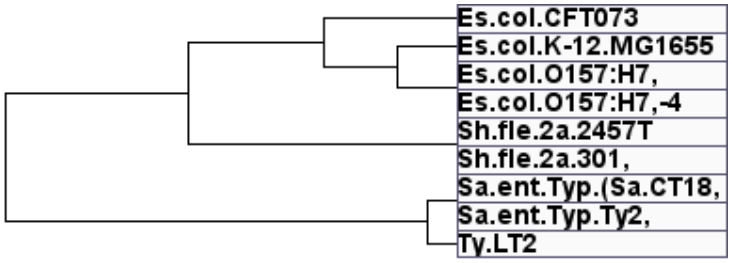


Figure 28: Dendrogram showing the classification of nine closely related enteric bacteria. The primer set was generated using e 0.0 E 0.9 c4 c5 f2, giving 2411 primers. The data were clustered using UPGMA and **Pearson correlation**. All species were included in the primer selection set.

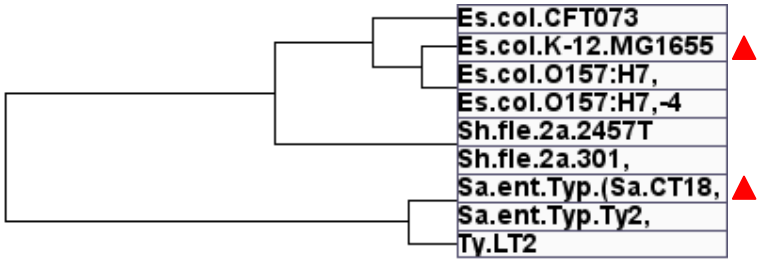


Figure 29: Dendrogram showing the classification of nine closely related enteric bacteria. The primer set was generated using e 0.15 E 0.581 c4 c5 f1, giving 4636 primers. The data were clustered using UPGMA and **Pearson correlation**. Species marked with a red triangle are included in the primer selection set.

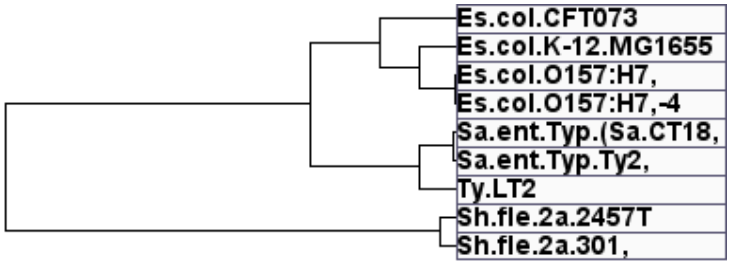


Figure 30: Dendrogram showing the classification of nine closely related enteric bacteria. The primer set was generated using e 0.0 E 0.9 c4 c5 f2. The data were clustered using UPGMA with **Euclidian** distance measures. All species were included in the primer selection set.

The first two trees are identical, only having slightly different branch length, and correspond perfectly to the phylogenetic reference trees. Remembering that these trees are made by different primer sets generated from different settings and species, this is a positive result. The third tree is different, placing *Escherichia coli* and *Salmonella* strains in the same cluster, which is not in correspondence with the reference trees. This faulty clustering is caused by shortcomings in the Euclidian distance measures. While Euclidian distance is a measurement of the distance between two profiles, Pearson correlation is a similarity measure and probably more suitable for our analysis (for further details see 3.2.7.1).

