Universität Bielefeld

Faculty of Technology

Center for Biotechnology (CeBiTec)

TAXONOMIC CLASSIFICATION OF GENOMIC SEQUENCES: FROM WHOLE GENOMES TO ENVIRONMENTAL GENOMIC FRAGMENTS

By

Naryttza Namelly Díaz Solórzano

SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

DOCTOR RERUM NATURALIUM

AT

FACULTY OF TECHNOLOGY

BIELEFELD UNIVERSITY

BIELEFELD, GERMANY

April 25, 2010

Naryttza Namelly Díaz Solórzano

Chemin du Cèdre 23

1030 Bussigne-prés-Lausanne

Switzerland

ndiaz@CeBiTec.Uni-Bielefeld.DE

Supervisors: JunProf. Dr. Ing. Tim W. Nattkemper

Prof. Dr. Karsten Niehaus

Printed on acid-free paper.

To my mother my biggest source of inspiration

Acknowledgments

Now at the end of my studies I have learned one important thing: It would have been almost impossible to get here without the encouragement and support of lots of people. It is my pleasure to thank those who directly or indirectly made this work possible.

First, I am very grateful to my supervisors Tim Nattkemper and Karsten Niehaus for all their scientific advise invaluable comments and accompanying me throughout my PhD project. I am also grateful to the members of my evaluation committee Susanne Schneiker and specially Jens Stoye for helping me to come back to Bielefeld University to pursue my PhD study as well as his interest and unconditional support to my scientific academic career.

I was able to focus in my work and enjoy my time in Germany due to the funding obtained from the Deutscher Akademischer Austausch Dienst (DAAD). A great support to develop my scientific career was also given by the International Graduate School in Bioinformatics and Genome Research at Bielefeld University.

Special thanks also go to my colleagues and friends from the Bioinformatic Resource Facility in particular the support team Torsten Kasch, Ralf Nolte, Achim Neumann, Volker Tölle. A special thank you goes to Björn Fischer who helped me in the process of printing and handing in my thesis. Without all of you this work would have been almost impossible. Again thank you all for your patient and not throwing me out of the window whenever I tried -believe me never on purpose- to bring the cluster to its knees. I am also in great debt to Alex Goesmann whom always demonstrated truly interest for my PhD project and supported me in many ways to complete this thesis.

I was extremely lucky to find not only great colleagues but also exceptional friends during my time in Bielefeld so my deepest gratitude goes to Heiko Neuweger, Stefan Albaum (Alu), Michael Dondrup, Justina Krawczy, Claudia Rubiano, Jomuna Choudhuri, Martina Mertens, Britta Seefeld, Jan Reinkensmeier, Magdalene Kutyniok (Magga), Christian Martin, Jan-Frederic Meier and Julia Köhler for the sincere friendship and constant support. A very special acknowledgement goes to Diego Rojas who gave me all his support to come to Germany and complete my postdoctoral studies. To my life time friend Yuleima Diaz who has been always there for the good and the bad.

I was able to discover Germany from a very different perspective thanks to two special women: Renate Krause and Imgard Rudolph. Thank you also for the strong moral support during these past years.

Finally, this work would have not been possible without the support of Lutz Krause. Your personal and scientific input was pivotal to succeed in obtaining my PhD. Your help was essential to get me through the good and the bad times that one inevitable encounters during a PhD project. Thanks again for giving me the courage to get to the end.

List of publications

- **Diaz NN**, Krause L, Goesmann A, Niehaus K, Nattkemper TW. TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. 2009. BMC Bioinformatics10:56.

- Kröber M, Bekel T, **Diaz NN**, Goesmann A, Jaenicke S, Krause L, Miller D, Runte KJ, Viehöver P, Pühler A, Schlüter A. Phylogenetic characterization of a biogas plant microbial community integrating clone library 16S-rDNA sequences and metagenome sequence data obtained by 454-pyrosequencing. 2009. J Biotechnol. 1;142(1):38-49.

- Krause L, **Diaz NN**, Edwards RA, Gartemann KH, Krömeke H, Neuweger H, Pühler A, Runte KJ, Schlüter A, Stoye J, Szczepanowski R, Tauch A, Goesmann A. Taxonomic composition and gene content of a methane-producing microbial community isolated from a biogas reactor. 2008. J Biotechnol. 136(1-2):91-101.

- Schlüter A, Bekel T, **Diaz NN**, Dondrup M, Eichenlaub R, Gartemann KH, Krahn I, Krause L, Krömeke H, Kruse O, Mussgnug JH, Neuweger H, Niehaus K, Pühler A, Runte KJ, Szczepanowski R, Tauch A, Tilker A, Viehöver P, Goesmann A. The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-pyrosequencing technology. 2008. J Biotechnol. 136(1-2):77-90.

- Martin C, **Diaz NN**, Ontrup J, Nattkemper TW. Hyperbolic SOM-based clustering of DNA fragment features for taxonomic visualization and classification. 2008 Bioinformatics. 24(14):1568-74.

- Krause L, **Diaz NN**, Goesmann A, Kelley S, Nattkemper TW, Rohwer F, Edwards RA, Stoye J. Phylogenetic classification of short environmental DNA fragments. Nucleic Acids Res. 2008. 36(7):2230-9.

- Krause L, **Diaz NN**, Bartels D, Edwards RA, Pühler A, Rohwer F, Meyer F, Stoye J. Finding novel genes in bacterial communities isolated from the environment. 2006. Bioinformatics. 22(14):e281-9.

- Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, de Crécy-Lagard V, **Diaz N**, Disz T, Edwards R, Fonstein M, Frank ED, *et al.* The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res. 2005. 33(17):5691-702.

Summary

The main goal of this dissertation is to develop a classifier for assigning environmental genomic fragments to the closest known source organism. This has been achieved by the development of a novel method for the TAxonomic COmposition Analysis – TACOA– of environmental genomic fragments using a kernelized nearest neighbor approach. A combination of machine learning techniques has been employed to realized a classifier that exploits the wealth of knowledge deposited in public databases. The developed classifier uses as features oligonucleotide frequencies which carry the so called *genomic signature*. A key advantage of the use of genomic signatures is that enable sequence comparison without alignment. A central assumption of the genomic signature is that oligonucleotide compositions of DNA sequences from the same or closely related organisms are prone to be more similar than those from far related ones.

This work embodies one of the first attempts to tackle the problem of taxonomic classification of metagenomic data. Moreover, it is the first of its kind using a kernelized nearest neighbor approach. The use of the *k*-nearest neighbor algorithm in the TACOA strategy assures that the realized classifier is in its nature multi-class. In addition, this approach has the advantage of not making any assumptions about the distribution of the input data and the classification results can easily be interpreted. However, the traditional *k*-NN algorithm has the drawback of running into problems when dealing with high dimensional input data (called curse of dimensionality). In the kernelized extension presented herein, this problem is overcome by the incorporation of a Gauss kernel into its architecture.

Furthermore, the developed software can easily be installed and run on a desktop computer offering more independence in the analysis of metagenomic data sets. The reference set used by the proposed classifier can be easily updated with newly sequenced genomes, a very desirable feature on this situation of continuing expansion of genomic databases.

The novel strategy presented was extensively evaluated using genomic fragments of variable length (800bp – 50Kbp) from 373 completely sequence genomes. As a whole, the classification accuracy at five different taxonomic ranks was evaluated: superkingdom, phylum, class, order and genus. TACOA is able to classify genomic fragments of length 800bp and 1Kbp with high accuracy until rank class. For fragments longer than 3Kbp accurate predictions are made even at deeper taxonomic ranks (order and genus). TACOA compares well to the latest intrinsic classifier PhyloPythia. For fragments of length 800bp and 1Kbp the overall accuracy of TACOA is higher than that obtained by PhyloPythia across all taxonomic ranks. For all fragment lengths, both methods achieved comparable high specificity results up to rank class and low false negative rates.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

The massive amount of available data in all fields of knowledge has experienced a blistering surge in the past decade. In particular, microbiology has recently undergone a revolution, comparable to the invention of the microscope. Technical advances have motorized extraordinary improvements in the field of DNA sequencing. A boost in speed and efficiency, together with persistent reducing costs, is making possible to deliver even more sequence data into public databases. In particular, the young field of metagenomics has benefited from this breakthrough in sequencing technology. In turn, *metagenomics* or the genomic analysis of co-occuring species in a community is reshaping the landscape of microbiology, ecology, evolution and medicine. Transcending individual genes and genomes, metagenomics offers access to all genomes of a microbial community revealing the secrets of the ”uncultured world”, i.e.the enormous number of microbial species that currently cannot be isolated into pure culture. In the near future, metagenomics will expand our ability to discover and benefit from microbial capabilities, improve our understanding of microbial communities and promises a lead to major advances in medicine, agriculture, energy production and bioremediation.

One key step in the metagenomic approach is to directly sequence the DNA collected from an ecosystem, which may contain thousands of species. After sequencing, typical metagenomics data comprises a vast collection of small fragments that has not association to the organisms from which they were derived. Thus, the first major task imposed by this type of data is to phylogenetically classify raw sequence fragments into related taxonomic groups. The classification step is frequently a *syne qua non* condition for the recovery of complete genomes or assessing the biological diversity of a sample.

Computational challenges come hand by hand with the vast amount of data; their complexity and multi-dimensionality are strongly pushing forward the development of new methods and technologies. They should be able to contribute to the analysis of the data in a high-throughput and ”intelligent” manner, in the hope that new well-founded knowledge can be extracted from the raw data. In this context, *machine learning* methods are employed to unveil valuable information from the data by mining, visualizing and revealing hidden correlations. Two major approaches in machine learning can be recognized:

- *Supervised learning.* In this case a classification function is learned from a reference set of items with known class labels, This process is known as training. Subsequently, the trained classifier, i.e., the learned function is applied to classify new items with unknown class affiliation. An example of a classifier using this type of approach is the Support Vector Machine (SVM) algorithm (Hastie *et al.*, 2002). A second example for a supervised classifier is the *k*-nearest neighbor approach (*k*-NN) (Cover and Hart, 1967). However, in this approach, the classification function is not learned during an explicit training step, but online during the classification phase.

- *Unsupervised learning.* The classifier is not provided with prior knowledge of existing classes. It classifies items based on patterns found in the input data. A classical example of unsupervised learning is the self organizing maps (Kohonen, 1982).

In general, items are classified based on intrinsic features. As features, the taxonomic classifier presented in this work employs the concept of genomic signatures (Karlin and Burge, 1995) which allows alignment-free sequence comparisons. It is based on the postulate that oligonucleotide composition of DNA fragments from the same species or phylogenetically close relatives are prone to be more similar to each other than those from distantly related species. This basic idea has already been used to detect horizontal gene transfers (Karlin, 2001; Merkl, 2004; Dufraigne *et al.*, 2005) and study the evolution of viruses and plasmids (Campbell *et al.*, 1999; Karlin and Mrázek, 2007). In this work, the genomic signature notion was used to taxonomically classify whole genomes as part of the exploratory analysis to evaluate the suitability of employing the genomic signature as a feature.

Traditional genome sequencing and analysis approaches where single-species is studied at a time have generated an immense valuable knowledge ready to be exploited. Completely sequenced genomes, which could be used as references for the taxonomic classification of metagenomic sequences, become available at an exponential rate. Therefore, the taxonomic classification of metagenomic data will greatly benefit from supervised methods that can be instantaneously updated when new genomes are made available.

The work developed in this dissertation can be count within one of the first attempts to tackle the problem of taxonomic classification of genomic fragments from metagenomic data. A novel classifier able to predict the taxonomic origin of environmental genomic fragments of variable length in a supervised manner is presented. As one of the main outcome of this work, the TAxonomic COmposition Analysis method –TACOA– developed was implemented in a software. Furthermore, the developed software can be easily installed and run on a desktop computer offering more independence in the analysis of metagenomic data sets. The reference set used by the proposed classifier can easily be updated with newly sequenced genomes, a very desirable feature in this time of constantly expansion of genomic databases.

TACOA applies the intuitive idea of the *k*-nearest neighbor (*k*-NN) approach (Cover and Hart, 1967) and combines it with a smoother kernel function (Hastie *et al.*, 2002; Tran *et al.*, 2006). Compared to other less intuitive and more complex approaches, *k*-NN based methods have proven to yield competitive results in a large number of classification problems (Berrar *et al.*, 2006; Saha and Heber, 2006; Yao and Ruzzo, 2006; Zhu *et al.*, 2007). In particular, when the classification problem to be solved has a multi-class nature. The kernelized *k*-NN approach used in TACOA allows to realize an accurate multi-class classifier. In general, *k*-NN is intuitive, does not make any assumptions about the distribution of the input data and the reference set can be easily updated. For a wide range of practical applications it approximates the optimal classifier if the reference set is large enough. A further advantage is that the classification results can be easily interpreted. However, the traditional *k*-NN algorithm runs into problems when dealing with high dimensional input data (called curse of dimensionality) (Hastie *et al.*, 2002). In our extension of the *k*-NN algorithm, the introduction of a Gaussian kernel helps to alleviate this problem. (Hastie *et al.*, 2002). By using a smoother kernel function the complete reference set is considered during the classification procedure instead of a strict neighborhood. The presented kernelized *k*-NN approach provides an alternative to solve the problem of taxonomically classifying environmental genomic fragments derived with sequencing technologies producing fragments that are at least 800bp long.

Another aspect regarding the analysis of metagenomic data relates to processing of very short genomic fragments, as well as visualization of metagenomic data. These two aspects were explored within collaborations. For the analysis of very short genomic fragments a framework (in cooperation with Lutz Krause) was developed to identify fragments bearing a partial protein family domain. Subsequently, with the help of a phylogenetic tree the taxonomic origin of fragments bearing a partial protein family domain is assigned to its taxonomic source. Visualization of the pre-clustered data is possible using SOMs, or more precisely with the Poincaré projection of a trained H^{2}SOM. This allows detecting groups of genomic feature vectors having either a low or high variation in feature space in a graphical manner. The visualization work was done as part of a cooperation with Christian Martin.

1.1 Overview of this dissertation

This dissertation is structured as follows:

- Chapter 1 gives a broad overview on all different topics covered along this dissertation. General concepts developed in later chapters are presented, as well as the biological and computational motivations and goals founding this work.

- Chapter 2 introduces in more detail basic concepts and terminology related to the biology and computational aspects used in this dissertation. The notion of metagenomics is presented together with the description of the approach. From the computational perspective, two approaches for the analysis of metagenomic data are presented: first, similarity based and second, compositional based. Important notions of the machine learning algorithms used and evaluated in this work are also reviewed, as well as methods widely used to assess the classification accuracy of the classifiers presented herein.

- Chapter 3 reviews existing approaches directed to sove the problem of taxonomic classification of environmental genomic fragments. Existing methods such as Bayesian classifier, TETRA, self organizing maps, and the support vector machine based PhyloPhytia are discussed. In addition, accuracy results obtained by the above mentioned methods are given.

- Chapter 4 describes the data sets employed in the exploration analysis undertaken in this work, as well as the data set used to evaluate the TACOA classifier. The data sets used for the comparison analysis of TACOA and PhyloPhytia are also explained. Finally, the vector representation of the oligonucleotide features used throughout this dissertation is developed.

- Chapter 5 is a pivotal part of this dissertation presenting the body of results obtained in this work. In section 5.2 the outcome of the exploratory analysis of the features used in the TACOA classifier is given. Following, results from the exploratory classification using a novel implemented SVMs strategy are given. The main contribution of this thesis, the TAxonomic COmposition Analysis method –TACOA–, is presented in section 5.3, as well as the classification accuracy obtained. Section 5.4 focuses on the comparison, in terms of accuracy, between TACOA and the svm-based PhyloPyhtia. In each one of the above mentioned sections, the corresponding strategy, implementation, and evaluation of the method is presented. Furthermore, in section 5.5 the influence of horizontally transfered DNA chunks on the classification accuracy of a composition based classifier is assessed using two case of study. The last section of this chapter, highlights some results obtained as part of a collaboration made within other metagenomic related projects. One of then relates to the analysis of short environmental fragments and the other to the visualization of metagenomic data.

- Chapter 6 discusses particular aspects associated with the results obtained for the TACOA classifier and its classification. The accuracy obtained in the comparison analysis between both classification approaches (TACOA and PhyloPyhtia) is examined. This chapter also considers the manner in which a kernelized *k*-NN strategy can give competitive results when compared to an svm-based approach. A detailed discussion on an adequate interpretation of the accuracy measures used in this work, in the context of multiclass classification, is also given. Finally, the influence of horizontal transfer events on the classification performance of a composition based classifier is interpreted.

- Chapter 7 outlines the main contributions of this dissertation.

- Chapter 8 presents and discusses possible future directions of new aspects to be explored in follow up research.

CHAPTER 2

Background

Overview

This chapter presents fundamental concepts that will be used throughout this dissertation. First, the biological basics of the problem treated herein are stated. Second, statistical techniques employed in this dissertation are introduced. The biological and computational motivations and goals founding this work are also provided.

2.1 Metagenomics

Metagenomics is a new field of research that has recently emerged from genomics. In principle, genomics and metagenomics are devoted to deciphering the DNA sequence or genetic code that serves as the blueprint of life for every living organism and many viruses. In *genomics*, each genome from a single organism is cultured in a lab, subsequently sequenced and finally analyzed. In contrast, in *metagenomics* the collective genomes of all organisms inhabiting a common environment are simultaneously sequenced and analyzed. Metagenomics offers researchers to change the genome-centric paradigm, which focussed on sequencing single species at a time by directly sequencing all genomes sampled from an environment. Therefore, the metagenomics approach allows to bypass the isolation and cultivation procedures, which are estimated to capture only 1% of the microbial and viruses diversity (Rappe and Giovannoni, 2003). In particular, this has been possible with the development of new sequencing techniques that do no require a cultivation step before sequencing.

2.1.1 Sequencing a metagenome

Prior to sequencing, the genomic DNA from organisms collected in an environmental sample, i.e., sample directly extracted form the environment, needs to be extracted. Subsequently, the extracted DNA is sequenced, which is mainly carried out using the whole genome shotgun (WGS) approach (Venter *et al.*, 1998) (Figure 2.2). In the WGS approach, the environmental DNA is directly fragmented into small pieces of variable length, which are later sequenced (Figure 2.1).

An essential step in most sequencing protocols is to generate numerous copies of a DNA fragment, i.e., DNA *amplification*, which can be undertaken *in vivo* or *in vitro*. *In vivo* amplification uses the replicative machinery of a living system (e.g. bacteria) to make copies of a DNA fragment. In the conventional Sanger method (Sanger *et al.*, 1997), this is achieved by cloning a DNA fragment into vectors, i.e., plasmids or fosmids (step 3 in Figure 2.1). These cloning vectors provide the replicative ability that enables the cloned DNA fragment to be copied *in vivo* using a host cell (commonly *Escherichia coli*). In the Sanger method, *in vitro* amplification can also be applied using the *polymerase chain reaction* (PCR). Despite a few advantages (e.g. gain in speed or avoid bias), the use of PCR has not completely substituted traditional cloning in the Sanger procedure.

Following the amplification step, the DNA to be sequenced is put together with enzymes that copy the DNA (i.e. DNA polymerase) and a mixture of standard and modified (fluorescent dye-labelled terminators) nucleotides (Sanger *et al.*, 1997; Fleischmann *et al.*, 1995). The standard nucleotides allow to incorporate other contiguous nucleotides while the modified ones terminate the copying process. As a result, a collection of many prematurely terminated strands (all differing by one nucleotide) is obtained, which are then separated and read using a device that separates the strands by length differing in a single-base-pair. As fragments of each discrete length pass through a special device (capillary electrophoresis instrument) the fluorescent labeled nucleotide can be detected and interpreted by a computational component (Shendure *et al.*, 2008; Lindsay, 2008).

The Sanger technique (Sanger *et al.*, 1997; Fleischmann *et al.*, 1995) produces sequenced DNA genomic fragments of high quality (99.5% accuracy) with a fragment length ranging between 750 and 1,000 base pairs (bp) (Tyson *et al.*, 2004). Despite the high accuracy of the Sanger technique, *amplification bias* can be introduced by the use of *in vivo* cloning. The amplification bias is due to the fact that not all DNA stretches can be successfully amplified in a nonnative living system. Disruption of amplification relates to toxic compounds or intrinsic physical properties, originating form the foreign DNA fragment, that are not compatible with the bacterial host used (Hall, 2007).

In the so called ”next generation” technologies (e.g. 454 *Life Sciencies* or Illumina–Solexa) the DNA is also extracted and fragmented (step 2 Figure in 2.1) as it would be done for the traditional cloning into plasmids. Following, each DNA molecule is attached to short specific oligonucleotide sequences called *adapters*, which are then immobilized on a solid support (beads in 454 or a glass slides in Solexa). A key issue in this sequencing technology, is that only one DNA fragment is attached to one bead (454) or bridged on a glass surface (Solexa) allowing the amplification of individual DNA molecules using PCR. Since the amplification of the DNA fragments is aided by beads or a planar support, on which clusters of identical sequences are formed, it is regarded as *in vitro* cloning (step 3 Figure in 2.1). Therefore, no bacterial cloning step is required to amplify the genomic DNA. To decipher the DNA sequence, each base is interrogated as each fluorescently labeled nucleotide is incorporated by a polymerase, which is another key ingredient of these new technologies. Moreover, this process is carried out simultaneously enabling a higher degree of parallelization compared to the conventional capillarity sequencing, exceeding by far the sequencing capacity of the conventional Sanger method (Shendure and Ji, 2008; Hudson, 2008).

In particular, the use of *in vitro* amplification, that circumvents amplification biases, in the ”next generation” sequencing technologies makes it possible to have a better coverage of the number of different DNA fragments that can be amplified. For the most popular high-throughput technology developed (454 *Life Sciencies*), the average read length has already improved from 100 bp to 400 bp, since it appeared in the market in 2005.

2.2 Computational analysis of metagenomic data

In metagenomic sequencing projects, a basic step following sequencing is the *assembly* of raw reads into longer contigs to gain insight into their taxonomic distribution or functional attributes of the source community inhabiting an specific environment. *Assembly* refers to the process of merging raw reads into contiguous stretches of DNA called *contigs*. A consensus composite contig is produced based on the highest-quality score (low probability of calling a nucleotide incorrectly at that position) or based on a majority rule (the most frequently found nucleotide at each position). The assembly of metagenomic data is a challenging task due to fluctuating read depth produced by the unequal species distribution and the possible co-assembly of reads originating form different species (chimeras) or closely related ones. All these elements contribute to the final quality of assembled contigs to be deposited in public databases. Reads showing high sequence similarity because they stem from closely related species or from highly conserved regions across distantly related species are prompt to be co-assembled.

According to Kunin *et al.* (2008), the performance of assembly programs on metagenomic data is highly variable. A reason for this is that all of them were designed to assemble reads stemming from one genome and not from collective genomes (Mavromatis *et al.*, 2007). A tool called AMOS has been developed to assemble metagenomic

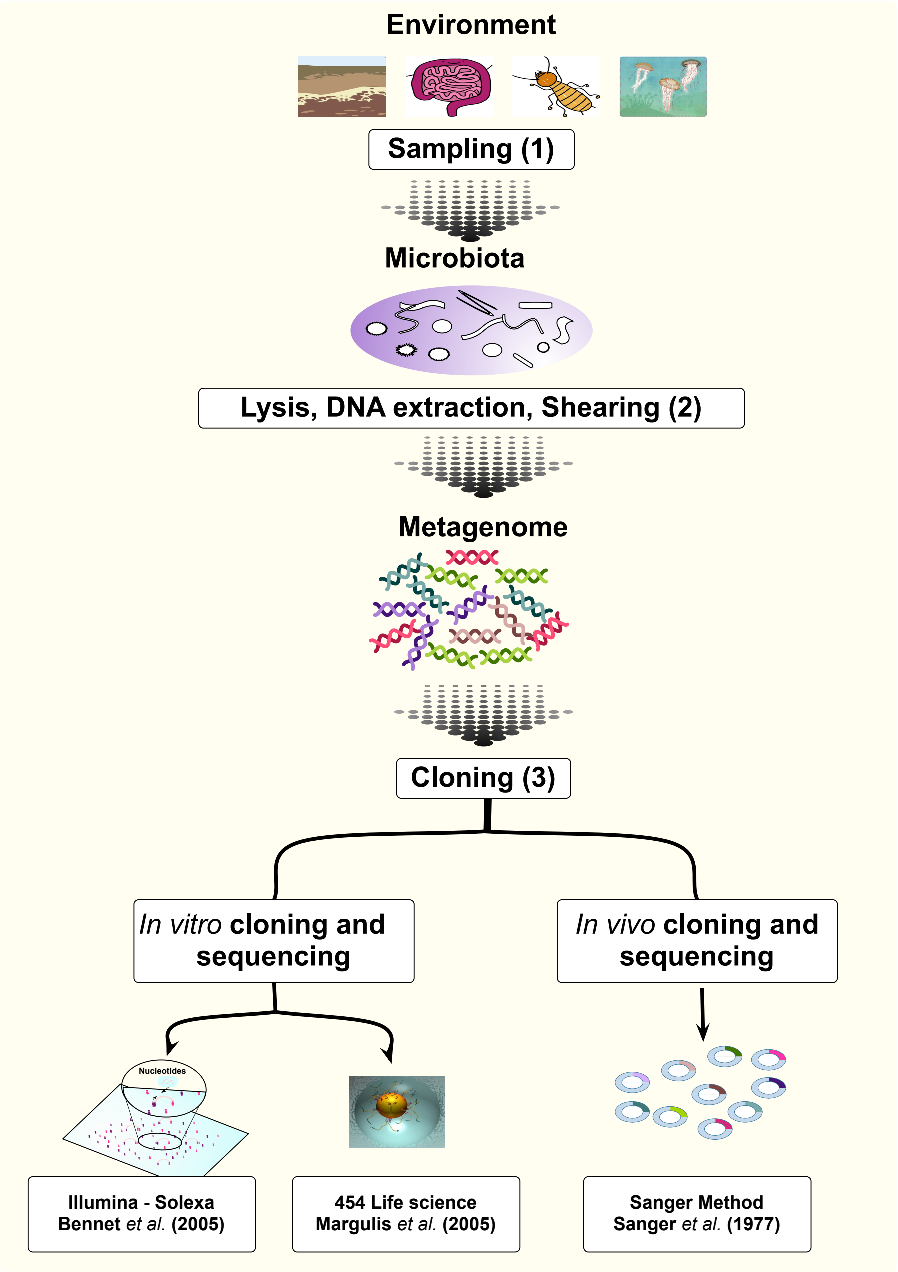


Figure 2.1: Steps carried out in the sequencing process of environmental DNA samples. Steps 1 to 3 are common despite the sequencing technique used. The first common step (1) is to extract the DNA from the organisms by means of lysis. Subsequently, all extracted DNA molecules are mechanically sheared using the shotgun approach. The cloning step (3) can be carried out either *in vivo* (Sanger) or *in vitro* (454 *Life Sciencies* and Illumina - Solexa). *In vivo* cloning refers to the use of modified organisms into which a foreign DNA fragment can be inserted and copied numerous times. Illumina - Solexa and 454 *Life Sciencies* techniques perform the amplification step *in vitro* by means of agarose beads (454 *Life Sciencies*) or bridge amplification on a glass surface device.

data based on a comparative approach (Pop *et al.*, 2004). The AMOS assembler uses reference genomes meaning that only those genomes that have been sequenced can be assembled.

Raw metagenomic data (i.e. unassembled reads) can also be analyzed using the so called *gene-centric approach* by means of mapping each read to a functional category without the need of assembly. Each read having a hit to a functional category is called *environmental gene tag* (EGT) (Tringe and Rubin, 2005). The gene-centric approach is focussed on interpreting the over and under-representation of genes in the studied community, thus treating the community as an aggregate and deliberately obviating the contribution of individual species. The idea behind this reasoning is that in an environment, genes with high frequencies confer beneficial traits to the members of the community embracing them. Relative abundances of gene families permit to focus on prominent functional differences or what the organisms are doing in the studied community. The exploration of how these beneficial genes are interacting with each other can be performed at a higher level, by looking at them as part of broader functional units such as metabolic pathways (Tringe and Rubin, 2005).

To be able to draw hypotheses about the environment from which an environmental sample was taken, it would be desirable to assess the taxonomic information of co-occurring organisms and their genes. The process of predicting the taxonomic affiliation of reads or contigs in a metagenomic sample is called *binning* or classification. The prediction of the taxonomic origin of reads or contigs is an important ingredient to support three major different steps in the analysis of metagenomic data: (i) It facilitates the assembly of highly diverse communities containing a small number of dominant species. For example, a high complexity metagenome can be partitioned into groups or *bins* according to broad phylogenetic relatedness, thereafter each bin is assembled separately. (ii) To reconstruct the taxonomic composition of the studied sample, which helps to derive important community and population-related parameters to understand natural living systems. (iii) In linking interesting gene functions identified in metagenomic reads or contigs to members of the community. For instance, an example often mentioned is the discovery of rhodopsin-like proteins in the bacterial linage. This finding has been a breakthrough for understanding the flux of carbon and energy in the photic zone of oceans worldwide, which is considered a relatively nutrient poor environment. In this case, it took several additional experimental steps to be able to link the rhodopsin-like gene to its phylogenetic source (Béjà *et al.*, 2000).

For the analysis of metagenomic data, computational methods are particularly needed due to the vast amount of information that must be processed. As it was mentioned before, metagenomic data is highly fragmented, thus imposing an additional challenge to bioinformaticians in the process of making sense of the data. So far, large efforts have been devoted to characterize the data in terms of genes, phylotypes, protein domains, and metabolic pathways. The analysis of metagenomic data relates to an important bioinformatics branch: sequence analysis. Without prior modifications, existing traditional computational methods for sequence analysis have difficulties when dealing with these fragmented data. In such a scenario of lack of tailored tools, the contribution that novel computational approaches can provide is of crucial importance.

From the perspective of sequence analysis, two major approaches exist to taxonomically classify metagenomic data: (i) similarity–based methods focus on identifiying genes, domains, conserved gene families using traditional sequence homology methods. (ii) Compositional–based methods aiming to predict the source organism of environmental genomic fragments using intrinsic characteristics directly computed from the genomic sequences.

2.2.1 Similarity–based analysis

Similarity–based analysis makes use of approaches traditionally employed in genomics to search for homology. Similarity-based-methods depend on a sequence-comparison with a reference set of genomic sequences. Similarity-based methods directly align metagenomic sequences to known sequences in a database using the BLAST algorithm (Altschul *et al.*, 1997). Some tools have been developed to build searchable databases suited to annotate and analyze metagenomic data (Huson *et al.*, 2007; Markowitz *et al.*, 2006). The use of BLAST homology searches has been successfully applied for the taxonomic classification of genomic fragments originating from closely related organisms already represented in databases (Kunin *et al.*, 2008) but this may not be always the case for organisms contained in an environmental sample. Although, these databases provide an emerging infrastructure for the analysis of metagenomic data, their practical use is limited given the large number of unknown proteins, and bias towards cultured organisms.

