DTU Library

Determining and comparing protein function in Bacterial genome sequences

Vesth, Tammi Camilla

Publication date:
2014

Document Version
Peer reviewed version

Link back to DTU Orbit

Citation (APA):
Vesth, T. C. (2014). *Determining and comparing protein function in Bacterial genome sequences*. Technical University of Denmark.

Determining and comparing protein function in Bacterial genome sequences

Tammi Camilla Vesth

January 29, 2014

Information is not knowledge

- Albert Einstein

Contents

Preface

This thesis was prepared at the Department of Systems Biology, the Technical University of Denmark, in fulfillment of the requirements for acquiring a Ph.D degree. This thesis describes technical advances in the characterization and removal of measurement error from gene expression profiles. This work was carried out at the Center for Biological Sequence Analysis (CBS) under the supervision of David W. Ussery and Karin Lagesen (University of Oslo). The thesis consists of four background chapters and a collection of three research papers written during the period 2011-2013.

Data used for this project can be obtained from: http://cmgfunc.20gbfree.com/ and online documentation of software and approaches can be found on http://github.com/cmgvesth/cmgfunc and http://biotoolscmg.wikia.com/.

Lyngby, January 2014

Tammi Vesth

Abstract

In November 2013, there was around 21.000 different prokaryotic genomes sequenced and publicly available, and the number is growing daily with another 20.000 or more genomes expected to be sequenced and deposited by the end of 2014. An important part of the analysis of this data is the functional annotation of genes – the descriptions assigned to genes that describe the likely function of the encoded proteins. This process is limited by several factors, including the definition of a *function* which can be more or less specific as well as how many genes can actually be assigned a function based on known functions.

This thesis describes the development of new tools for comparative functional annotation and a system for comparative genomics in general. As novel sequenced genomes are becoming more readily available, there is a need for standard analysis tools. The system CMG-biotools is presented here as an example of such a system and was used to analyze a set of genomes from the *Negativicutes* class, a group of bacteria closely related to Gram positives but which has a different cell wall structure and stains Gram negative, as the name indicates. The results of this work show that genomes of this class have very little homology to other known genomes making functional annotation based on sequence similarity very difficult.

Inspired in part by this analysis, an approach for comparative functional annotation was created based public sequenced genomes, CMGfunc. Functionally related groups of proteins were clustered based on sequence domains so that each group represented a protein function. Each function was then modeled using Artificial Neural Networks (ANN) and the model was evaluated based on its ability to identify true positives and negatives, that is proteins with or without the function of the model. The models were used to annotate a number of proteins without functional annotations and predicted functions for 98% of the genes. Evaluation of the precision of the method was performed, using data from the Critical Assessment of Functional Annotation (CAFA) project, and correct predictions were made in about 60% of the cases.

This project has highlighted the difficulties and challenges in functional annotation and computational analysis of sequence data. It has provided possible solutions for creating reproducible pipelines for comparative genomics as well as constructed a number of functional models not based on sequence similarity. Although much work is still left to be done, resources are flowing into the area of sequence analysis and progress is being made every day. As such, many different approach are being tried out and tested which will, in time, improve the knowledge gained from sequencing genomes.

Dansk resume

I november 2013 var der omkring 21.000 prokaryote genom sekvenser i offentlige databaser og tallet vokser dagligt og det forventes at endnu 20.000 sekvenser vil blive publiceret i løbet af 2014. En essential del af analysen af alle disse genom sekvenser er den funktionelle annotering, den beskrivelse som forklarer hvilken funktion genet har i cellen. Denne process ebregrænses af flere forskellige faktorer, her iblandt definitionen af *funktion*, en definition der kan være mere eller mindre specifik, samt hvor mange gener der rent faktisk kan tilknyttes en kendt cellulær funktion.

Dette speciale beskriver udviklingen af nye værktøjer til funktionel annotering og sammenligning af disse annoteringer mellem flere genomer samt general sammenligning af genom sekvenser. I takt med at der foreligger flere og flere sekvenser vokser behovet for standard procedure i analysen af disse data. Systemet CMG-biotools presenteres her som et exemple på et sådant system og blev brugt i analysen af genomer fra klassen *Negativicutes*, en gruppe organismer tæt beslægtet med Gram positive organismer men har en anderledes cell væg komposition der gør dem Gram negative, som navnet også antyder. Resultatet af denne analyse viste at denne klasse har meget lidt sekvens homologi med andre sekventerede genomer, og da funktionel annotering ofte hviler på netop sekvens homologi, besværligører dette annoteringen.

Inspireret af denne problematik, startede udviklingen af an metode til funktionel annotering baseret på publiceret data, CMGfunc. Funktionelt relaterede grupper af proteiner blev defineret ud fra konserverede protein domæner således at hver gruppe representerede en protein funktion. Hver funktion blev derefter modeleret ud fra en neural netværk model og hver model blev evalueret ud fra dens even til at genkende sande positiver og negativer, det vil sige, proteiner der faktisk besad eller ikke besad den funktion some modellen representerede. Modellerne blev brug til at annotere en række protein uden kendt funktion og annoterede 98% af disse. Præcessionen af de funktionelle modeller blev estimeret ud fra data fra projektet Critical Assessment of Functional Annotation (CAFA), og in 60% af tilfældende forudsage CMGfunc modellerne den rigtige funktion.

Dette projekt har belyst mulige løsninger på udfordringerne indenfor sammenlignelig funktionel annotering og computer analyse af sekvens data. I dette arbejde er der blevet udviklet en række forslag til reproducerbare analyse metoder samt modeller til funktionel annotering. Selvom der stadig er udfordringer der skal takles indenfor disse felter tilføres store resourcer til netop denne type analyse og det er klart fremskridt i syne i takt med at mange nye metoder testes.

Publications

A number of papers were written during the course of this project. Three of these are included in this thesis while two papers and a book chapter is not. One paper is a manuscript ready for submission while the remaining are published.

Papers included

CMG-biotools, a free workbench for basic comparative microbial genomics

Vesth T, Lagesen K, Acar Ö, Ussery D. PLoS One. 2013;8(4):e60120. doi: 10.1371/journal.pone.0060120. Epub 2013 Apr 5. PMID: 23577086

Veillonella, Firmicutes: Microbes disguised as Gram negatives

Vesth, T, Özen, A, Andersen, S, Lukjançenko, O, Kaas, R, Nookaew, I, Bohlin, J, Wassenaar, T, Ussery, D. Standards in Genomic Sciences, North America, 9, dec. 2013. Available at: www.standardsingenomics.org/index.php/sigen/article/view/sigs.2981345/1064.

CMGfunc, comparative functional annotation of microbial genomics

Vesth T, Lagesen K, Ussery D. F1000research. Manuscript, not submitted

Papers not included

Comparative Genomics, book chapter

Özen A, Vesth T, Ussery DW. Book: The Prokaryotes, chapter, page 209-227. Editor: Rosenberg, Eugene and DeLong, EdwardF. and Lory, Stephen and Stackebrandt, Erko and Thompson, Fabiano. doi: 10.1007/978-3-642-30194-0_11. publisher: Springer Berlin Heidelberg.

Amino acid usage is asymmetrically biased in AT- and GC-rich microbial genomes

Bohlin J, Brynildsrud O, Vesth T, Skjerve E, Ussery DW. PLoS One. 2013 Jul 26;8(7): e69878. doi: 10.1371/journal.pone.0069878. Print 2013. PMID: 23922837.

Bayesian prediction of bacterial growth temperature range based on genome sequences

Jensen DB, Vesth TC, Hallin PF, Pedersen AG, Ussery DW. BMC Genomics. 2012;13 Suppl 7:S3. doi: 10.1186/1471-2164-13-S7-S3. Epub 2012 Dec 13. PMID: 23282160.

Acknowledgements

I first set my feet into CBS as a bachelor student in 2005 and I would like to thank all of CBS, for great education, research opportunities and support of students with a curious mind and a will to work hard.

I would like to express my heartfelt gratitude to my supervisor and mentor Prof. David Ussery. Through this project he has been my constant source of inspiration and motivation as well as smiles, fun and chocolate. As he left his position in the final stages of my project, he has continued to provide support and input when it was the most needed and I am very grateful to him. He has taught me much about being a scientist, especially on how to identify a good idea and how to collaborate to create great projects from such ideas.

Another important person in this project has been my co-supervisor Karin Lagesen. Her always keen eye and attention to detail has been the source of much joy and frustration and I am a better scientist because of her constructive criticism and feedback. I have learned much about collaboration and the benefits of it from her. I am deeply grateful to her for all her work and time and hope to get the chance to work with her again.

I would also like to extend my thanks to the researchers at the universities in Alicante, Spain and Khatmandu, Nepal, where I was given the opportunity and great honor of teaching workshops. These experiences has taught me much about the gap between bioinformaticians and biologists and how these can be addressed. They also showed me how important it is to put sequence analysis into context and base new ideas on exists challenges. I thank all my students over these three years for all they have taught me back.

I would like to thank my Ph.D colleagues at CBS, for laughs, late nights and lots of coffee! Thanks also goes to all my friend and especially to Juliet, Ida, Steven, Rachita and Aslı for input, ideas and support in times of need. A very special thanks goes to Öncel Acar, for being there through it all.

Thesis aim and structure

This project was inspired from the author's experience as a bachelor and then master student in the comparative microbial genomics group (CMG) at the Center for Biological Sequence analysis (CBS) at the Technical University of Denmark (DTU). As the number of available genome sequences increases, roughly doubling every year, questions are becoming more complex and the requirements for analysis are changing. Historically, with only a few genomes sequenced, general comparisons was the most common analysis, often with one genome per species; however, with the explosion of sequence data, new expectations included distinguishing one isolate from another and identifying why one strain of a bacteria is pathogenic while another is not. Further complicating the analysis was the observation that often the strain-specific genes have no known function, based on sequence homology with proteins of known function in the current databases. From this background came some general questions: how can protein functions be compared across many genomes? How can proteins with no sequence homology be functionally annotated?

The primary goal of this project has been to develop methods for analyzing sequences without homology to known and annotated sequences. Early on, it was necessary to consolidate many of the tools that had been developed by the Comparative Microbial Genomics research group at CBS over the years. No system for gathering these into a coherent system and working environment was available. Thus, the CMG-biotools was developed and was published in 2012 (see Section 5.1 on page 31). This system was designed to include a wide range of tools for comparative genomics of bacteria and to make these tools freely available and user-friendly for researchers with no bioinformatic background. The project was also a way of testing how computational analysis can be made reproducible and publicly available for evaluation when publishing and different aspects are taken into consideration, such as accessibility, reproducibility, speed and difficulty of use. In summary, an important part of this project has been to focus on how to create a user-friendly and computationally powerful bioinformatics pipelines.

The functional potential of bacteria can be used to group evolutionary similar organisms. The next project was carried out using a diverse set of publicly available genome sequences with focus on the class *Negativicutes*, a group of bacteria closely related to Gram positives but which has a different cell wall structure and stains Gram negative, as the name indicates. A set of 24 *Negativicutes* were compared to a wide range of other bacterial genome sequences using sequence similarity methods such as BLAST and 16S rRNA alignments as well as feature based methods such as Composition Vector Trees and DNA tetramer frequencies. The metabolic potential of each genome was analyzed using the Kyoto Encyclopedia of Genes and Genomes (KEGG) in combination with Hidden Markov Models (HMMs) (see Section 5.2 on page 48).

Construction of a pipeline for functional annotation of bacterial proteins is a major part of this thesis. This involved evaluation of available public data, construction of a functional scheme, clustering of functionally related proteins and modeling of functions using Artificial Neural Networks (ANNs). The aim of this work was to use a set of defined functions with a controlled vocabulary to investigate the patterns found in large sets of data, and to model these functions. Data was collected from public sources and as much data as possible was used in the further modeling. This was of importance in order to ensure annotation of less common sequences, a shortcoming of many sequence based methods. Proteins were clustered based on shared functional or structural domains as obtained from the Pfam-A database. Each cluster represent a function described by its domain descriptions, GO terms and Pfam clan descriptions. Each sequence was translated into a number of sequence and pattern based features, including chemical properties, signaling patterns and secondary structure. The aim of this process was to add information to the domain identification already established in the clustering, possibly improving the coverage of the modeling. The features of each sequence in each functional cluster were used to train and test an ANN model for each function. This work is presented in the manuscript in Section 5.3 on page 67.

This thesis begins with an introduction (Chapter 1), describing the field of computational functional annotation, its challenges, advances and approaches. The concept of *function* is discussed leading to the next chapter (Chapter 2) where ways of describing a proteins function are presented. The section includes the detection of genes in DNA sequences, protein domain models, proteins features such as biochemical properties and chemical composition and gene ontologies. Chapter 3 highlights different mathematical methods in prediction of function with a focus on Artificial Neural Networks (ANN). A short discussion on data sharing, management and distribution of work and research in bioinformatics is presented in Chapter 4. Chapter 5 includes three manuscripts published during this project illustrating different aspects in comparative genomics, functional annotation and data handling in a the new Genomic Era. The final chapter (Chapter 6) presents concluding remarks on the presented work as well as future perspectives.

1 Introduction

The field of genomics has undergone tremendous changes since GenBank was first established in 1982. In the early 1980s, it would have taken more than a thousand years to sequence the DNA (deoxyribonucleic acid) of an *Escherichia coli* genome, and it would have taken several million years to sequence a human genome. Thus at that time, due to the unreasonable time limit, sequencing the human genome was not considered possible. In 1984, an effort was made to solve this problem when the U.S. Department of Energy decided to invest $200 million per year, over 20 years, to increase the speed of DNA sequencing. This was more money than previously spent on the Apollo space program, all going to research to improve DNA sequencing speed. After only 10 years of investment, the speed of sequencing had improved sufficiently to allow for the first two bacterial genome sequences to be finished and published; *Haemophilus influenzae* [1] and *Mycoplasma genetalium* [2], and a mere five years later, a draft of the human genome was published [3]. The speed of DNA sequencing has continued to increase over the past ten years, such that sequencing a bacterial genome is now both fast and inexpensive. As of November 2013, there are about 21.000 different prokaryotic genomes sequenced and publicly available, and the number is growing daily with more than 10.000 bacterial genomes expected to be sequenced and deposited in public databases through 2014. Some projects focus on exploring the diversity of bacteria, examples are the Human Microbiome project[1] and the Earth Microbiome project[2]. However, many projects still focus on sequencing the same organism over and over again, in order to infer functional and evolutionary knowledge; examples of this include the 1.000 human genomes project[3] or the Microbial Genome Program of U.S. Department of Energy[4] [4]. Complete genome sequences are available from a number of online sources including NCBI GenBank [5], UniProt [6, 7] and the The Genomes On Line Database [8]. These advances and initiatives will greatly increase the amount of sequence data in the immediate future and though the results of automated functional annotation systems should preferably be subject

[1] http://commonfund.nih.gov/hmp/index

[2] http://earthmicrobiome.org

[3] http://1000genomes.org

[4] http://microbialgenomics.energy.gov

to some level of human review, the development of tools and standards in the field will greatly help in setting up reasonable hypothesis for protein functions.

1.1 Functional annotation

Genome annotation can broadly be defined as the extraction of biological knowledge from DNA sequences. Most of the DNA (roughly 90% or more) in a bacterial genome encodes proteins, and it is the *function* of proteins which defines much of the activity of an organism, either directly (such as enzymes) or indirectly (*e.g.*, structural proteins). A major part of the process is first locating the genes - that is, the prediction of protein encoding genes, as well as small non-coding RNAs (ribonucleic acid), tRNAs, rRNAs and repeats. Once these genes have been identified, then their sequences are used as input for functional annotation. Functional annotation has become a greater challenge as sequencing technology has moved the emphasis of computational biology away from data production to data analysis. With the sequence of each new genome, it has been estimated that anywhere from 500 to 1.000 new genes of unknown function will be detected [9]; thus the 10.000 bacterial genomes sequenced in the year 2014 will result in 5 to 10 million new protein sequences deposited in the public databases, with no known function. The lack of knowledge about protein function and structure greatly limits the knowledge gained from high-throughput sequencing technologies and presses the bioinformatics field to develop new methods and standard procedures for functional annotation. Furthermore, shortcomings in automated systems to reliably replace manual curation of single genes and the difficulties in combining available scientific data into meaningful functional descriptions is holding back the generalization of genome annotation. Another limitation in functional annotation is the accumulation of faulty annotations in databases still being used [9]. The main reason for this is the commonly used tactic of relying on detecting sequence similarities to already annotated genes when assigning a function to a newly sequenced gene. This process is sometimes described as ''guilt-by-association'' and involves the transfer of annotation between genes with evidence of sequence similarity. In trying to annotate a single gene using sequence similarity, often the best match will be to a *putative*, *probable*, *unknown* or *uncharacterized* function, yielding the search unhelpful in further investigation [10, 11]. Collectively, these issues represent the challenges and current status of large scale and comparable functional annotation, and these must be addressed in order to gain the most information from sequencing data.