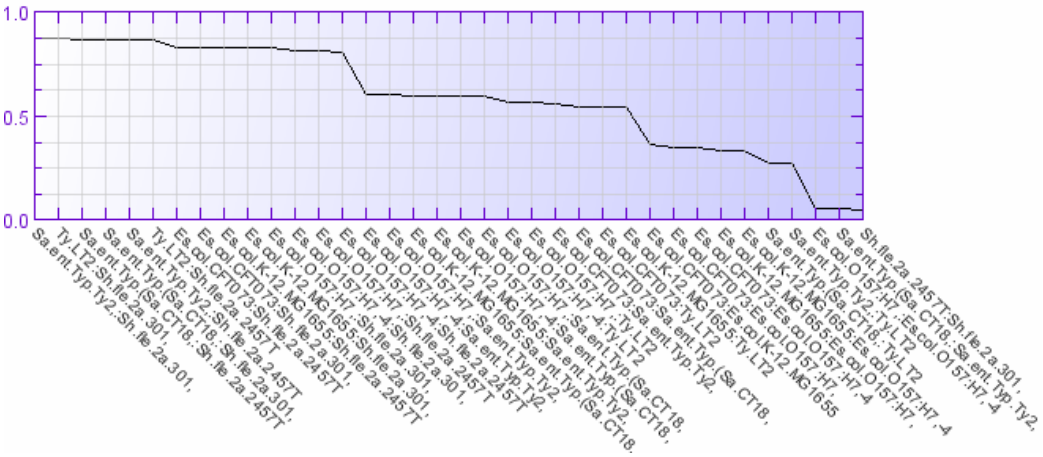


Figure 31: Chart showing pairwise Euclidian distances between genomes in the EcoSalmoFlex set, made by sorting and visualizing the output from Testprimers in J-Express. The comparison was made using the same primer set used in creating the dendrograms in **Figure 28** and **Figure 30**, e 0.0 E 0.9 c4 c5 f2.

Problems occurring from using Euclidian distance measures can also be seen in the chart revealing pairwise comparison of species in the EcoSalmoFlex set. Here the program Testprimers computes shorter distances between *Escherichia coli* and *Salmonella* than between *Escherichia coli* and *Shigella flexneri,* finally leading to faulty grouping (see **Figure 31).**

The array plot constructed in **Figure 25** shows *Shigella flexneri* 301 vs. *Shigella flexneri* 2457T, both serotype 2a. Two genomes from the same serotype, but geographically and temporally separated. The 301 strain (Jin *et al*., 2002) is 7.85 kb larger than the 2457T strain, which is largely accounted for by differences in IS complement (Wei *et al*., 2003). There are more than 1400 single-nucleotide differences between them, but this is a small number compared to their total genome size. Even though the output from

Testarray contains detectable variations they are very limited and probably only distinguishable in an *in silico* experiment. If these two serotypes, or any other two species being equally related, are to be distinguished, the primer set probably has to be more specialized. By designing a set for a certain group of species or strains, and by using a more sophisticated algorithm for primer selection, a sufficient resolution should be achievable. It is also interesting to note that both *Shigella flexneri* strains are placed in a separate group next to *the Escherichia coli* group, while these two species tends to mix in the reference trees. The inconsistent placements of these species in the reference trees are probably due to a very limited number of differences in the specific genes, making phylogenetic classification difficult. Looking at the profiles in the array plot above confirms the degree of dissimilarity between these species, thus explaining their placement in separate clusters. Comparing *Shigella flexneri* to four other strains of *Escherichia coli* reveals a remarkable number of differences detected by the array, see **Figure 26**. Looking at the dendrograms in **Figure 28** and **Figure 29**, and bearing in mind that *Shigella* strains are probably clones of *Escherichia coli* (Jin *et al*., 2002; Lan and Reeves, 2002), having nearly 3.000 ORFs in common (Wei *et al.*, 2003), see **Figure 32**) these results clearly reflects the potential of the oligonucleotide microarray classification method.

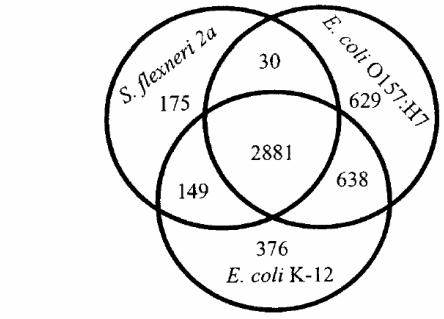


Figure 32: Venn diagram showing the distribution of common and unique ORFs among *S. flexneri 2a*, *E. coli K-12*, and *E. coli O157: H7.* Only complete protein-coding ORFs, including hypothetical unknowns, are included. IS element and phage ORFs, as well as pseudogenes, are excluded. Figure taken from (Wei *et al*., 2003).

The Venn diagram above indicates a closer relationship between *Escherichia coli K-12* and *Shigella Flexneri,* than between *Shigella flexneri* and *Escherichia coliO157:H7*. The diagram has been made by comparing the complete genome sequences of these three genomes. The same conclusion, regarding the relationship between these three organisms, has been reached in other studies involving complete genome comparison

(Jin *et al*., 2002; Lan and Reeves, 2002; Wei *et al*., 2003). All these results are in perfect correspondence to our results, generated by the oligonucleotide method, as can be seen in **Figure 33**, where *Escherichia coli K-12* is found to be the most similar bacteria compared to *Shigella flexneri.* The same results were reached using other primer sets (those used in Figure 28 and **Figure 29**) and distance measures (Euclidian and Manhatten).