Furthermore, similarity–based methods are also employed to characterize the functional capabilities of a community. Mostly, homology searches are performed against databases such as NCBI cluster of orthologous groups (COG's), Kyoto Encyclopedia of Genes and Genomes (KEGG), the Pfam protein family database to identify the genes present in the community from which the sample was taken. Simple BLAST searches allow to allocate $\approx$ 25-50% of known proteins in a metagenome (Raes *et al.*, 2007). However, this percentage raises to $\approx$ 50-80% when more sophisticated methods are used such as modeling protein domains and building profiles that are later used to to search for protein modules in domain databases (Finn *et al.*, 2008; Letunic *et al.*, 2006; Mulder *et al.*, 2007).

2.2.2 Composition–based analysis

On the other hand, composition-based analysis relies on characteristics which can be extracted directly from nucleotide sequences (e.g. oligonucleotide frequencies, GC-content, codon usage). It has been suggested that sequence composition of genomes reflects environmental constraints (Foerstner *et al.*, 2005).

In absence of a phylogenetic anchor (e.g. rRNA genes) taxonomic classification of genomic fragments can be achieved using nucleotide frequencies. Different cellular processes such as codon usage, DNA base-stacking energy, DNA structural conformation or DNA repair mechanisms can produce *sequence composition signatures* that are species–

specific (Karlin *et al.*, 1997; Campbell *et al.*, 1999). This global statistical property of sequence composition among genomes can be used to determine the taxonomic origin of a genomic fragment (Sandberg *et al.*, 2001; Teeling *et al.*, 2004a; Abe *et al.*, 2005; McHardy *et al.*, 2007; Chan *et al.*, 2008) and to identify atypical genomic regions produced by horizontal gene transfer (HGT) events (Bohlin *et al.*, 2008a; Zhang and Ya-Zhi, 2008). Nucleotide frequencies are a measure of occurrences of words of fixed size in a genomic fragment. The word size routinely used ranges from 1 (GC content) and is not longer than 8 nucleotides (Kunin *et al.*, 2008). These words are known as di-, tri-, tetra-, penta-, hexa-, septa- or octa-nucleotides. In general, longer words produce better taxonomic resolution but due to the highly fragmented nature of metagenomic data their use is not recommendable (Bohlin *et al.*, 2008a). Longer words are not only computationally expensive but they also need longer DNA fragments such that all possible word combinations are sufficiently represented. Most commonly used word sizes producing good results range between 3 and 6 nucleotides (Kunin *et al.*, 2008).

2.3 Application of the metagenomic approach

Since decades, microbiologists have been intrigued with answering classical questions in their field, such as ”Who is out there?” (microbial diversity) and ”What are they doing” (metabolic or functional capacity) (Amann, 2000). With the advent of metagenomics, this hope seems to have materialized. An example of the colossal genetic diversity is given by the Global Ocean Sampling Expedition (Rusch *et al.*, 2007), in which six million proteins (nearly twice the number of proteins present in current public sequence databases) are reported. Furthermore, 1,700 new protein families were discovered with more than 20 representatives per family (Yooseph *et al.*, 2007). These results reported by Yooseph *et al.* (2007) are not surprising if they are taken in light of recent estimates of microbial diversity, which suggest to be in the hundreds of millions to billions microbial species globally (Hugenholtz and Tyson, 2008).

Despite its infancy, metagenomics has already contributed to broaden our understanding of microbial communities and their functional capabilities (Figure 2.2). For example, Tyson and colleagues (Tyson *et al.*, 2004) showed the possibility to reconstruct five genomes (two of them non-culturable species) from the dominant organisms of the acid mine drainage habitat at Iron Mountain, California (USA). Moreover, the authors could bring together the metabolic capabilities of the community inhabiting this extremely acidic effluent (pH between 0.5 and 0.9) and link them to specific strains. Data analysis from the archaea populations in the same acid mine drainage showed that genetic recombination occurs at a much higher rate than previously predicted and is the primary force of evolution in these populations (Eppley *et al.*, 2007). This example shows, that relative simple communities can be explored in all their components using metagenomics. Moreover, metagenomics has astonished the scientific community with striking discoveries, e.g., the presence of proteorhodopsin proteins (light-driven proton pumps) in members of the bacterial domain. These types of proteins were previously thought to be specific to archaeal species. Similarly, the study of the human and mouse gut microbiota has helped to shed light on the mechanisms underlying biomass conversion in these species (Turnbaugh *et al.*, 2006, 2007, 2008, 2009). This emerging area of research holds the potential to provide a better understanding of our ecosystems and the impact of microbes on human health.

The increase in popularity and impact of metagenomics has been facilitated not only by the development of massive sequencing capacity, but also by the assistance of bioinformatics. High-throughput sequence technologies (454, Illumina and ABI) have the capacity to deliver huge amounts of sequence more than 1Gb per run that is vastly more than capillary-based technology can produce (Cardenas and Tiedje, 2008; Wold and Myers, 2008). This ever increasing bulk of data is posting new challenges to the field of informatics (e.g. in issues such as data handling and storage) as well as to bioinformaticians who are urged to develop new tools and methods for the analysis of these data.

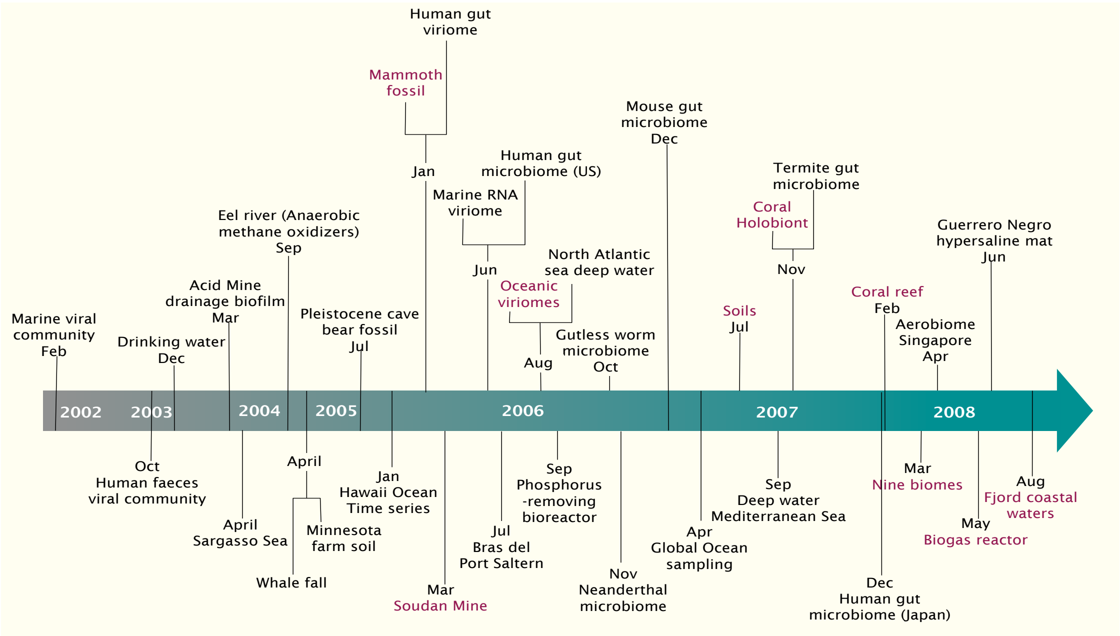

Human gut viriome
Mammoth fossil
Jan
Mouse gut microbiome Dec
Human gut microbiome (US)
Marine RNA viriome
Jun
Termite gut microbiome
Coral Holobiont
Nov
Eel river (Anaerobic methane oxidizers) Sep
North Atlantic sea deep water
Oceanic viriomes
Aug
Guerrero Negro hypersaline mat Jun
Acid Mine drainage biofilm Mar
Pleistocene cave bear fossil Jul
Gutless worm microbiome Oct
Soils Jul
Coral reef Feb
Aerobiome Singapore Apr
Marine viral community Feb
Drinking water Dec
2002
2003
2004
2005
2006
2007
2008
Oct Human faeces viral community
April
April Sargasso Sea
Minnesota farm soil
Whale fall
Jan Hawaii Ocean Time series
Mar Soudan Mine
Jul Bras del Port Saltern
Sep Phosphorus -removing bioreactor
Nov Neanderthal microbiome
Apr Global Ocean sampling
Sep Deep water Mediterranean Sea
Dec Human gut microbiome (Japan)
Mar Nine biomes
May Biogas reactor
Aug Fjord coastal waters

Figure 2.2: Timeline of metagenomic projects and the variety of habitats sampled. The different sequencing technologies mainly use in metagenomic projects are: Sanger dye-terminator (black) and pyrosequencing (red). This information was extracted from the metagenomic projects present at www.genomesonline.org until January 2009. The Soils represent the microbiomes of four different geographical locations. The nine biomes include samples from: stromatolites, fish gut, fish ponds, mosquito viriome, human lung viriome, chicken gut, marine viriome and saltern microbial. (Figure adapted from Hugenholtz and Tyson, 2008)

Although computational methods to analyze the immense amount of metagenomics data are in their infancy, these have already helped in giving some interesting glimpses into our natural world. The amount of interpretable information, regarding taxonomic composition or metabolic capacities, that can be extracted from a sequenced metagenome highly depends on the complexity of the underlying community, being soil one of the most complex communities studied up to date (Hugenholtz and Tyson, 2008).

Metagenomic sequences from low complexity communities can be used to reconstruct nearly-complete composite microbial genome sequences (Tyson *et al.*, 2004; García Martín *et al.*, 2006). Variability in genomic sequences that contribute to the composite genome can be used to evaluate population heterogeneity within a given microbial community (Tyson *et al.*, 2004). For low complexity communities, in which deep sequence-read coverage of individual populations is possible, metagenomics is able to provide an exquisite unique insight into evolutionary processes shaping population of natural microbial systems. An excellent example, refers to the detection of discrete archaeal sequence clusters in acid mine drainage biofilms, related to Ferroplasma types I and II. Additionally, the reduced rate of genetic exchange seen between recently diverged Ferroplasma types I and II relative to the high rates within each population provides support for the concept that the breakdown of homologous recombination in these archaea serves as a species boundary (Eppley *et al.*, 2007). The authors suggested sympatric speciation (i.e. speciation without a physical barrier) as a possible mechanism to explain these observations (Eppley *et al.*, 2007). A highly debated issue in evolutionary theory has been sympatric evolution due to the limited amount of evidence. Although some studies have revealed its existence, these are mostly related to large eukaryotes (e.g. mammals, reptile, fish) (Niemiller *et al.*, 2008; Lodé, 2001; Barluenga *et al.*, 2006). With the advent of metagenomics, evidence for sympatric speciation in bacteria and archaea is emerging as is the case reported by Eppley *et al.* (2007) in which the archaeal populations showed a limited genetic exchange despite inhabiting the same biofilm sampled in an acid mine drainage. For microorganisms, this type of evidence has been missing partially due to their small size, which makes it difficult to differentiate phenotypes among isolated populations. In addition, the identification of potential geographical barriers that prevent gene flow and migrations among microorganisms have been particularly difficult (Whitaker, 2006; Hanage *et al.*, 2006).

Perhaps one of the most important contribution that metagenomics can provide, relates to the highly disputed issue of microbial species definition. Speciation arises either via genetic divergence of coexisting populations or via a geographical barrier that separates populations into discontinuous lineages over time. With the high-throughput techniques coming into play, it is now possible to unveil such patterns of individual-level variation in microorganisms (Whitaker, 2006). Metagenomic analysis can ultimately offer a culture independent way of addressing the number of genome variants present in a community which has been shown to greatly exceed the number of 16s rRNA phylotypes. Even close relative microbial species sharing $\geq$97% of sequence identity of their 16srRNA can display a surprisingly high amount of variability in their proteomes. For example, a survey of 32 strains belonging to *Escherichia coli* and *Shigella* (Willenbrock *et al.*, 2007) as well as the analysis of six strains of *Streptococcus agalactiae* (Tettelin *et al.*, 2005) revealed a large set of ”disposable” genes found in a subset of genomes of each ”species”. Such findings are driving microbiologist and evolutionary biologist to ”re-think” the definition of ”species” and the underlying mechanism that originates it. A more unifying concept of species might be delineated not only by marker genes, e.g. 16s rRNA, but also include the functions encoded by the set or core-of-genes present in the ”pan-genome” representing the sum of all genes found in 16s rRNA-related phylotypes.

Despite its complications, analysis of metagenomic sequence data from complex communities is possible by means of associating metabolic processes to members of the community. For example, Turnbaugh *et al.* (2006, 2008, 2009); Gill *et al.* (2006) and Kurokawa *et al.* (2007) have elucidated connections between the biomass conversion and the underlying microbiome in a variety of natural bioreactors such as the gut of mice and humans. In a recent work, Turnbaugh *et al.* (2009) investigated the gut microbiomes of adult monozygotic and dizygotic twin pairs. The authors could identify a ”core microbiome” at the gene level, rather than at the organismal lineage. Turnbaugh *et al.* (2009) concluded that a diversity of organismal assemblages can still yield a core microbiome at a functional level, and that deviations from this core are associated with different physiological states such as obese or lean.

2.4 Machine learning for classification

The task of predicting the taxonomic origin of a DNA fragment can be regarded as a multiclass classification problem. Given a DNA fragment, the goal is to decide from which of the multiple possible taxonomic classes (e.g. Firmicutes, Chlamydia, Actinobacteria, etc.) the fragment stems. This task can be addressed using statistical classification techniques from the field of machine learning.

In general, the goal of statistical classification is to categorize individual items x from an input space $\mathcal{X}$ into groups based on quantitative information on one or more attributes inherent to the items. Statistical classification can be divided into two major approaches: *supervised* and *unsupervised* (Hastie *et al.*, 2002; Duda *et al.*, 2001; Tarca *et al.*, 2007). In *supervised* classification, a classification function is learned from a so called *training set* of items with known class labels. Formally, let Y be the set of class labels and $\mathcal{X}$ the input space. Given a training set $\{x_j, y_j\}$, $1 \leq j \leq N$ of items $x_j \in \mathcal{X}$ with known class label $y_j \in Y$, the goal in supervised classification is to learn a classification function $f : \mathcal{X} \rightarrow Y$ that assigns a class label $y \in Y$ to each item x_j from the training set. This process is called *training* of the classifier. Subsequently, the trained classifier, i.e., the learned classification function is applied to classify items with unknown class affiliation. In this process, a class label $y \in Y$ is assigned to new items $x \in \mathcal{X}$, which in the following are called *test items*. Let $f_t : X \rightarrow Y$ be the classification function that assigns the correct class label $y \in Y$ to each $x \in \mathcal{X}$. A main goal of supervised classification is to learn a classification function f based on the training set that minimizes the classification error, this is commonly measured by the mean-squared error:

$$MSE(f) = \sum_{x \in \mathcal{X}} (f(x) - f_t(x))^2 \tag{2.1}$$

In the context of this work, a training set is built from DNA fragments (x) of known taxonomic origin (y). During training, a classifier is trained to discriminate between fragments from different taxonomic classes. In other words, a classification function is learned that assigns the respective taxonomic class to each fragment of the training set. Subsequently, the learned classifier could be employed to predict the taxonomic class of new metagenomic DNA fragments of unknown origin.

If a learned classification function is able to nicely reproduce the class labels of the training set it is called well fit to the training data. More formally, a well fit classifier has a low mean-squared error for the training set. The ability of a classifier to correctly predict the class labels of so far unseen items, which were not contained in the training set is called *generalization*. A non trivial task in machine learning is to find a good trade-off between a classifier that is well fit to the training data and at the same time has a good generalization ability. For example, if a training set contains outliers (e.g. items with wrong class labels) a complex, perfectly fitted classifier might achieve only a poor generalization ability. Such a classifier is then called *overfitted* or *overtrained*.

A similar concept in machine learning theory is the bias-variance trade-off. In brief, the bias measures how well a classifier is fit to the training data, i.e., a well fit classifier has a low bias. On the other hand, the variance measures how much the learned classification function depends on the selected training set, i.e., how consistent the learned function is for different training sets. The mean-squared error of a classifier can be expressed as the sum of the bias and variance (Hastie *et al.*, 2002) and hence an optimal classifier should have both a low variance and bias. A complex classifier (e.g. many parameters or high power) might have a low bias but a high variance and hence a poor generalization ability. On the other hand, a too simple classifier may have a low variance but a high bias.

Unsupervised classification methods do not require labeled training data but are able to directly group individual items without a prior knowledge of existing classes. These methods are used to identify patterns in the input data or how the data is organized (e.g. PCA, ICA or SOM). For instance, all metagenomic DNA fragments with a high pairwise sequence identity could be grouped together. The resulting groups would give insights into the diversity and structure of the underlying microbial community.

If the boundaries (*called decision boundaries*) between the learned classes in the input space are linear, a classifier is called linear, otherwise non-linear.

2.4.1 *k*-Nearest Neighbor

The *k*-Nearest Neighbor (*k*-NN) approach was developed by Cover and Hart (1967) and is one of the oldest and simplest methods for statistical classification. A *k*-Nearest Neighbor classifier is a case-based reasoning strategy, which accesses training items at the same time when a new case needs to be classified. Thus, this method does not require an explicit training step. A new item is classified by a majority vote of its neighbors, with the item being assigned to the most common class among its *k* nearest neighbors. In this approach, three key elements can be identified: First, the need of a set of labeled training items, e.g. DNA fragments. Second, a distance function to compute the distance between the labeled items and the test item. Third, the number of *k* nearest neighbors to be considered in the classification step. Formally, let

$$(x_j, y_j) \text{ with } x_j \in \mathcal{X}, y_j \in Y,\ j = 1, \ldots, N \tag{2.2}$$

be the training set (**ref**$_{set}$), where y_j denotes the class membership of each training item x_j. The computation of the nearest neighbors is based on a distance function (commonly Euclidean distance) $d(x, x_j)$.

Let $N_k(x)$ denote the *k*-neighborhood of a test item *x*, which is defined as the set of *k* training items x_j with the smallest distance to *x*. Then *x* is classified into the class y^* with

$$y^* = argmax_{\ y \in Y} |\{x_j | x_j \in N_k(x) \text{ and } y_j = y\}| \tag{2.3}$$

The best choice of the parameter *k* depends upon the classification problem. In general, larger values of *k* will increase the bias and reduce the variance of the classifier and vice versa. Small values of *k* result in decision boundaries with higher variance that well-fit the training set, while large values achieve smooth and stable decision boundaries that avoid overfitting and are more robust (Hastie *et al.*, 2002).

The k-NN algorithm is easy to understand, implement and despite its simplicity, it performs well in many classification tasks. Furthermore, k-NN based methods have provided competitive results in a large number of classification problems (Berrar *et al.*, 2006; Saha and Heber, 2006; Yao and Ruzzo, 2006; Zhu *et al.*, 2007). In particluar, if the classification problem has a multi-class nature. It has also been shown that the error rate of the k-NN algorithm is upper-bounded above by twice the Bayes error, which is the minimal achievable error rate given the distribution of the data (Cover and Hart, 1967). The k-NN is a non-parametric estimation approach, i.e., it does not assume an underlying distribution of the data (Hastie *et al.*, 2002; Duda *et al.*, 2001). The k-NN has the advantage to approximate the optimal classifier if the training set is large enough, however, it runs into problems with high dimensional data (Hastie *et al.*, 2002; Duda *et al.*, 2001).

2.4.2 Kernel functions

Kernel functions $k(x,x')$ are similarity measures $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ between two items x and $x' \in \mathcal{X}$ that can be regarded as computing the dot-product of x and x' in a higher dimensional feature space $\mathcal{F}$:

$$k(\mathbf{x},\mathbf{x}') =< \phi(\mathbf{x}), \phi(\mathbf{x}') >, \tag{2.4}$$

where $\phi : \mathcal{X} \rightarrow \mathcal{F}$ is a mapping function that maps each item of $\mathcal{X}$ into $\mathcal{F}$ (Boser *et al.*, 1992; Schoenberg, 1938). A key concept of kernel functions is that they can time efficiently compute the dot product in the feature space without explicitly mapping the data into that space.

Any learning algorithm that accesses the input data only via dot products can rely on the implicit mapping offered by kernel functions. This is achieved by simply replacing the dot product $< \mathbf{x}, \mathbf{x}' >$ by a kernel function $k(\mathbf{x}, \mathbf{x}')$, which is called the *kernel trick*. In this manner, learning methods can easily be adapted to different problems without changing the underlying algorithm.

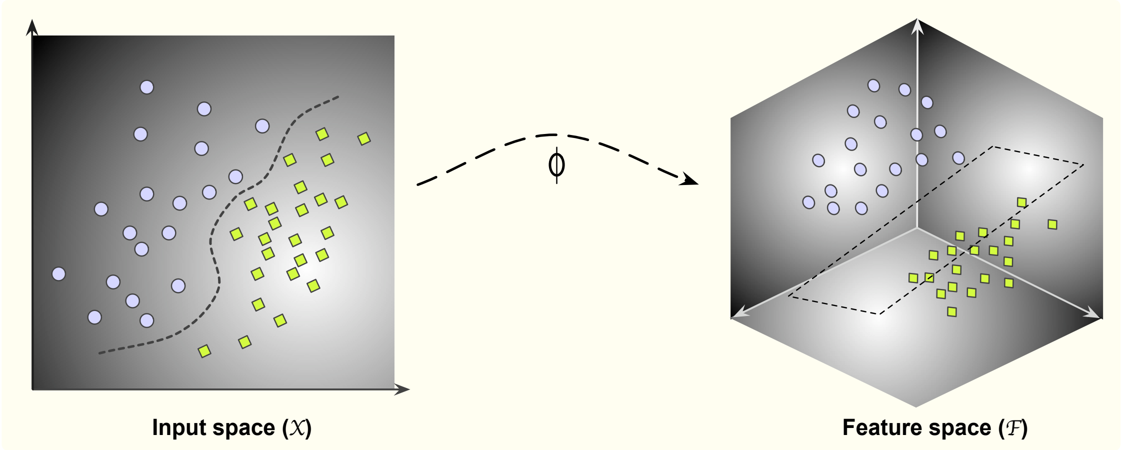


Figure 2.3: Graphic representation of the mapping of kernel functions.
Two classes of objects are depicted (circles and diamonds). On the left the data points for each class are represented in the input space $\mathcal{X}$. After mapping into a higher dimensional feature space via the mapping function ϕ the items become linearly separable (on the right). By learning a linear decision boundary in the feature space, a non-linear decision boundary can be realized in the original input space (dotted lines).

In the context of statistical classification, the kernel trick can be applied to transform a linear classifier into a non-linear one. Assume a given classification problem that is not linear-separable in the input space $\mathcal{X}$ (left side of Figure 2.3). Frequently, a non-linear mapping function $\phi : \mathcal{X} \rightarrow \mathcal{F}$ exists, such that the data becomes linearly separable in $\mathcal{F}$ Figure 2.3). Hence, if the input data is not linearly separable in the input space, a linear classifier can be employed that makes use of a non-linear kernel function $k(\mathbf{x},\mathbf{x}')$ (i.e. the respective ϕ of $k(\mathbf{x},\mathbf{x}')$ is non-linear). Then by learning a linear decision function in the respective feature space defined by $k(\mathbf{x},\mathbf{x}')$, a non-linear classifier can be achieved in the original input space (Figure 2.3).

A kernel based classifier contains two modules: (i) a module that performs the implicit mapping into the feature space via a kernel function and (ii) a linear classifier to discriminate between classes. In this modular context, the feature space can be redefined by changing the kernel without modifying the classification algorithm itself.

The most commonly used kernel function in real world application is the Gaussian kernel. In the following, the Gaussian kernel (K_λ) will be presented in detail since it is used within this dissertation. The Gaussian kernel is defined as:

$$K_\lambda(\mathbf{x},\mathbf{x}') = e^{\left(-\frac{d(\mathbf{x},\mathbf{x}')^2}{2\lambda}\right)}, \tag{2.5}$$

where d is the Euclidean distance and $\lambda > 0$ is a parameter that controls the width of the Gaussian function. The λ parameter relates directly to the bias-variance trade-off of a kernel based classifier. Small values of λ (narrow width of the Gaussian) result in a high variance and a low bias. Conversely, large values of λ (wide width of the Gaussian) lead to a low variance but a high bias. The Gaussian kernel (Equation 2.5) is a decreasing function of the Euclidean distance between points, implying that the larger the kernel $K_\lambda(\mathbf{x},\mathbf{x}')$, the closer the points $\mathbf{x}$ and $\mathbf{x}'$ in $\mathcal{X}$. On the other hand, the Gaussian kernel uses weights that decrease smoothly to zero with increasing distance from the item x to be classified or *target point*. As result, the contribution of items close to the target point is bigger than those located far away. This property is beneficial if the training data is sparsely distributed in the input space.

The selection of an appropriate kernel function for a particular problem is in itself an area of research. Unfortunately, no recipe of how to choose an optimal kernel exists and the choice is usually made based on the trial and error approach. That is, several kernel functions are selected, subsequently the accuracy of the resulting classifier is evaluated. Finally, the kernel function that results in the highest accuracy for the analyzed data is selected. Custom based kernel functions can be developed to incorporate, for example, prior knowledge about the data or to adapt a learning algorithm to different types of input data (e.g. DNA fragments). Although the selection of an optimal kernel function is a demanding task, the popularity of kernel methods in the area of pattern analysis is flourishing.

2.4.3 Support Vector Machine

The Support Vector Machine (SVM) algorithm is a supervised learning method that was initially proposed by Boser *et al.* (1992) and later exhaustively studied by Vapnik (1995, 1998). The SVM algorithm was developed as a binary (two class) classifier and has a strong mathematical foundation and high generalization abilities. The usage of the SVM algorithm has become very popular in recent years. For instance, SVMs have been successfully applied for handwritten digit recognition and also to a variety of biological applications (e.g. gene prediction).

Four key concepts of SVMs can be identified: (i) *separating hyperplane*, (ii) *maximum margin hyperplane*, (iii) *soft margin* and (iv) *kernel function*.

The *separating hyperplane* is a hyperplane which separates the items of two classes (right side of Figure 2.3). An infinite number of hyperplanes separating the items of two classes exists but SVMs select the *maximum margin hyperplane*, which optimally separates two classes, that is, it maximizes the distance between the hyperplane and the nearest data point of each class. By selecting the maximum margin hyperplane, SVMs achieve high generalization abilities for so far unseen items.

In real case scenario outliers may exist in the training set, or items from one class may even be embedded among items of the other class. Then a *soft margin* SVM allows for misclassifications of some training items avoiding overfitting, thus, improving the generalization ability.

In cases where the input data is not linearly separable in input space, SVMs can be combined with non-linear kernel functions. Then by learning an optimal separating hyperplane in the respective feature space, a non-linear classifier can be realized in the original input space (Noble, 2005) (Figure 2.3).

For simplicity, in the following SVMs are introduced in more detail for the case in which the input space X equals $\mathbb{R}^M$. Considering a set of training vectors x_j $(1 \leq j \leq N)$ with known class labels $y_j \in \{+1, -1\}$. Further, let $\mathcal{H}$ be a vector space in which the dot product $< \mathbf{x}, \mathbf{x}' >$ is defined. In the context of SVMs $\mathcal{H}$ is called *feature space* into which the input items are implicitly mapped using kernel functions. Furthermore, the *separating hyperplane* in $\mathcal{H}$ is given by a vector $\mathbf{w} \in \mathcal{H}$ and a scalar $b \in \mathbb{R}$ and is defined as:

$$\{\mathbf{x} \in \mathcal{H} \mid < \mathbf{w}, \mathbf{x} > + b = 0\} \tag{2.6}$$

The *separating hyperplane* that is learned during training separates the vectors of the two training sets. The vector **w** that defines the *separating hyperplane* can be expressed as a linear combination of weighted training vectors:

$$\mathbf{w} = \sum_{j=1}^{N} \alpha_j y_j \mathbf{x_j}, \tag{2.7}$$

where α_j are weights that are assigned to each $\mathbf{x_j}$ during the training phase. The subset of training items $\mathbf{x_j}$ with $\alpha_j \neq 0$ are called *support vectors* (Chen *et al.*, 2005).

The unique *maximum margin hyperplane*, which maximizes the distance between the hyperplane and the nearest data point of each class, allows for improvement in the classification accuracy of new test items with unknown class labels. Given a learned hyperplane, a test item x is classified depending on the side of the hyperplane where it is located. This is done using the following decision function:

$$f(x) = \operatorname{sgn}(< \mathbf{w}, x > + b) \ = \operatorname{sgn}(\sum_{j=1}^{N} \alpha_j y_j < \mathbf{x}, \mathbf{x_j} > + b). \tag{2.8}$$

The item x is classified into the class with the +1 label if $f(x)$ is above 0, otherwise into class with the -1 label. Note that during classification only the support vectors are taken into account (Chen *et al.*, 2005; Ben-Hur *et al.*, 2008).

In practice, a *separating hyperplane* often does not exist if the distributions of the training items from the two classes overlap. As mentioned above, the solution to this problem is to allow the misclassification of some of training items. For this purpose, the key concept of *soft margin* is introduced (Figure 2.4). Soft margin hyperplanes are accomplished by imposing upper bounds to the weights α_j learned during training by a constant $\mathcal{C}$ (Chen *et al.*, 2005; Ben-Hur *et al.*, 2008). The constant $\mathcal{C}$ permits to control the bias and variance of the SVM classifier. If the value of $\mathcal{C}$ is small, outlier items are

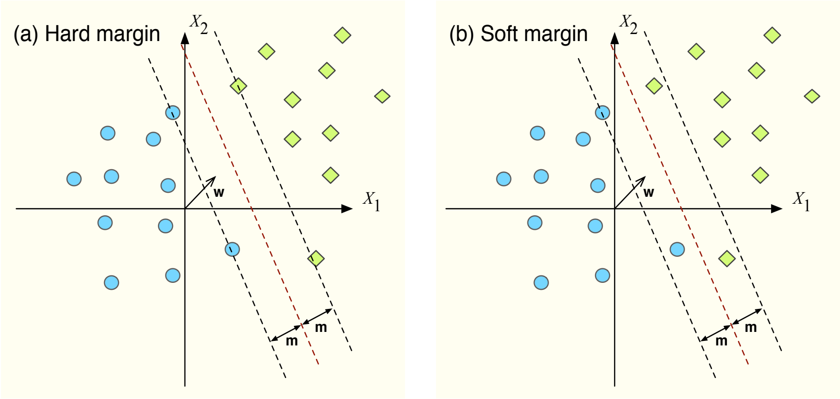


Figure 2.4: Representation of a hard and soft margin SVM.
A hyperplane separates (red dotted line) two classes of items (diamonds and circles). A hard margin (a) does not allow misclassifications of outliers. Compared to a hard margin, a soft margin (b) is wider allowing that outlier items to be misclassified.

misclassified and the margin of the hyperplane w.r.t. the remaining correctly classified vectors increases. In this case, the resulting classifier has a high bias and low variance. Conversely, a large value of $\mathcal{C}$ assigns a large penalty to ”errors”, thus allowing only a low number of misclassifications. Such a classifier will have a low bias but high variance. An SVM classifier using a finite value of the parameter $\mathcal{C}$ is called a *soft margin SVM classifier* (Hastie *et al.*, 2002).