An often overlooked issue in annotation is the actual definition of the concept *function*. As Freidberg described in 2006, the definition of a function varies based on the context in which it is used [12]. Another way of looking at the ambiguity of the protein function concept was laid forth by Galperin in 2010 [13]. He described the problem as a matter of defining the word *understand*. As the aim of sequencing DNA is to get greater understanding of biology, the process of assigning a function to a gene or protein, becomes a question of what we understand by the word *function*. The rest of this chapter will discuss the concept of protein function as well as different levels of function.

1.2 Protein function

Several different systems for functionally annotating and grouping proteins have been suggested. In 1993, Riley *et al.* proposed a six level system for functional annotation of *Escherichia coli* [14], with functional groups like *Intermediary metabolism*, *Cellular processes* and *Cell structure*. This system had the advantage of being very high level and was used in other contexts for both bacteria and eukaryotes [15, 16]; however, this method was developed before the first bacterial genome was sequenced. In contrast, the Clusters of Orthologous Groups of proteins (COGs) is a database of protein groups made from protein sequences encoded in complete genomes [17]. Each COG consists of individual proteins from at least three lineages and attempts to model domains conserved across evolution. COGs are organized into functional categories such as *RNA processing and modification* and *Carbohydrate metabolism and transport* with fifteen different levels of function. Unfortunately, COGs is no longer being updated. In spite of this, COGs has been used in different contexts [18], such as in the MeGa system where is was used for functional annotation in combination with InterPro [19].

The Enzyme Commission was set up by the International Congress of Biochemistry in 1955 and constructed a standardized system for enzymes, based on the chemical reactions they are involved in[5]). The EC numbers have six higher levels with examples such as *Oxidoreductase reactions*, *Transferase reactions* and *Isomerase reactions*. The system does not cover all protein functions but does illustrate how, early on, there was a focus on a standardized system for functions with identifiers as well as curated descriptions. Building on the same idea as EC, the Gene Ontology Consortium[6] was established in 2000 [20] and presented an

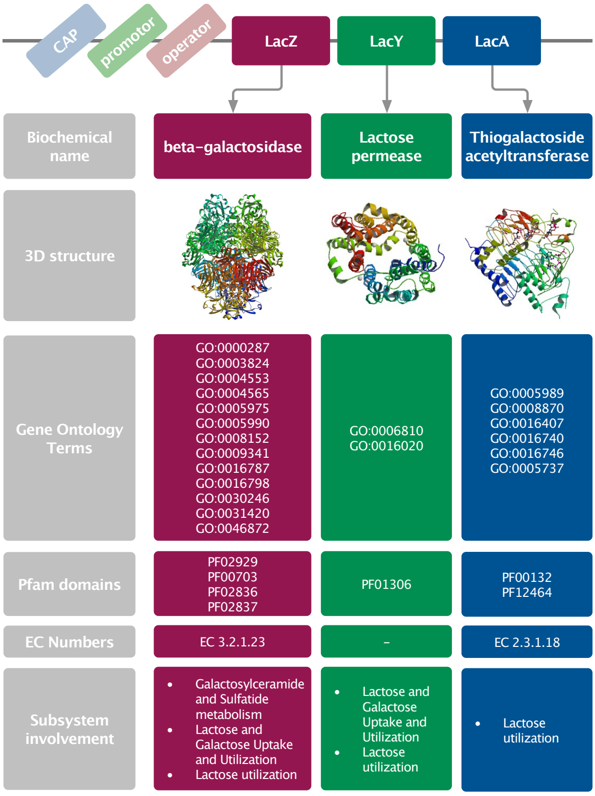
[5] http://www.chem.qmul.ac.uk/iubmb/enzyme/rules.html

[6] http://www.geneontology.org

even more elaborate system of functional levels and descriptions. The consortium was originally launched as a collaborative project between three eukaryotic model organism databases, but has since expanded to include many microbial data sources as well [20, 21]. The ontology consists of three structured, controlled vocabularies (ontologies) of functional descriptions (GO terms) as well as unique identifiers constructed through manual annotation and combines data from several databases and scientific literature. The different vocabularies cover three aspects of gene product function: molecular function, biological process and cellular components. The descriptions in GO are organized in a relational manner with a "child-parent" relationships between different terms and this hierarchy allows for annotation at varying levels of specificity. The system is currently the dominant approach for computational work in functional annotation and is widely used in annotation pipelines [22, 23, 24]. Systems like GO and EC are invaluable in the automatic prediction of function but the "guilt-by-association" method often employed in the use of these systems continues to add questionable predictions to the annotation pool [25, 26]. These different annotation systems illustrate the difficulties in putting general terms and words to functions that were historically analyzed and documented by laboratory experiments.

Another aspect of function is the context in which it is used. One use of *protein function* is the inference of evolutionary relationships between organisms. Accumulated sequence mutations have historically been used to investigated evolutionary relationships, like using DNA sequences such as 16S rRNA, but in the field of phylogenomics, conserved functions are used to infer these relationships [27, 28]. In this field, comparative and automatic functional annotation is crucial as evolutionary relationships are determined based on shared functions which can only be accomplished when the annotation is standardized across genomes.

One example of a cellular function description in *Escherichia coli* K12 is the ability to break down lactose. This *function* (using lactose rather than glucose or another sugar as an energy source) is actually a combination of actions, requiring the activation of a set of genes, including an enzyme to cleave the lactose sugar, and a transporter to bring in more lactose. In *E. coli*, this function is the result of a classic example of the operon structure, first proposed by François Jacob and Jacques Monod more than 50 years ago [29]. The *lac* operon contains genes encoding three structural proteins, and the genes are arranged in a series which are co-transcribed [30]. Such an operon represents an overall *function* (use of lactose and hence the name *lac* operon), although each of the three proteins encoded in the operon have their own function, that can be described separately



	LacZ	LacY	LacA
Biochemical name	beta-galactosidase	Lactose permease	Thiogalactoside acetyltransferase
3D structure			
Gene Ontology Terms	GO:0000287 GO:0003824 GO:0004553 GO:0004565 GO:0005975 GO:0005990 GO:0008152 GO:0009341 GO:0016787 GO:0016798 GO:0030246 GO:0031420 GO:0046872	GO:0006810 GO:0016020	GO:0005989 GO:0008870 GO:0016407 GO:0016740 GO:0016746 GO:0005737
Pfam domains	PF02929 PF00703 PF02836 PF02837	PF01306	PF00132 PF12464
EC Numbers	EC 3.2.1.23	-	EC 2.3.1.18
Subsystem involvement	• Galactosylceramide and Sulfatide metabolism • Lactose and Galactose Uptake and Utilization • Lactose utilization	• Lactose and Galactose Uptake and Utilization • Lactose utilization	• Lactose utilization

Figure 1.1: *Escherichia coli* str. K-12 substr. MG1655 *lac* operon annotations on different levels. Each annotation type covers a different way of describing the functions of the gene. Annotations were obtained from the PDB, UniProt and SEED databases.

[26]. Figure 1.1 shows the operon from the *Escherichia coli* K-12 MG1655 genome, and the different levels of annotation given by Gene Ontology, Pfam and the Protein Data Bank (Uniprot P00722: LacZ[7,8], Uniprot C9QQT5: LacY[9,10], Uniprot

[7] http://www.pdb.org/pdb/images/1jz2_bio_r_500.jpg?bioNum=1

[8] http://pubseed.theseed.org/?page=Annotation&feature=fig|83333.1.peg.341

[9] http://www.pdb.org/pdb/images/1KQA_bio_r_250.jpg?bioNum=1

[10] http://pubseed.theseed.org/?page=Annotation&feature=fig|83333.1.peg.340

P07464: LacA[11,12]). The beta-galactosidase protein contains four Pfam-A functional domains: a sugar binding domain (PF02837), a TIM barrel (PF02836), the small chain of dimeric beta-galactosidases (PF02929) and a domain family of glycosyl hydrolases (PF00703). Using the search domain architecture function on the Pfam-A web page[13] reveals that although this domain combination is quite common (1.877 sequences in Pfam-A[14] the more general combination of three domains - sugar binding, barrel and glycosyl hydrolase is more common (3.034 sequences in Pfam-A[15]). The other two proteins in the *lac* operon, shown in Figure 1.1 on the preceding page also contain functional domains, and it is the combination of functional domains which determine the functional capabilities. Another example of function and domain relationships is the RuBisCo enzyme, one of the most abundant enzyme in the world [31, 32]. The enzyme catalyzes the reaction in the Calvin-Benson (CB) cycle, where the CO_2 is condensed with a five-carbon molecule [33]. So far, four different domain configurations exist for RuBisCo. Each form has a different structure combining three different Pfam domains[16]). The change in the number and size of subunits affects the catalytic activity, however the function remains the same [33]. This multi-functionality is especially difficult capture in the automatic annotation as there is no manual decision making on which function is right or wrong for the current interest area. If *function* depends on which question is being asked, then evaluating whether the correct function is being predicted is difficult. These kind of problems require systems like GO which allow for multiple functional descriptions for one protein and describes different aspects of the functions such as where the protein works and in which pathway it is involved.

Though the link between protein sequence alignment and the function of a reference protein is a fact, predicting function solely on sequence can be quite unreliable [34]. A property of proteins that when known significantly strengthens the prediction of function is the structure, which is obviously connected to protein function [35, 31]. While similar sequences produce similar structures, similar structures might also be formed from very different sequences [35]. High resolution protein structures are traditionally obtained from experimental methods with purified proteins. Obtaining structures through such methods as crystallography and NMR are labour intensive and time consuming. However, predicting pro-

[11] http://www.pdb.org/pdb/images/1jz2_bio_r_500.jpg?bioNum=1

[12] http://pubseed.theseed.org/?page=Annotation&feature=fig|83333.1.peg.339

[13] http://pfam.sanger.ac.uk/search#tabview=tab3

[14] http://pfam.sanger.ac.uk/protein/BGAL_ECO57

[15] http://pfam.sanger.ac.uk/protein/Q19U06_9LACO

[16] http://pfam.sanger.ac.uk/family/PF00016

tein structure computationally has proved to be equally difficult [36]. Methods for predicting secondary structure include PSIPRED [37] and CPHmodels [38], while methods for predicting local structural components include the prediction of transmembrane helices (TMHMM [39]) and coiled-coils (COILS [40]). Nevertheless, combining structure and sequence based methods seem a promising approach for improving computational functional annotation [36].

1.3 Comparative genomics

The advances in computational predictions of functions have been spurred on by the increased amount of sequence data, driven by technological advances, however, further advances rely on the sequencing of multiple genomes of the same or very similar organisms. The field of comparative genomics has shed new light on the relationship between sequence and function. When comparing the genes or proteins of similar organisms, differences and similarities highlight which parts of the genomes are essential for the functions observed. As DNA sequencing and computational methods yield significantly less functional information than could be obtained by traditional experimental work, new information relies on detecting the same pattern many times in many organisms.

Comparing sequences has also given rise to a number of databases and resources containing functionally annotated groups of proteins found across multiple genomes or even distinct genera [41, 42]. The HAMAP (High-quality Automated and Manual Annotation of Microbial Proteomes) [43] resource is such a collection, where protein families from many genomes are collected and manually annotated based on strict similarity. This type of annotation has the ability to describe well-characterized clusters of proteins as well as clusters of commonly identified proteins without known functions. By comparing many genomes it is also possible to identify proteins needed for specific pathways, as the same process might occur in different organisms using a slightly different set of proteins and as such a more general model for the construction of that pathway can be build.

Some functions are expected to be essential for microbial life; this includes genes involved in replication and transcription as well as specific enzymes for essential chemical processes. Though work is still being done trying to find computational methods for identifying such a set of universal genes [44], some functions are proving to be more abundant than others with transposases being the most abundant [31]. When working towards automated functional annotation it is important to keep these biases in mind as mathematical models often assume a randomness which is not found in nature.

1.4 Resources

Many databases are available for obtaining complete genome sequences and individual protein sequences (SWISS-PROT [45], NCBI GenBank and RefSeq [5, 46], DDBJ [47], HAMAP [43]), protein domains (Pfam [48], InterPro [49], Prosite [50], SMART [42]), clusters (COG [17], TIGRFAMs [51], PANTHER [52]), structures(SCOP [53], SeqHound [54]) and many other types of information. Many of these are self-contained and not compatable with others, while other systems integrate the information from several sources with different levels of manual inspection and curation. These resources offer a wide range of information for further analysis and modeling, making them an important addition in the field of computational biology. A process in getting to the full understanding of genetics will commonly include the analysis of information from these databases.

In summary, what is *protein function* exactly? Based on the scientific publications and the variations in annotation systems and upper level functional categories, function clearly depends on the context in which it is used. On the other hand, computational analysis of proteomes continues to find orthologous groups of proteins conserved across many genera, and higher levels. Thus, functional descriptions can be targeted towards these proteins.

2 Describing proteins

The Oxford dictionary defines a protein to be "*any of a class of nitrogenous organic compounds which have large molecules composed of one or more long chains of amino acids and are an essential part of all living organisms, especially as structural components of body tissues such as muscle, hair, etc., and as enzymes and antibodies*".

Although this definition serves its purpose in many discussions, the description is not sufficient in the area of biochemistry and microbiology. A protein, in these fields, can be described by any number of chemical, structural, and physical parameters, as well as its influence on larger reactions, pathways and other proteins. In fact, proteins were described this way up until recently when sequencing became a tool in biochemistry. A protein, in bioinformatics, is the sequence of amino acids encoded by a gene - a deoxyribonucleic acid (DNA) sequence in a genome (this can be chromosomal, viral, or plasmid). This definition adds additional insecurity as it involves the identification of a gene sequence in a larger DNA sequence, a process that has its own errors and pitfalls. Once a gene has been identified, the translation must be considered, as some organisms use different translation tables than others, and there can be more than one place where the protein encoding sequence can start and also end. The way proteins are used in bioinformatics often requires more knowledge about the molecule than just the sequence. Depending on the questions at hand, biochemical, structural and interaction information might be required. Common topics studied using genome sequences include: detection of genes involved in pathogenicity or disease [55, 56], detecting resistance genes [57], identification of mutations resulting in specific phenotypes [58] and determining species, genera or other group specific genes [59]. Each of these examples illustrates how a single gene or protein can be used and described in a different way depending on the situation, a description different from the actual sequence of the protein.

This project uses a set of pipelines and parameters to describe and cluster protein sequences as accurately as possible. In an attempt to create a functional annotation level that can be compared across several genomes, the first step is to group the functionally related protein sequences and gather biologically useful descriptions and quantifiable measures for their comparison and modeling. Hence the following sections will describe a number of different ways in which a protein can be characterized and defined. The following sections describe a number of different ways in which a protein can be characterized and defined.

2.1 Gene finding

Before any type of functional annotation can be assigned to a gene, the gene must first be identified. This makes gene finding an essential part of the analysis of proteins from DNA sequencing. Different approaches have been employed to accomplish this task resulting in a range of algorithms including Glimmer [60], GenemarkHMM [61], Prodigal (PROkaryotic DYnamic programming Gene-finding ALgorithm) [62], Easygene [63] and Multivariate Entropy Distance (MED) [64]. Using different statistical models, such as Hidden Markov Models (HMMs), dynamic programming and entropy density profile models, they all rely on the modeling of gene related signals such as the Pribnow box, Shine-Dalgarno sequence, transcription factor binding sites, start (*ATG*, *GTG*, or *TTG*) and stop (*TAG*, *TAA* or *TGA*) codons and codon potential. This approach is highly useful for bacterial genomes as their coding density is very high, around 90% [60], making the most limiting factor in gene prediction the identification of the correct reading frame. Although most current gene prediction methods work relatively well in genomes of low GC content, high GC genomes contain fewer stop codons and more false open reading frames [62]. Because of this, a common mistake is the prediction of too many genes. Furthermore, longer open reading frames in high GC genomes contain more potential start codons creating a drop in accuracy of the translation initiation site [62]. This work will not compare the performance of these methods but will explain the method used in this work.