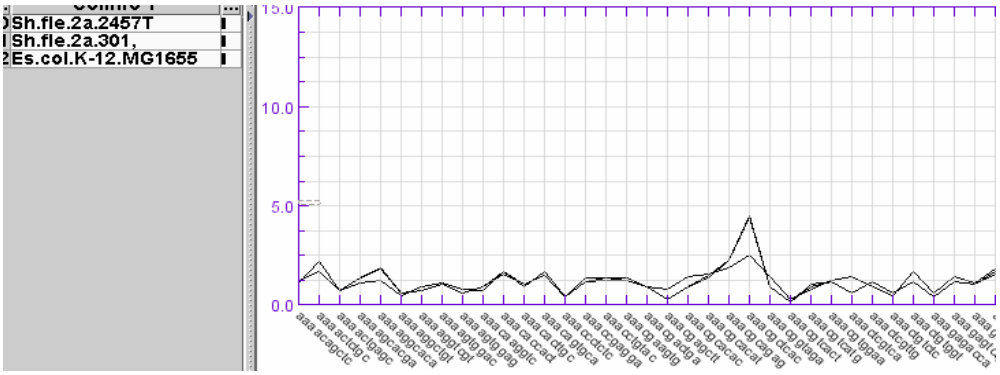


Figure 33: This chart shows the three most similar species when *Shigella flexneri 2a 2457T* is used as starting point using Pearson correlation. The primer set was generated using e 0.15 E 0.581 c4 c5 f1, and only two species were used in the primer selection.

Two MUMmer plots were made by aligning two complete genome sequences, using a 100 bp alignment frame (see **Figure 34**). Looking a these figures there seems to be a little more genome rearrangements between *Shigella flexneri 2a 2457T* and *Escherichia coli O157:H7 EDL933* than between *Shigella flexneri 2a 2457T* and *Escherichia coli K12-MG1655.* Even though no large differences can be seen between the two plots, they confirm the results reached by other methods and the oligonucleotide method.

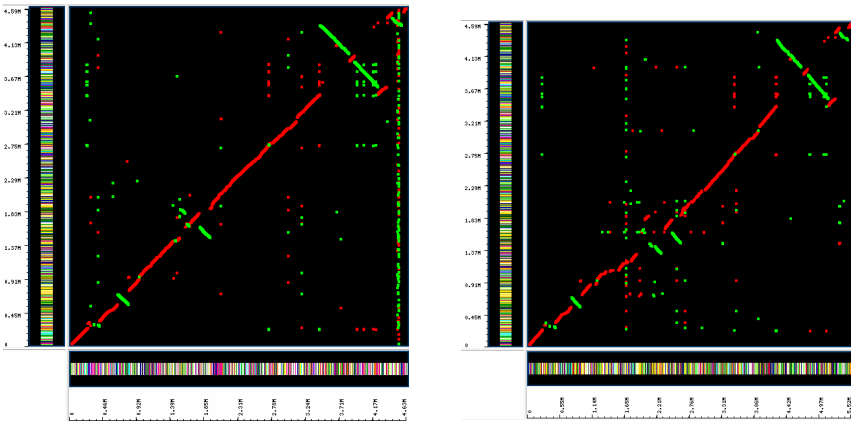


Figure 34: The MUMmer plot to the left shows *Shigella flexneri 2a 2457T* (Y-axis) vs. *Escherichia coli K12-MG1655* (X-axis). The plot to the right shows *Shigella flexneri 2a 2457T* (Y-axis) vs. *Escherichia coli O157:H7 EDL933* (X-axis). Both plots use 100 bp as minimum alignment length.

6 Conclusion

The object of this study has been to evaluate an *in silico* method for bacterial classification, using a set of 4000 oligonucleotides, selected according to their entropy. To evaluate the outcome of this method, visualized as dendrograms, a wide range of phylogenetic trees were constructed using well known techniques, involving 4 different genes and 4 different algorithms, giving us a total of 11 phylogenetic trees for comparison. Although the ultimate goal of this technique is to construct a microarray for classification purposes, this study has only been concentrated on testing the possibilities of this method *in silico*.