As previously mentioned, in cases where the data set is not linearly separable in the input space, an appropriate transformation of the data into a higher dimensional feature space $\mathcal{H}$ may enable a linear separation in $\mathcal{H}$. In the context of SVMs, this can implicitly be achieved by combining the SVM with an adequate *kernel function* (Noble, 2005; Chen *et al.*, 2005; Ben-Hur *et al.*, 2008). This transformation is performed implicitly by replacing the dot product $< \mathbf{x}, \mathbf{x_j} >$ in Equation 2.8 by a kernel function $k(\mathbf{x}, \mathbf{x_j})$. Non-linear kernel functions can be effectively used by an SVM to learn complex and non-linear decision functions in the original input space. Although kernel functions greatly help in complicated classification problems, the choice of the optimal kernel is troublesome. An adequate kernel function can be determined for example by trying different standard kernel functions and subsequently asessing the classification accuracy of the resulting classifier using cross-validation (see section 2.4.1) (Noble, 2005; Chen *et al.*, 2005; Ben-Hur *et al.*, 2008).

Although SVMs were originally developed for binary classification problems, several extensions of SVMs habe been devised for multiclass classification problems (Vapnik, 1998; Crammer and Singer, 2002). Most methods for multiclass SVM decompose the data set into several binary problems. Two main approaches have been used for the extension to a multiclass SVM: (i) ”one-against-one” and (ii) ”one-against-all”. The ”one-against-one” strategy trains a seperate binary SVM for each combination of two classes. In consequence, for a k-class problem, $k(k-1)/2$ SVMs are trained. In the classification phase, a voting scheme is used: a test item x is classified to the class obtaining the maximum number of votes. On the other hand, in the ”one-against-all” approach a binary SVM is trained for each class to separate vectors of that class from vectors of all remaining classes (Hastie *et al.*, 2002). This strategy builds as many k binary SVMs as there are different classes. The output label of the multiclass classifier for new items is given by the class whose associated classification score is maximal (Hastie *et al.*, 2002).

2.4.4 Self Organizing Maps

Self Organizing Maps (SOMs) are based on *competitive learning* in which *units* or artificial neurons compete to represent a pattern of the input space. This representation is realized on a one-, two- or multi-dimensional map or *topological map*, which is predefined and organized in a grid (Figure 2.5) (Hastie *et al.*, 2002). Self organizing maps are a unique kind of artificial neural networks because they employ a neighborhood function to preserve topological attributes of the input space, i.e., the relative distances of items in input space (Kohonen, 1982). The main task of the SOM algorithm is to find a way to distribute the data on the grid while preserving topological attributes of the input space, this is achieved in feature space by modifying the neurons while the data items remain fixed (Kohonen, 1982). In consequence, SOMs can be applied as a visualization or a clustering tool of high-dimensional data, e.g., by mapping multidimensional data onto a two dimensional map. SOMs can be extended to work in hyperbolic space by applying

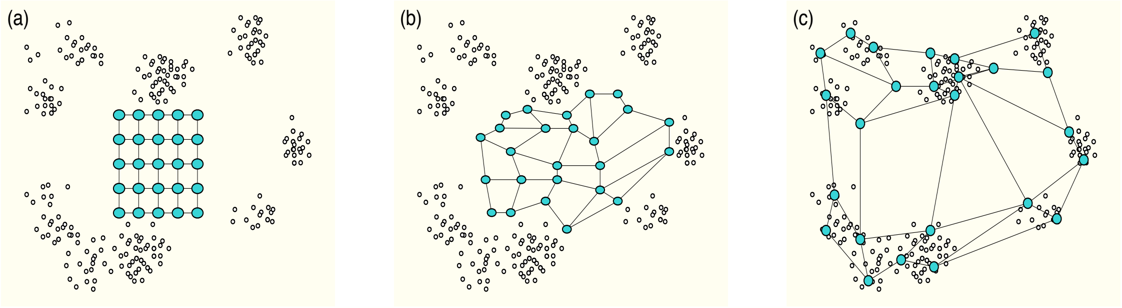


Figure 2.5: Training of a Self-Organizing map (SOM).
Each unit or prototype is represented by a blue circle. (a) Initially, the SOM starts with a rectangular grid. During the learning process the weight vectors approximate the distribution of the data items while keeping their grid structure (b). At the of the training, each unit canbe regarded as a prototype vector of a small region of the input space (c). The map itself is two dimensional but the units in the map have the same number of dimensions as the feature space.

a tree-like grid in that space (Ontrup and Ritter, 2006). Some advantages of hyperbolic SOMs is that they can model hierarchically organized data and allow a very fast training of the algorithm (Ontrup and Ritter, 2006).

A key notion of the SOM algorithm are the *units*, which can adapt to different regions of the input space. Theoretically, *units* can be arranged on a one-, two-, or multidimensional grid and they have lateral connections to neighboring units. However in practice, two-dimensional grids are used in visualization applications. The grid can have a rectangular form but other forms are also used (e.g. hexagonal). The SOM algorithm employes a neighborhood function η, whose value $\eta(u_i, u_m)$ represents the strength of the coupling between unit u_i and unit u_m during the training process. A simple choice is defining $\eta(u_i, u_m) = 1$ for all units u_i in a neighborhood of radius r of unit u_m and $\eta(u_i, u_m) = 0$ for all other units (Rojas, 1996).

Let $\mathcal{S} = x_j$ be the data set of items (in the context of SOMs called observations) in an n-dimensional input space $\mathcal{X}$. Furthermore, the *units* are arranged in a two dimensional rectangular grid of Z *computing units*. Each *unit* u_i, $1 \leq i \leq Z$ is associated with an n-dimensional weight vector $\mathbf{v_i}$. The learning of a SOM proceeds in four steps and over a preset number of iterations:

Initialization: The n-dimensional weight vectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \ldots, \mathbf{v}_Z$ of the Z *computing units* are randomly selected. An initial radius r, a learning rate ρ, and a neighborhood η are chosen.

step 1:

Randomly select an input vector $\xi \in \mathcal{S}$

step 2:

Select the unit u_m with minimal distance between $\mathbf{v}_m$ and ξ, $m = 1, \ldots, Z$

step 3:

The weight vectors are updated using the update rule and neighborhood function

$$\mathbf{v}_i \leftarrow \mathbf{v}_i + \rho\ \eta\ (u_i, u_m)(\xi - \mathbf{v}_i), \text{for } i = 1 \ldots, Z$$

step 4:

Stop if the maximum number of iterations has been reached; otherwise continue with step 1. Notably, the learning rate ρ is linearly reduced with each iteration.

In each iteration, the weight vectors $\mathbf{u}_m$ and the units in the neighborhood of $\mathbf{u}_m$ are attracted in the direction of the observations ξ. The neighborhood function is larger for units that are closely together (especially the same unit) and decreases as the distance between the two units increases. It is important to note that neighbourhood refers to the proximity of the units on the grid not how close their weight vectors are in feature space. The final *topological map* is deformed (compared to the original) to adapt to input data (see sketch (a) and (b) in Figure 2.5) and neighboring units remain together after adaptation (Rojas, 1996).

2.5 Methods for evaluating the classification accuracy

An essential issue in machine learning relates to judge *generalization capability* or its ability of correctly predicting unseen examples of the learning method. The degree of generalization capability is evaluated by the ”closeness” between the learned function and the true function, measure by the *generalization error*. In machine learning problems, a good model or classifier is one that minimizes the generalization error (produces good predictions) and not the training error on a particular data set. Assessment of the *generalization error* can be achieved by employing *analytical methods* that seek for the model having minimal generalization error. On the other hand, *empirical methods* based on efficient sample re-use are also applied, e.g., cross-validation and bootstrap (Hastie *et al.*, 2002). Among all existing *empirical methods*, cross-validation is one of the most simple and widely used method for estimating the *generalization error*.

2.5.1 Leave one out cross validation strategy

In this dissertation the Leave-One-Out Cross-validation (LOOC) strategy is employed, which is a special case from the K-fold cross-validation. The idea behind the K-fold cross-validation is to divide the training data into K-parts, then to use part (K-1) of the available data to fit the model, and a different part (K-*th*) to test it. Subsequently, the average error across all K trials is quantified. Ideally, the data set should be partitioned into K parts of equal size. The case in which K equals the size of the data set is known as *leave-one-out* cross-validation. In the special case of LOOC, a single item is removed from the training set and the classifier is trained with the remaining training data. Subsequently, in each step each single item is classified and the generalization error measured.

2.5.2 Measurements for assessing the classification accuracy

A crucial step for evaluating the predictions of a classifier relates to quantifying misclassifications or the times a classifier wrongly predicts the class of an example. The intuition behind evaluating the classification accuracy is to compare the predicted class of an item with its known class label. The task of predicting the taxonomic origin of environmental fragments is in essence a multi-class classification problem because several taxonomic classes exist into which a genomic fragment can be assigned.

The binary classification version of sensitivity and specificity proposed by Baldi *et al.* (2000) was employed and adapted it to a multi-class classification problem. The classification accuracy was evaluated for each taxonomic class. Let the *i*-th class be denoted as class i. Further, let Z_i be the total number of items from class i, the true positives (TP_i) the number of items correctly assigned to class i, the false positives (FP_i) the number of items from any class $j \neq i$ that is wrongly assigned to i. The false negative rate (FN_i) is defined as the number of items from class i that is erroneously assigned to any other class $j \neq i$. For an item whose class label cannot be inferred, the algorithm classifies it as "unclassified". The unclassified items (U_i) are the number of items from class i that cannot be assigned to any known class, so $Z_i = TP_i + FN_i + U_i$.

The sensitivity ($\mathbf{Sn}_i$) for class i is defined as the percentage of items from class i correctly classified and it is computed by:

$$\mathbf{Sn}_i = \frac{TP_i}{Z_i} \tag{2.9}$$

The reliability (expressed in percentage) of the predictions made by the classifier for class i is denoted as specificity ($\mathbf{Sp}_i$) and it is measured using the following equation:

$$\mathbf{Sp}_i = \frac{TP_i}{TP_i + FP_i} \tag{2.10}$$

Note that the specificity for class i is undefined for those cases when the terms TP_i and FP_i are both zero. Herein, the average specificity is computed over those classes on which a specificity can be mathematically defined.

We make use of the false negative rate ($\mathbf{FNr}_i$) to measure the percentage of items from class i that is misclassified into any class $j \neq i$, which is given by:

$$\mathbf{FNr}_i = \frac{FN_i}{Z_i} \tag{2.11}$$

CHAPTER 3

Existing methods for the taxonomic classification of environmental genomic fragments

Overview

This chapter will introduce existing methods developed to address the problem of taxonomically classifying environmental DNA fragments. From a machine learning point of view composition- and similarity-based methods can be further divided into supervised and unsupervised. In the context of this work, supervised methods require a reference set of genomic sequences with known taxonomic origin. Supervised composition-based methods employ the reference set to learn sequence characteristics of each taxonomic class during a training phase. Methods such as a Bayesian classifier (Sandberg *et al.*,

2001) and PhyloPythia (McHardy *et al.*, 2007) fall into the supervised composition-based category.

On the other hand, unsupervised approaches do not depend on reference sequences for classification, instead characteristics are directly learned from the data set being analyzed. In the context of metagenomics, unsupervised methods are used to group genomic sequences such that all sequences originating from the same taxon are grouped into one cluster. Several unsupervised methods have been developed for the analysis of metagenomic data. TETRA (Teeling *et al.*, 2004a,b) was a pioneering study, followed by the work of Abe *et al.* (2005, 2006) who employed a self-organizing map (SOM). In this chapter, all existing methods for the taxonomic classification of environmental genomic fragments will be discussed in more detail, as well as important results drawn.

3.1 Bayesian classifier

The work of Sandberg *et al.* (2001) was the first to explore the feasibility to use oligonucleotide frequencies to classify DNA fragments. However, it was casted on the challenge of identifying the taxonomic origin of horizontally transfered regions and not for metagenomic data. The authors analyzed DNA fragments of six different lengths: 35, 60, 100, 200, 400, and 1000 bp, using nine different oligonucleotide lengths (between 1 and 9 nucleotides). The classifier developed by the authors is based on Bayesian statistics, which relates the conditional and marginal probabilities of two random events. That is, given that event A occurred, how likely (P) is event B to occur, $P(B \mid A)$ (Sandberg *et al.*, 2001; Langley *et al.*, 1992). In the context of taxonomic classification of genomic sequences, the probability of finding a sequence, S, in a genome, G, can be use to calculate the probability of a sequence to belong to a certain genome, $P(G \mid S)$ (Sandberg *et al.*, 2001). Using a set of 25 different organism, including bacteria and archaea, the authors taxonomically classified DNA fragments as shorts as 400 bp with 85% accuracy (correctly classified). Sandberg *et al.* (2001) showed that classification accuracy increases when longer oligonucleotide length were analyzed, the highest accuracy obtained was using oligonucleotides of size 8 and 9. Furthermore, the authors speculate that with some improvements their proposed naive Bayesian classifier could be employed to analyze the microbial species composition from environmental samples (Sandberg *et al.*, 2001).

3.2 TETRA

A pioneering study to investigate the problem of taxonomic classification of genomic fragments was carried out by Teeling *et al.* (2004a). Tetra-oligonucleotides were employed to recover the phylogenetic signal of genomic fragments of length $\approx$ 40 Kbp. Correlations, i.e. z-scores, were used to measure the divergence between observed and expected oligonucleotide frequencies. To predict the taxonomic origin of a given DNA fragment, Pearson correlation coefficients were employed. The authors evaluated a data set of 118 different genomes including archaea and bacteria, from which a synthetic fosmid-size ($\approx$ 40 Kbp) library was generated. In addition, fosmid inserts for two real fosmid libraries from methane-rich habitats were analyzed.

The TETRA method provided reliable assignments (with correlation coefficients close to 1) when closely related species (at the taxonomic level of species) exist in the reference set. However, these correlation coefficients deteriorate when higher taxonomic levels are considered (e.g. order, class, phylum, and superkingdom). A main limitation of this approach is that it is not suited for large metagenomic data set because the all-versus-all pairwise matrix of all test fragments becomes quickly intractable (McHardy *et al.*, 2007). Another limitation is that the minimal genomic fragment length required to get reliable results with this method is 20Kbp (Teeling *et al.*, 2004b).

3.3 Self Organizing Maps

Abe and coworkers (Abe *et al.*, 2005, 2006), showed the feasibility to accurately classify environmental genomic fragments with minimal length of 5Kbp in an unsupervised manner using a self-organizing map (SOM). The authors also classified genomic fragments of length 1 Kbp but with low accuracy (69%). The SOM method (Abe *et al.*, 2005) is able to recognize in a DNA fragment key combinations of short oligonucleotide frequencies that are the signature of each genome. By using these signatures the SOM separated DNA fragments into species-specific clusters without prior information about the species (Abe *et al.*, 2005). In this work, several SOMs were trained using oligonucleotides of length 2, 3, and 4. To evaluate the SOM based classifier, 81 completely sequenced genomes were selected, in this data set only one representative per species was chosen. The highest clustering power was achieved when the classifier was trained using 5 Kbp long sequences and tetra-nucleotide frequencies.

3.4 PhyloPythia

PhyloPythia, a supervised composition-based method, uses over-represented oligonucleotide patterns as features to train a hierarchical collection of Support Vector Machines (SVMs). The trained SVMs are subsequently used to predict the taxonomic origin of genomic fragments as short as 1 Kbp (McHardy *et al.*, 2007). PhyloPythia also makes use of the genomic signatures, with the exception that the evaluated oligonucleotides of a fixed length are not always literal strings of nucleotides. In more detail, the oligonucleotide patterns used in this framework are regular expressions, which may contained ”gaps” or ”wild cards”, that is, nucleotides can be ”ignored” at certain positions in the analyzed oligonucleotide of a fixed length.

Due to the multiclass nature of the taxonomic classification problem, PhyloPythia uses a hierarchical collection of SVMs. In this framework the ”all-versus-all” technique is employed to extend the SVM to multiclass using a total of $k(k-1)/2$ distinct binary classifiers, where N is the number of classes at a given taxonomic rank. At every taxonomic rank, several SVMs are used, one for each pair of taxonomic classes. The taxonomic class prediction is made based on a majority vote scheme and in case of a tie the assignment to a class is made at random. Additionally, in a post processing step the prediction is confirmed or rejected using a binary ”one-versus-all” SVM (McHardy *et al.*, 2007).

PhyloPythia was evaluated using a total of 340 completely sequenced genomes from archaea, bacteria and eukarya. At each taxonomic level, severeal classifiers are employed, each one trained on genomic fragments of certain length. The different DNA fragment lengths used to train the classifiers were: 1, 3, 5, 10, 15, and 50 Kbp. During the evaluation of this framework, the authors tested two different kernel functions (linear and Gaussian) and showed that the Gaussian kernel is better suited for the taxonomic classification problem (McHardy *et al.*, 2007).

Support Vector Machines demonstrated to achieve a high classification accuracy for fragments of length $\geq$ 3 Kbp and moderate for 1 Kbp long fragments. However, the complete collection of classifiers built into the framework of PhyloPythia need to be retrained (a computationally expensive procedure) when newly sequenced genomes are added to the training set.

CHAPTER 4

Data

Overview

This chapter describes the hierarchical biological nature of the data used in this work. In addition, a description of the data sets employed in the exploration analysis is given. The data set used in the comparison of TACOA and PhyloPhytia is also introduced. Finally, an explanation on how the data is represented as feature vectors, which are used throughout this dissertation, is provided.

4.1 Data sets

4.1.1 Data sets used in the exploratory analysis

For feature exploration

The feature exploration was carried out using a data set comprising 350 genomes downloaded from the SEED database (Overbeek *et al.*, 2005). The selected genomes represent 2 Superkingdoms, 11 Phyla, 20 Classes, 41 Orders, and 59 Genera. The taxonomic information for this data set was collected from the taxonomy database located at the US National Center for Biotechnology Information (NCBI) (Wheeler *et al.*, 2002). Some of the genomes downloaded from SEED were unfinished and present as several contigs. In this case, all contigs of each genome were arbitrarily joined together. The exploration of the features used in the different classifiers was carried out on the whole genomes of the 350 genomes data set named in the following 350-genomes.

For the exploratory classification experiment

In the course of the exploratory classification experiment a novel multiclass SVM was developed and its classification accuracy was tested by using two benchmark data sets. Namely, the iris and wine data sets from the UCI machine learning repository (Asuncion and Newman, 2007).

Iris data set This data set contains information regarding different types of Iris flowers. It contains 50 examples from each of the three type of flowers (Setosa, Versicolor and Virginica). The data set has 4 numerical features: sepal length,sepal width, petal length, and petal width. One class is linearly separable from the other two and the remaining two have some overlap.

Wine recognition data set This dataset contains instances of three categories of wine derived from a chemical analysis of wines grown in a region in Italy but derived from three different cultivars. The datset contains the chemical analysis of these wines in the form of 13 numerical attributes, 59 instances of wine with the following class distribution class 1: 59 examples class 2: 71 examples class 3: 48 examples.

The complete set of 350 genomes previously used for the feature exploration was employed to examine the classification accuracy of an SVM classifier using the selected features. To assess the classification accuracy of the novel devised multiclass SVM four different data set were generated. The use of these four different data sets will allow to examine the influence of the number of contained classes in the performance. Each data set contains 6, 12, 18 and 29 different taxonomic classes at rank order. At this rank, a taxonomic class was included in the test data set if at least three different organisms for that class existed. Each data set was partitioned as follows:

6-orders data set comprises organisms from the orders: Actinomycetales, Chlamydiales, Bacillales, Lactobacillales, Rhizobiales, and Enterobacteriales.

12-orders data set contains all orders from the 6-orders data set plus: Rickettsiales, Burkholderiales, Campylobacterales, Alteromonadales, Pseudomonadales, and Thermoplasmatales.

18-orders data set contains all orders from the 12-orders data set plus: Neisseriales, Pasteurellales, Xanthomonadales, Spirochaetales, Methanosarcinales, and Thermococcales.

29-orders data set contains all the following orders: Actinomycetales, Bacteroidales, Chlamydiales, Chroococcales, Nostocales, Prochlorococcales, Bacillales, Lactobacillales, Clostridiales, Mycoplasmatales, Rhizobiales, Rickettsiales, Burkholderiales, Neisseriales, Desulfuromonadales, Campylobacterales, Alteromonadales, Enterobacteriales, Legionellales, Pasteurellales, Pseudomonadales, Thiotrichales, Vibrionales, Xanthomonadales, Spirochaetales, Sulfolobales, Methanosarcinales, Thermococcales, and Thermoplasmatales.

4.1.2 Data sets used to evaluate two different classifiers

Data set used in the kernelized k-nearest neighbor classifier

As a proof of concept the kernelized k-nearest neighbor method was evaluated on a data set containing 373 completely sequenced genomes and named in the following 373-genomes. The 373-genomes data set comprised a vast majority of members from the archaeal and bacterial phyla. All completely sequence genomes available up to March 2008 were downloaded from the SEED database (Overbeek *et al.*, 2005). The selected genomes represent 2 Superkingdoms, 11 Phyla, 21 Classes, 45 Orders, and 61 Genera. The taxonomic information for this data set was collected from the taxonomy database located at the US National Center for Biotechnology Information (NCBI) (Wheeler *et al.*, 2002). Some of the genomes downloaded from SEED were unfinished and present as several contigs. In this case, all contigs of each genome were arbitrarily joined together.

Data set used for the comparison of TACOA and PhyloPythia classifiers

A set of 63 completely sequenced genomes was downloaded from the NCBI genome database (Wheeler *et al.*, 2002). In the following, this data set is named as the 63-genomes data set. It comprises completely sequenced genomes from 2 Superkingdoms, 12 Phyla, 22 Classes, 38 Orders, and 54 Genera. The taxonomic information for the 63-genomes data set was collected from the taxonomy database located at the US National Center for Biotechnology Information (NCBI) (Wheeler *et al.*, 2002).

4.2 Hierarchical taxonomic organization

Taxonomic classification refers to the categorization of organisms into groups reflecting the principle of common descendent proposed by Darwin in his origin of species work (Darwin, 1859). Before the DNA molecule was discovered organisms were grouped using morphological characteristics (e.g. number of legs, presence-absence of hair, etc.). With the advent of molecular methods the taxonomic classification of organisms is being constantly revised and modified. Currently, the most widely accepted scheme is the three domain system proposed by Woese *et al.* (1990). In microbial taxonomy, eight *taxonomic levels* or *ranks* exist but five are the most commonly used, namely: Superkingdom, Phylum, Class, Order and Genus.

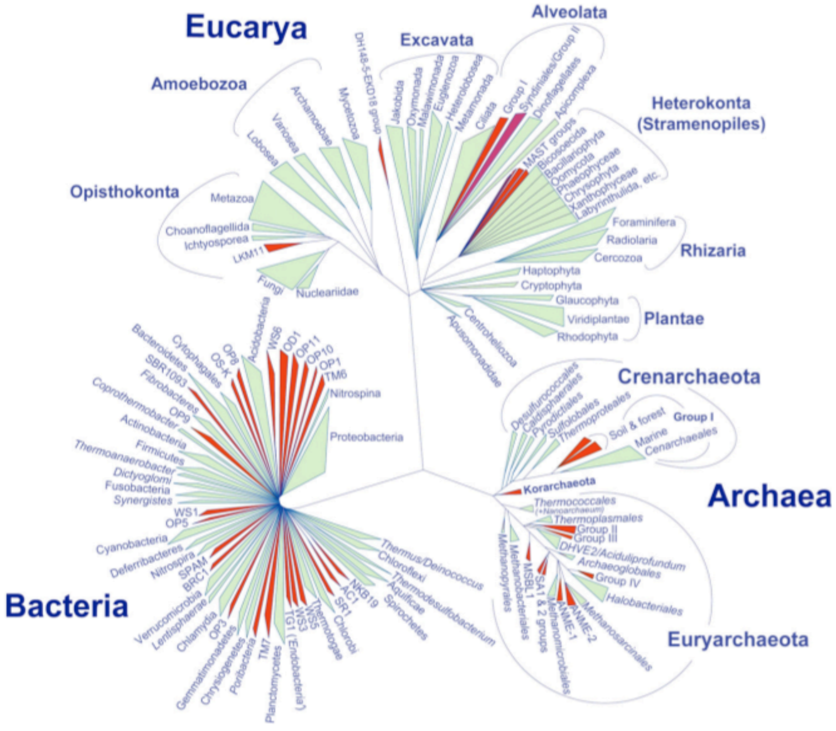


Figure 4.1: Schematic representation of the three-domain system.

This tree is based on molecular evidence. The broader taxonomic groups or ranks are represented by archaea, bacteria and eucarya. Triangles depict the organismal groups at the taxonomic rank of phylum. Green color represent taxonomic groups for which at least one member has been cultivated. Red triangles depict highly divergent or candidate divisions for which no member has been cultivated. Tree of life figure taken from (López-García and Moreira, 2008).

4.3 Vector representation of features

The advantage of using oligonucleotide frequencies as features to taxonomically classify genomic fragments is that a given DNA fragment can be represented in a vector. The entries of a vector are all possible oligonucleotide frequencies of fixed length. This vector representation of genomic fragments implies that they can be directly used by the kernel module of a kernel based classifier. Each DNA fragment is represented in a n-dimensional space in which each oligonucleotide of a fixed length constitute one axis and its frequency in the analyzed genomic fragment (Salton *et al.*, 1975).

4.3.1 Computation of genomic feature vectors using the oligonucleotide frequency deviation

In the following, the computation of genomic feature vectors (GFVs) used throughout this dissertation is described in detail. Computation of the GFVs is performed for each genome in the reference set as well as for each genomic fragment (read or contig) to be classified.

An oligonucleotide o is defined as a string over the alphabet $\sum = \{a,t,c,g\}$. The total number of possible oligonucleotides of length l is given by 4^l, e.g. for l = 3 oligonucleotides can take the form of $o^{[1]} = aaa, o^{[2]} = aat, \ldots, o^{[64]} = ggg$. To generate a GFV for a genomic fragment, the oligonucleotide deviation score is computed for each oligonucleotide. Given the GC-content of the analyzed fragment, the oligonucleotide deviation score is defined as the ratio between the observed oligonucleotide frequency in the fragment and the expected oligonucleotide frequency in that fragment. The GC-content should be subtracted because it has a profound impact on the sequence composition of genomes but a low phylogenetic signal. It has been shown that closely related organisms coming from different environments may show profound differences in GC-content (Foerstner *et al.*, 2005).

Formally, given a genomic fragment s, for each oligonucleotide $o^{[y]}(y = 1, \ldots, 4^l)$ the number of occurrences of $o^{[y]}$ in s is counted. The computation of the oligonucleotide frequencies is conducted in a sliding window approach with step size of 1 and window size l. This approach is carried out on the forward and reverse DNA strand.

In order to more efficiently recover the phylogenetic signal contained in the oligonucleotide frequency deviation, biases introduced by the GC-content of the genomic fragments are corrected. The expected frequency for a certain oligonucleotide o in a genomic fragment s can be estimated by:

$$E[o] \approx |s| - (l-1) \prod_{q=1}^{|o|} p(o_q) \tag{4.1}$$

The length of a genomic fragment is defined as $|s|$ and $|o|$ is the length of an oligonucleotide. Let $O[o_q]$ be the observed occurrence of oligonucleotide o_q in the analyzed genomic fragment, then $p(o_q)$ is estimated by $p(o_q) = \frac{O[o_q]}{|s|}$. For each oligonucleotide o, a GC-normalized deviation score $g(o)$ is computed in a given genomic fragment. The deviation score $g(o)$ resolves for under-represented (negative value) and over-represented (positive value) oligonucleotide frequencies in a genomic fragment. The deviaton score $g(o)$ is given by:

$$g(o) = \begin{cases} 0 & \text{if } O[o] = 0 \\ \frac{O[o]}{E[o]} & \text{if } O[o] > E[o] \\ -\frac{E[o]}{O[o]} & \text{if } O[o] \leq E[o] \end{cases} \tag{4.2}$$

The computed $g(o)$ for each possible $o^{[y]}$ of length l in a given genomic fragment is summarized in a GFV $\mathbf{x}$ (Equation 4.3), this approach is also referred to as the vector representation model (Salton *et al.*, 1975).

$$\mathbf{x} = \left(g(o^{[1]}), g(o^{[2]}), \ldots, g(o^{[4^l]}) \right)^T \tag{4.3}$$

CHAPTER 5

Results

Overview

This chapter presents the body of results obtained in the course of this work. In section 5.1 the outcome from the exploratory analysis of the features employed in the classifiers is given. Results from the exploratory classification analysis using a novel SVM approach is provided. The main contribution of this thesis the TAxonomic COmposition Analysis method –TACOA– is presented in section 5.2, together with its classification accuracy. Section 5.3 focuses on comparing the accuracy obtained by TACOA to the SVM-based PhyloPyhtia. For each of the methods presented in this chapter, the corresponding strategy, implementation, and evaluation is given. Furthermore, by using two case study, the influence of horizontally transfered chunks of DNA on the classification accuracy of a composition based classifier is assessed. The last section of this chapter, describes some results obtained as part of a collaboration made within two metagenomic related projects. One of them investigate the classification of very short genomic fragments (80 - 120bp), and the other one examines how to visualize metagenomic data. Major results obtained within these collaborations are briefly presented.

5.1 Exploratory analysis

5.1.1 Feature exploration

A key part for any classification task are the features describing the items to be classified. For the taxonomic classification of environmental genomic fragments it is desirable that the selected features reflect the taxonomic relatedness of organisms. As described before, sequence similarities have been traditionally used to explore the taxonomic or phylogenetic relationships among organisms. Other features that have also been employed include domain content (Yang *et al.*, 2005), concatenated proteins (Brown *et al.*, 2001; Baldauf *et al.*, 2000), gene content (Fitz-Gibbon and House, 1999; Wolf *et al.*, 2002; Snel *et al.*, 1999), gene order (Dandekar *et al.*, 1998; Korbel *et al.*, 2002; Wolf *et al.*, 2001) and the distribution of structural folds (Gerstein, 1998; Gerstein and Hegyi, 1998; Wolf *et al.*, 1999). These features can be categorized as similarity-based since they depend on prior identification of a functional region (e.g. genes). The use of similarity based features represent a disadvantage for the analysis of metagenomic data because a vast part of the data stems from not yet sequenced or from genomic regions that have not been functionally characterized, thus similarity based features are of limited used for this type of data.