The gene finder used in this work is Prodigal version 2.0 (March 2010)[62]. The program was designed for prokaryotic genes, that is, bacterial and archaeal genes. The algorithm was tested on the experimentally verified Ecogene dataset and correctly identified the 3' end of every single gene (excluding intron containing genes). Another feature of this genefinder, especially valuable when working with translated genes, is its accuracy at predicting translation initiation sites (96% of the 5' ends in the Ecogene data set) and its low false positive rate, usually below 5%. Last but not least, Prodigal is easy to use and is published under the General Public License (GPL).

The genome sequences used in this project were all obtained from National Center for Biotechnology Information (NCBI) GenBank and some contained unknown bases. These bases are a product of sequencing, where the actual base cannot be identified but the length of the DNA stretch can still be established. In these cases, a base can be represented by a number of unknown letters; X or N is used to describe a completely unknown base (G, A, T or C) while other letters such as the letter B means C, G or T, and the letter M can be either an A or C. The ambiguous alphabet includes all possible 2 and 3 base combinations. Prodigal includes an option that ignores reading frames across unknown DNA bases[1] (option $-m$) and this was used in this project. The advantage of this choice is the higher reliability of the genes obtained, as they will be completely described by their sequence. The drawback is that some proteins might be left out. In this project, protein sequences are used to describe the function of a protein, making the actual sequence very important. For this reason, it was decided to not include genes with ambiguous bases.

2.2 Sequence features

Recent developments have explored the option of classifying and predicting protein function independently of sequence or structural alignments [65, 66, 67]. Instead of making predictions based on actual amino acid sequence similarities, these approaches use various *sequence features* to predict protein function or to cluster proteins. Such features include parameters such as protein length, molecular weight, number of atoms, amino acid composition, predicted secondary structures, subcellular location, sequence motifs or highly conserved regions. Feature based prediction was used as the backbone in a method presented by Lee *et al.* using support vector machines and random forests to predict different protein functions [68]. In this approach, 484 features were used and features were selected independently for different functions. The approach suggested a number of new features and proved highly successful in categorizing proteins into 11 different functional classes (94-100%).

Sequence features is a term which here is used as any parameter which describes a biochemical, structural or component of a protein sequence. This way of describing a protein offers the opportunity to compare very different protein sequences by only looking at specific characteristics - features - in them. In this work, the features used include a number of simple amino acid calculations, like percentage of charged and aromatic amino acids as well as mathematical estimations of physical characteristics like the extinction coefficient, which indicates how

[1] http://code.google.com/p/prodigal/source/browse/README

much light a protein absorbs at a certain wavelength. The simpler features (such as length and molecular weight of the protein) were calculated using the ExPaSy ProtParam [69] pipeline (implemented in biopython[2]). Other features were calculated using more sophisticated models, such as Psort [70], which identifies signal peptides and estimates the cellular location of the protein. Another set of features were calculated using SignalP [71] which predicts the presence and location of signal peptide cleavage sites in a amino acid sequence. The last set of features came from the SEG procedure [72], identifying high and low complexity regions in amino acid sequence. Figure 2.1 shows how information flows from amino acid sequences into numerical parameters and are normalized by passing a protein sequence through a number of prediction programs and models to yield a feature description of the protein sequence.

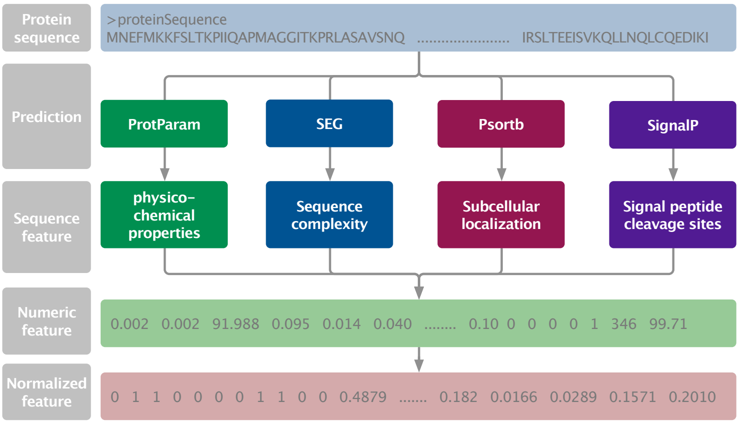


Figure 2.1: Sequence feature flow. Protein sequences, through series of feature calculation programs, numeric vector and normalized vector.

The feature were selected based on their supposed connection with protein function, their speed in calculation and numeric properties. They include both high and low level features, that is parameters calculated using high-level models and motifs as well as more simple counts and measures. The ProtParam features include a number of simple counts, like amino acids, but also highlights potential structures based on amino acids and half life prediction. Structure is known to

[2] http://biopython.org/w/index.php?title=ProtParam

be connected to functional properties and the amount of time a protein is available after being produced also seems likely to have an influence on function. The SignalP properties uses artificial neural networks to predict signal peptides in protein sequences. These signals are involved in moving proteins around to different organelles in the cell, a property that could very well be related to the function a protein serves. While SignalP returns value related to the potential of signal peptides in a sequence, Psortb is used to estimate where the protein will be transported. The features of both Psortb and SignalP are calculated for Gram negatives and positives separately as the organelles structure of the two cell types are different. Biological protein sequences are very different from random strings of amino acids [72] and contain repeated sequences or clusters of specific amino acids. A measure of such patterns adds to the description of a sequence by defining how much of the sequence will be important or unimportant for folding and function prediction, as random sequences are more likely to be unimportant.

2.3 Sequence domains - Pfam

Protein domains have long been of interest in the field of functional annotation. The reason for this interest partly stems from the functional nature of proteins. The key to a protein's function is often to be found not in the complete amino acid sequence, but in a part, a domain, of the sequence. The domain might serve as a specific binding site, or create a specific secondary or tertiary structure essential for the proteins function. These characteristic domains have been used to build models of functional or structural domains of proteins. These models make it possible to compare essential sections of a protein, excluding sequence that might add noise to a global alignment but serves no actual function. The construction of such models have been done by a large number of database resources and are rapidly becoming more used in protein clustering and functional annotation.

Protein motifs and domains identified by various methods are made accessible in a variety of data collections. Most of these databases organize proteins into families according to their motifs and domains. Important examples include PFAM [73], Superfamily [74], TIGRFAMs [75], Prosite [50], SMART 7 (Simple Modular Architecture Research Tool) [42] and PANTHER (Protein Annotation Through Evolutionary Relationship) [52]. Pfam, Superfamily, PANTHER, SMART and TIGRFAM are based on Hidden Markov Models (HMMs) while Prosite consists of weight matrices and short regular expressions corresponding to functionally or structurally important residues.

Hidden Markov Models (HMMs) are widely used in bioinformatics for identifying sequence patterns. A hidden Markov model is a statistical model used to estimate the probability of a specific pattern, here a sequence, given the model. The most straightforward way of identifying a sequence pattern is to identify an exact match, as when using the "Find" function in a text editor. This approach has the drawback of identifying only the exact match and does not allow for any deviation. If you search for $ATCGTGA$ the search will only return that one pattern. A slightly more general pattern can be constructed using regular expressions, which allow for a specific position in a pattern to take on several values, say $AT[CG]GTGA$, indicating that position 3 could be C or G. This approach can identify a slightly wider range of patterns than the exact match search. At this point, it becomes relevant to talk about how, a search pattern is identified in bioinformatics. Although exact match searches can be useful in the analysis of biological sequences, most often the patterns have too much variation to be identified in this manner. Instead, multiple sequences with a desired pattern are compared using multiple alignments and the resulting alignment describes the pattern on a position to position basis. This approach indicates positions in the pattern which are always conserved, some that can deviate to some extent and some that have very little importance. When the patterns start taking this kind of complexity, models are needed with a higher complexity level than regular expressions. Such patterns can be described using Position Specific Scoring Matrices (PSSMs) which, based on a multiple alignment, makes a probability distribution of each position in the pattern. Although these models can describe parts of a pattern as irrelevant, they are not good at detecting insertions and/or deletions in the patterns. For this purpose, HMMs are used. On top of the probabilities of specific letters at specific positions, an HMM can also detect patterns which have been interrupted by inserted letters.

The HMM consists of two processes, namely, an invisible process of hidden *states* and a visible process of observations. The hidden states form a Markov chain, and the probability of the observation depends on the underlying state. Modeling observations in one visible and another invisible layer, is commonly used for many problems dealing with classifying observations into groups. Using handwriting recognition as an example, the interest is to predict the desired letter from the written structure. The model tries to recognize different parts of the letter (states) that make up the entire letter (the observation) [76]. Since the actual written letter varies substantially, the actual letter cannot be observed and must be predicted from the written representation [77]. HMMs have traditionally been used for problems of pattern recognition, such as handwriting, speech recognition and in bioinformatics to model specific sequence patterns. The method is a mathematical approach which can be used to solve certain types of problems:

A) given the model, find the probability of the observations
B) given the model and the observations, find the state transition path; or
C) maximize either A or B by adjusting the models parameters.

Using HMMs to locate sequence patterns is a problem of type A. The Pfam database is widely used in many areas of bioinformatics and are part of several annotation pipelines including the Institute for Genome Sciences (IGS) Standard Operating Procedure for Automated Prokaryotic Annotation [22] and the J. Craig Venter Institute (JCVI) standard operating procedure for annotating prokaryotic metagenomic shotgun sequencing data [24]. Systems like FACT, functional annotation transfer between proteins with similar feature architectures [12] and CD, Conserved Domain Database for the functional annotation of proteins [41] use Pfam for clustering proteins according to functions and structures.

In the work presented here, Pfam is used for domain identification, protein clustering and description. Pfam is a collection of multiple sequence alignments and profile hidden Markov models (HMMs) and each HMM represents a protein family or domain. Pfam consists of two different databases, Pfam-A and Pfam-B. Pfam-A is a collection of protein family alignments which are constructed semi-automatically using HMMs. Sequences that are not covered by Pfam-A are clustered and aligned automatically, and included in Pfam-B. These families have no annotation or reference and their alignments are not manually checked by a curator. Pfam-A families have permanent accession numbers and contain functional annotation and cross-references to other databases, while Pfam-B families are regenerated at each release and are unannotated. Pfam-A (version 26.0, November 2011) contains a total of 13.672 domains constructed from 12.650.879 sequences obtained from SWISS-Prot and SP-TrEMBL and Pfam-A covers 77% of NCBI data (11.087.249 sequences).

A Pfam-A family HMM is based on a subset of proteins belonging to that group, the full set is called the *full alignment* while the subset is referred to as the *seed alignment* [78]. Initially, protein families were manually constructed from several data sources including Swissprot, Prosite, BLAST results or published alignments. A selection of sequences from each family was used to build a seed alignment for each group and a HMM was constructed (using HMMER3 software) and used to search SwissProt. If the search returned all the members in the initial set, a full alignment was constructed and evaluated. The alignment would have to conserve all the known features for the initial protein family to be considered well performing. If the full alignment was found good both seed and full were stored for the family. If the seed HMM did not locate all the initial members, the missing sequences were added to the seed to make sure that the HMM would pick up those sequences as well. Maintaining Pfam includes the update of full alignments by automatically collecting sequences which match the HMM with a set threshold.

Some Pfam-A families represent structural domains, like a helix-turn-helix motif, while others describe a binding site for a specific molecule, like a cellulose binding domain. It is therefore not surprising that one protein may contain several Pfam domains. The function of a protein is in our work determined by the combination of domains it contains, this combination is here called an *architecture*. The Pfam database also describes the concept of an architecture, and defines it as the exact order of domains found in a protein. In this work, this definition was changed slightly, disregarding the order and looking only at the presence or absence of a domain in a protein. Another possibility for domain combinations, is the repetition of a domain in a sequence; again, these duplicates are included in the Pfam definition, but excluded in this work. Figure 2.2 shows the definition of an architecture as defined for this work.

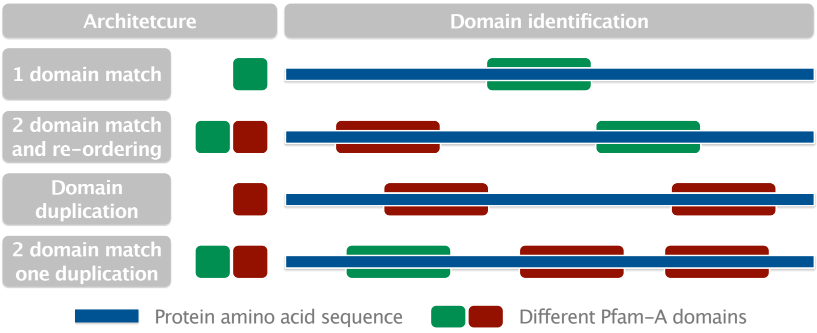


Figure 2.2: Different types of domain architectures as defined in this work.

Pfam-A contains manually curated and verified data which offers a set of functional domains with a very high information level. This information has made it possible to create Pfam *clans*. Clans are defined as domains likely to share a common origin, with evidence in tertiary structure or common sequence motifs. As such, these domains are separate sequence patterns, but likely serve similar functions. Although Pfam does offer clan HMMs, these were not used in this work; instead Pfam clans were used to further classify proteins after the prediction of Pfam domains. Architectures were connected to clans by single linkage. Thus, if a domain in the architecture could be connected to the clan the architecture was connected to that clan. When an architecture could be connected to several clans, the architecture was assigned to both clans. The Pfam domains and clans were used in this project for its high level information quality due to manual curation as well as connections to other databases and short but precise functional descriptions, useful for biological analysis though not essential for computer analysis of annotations.

2.4 Ontologies - Gene Ontology

The amounts of data being collected due to improved sequence technology has put pressure on developing standardized methods and terms to describe the data. Large amounts of data require automatic information retrieval and merging to be of use, and sets of large data must be described in similar ways in order to compare one set to another. This is not only important in terms of extrapolating biological information from sequences but also for evaluating the sequencing technologies. Sources like InterPro [49] and UniProt [7] try to incorporate data from many different sources, creating a nuanced and comprehensive picture of each sequence in the database. The resource can be useful for individual gene investigation, but fall short when comparing many proteins. The Gene Ontology (GO) system [20] is a widely used option for comparing functional categories and is used by several annotation pipelines (IGS [22], SIFTER [79], DOE-JGI [23], JCVI [24]) as well as being part of both InterPro and UniProt entries.

The structure of GO is a type of graph and each GO term is connected to other terms through a relationship. The relationships are directed, for example, a mitochondrion "is an" organelle, but an organelle is not a mitochondrion. The GO structure does not allow for cyclic relationships and is therefore called an acyclic graph. Like a hierarchy, each *child* term is more specific and *parent* terms less precise, but unlike a hierarchy, a term may have more than one parent term. Another key feature of GO is its description of function as having three distinct aspects, incorporated in three different ontologies: cellular component, molecular function, and biological process. Molecular function terms represent activities, rather than molecules or complexes, that perform the actions, and do not specify where the action takes place. Examples of individual molecular function terms are the broad concept *kinase activity* and the more specific *6-phosphofructokinase activity*, which represents a subtype of kinase activity. Molecular function terms are often the sort of functional groupings that molecular biologists refer to as the *function of a protein*. In contrast, the Biological Process (BP) ontology describes biological operations carried out by one or more of molecular functions. High-level processes such as *cell death* can have both subtypes, such as *apoptosis*, and subprocesses, such as *apoptotic chromosome condensation*. The final category, cellular component, is not a function in the common sense, but is highly related to the function of a protein and describes location, at the level of sub-cellular structures and macromolecular complexes. Examples include *nuclear inner membrane*, with the synonym *inner envelope*. Tools like BLAST2GO [80] offers a graphical interface to the GO and can both be used to assign GO terms and to create graphs and charts of different levels of terms. Systems like GO are useful because they allow for comparisons and for computational analysis of data. While a human is capable of understanding how two different descriptions cover the same function, computers are not capable of making such inferences.