Oligonucleotides as a tool for bacterial classification tends to meet some problems at the genus level, but has proven to give a remarkably high resolution at the species and strain level. In fact the same classification results were obtained by using our method as with by whole genome comparison. Most certainly this method can also be applied to distinguish other closely related and pathogenic species such as strains of *Bacillus anthracis* or *Staphylococcus aureus*. Probably it should be possible to obtain even higher resolution by tuning the primer selection method and by designing custom made oligonucleotide arrays for a certain group of species or strains. The method should also be improved in order to handle peaks in the array data, either by filtering or by using a distance measures that is unaffected by obstacles.

It is important to remember that this study has been conducted solely *in silico*. In a real life experiment the data scanned from the microarray are inaccurate and contain noise, thus leading to a lower resolution. Since the minimum sequence length used in oligonucleotide microarrays is approximately 25 bases, the single base extension technique will probably be employed (Nikiforov *et al*., 1994) in conducting the experiment. The extracted 10-mer oligonucleotides will be used as primers in a single base extension reaction, providing a fluorescent or radioactively signal for detection. In order to immobilize the primers and to facilitate the enzymatical reaction the array can be covered with a polyacrylamide gel (Strizhkov *et al*., 2000; Vasiliskov *et al*., 1999).

7 Bibliography

Amaratunga, D. and Cabrera, J. (2003) *Exploration and Analysis of DNA Microarray and Protein Array Data*. Wiley-Interscience.

Gibson, G. and Muse, S.V. (2002) *A Primer of Genome Science*. Sinauer Associates, Inc., Sunderland, Massachusetts.

Hall, B.G. (2001) *Phylogenetic Tress Made Easy*. Sinauer.

Lengeler, W., Drews, G. and Schlegel, H. (1999) *Biology of the Prokaryotes*. Blackwell Science.

Madigan, M.T., Martinko, J.M. and Parker, J. (2003) *Brock Biology of Microorganisms*. Pearson Education.

Mount, D.W. (2001) *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press, New York.

8 References

Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson, J., Jr., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Botstein, D., Brown, P.O. and Staudt, L.M. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503-511.

Brazma, A. and Vilo, J. (2000) Gene expression data analysis. *FEBS Lett*, **480**, 17-24.

Brochier, C., Philippe, H. and Moreira, D. (2000) The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome. *Trends Genet*, **16**, 529-533.

Brown, J.R. and Doolittle, W.F. (1997) Archaea and the prokaryote-to-eukaryote transition. *Microbiol Mol Biol Rev*, **61**, 456-502.

Brown, J.R., Douady, C.J., Italia, M.J., Marshall, W.E. and Stanhope, M.J. (2001) Universal trees based on large combined protein sequence data sets. *Nat Genet*, **28**, 281-285.

Carver, T. (2000) Cons. EMBOSS, Cambrigde.

Castresana, J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evelution*, **17**, 540-552.

Daubin, V., Gouy, M. and Perriere, G. (2001) Bacterial molecular phylogeny using supertree approach. *Genome Inform Ser Workshop Genome Inform*, **12**, 155-164.

Daubin, V., Gouy, M. and Perriere, G. (2002) A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Res*, **12**, 1080-1090.

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

Delcher, A.L., Phillippy, A., Carlton, J. and Salzberg, S.L. (2002) Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res*, **30**, 2478-2483.

Doolittle, W.F. (1999) Phylogenetic classification and the universal tree. *Science*, **284**, 2124-2129.

Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P. and Trent, J.M. (1999) Expression profiling using cDNA microarrays. *Nat Genet*, **21**, 10-14.

Dysvik, B. and Jonassen, I. (2001) J-Express: exploring gene expression data using Java. *Bioinformatics*, **17**, 369-370.