On the other hand, as introduced before oligonucleotide frequencies carry a phylogenetic signal that is species-specific(Karlin *et al.*, 1997; Karlin, 1998). These two important aspects: 1) not need of prior identification of functional regions, and 2) the species-specific signal, make the oligonucleotide frequency a very appealing feature to be used in the taxonomic classification of environmental genomic fragments. In this dissertation, the oligonucleotide frequencies are used in the oligonucleotide deviation score (*ODS*) measure that accounts for their over- and under-representation in a DNA fragment (as

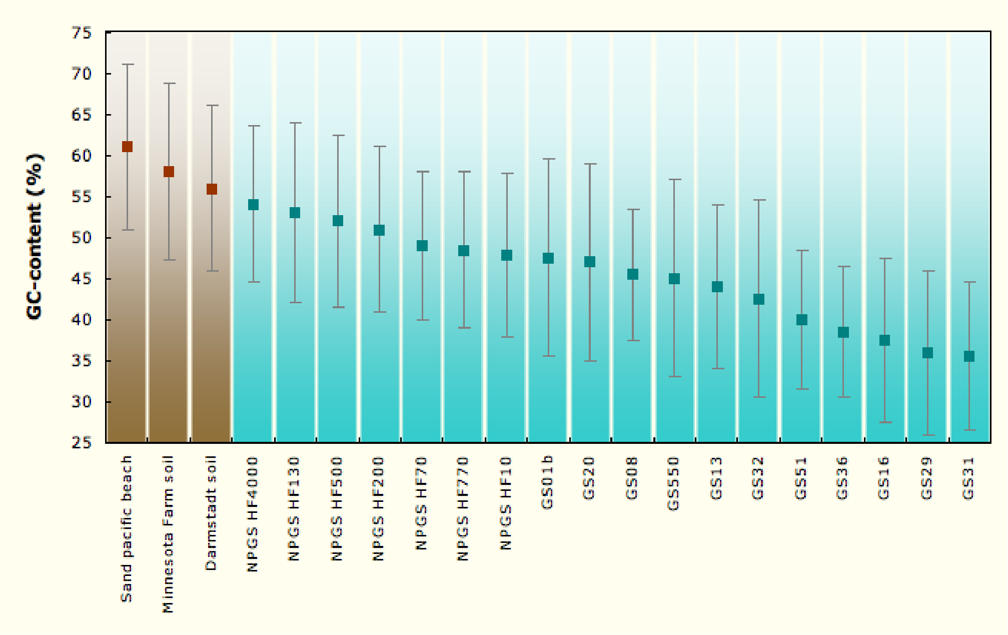


Figure 5.1: Average GC-content of the microbiota from soil and oceanic water samples. GC-content distribution measured for two different type of complex environments: soil and sea water. GC-content distribution was ordered from high to low percentage. NPSG, North Pacific Subtropical Gyre. GS, Global Sampling of oceanic waters. Data collected from the Gold database.

described in chapter 4).

The oligonucleotide frequency patterns found in a genomic fragment are strongly influenced by the genomic GC-content (Noble *et al.*, 1998; Reva and Tümmler, 2004, 2005; Bohlin *et al.*, 2008b). However, it has been proposed that the genomic GC-content is more the result of the environment in which an organism lives, and in many cases correlates poorly to its taxonomic group (Foerstner *et al.*, 2005; Chen and Zhang, 2003). Foerstner *et al.* (2005) showed that samples from soil and ocean surface waters have narrower GC-content distribution than theoretically expected, despite of being very complex communities with more than 1,000 non-abundant species (Venter *et al.*, 2004; Tringe *et al.*, 2005; Rusch *et al.*, 2007).

To explore the distribution of the GC-content from the microbiota sampled in different environments, namely soil and oceanic waters, the mean value was compared in Figure 5.1. In average, the GC-content found in an environmental sample from soil is higher than that obtained for oceanic waters samples (Figure 5.1).

Oligonucleotide length	*r*	
	GC-normalized	*GC-non-normalized*
3	0.438*	0.970*
4	0.540*	0.980*
5	0.545*	0.982*

Table 5.1: Pearson correlation values obtained between the GC-content and the *ODS* feature.

Pearson correlation values obtained between GC-content and the oligonucleotide deviation score *ODS* on different oligonucleotide lengths. *ODS* for oligonucleotide of length 3, 4, and 5 were chosen as representative to test the correlation of the *ODS* feature vs. the GC-content. All values with (*) are highly significant at *p*-level of 0.01.

If the GC-content is similar among organisms co-occurring in an environment, then it is valid to assume that it is not a good discriminatory feature (i.e. noise) for the taxonomic classification problem this thesis deals with. Therefore, a pearson correlation (*r*) was employed to evaluate the correlation between the GC-content of the organisms used in this study and the *ODS* feature used herein. Table 5.1 shows that the strength of the correlation between the genomic GC-content and the oligonucleotide frequency is significantly higher (*p*-level 0.01) if the score is not normalized. The GC-normalization of the *ODS* reduced the noise introduced by the GC-content considerably, however it does not remove it all (Table 5.1). Despite the fact that the correlation between the GC-normalized *ODS* and the GC-content is still significant, this is weaker than the one obtained for the GC-non-normalized *ODS* (Table 5.1).

The 350-genomes data set was used to explore if the *ODS* feature account for the taxonomic relationship among the test organisms. The *ODS* were computed for each of the 350 genomes and their respective scores (*ODS*) were used to build the oligonucleotide feature vectors (GFVs) and hence each organism is described by a GFV. By normalizing each vector to unit length differences in genomic vector lengths are corrected.

Pairwise dot products were used to compute the similarities between pairs of GFVs (all against all comparison) and subsequently stored in a symmetric matrix called the *phylo-matrix* (Figure 5.3). A phylo-matrix graphically summarizes all pairwise similarities between organisms. The degree of similarity between two organisms is visualized using a color code from red (highly similar), over white, to black (very dissimilar). All entries of the phylo-matrix are symmetric with respect to the main diagonal, thus, entries on the upper half are the same as those in the lower half (Figure 5.3). The main diagonal represents the highest similarity possible between two GFV, since each of this entry is the pairwise comparison of a GFV to itself.

In the phylo-matrix all entries are ordered in a nested manner according to their respective taxonomic information. First, GFVs are ordered according to their broadest taxonomic rank of superkingdom, followed by phylum and so on until the deepest taxonomic rank of species. In consequence, entries around the main diagonal always represent closely related species while distant related ones correspond to entries located further away from the main diagonal. It is important to note that the GFVs computation is independent of the matrix re-ordering which can be considered a post-processing step. A phylo-matrix is generated for each oligonucleotide length evaluated in this work and they were employed to explore four different oligonucleotide lengths, namely di-, tri-, tetra- and penta-oligonucleotides.

After reordering the phylo-matrices, the underlying structure reflecting the relatedness among different taxonomic groups became apparent and the different color shades are not randomly arranged (Figure 5.3). Those entries colored with a dark red shade are localized around the main diagonal, indicating the high degree of similarity among the respective GFVs. Moreover, lighter red and grey shades can be seen in entries further away from the main diagonal where the similarity of the GFVs becomes smaller. This behavior is expected since those entries further away from the main diagonal represent distantly related organisms.

In general, a clearer pattern of closely related GFVs was obtained for oligonucleotides of longer length (Figure 5.3). Differences between the GC-normalized and non-normalized *ODS* feature is clearly seen in Figure 5.3. The GC-normalized *ODS* feature better discriminate among taxonomic groups. Although the *ODS* feature helps to unveil some structure in the data, this is not strikingly clear, showing the complexity of the classification problem being addressed in this work. The phylo-matrices help in

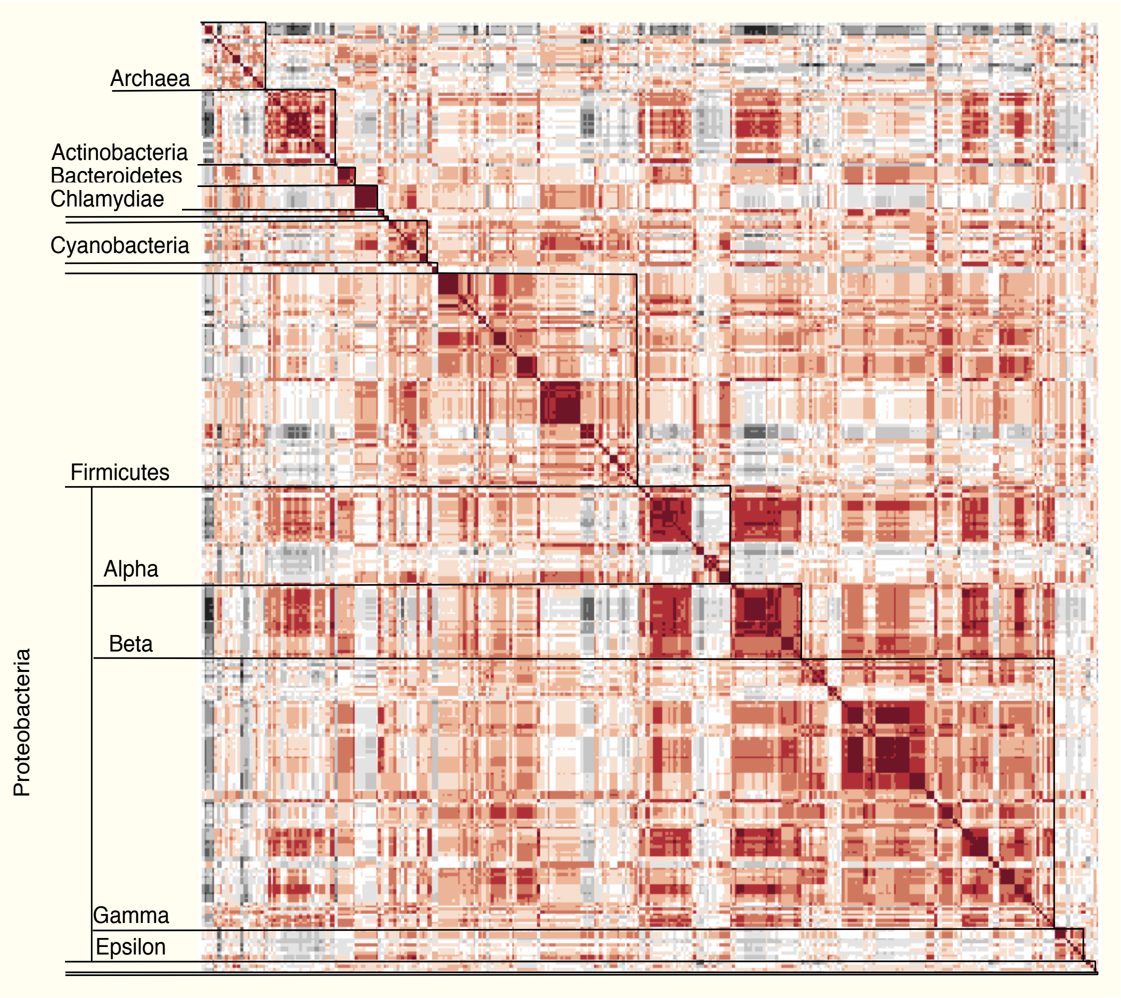


Figure 5.2: Visualization example of a phylo-matrix.
A phylo-matrix is a symmetric matrix whose entries are all pairwise similarities between genomic feature vectors (GFVs). The similarity with respect to the entries of the main diagonal diminishes from top to right and from top to bottom. The highest similarity possible is found in the diagonal since these entries represent the pairwise comparison of a GFVs to itself. The color scheme is defined by dark red shades (high similarity), over white, to black (low similarity). Highly similar GFVs (phylogenetically closer organisms) are clearly spotted with a dark red tone while far related organism have a pale tone and very dissimilar GFVs will have a dark grey tone. Taxonomic groups with highly homogenous GFVs are easily detected (e.g. Chlamydiae) as well as more heterogeneous one (e.g. Proteobacteria). This phylo-matrix example is based on *ODS* computed for oligonucleotides of length 5.

exploring the variability of the *ODS* feature among taxonomic groups and across all taxonomic ranks, therefore giving hints on which taxonomic groups will be more difficult to classify.

An example can be visualized using Figure 5.2, organisms belonging to Chlamydia group do not show high variability in their GFVs. This observation contrast to the GFVs computed for the Cyanobacteria group. Thus, problems in classifying genomic fragments into the latest taxa can be anticipated. Additional to detecting variation of the *ODS* feature inside a taxonomic group, the phylo-matrices are also helpful in detecting outliers inside a taxonomic group. Outliers are easily recognized because their dot product clearly differ in magnitude with respect to its relatives, thus displaying a lighter color shade in a dark shaded area. Large homogeneous and heterogeneous group of GFVs can be detected in entries relative far away from the main diagonal making evident that problems in the classification task might arise for those items. In summary, the phylo-matrices clearly expose that the taxonomic classification problem addressed in this work is far from trivial.

5.1.2 Exploratory classification of fragments of variable length using the oligonucleotide feature score

5.1.3 Strategy

In order to explore the feasibility to classify DNA fragments of variable length using the *ODS* as features, a soft margin SVM was employed. The classification task was performed only at rank order, which represents a good balance between the number of taxonomic classes at that rank and the number of genomes per class. In this work, the novel ”one-against-random” multiclass SVM was developed, which is a modification of the ”one against all” strategy. This newly proposed strategy uses a random selection of items from the complete training set for each binary SVM. It restricts the size of the training set (the ”all-part”) to the size of the test set for the analyzed class (the ”one-part”), by

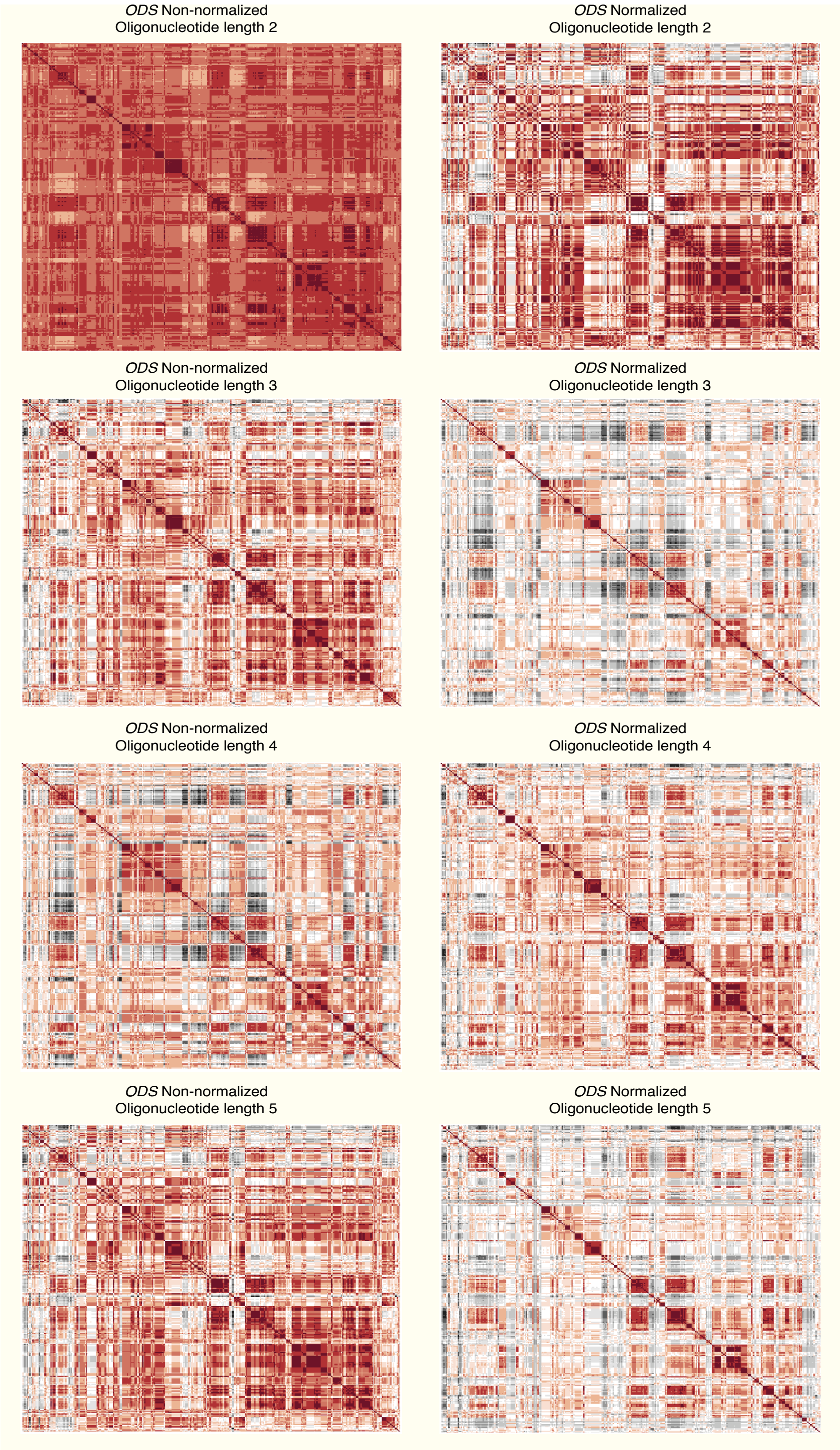


Figure 5.3: Phylo-matrices obtained using four different oligonucleotide lengths. Phylo-matrices (from top to bottom) depict GFVs computed for the *ODS* feature using different oligonucleotide lengths (i.e. 2, 3, 4, and 5). Taxonomically related groups are more evident (dark red tone squares) around the main diagonal. Differences in the phylo-matrices generated using the GC-normalized oligonucleotide deviation scores (right side) than the non-normalized matrices (left side). A clearer structure can be seen for longer oligonucleotide lengths as well as for the normalized *ODS* feature.

picking as many random items for the negative class as there are items for the positive class. As for the ”one-against-all” approach, the prediction function of the ”one-against-random” method returns the class corresponding to the decision function with highest score. If there is not an exact solution, e.g., the sample cannot be classified into a single class or certain threshold is not overcome the item is labeled as *unclassifiable*. An advantage of the ”one-against-random” strategy is that builds only as many binary classifiers as there are existing classes, while the ”one-against-one” needs to build $k(k-1)/2$, where k is the number of existing classes in the entire data set. This is of particular use when the number of classes is large, as is the taxonomic classification problem addressed in this work. In which the number of taxonomic classes explode (up to 60 classes) at deeper ranks (class, order, species)

5.1.4 Implementation

The SVM used in the exploratory analysis was implemented in JAVA using an object orientated approach. The classification methods are based on a JAVA version of the libSVM (Chang and Lin, 2001). This version is provided with the ”one-agaisnt-one” multiclass SVM implementation, which was extended to the ”one-agaisnt-random” approach used herein. In addition, the BioJava library (Holland *et al.*, 2008) library was employed to compute the features on the DNA fragments. The framework is divided into classes which are briefly explained in the following:

SVMClassifier is the main class containing methods for parameter selection using an iterative grid search, training, and classification of new items.

Features contains object oriented methods to compute features from sequences stored in the BioJava object SequenceDB. The resulting features are store in vectors that can be exported as `svm_problem` rows.

SVMData build an `svm_problem` from the given `svm_problem` rows. Two different modes are provided: (i) a training mode with labels and (ii) a testing mode that

Data set	*Strategy*		
	One-Against-Random	*One-Against-All*	*One-Against-One*
Iris (C;γ)	95.34% (0.84;4.76)	96.88% (1.76;2.94)	97.33% (32;0.42)
Wine (C;γ)	98.31% (9.51;4.76)	98.91% (12.71;2.94)	98.88% (0.84;4.76)

Table 5.2: Accuracy evaluation of the One-Against-Random strategy. The number of correctly classified items is expressed in percentage using two benchmark data sets (Iris and Wine). Average classification rates obtained using a ten-fold cross-validation, the corresponding cost (C) and Gaussian kernel width (γ) values are given in parenthesis.

obviate class labels.

IOTools provides input, output, and parsing methods for the main objects and file formats.

libSVMmod is a modified SVM library from the libSVM containing the new ”one-against-random” strategy.

5.1.5 Evaluation

To evaluate the classification accuracy of the proposed ”one-against-random” strategy, the new strategy was compared to the traditional ”one-against-all” and the most comprehensive ”one-agaisnt-one” using two well known benchmark data sets: iris and wine data sets (Asuncion and Newman, 2007). The ”one-against-random” strategy achieves comparable results when contrasted to the ”one-against-all” and to the ”one-against-one” approach for the iris and wine data set using a ten-fold cross-validation (Table 5.2).

The comparison between the less complex ”one-against-random” and the ”one-against-one” as well as the traditional ”one-against-all” showed that the new proposed strategy achieves comparable classification accuracies (Table 5.2), while being more time efficient particularly when a high number of classes exist as shown in Table 5.3 for the 18-orders data set. The newly devised multiclass SVM strategy ”one-against-random” was employed to perform the exploratory classification analysis for the 6-, 12-, 18- and 29-orders dat sets.

Strategy	*ODS length*	*Runtime (secs)*		
		6-Orders	*12-Orders*	*18-Orders*
One-Against-Random	2	550	1540	1661
One-Against-Random	3	856	2173	3022
One-Against-Random	4	3030	5737	9548
One-Against-One	2	573	3775	2167
One-Against-One	3	870	5718	5343
One-Against-One	4	2921	9676	15002

Table 5.3: Average runtimes for three different data set sizes employing two strategies. The *ODS* length refers to the length of the oligonucleotide chosen to calculate the oligonucleotide deviation score. The runtime is measures in seconds (secs).

At rank order, all taxonomic classes having at least 3 different genomes were selected and included in the data set to be used for evaluation. An SVM classifier was employed to classify DNA fragments included in the data set for evaluation. To assess the accuracy of the SVM classifier, 200 DNA fragments were extracted from each completely sequenced genome present in the analyzed data set. A ten-fold cross validation strategy was used to estimate the classification accuracy. For this purpose, the complete data set containing genomic fragments of all selected genomes was randomly partitioned into 10 subsets of equal size. Fragments of lengths 100bp and 200bp were used to simulate reads produced by the 454 sequencing technology. In addition, DNA fragment lengths of 800bp and 1Kbp were chosen to represent reads obtained by the Sanger (dye-terminator) technique. Moreover, contigs were simulated by DNA fragments of length 3, 5, and 15Kbp. The DNA fragments used in this exploratory analysis were generated by extracting, from each completely sequenced genome, subsequences of a fixed length at random positions. This procedure simulates the randomly fragmentation step prior the sequencing of a metagenome (Figure 2.1 step 2)

Parameter optimization

The grid search method (Staelin, 2003) was employed to optimize parameters of the SVM classifier, in which a combination of parameters are simultaneously optimized. The parameter C (misclassification cost) and γ (width of the Gaussian) required for the soft margin SVM classifier were optimized prior the training step. For this purpose, the algorithm developed by Staelin (2003) was employed, which first uses a coarse grid and later the search is refined based on the ”best” grid region found. This procedure is performed over several iterations. After each one, the performance of every parameter combination is measured and the search space is centered around the best point. This process is repeated until a previously defined number of iterations is reached and subsequently those parameters with best performance are chosen (Staelin, 2003).

5.1.6 Classification accuracy obtained using an SVM classifier for genomic fragments of variable length

A support vector machine was employed to evaluate the feasibility to taxonomically classify genomic fragments based on the *ODS* feature selected during the feature exploration as described in section 5.1.1. This exploratory experiment was performed at rank order which is one of the deepest rank of the taxonomic hierarchy. A total number of 29 different classes constitute the rank order representing a challenging multiclass classification task. To investigate the influence of the number of taxonomic classes consider in the multiclass classification problem, the entire 350-genomes dat set was partition in four different smaller subsets. Namely, -6,-12, -18, and 29-orders data set. Three different oligonucleotide length were used to compute the *ODS* feature, namely di-, tri-, and tetra-nucleotide. Simulated reads of length 100, 200 and 800 bp were used to explore the feasibility to taxonomically classify short and very short genomic fragments. On the other hand, genomic fragments of length 1, 3, 5, and 15 Kbp were used to examine the classification accuracy of simulated contigs.

For all genomic fragment length analyzed, the *ODS* feature based on oligonucleotides of length 4 achieved a better accuracy in terms of specificity, sensitivity and false negative rate compared to *ODS* of length 2 (Figure 5.4). A general observed trend was the reduction in accuracy (sensitivity and specificity) as the number of consider classes increased (Figure 5.4). Conversely, the opposite trend is detected for the false negative

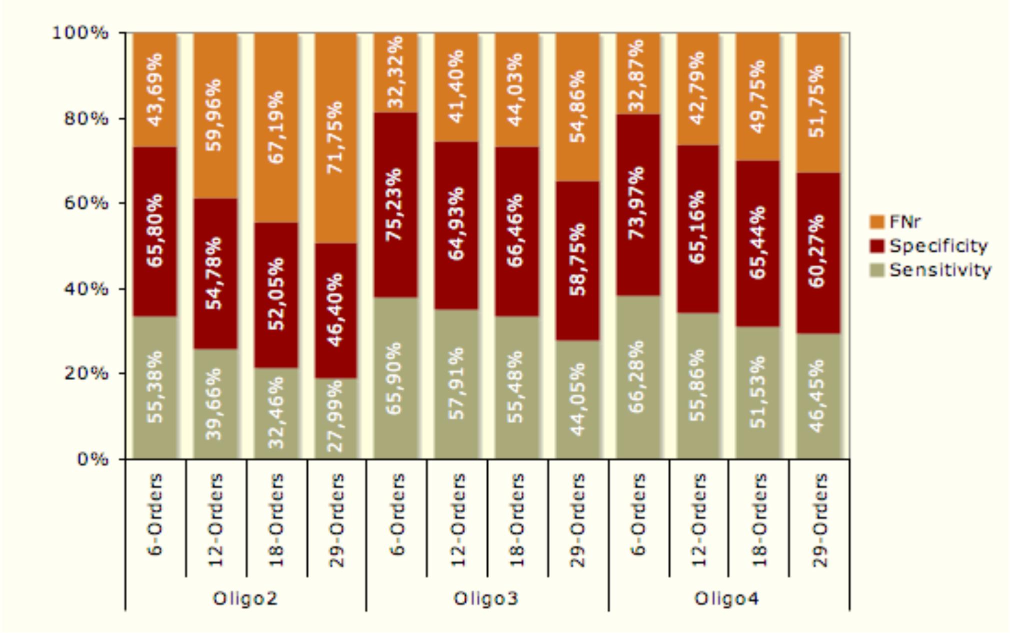


Figure 5.4: Overall classification accuracy of the SVM classifier for three different oligonucleotide lengths.

Accuracy was measured by sensitivity, specificity and false negative rates (FNr) using four different data sets at rank order. The numbers 6-, 12-, 18-, and 29 refers to the number of existing classes in the analyzed rank. *ODS* were computed using oligonucleotides of length 2, 3, and 4.

rate (increments with the number of analyzed classes), revealing that the classification problem turns more complex as the number of classes increases (Figure 5.4).

When comparing different data sets in terms of the number of taxonomic classes, fluctuations as large as 20% in sensitivity and specificity are detected for *ODS* corresponding to oligonucleotides of length 2. Moreover, *ODS* based on oligonucleotide of length 2 showed poor taxonomic resolution even for longer DNA fragments (5 Kbp and 15 Kbp) analyzed (Figure 5.5). Furthermore, *ODS* of length 2 achieved the highest number of misclassifications (up to 90% as seen by the FNr Figure 5.5) for all genomic fragment lengths evaluated. This trend is more clear for those genomic fragments of larger size. These observations suggest that *ODS* based on oligonucleotide of length 2 is not a good discriminatory feature to taxonomically classify genomic fragments. For all *ODS* and DNA fragment lengths evaluated, the FNr increased when new classes are included in the data set. Minimal FNr values are registered for the 6-orders data set while maximal were obtained for the data set with 29-orders (Figure 5.5).

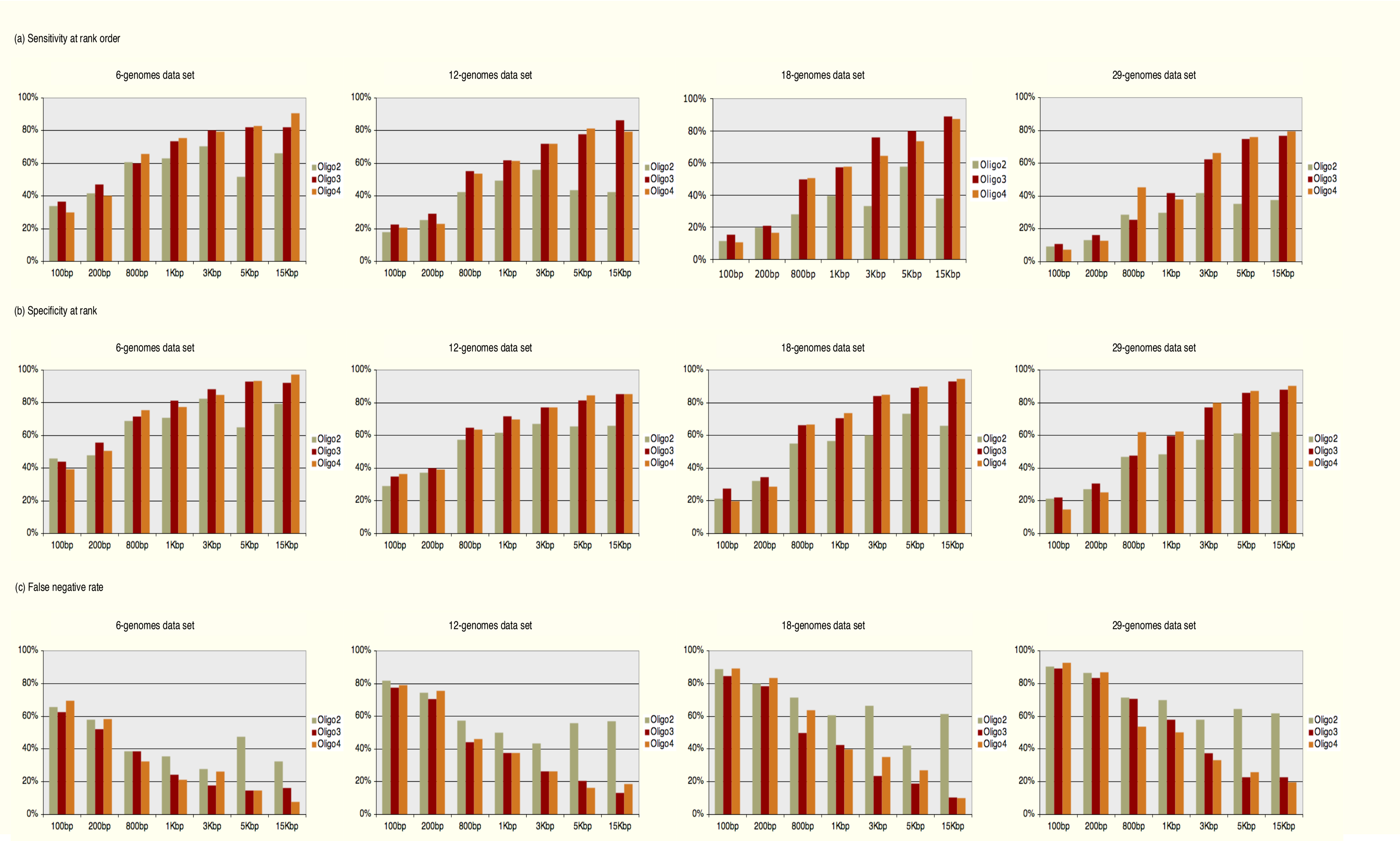


Figure 5.5: Overall classification accuracy for three different oligonucleotide lengths.
Accuracy was measured by sensitivity, specificity and false negative rates (FNr) using four different data sets at rank order. The numbers 6-, 12-, 18-, and 29- refers to the number of existing classes at rank order. *ODS* were computed using oligonucleotides of length 2, 3, and 4. Results are shown for simulated reads (100 - 800 bp) and contigs (1 - 15 kbp).