Assigning functions according to the Gene ontology (GO) has the drawback of being difficult to generalize because of the graph structure of the GO system, called a directed acyclic graph (DAG). This means that any term might have multiple parent terms and cannot be generalized to one *upper* level function. This has given rise to different approaches for comparing terms, as one specific term can be assigned two different upper level terms (parents). Another problem with using GO for prokaryotic annotation is that the system was originally and is still, maintained by eukaryotic experts. This might bias the GO terms and overlook pathways and systems specific to or more frequently found in prokaryotes. While the systematic nature and structure of GO is on one hand an advantage, it is also the drawback of the system as manually literature derived keywords hold more biological information than GO terms [81].

The use of GO in this project was based on several considerations. First and foremost, using GO terms as annotations made it possible to automate the procedure of comparing predictions to known functions. Another consideration is that users of an annotation system will want to compare the system to other systems themselves, as such, the user might desire a comparable result base. Not all domains can be described with GO, therefore the combination of using GO terms and Pfam domain descriptions where no GO term is available might offer a greater coverage of protein annotation.

3 Predicting protein function

Computational prediction of protein function has developed dramatically in the last few years. Current methods include approaches relying on identification of similarity between sequence and/or structure or similarity to defined protein families. Although many variations of annotation system exists, they all employ some version of these setups.

The Institute for Genome Sciences (IGS) system [22] uses a wide range of predictive tools to annotate genomes. The first step includes identification of protein coding and non-coding genes using Glimmer3 and non-coding RNA sequences are identified using RNAmmer [82]. The protein coding genes are then compared to UniRef100 using BLASTx [83] followed by comparisons to TIGRfams [51] and Pfam [48] domain databases. Then follows detection of motifs using SignalP [71], LipoP [84], TMHMM [39] and PROSITE [50]. A number of selection criteria specifies how each protein is to be annotated and results are returned as graphical and text based files. The methods used here include different implementations of Markov models (Glimmer3, RNAmmer, LipoP, TIGRfam and Pfam) as well as artificial neural networks (SignalP) and weight matrix (PROSITE) approaches. The Department of Energy Joint Genome Institute (DOE-JGI) Microbial Annotation Pipeline [23] also uses RNAmmer for RNA detection but uses a combination of GeneMark [61], Prodigal [62] and Metagene [85] for protein coding genes. While GenMark uses HMMs for prediction, Metagene uses GC content and codon frequencies predict protein coding genes. Functional annotations are created from comparisons to Pfam, TIGRfam, KEGG (Kyoto Encyclopedia of Genes and Genomes)[86] and COG (Clusters of Orthologous Groups of proteins)[87]. The JCVI standard operating procedure for annotating prokaryotic metagenomic shotgun sequencing data [24] uses similar approaches for coding and non-coding gene finding as the systems described above and annotation is performed using Pfam and TIGRfam, TMHMM and LipoP in addition to homology search using BLAST and the enzyme profile database, PRIAM [88] which uses position specific scoring matrices.

As illustrated above, annotation systems include a variety of different data and model sources in order to annotate as many proteins as possible and as precisely as possible. Any one of the approaches used in these pipelines can stand alone as they describe a specific part of protein function. The InterPro database [49] has collected a large amount of data into one single entity connecting all relevant information to each entry and has created a method for searching and combining results from all the databases, InterProScan [89]. The pipelines described above predict protein function with a balance between high coverage and high specificity in annotation. This is often the desired level of annotation for publication of sequence data and identification of individual traits of a genome. However, the requirements for annotation in a comparative context are slightly different. Here the criteria for specificity is lowered as the need for coverage is prioritized. This work describes the construction of a functional scheme based on Pfam-A and the training of Artificial Neural Networks (ANNs) using sequence features, was used to create models for comparative functional annotation. The models are incorporated into a pipeline called CMGfunc, Comparative Microbial Genomic functions, and the performance of the pipeline is described in the paper ”CMGfunc, Comparative functional annotation of bacterial proteins using artificial neural networks and proteins domains”, Section 5.3 on page 67.

3.1 CAFA experiment

As the number of computational prediction methods increases it becomes necessary to construct a system for evaluating the performance of these methods in a standard framework. Efforts and focus on the need for this kind of assessment of computational predictions dates back to the first Critical Assessment of Structure Prediction (CASP, 1994), which aims to determine the progress in protein structure prediction [90]. The assessment project has proven highly successful and 20 years later CASP has a key role in the field of protein structure prediction [91]. The Critical Assessment of Functional Annotation experiment (CAFA) aims to become the functional equivalent of CASP by improving the performance and evaluation of functional annotation of proteins [92]. The project constructs a functionally unknown dataset from public data (Swiss-Prot and the Enzyme Function Initiative[93]) and research groups sign up to attempt to assign functions to the data. After a time (6 months to a year), predicted functions are compared to accumulated experimental functions and performance is evaluated. The project ran for the first time in 2006 and has collected many useful approaches to functional annotation. Furthermore, the experiment has highlighted the slow progress in experimental verification, further supporting the need for computational methods.

Other initiatives include the COMBREX[1], a project focused on the increase of speed in functional annotation and consists of a database of computational functional predictions and a system for experimentalists to validate the predictions. The project also offers small grants to support experiments [9].

3.2 Artificial Neural Networks

Artificial Neural Networks (ANNs) have proved useful as classification tools for different aspects of protein function such as protein fold [94], enzyme classes [66] and transmembrane proteins [95]. Non-homologous function prediction using features and ANNs was first implemented in the ProtFun method for human proteins [66]. The basic concept of an artificial neural network (ANN) consists of a large number of independent connected units called *nodes*. The connections between the nodes, the weights, hold the actual pattern of the model. Nodes are essentially equations which calculate an output value based on their input. The neurons function as switches which output a value, based on an activation function and the sum of the input values.

Usually all nodes in a layer will have the same activation function. If not else stated, equations presented in this section are as described by Baldi et. al. [96] and Lund et. al. [97]. Nodes are commonly arranged into connected layers, with one input and one output layer. The output layer commonly consists of only one node. The arrangement of layers is called the architecture of the ANN, and a commonly used architecture is the layered feedforward architecture, consisting of visible and hidden layers [96]. In a feedforward network, information flows only in one direction, that is, from input to output. The nodes in the input layer receive input from a real number vector and do not compute any values. The function of the input layer is to store data. The hidden layer nodes receive the data from the input layer, calculate an output from this input and send the output to the next layer. Figure 3.1 on the following page shows the structure of a node which receives input and computes output. This structure applies for both nodes in the hidden and the output layers. A node in a layer receives input from all the nodes connected to it, which is commonly all nodes in the preceding layer. The sum of the weighted ($w_{j,i}$) inputs is passed into the node x_i and the produces an output $z_i = g(x_i)$, where g is the activation function of the node. The total input to a

[1] www.combrex.bu.edu

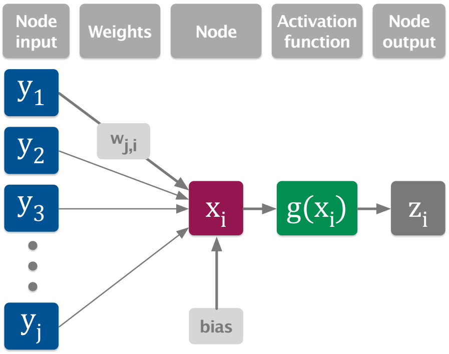


Figure 3.1: General structure of an artificial node, applies to nodes in hidden layers. The inputs values y_j, multiplied by their weights $(w_{j,i})$ and then summed up in the node, passed through the activation function and produces a output value, z_i.

node is a weighted $(w_{j,i})$ sum of outputs from the previous layer

$$x_i = \sum_j w_{j,i} * y_j + bias \tag{3.1}$$

where j is the index number of the node in the previous layer and bias is the bias or threshold of the unit which is the same for all nodes in a layer. The output from a node is then

$$z_i = g(x_i) = g\left(\sum_j w_{j,i} * y_j + bias\right) \tag{3.2}$$

There are a number of different activation functions in use. If the function is a threshold function, such as

$$g(x) = \{1 \text{ if } x > 0 \quad and \quad 0 \text{ if } x \leq 0\} \tag{3.3}$$

the node works as a threshold gate. Here, a logistic sigmoid activation function was used as, with the function given as

$$g(x) = \frac{1}{(1 + e^{-x})} \tag{3.4}$$

This function has a monotonic S shape which is shown in Figure 3.2. This function is especially useful for estimating the probability of binary events, though other activation functions can lead to essentially the same results [96].

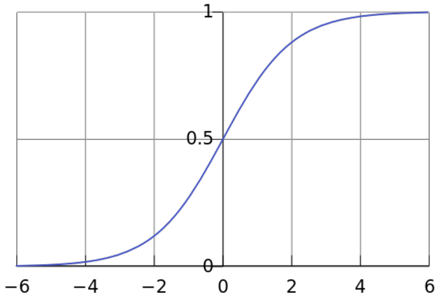


Figure 3.2: The logistics sigmoid function.

A network can have multiple hidden layers and the number of nodes in the hidden layer can also vary. The network can have multiple output nodes if this fits the purpose. An example of a neural network architecture consisting of 75 inputs, 2 hidden layers and 30 nodes in each and a single output node is shown in Figure 3.3 on the following page.

Testing and training, back- and forward-propagation

Artificial neural networks are not programmed in the traditional sense. Instead, a network gets its information from a *training* procedure where the model is exposed to input data with known output values. The network models the pattern in the data by adjusting the weights to fit as much of the data as best as possible. This is done by evaluating the error between the computed output (O_i) by the network and the true output (T_i) associated with the input. A commonly used error measure is to let the error E be proportional to the sum of the squared difference between known and computed output

$$E = \frac{1}{2}\sum_i (O_i - T_i)^2 \tag{3.5}$$

The goal of the training procedure is to minimize this error. One way of doing this is by using *back-propagation* to update the weights. The back-propagation

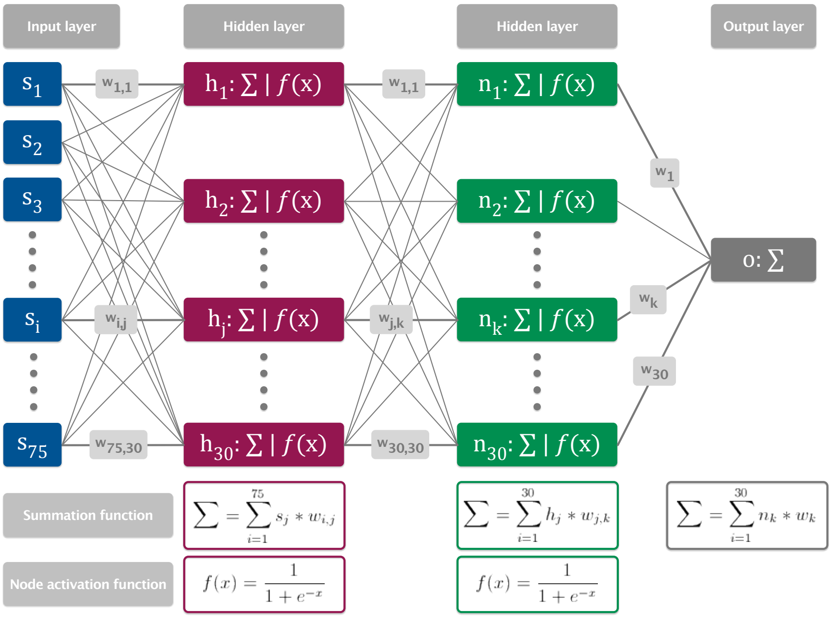


Figure 3.3: Example of feed-forward artificial neural network architecture. The input layer has 75 inputs and is fully connected to two hidden layers with 30 neurons in each and one output layer. The sigmoid activation function is used between all layers.

algorithm is a “steepest descent” method, which is used for finding the local minimum of a given function, such as the error function in this case and updates the weights using a defined step size. The step size is called the *learning rate* and the setting for this is is application-dependent and is typically chosen by experimentation. The error, the difference between the true and predicted output value, is back-propagated after each training set has been presented. With each back-propagation of the error the weight in the network is changed, and another forward-propagation is initiated by presenting a new set of inputs.

The neural network is trained, the weights are updated to reflect the pattern in the sequences, by presenting a number of examples of the pattern to the network. These are referred to as positive examples. Positive examples are essential for proper training, but it is also essential that the model can correctly identify true negatives, sequences that do not fit the model. These are called negative examples, and should be representative of the kind of negative data the model will encounter in true data. If the model needs to identify a very small fraction of proteins, it has to be trained with a wide variety of negative examples as these represent a much wider pattern than the positive examples. The training of artificial neural networks is the key step in using these models for prediction and great effort should be made in setting up the architectures and training data.

This chapter has described how protein function can be predicted with special emphasis on using artificial neural networks, how they function and how then can be trained. The paper ”CMGfunc, Comparative functional annotation of bacterial proteins using artificial neural networks and proteins domains”, Section 5.3 on page 67 illustrates how these concepts are used to create a set of functional models for bacterial proteins.

4 Data Sharing

In the work presented here, free and open source software was where possible used for analysis. Focus was kept on sharing data and procedures for results presented in this work for future reproduction. Furthermore, results have been published in open access journals to ensure accessibility to the work, and the software has been stored in public depositories for anyone to download and use.

4.1 Reproducibility

As bioinformatics becomes widely used and evolves, computational systems become more advanced and specific. Reproducing results is not only a matter of access to the right data but also requires access to the actual setup/pipeline/procedure used to generate figures/tables or key numbers. Creating portable and local running systems that can be distributed is key in this process and also ensures privacy when working with confidential data, which is not suited for online upload and analysis [98]. Such efforts are becoming more and more important as the data behind each publication grows too large to be shared in supplementary files or send via email. As a result, many published scientific publications cannot be reproduced, as for example microarray experiments where 10 of 18 tested publications could not be reproduced [99]. This further raises the question of responsibility – who should be in charge of ensuring that results can be reproduced by peers? Currently, the duty is being passed between the scientists and the journal editors, with no clear decision on the problem [100]. Only about 20% of large scale published datasets are used again by others in future publications [101]. Phrased another way, 80% of the papers with large datasets get published and the underlying data is ignored, never cited by others.

However difficult it is to place the responsibility, efforts to ensure reproducibility are becoming part of grant requirements, with applicants being asked to make plans and provide documentation for how to share and publish data created during the project (Research Councils UK[1]). Some journals are taking a similar approaches, requiring authors to deposit supporting data either with the journal

[1] www.uk.sagepub.com/repository/binaries/pdf/Library-OAReport.pdf

itself or with recognised data repositories (American Economic Review, the Journal of Evolutionary Biology, and Clinical Infectious Diseases). Initiatives like the The Digital Curation Centre (DCC) works on capacity, capability and skills for research data management in the research community[2]). Although requirements for publication of data is key to reproducibility, major problems arise when scientific credit, ownership and confidentiality comes into play [102]. Such problems require explicit licensing and in the case of confidentiality specific permits have to be retrieved before publication[3]. Several benefits come from working towards reproducible research, the first one being that much research is funded by state money, and the people have a right to access and have the research evaluated. Other effects include improvement in work habits which leads to better collaborations across research institutions and countries. Finally, working reproducibly increases the impact of the results, as it leads to less competition and more acknowledgement, when data can be reproduced, less people end up doing something almost identical and it is more easy to use and acknowledge the published work [103].