Eisen, J.A. (2000) Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. *Curr Opin Genet Dev*, **10**, 606-611.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, **95**, 14863-14868.
Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*, **17**, 368-376.
Felsenstein, J. (1993) Phylogeny Inference Package.
Fitch, W.M. (1970) Distinguishing homologous from analogous proteins. *Syst Zool*, **19**, 99-113.
Fitch, W.M. and Margolia.E. (1987) Construction of Phylogenetic Trees. *Science*, **155**, 279-&.
Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M. and et al. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, **269**, 496-512.
Gilbert, D. (1999) Readseq. Bloomington, Indiana.
Gogarten, J.P., Starke, T., Kibak, H., Fishman, J. and Taiz, L. (1992) Evolution and isoforms of V-ATPase subunits. *J Exp Biol*, **172**, 137-147.
Grimont, F. and Grimont, P.A. (1986) Ribosomal ribonucleic acid gene restriction patterns as potential taxonomic tools. *Ann Inst Pasteur Microbiol*, **137B**, 165-175.
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.
Higgins, D.G., Thompson, J.D. and Gibson, T.J. (1996) Using CLUSTAL for multiple sequence alignments. *Methods Enzymol*, **266**, 383-402.
Huelsenbeck, J.P. and Ronquist, F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17**, 754-755.
Huelsenbeck, J.P., Ronquist, F., Nielsen, R. and Bollback, J.P. (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, **294**, 2310-2314.
Jain, R., Rivera, M.C. and Lake, J.A. (1999) Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci U S A*, **96**, 3801-3806.
Janssen, P., Coopman, R., Huys, G., Swings, J., Bleeker, M., Vos, P., Zabeau, M. and Kersters, K. (1996) Evaluation of the DNA fingerprinting method AFLP as an new tool in bacterial taxonomy. *Microbiology*, **142 (Pt 7)**, 1881-1893.
Jin, L. and Nei, M. (1990) Limitations of the evolutionary parsimony method of phylogenetic analysis. *Mol Biol Evol*, **7**, 82-102.
Jin, Q., Yuan, Z., Xu, J., Wang, Y., Shen, Y., Lu, W., Wang, J., Liu, H., Yang, J., Yang, F., Zhang, X., Zhang, J., Yang, G., Wu, H., Qu, D., Dong, J., Sun, L., Xue, Y., Zhao, A., Gao, Y., Zhu, J., Kan, B., Ding, K., Chen, S., Cheng, H., Yao, Z., He, B., Chen, R., Ma, D., Qiang, B., Wen, Y., Hou, Y. and Yu, J. (2002) Genome sequence of Shigella flexneri 2a: insights into pathogenicity through comparison with genomes of Escherichia coli K12 and O157. *Nucleic Acids Res*, **30**, 4432-4441.
Jones, D.T., Taylor, W.R. and Thornton, J.M. (1992) The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci*, **8**, 275-282.
Jukes, T.H. and Cantor, C.R. (1969) Evolution of protein molecules. *Mammalian Protein Metabolism*, 21-32.
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.
Kostman, J.R., Edlind, T.D., LiPuma, J.J. and Stull, T.L. (1992) Molecular epidemiology of Pseudomonas cepacia determined by polymerase chain reaction ribotyping. *J Clin Microbiol*, **30**, 2084-2087.
Lan, R. and Reeves, P.R. (2002) Escherichia coli in disguise: molecular origins of Shigella. *Microbes Infect*, **4**, 1125-1132.
Lawrence, J.G. and Ochman, H. (1997) Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol*, **44**, 383-397.
Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. and Brown, E.L. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*, **14**, 1675-1680.
MacQueen, J.B. (1967) Some methods for classification and analysis of multivariate observations. *Proc. Fifth Berkeley Symp. Mathematical Statistics and Probability*, **1**, 281-297.
Martin, W. (1999) Mosaic bacterial chromosomes: a challenge en route to a tree of genomes. *Bioessays*, **21**, 99-104.
Mau, B., Newton, M.A. and Larget, B. (1999) Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics*, **55**, 1-12.
Nikiforov, T.T., Rendle, R.B., Goelet, P., Rogers, Y.H., Kotewicz, M.L., Anderson, S., Trainor, G.L. and Knapp, M.R. (1994) Genetic Bit Analysis: a solid phase method for typing single nucleotide polymorphisms. *Nucleic Acids Res*, **22**, 4167-4175.

Posada, D. and Crandall, K.A. (1998) MODELTEST: testing the model of DNA substitution. *Bioinformatics*, Vol. 14, pp. 817-818.

Quackenbush, J. (2001) Computational analysis of microarray data. *Nat Rev Genet*, **2**, 418-427.

Rannala, B. and Yang, Z. (1996) Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J Mol Evol*, **43**, 304-311.