On the other hand, *ODS* of length 4 showed poor values of sensitivity and specificity as well as high false negative rates for the two shortest DNA fragments evaluated (100 bp and 200 bp). The accuracy obtained using *ODS* of length 3 and 4 are comparable for the four data sets tested. However, a small decrease in sensitivity is observed for the 29-genomes data set for all oligonucleotide length evaluated when compared to the 18-genomes data set. As a whole, the high number of misclassifications suggests that this complete strategy needs further evaluation, exploration and fine tuning of parameters to achieve better results. However, the accuracy results in terms of sensitivity and specificity also indicate that the complete strategy can be improved.

This exploratory classification experiment showed that the number of misclassifications rises dramatically when the number of taxonomic classes are increased (Figure5.5). This observation is valid for all *ODS* length evaluated, although to a lesser extent for *ODS* of length 3 and 4 (Figure 5.4). Figure 5.4 shows how the overall accuracy (sensitivity, specificity and FNr) deteriorates as the number of taxonomic classes increases. This observation confirms the intuitive idea that a classification task gets more complex as the number of taxonomic classes increases.

As an overall trend, the average sensitivity and specificity increased as longer genomic fragments were considered. A general observation is that the number of misclassifications decreases as the length of the genomic fragment increases (Figure 5.5). The overall accuracy of *ODS* of length 4 is comparable for longer genomic fragments.

To compared the classification results of the ”one-against-random” with the ”one-agaisnt-one” strategy using metagenomic simulated data, an SVM for each strategy was trained to analyzed the 18-orders data set. This data set was selected because represents a good balance between the number of classes and number of members per class. As with the benchmark data set, the obtained accuracies (sensitivity, specificity and false negative rates) are comparable (Figure 5.6). Moreover, the run time of the ”one-against-random” strategy is smaller (9,548 sec) in one order of magnitude compared to the ”one-agaisnt-one” (15,002 sec) for the 18-order data set. This example highlights a key advantage of the ”one-against-random” strategy.

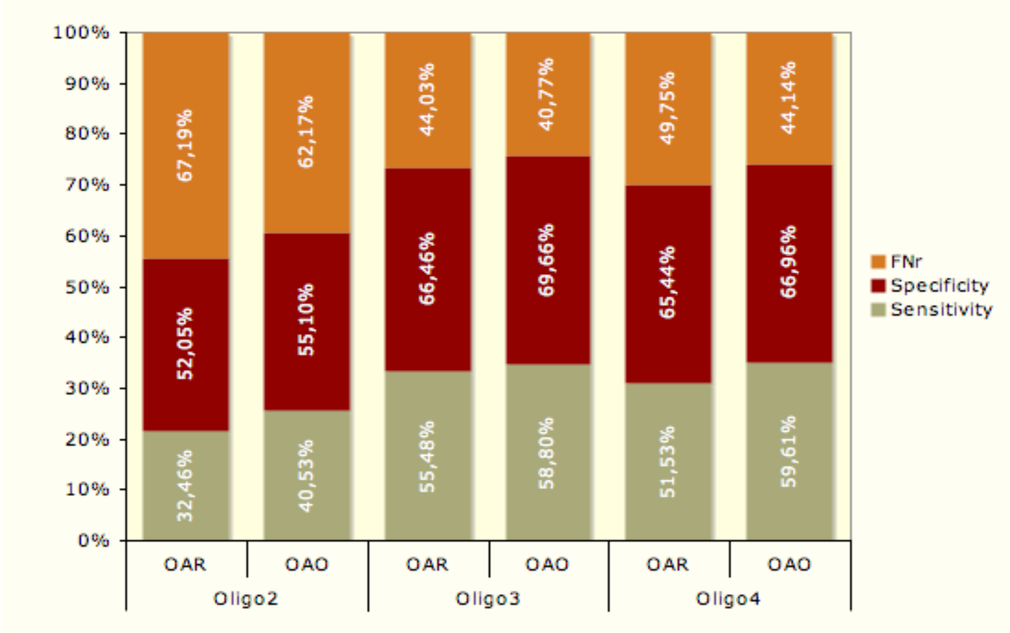


Figure 5.6: Overall classification accuracy for three different oligonucleotide lengths using two different strategies.

Average values of sensitivity, specificity and false negative rates (FNr) obtained using the ”One-Against-Random” and the ”One-Agaisnt-One” strategy. The classification accuracy was measured for three different ‘oligonucleotide length using the 18-orders data set.

Despite the encouraging classification results obtained with the SVM based classifier, this first approximation to the problem showed the complexity of the classification task this work deals with. This observation can also be found in the work of McHardy *et al.* (2007), the authors opted for the expensive ”one-against-one” but accurate strategy and combined it with several additional SVMs using the ”one-against-all” strategy to confirm or reject the predictions made in the classification phase. As result the developed classifier is a hierarchical collection of SVMs, incrementing the complexity of the developed classifier. Some pitfalls that were detected in this exploratory analysis are: need of costly retraining procedures, exploration of more intricate features to achieve better performance, complexity of the classifier as the number of classes grows. The issues revealed in this exploratory analysis motivated the search for a more straightforward solution without sacrificing classification power and that could still yield competitive results.

5.2 TACOA – A novel classification approach of environmental genomic fragments

5.2.1 Strategy

In this study, a genomic fragment is defined as a DNA sequence of a given length (note, that a completely sequenced genome can be regarded as a genomic fragment). The total number of oligonucleotides of length l, from the alphabet $\sum = \{a,t,c,g\}$ is given by 4^l, where 4 represents all possible nucleotides. Each genomic fragment is represented as a vector (i.e. GFV) using the Vector Space Model (Salton *et al.*, 1975) as described in chapter 4. To predict the taxonomic origin of a query GFV, TACOA compares that query GFV to the reference GFVs. The reference GFVs are computed from all 373 completely sequenced reference genomes used in this study. In the following, the set of all reference GFVs is named reference set $\mathbf{ref}_{set}$. In this study the 373-genomes was used as reference set.

More formally, let $\mathbf{ref}_{set} = \{\mathbf{x}_j\}$ with $1 \leq j \leq T$ be the set of reference GFVs, where each $\mathbf{x}_j$ represents a GFV computed from a completely sequenced reference genome. Let $\mathbf{x}$ be a query GFV representing a genomic fragment to classify. The multi-class classification problem addressed herein, resides in deciding to which of all different taxonomic classes, at rank r, $\mathbf{x}$ belongs to.

For each taxonomic rank r out of superkingdom, phylum, class, order and genus and for each taxonomic class i at that rank, the algorithm computes a discriminant function $\delta_i(\mathbf{x})$, and then classifies $\mathbf{x}$ into that class with the highest value for its discriminant function. More precisely, for a given taxonomic rank r, let i be that class with the highest discriminant function $\delta_i(\mathbf{x})$. Then, $\mathbf{x}$ is classified into class i if $\delta_i(\mathbf{x})$ is at least half as large as the value of the second highest discriminant function on rank r, otherwise $\mathbf{x}$ is classified as ”unclassified”. This optimal cut-off value for the discrimination function at each taxonomic rank r was identified in a grid search. The discriminant function for a taxonomic class i is computed by:

$$\delta_i(\mathbf{x}) = \sum_{\mathbf{x}_j \in \mathbf{ref}_i} K_\lambda(\mathbf{x}, \mathbf{x}_j) \tag{5.1}$$

where $\mathbf{ref}_i = \{\mathbf{x}_j | \mathbf{x}_j \in \mathbf{ref}_{set}$ and $\mathbf{x}_j$ stems from class $i\}$ is the set of all reference GFVs from class i. The smoother kernel $K_\lambda(\mathbf{x}, \mathbf{x}_j)$ is based on the Gaussian density function that exponentially decreases with Euclidian distance from $\mathbf{x}$:

$$K_\lambda(\mathbf{x}, \mathbf{x}_j) = e^{\left(-\frac{d_w(\mathbf{x}, \mathbf{x}_j)^2}{2\lambda}\right)} \tag{5.2}$$

where $d_w(\mathbf{x}, \mathbf{x}_j)$ is a weighted distance function as defined later in Equation (5.4) and λ controls the neighborhood width around $\mathbf{x}$ in the kernel function. Small values of λ result in decision boundaries with higher variance that well-fit the reference set while large values achieve smooth and stable decision boundaries that avoid overfitting and are more robust (Hastie *et al.*, 2002).

To estimate how much a query GFV $\mathbf{x}$ differs from a reference GFV the distance between the two vectors is determined. By normalizing each vector to unit length differences in genomic vector lengths are corrected. The similarity between a query GFV $\mathbf{x}$ and each reference GFV $\mathbf{x}_j$ is computed using the dot-product between the normalized query (GFV $\hat{\mathbf{x}}$) and the normalized reference (GFV $\hat{\mathbf{x}}_j$). The term similarity is expressed as the dot product $< \hat{\mathbf{x}}, \hat{\mathbf{x}}_j >$. The similarity can be easily transformed into a distance by subtracting $1 - similarity$. The distance is then expressed by:

$$d(\mathbf{x}, \mathbf{x}_j) = 1 - < \hat{\mathbf{x}}, \hat{\mathbf{x}}_j > \tag{5.3}$$

The distance d was weighted in order to account for the imbalanced reference set used in this study, where majority classes and minority classes are present, e.g. the phylum bacteria was over-represented compared to the archaea in a proportion of 10:1.

The weighted distance function is denoted as d_w and the weights are assigned using the following weighting scheme. Let $\mathbf{x}_j$ originate from class i and let n_i be the number of genomes in class i. Furthermore, let T be the number of genomes constituting the reference set. The weighted distance function d_w is given by:

$$d_w(\mathbf{x},\mathbf{x}_j) = \frac{T}{n_i} d(\mathbf{x},\mathbf{x}_j) \tag{5.4}$$

This weighting scheme assigns small weights to the GFVs belonging to the majority classes and a relative larger weight for GFVs member of the minority classes.

5.2.2 Implementation

The TACOA classifier was implemented in PERL in object orientated manner. The classifier program is composed of separate objects:

DotProduct.pm Module dedicated to the computation of the dot product between the oligo vectors.

KernelNN.pm In this module the kernelized version of the k-NN method is implemented.

OligoVectors.pm This modules computes vectors from a DNA sequence which entries are oligonucleotide frequencies of a given length.

DNAutils.pm This module contains various functions and methods frequently used to process DNA sequences (e.g. subset of genetic codes, translation of DNA sequences).

Configs.pm In this module all system parameters needed to be adapted to local systems are found as well as the creation of required directories for storage of the outputted and inputted data. This modules also executes checks to assure that required software and adequate version is installed.

Common.pm Module dedicated to the process of DNA sequence to be used in the generation of the genomic feature vectors.

This modularity brings flexibility to the program allowing modification of single components without changing the remaining system. The main PERL script is executed via a bash wrapper script which sets up the environment for TACOA classifier. In particular, automatically sets up the paths to locate the libraries, default reference genomes and other components needed. This ensures that the installation is extremely simple with minimal user intervention and it can immediately be executed via the wrapper script.

5.2.3 Evaluation

The classification accuracy of the presented method was assessed using the leave-one-out cross-validation strategy. In the leave-one-out cross validation, one genome is used to generate fragments of a fixed length and thereafter the taxonomic origin of each fragment was predicted using the remaining 372 genomes and used as the reference set (Figure 5.7). This procedure was repeated for each genome out of the 373 completely sequenced genomes present in the data set (Figure 5.7).

This simulates the case when the taxonomic origin of DNA fragments is predicted that stem from genomes that are not yet represented in the public genome databases. In a second experiment, the classification accuracy of the method with the test set included in the reference set was evaluated. In this case the fragments of each genome were taxonomically classified using all 373 genomes as reference.

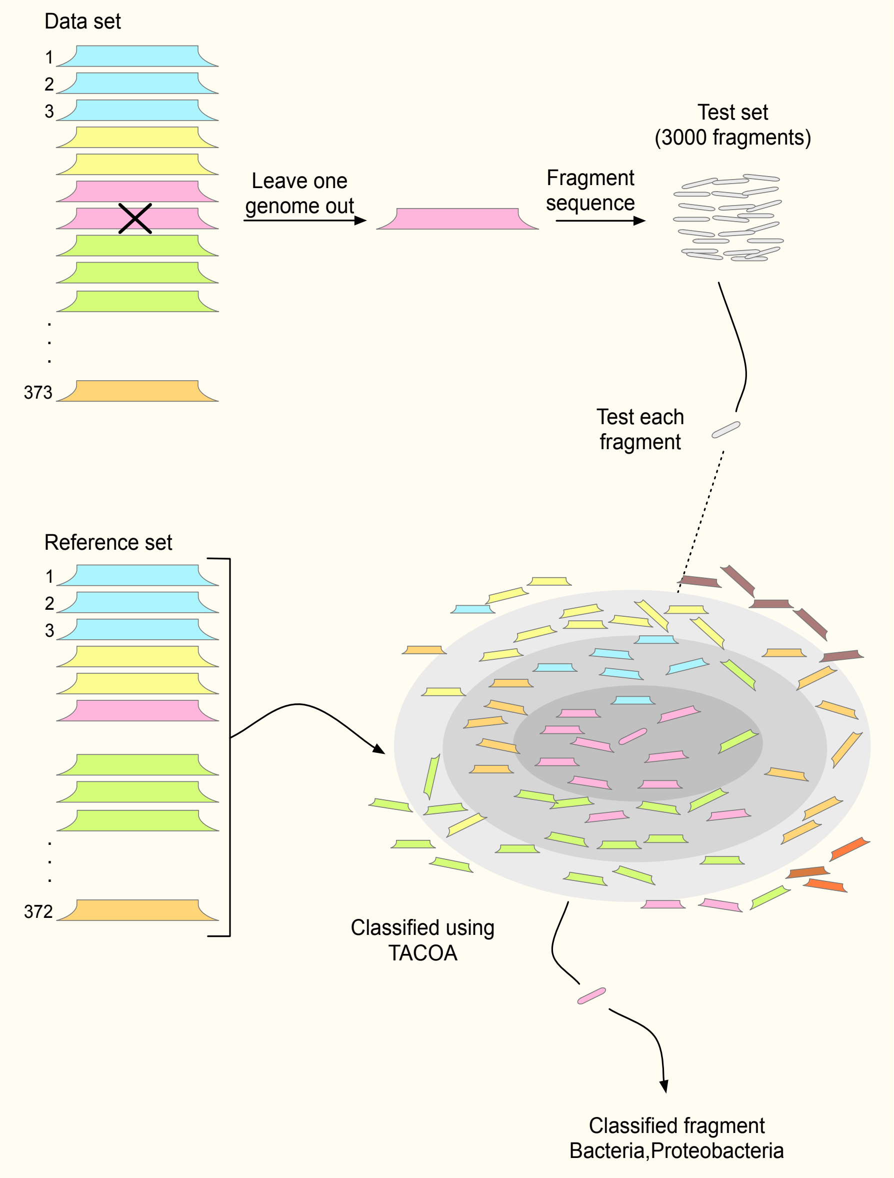


Figure 5.7: Sketch of the leave-one-out cross validation (LOOCV) strategy adopted in this study is depicted.

A genome is selected from the data set comprising 373 genomes and fragmented subsequently. The collection of genomic fragments is regarded as the test set from which each fragment is drawn and classified afterward. Classification of each test fragment is carried out using the remaining 372 organisms as a reference.

The accuracy evaluation carried out in this study requires the existence of at least two different genomes per taxonomic class. This criteria responds to the need of having a genome as reference and another one for testing. Thus, one genome is used to generate fragments of fixed length and thereafter predict the taxonomic origin of each fragment while the second is part of the 372 genomes and used as reference set. At the same time, the classification accuracy for fragments originating from genomes with only one representative per taxonomic class was also evaluated, namely a reference to those genomes do not exist in the reference set. The purpose of the latter evaluation was to assess the classification accuracy of TACOA in the situation when the taxonomic origin of a genomic fragment stemming from a taxonomic group that has not yet been sequenced needs to be predicted. Conversely, also already sequenced genomes may be present in real metagenomic data sets. Thus, in a second experiment the classification accuracy of the method having the test set included in the reference set was also evaluated.

For both experiments, different genomic fragment lengths to simulate DNA fragments obtained in real metagenomic sequencing projects were selected. Genomic fragments of length 800bp and 1Kbp were chosen to resemble single reads derived by the Sanger technology. Assembled contigs were simulated choosing fragment lengths of 3, 10, 15, and 50Kbp. Genomic fragment generation was executed in the following manner: For each completely sequenced genome and for each chosen genomic fragment length, 3000 non-overlapping fragments were randomly retrieved from the selected genome and subsequently included into the test set.

Parameter optimization

An extensive investigation of the oligonucleotide length parameter choosing different values of l ($2 \leq l \leq 6$) and detected the length with maximal classification accuracy. For short fragment lengths only small values of l were considered to guarantee that all possible oligonucleotides occur sufficient times, i.e. $4^l < |s|$ in a considered genomic fragment s as mentioned in Section 3.4. The optimal oligonucleotide length l was determined for each genomic fragment length at each taxonomic rank.

Oligonucleotides of length 4 achieved the highest classification rates for genomic fragments of length 800bp, 1Kbp and 3Kbp. For genomic fragments of length 10, 15 and

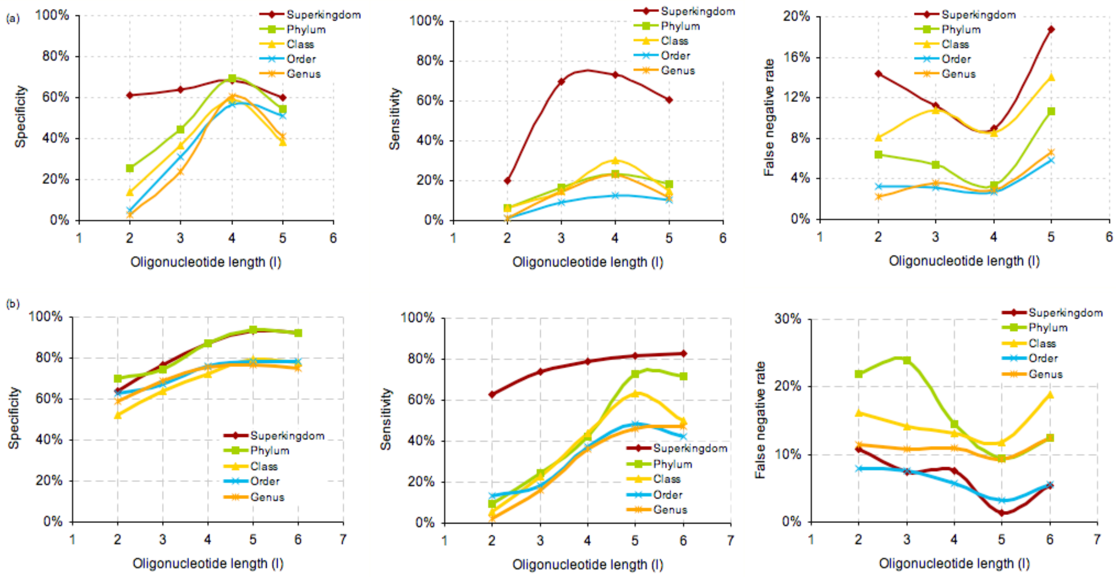


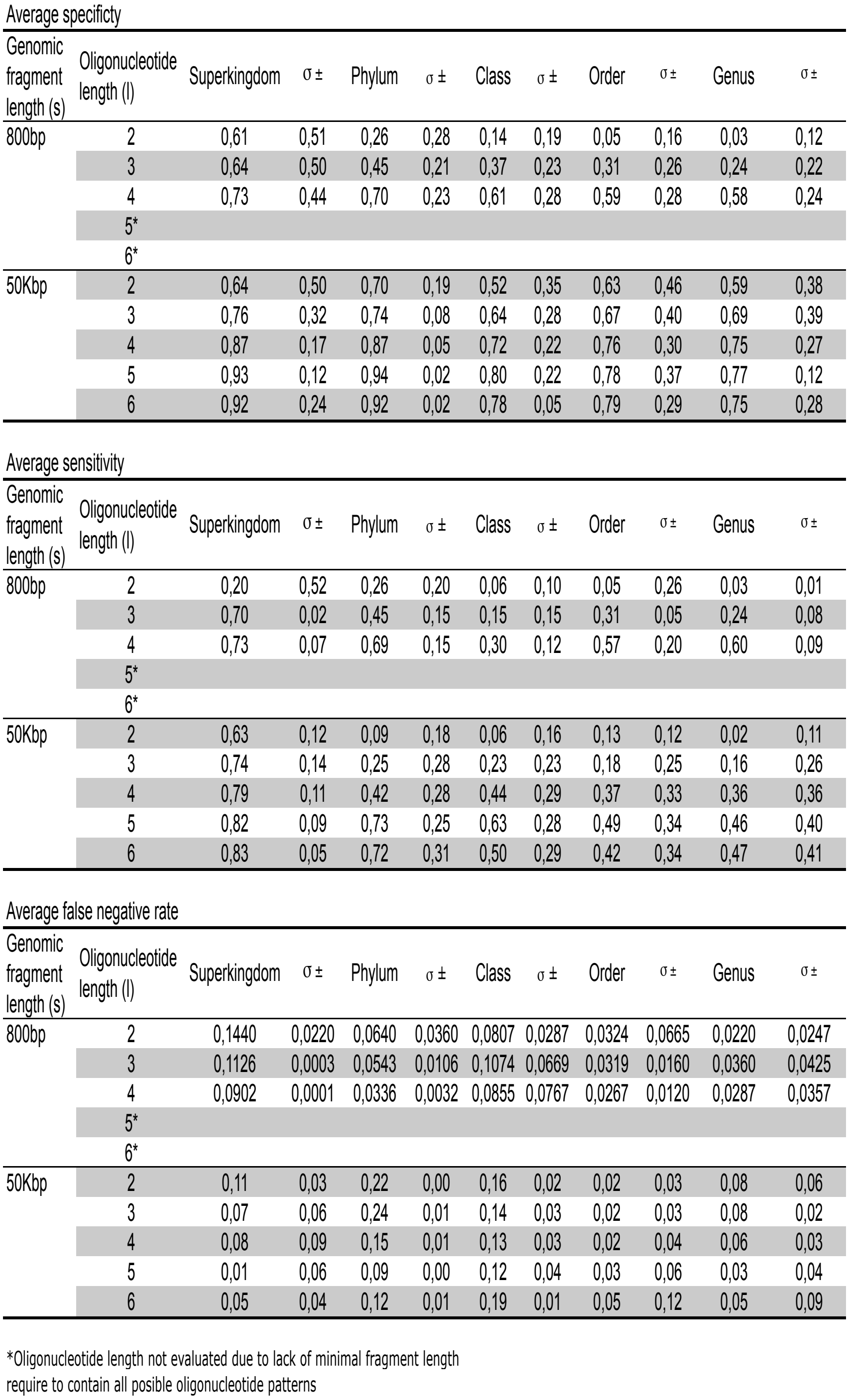
Figure 5.8: Oligonucleotide length-dependent performance for two different genomic fragment length.

Achieved specificity (left), sensitivity (middle) and false negative rate (right) for different oligonucleotide lengths in genomic fragments of length 800bp (a) and 50Kbp (b). For clarity the standard deviation was not depicted in these figures, instead is given in Figure 5.9.

50Kbp, oligonucleotides of length 5 were best suited for classification. A general trend for all genomic fragment lengths was that both average specificity and average sensitivity dropped when oligonucleotides longer than 5 were analyzed. In Figure 5.8 the oligonucleotide length-dependent trend is exemplified with sequence of length 800bp and 50Kbp. Conversely, the false negative rate increased when longer oligonucleotide lengths were considered (Figure 5.8). A detailed table summarizing average accuracy values and standard deviations for the two different fragment length (800bp and 50Kbp) and for each oligonucleotide length analyzed is given in Figure 5.9.

The kernel parameter λ governs the width of the local neighborhood, thus influencing the local behavior of the decision boundary allowing the search of an optimal trade-off between a well-fitted and a more generalized classifier.

A grid search ($2 \leq \lambda \leq 1000$) was employed to detect optimal values of λ achieving maximal accuracy ($\lambda_{\mathbf{opt}}$). In general, $\lambda_{\mathbf{opt}}$ is smaller at lower taxonomic ranks (Table 5.4). This observation may be explained by the drastically increased number of taxonomic



Average specificty

Genomic fragment length (s)	Oligonucleotide length (l)	Superkingdom	σ ±	Phylum	σ ±	Class	σ ±	Order	σ ±	Genus	σ ±
800bp	2	0,61	0,51	0,26	0,28	0,14	0,19	0,05	0,16	0,03	0,12
	3	0,64	0,50	0,45	0,21	0,37	0,23	0,31	0,26	0,24	0,22
	4	0,73	0,44	0,70	0,23	0,61	0,28	0,59	0,28	0,58	0,24
	5*										
	6*										
50Kbp	2	0,64	0,50	0,70	0,19	0,52	0,35	0,63	0,46	0,59	0,38
	3	0,76	0,32	0,74	0,08	0,64	0,28	0,67	0,40	0,69	0,39
	4	0,87	0,17	0,87	0,05	0,72	0,22	0,76	0,30	0,75	0,27
	5	0,93	0,12	0,94	0,02	0,80	0,22	0,78	0,37	0,77	0,12
	6	0,92	0,24	0,92	0,02	0,78	0,05	0,79	0,29	0,75	0,28

Average sensitivity

Genomic fragment length (s)	Oligonucleotide length (l)	Superkingdom	σ ±	Phylum	σ ±	Class	σ ±	Order	σ ±	Genus	σ ±
800bp	2	0,20	0,52	0,26	0,20	0,06	0,10	0,05	0,26	0,03	0,01
	3	0,70	0,02	0,45	0,15	0,15	0,15	0,31	0,05	0,24	0,08
	4	0,73	0,07	0,69	0,15	0,30	0,12	0,57	0,20	0,60	0,09
	5*										
	6*										
50Kbp	2	0,63	0,12	0,09	0,18	0,06	0,16	0,13	0,12	0,02	0,11
	3	0,74	0,14	0,25	0,28	0,23	0,23	0,18	0,25	0,16	0,26
	4	0,79	0,11	0,42	0,28	0,44	0,29	0,37	0,33	0,36	0,36
	5	0,82	0,09	0,73	0,25	0,63	0,28	0,49	0,34	0,46	0,40
	6	0,83	0,05	0,72	0,31	0,50	0,29	0,42	0,34	0,47	0,41

Average false negative rate

Genomic fragment length (s)	Oligonucleotide length (l)	Superkingdom	σ ±	Phylum	σ ±	Class	σ ±	Order	σ ±	Genus	σ ±
800bp	2	0,1440	0,0220	0,0640	0,0360	0,0807	0,0287	0,0324	0,0665	0,0220	0,0247
	3	0,1126	0,0003	0,0543	0,0106	0,1074	0,0669	0,0319	0,0160	0,0360	0,0425
	4	0,0902	0,0001	0,0336	0,0032	0,0855	0,0767	0,0267	0,0120	0,0287	0,0357
	5*										
	6*										
50Kbp	2	0,11	0,03	0,22	0,00	0,16	0,02	0,02	0,03	0,08	0,06
	3	0,07	0,06	0,24	0,01	0,14	0,03	0,02	0,03	0,08	0,02
	4	0,08	0,09	0,15	0,01	0,13	0,03	0,02	0,04	0,06	0,03
	5	0,01	0,06	0,09	0,00	0,12	0,04	0,03	0,06	0,03	0,04
	6	0,05	0,04	0,12	0,01	0,19	0,01	0,05	0,12	0,05	0,09

*Oligonucleotide length not evaluated due to lack of minimal fragment length
require to contain all posible oligonucleotide patterns

Figure 5.9: Standard deviation for average accuracy and false negative rate obtained for fragments of length 800bp and 50Kbp.

Standard deviation and average specificity, sensitivity and false negative rate is given for all oligonucleotide length and taxonomic ranks evaluated.

	$\lambda_{\mathbf{opt}}$				
Fragment length	**S**	**P**	**C**	**O**	**G**
800bp	500	300	100	25	100
1Kbp	500	300	200	100	100
3Kbp	500	300	300	500	400
10Kbp	300	400	300	100	90
15Kbp	400	300	500	200	100
50Kbp	500	1000	400	500	80

Table 5.4: Optimized λ parameter obtained for each genomic fragment length at each taxonomic rank.
Optimal lambda parameter ($\lambda_{\mathbf{opt}}$) is shown for each genomic fragment length at each taxonomic rank: Superkingdom (S), Phylum (P), Class (C), Order (O), and Genus (G).

classes at deeper ranks. If a large number of taxonomic classes occur at deeper ranks the neighborhood to be considered in the classification task needs to be smaller (small λ) than in broader taxonomic ranks. If a large λ is considered and a large number of classes exists, the respective neighborhood of a query genomic vector may cover too many reference vectors from diverse taxonomic classes having a negative impact on the classification accuracy. On the other hand, if the reference GFVs from a taxonomic class are sparsely distributed with respect to the query GFVs, a bigger neighborhood (large λ) needs to be considered. This may explain those cases where a large $\lambda_{\mathbf{opt}}$ is obtained.

During the optimization procedure, optimal parameters were chosen based on average accuracy values over all taxa at each taxonomic rank, therefore it may occur that the optimal parameters chosen are suboptimal for some taxonomic classes at a given rank. In consequence, the accuracy for some taxonomic classes can drop dramatically, this situation can be seen as ”gaps” in Figure 5.10.

From a practical perspective, in this work was regarded as more valuable to be able produce a low number of highly reliable predictions than a large number of predictions with low reliability. Therefore in this study parameters producing high specificity values over high sensitivity were favored.