In this work, initiatives were taken to ensure reproducibility. Considerations like these were part of the decision process when creating the CMG-biotools and CMGfunc systems. Both of these systems were created as virtual computers giving the user access to all code and data. For CMGfunc, the training data used to create the models is also made available via the GitHub repository (repository name *cmgfunc*). The CMG-biotools system has no additional data and is as such, self contained. Both of these systems have the additional advantage of being graphical as well as command line based, giving the user an easy entry to simple command line use as well as a well known interface for handling files. The virtual computers are based on Xubuntu, a graphical but reduced version of the Ubuntu system and the virtual implementation means that they can run on all types of operating systems. Ensuring reproducibility is not only a matter of ensuring access to data but also access to procedures and methods used to analyze the data. In this project, data was stored in a MySQL database which can be used to verify the results. Although data in a MySQL database requires some practice to retrieve, using so called *queries*, the queries used to obtain the data can be made available with the database file itself. Using a database and queries to make data public is becoming more common, with a key example being the Pfam database[4]. Another example is the InterPro database, which can be downloaded

[2] www.dcc.ac.uk

[3] http://blogs.biomedcentral.com/bmcblog/files/2010/09/opendatastatementdraft.pdf

[4] http://pfam.sanger.ac.uk/help#tabview=tab11

as a local installation, containing an offline version of all the information of the system. InterPro is however not a SQL database and must be searched using specially designed programs in the install package. All data used for the CMGfunc project has been made available through GitHub in an effort to ensure future reproducibility of the results presented.

4.2 Publication

Another aspect of sharing data is the publication of results. The number of scientific articles is skyrocketing, with more than 25.000 new articles on Web of Science every week of 2011. A small fraction of these will be retracted [100]. Of course, retractions are a bad thing, and it is not desired to have a high rate of retractions; however, with the large increase in publications, and some constant level of error and misconduct, it is reassuring that some bad publications get detected and retracted. Although the peer review process has been a good way of ensuring relevance of scientific publication, the nature of the process might not always ensure the quality of publications. With much work being published based on computational analysis it is very difficult for reviewers to verify the published results. It is likely that changes must be made to the review process as both science and publications requirements change. Some initiatives have already been made in this direction, such as the F1000Research journal. The journal was started in 2010 and was indexed in PubMed in 2012. The goal is to improve the way research is communicated and introduces a new setup for publication and data access[5]. This journal runs on a very different scheme than other journals. The traditional approach to scientific publishing has been that the editor is the first to view the paper after which it may or may not be send for review. When the reviewers return their comments they also give a verdict, *approved*, *approved with revision* or *not approved.* The author is then given the chance to make revisions and resubmit within a timeline. This procedure is often very long, and many iterations might be required before publication is reached. Furthermore, the paper is not public or citable through this process, making it impossible for the authors to publish new work based on the first paper. The F1000 has a very different submission system. A paper is initially evaluated by a editorial team before being published as *awaiting peer review*, at which point the paper is made available on the F1000 webpage. Referees are selected and the comments are publically displayed with the online version of the paper. All communication between referee and scien-

[5] http://f1000research.com/about

tist takes place openly and when the revisions have been meet the paper is given the status *approved* and is indexed in Pubmed, Google Scholar and a number of other resources. Although some debate is going on about the review process of F1000research, some calling it a *non-journal* and referring to the review process as incomplete and messy, the project does illustrate the need for changes in the publication process as results are generated faster and based on larger datasets.

The work presented in this thesis is based on large amounts of data with complex analysis pipelines and setups, and although all data was obtained from public sources, the road to reproducibility is still long. Hopefully, procedures for data and analysis sharing will become standard soon and increase the knowledge which can be found in data.

5 Articles

5.1 CMG-Biotools, basic Comparative Microbial Genomics

The paper presented here represents a project focused on how to compare bacterial genomes and how to make the tools available and useful for other scientists. The work started as the development of teaching material but proved to be of interest to people outside the classroom as well. The paper describes a self-contained package that can be run on almost any modern portable laptop computer. The virtual machine creates a user-friendly and computationally powerful bioinformatics pipeline taking accessibility, reproducibility, speed and difficulty of use into consideration. The package has been used in one week introductory workshops in Spain, Norway, Morocco, Thailand, Nepal, and at the Center for Disease Control in Atlanta, Georgia (USA) as well as being used for several years in the semester course, Comparative Microbial Genomics at the Technical University of Denmark.

The paper presents CMG-biotools, a free workbench for Comparative Microbial Genomics which is a virtual computer that can be installed on all platforms. The workbench includes tools for formatting and handling of data types, gene prediction, proteome comparisons using sequence and feature based methods as well as structural analysis using circular DNA plots. The individual tools in this workbench have been published previously and were not developed for this project. The workbench, however, is the first time these tools are made available for local installation and use. The workbench is based on the Xubuntu operating system, which is similar to systems like Windows and Mac OSX. The workbench requires some introduction to Unix, and the user will be able to get more out of the tools once a basic level of Unix has been achieved. However, the system has proven a great way to introduce command line tools to people with no prior experience.

The paper was published in PLOS ONE in April 2013 and has been viewed over 1.800 times and based on the correspondence directed at the authors, the workbench is being used around the world.

CMG-Biotools, a Free Workbench for Basic Comparative Microbial Genomics

Tammi Vesth[1], Karin Lagesen[2], Öncel Acar[3], David Ussery[1]*

1 Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Kgs. Lyngby, Denmark, **2** Centre for Ecological and Evolutionary Synthesis, Department of Biology, University of Oslo, Oslo, Norway, **3** Department of Electrical Engineering, Technical University of Denmark, Kgs. Lyngby, Denmark

Abstract

Background: Today, there are more than a hundred times as many sequenced prokaryotic genomes than were present in the year 2000. The economical sequencing of genomic DNA has facilitated a whole new approach to microbial genomics. The real power of genomics is manifested through comparative genomics that can reveal strain specific characteristics, diversity within species and many other aspects. However, comparative genomics is a field not easily entered into by scientists with few computational skills. The CMG-biotools package is designed for microbiologists with limited knowledge of computational analysis and can be used to perform a number of analyses and comparisons of genomic data.

Results: The CMG-biotools system presents a stand-alone interface for comparative microbial genomics. The package is a customized operating system, based on Xubuntu 10.10, available through the open source Ubuntu project. The system can be installed on a virtual computer, allowing the user to run the system alongside any other operating system. Source codes for all programs are provided under GNU license, which makes it possible to transfer the programs to other systems if so desired. We here demonstrate the package by comparing and analyzing the diversity within the class *Negativicutes*, represented by 31 genomes including 10 genera. The analyses include 16S rRNA phylogeny, basic DNA and codon statistics, proteome comparisons using BLAST and graphical analyses of DNA structures.

Conclusion: This paper shows the strength and diverse use of the CMG-biotools system. The system can be installed on a vide range of host operating systems and utilizes as much of the host computer as desired. It allows the user to compare multiple genomes, from various sources using standardized data formats and intuitive visualizations of results. The examples presented here clearly shows that users with limited computational experience can perform complicated analysis without much training.

Citation: Vesth T, Lagesen K, Acar Ö, Ussery D (2013) CMG-Biotools, a Free Workbench for Basic Comparative Microbial Genomics. PLoS ONE 8(4): e60120. doi:10.1371/journal.pone.0060120

Editor: Valerie de Crécy-Lagard, University of Florida, United States of America

Received October 15, 2012; **Accepted** February 24, 2013; **Published** April 5, 2013

Funding: This work was supported in part by grants from the Danish Research Council, and from The Technical University of Denmark. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: dave@cbs.dtu.dk

Introduction

The number of microbial genome sequences has exploded due to the lower cost of sequencing facilitated by advances in sequencing technology making these services easier and faster. There are now roughly a hundred times as many sequenced prokaryotic genomes available as in 2000. The National Center for Biotechnology Information (NCBI) has an online list of genome sequences, complete and in progress. In 2000, 42 sequenced genomes were available on the NCBI list, and this number had grown to 4 189 in February 2012 (www.ncbi.nlm.nih.gov/genomes/lproks.cgi). Further, recently a single study [1] has compared genome sequences from 2 348 *Mycobacterium tuberculosis* isolates, and there are many more studies in progress where thousands of bacterial genome sequences are compared. As a consequence, more experimental biologists with little to no experience with bioinformatics find themselves in possession of an enormous amount of sequencing data and in need of tools necessary for analysis.

Analyzing the sequence of a single genome can confer a wide range of knowledge [2,3]. It is possible to use alignment tools to find a specific gene in a genome within seconds, for example to identify a genetic marker for a specific phenotype. DNA structure analyses can pinpoint chromosomal regions that lend themselves to certain genes and genomic elements. Regions that show distinct structural properties along the chromosome include clusters of genes encoding surface-proteins (usually more AT rich), possible phage insertions, regions likely to contain highly expressed genes as well as potential genomic islands [4–6]. Based on the annotation of a genome it is also possible to find the gene neighbors of a specific gene, thus possibly identifying functionally connected genes. The sequencing of individual genomes has facilitated a whole new approach to wet lab experiments that until recently were not possible. There is an enormous amount of information just in a single genome sequence.

However, the real power of genomics is manifested through comparative genomics. Even within a species, comparative genomics has highlighted a diversity that would not have been detected otherwise. The diversity within *Escherichia coli* was illustrated in a study from 2009, where the number of gene families, in *Escherichia coli* was estimated to be 43 000 [7]; this number is expected to become larger as more genomes are sequenced. Another example of the power of comparative genomics, this time within low diversity genomes, can be found in a study of two *Bacillus* species, *B. anthracis* and *B. cereus*. These are difficult to differentiate based on chromosomal markers [8], and the difference in pathogenicity is solely determined by the strict presence of two virulence plasmids, which both are required for anthrax. The diversity of a species can be estimated by multiple sequence comparisons across genomes calculating the pan genome (all genes found in genomes) [9]. Comparative microbial genomics (CMG) also allows for fast and inexpensive analyses, for example phylogenetic relationships between organisms. Further, it is possible to build up data from known organisms that would allow for quick classification of an isolate of an unknown organism, just from its genome sequence.

The CMG-biotools package presented here is designed for microbiologists with limited knowledge of computational analyses and comes with a basic introduction to Unix. Within this package it is possible to do phylogenetic analysis, proteome comparisons, DNA structure analysis and much more, just with a list of genomes. Most of the analyses can be performed on FASTA formatted DNA sequences from unpublished projects as well. The CMG-biotools system presents a stand-alone interface for comparative microbial genomics. The package is a installable operating system, based on Xubuntu 10.10 available through the open source Ubuntu project (www.xubuntu.org/get). This setup overcomes problems with dependencies and platform specificity allowing for all users to work in the same environment. Ubuntu is a widely used, free of charge and open source operating system with a large user community and thousands of free applications. As of 2012, Ubuntu is the second most popular Linux distribution, only surpassed by Mint [10]. It is a stand-alone operating system and can be installed directly onto a local computer or on a virtual computer using virtualization software. The CMG-biotools operating system has been tested on a free virtualization application, VirtualBox (www.virtualbox.org). This system addresses the problem of working with large amounts of data, allowing for comparative analyses of multiple genomes, thereby making use of the vast amount of sequence information that is now available in laboratories all over the world.

Results and Discussion

To demonstrate the capabilities of the CMG-biotools (Comparative Microbial Genomics), analyses are performed on a set of genomes from the class *Negativicutes*. The CMG-biotools operating system was installed on an 8 Gigabyte virtual computer using VirtualBox (www.virtualbox.org). Figure 1 illustrates the work and data flow of the analyses.

Data Collection and Assessment

The first step of the analyses is to obtain genome data for a set of organisms. In the example presented here, we obtain data from the GenBank database [11] at the National Center for Biotechnology Information (NCBI, www.ncbi.nlm.nih.gov/genome/browse/) This database is part of the International Nucleotide Sequence Database Collaboration (INSDC) and contains more than 3000 bacterial genome projects. For the example, organisms of the class *Negativicutes* were identified from NCBI genomes list (www.ncbi.nlm.nih.gov/genome/browse/, "Prokaryotes", *Negativicutes* (taxid:909932)) and GenBank INSDC numbers or whole genome sequence numbers (WGS) were obtained. The genome sequences of 6 complete (NCBI Genomes list, status: "Complete") and 25 assembly genomes (NCBI Genomes list "Scaffolds/contigs") were identified. NCBI GenBank INSDC numbers were used for complete genomes while WGS numbers were used for draft sequences. Using the program getgbk and the INSDC/WGS numbers, each genome was downloaded in the NCBI GenBank format (Figure 1, Step 1). A list of genome names and INSDC/WGS numbers is found in Table 1. DNA sequences were extracted from GenBank files and saved in FASTA format(*saco_convert* [12], Figure 1, Step 2B).

Basic statistical parameters were calculated for the 31 genomes (Figure 1, Step 3B), using whole genome DNA FASTA files as input, and the results are shown in Table 2. The AT content varied from 42 to 66% and the genome size ranged from 1.26 to 2.89 Mega bases (Mb). The percentage of unknown bases refers to letters in the DNA code that are not A, C, T or G. These bases might be the result of an assembly process or errors in sequencing. Of the 31 genomes analyzed, 8 had non-canonical base letters in the DNA sequences, ranging from 0.0001%. to 3.6%. The fraction of the largest contig will be 100% for genomes with one chromosome and therefore this measure is more useful for identification of incomplete sequences. For the non-complete genomes, the fraction made up by the largest sequence varied from 5% to 30%. It is seen the the fraction correlates with the number of contigs, if the genome sequence is in many contigs, then the largest sequence covers a small fraction of the entire genome. These findings show a large variation in the dataset, both in the context of biology (AT content and size) and data quality (number of contigs and percentage of unknown bases).

Gene Finding

The next step in the analysis is to identify coding regions in DNA sequences. Some genome projects have manually curated and high quality annotations while others have no annotations at all. Again others have been annotated using a genefinding algorithm without any additional evaluation of the findings. The CMG-biotools uses the program Prodigal [13] for genefinding and has been incorporated into a pipeline called prodigalrunner. This pipeline takes a genome DNA GenBank or FASTA file as input (Figure 1, Step 2C) The output from prodigalrunner is a preliminary GenBank file (.gbk), a general feature format file (.gff), a FASTA formatted open reading frame file (genes,.orf.fna) and a FASTA formatted protein file which contains the translations of the genes (orf.fsa). Table 3 shows the number of published genes compared to the number of genes found when using Prodigal for genefinding.

This genefinder found between 1 206 (*D. micraerophilus* DSM 19965) and 2 886 (*Thermosinus carboxydivorans* Nor1) proteins in the 31 genomes. Compared to the published proteins from GenBank, Prodigal finds roughly the same number of genes, except for two genomes which did not have any published annotations. The advantage of using an independent gene finder for all genome sequences in an analysis is that the difference introduced by annotators will be removed. As information on how genefinding was performed is rarely available, doing local genefinding might eliminate badly annotated projects. Whether to use published annotations is up to the individual user but for obvious reasons, genefinding will have to be done for projects with no published annotations. For the remaining analysis in this paper, proteomes predicted using prodigalrunner will be used.