Read, T.D., Peterson, S.N., Tourasse, N., Baillie, L.W., Paulsen, I.T., Nelson, K.E., Tettelin, H., Fouts, D.E., Eisen, J.A., Gill, S.R., Holtzapple, E.K., Okstad, O.A., Helgason, E., Rilstone, J., Wu, M., Kolonay, J.F., Beanan, M.J., Dodson, R.J., Brinkac, L.M., Gwinn, M., DeBoy, R.T., Madpu, R., Daugherty, S.C., Durkin, A.S., Haft, D.H., Nelson, W.C., Peterson, J.D., Pop, M., Khouri, H.M., Radune, D., Benton, J.L., Mahamoud, Y., Jiang, L., Hance, I.R., Weidman, J.F., Berry, K.J., Plaut, R.D., Wolf, A.M., Watkins, K.L., Nierman, W.C., Hazen, A., Cline, R., Redmond, C., Thwaite, J.E., White, O., Salzberg, S.L., Thomason, B., Friedlander, A.M., Koehler, T.M., Hanna, P.C., Kolsto, A.B. and Fraser, C.M. (2003) The genome sequence of Bacillus anthracis Ames and comparison to closely related bacteria. *Nature*, **423**, 81-86.

Rychlik, W. (1995) Selection of primers for polymerase chain reaction. *Mol Biotechnol*, **3**, 129-134.

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

Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467-470.

Schmidt, H.A., Strimmer, K., Vingron, M. and von Haeseler, A. (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, Vol. 18, pp. 502-504.

Sokal, R.R. and Michener, C.D. (1958) A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.*, **38**.

Southern, E.M. (1975) Detection of Specific Sequences among DNA Fragments Separated by Gel-Electrophoresis. *Journal of Molecular Biology*, **98**, 503-&.

Strizhkov, B.N., Drobyshev, A.L., Mikhailovich, V.M. and Mirzabekov, A.D. (2000) PCR amplification on a microarray of gel-immobilized oligonucleotides: detection of bacterial toxin- and drug-resistant genes and their mutations. *Biotechniques*, **29**, 844-848, 850-842, 854 passim.

Swofford, D.L. (1993) Paup - a Computer-Program for Phylogenetic Inference Using Maximum Parsimony. *Journal of General Physiology*, **102**, A9-A9.

Swofford, D.L. (1998) Phylogenetic Analysis Using Parsimony. Sinauer Associates, Sunderland Massachusetts.

Tajima, F. and Nei, M. (1984) Estimation of evolutionary distance between nucleotide sequences. *Mol Biol Evol*, **1**, 269-285.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S. and Golub, T.R. (1999) Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences of the United States of America*, **96**, 2907-2912.

Tamura, K. and Nei, M. (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol*, **10**, 512-526.

Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. and Church, G.M. (1999) Systematic determination of genetic network architecture. *Nat Genet*, **22**, 281-285.

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

Toronen, P., Kolehmainen, M., Wong, C. and Castren, E. (1999) Analysis of gene expression data using self-organizing maps. *Febs Letters*, **451**, 142-146.

Vasiliskov, A.V., Timofeev, E.N., Surzhikov, S.A., Drobyshev, A.L., Shick, V.V. and Mirzabekov, A.D. (1999) Fabrication of microarray of gel-immobilized compounds on a chip by copolymerization. *Biotechniques*, **27**, 592-594, 596-598, 600 passim.

Wang, D.G., Fan, J.B., Siao, C.J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M.S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T.J., Lander, E.S. and et al. (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, **280**, 1077-1082.

Wei, J., Goldberg, M.B., Burland, V., Venkatesan, M.M., Deng, W., Fournier, G., Mayhew, G.F., Plunkett, G., 3rd, Rose, D.J., Darling, A., Mau, B., Perna, N.T., Payne, S.M., Runyen-Janecky, L.J., Zhou, S., Schwartz, D.C. and Blattner, F.R. (2003) Complete genome sequence and

comparative genomics of Shigella flexneri serotype 2a strain 2457T. *Infect Immun*, **71**, 2775-2786.
Welsh, J. and McClelland, M. (1990) Fingerprinting genomes using PCR with arbitrary primers. *Nucleic Acids Res*, **18**, 7213-7218.
Williams, G. (1999) Revseq. Genome Campus, Hixton, Cambridge.
Williams, G. (2000) Comseq. Genome Campus, Hinxton, Cambrigde.
Woese, C.R. and Fox, G.E. (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A*, **74**, 5088-5090.
Woese, C.R., Fox, G.E., Zablen, L., Uchida, T., Bonen, L., Pechman, K., Lewis, B.J. and Stahl, D. (1975) Conservation of primary structure in 16S ribosomal RNA. *Nature*, **254**, 83-86.
Zuckerkandl, E. and Pauling, L. (1965) Molecules as documents of evolutionary history. *J Theor Biol*, **8**, 357-366.