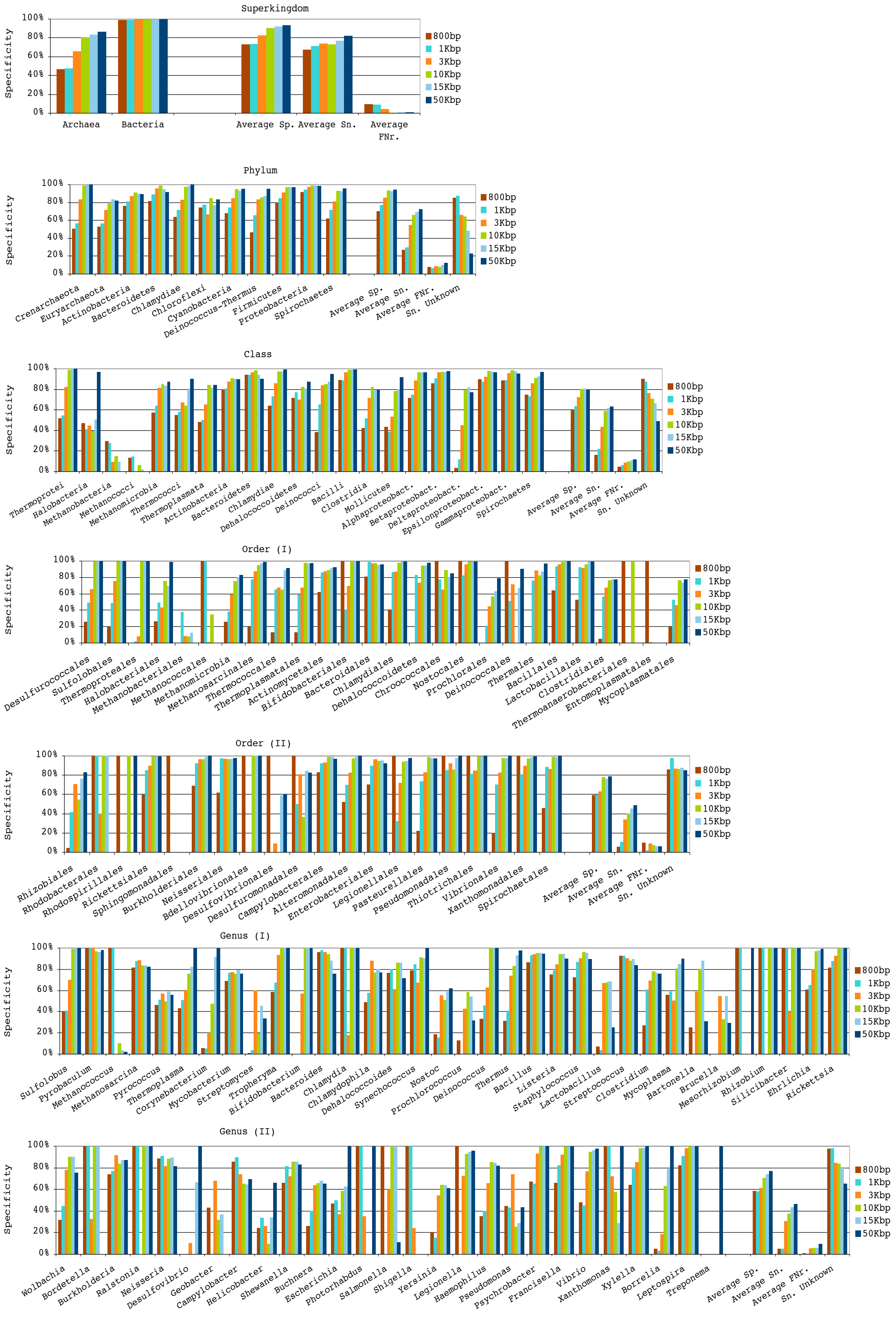


Figure 5.10: Classification accuracy achieved for genomic fragments of different length.

Bars depict detailed specificity and average values for specificity (Sp.), sensitivity (Sn.) and false negative rate (FNr.) for each fragment length on different taxonomic ranks. Each color represents a genomic fragment length.

5.2.4 Classification accuracy obtained by TACOA for genomic fragments of variable length

The classification accuracy of TACOA was evaluated on genomic fragments of lengths ranging from 800bp to 50kbp. A total of 11,730,382 genomic fragments from 373 different species were analyzed, comprising ≈42 Mb of sequence data. The classification accuracy for all different evaluated genomic fragment lengths, taxonomic ranks, and taxa is given in detail in Figure 5.10.

A high proportion of contigs (genomic fragments of length 3Kbp, 10Kbp, 15Kbp, and 50Kbp) was correctly classified with an average sensitivity between 76% at rank superkingdom and 39% at rank genus (Figure 5.11). At the same time, less than 10% of contigs were misclassified (false negative rate) at all taxonomic ranks. For the remaining contigs the taxonomic origin could not be inferred and hence these were assigned to the ”unclassified” class. Overall, reliable predictions were obtained with an average specificity ranging from 89% at superkingdom to 71% at rank genus. For the longest analyzed contig length (50Kbp), TACOA achieved an average sensitivity of 82% at superkingdom and 46% at genus, and specificity of 93% (superkingdom) and 77% (genus) (Figure 5.10, 5.12). Moreover, also for shorter contigs a high classification accuracy was obtained. For example, 74% of the contigs of length 3Kbp were correctly classified at rank superkingdom and 31% at rank genus (Figure 5.10, 5.12), the specificity for contigs of length 3kbp reached values between 74% (superkingdom) and 31% (genus).

In this evaluation, single reads were represented by genomic fragments of length 800bp-1Kbp. TACOA is capable of accurately predicting the taxonomic origin of single reads up to the rank of class, despite the limited information contained in these short sequences. A high proportion of reads was correctly classified. For reads of length 800bp, the average sensitivity was between 67% at superkingdom and 16% on rank class and for reads of length 1Kbp, it ranged from 71% to 22%. Furthermore, in average only between 9% (superkingdom) and 5% (class) of reads were misclassified. Overall, reliable predictions were obtained, with an average specificity ranging from 73% (superkingdom) to

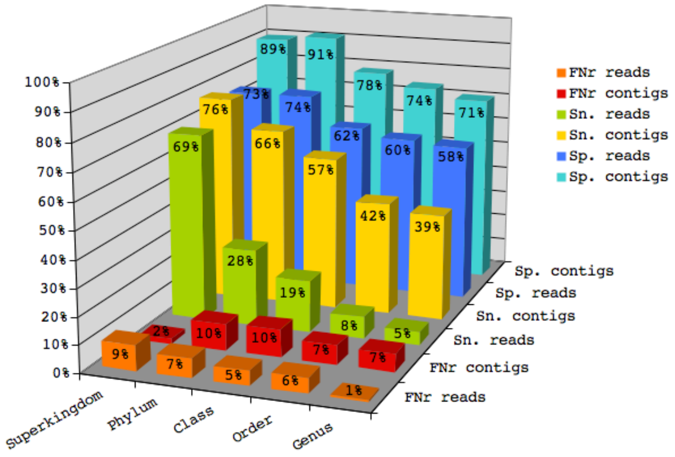


Figure 5.11: Overall performance achieved by TACOA at each taxonomic rank for reads and contigs.

Bars depict the average sensitivity (Sn.), specificity (Sp.), and false negative rate (FNr.) achieved for reads and contigs at each taxonomic rank.

62% (class) for 800bp reads and between 73% and 64% for reads of length 1Kbp. In light of the limited information contained in fragments of length 800bp - 1Kbp and the complexity of the classification problem (e.g. 62 classes on rank genus), TACOA also achieves a surprisingly good performance for single reads at rank order and genus (Figure 5.11). However, in practice it is not recommended to interpret classification results of single reads at these ranks because only a small number of organisms may be represented in the currently available sequenced genomes employed as references.

In real metagenomic data sets, already sequenced organisms may be contained in the studied sample. Therefore, the classification accuracy of TACOA was also assessed for fragments stemming from organisms included in the reference set (Figure 5.12). As expected, this has a markedly positive impact on the accuracy at all taxonomic ranks. An increase in sensitivity of up to 30% was observed. Furthermore, the specificity substantially increased while the false negative rate was reduced (Figure 5.12).

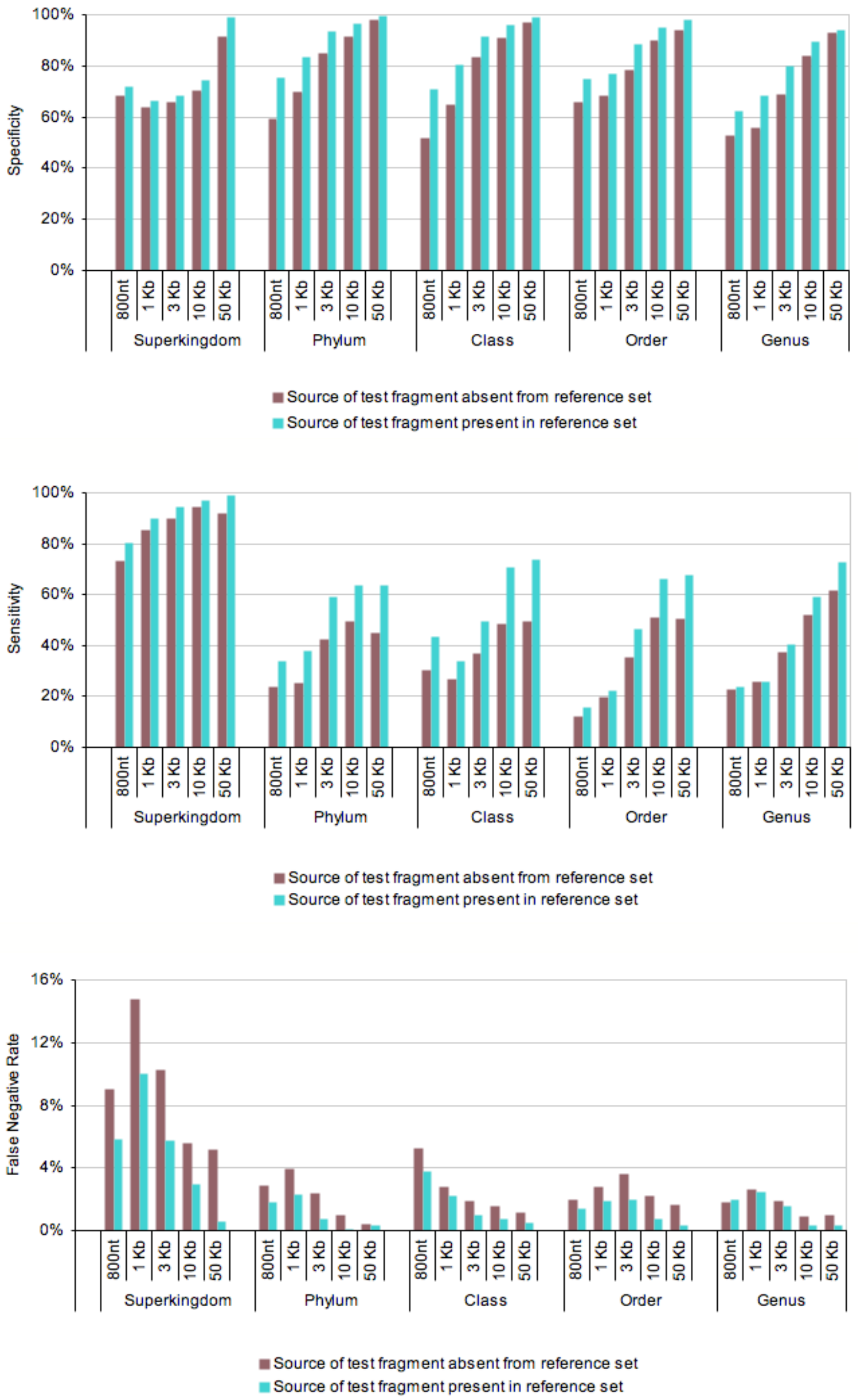


Figure 5.12: Classification accuracy achieved using two different reference sets. Each colored bar depicts the accuracy achieved by TACOA with two different reference sets. The label ”Taxonomic organism of test fragment absent from reference set” refers when the test fragment is classified using a reference set not containing the source organism from which the test fragment originates from.

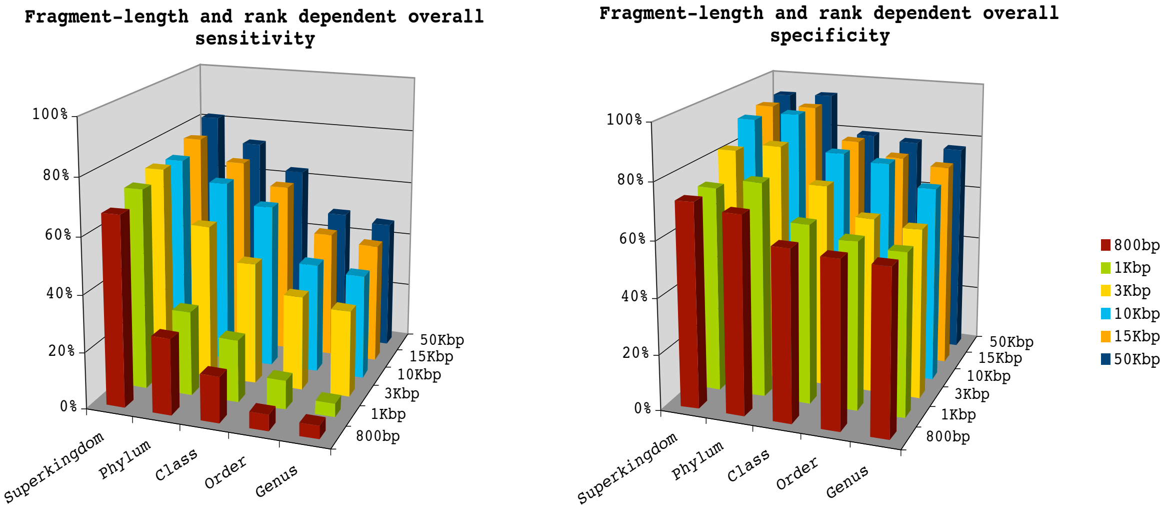


Figure 5.13: Fragment length and rank dependent performance.
Sensitivity (left) and specificity (right) achieved by TACOA for each genomic fragment length and taxonomic rank evaluated. Single single read lengths were simulated using genomic fragments of 800bp and 1Kbp long. Contigs were simulated by fragment lengths raging between 3Kbp and 50Kbp.

As a general trend, the accuracy improves when longer genomic fragments were classified (Figure 5.13). For example, on rank superkingdom the sensitivity increased from 67% (800bp reads) to 82% (50Kbp contigs) and at rank genus from 5% to 46% respectively. Conversely, the accuracy decreases as deeper taxonomic ranks were examined (Figure 5.11). In general, it is easy to predict classes that are well represented in the reference set, while detecting the underrepresented taxonomic groups is more challenging (Figure 5.10). TACOA is capable of detecting a remarkably high number of different taxonomic classes (Figure 5.14), for example for contigs of length 3Kbp, TACOA achieved a sensitivity above 20% for all 11 phyla, for 18 of the 21 classes, for 30 of the 45 order, and for 33 of the 61 genera represented in our test set (Figure 5.15).

5.3 Assessing the classification accuracy of TACOA and PhyloPythia for genomic fragments of variable length

TACOA was compared to the SVM-based PhyloPythia, which is one of the state-of-the-art, most accurate existing methods for the taxonomic classification of environmen-

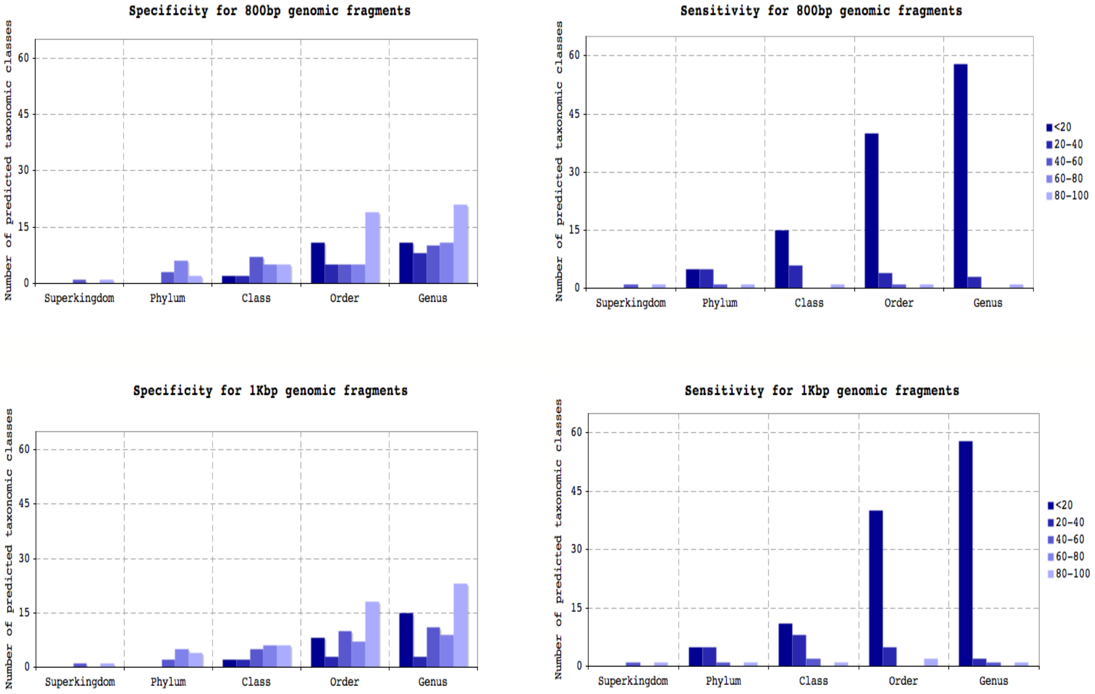


Figure 5.14: Specificity and sensitivity intervals for predicted taxonomic classes and reads.

Each bar depicts the sensitivity (right) and specificity (left) of predicted taxonomic classes for reads. Classification accuracy intervals for genomic fragments of length 800bp (top) and 1Kbp (bottom) is given. Per taxonomic rank, the distribution of number of predicted taxonomic classes at each interval is shown.

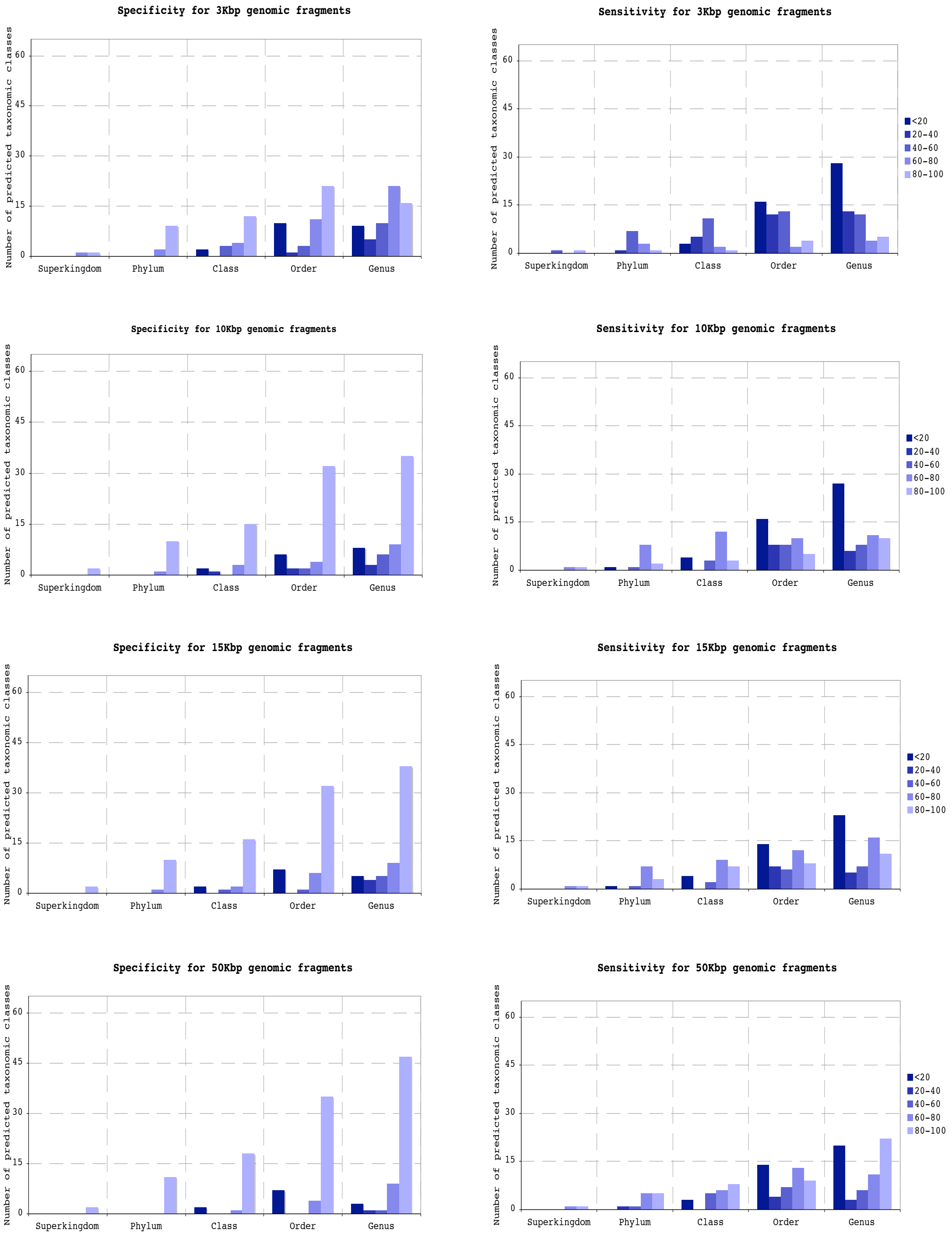


Figure 5.15: Specificity and sensitivity intervals for predicted taxonomic classes and contigs.

Each bar depicts the sensitivity (right) and specificity (left) of predicted taxonomic classes for contigs. Classification accuracy intervals for genomic fragments of length 3; 10; 15; and 50Kbp (from top to bottom) is given. Per taxonomic rank, the distribution of number of predicted taxonomic classes at each interval is shown.

tal genomic fragments. The accuracy of both classifier was assessed on a synthetic metagenome generated from 63 completely sequenced genomes.

5.3.1 Measuring the classification accuracy in the comparison of PhyloPythia and TACOA

The set of completely sequenced genomes used for comparison was selected as follows: at rank class, two different genomes were randomly chosen from each taxa guaranteeing that the data set used in the comparison is as unbiased as possible. This procedure yielded a set of 63 genomes that were downloaded from the NCBI genome database (Wheeler *et al.*, 2002). For each evaluated fragment length and for each selected genome, ten non-overlapping genomic fragments were randomly extracted for classification. Both classification strategies were evaluated at five different taxonomic ranks using three different genomic fragment lengths: 800bp, 1Kbp, and 10Kbp. The PhyloPythia web server with the built-in generic model was employed to predict the taxonomic origin of genomic fragments generated from the 63 selected genomes. TACOA was executed with default parameters to predict the taxonomic origin of genomic fragments from the same data set. Notice that this evaluation aims to investigate the performance that a researcher should expect when analyzing their metagenomic data.

The accuracy of both classifiers was assessed using the sensitivity, false negative rate and specificity. Values of sensitivity, specificity and false negative rate were computed as previously described in Section 1.5.2. As measures of accuracy the sensitivity and the false negative rates (FNr or misclassifications) was chosen to compare the PhyloPythia and TACOA classifier to account for possible compositional biases of the data set. The sensitivity and the FNr measured for one class do not depend on the composition of the remaining classes (since the term false positive is absent in the equations of sensitivity 2.9 and FNr 2.11). Hence, the sensitivity and FNr measured for each taxonomic group is not affected by possible biases of the test set. Contrastingly, the specificity measured for a class is strongly affected by the composition of the test set since it includes the false positives obtained from other classes.

5.3.2 Accuracy obtained by TACOA and PhyloPythia

The classification accuracy of the proposed kernelized *k*-NN classification method TACOA was compared to PhyloPythia using sensitivity, specificity and false negative rate as described in chapter 2, section 2.5.

In general, TACOA and PhyloPythia achieved quite comparable classification accuracies, but TACOA had a slightly improved performance for the classification of short DNA fragments. For the classification of reads of length 800bp and 1Kbp, TACOA has a larger sensitivity while both tools achieve a comparable false negative rate and specificity values (Figure 5.16). Remarkably, on ranks order and genus TACOA is still able to correctly classify between 3% and 17% of short fragments (sensitivity), while PhyloPythia cannot infer the taxonomic origin of any of the genomic fragments and thus has an average sensitivity of 0%. For longer contigs (DNA fragments of length 10Kbp) PhyloPythia is more sensitive on higher taxonomic ranks (superkingdom, phylum and class). In contrast, TACOA produces less misclassifications (false negative rate) making its prediction more reliable. On lower taxonomic ranks (genus and order), TACOA is able to correctly infer the taxonomic origin of about 10% to 17% of all contigs, while PhyloPythia has a sensitivity of 0% for all taxonomic groups at these ranks.

Across ranks superkingdom, phylum and class TACOA achieved sensitivity values of 71% to 3% for 800bp fragments and 76% to 11% for 1Kbp fragments. On the other hand, at the same ranks, PhyloPythia obtained a slightly lower sensitivity of 66% to 6% for 800bp fragments and 75% to 9% for 1Kbp fragments. At deeper ranks order and genus, TACOA was able to correctly classify between 3% and 7% of all short fragments (sensitivity), while only between 1% and 2.43% of fragments were misclassified (false negative rate). In contrast, PhyloPythia was not able to predict any taxa resulting in a sensitivity of 0% for all groups on these two ranks. According to the authors of PhyloPythia (personal communication), the stand alone classifier, which is not pub-

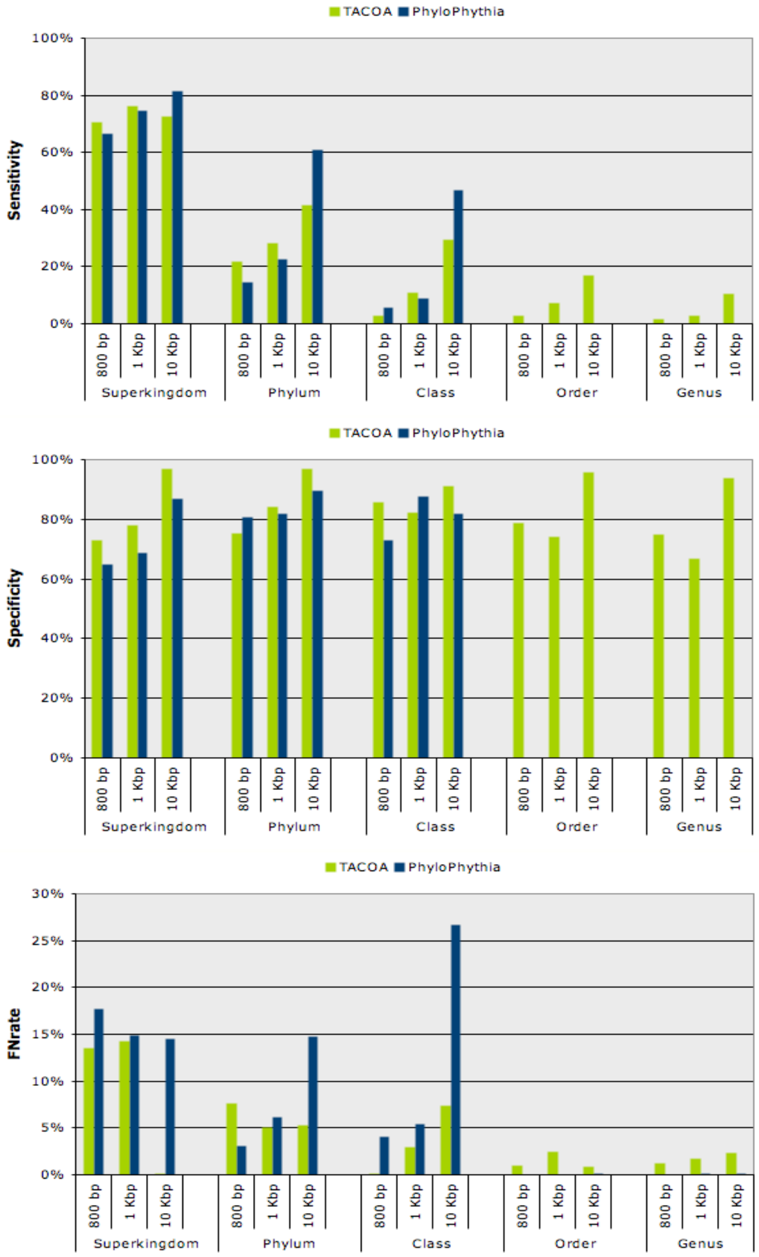


Figure 5.16: Classification accuracy obtained for TACOA and PhyloPythia. Sensitivity (top), specificity (middle) and false negative rate (bottom) achieved by TACOA and PhyloPythia for three different genomic fragment lengths and taxonomic ranks evaluated. Single read lengths are represented by fragments of length 800bp and 1Kbp and contigs by 10Kbp long fragments. The accuracy achieved is depicted using green bars for TACOA and blue bars for PhyloPythia. The sensitivity and specificity charts are scaled between 0 –100% and the false negative rate is scaled between 0 –30%

licly available is able to make predictions at ranks order and genus. Conversely, the web server available to the general user intentionally does not report prediction for these two lower taxonomic ranks explaining why the sensitivity achieved by PhyloPythia was 0%. However, this situation does not change anything on the fact that the standard user that employs PhyloPythia via the web server will only get predictions until taxonomic rank class while TACOA is able to provide predictions until rank genus.

In general, for short fragments TACOA is more sensitive at almost all taxonomic ranks, in particular at ranks order and genus. The only exception is at rank class, at which PhyloPythia is more sensitive for the classification of 800bp fragments. At the same time, for the classification of short fragments TACOA has a slightly lower false negative rate for almost all taxonomic ranks. Excepting rank phylum at which PhyloPythia has a lower false negative rate for 800bp fragments.

For the classification of contigs of length 10Kbp, TACOA achieved a sensitivity between 73% and 30% at ranks superkingdom to class, while PhyloPythia correctly classified between 82% and 47%. According to these results PhyloPythia was between 9% and 17% more sensitive than TACOA. But for the same contig length and ranks, TACOA was between 10% and 9% more specific than PhyloPythia. In addition, a high percentage of misclassifications was also observed for PhyloPythia (18.64% in average) in contrast to that achieved by TACOA (4.30% in average). At lower taxonomic ranks, TACOA achieved average sensitivity values between 17% (order) and 10% (genus) for the classification of 10Kbp contigs, while PhyloPythia was not able to predict any taxa for these long contigs, thus obtaining a sensitivity of 0% (Figure 5.16). Although PhyloPythia was not able to make predictions at ranks order and genus, a marginal misclassification rate was observed (0.14% at rank order and 0.10% at rank genus) for fragment length of 10Kbp.

5.4 Influence of horizontal gene transfer on the classification accuracy

The classification accuracy of methods using composition-based features might be influenced by an heterogeneous nucleotide composition present in the DNA sequence of the analyzed genomic fragment.

Although differences in the nucleotide composition of DNA sequences can be linked to a number of genomic attributes, including codon usage, DNA base-stacking energy, DNA structural conformation, strand asymmetry and even relic features of the primary genetic information, horizontal gene transfer events (HGT) is one of the most common cause (Bohlin *et al.*, 2008a; Zhang and Ya-Zhi, 2008). The work of Brown *et al.* also suggest that despite the rapid changes on the nucleotide composition of recent transferred DNA chunks, the phylogenetic signal from the donor can still be detected if the HGT event is recent, rather than ancient (Brown, 2003). Since the importance of HGT events has been gaining increasing attention lately (Keeling and Palmer, 2008), its influence in the accuracy of the intrinsic-based classifier TACOA was investigated.