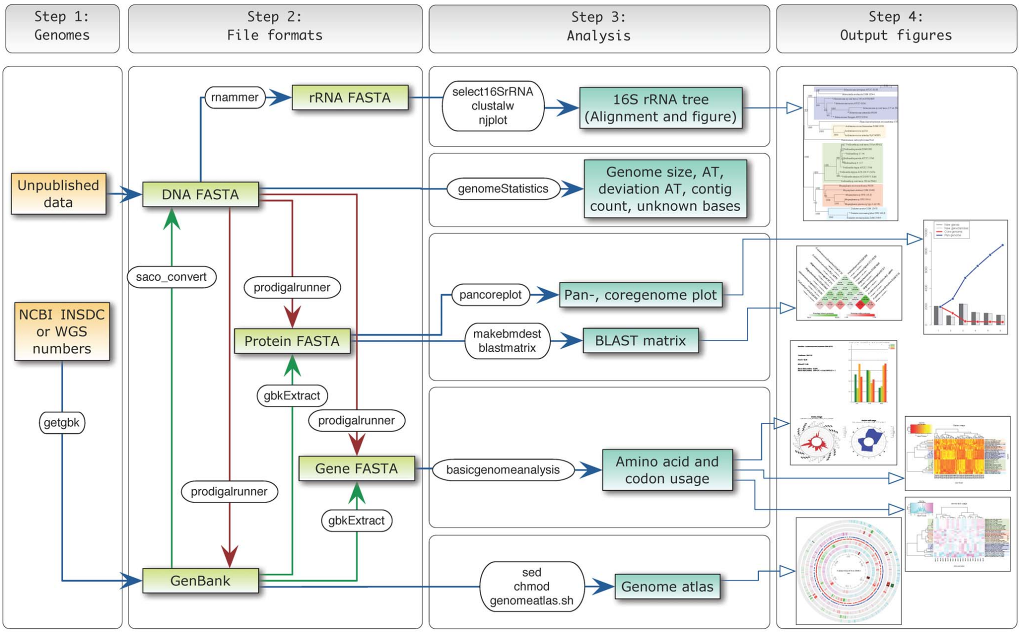


Figure 1. Analysis workflow. Visual representation of the data flow through each of the steps in the CMG-biotools system. The figure shows the analysis input and program name along with the analysis output type. Green arrows indicate data extraction from a GenBank file format, this data needs to be available in the file for these steps to work. Red arrows indicate local genefinding which results in gene FASTA, protein FASTA and GenBank files.
doi:10.1371/journal.pone.0060120.g001

Phylogenetic Analysis

The chromosomal DNA sequence, as extracted from the GenBank files (FASTA format) is used as input for this analysis, as illustrated in Figure 1, Step 2A. The whole genome DNA sequence is searched for rRNA sequences using RNAmmer [14] and a sequence from each genome is extracted (select16SrRNA, Figure 1, Step 3A). The selection criteria for the extraction process defaults to the highest scoring sequence found with a length between 1 400 and 1 800 base pairs. This selection is not necessarily the most correct way of selecting a 16S rRNA sequence for phylogenetic analysis, but offers the opportunity to compare genomes based on a single sequence. The alignment program ClustalW [15] is used for multiple sequence alignment of the sequences. From the alignment, a distance tree is constructed, using 1 000 bootstraps [16] to find the best fitting distance tree (the output is a file Phylip tree format.phb). Each node of the tree is shown with a bootstrap value between 0 and 1 000, the number indicating how many times this branching is seen out of 1 000 re-samplings. The higher the number the more reliable the branching. The visualization of the tree was done using njplot [17] and is shown in Figure 2.

The results of the RNAmmer analysis yielded no rRNA sequences for two genomes (*Centipeda periodontii* DSM 2778, 72 contigs, and *Megamonas hypermegale* ART12 1, 1 replicon). Sequences from 6 genomes had lengths outside the default thresholds - length between 1 400 and 1 800 base pairs (Table 4, 16S rRNA length and score for each genome). For this analysis the thresholds were changed to include these 6 genomes (lower threshold for sequence length was changed to 1 100 base pairs). The genome of *Megamonas hypermegale* contains a large number of unknown bases (found in 99 stretches of lengths between 141 and 1780 nucleotides, calculated using countUnknowns.pl). The average length of these stretches was 804 nucleotide positions, roughly half the length for a 16S rRNA sequence. It is here hypothesized that such unknown base stretches can prevent rnammer from identifying ribosomal RNA sequences, because parts of the 16S rRNA sequence might be missing. The sequence of *Centipeda periodontii* DSM 2778 does not contain any unknown bases, but still no rRNA sequences were found in this sequence. The genome is in 72 contigs and the largest sequence is 8.5% of the total, numbers that are not extreme compared to other genomes in this analysis (Table 3). It can be hypothesized that the lack of 16S rRNA sequences in this genome might be a result of the sequence assembly. Since ribosomal RNA sequences often are repeated sequences, the assembly process might not be able to conclusively place the rRNA in the DNA, and might discard the sequences all-together.

The 16S rRNA tree (Figure 2) has been manually colored by genus, where multiple genomes per genus was available. The genomes show a general tendency to cluster within their taxonomical groups. Furthermore, the tree shows three main clusters with *Acidaminococcus* and *Selenomonas* as separate clusters (cluster II and III). The last cluster contains the genomes of *Veillonella*, *Megasphaera* and *Dialister*, all clustered in subgroups according to taxonomy. The clustering of genomes according to genera is expected since the taxonomic naming is based on 16S

Table 1. Genome information.

Tax	Organism	INSDC	WGS	WGS for download	Status
591001	*Acidaminococcus fermentans* DSM 20731	CP001859	–	–	Complete
568816	*Acidaminococcus intestini* RyC-MR95	CP003058	–	–	Complete
563191	*Acidaminococcus sp* D21	–	ACGB01	ACGB00000000	Scaffolds/contigs
888060	*Centipeda periodontii* DSM 2778	–	AFHQ01	AFHQ00000000	Scaffolds/contigs
592028	*Dialister invisus* DSM 15470	–	ACIM02	ACIM00000000	Scaffolds/contigs
888062	*Dialister micraerophilus* DSM 19965	–	AFBB01	AFBB00000000	Scaffolds/contigs
910314	*Dialister microaerophilus* UPII 345-E	–	AENT01	AENT00000000	Scaffolds/contigs
158847	*Megamonas hypermegale* ART12 1	FP929048	–	–	Complete
907	*Megasphaera elsdenii* DSM 20460	HE576794	–	–	Complete
699218	*Megasphaera genomosp* type 1 str 28L	–	ADGP01	ADGP00000000	Scaffolds/contigs
706434	*Megasphaera micronuciformis* F0359	–	AECS01	AECS00000000	Scaffolds/contigs
1000569	*Megasphaera sp* UPII 135-E	–	AFUG01	AFUG00000000	Scaffolds/contigs
1000568	*Megasphaera sp* UPII 199-6	–	AFIJ01	AFIJ00000000	Scaffolds/contigs
500635	*Mitsuokella multacida DSM 20544*	–	ABWK02	ABWK00000000	Scaffolds/contigs
626939	*Phascolarctobacterium succinatutens* YIT 12067	–	AEVN01	AEVN00000000	Scaffolds/contigs
749551	*Selenomonas artemidis* F0399	–	AECV01	AECV00000000	Scaffolds/contigs
638302	*Selenomonas flueggei* ATCC 43531	–	ACLA01	ACLA00000000	Scaffolds/contigs
585503	*Selenomonas noxia* ATCC 43541	–	ACKT01	ACKT00000000	Scaffolds/contigs
879310	*Selenomonas sp* oral taxon 137 str F0430	–	AENV01	AENV00000000	Scaffolds/contigs
864563	*Selenomonas sp* oral taxon 149 str 67H29BP	–	AEEJ01	AEEJ00000000	Scaffolds/contigs
546271	*Selenomonas sputigena* ATCC 35185	CP002637	ACKP02	ACKP00000000	Complete
401526	*Thermosinus carboxydivorans* Nor1	–	AAWL01	AAWL00000000	Scaffolds/contigs
866776	*Veillonella atypica* ACS-049-V-Sch6	–	AEDR01	AEDR00000000	Scaffolds/contigs
866778	*Veillonella atypica* ACS-134-V-Col7a	–	AEDS01	AEDS00000000	Scaffolds/contigs
546273	*Veillonella dispar* ATCC 17748	–	ACIK02	ACIK00000000	Scaffolds/contigs
686660	*Veillonella parvula* ATCC 17745	–	ADFU01	ADFU00000000	Scaffolds/contigs
479436	*Veillonella parvula* DSM 2008	CP001820	–	–	Complete
457416	*Veillonella sp* 3 1 44	–	ADCV01	ADCV00000000	Scaffolds/contigs
450749	*Veillonella sp* 6 1 27	–	ADCW01	ADCW00000000	Scaffolds/contigs
879309	*Veillonella sp* oral taxon 158 str F0412	–	AENU01	AENU00000000	Scaffolds/contigs
944564	*Veillonella sp* oral taxon 780 str F0422	–	AFUJ01	AFUJ00000000	Scaffolds/contigs

Table listing the genomes used in the analysis. Data was downloaded from NCBI GenBank database. Abbreviations: *Tax*: NCBI taxonomy id number, *Organism*: Name of organism, *INSDC*: NCBI GenBank Accession number, *WGS*: NCBI Whole Genome Sequence Project number, *Status*: status of sequencing project. The WGS number can be used for downloading whole genome sequencing projects by removing the last two numbers and adding 6 zeros (ACGB01 is downloaded using the number ACGB000000).

doi:10.1371/journal.pone.0060120.t001

rRNA comparison [18]. It should be noted that the resulting trees shown here should be considered as preliminary classification.

Genome Atlases (Structural DNA Atlas)

Genome atlases were constructed for each of the 6 complete genomes using GenBank files generated by prodigalrunner(Table 1 and Figure 3, high resolution figure as supplemental Figure S1). The input to this analysis is a GenBank file containing one replicon of a genome (a single chromosome or plasmid, Figure 1, Step 3E). The analysis is performed using the program genomeAtlas, which is a collection of scripts that utilizes the GeneWiz program [6]. The genome atlas shows three types of information: base composition (AT content, GC skew), global repeats within the replicon (direct and inverted), and DNA structural properties (position preference, DNA stacking energy, and curvature). Genes (blue for leading and red for lagging strand), rRNAs and tRNAs are displayed as found in the GenBank annotation. The DNA is used for simple base count information includes AT content and GC skew. The atlas also shows a visual representation of structural properties of the DNA molecule (inverted and direct repeats, position preference [19], stacking energy [20] and intrinsic curvature [21,22]). These different structures can potentially influence gene expression, likelihood of gene rearrangement and even evolutionary hotspots. The atlases in Figure 3 show a range of different DNA structure properties. Arrows and colors mark different important regions on each atlas (added to the atlases manually).

Mobile elements sometimes have different base composition, and can be indicated by areas of different curvature, stacking energy and position preference, compared to the chromosomal average (grey), as seen from the atlas of *Acidaminococcus fermentans*. Highly expressed regions are sometimes regions which will not

Table 2. Genome statistics.

Organism	bp	AT	Std. AT	Contig	Unknown	Largest	N50
Acidaminococcus fermentans DSM 20731	2 329 769	44,16	–	1	–	100	2 329 769
Acidaminococcus intestini RyC-MR95	2 487 765	49,98	–	1	–	100	2 487 765
Acidaminococcus sp D21	2 238 973	49,80	0,03	79	–	6,2	43 082
Centipeda periodontii DSM 2778	2 650 230	44,02	0,04	71	–	8,4	72 349
Dialister invisus DSM 15470	1 895 860	54,50	0,03	2	–	99,9	1 894 898
Dialister micraerophilus DSM 19965	1 256 198	64,69	0,05	32	–	17,9	90 852
Dialister microaerophilus UPII 345-E	1 395 825	64,35	0,07	32	–	15,4	122 970
Megamonas hypermegale ART12 1	2 209 938	65,89	–	1	3,602	100	2 209 938
Megasphaera elsdenii DSM 20460	2 474 718	47,01	–	1	0,397	100	2 474 718
Megasphaera genomosp type 1 str 28L	1 726 197	53,95	0,03	34	–	12,2	156 177
Megasphaera micronuciformis F0359	1 765 374	54,56	0,04	49	–	24,8	142 252
Megasphaera sp UPII 135-E	1 440 762	61,19	0,04	46	0,001	12,0	63 822
Megasphaera sp UPII 199-6	1 242 998	53,26	0,04	38	–	12,0	96 055
Mitsuokella multacida DSM 20544	2 204 718	41,89	0,04	28	–	19,5	321 943
Phascolarctobacterium succinatutens YIT 12067	2 122 261	52,36	0,05	118	–	5,1	43 220
Selenomonas artemidis F0399	2 209 623	42,75	0,06	66	–	19,7	89 528
Selenomonas flueggei ATCC 43531	2 157 862	44,03	0,04	33	–	12,2	125 841
Selenomonas noxia ATCC 43541	2 039 467	44,13	0,05	56	–	14,2	106 401
Selenomonas sp oral taxon 137 str F0430	2 475 066	43,27	0,05	15	–	22,1	306 540
Selenomonas sp oral taxon 149 str 67H29BP	2 429 414	43,20	0,05	56	–	7,8	95 526
Selenomonas sputigena ATCC 35185	2 568 361	42,89	–	1	–	100	2 568 361
Thermosinus carboxydivorans Nor1	2 889 774	48,50	0,03	49	–	12,1	108 262
Veillonella atypica ACS-049-V-Sch6	2 053 871	61,03	0,04	63	–	10,3	80 793
Veillonella atypica ACS-134-V-Col7a	2 151 913	61,02	0,04	70	–	9,8	74 331
Veillonella dispar ATCC 17748	2 116 567	61,14	0,06	25	–	30,4	498 249
Veillonella parvula ATCC 17745	2 163 473	61,43	0,04	19	–	26,9	416 853
Veillonella parvula DSM 2008	2 132 142	61,37	–	1	–	100	2 132 142
Veillonella sp 3 1 44	2 156 561	61,36	0,04	31	–	18,0	282 953
Veillonella sp 6 1 27	2 169 785	61,33	0,04	22	–	15,8	257 597
Veillonella sp oral taxon 158 str F0412	2 176 752	61,05	0,04	21	–	19,3	366 615
Veillonella sp oral taxon 780 str F0422	1 731 014	60,55	0,03	75	–	14,0	73 892

Basic genome statistics for genome DNA sequences. Values of zero are marked by "–". Abbreviations: *Organism*: Name of organism. *Status*: sequencing status of published project. *bp*: total number of base pairs in all DNA. *AT*: Percent of AT in DNA. *Std. AT*: Standard deviation in AT across DNA fragments. *Contig*: number of DNA fragments corresponding to replicons or contigs. *Unknown*: percentage of unknown bases (not A, T, C or G). *Largest*: size of largest contig as a percentage of total length. *N50*: weighted median statistic such that 50% of the entire assembly is contained in contigs or scaffolds equal to or larger than this value.

doi:10.1371/journal.pone.0060120.t002

easily condense around chromatin proteins (See atlas for *Acidaminococcus intestini* RyC-MR95, very low position preference, average stacking energy and position preference). Some regions are often associated with rRNA sequences and these patterns are also thought to correlate with high gene expression (See atlas for *Megasphaera elsdenii* DSM 20460, less negative stacking energy (red, melt easy) and low position preference (flexible)). Regions with high curvature and stacking energy indicate a strongly curved region with tendency to melt (See atlas for *Selenomonas sputigena* ATCC 35185). This structure might be involved in a special DNA structure, maybe where the chromosome attaches to the bacterial cell membrane. On the chromosome of *Veillonella parvula* DSM 2008 are several regions with high curvature, stacking energy and position preference, suggesting this region to be curved, rigid and easily melted. The genes in this region might be highly expressed but controlled by histone-like proteins that preferentially bind to curved DNA. The draft chromosome of *Megamonas hypermegale* ART12 1 is slightly different from the other atlases. For five of the six atlases in Figure 3, the GC skew indicates the location of the origin and terminus of replication, and changes from most G's (blue) to more C's (pink). For most bacterial genomes, G's are biased toward the leading strand [23]. Note how the number of genes on leading/lagging strand changes along with the GC skew (more G's, more minus strand genes). For the genome of *Megamonas hypermegale* ART12 1, the GC skew lane is a mixture of pink and blue, likely because this is a draft genome sequence. The genome is also highly AT rich (66%) and contains three regions with DNA structural patterns different from the rest of the genome.

Table 3. Genefinding and published genes.