One of the finding of this work is that tetranucleotides were best suited to analyzed genomic fragments $\leq$ 3Kbp. But it has been reported that tetranucleotide frequencies are a good measure to detect horizontally transferred regions (Bohlin *et al.*, 2008b). Therefore, any classifier aiming to predict the taxonomic origin of genomic fragments based in a tetranucleotide feature is susceptible to ”wrongly” classify to the donor taxonomic class a genomic fragment obtained via HGT. To explore the influence of HGT events in the classification accuracy of TACOA, fragments of length 1Kbp from two genomes (one archaeal and one bacterial) were selected. Several studies (Koonin *et al.*, 2001; Podell and Gaasterland, 2007; Ruepp *et al.*, 2000; Garcia-Vallve *et al.*, 2000) have reported acquisition of large stretches of DNA via HGT events for *Thermoplasma acidophilum* (archaea) and for *Thermotoga maritima* (bacteria).

In particular, the archaeal genome of *Thermoplasma acidophilum* has been reported to acquire ≈12% of its genome via HGT. The main donors seem to belong to bacterial organisms, but also some archaeal species have been proposed (Koonin *et al.*, 2001; Podell and Gaasterland, 2007). It has been suggested that *T. acidophilum* has received genes via HGT from *Sulfolobus solfataricus*, a distantly related crenarchaeota living in the same ecological niche (Ruepp *et al.*, 2000; Podell and Gaasterland, 2007). The sensitivity achieved by TACOA for *T. acidophilum* was 43% for reads 800bp long and 51% for reads of length 1Kbp.

In order to evaluate the taxonomic distribution of misclassifications for *T. acidophilum* genomic fragments, its genome was fragmented into pieces of length 1Kbp and predicted their taxonomic origin. For the 1,564 fragments analyzed, 1% (16 from 1,564) were misclassified into the order sulfolobales, another 3% (47 from 1,564) into other members of the euryarchaeota group, 7% (110 from 1,564) to a variety of members from the bacterial group, and 38% (601 from 1,564) could not be classified (Figure 5.17). From the proportion of genomic fragments that were ”erroneously” misclassified, the largest fraction (7%) was placed into the sulfolobus group. The results of the taxonomic distribution of ”misclassifications” made by TACOA for *T. acidophilum* are in close agreement to previous studies made by Koonin *et al.* (2001); Podell and Gaasterland (2007). Hence, the low number of correctly classified fragments obtained for *T. acidophilum* at rank genus may be partially explained by the lateral transfered DNA from other species.

The same analysis was performed for the bacterial genome of *Thermotoga maritima*, which is another organism with a high number of candidate genes that have been presumably acquired from archaea via HGT (Koonin *et al.*, 2001). A total of 1,860 genomic fragments of length 1Kbp each were classified using TACOA and analyzed (Figure 5.18). A high number of misclassified genomic fragments were ”wrongly” assigned to the archaeal group (91 from 1,860), a small fraction (27 from 1,860) was erroneously assigned to the sulfolobus group and 27% (503 from 1,860) could not be classified. Conversely to *T. acidophilum*, the genome *T. maritima* seems to be recipient of DNA originating mainly from archaeal species as already suggested by other authors (Koonin *et al.*, 2001; Podell and Gaasterland, 2007; Ruepp *et al.*, 2000; Garcia-Vallve *et al.*, 2000).

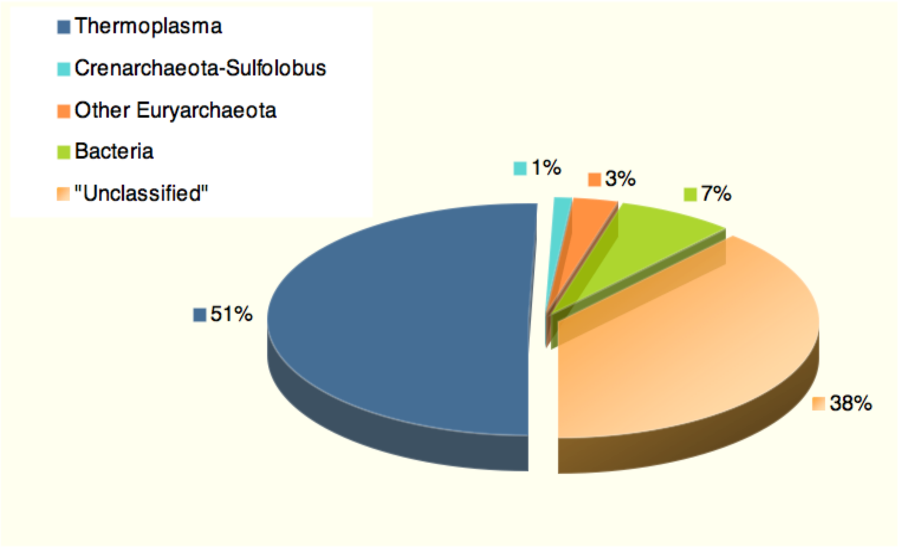


Figure 5.17: Distribution of taxonomic assignments for *Thermoplasma acidophilum*. Proportions of genomic fragments originating from the *T. acidophilum* genome that are misclassified into other taxonomic groups.

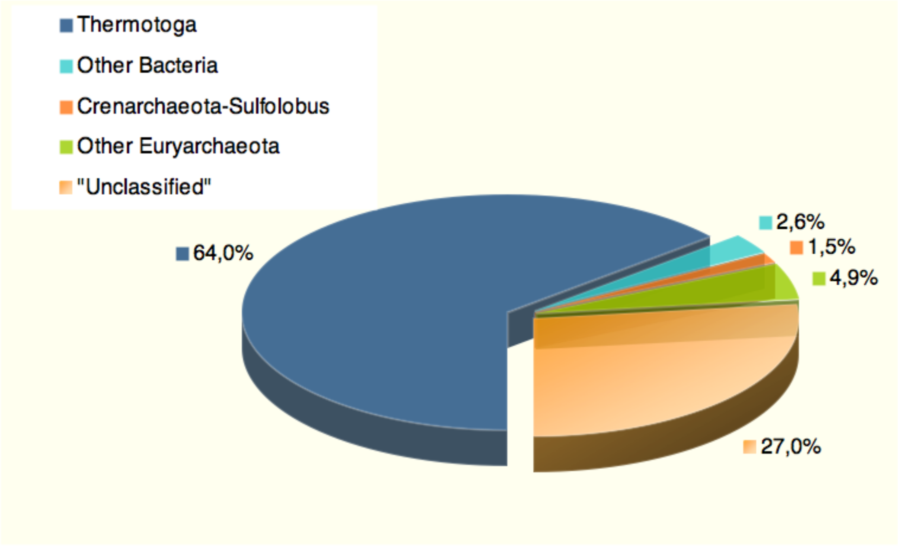


Figure 5.18: Distribution of taxonomic assignments for *Thermotoga maritima*. Proportions of genomic fragments originating from the *T. maritima* genome that are misclassified into other taxonomic groups.

These two case of studies strongly suggest that horizontally transfered stretches of DNA can affect the classification accuracy of a classifier using compositional based features to infer the taxonomic origin of genomic fragments. A possible explanation for this observation is that the nucleotide composition of transferred DNA chunks still carry phylogenetic signals from the donor genome after the HGT event has occurred as suggested by Brown (2003).

5.5 Cooperation in other metagenomic related projects

5.5.1 Overview

Nowadays high-throughput technologies are low in cost, fast and cloning biased free but the size of the produced sequences is small when compared to the traditional Sanger sequencing. One of the many open questions in the field of metagenomics relates to the taxonomic classification of very short fragments (ranging from 100 to 400bp). A short length of reads means that the information contained in it is very limited. The first cooperation described in the following sections investigated the problem of taxonomically classifying these short genomic fragments. The most important results obtained in this cooperation (the method developed and its classification accuracy) are given in subsection 5.6.2.

Another important aspect of metagenomic projects is the visualization of the analyzed data. After running the computational pipelines is desirable to summarize the information in a graphical manner permitting an easier interpretation. Section 5.6.3 reviews some results obtained within a cooperation, which one of the goals was to develop a method to reveal patterns on metagenomics data and visualize them.

5.5.2 Classification of short DNA fragments – CARMA

CARMA is an algorithm that employes Pfam protein families as phylogenetic markers to classify short read fragments based on a highly sensitive sequence similarity method.

This classifier incorporates two components: the first, dedicated to identify protein family fragments using profile hidden Markov models (pHMMs). Each existing protein family is modeled using pHMMs derived from multiple alignments of all members of each family. In CARMA, these pHMMS are used to identify new family members. The second, relates to the reconstruction of a phylogenetic tree per matching Pfam family. The taxonomic classification of an unknown read is based on its phylogenetic relationship to family members with known taxonomic affiliation, derived from the reconstructed phylogenetic tree. All gene fragments encoding a protein family are regarded as environmental gene tags (EGTs).

Identification of EGTs is carried out using the comprehensive and manually curated Pfam database. Multiple alignments of protein families as well as their corresponding pHMMs are deposited in the Pfam data base (Finn *et al.*, 2008). The taxonomic origin of each member of the protein family is also stored in the database. The highly sensitive pHMMs are employed to detect partial domains and protein families in short length read sequences.

All EGTs carrying a complete or partial protein family are aligned to the multiple alignment of the matching Pfam family. Subsequently, a phylogenetic trees is reconstructed using the multiple sequence alignment containing all members of the protein family and the corresponding matching EGTs. To reconstruct a phylogenetic tree for the EGTs and the members of the matching protein family, the pairwise distances of all members of the protein family matching an EGT is employed. The pairwise distance is computed using the fraction of identical amino acids contained in the aligned region. All phylogenetic trees reconstructed by CARMA are unrooted. An unrooted phylogenetic tree illustrate the relatedness of the sequences used but without assuming a common ancestry. Unrooted phylogenetic trees are reconstructed using the neighbor-joining clustering method (Saitou and Nei, 1987).

To evaluate the classification accuracy of the developed algorithm, 77 completely sequenced genomes were used to build a synthetic metagenome covering the archaea and bacteria phylum. The generated test set contained fragments of short length averaging

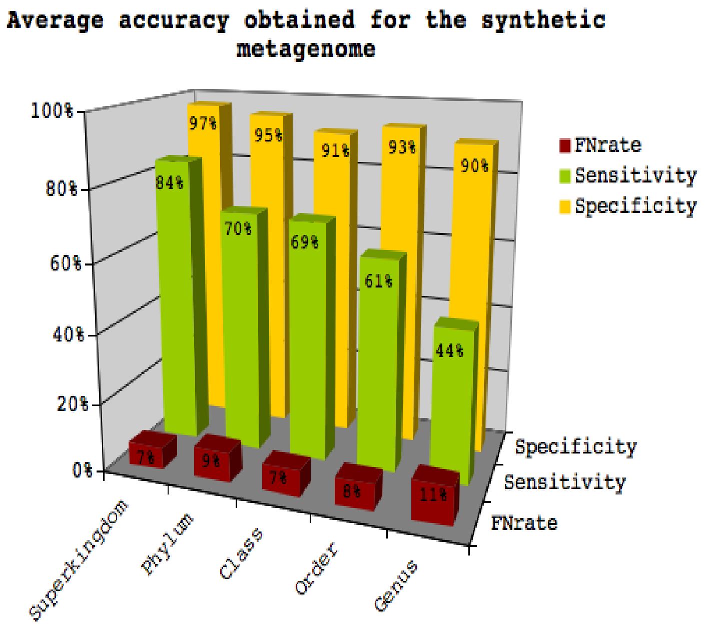


Figure 5.19: Average accuracy obtained by CARMA on a synthetic metagenome. The average accuracy obtained for assignments of short genomic fragments (80-120 bp) at five different taxonomic ranks are depicted as bars. Colors represent three different accuracy measures (sensitivity, specificity and False Negative rate).

100bp representing 10 phyla, 11 classes, 29 orders and 62 genera. Short sequence reads were simulated using all selected genomes and generating genomic fragments ranging between 80 – 120 bp in length. Artificial sequencing errors were introduced at homopolymers to better recreate the short reads generated with a GS 20 system (Margulies *et al.*, 2005).

In general, a high classification accuracy was obtained considering the short length of the genomic fragments analyzed. Approximately 15% of the ≈ 2.7 million genomic fragment analyzed was labeled as having an EGT. On average, the sensitivity ranged from 84% (superkingdom) to 44%(genus) while the specificity reached 97% at rank superkingdom and 90% at rank genus (Figure 5.19). Conversely, the number of misclassifications was kept low fluctuating between 7% (superkingdom) to 11% (genus). On the other hand, the number of EGTs that cannot be assigned to a known taxonomic group increased from 10% (superkingdom) to 45% (genus).

5.5.3 Using a hierarchically growing hyperbolic SOM to cluster and visualize taxonomic hierarchical data

The main idea behind the use of a hierarchically growing hyperbolic SOM (H^2SOM) was to take advantage of its capacity to visualize large and hierarchically structured data sets. H^2SOM employs hyperbolic spaces which are characterized by an uniform negative curvature, meaning that the size of the neighborhood around a given point increases exponentially with its radius (Ontrup and Ritter, 2006; Martin *et al.*, 2008). This exponential behavior enables to produce visualizations that are easy to explore, for example in Ontrup and Ritter (2006) the Poicaré disk was employed to project trained H^2SOM. This projection enables to selectively bring to the center (focal area) while still keeping on sight its surrounding (Ontrup and Ritter, 2006). Moreover, the use of H^2SOM allows to explore, cluster and visualize the analyzed data simultaneously in an unsupervised manner. To explore and classify metagenomic data the H^2SOM can be provided with pre-labeled data or by identifying highly conserved sequences using the 16sRNA to link each node to its taxonomic affiliation.

An example is given in Figure 5.20, which was generated using genomic feature vectors (GFVs) derived from 350 genomes from the archaea and bacteria phylum. GFVs were computed using oligonucleotides of length 4. In this visualization, the objects represent the most abundant taxonomic group, in terms of number of sequences. Objects are colored according its taxonomic affiliation at rank superkingdom, red and orange depict the archaeal group while yellow, green cyan, and blue the bacteria superkingdom. The labels of each node outlines its content, by numerical and taxonomically means. The inner or first ring of the trained H^2SOM represents the taxonomic rank superkingdom and is labeled with its representing groups and their respective counts. The trained H^2SOM continuos growing towards the periphery with outer rings each one of them representing deeper taxonomic ranks (i.e. phylum, class, order and genus). The background color helps to visualize the node distances in feature space, thus easily revealing areas of high or low variation among GFVs. Blue areas corresponds to large node distances while red ones signify that GFVs are close to each other in feature space.

The trained H^{2}SOM shown in Figure 5.20 reveals that GFVs are clustered in a biological meaningful manner, that is taxonomically related organisms are often localized close to each other. In general, nodes to which closely related species were assigned have lower variation, thus small distance (red background), reflecting their similar GFVs. Conversely, variation increases in areas located towards the inner ring regions indicating that GFVs contained in those nodes are more dissimilar (blue background). The high variation observed for inner rings with respect to the periphery is reasonable due to the wider phylogenetic spectrum to which the contained GFVs belong to. Analogously, Figure 5.20 shows that GFVs obtained from archaeal genomes display a higher variance than those computed from the superkingdom bacteria.

The case of study Thermotoga reviewed using TACOA was also found in the unsupervised clustering accomplished by the H^{2}SOM. In section 5.5, it was shown that parts of the Thermotoga genome was misclassified into the euryarchaeota phyla. In Figure 5.20 this case is easily spotted, Thermotoga despite of being a bacteria is located in a node with members of the euryarchaeota (archaea) phyla. The H^{2}SOM clustering results independently confirm the effect of horizontally gene transfer events in the classification problem when using compositional based features.

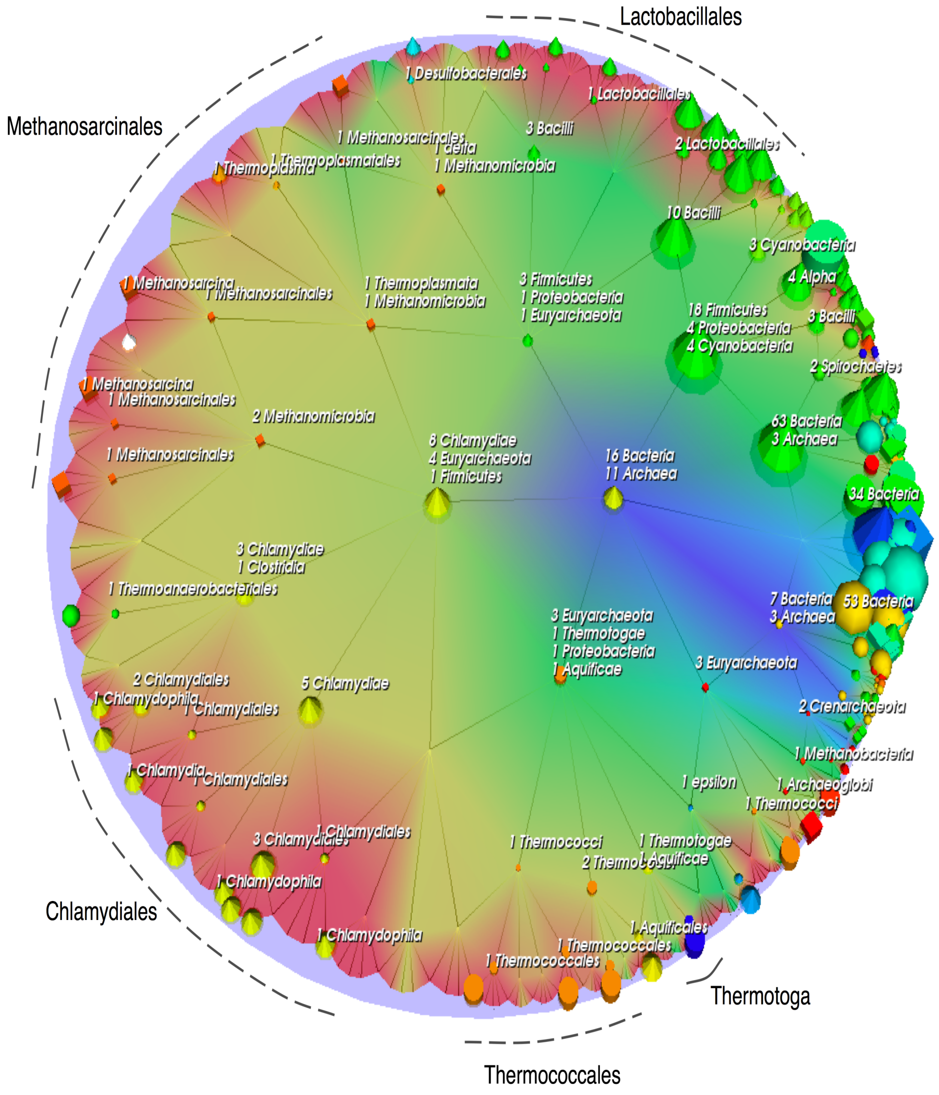


Figure 5.20: Graphical spotting of misclassified organisms.
A projection showing outer nodes and the taxonomic groups therein clustered. The projection has been dragged such that the methanosarcinales, chlamydiales, thermotogales, thermococcales and lactobacillales are on focus towards the edge of the graphic. At each node the most represented taxonomic group is displayed by a colored object red and orange for archaea and yellow, green, cyan and blue for bacteria. In this projection, the blue (high variation)-red (low variation) spectrum is more evident indicating the degree of variance of the groups affiliated to the exhibited nodes. It is readily seen that a blue object (bacteria - thermotoga) is co-localized within a group that is mainly composed by euryarchaeota.

CHAPTER 6

Discussion

TACOA, a novel method developed in this work is able to accurately predict the taxonomic origin of genomic fragments from metagenomic data sets by combining the advantages of the *k*-NN approach with a smoothing kernel function. The reference set used by the proposed method can be easily updated by simply adding the Genomic Feature Vectors (GFVs) of new genomes to the reference set without the need of retraining. TACOA is a standalone tool, which can be easily installed and can be run on a desktop computer. Therefore allowing researchers to locally analyze their metagenomic sequence data or integrate the program into their computational pipelines.

Analogous to PhyloPythia, sample specific-models of particular organisms can be easily integrated into TACOA framework and hence supporting the identification of organisms of special interest. Sample specific-models can be easily incorporated using the following approach: Genomic fragments carrying phylogenetic marker genes (such as rRNA genes) or fragments with high similarity to reference sequences of known origin (e.g. identified using a blast search) can be taxonomically annotated in a pre-processing step. Subsequently, these annotated fragments can be added to the reference set of TACOA. This can be easily achieved using the ”addReferenceGenome” program provided in TACOA. The use of sample-specific models can improve the accuracy of the classifier for those species that are represented in public databases (because the test set is contained in the reference set). It was demonstrated in this work that having the test set in the reference set substantially improves the sensitivity and specificity (up to 30%) and at the same time a decline on the false negative rate is observed.

As a whole, the classification accuracy at five different taxonomic ranks was evaluated: superkingdom, phylum, class, order, and genus. TACOA is able to correctly classify genomic fragments as short as 800bp up to rank class. It can be applied to predict the taxonomic origin of genomic fragments obtained using any sequencing technology able to produce fragments $\geq$ 800bp. The TACOA strategy also produced reliable predictions for genomic fragments originating from taxonomic groups that are absent from the reference set (simulating fragments stemming from genomes not yet sequenced). On average and over all taxonomic ranks, 77% of these fragments were correctly classified as ”unknown”.

TACOA compares well to PhyloPyhtia the current most sophisticated taxonomic classifier for environmental fragments. In terms of percentage of correctly classified fragments (sensitivity) TACOA slightly outperforms PhyloPythia for reads of length 800bp and 1Kbp at all taxonomic ranks evaluated, except for reads 800bp at rank class. But the very low false negative rate (0.16%) and the high specificity (86%) of TACOA makes the accuracy for reads of length 800bp (at rank class) comparable to that obtained by PhyloPythia. Compared to TACOA, the overall reduced sensitivity obtained by PhyloPythia (evident for the analyzed read lengths) is partially due to the absence of the phylum Chloroflexi and Thermatogae from its training set. This example illustrates the positive effect of an updated training or reference set in the prediction of known taxonomic classes.

For contigs of length 10Kbp, TACOA achieved a lower sensitivity, lower false negative rate and higher specificity values than PhyloPyhtia. Although PhyloPythia achieves higher sensitivity values for contigs of length 10Kbp the overall performance is comparable for both classifiers at ranks superkingdom, phylum and class.

At deeper taxonomic ranks (order and genus), for all evaluated lengths TACOA was still able to provide correct classifications for several taxonomic classes (average sensitivity of about 7%) while PhyloPythia failed in making any taxonomic assignments (sensitivity of 0%). With an average sensitivity of 17% (order) and 10% (genus), an average false negative rate of 1.45% (order) and 2.29% (genus), TACOA can provide a more detailed view of the taxonomic composition of an environmental sample. Notice that in practice it is not recommended to draw conclusions at such deep ranks for reads $\leq$ 1Kbp because only a small number of fragments may be represented in the currently available sequenced genomes.

The accuracy of both classifiers was assessed using the sensitivity, false negative rate and specificity. To rule out any differences on performance due to possible compositional biases of the data set used in the comparison between PhyloPythia and TACOA, only the sensitivity and false negative rates (FNr) were given emphasis. The reasoning behind this criteria is that the sensitivity and FNr measure for a taxonomic class is independent of the composition of remaining classes. The term false positive is absent in the equations used to compute the sensitivity and FNr (chapter 2). Hence, the sensitivity and FNr measured for each taxonomic group is not affected by possible biases of the test set. Contrastingly, the specificity measured for a class is strongly affected by the composition of the remaining test set since it includes the false positives obtained from other classes. To better illustrate this issue two cases are given:

First case: Lets assume that at rank phylum only three different taxonomic classes exist: Proteobacteria, Cyanobacteria and Chloroflexi. Lets also assume that the number of DNA fragments representing each class is biased: Proteobacteria has 10,000 fragments while Cyanobacteria and Chloroflexi have 100 fragments each. Now lets assume that the FNr is constant at 10% per class. In consequence, each class contributes to the false positives (FPs) of other classes unequally: 1,000 fragments from Proteobacteria will be wrongly assigned between Cyanobacteria and Chloroflexi. On the other hand, only 10 fragments from Cyanobacteria will be wrongly assigned between Proteobacteria or

Chloroflexi. Additionally, only 10 fragments from Chloroflexi will be assigned between Proteobacteria and Cyanobacteria. Furthermore, lets assume that the false positives are assigned at random to one of the wrong classes. In this case, a high specificity for Proteobacteria will be measured and a low specificity for Cyanobacteria and Chloroflexi.

Second case: Now lets assume that class Protebacteria is represented by only 10 fragments and Cyanobacteria and Chloroflexi are both represented again by 100 fragments. In this case, a low specificity for Proteobacteria but a high specificity for Cyanobacteria and Chloroflexi would be measured.

These two cases clarify the issue of how the specificity is influenced by the composition of the entire data set. On the other hand, the sensitivity and FNr are not affected by the false positives thus their values for each of the three classes should be very similar in both cases. This example clearly illustrates that in a multi-class classification problem, the sensitivity and FNr measured for each class is not impacted by the composition/bias of the entire test set but the specificity is. However, the specificity for TACOA and PhyloPythia was also given because it provides an idea on how reliable the predictions are, despite of being influenced by possible biases of the test set.

It has already been reported for many practical examples that simple traditional classification algorithms such as K-NN can achieve competitive results when compared to more sophisticatted techniques such as SVMs (Zhu *et al.*, 2007). Moreover, several works have already shown that a boost in the performance of the K-NN algorithm can be achieved by introducing modifications such as: a) a weight adjusted scheme (Song *et al.*, 2007), K-NN-kernel (Hotta *et al.*, 2004), b) modified distance metrics – adaptive metrics – (Domeniconi *et al.*, 2002), c) large margin (Weinberger *et al.*, 2006) among others. Furthermore, for some practical applications the performance of a modified K-NN algorithm demonstrated to be competitive or even outperform more sophisticated machine learning techniques such as SVMs (Song *et al.*, 2007; Okum, 2006; Yao and Ruzzo, 2006; Saha and Heber, 2006; Berrar *et al.*, 2006; Zhu *et al.*, 2007). In this work a modified K-NN-approach was used, namely a kernelized K-NN, which combines the advantages of K-NN with those of kernel methods (also used by the SVM technique).

In the comparison analysis of TACOA and PhyloPythia similar results were obtained. The slightly higher accuracy achieved by TACOA does not stem from the method used, in turn it is valid to assume that the higher accuracy is the result of the entire strategy implemented in the classifier presented in this thesis. A possible explanation could be that the features used in TACOA probably have a better discrimination power among the different taxonomic classes. TACOA uses a ratio between the observed and the expected frequency of each oligonucleotide as features which can be considered as a more elaborated measure of the oligonucleotide frequencies since it considers over and underrepresentation of a given oligonucleotide. This ratio is computed for each possible oligonucleotide of a fixed length. In contrast, PhyloPythia uses plain frequencies of a given set of patterns in a sequence. For example, in McHardy *et al.* (2007) it was reported that oligonucleotide patterns that best separate taxonomic classes have a dual dependency: rank and genomic fragment length dependency. Conversely, the best separating feature in TACOA depends only on the genomic fragment length. It is likely that the features chosen for the PhyloPythia classifier would require longer fragments. This would explain the higher sensitivity achieved by TACOA for smaller genomic fragments.

The fact that TACOA and PhyloPythia obtained comparable results can also be partially explained by the following reasons: a) both approaches compare the query vector to reference vectors via a Gaussian kernel function. b) Both approaches are able to learn complex and disjoint decision functions. c) Both approaches can deal with high-dimensional input-data (using a Gaussian kernel), with unbalanced training-data (TACOA using its weighting scheme) and are able to perfectly separate the classes in most practical applications. The key difference between both classifiers is that the SVM based PhyloPythia is able to maximize the margin between the learned hyperplane and the two classes which is not the case for TACOA. On the other hand, the kernelized approach in TACOA has the advantage to be a natural multi-class classifier, in contrast to a collection of binary classifiers.

Based on the comparable results obtained for PhyloPythia and TACOA, it can be drawn that for this practical application the maximization of the margin (in the case of SVM) does not have a great influence in the overall accuracy achieved. This may be explained by the sparseness of the data in the input space. If the classes are distant enough (not too many examples are located in the class boundaries) then maximizing the margin may not have a strong effect on the performance of the classifier.

According to the arguments presented above it is reasonable to obtain comparable results. Notably, the idea behind the comparison between the two classifiers was to evaluate the strategy as a whole and not aimed to compare the performance of SVMs, K-NN and kernelized K-NN. It can be said that the comparison of the classifier made in this work was user-oriented or what an end user will expect when utilizing the available classifiers. In the end, users are interested on the overall performance of the tools available for the taxonomic classification of environmental DNA fragments.

An interesting observation made during this work was that the classification of genomic fragments is possible using only GFVs computed from complete sequenced genomes rather than computing the vectors on fragments derived from complete genomes. Similar observations have already been made by Abe *et al.* in 2005 and 2006 and more recently by McHardy *et al.* in 2007, where the developed classifiers were trained with genomic fragments longer than those being tested. In addition to these findings, this work demonstrated that complete genomes can also be used as reference to classify environmental genomic DNA fragments.

This study supports the findings that frequencies of short length oligonucleotides (i.e. tetra- and penta-oligonucleotides) are best suited to capture taxon-specific differences among prokaryotic genomes (Abe *et al.*, 2005, 2006; Sandberg *et al.*, 2001; Teeling *et al.*, 2004a). Moreover, our parameter search analysis strongly suggests that tetra- or penta-oligonucleotides frequencies are optimal features for TACOA to classify environmental genomic fragments as short as 800bp. This observation is in accordance to those reported by Bohlin *et al.* Bohlin *et al.* (2008a) who already proposed that little increase in information potential about phylogenetic relationships is gained when using oligonucleotide sizes larger than hexa-nucleotides.

Parts of this work demonstrated that recent events of HGT can affect the accuracy of a composition-based classifier. The correct classification of horizontally transferred regions into its ”current” taxon is difficult if these still carry a strong phylogenetic signal from the donor genome. This was illustrated by classifying fragments of length 1Kbp from the archaea *T. acidophilum* and the bacteria *T. maritima*. Notably, HGT is not the only phenomena causing variations in the oligonucleotide frequencies within genomes and hence affecting the classification performance.

The method developed in this work, TACOA, combines the ability of predicting the taxonomic origin of genomic fragments with high accuracy and the advantages of being a tool that can easily be installed and used on a desktop computer breaking any dependency and limitations that web server services may bring. Altogether, it strongly suggests that TACOA offers a great potential to assist on the exploration of the taxonomic composition of metagenomic data sets.

CHAPTER 7

Conclusions

One of the foremost contribution of this dissertation is a novel strategy targeting the problem of taxonomic classification of genomic fragments. Furthermore, the strategy presented in this work uses features that do not require sequence homology thus making possible the classification of genomic fragments by means of comparison of statistical properties derived directly from the DNA of taxonomic related organisms. Its contributions include:

- **Development of a stand alone tool named TACOA to taxonomically classify genomic fragments.** The strategy implemented in TACOA classifies a genomic fragments based on a ratio that is able to measure under- and over representation of all oligonucleotides on a DNA sequence. The classifier itself combines the simplicity of the *k*-NN algorithm with a kernel function. From a practical perspective, this combination, together with the selected features, made possible to develop a novel strategy for the classification of genomic fragments. Moreover, TACOA compares very well with other approaches employing more sophisticated methods being at the same time easy to understand, apply and implement. Moreover it was demonstrated in this work that the methodology is very accurate for ranks: superkingdom, phylum and class while the modest accuracy obtained at lower ranks will be soon overcome as the genomic fragment length growths in length.