Genome name	GenBank	Prodigal	ID
Acidaminococcus fermentans DSM 20731	2 026	2 063	CP001859
Acidaminococcus intestini RyC-MR95	2 404	2 372	CP003058
Acidaminococcus sp D21	2 005	2 105	ACGB00000000
Centipeda periodontii DSM 2778	2 559	2 440	AFHQ00000000
Dialister invisus DSM 15470	1 954	1 765	ACIM00000000
Dialister micraerophilus DSM 19965	1 243	1 206	AFBB00000000
Dialister microaerophilus UPII 345-E	1 310	1 308	AENT00000000
Megamonas hypermegale ART12 1	2 118	2 759	FP929048
Megasphaera elsdenii DSM 20460	2 220	2 222	HE576794
Megasphaera genomosp type 1 str 28L	1 610	1 560	ADGP00000000
Megasphaera micronuciformis F0359	1 774	1 724	AECS00000000
Megasphaera sp UPII 135-E	1 310	1 291	AFUG00000000
Megasphaera sp UPII 199-6	1 151	1 112	AFIJ00000000
Mitsuokella multacida DSM 20544	2 142	1 942	ABWK00000000
Phascolarctobacterium succinatutens YIT 12067	2 150	2 012	AEVN00000000
Selenomonas artemidis F0399	2 195	2 024	AECV00000000
Selenomonas flueggei ATCC 43531	2 117	2 045	ACLA00000000
Selenomonas noxia ATCC 43541	2 020	1 955	ACKT00000000
Selenomonas sp oral taxon 137 str F0430	2 395	2 341	AENV00000000
Selenomonas sp oral taxon 149 str 67H29BP	2 407	2 313	AEEJ00000000
Selenomonas sputigena ATCC 35185	2 255	2 283	CP002637
Thermosinus carboxydivorans Nor1	2 750	2 886	AAWL00000000
Veillonella atypica ACS-049-V-Sch6	1 840	1 865	AEDR00000000
Veillonella atypica ACS-134-V-Col7a	1 903	1 923	AEDS00000000
Veillonella dispar ATCC 17748	1 954	1 941	ACIK00000000
Veillonella parvula ATCC 17745	1 929	1 945	ADFU00000000
Veillonella parvula DSM 2008	1 844	1 865	CP001820
Veillonella sp 3 1 44	0	1 922	ADCV00000000
Veillonella sp 6 1 27	0	1 936	ADCW00000000
Veillonella sp oral taxon 158 str F0412	2 000	2 029	AENU00000000
Veillonella sp oral taxon 780 str F0422	1 588	1 605	AFUJ00000000

Table listing genome name, number of published proteins (*GenBank*) and number of proteins found using Prodigal for genefinding (*Prodigal*). The column labeled *"ID"* refers to the INSDC or WGS id number as described in Table 1.

doi:10.1371/journal.pone.0060120.t003

Amino Acid and Codon Usage

The input to the analysis of codon usage and bias in third codon position is a gene FASTA file (DNA). The amino acid usage can be performed on any set of proteins in FASTA format using aminoacidUsagePlot (Figure 1, Step 3D). Here, both analyses were run using the genes and proteins identified by prodigalrunner (Figure 1, Step 2C). The program basicgenomeanalysis calculates the bias in third position, codon and amino acid usage and the output is a text file containing the values along with a PDF file with plots. The bias is defined as −1 in the case of 100% A or T in third position, +1 is the case of 100% G or C.

The bias in third position was analyzed and visualized for the 6 complete genomes (Figure 4). The genomes of *V. parvula* DSM 2008 and *M. hypermegale* ART12 1 have a high bias towards A/T in third position (bias score, −0.3906 and −0.6256, respectively) and also a very high AT content (66% and 61%, respectively). The genomes of *S. sputigena* ATCC 35185 and *A. fermentans* DSM 20731, have low AT content and a bias towards G/C in third position (bias score, 0.4719 and 0.4276, respectively). *M. elsdenii* DSM 20460 and *A. intestini* RyC-MR95 have average AT content but *M. elsdenii* has a clear bias in third position towards C (bias score, 0.3175). This analysis shows the diversity of AT content between these genomes and also illustrates how AT content correlates with the nucleotide bias in third codon position.

The codon and amino acid usage was calculated for all 31 genomes and visualized in heatmaps created in R (Figure 5, genera colors were added manually). The genera of *Veillonella* and *Selenomonas* cluster together showing that each species have a unique use of both codons and amino acids. The genomes belonging to *Megasphaera*, *Acidaminococcus* and *Dialister* are less conserved, and do not consistently cluster together. These two trees show a different relationship than the 16S rRNA tree (Figure 2). The amino acid usage tree shows three main clusters with *Selenomonas* and *Dialister* forming their own clusters (cluster II and III). The last cluster (cluster I) consists of *Veillonella*, *Megasphaera* and *Acidaminococcus*. This is significantly different from the codon

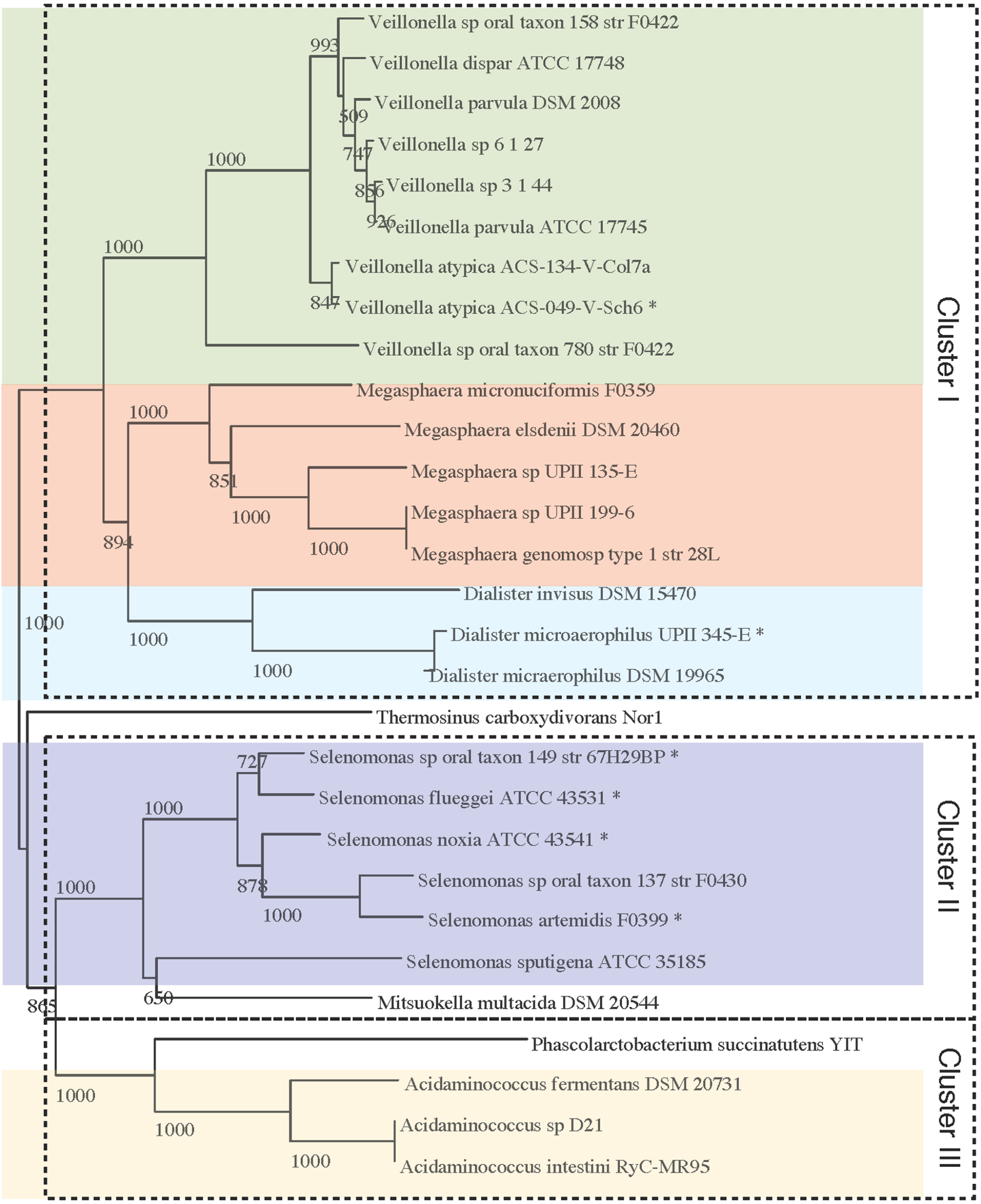

Veillonella sp oral taxon 158 str F0422
Veillonella dispar ATCC 17748
Veillonella parvula DSM 2008
Veillonella sp 6 1 27
Veillonella sp 3 1 44
Veillonella parvula ATCC 17745
Veillonella atypica ACS-134-V-Col7a
Veillonella atypica ACS-049-V-Sch6 *
Veillonella sp oral taxon 780 str F0422
Megasphaera micronuciformis F0359
Megasphaera elsdenii DSM 20460
Megasphaera sp UPII 135-E
Megasphaera sp UPII 199-6
Megasphaera genomosp type 1 str 28L
Dialister invisus DSM 15470
Dialister microaerophilus UPII 345-E *
Dialister micraerophilus DSM 19965
Thermosinus carboxydivorans Nor1
Selenomonas sp oral taxon 149 str 67H29BP *
Selenomonas flueggei ATCC 43531 *
Selenomonas noxia ATCC 43541 *
Selenomonas sp oral taxon 137 str F0430
Selenomonas artemidis F0399 *
Selenomonas sputigena ATCC 35185
Mitsuokella multacida DSM 20544
Phascolarctobacterium succinatutens YIT
Acidaminococcus fermentans DSM 20731
Acidaminococcus sp D21
Acidaminococcus intestini RyC-MR95
993
509
747
856
926
1000
1000
847
1000
851
1000
1000
894
1000
1000
1000
727
1000
878
1000
1000
650
865
1000
1000
1000
Cluster I
Cluster II
Cluster III

Figure 2. 16S rRNA tree. Each genome sequence was searched for 16S rRNA patterns and candidate sequences were extracted. The best sequence from each genome was selected. For two genomes, no sequences were found, *Centipeda periodontii* DSM 2778, *Megamonas hypermegale* ART12 1. For 6 additional genomes, the located sequences were shorter than the default acceptable length. The short sequences sequences are marked with a ''*''. Length criteria was changed from minimum 1 400 to 1 100 and maximum 1 800 unchanged. The distance tree was made with 1 000 bootstraps.
doi:10.1371/journal.pone.0060120.g002

Table 4. Ribosomal RNA analysis using RNAmmer.

Organism	Status	Score	Length (bp)	Total seq.
Acidaminococcus fermentans DSM 20731	Complete	1 910.8	1 545	6
Acidaminococcus intestini RyC-MR95	Complete	1 920.1	1 545	3
Acidaminococcus sp D21	Scaffolds/contigs	1 920.1	1 545	1
Centipeda periodontii DSM 2778	Scaffolds/contigs	–	–*	–
Dialister invisus DSM 15470	Scaffolds/contigs	1 836.1	1 557	3
Dialister micraerophilus DSM 19965	Scaffolds/contigs	1 878.8	1 555	1
Dialister microaerophilus UPII 345-E	Scaffolds/contigs	1 197.2	1 325*	1
Megamonas hypermegale ART12 1	Complete	–	–*	–
Megasphaera elsdenii DSM 20460	Complete	1 842.0	1 552	7
Megasphaera genomosp type 1 str 28L	Scaffolds/contigs	1 860.0	1 557	1
Megasphaera micronuciformis F0359	Scaffolds/contigs	1 816.0	1 550	1
Megasphaera sp UPII 135-E	Scaffolds/contigs	1 887.4	1 556	1
Megasphaera sp UPII 199-6	Scaffolds/contigs	1 868.7	1 556	1
Mitsuokella multacida DSM 20544	Scaffolds/contigs	1 915.8	1 549	2
Phascolarctobacterium succinatutens YIT 12067	Scaffolds/contigs	1 907.9	1 646	1
Selenomonas artemidis F0399	Scaffolds/contigs	6.368	1137*	1
Selenomonas flueggei ATCC 43531	Scaffolds/contigs	1 089.7	1 216*	1
Selenomonas noxia ATCC 43541	Scaffolds/contigs	1 364.8	1 296*	1
Selenomonas sp oral taxon 137 str F0430	Scaffolds/contigs	1 830.8	1 532	4
Selenomonas sp oral taxon 149 str 67H29BP	Scaffolds/contigs	1 252.5	1 258*	1
Selenomonas sputigena ATCC 35185	Complete	1 861.4	1 543	4
Thermosinus carboxydivorans Nor1	Scaffolds/contigs	1 898.8	1 549	7
Veillonella atypica ACS-049-V-Sch6	Scaffolds/contigs	1 512.8	1 369*	1
Veillonella atypica ACS-134-V-Col7a	Scaffolds/contigs	1 871.2	1 551	1
Veillonella dispar ATCC 17748	Scaffolds/contigs	1 870.5	1 551	3
Veillonella parvula ATCC 17745	Scaffolds/contigs	1 848.5	1 553	1
Veillonella parvula DSM 2008	Complete	1 859.5	1 551	4
Veillonella sp 3 1 44	Scaffolds/contigs	1 861.6	1 553	1
Veillonella sp 6 1 27	Scaffolds/contigs	1 862.2	1 551	1
Veillonella sp oral taxon 158 str F0412	Scaffolds/contigs	1 860.5	1 552	4
Veillonella sp oral taxon 780 str F0422	Scaffolds/contigs	1 877.1	1 550	4

The total number of identified 16S rRNA sequences is shown for each genome sequence. Length of highest scoring sequence and corresponding RNAmmer score is given. Default settings is to select the sequence with the highest RNAmmer score and a length between 1 400–1 800 bases. For this analysis the criteria were changed to a length range of 1 100–1 800, to include sequences from all genomes with 16S rRNA matches. Sequences with lengths shorter than the default acceptance threshold are marked with a "*". Two organisms did not have any hits to the RNAmmer models, values of zero are marked by "−".
doi:10.1371/journal.pone.0060120.t004

usage tree which creates a cluster consisting of *Veillonella* and *Dialister* with a single *Megasphaera* genome (cluster III), another cluster of *Selenomonas*(cluster II) and the last cluster of *Megasphaera* and *Acidaminococcus* (cluster I). None of the two methods makes a consistent clustering of the *Megasphaera* genomes as the 16S rRNA tree. In accordance, none of the three trees show the same general clusters, however they all manage to cluster closely related genomes, with the single exception of *Megasphaera*.

Proteome Comparisons Using BLAST

For this analysis, proteomes were constructed for all 31 genomes using prodigalrunner for genefinding. Presented here are two different types of proteome comparisons, both based on the BLAST algorithm (Basic Local Alignment Search Tool) [24,25]. The first method is a BLAST matrix and shows a pairwise proteome comparison by using BLAST to identify whether two proteins are shared between genomes [26]. Two proteins are considered to be in the same family if 50% of the alignment consists of identical matches and the length of the alignment is 50% of the longest gene. The main part of the matrix consists of pairwise genome comparisons; with fractions of shared proteins shaded in green (more green, more protein families shared). The row that would reflect self-comparison indicates internal homologs (internal paralogs, shaded red) which are defined as a significant hit within a genome to a protein other than the query protein itself.