- **Frequencies of short length oligonucleotides are best suited to capture taxon-specific differences among prokaryotic genomes.** The parameter search analysis performed in this work strongly suggests that tetra- and penta-oligonucleotides frequencies are optimal features for the strategy implemented in TACOA to classify environmental genomic fragments as short as 800bp.

- **Phylogenetic signal of complete genomes can still be traced to genomic fragments.** In general, close related organisms have the same pattern of over- and under-represented set of oligonucleotides and this oligonucleotide patterns are taxon specific. The phylogenetic signal is strong enough to be traced even in DNA chunks that have been recently transfer from one specie to other, despite the rapid adaption of the foreigner DNA to its new ”host”.

- **Complete genomes can be used as reference for *de novo* classification.** This work demonstrated that even complete genomes can be used as reference to classify environmental genomic DNA fragments. So far the taxonomic classification of genomics fragments was made using as reference also genomic fragments of a longer length than the one being analyzed.

- **Recent events of horizontal gene transfer can affect the accuracy of a composition based classifier.** Genomic fragments that have just being exchange between species will be ”misclassified” into their taxon of origin only because they still carry the phylogenetic signal of their source organism. The classification accuracy for such fragments is jeopardized and the overall accuracy for a taxon will highly depend on how much foreigner DNA it has received.

- **Use of sample specific models can greatly aid in identifying taxonomic groups of interest.** In this work it was demonstrated that guided searches can be performed. The use of high quality (e.g. fragments carrying rRNA genes) fragments from a taxonomic group of interest can greatly help in identifying genomic fragments from the same taxonomic group.

- **For multiclass classification problems the sensitivity and false negative rate are unbiassed measures.** In a multi-class classification problem, the sensitivity (Sn) and False Negative rate (FNr) (as defined in this work) measured for each class does not depend on any composition or bias of the test set, but the specificity does. This indicates that comparisons between methods should be made on the basis of the Sn and FNr, for those cases where another method (e.g. ROC analysis) can not be made.

- **First successful application of a kernelized *k*-NN for the problem of taxonomic classification of metagenomic data.** TACOA can be successfully apply to metagenomic data set that have a minimum genomic fragment size of 800bp. It was demonstrated in this work that the presented strategy can deal with high dimensional data owing to the integration of a Gaussian kernel into the *k*-NN algorithm.

CHAPTER 8

Future directions

The work presented in this dissertation is a first step towards answering one of the many questions posed by the emerging field of metagenomics. But as one of the pioneering work offers a solution to an immediate need, that is the problem of taxonomically classifying environmental genomic fragments. At the same time, it opens a myriad of new challenges that can be explored as a natural extension of this work. In this chapter an overview of possible directions of research is given.

There are two major areas of research that are immediately foreseen, the first concerns to the improvement of the classification strategy itself. The second relates to the visualization of the ever increasing amount of metagenomic data.

- *Improvements to the classification strategy.*

  One aspect to improve in the classification strategy is to perform a deeper analysis of the features (i.e. oligonucleotide ratio) used in the strategy of the TACOA classifier. It will be undoubtedly helpful to investigate wether the classification strategy would benefit from a feature selection preprocessing step. Or even further, if problematic taxonomic groups, in terms of classification, can benefit from it. For example it is possible to use feature selection techniques such as the Wrapper technique (Kohavi and John, 1998) which incorporates class information by evaluating sets of features according to the performance of the classifier. Therefore, the set of selected features are specifically tuned to a given classification method in this case to TACOA.

It was already shown, as part of the collaboration work done with Christian Martin for the visualization of metagenomic data, that feature selection of the oligonucleotide frequency helped in improving the clustering step. Feature selection could also positively impact the speed performance of the classifier presented in this work. If the size of the genomic feature vector is reduced then the size of the whole data set will also be reduced. As consequence, the amount of data to be stored and the computational time required in the classification step will be reduced.

Another interesting direction of research is to explore if techniques such as ”condensing” (Hart, 1968) or ”editing” (Wilson, 1972) can speed up the classification step of new items. The main advantage of ”condensing” is to reduce the training set by eliminating many reference items that need to be stored, but considerably retaining the decision boundary (DB). the idea behind condensing is that references items close to the DB are essential for the *k*-NN classification while those far from the DB do not impact the decision. Thus, deletion of this inefficient reference items aids in reducing the computational time (Fayed and Atiya, 2009). On the other hand, complete removal of reference items, for example, outlier reference items that are embedded by items from other classes or ”editing” can improve generalization capabilities. However, the effect of editing is already a feature implemented in TACOA by the using the Gaussian kernel but it could help in a better tuning of the lambda parameter.

- *Improvements for the visualization.*

> Another natural extension of the presented framework is to develop a visualization module able to graphically represent the discriminant function used to decide to which taxonomic class to a genomic fragment is more likely to be affiliated. An immediate benefit would be to explore the degree of association of the analyzed fragment to the different taxonomic classes. Thus, can also assist the researcher in deciding which is the taxonomic rank with stronger support for the proposed or conflicting classification of special cases. As consequence a greater interpretability of the results can be achieved. In this context, the methodology presented in this work is fully transparent since classification details as the one discussed above are completely traceable. This is not possible in black box strategies such as support vector machines.

The evaluation made in this dissertation demonstrated that TACOA can accurately predict the taxonomic origin of reads and contigs until rank genus. Even for reads as short as 800bp these predictions are reliable until rank order. As described before, next generation sequencing technologies have revolutionized the field of genome research owing to high throughput and low sequencing costs. However, the main draw back is the short read length of about 350bp for pyrosequencing and 120bp for Solexa. Therefore, these reads are too short to be directly taxonomically classify using any of the intrinsic classifiers including TACOA. However, sequencing technologies are rapidly advancing and the problem of short length reads will soon be overcome. This will make these sequencing technologies even more valuable for metagenomics as in metagenomics read length clearly does matter.

On the other hand, if in the future is possible to sequence reads of several thousand base pairs many of the current challenging problems in metagenomics will get much easier or even completely resolved, including taxonomic classification, reconstruction of complete genomes (assembly), and functional annotation.

Finally, another exciting direction of research is to bridge the knowledge gain by identifying ”who is out here” (metagenomics) to ”what they are doing” (functional genomics). The motivation for bridging focuses in the quest to go one step further: from characterization or cataloging, first necessary and important step, to putting into a community context ”what we see”. Furthermore, another ideal goal is to investigate the community dynamics, in other words how ”what we see” changes over time and range of conditions. However, to successfully achieve this objective is of great importance to accurately know ”what is out there” and ”what they are doing”.

Bibliography

Abe T., Sugawara H., Kanaya S., Ikemura T.: A novel bioinformatics tool for phylogenetic classification of genomic sequence fragments derived from mixed genomes of uncultured environmental microbes. *Polar Biosci*, 20:103–112, (2006).

Abe T., Sugawara H., Kinouchi M., Kanaya S., Ikemura T.: Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples. *DNA Res*, 12:281–290, (2005).

Altschul S. F., Madden T. L., Schaffer A. A., Zhang J., Zhang Z., Miller W., Lipman D. J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25:3389–3402, (1997).

Amann R.: Who is out there? Microbial aspects of biodiversity. *Syst. Appl. Microbiol.*, 23:1–8, (2000).

Asuncion A., Newman D.: UCI machine learning repository (2007).

Baldauf S. L., Roger A. J., Wenk-Siefert I., Doolittle W. F.: A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science*, 290:972–977, (2000).

Baldi P., Brunak S., Chauvin Y., Andersen C. A., Nielsen H.: Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16:412–424, (2000).

Barluenga M., Stölting K. N., Salzburger W., Muschick M., Meyer A.: Sympatric speciation in Nicaraguan crater lake cichlid fish. *Nature*, 439:719–723, (2006).

Béjà O., Aravind L., Koonin E. V., Suzuki M. T., Hadd A., Nguyen L. P., Jovanovich S. B., Gates C. M., Feldman R. A., Spudich J. L., Spudich E. N., DeLong E. F.: Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science*, 289:1902–1906, (2000).

Ben-Hur A., Ong C. S., Sonnenburg S., Schölkopf B., Rätsch G.: Support vector machines and kernels for computational biology. *PLoS Comput. Biol.*, 4:e1000173, (2008).

Berrar D., Bradbury I., Dubitzky W.: Instance-based concept learning from multiclass dna microarray data. *BMC Bioinformatics*, 7:73, (2006).

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

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

Boser B. E., Guyon I. M., Vapnik V. N.: A training algorithm for optimal margin classifiers. In: *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152, ACM Press (1992).

Brown J.: Ancient horizontal gene transfer. *Nature Reviews*, 4:121–132, (2003).

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

Campbell A., Mrázek J., Karlin S.: Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proc Natl Acad Sci U S A*, 96:9184–9189, (1999).

Cardenas E., Tiedje J. M.: New tools for discovering and characterizing microbial diversity. *Curr. Opin. Biotechnol.*, 19:544–549, (2008).

Chan C., Hsu A., Halgamuge S., Tang S.: Binning sequences using very sparse labels within a metagenome. *BMC Bioinformatics*, 9:215, (2008).

Chang C.-C., Lin C.-J.: *LIBSVM: a library for support vector machines* (2001), software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Chen L. L., Zhang C. T.: Seven GC-rich microbial genomes adopt similar codon usage patterns regardless of their phylogenetic lineages. *Biochem. Biophys. Res. Commun.*, 306:310–317, (2003).

Chen P.-H., Lin C.-J., Schölkopf B.: A tutorial on ν-support vector machines: Research articles. *Appl. Stoch. Model. Bus. Ind.*, 21(2):111–136, (2005).

Cover T., Hart P.: Nearest Neighbor Patter Classification. *IEEE Transactions*, 13:21–27, (1967).

Crammer K., Singer Y.: On the learnability and design of output codes for multiclass problems. *Machine learning*, 47:2–3, (2002).

Dandekar T., Snel B., Huynen M., Bork P.: Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.*, 23:324–328, (1998).

Darwin C.: *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, London (1859).

Domeniconi C., Peng J., Gunopulos D.: Locally adaptive metric nearest-neighbor classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(9):1281–1285, (2002).

Duda R., Hart P., Stork D.: *Pattern Classification*. Willey Interscience, USA (2001).

Dufraigne C., Lespinats B., Giron S., Dechavanne P.: Detection and characterization of horizontal transfers in prokaryotes using genomic signatures. *Nucleic. Acid Res.*, 33:e–6, (2005).

Eppley J. M., Tyson G. W., Getz W. M., Banfield J. F.: Genetic exchange across a species boundary in the archaeal genus ferroplasma. *Genetics*, 177:407–416, (2007).

Fayed H., Atiya A. F.: A novel reduction approach for the k-nearest neighbor method. *IEEE Trans on Neural Networks*, 20:890–896, (2009).

Finn R., Tate J., Mistry J., Coggill P., Sammut S., Hotz H., Ceric G., Forslund K., Eddy S., Sonnhammer E., Bateman A.: The Pfam protein families database. *Nucleic Acids Res.*, 36:D281–288, (2008).

Fitz-Gibbon S. T., House C. H.: Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Res.*, 27:4218–4222, (1999).

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.: Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, 269:496–512, (1995).

Foerstner K. U., von Mering C., Hooper S. D., Bork P.: Environments shape the nucleotide composition of genomes. *EMBO Rep*, 6:1208–1213, (2005).

García Martín H., Ivanova N., Kunin V., Warnecke F., Barry K. W., McHardy A. C., Yeates C., He S., Salamov A. A., Szeto E., Dalin E., Putnam N. H., Shapiro H. J., Pangilinan J. L., Rigoutsos I., Kyrpides N. C., Blackall L. L., McMahon K. D., Hugenholtz P.: Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat. Biotechnol.*, 24:1263–1269, (2006).

Garcia-Vallve S., Romeu A., Palau J.: Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res*, 10:1719–1725, (2000).

Gerstein M.: Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. *Proteins*, 33:518–534, (1998).

Gerstein M., Hegyi H.: Comparing genomes in terms of protein structure: surveys of a finite parts list. *FEMS Microbiol. Rev.*, 22:277–304, (1998).

Gill S. R., Pop M., Deboy R. T., Eckburg P. B., Turnbaugh P. J., Samuel B. S., Gordon J. I., Relman D. A., Fraser-Liggett C. M., Nelson K. E.: Metagenomic analysis of the human distal gut microbiome. *Science*, 312:1355–1359, (2006).

Hall N.: Advanced sequencing technologies and their wider impact in microbiology. *J. Exp. Biol.*, 210:1518–1525, (2007).

Hanage W. P., Spratt B. G., Turner K. M., Fraser C.: Modelling bacterial speciation. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 361:2039–2044, (2006).

Hart P.: The condensed nearest neighbor rule. *IEEE Trans Inform. Theory*, 14:515–516, (1968).

Hastie T., Tibshirami R., Friedman J.: *The Elements of Statistical Learning*. Springer-Verlag, New York (2002).

Holland R. C., Down T. A., Pocock M., Prlić A., Huen D., James K., Foisy S., Dräger A., Yates A., Heuer M., Schreiber M. J.: BioJava: an open-source framework for bioinformatics. *Bioinformatics*, 24:2096–2097, (2008).

Hotta S., Kiyasu S., Miyahara S.: Pattern recognition using average patterns of categorical k-nearest neighbors. In: *ICPR (4)*, pages 412–415 (2004).

Hudson M.: Technical review: Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources*, 8:3–17, (2008).

Hugenholtz P., Tyson G. W.: Microbiology: metagenomics. *Nature*, 455:481–483, (2008).

Huson D., Auch A., Qi J., Schuster S.: MEGAN analysis of metagenomic data. *Genome Res.*, 17:377–386, (2007).

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

Karlin S.: Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends Microbiol.*, 9:335–343, (2001).

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

Karlin S., Mrázek: Distintive features of large complex virus genomes and proteomes. *Proc. Natl. Acad. Sci.*, 104:5127–5132, (2007).

Karlin S., Mrázek J., Campbell A. M.: Compositional biases of bacterial genomes and evolutionary implications. *J Bacteriol*, 179:3899–3913, (1997).

Keeling P. K., Palmer J. D.: Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics*, 9:605–618, (2008).

Kohavi R., John G. H.: The wrapper approach. In: H. Liu, H. Motoda, eds., *Feature Selection for Knowledge Discovery and Data Mining*, pages 33–50, Kluwer Academic Publishers (1998).

Kohonen T.: Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69, (1982).

Koonin E. V., Makarova K. S., Aravind L.: Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol*, 55:709–742, (2001).

Korbel J. O., Snel B., Huynen M. A., Bork P.: SHOT: a web server for the construction of genome phylogenies. *Trends Genet.*, 18:158–162, (2002).

Kunin V., Copeland A., Lapidus A., Mavromatis K., Hugenholtz P.: A bioinformatician's guide to metagenomics. *Microbiol. Mol. Biol. Rev.*, 72:557–578, (2008).

Kurokawa K., Itoh T., Kuwahara T., Oshima K., Toh H., Toyoda A., Takami H., Morita H., Sharma V. K., Srivastava T. P., Taylor T. D., Noguchi H., Mori H., Ogura Y., Ehrlich D. S., Itoh K., Takagi T., Sakaki Y., Hayashi T., Hattori M.: Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res.*, 14:169–181, (2007).

Langley P., Iba W., Thompson K.: An analysis of bayesian classifiers. In: *In Proceedings of the tenth national conference on artificial intelligence*, pages 223–228, AAAI Press (1992).

Letunic I., Copley R. R., Pils B., Pinkert S., Schultz J., Bork P.: SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.*, 34:D257–260, (2006).

Lindsay S.: Genetic sequencing. *The bulletin of atomic science*, 64:5053, (2008).

Lodé T.: Genetic divergence without spatial isolation in polecat mustela putorius populations. *J. Evol. Bio.*, 14:228–236, (2001).

López-García P., Moreira D.: Tracking microbial biodiversity through molecular and genomic ecology. *Res. Microbiol.*, 159:67–73, (2008).

Margulies M., Egholm M., Altman W. E., Attiya S., Bader J. S., Bemben L. A., Berka J., Braverman M. S., Chen Y.-J., Chen Z., Dewell S. B., Du L., Fierro J. M., Gomes X. V., Godwin B. C., He W., Helgesen S., Ho C. H., Irzyk G. P., Jando S. C., Alenquer M. L. I., Jarvie T. P., Jirage K. B., Kim J.-B., Knight J. R., Lanza J. R., Leamon J. H., Lefkowitz S. M., Lei M., Li J., Lohman K. L., Lu H., Makhijani V. B., McDade K. E., McKenna M. P., Myers E. W., Nickerson E., Nobile J. R., Plant R., Puc B. P., Ronan M. T., Roth G. T., Sarkis G. J., Simons J. F., Simpson J. W., Srinivasan M., Tartaro K. R., Tomasz A., Vogt K. A., Volkmer G. A., Wang S. H., Wang Y., Weiner M. P., Yu P., Begley R. F., Rothberg J. M.: Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437:376–380, (2005).

Markowitz V. M., Ivanova N., Palaniappan K., Szeto E., Korzeniewski F., Lykidis A., Anderson I., Mavromatis K., Mavrommatis K., Kunin V., Garcia Martin H., Dubchak I., Hugenholtz P., Kyrpides N. C.: An experimental metagenome data management and analysis system. *Bioinformatics*, 22:e359–367, (2006).

Martin C., Diaz N., Ontrup J., Nattkemper T.: Hyperbolic som-based clustering of dna fragment features for taxonomic visualization and classification. *Bioinformatics*, 24:1568–1574, (2008).

Mavromatis K., Ivanova N., Barry K., Shapiro H., Goltsman E., McHardy A. C., Rigoutsos I., Salamov A., Korzeniewski F., Land M., Lapidus A., Grigoriev I., Richardson P., Hugenholtz P., Kyrpides N. C.: Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat. Methods*, 4:495–500, (2007).

McHardy A. C., Martin H. G., Tsirigos A., Hugenholtz P., Rigoutsos I.: Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods*, 4:63–72, (2007).

Merkl R.: Sigi: Score-based identification of genomic islands. *BMC Bioinformatics*, 5:22, (2004).

Mulder N. J., Apweiler R., Attwood T. K., Bairoch A., Bateman A., Binns D., Bork P., Buillard V., Cerutti L., Copley R., Courcelle E., Das U., Daugherty L., Dibley M., Finn R., Fleischmann W., Gough J., Haft D., Hulo N., Hunter S., Kahn D., Kanapin A., Kejariwal A., Labarga A., Langendijk-Genevaux P. S., Lonsdale D., Lopez R., Letunic I., Madera M., Maslen J., McAnulla C., McDowall J., Mistry J., Mitchell A., Nikolskaya A. N., Orchard S., Orengo C., Petryszak R., Selengut J. D., Sigrist C. J., Thomas P. D., Valentin F., Wilson D., Wu C. H., Yeats C.: New developments in the InterPro database. *Nucleic Acids Res.*, 35:D224–228, (2007).

Niemiller M. L., Fitzpatrick B. M., Miller B. T.: Recent divergence with gene flow in Tennessee cave salamanders (Plethodontidae: Gyrinophilus) inferred from gene genealogies. *Mol. Ecol.*, 17:2258–2275, (2008).

Noble P. A., Citek R. W., Ogunseitan O. A.: Tetranucleotide frequencies in microbial genomes. *Electrophoresis*, 19:528–535, (1998).

Noble W. S.: What is a support vector machine? *Nature Biotechnology*, 24(12):1565–1567, (2005).

Okum O.: K-local hyperplane distance nearest neighbor algorithm and protein fold. *Pattern Recognition and Image Analysis*, 6:19–22, (2006).

Ontrup J., Ritter H.: Large-scale data exploration with the hierarchically growing hyperbolic som. *Neural Netw.*, 19(6):751–761, (2006).

Overbeek R., Begley T., Butler R. M., Choudhuri J. V., Chuang H.-Y., Cohoon M., de Crecy-Lagard V., Diaz N., Disz T., Edwards R., Fonstein M., Frank E. D., Gerdes S., Glass E. M., Goesmann A., Hanson A., Iwata-Reuyl D., Jensen R., Jamshidi N., Krause L., Kubal M., Larsen N., Linke B., McHardy A. C., Meyer F., Neuweger H., Olsen G., Olson R., Osterman A., Portnoy V., Pusch G. D., Rodionov D. A., Ruckert C., Steiner J., Stevens R., Thiele I., Vassieva O., Ye Y., Zagnitko O., Vonstein V.: The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res*, 33:5691–5702, (2005).

Podell S., Gaasterland T.: DarkHorse: a method for genome-wide prediction of horizontal gene transfer. *Genome Biol*, 8:R16, (2007).

Pop M., Phillippy A., Delcher A. L., Salzberg S. L.: Comparative genome assembly. *Brief. Bioinformatics*, 5:237–248, (2004).

Raes J., Harrington E. D., Singh A. H., Bork P.: Protein function space: viewing the limits or limited by our view? *Curr. Opin. Struct. Biol.*, 17:362–369, (2007).

Rappe M., Giovannoni S. J.: The uncultured microbial majority. *Annu Rev Microbiol*, 57:369–394, (2003).

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

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

Rojas R.: *Neural Networks – A systematic introduction*. Springer-Verlag, New York (1996).

Ruepp A., Graml W., Santos-Martinez M., Koretke K., Volker C., Mewes H., Frishman D., Stocker S., Lupas A., Baumeister W.: The genome sequence of the thermoacidiphilic scavender *Thermoplasma acidophilum*. *Nature*, 407:508–513, (2000).

Rusch D. B., Halpern A. L., Sutton G., Heidelberg K. B., Williamson S., Yooseph S., Wu D., Eisen J. A., Hoffman J. M., Remington K., Beeson K., Tran B., Smith H., Baden-Tillson H., Stewart C., Thorpe J., Freeman J., Andrews Pfannkoch C., Venter J. E., Li K., Kravitz S., Heidelberg J. F., Utterback T., Rogers Y. H., Falcón L. I., Souza V., Bonilla-Rosso G., Eguiarte L. E., Karl D. M., Sathyendranath S., Platt T., Bermingham E., Gallardo V., Tamayo-Castillo G., Ferrari M. R., Strausberg R. L., Nealson K., Friedman R., Frazier M., Venter J. C.: The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol.*, 5:e77, (2007).

Saha S., Heber S.: In silico prediction of yeast deletion phenotypes. *Genetics and Molecular Research*, 5:224–232, (2006).

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

Salton G., Wong A., Yang C.: A vector space model for automatic indexing. *Communications of the ACM*, 18:613–620, (1975).

Sandberg R., Winberg G., Bränden C., Kaske A., Ernberg I., Cöster J.: Capturing whole-genome characteristics in short sequences using a naïve Bayesian classifier. *Genome Res.*, 11:1404–1409, (2001).

Sanger F., Nicklen S., Coulson A. R.: DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci*, 74:5463–5467, (1997).

Schoenberg I.: Metric spaces and positive definite fucntions. *Trans. Amer. Math. Soc.*, 44:522–236, (1938).

Shendure J., Ji H.: Next-generation DNA sequencing. *Nat. Biotechnol.*, 26:1135–1145, (2008).

Shendure J. A., Porreca G. J., Church G. M.: Overview of DNA sequencing strategies. *Curr Protoc Mol Biol*, Chapter 7:Unit 7.1, (2008).

Snel B., Bork P., Huynen M. A.: Genome phylogeny based on gene content. *Nat. Genet.*, 21:108–110, (1999).

Song Y., Huang J., Zhou D., Zha H., Lee Giles C.: Iknn: Informative k-nearest neighbor pattern classification. in: Machine learning: Ecml 2007, 18th european conference on machine learning, warsaw, poland, september 17-21, 2007, proceedings. In: J. N. Kok, J. Koronacki, R. L. de Mántaras, S. Matwin, D. Mladenic, A. Skowron, eds., *ECML*, vol. 4701 of *Lecture Notes in Computer Science*, pages 248–264, Springer (2007).

Staelin C.: *Parameters selection for support vector machines*. Ph.D. thesis, Technical report, HP Laboratories Israel (2003).

Tarca A. L., Carey V. J., Chen X.-w., Romero R., Drăghici S.: Machine learning and its applications to biology. *PLoS Comput Biol*, 3(6):e116, (2007).

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

Teeling H., Waldmann J., Lombardot T., Bauer M., Glöckner F. O.: TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics*, 5:163, (2004b).

Tettelin H., Masignani V., Cieslewicz M. J., Donati C., Medini D., Ward N. L., Angiuoli S. V., Crabtree J., Jones A. L., Durkin A. S., Deboy R. T., Davidsen T. M., Mora M., Scarselli M., Margarit y Ros I., Peterson J. D., Hauser C. R., Sundaram J. P., Nelson W. C., Madupu R., Brinkac L. M., Dodson R. J., Rosovitz M. J., Sullivan S. A., Daugherty S. C., Haft D. H., Selengut J., Gwinn M. L., Zhou L., Zafar N., Khouri H., Radune D., Dimitrov G., Watkins K., O'Connor K. J., Smith S., Utterback T. R., White O., Rubens C. E., Grandi G., Madoff L. C., Kasper D. L., Telford J. L., Wessels M. R., Rappuoli R., Fraser C. M.: Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial ”pan-genome”. *Proc. Natl. Acad. Sci. U.S.A.*, 102:13950–13955, (2005).

Tran T. N., Wehrens R., Buydens L. M.: Knn-kernel density-based clustering for high-dimensional multivariate data. *Computational Statistics & Data Analysis*, 51(2):513–525, (2006).

Tringe S. G., Rubin E. M.: Metagenomics: DNA sequencing of environmental samples. *Nat. Rev. Genet.*, 6:805–814, (2005).

Tringe S. G., von Mering C., Kobayashi A., Salamov A. A., Chen K., Chang H. W., Podar M., Short J. M., Mathur E. J., Detter J. C., Bork P., Hugenholtz P., Rubin E. M.: Comparative metagenomics of microbial communities. *Science*, 308:554–557, (2005).

Turnbaugh P. J., B´ackhed F., Fulton L., Gordon J. I.: Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. *Cell Host Microbe*, 3:213–223, (2008).

Turnbaugh P. J., Hamady M., Yatsunenko T., Cantarel B. L., Duncan A., Ley R. E., Sogin M. L., Jones W. J., Roe B. A., Affourtit J. P., Egholm M., Henrissat B., Heath A. C., Knight R., Gordon J. I.: A core gut microbiome in obese and lean twins. *Nature*, 457:480–484, (2009).

Turnbaugh P. J., Ley R. E., Hamady M., Fraser-Liggett C. M., Knight R., Gordon J. I.: The human microbiome project. *Nature*, 449:804–810, (2007).

Turnbaugh P. J., Ley R. E., Mahowald M. A., Magrini V., Mardis E. R., Gordon J. I.: An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, 444:1027–1031, (2006).

Tyson G. W., Chapman J., Hugenholtz P., Allen E. E., Ram R. J., Richardson P. M., Solovyev V. V., Rubin E. M., Rokhsar D. S., Banfield J. F.: Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428:37–43, (2004).

Vapnik V.: *The nature of statistical learning theory*. Springer–Verlag, New York (1995).

Vapnik V.: *Statistical learning theory*. Wiley, New York (1998).

Venter J. C., Adams M. D., Sutton G. G., Kerlavage A. R., Smith H. O., Hunkapiller M.: Shotgun sequencing of the human genome. *Science*, 280:1540–1542, (1998).

Venter J. C., Remington K., Heidelberg J. F., Halpern A. L., Rusch D., Eisen J. A., Wu D., Paulsen I., Nelson K. E., Nelson W., Fouts D. E., Levy S., Knap A. H., Lomas M. W., Nealson K., White O., Peterson J., Hoffman J., Parsons R., Baden-Tillson H., Pfannkoch C., Rogers Y. H., Smith H. O.: Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 304:66–74, (2004).

Weinberger K. Q., Blitzer J., Saul L. K.: Distance metric learning for large margin nearest neighbor classification. In: *In NIPS*, MIT Press (2006).

Wheeler D. L., Church D. M., Lash A. E., Leipe D. D., Madden T. L., Pontius J. U., Schuler G. D., Schriml L. M., Tatusova T. A., Wagner L., Rapp B. A.: Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Res*, 30:13–16, (2002).

Whitaker R. J.: Allopatric origins of microbial species. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 361:1975–1984, (2006).

Willenbrock H., Hallin P. F., Wassenaar T. M., Ussery D. W.: Characterization of probiotic Escherichia coli isolates with a novel pan-genome microarray. *Genome Biol.*, 8:R267, (2007).

Wilson D.: Asymptotic properties of nearest neighbor rules using edited data. *IEEE Trans Sys. Man Cyberne*, 2:408–420, (1972).

Woese C. R., Kandler O., Wheelis M. L.: Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. U.S.A.*, 87:4576–4579, (1990).

Wold B., Myers R.: Sequence census methods for functional genomics. *Nature Methods*, 5:19–21, (2008).

Wolf Y. I., Brenner S. E., Bash P. A., Koonin E. V.: Distribution of protein folds in the three superkingdoms of life. *Genome Res.*, 9:17–26, (1999).

Wolf Y. I., Rogozin I. B., Grishin N. V., Koonin E. V.: Genome trees and the tree of life. *Trends Genet.*, 18:472–479, (2002).

Wolf Y. I., Rogozin I. B., Grishin N. V, Tatusov R. L., Koonin E. V.: Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol. Biol.*, 1:8, (2001).

Yang S., Doolittle R. F., Bourne P. E.: Phylogeny determined by protein domain content. *Proc. Natl. Acad. Sci. U.S.A.*, 102:373–378, (2005).

Yao Z., Ruzzo W. L.: A regression-based k nearest neighbor algorithm for gene functions prediction from heterogeneous data. *BMC Bioinformatics*, 7 Suppl I:S11, (2006).

Yooseph S., Sutton G., Rusch D. B., Halpern A. L., Williamson S. J., Remington K., Eisen J. A., Heidelberg K. B., Manning G., Li W., Jaroszewski L., Cieplak P., Miller C. S., Li H., Mashiyama S. T., Joachimiak M. P., van Belle C., Chandonia J. M., Soergel D. A., Zhai Y., Natarajan K., Lee S., Raphael B. J., Bafna V., Friedman R., Brenner S. E., Godzik A., Eisenberg D., Dixon J. E., Taylor S. S., Strausberg R. L., Frazier M., Venter J. C.: The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol.*, 5:e16, (2007).

Zhang S.-H., Ya-Zhi H.: Characteristics of oligonucleotide frequencies across genomes: Conservation versus variation, strand symmetry, and evolutionary implications. *Nature Precedings*, npre.2008.2146.1:1– 28, (2008).

Zhu M., Zhang Z., Hirdes J. P., Stolee P.: Using machine learning algorithms to guide rehabilitation planning for home care clients. *BMC Medical Informatics and Decision Making*, 7:41, (2007).