The program performing this analysis is called blastmatrix and the input is an XML file (Figure 1, Step 3C). This file is created by the program makebmdest by inputting the name of a directory containing protein files. This program takes all the protein FASTA files in a given directory, extracts relevant information and formats it into an XML file which is read by the *blastmatrix* program. The

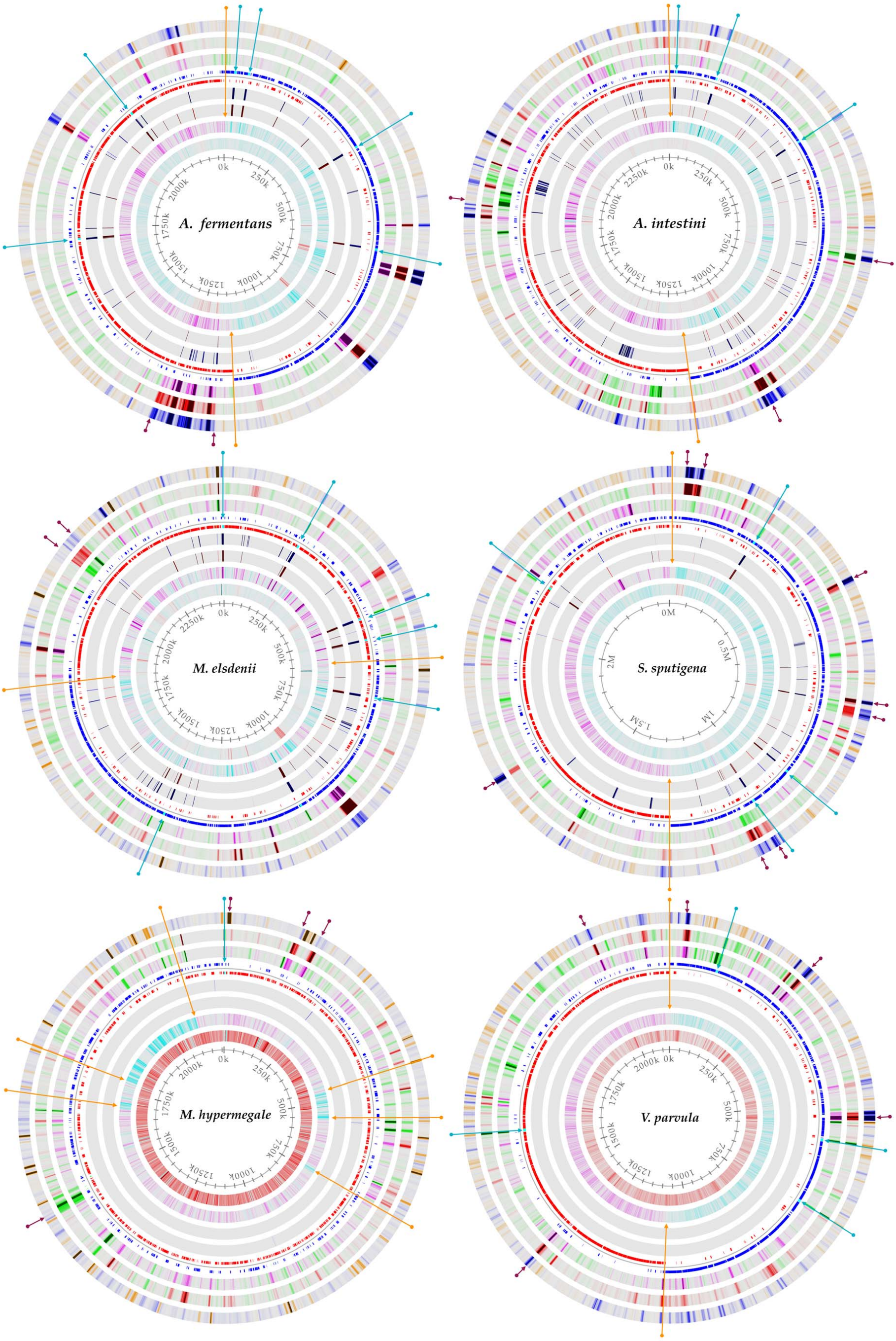
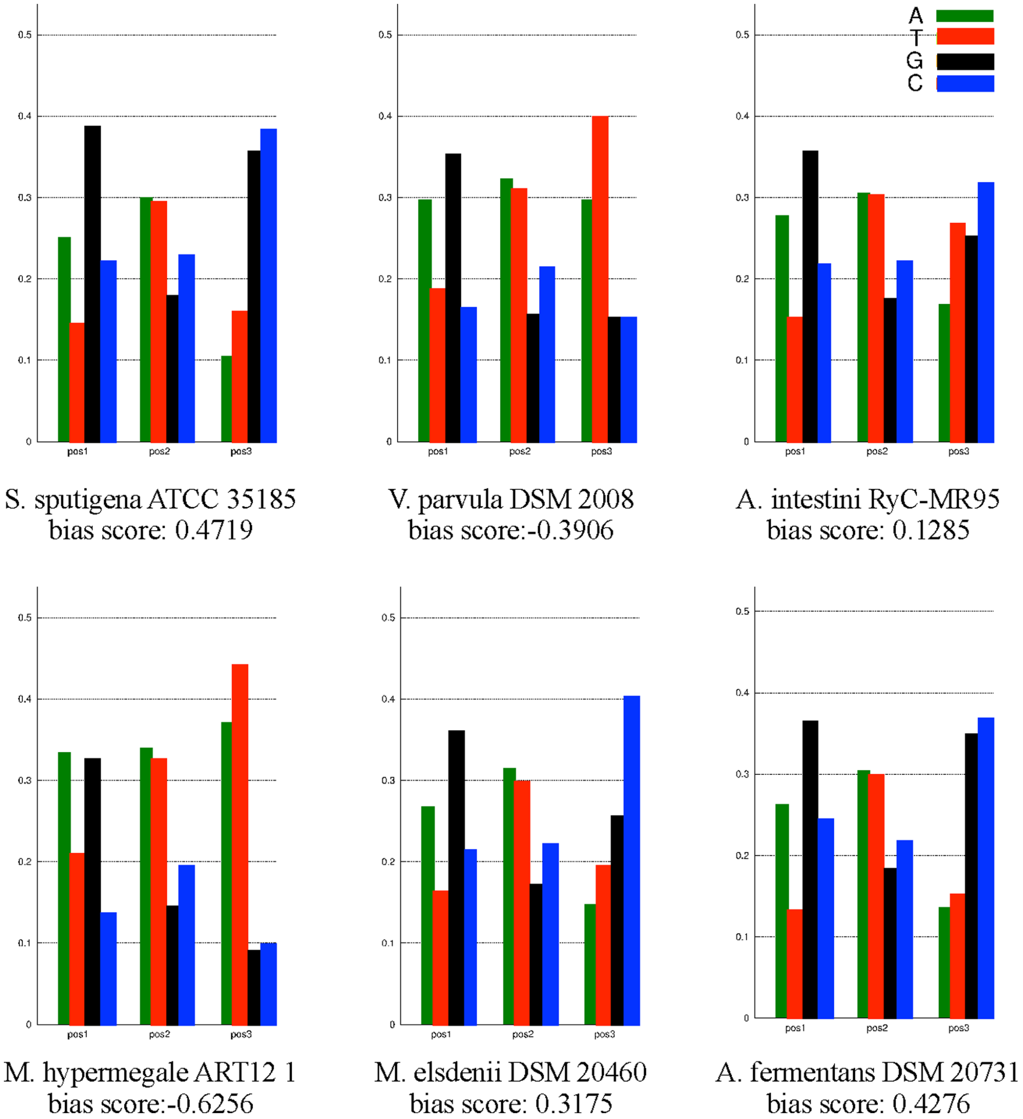

A. fermentans
0k
250k
500k
750k
1000k
1250k
1500k
1750k
2000k
A. intestini
0k
250k
500k
750k
1000k
1250k
1500k
1750k
2000k
2250k
M. elsdenii
0k
250k
500k
750k
1000k
1250k
1500k
1750k
2000k
2250k
S. sputigena
0M
0.5M
1M
1.5M
2M
M. hypermegale
0k
250k
500k
750k
1000k
1250k
1500k
1750k
2000k
V. parvula
0k
250k
500k
750k
1000k
1250k
1500k
1750k
2000k

Figure 3. Genome atlases, DNA structures. A DNA structural atlas was generated for each of the 6 complete genomes. DNA, RNA and gene annotations are from the published GenBank data. Each lane of the circular atlas shows a different DNA feature. From the innermost circle: size of genome (axis), percent AT (red = high AT), GC skew (blue = most G's), inverted and direct repeats (color = repeat), position preference, stacking energy and intrinsic curvature. Orange arrows indicate changes in the skew of G and C, which frequently indicate origin and terminus of replication. Blue arrows show the location of rRNA operons, as annotated in the GenBank file. Dark red arrows highlight areas of the genome that show significantly different DNA structures than the rest of the genome. A higher resolution pdf is available as a supplemental figure. A high resolution figure can be found as supplemental Figure S1.
doi:10.1371/journal.pone.0060120.g003



S. sputigena ATCC 35185
bias score: 0.4719

V. parvula DSM 2008
bias score:-0.3906

A. intestini RyC-MR95
bias score: 0.1285

M. hypermegale ART12 1
bias score:-0.6256

M. elsdenii DSM 20460
bias score: 0.3175

A. fermentans DSM 20731
bias score: 0.4276

Figure 4. Bias in third position. The bias in third codon position is visualized for each of the 6 complete genomes. The bias was defined as −1 in the case of 100% A or T in third position, +1 is the case of 100% G or C.
doi:10.1371/journal.pone.0060120.g004

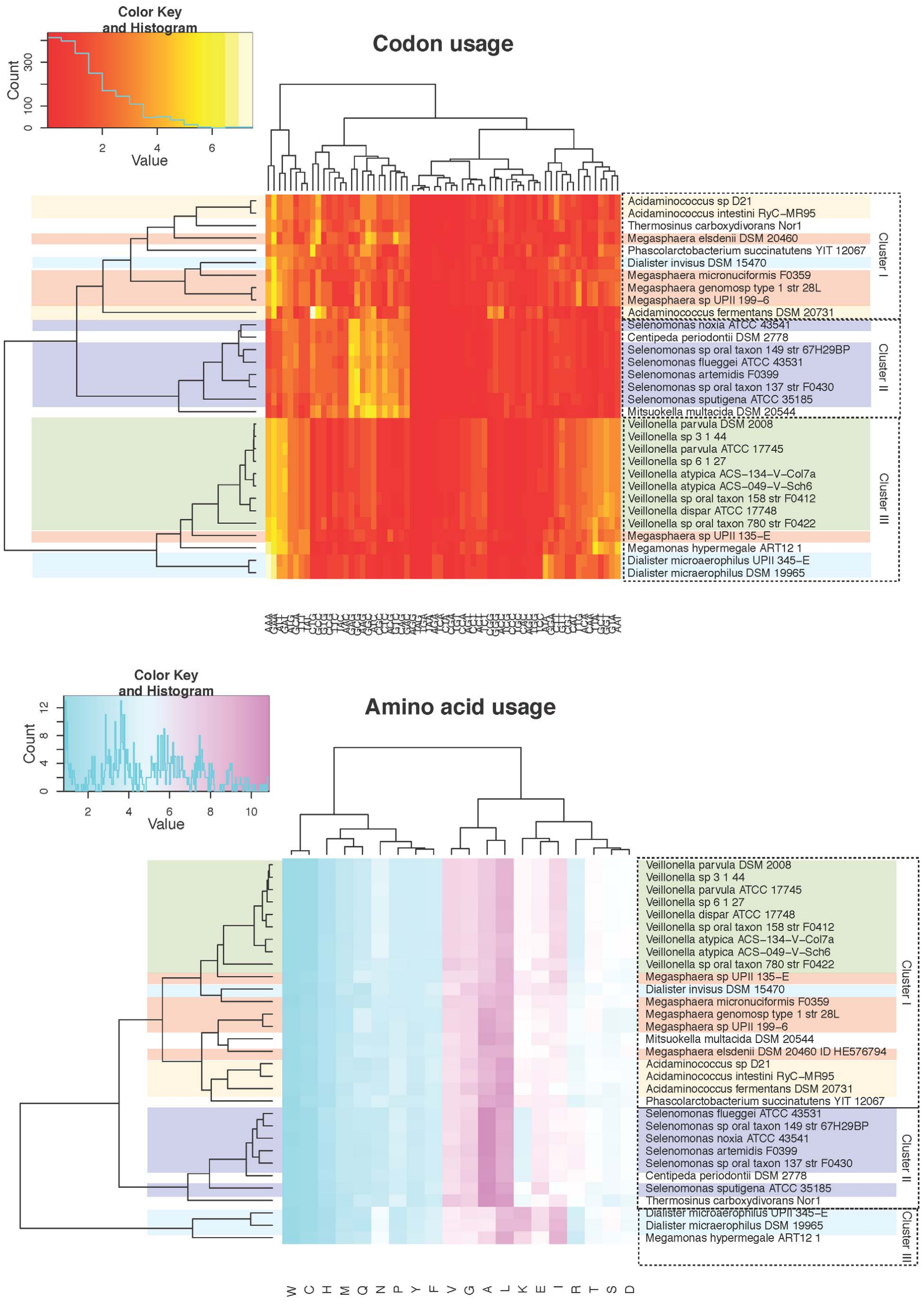

Color Key
and Histogram
Count
0
100
300
2
4
6
Value
Codon usage
Acidaminococcus sp D21
Acidaminococcus intestini RyC–MR95
Thermosinus carboxydivorans Nor1
Megasphaera elsdenii DSM 20460
Phascolarctobacterium succinatutens YIT 12067
Dialister invisus DSM 15470
Megasphaera micronuciformis F0359
Megasphaera genomosp type 1 str 28L
Megasphaera sp UPII 199–6
Acidaminococcus fermentans DSM 20731
Selenomonas noxia ATCC 43541
Centipeda periodontii DSM 2778
Selenomonas sp oral taxon 149 str 67H29BP
Selenomonas flueggei ATCC 43531
Selenomonas artemidis F0399
Selenomonas sp oral taxon 137 str F0430
Selenomonas sputigena ATCC 35185
Mitsuokella multacida DSM 20544
Veillonella parvula DSM 2008
Veillonella sp 3 1 44
Veillonella parvula ATCC 17745
Veillonella sp 6 1 27
Veillonella atypica ACS–134–V–Col7a
Veillonella atypica ACS–049–V–Sch6
Veillonella sp oral taxon 158 str F0412
Veillonella dispar ATCC 17748
Veillonella sp oral taxon 780 str F0422
Megasphaera sp UPII 135–E
Megamonas hypermegale ART12 1
Dialister microaerophilus UPII 345–E
Dialister micraerophilus DSM 19965
Cluster I
Cluster II
Cluster III
Color Key
and Histogram
Count
0
4
8
12
2
4
6
8
10
Value
Amino acid usage
Veillonella parvula DSM 2008
Veillonella sp 3 1 44
Veillonella parvula ATCC 17745
Veillonella sp 6 1 27
Veillonella dispar ATCC 17748
Veillonella sp oral taxon 158 str F0412
Veillonella atypica ACS–134–V–Col7a
Veillonella atypica ACS–049–V–Sch6
Veillonella sp oral taxon 780 str F0422
Megasphaera sp UPII 135–E
Dialister invisus DSM 15470
Megasphaera micronuciformis F0359
Megasphaera genomosp type 1 str 28L
Megasphaera sp UPII 199–6
Mitsuokella multacida DSM 20544
Megasphaera elsdenii DSM 20460 ID HE576794
Acidaminococcus sp D21
Acidaminococcus intestini RyC–MR95
Acidaminococcus fermentans DSM 20731
Phascolarctobacterium succinatutens YIT 12067
Selenomonas flueggei ATCC 43531
Selenomonas sp oral taxon 149 str 67H29BP
Selenomonas noxia ATCC 43541
Selenomonas artemidis F0399
Selenomonas sp oral taxon 137 str F0430
Centipeda periodontii DSM 2778
Selenomonas sputigena ATCC 35185
Thermosinus carboxydivorans Nor1
Dialister microaerophilus UPII 345–E
Dialister micraerophilus DSM 19965
Megamonas hypermegale ART12 1
Cluster I
Cluster II
Cluster III
W
C
H
M
Q
N
P
Y
F
V
G
A
L
K
E
I
R
T
S
D

Figure 5. Amino acid and codon usage heatmaps. Amino acid and codon usage were for all 31 genomes calculated based on the genes identified by gene finding (Prodigal). The percentage of codon and amino acid usage was plotted in two heatmaps using R. The heatmaps were clustered in 2D, thus reordering the organisms and the amino acids/codon to show the shortest distance between them. Dendograms were draw for both and can be used to visualize the difference in usage between organisms.
doi:10.1371/journal.pone.0060120.g005

protein FASTA file can be obtained by extracting proteins from a GenBank file (using saco_extract) or by using the Prodigal genefinder (extract DNA from GenBank, saco_convert, and find genes using prodigalrunner). A BLAST matrix comparison of the 31 *Negativicutes* genomes was calculated on the CMG-biotools system, using 4 processors (calculation time was 9 hours).

The BLAST matrix (Figure 6, high resolution figure as supplemental Figure S2) illustrates that the conservation between genomes is generally higher within a genus than between genera (for example *Selenomonas*, 53–57%, and *Megasphaera*, 33–81%). The *Selenomonas* strains also show a high similarity to the genome of *C. periodontii* DSM 2778, while the *Megasphaera* genus shows no higher similarity to other genera. For both the genomes of *Acidaminococcus* and *Dialister*, the similarity is varied with one comparison being very similar and the others not (31–45%). Within the *Veillonella* genus, the conservation is 64–84% with the exception of *Veillonella* species oral taxon 780 str F0422 (conservation 36–38% to other *Veillonella*). In comparison, a study performed on genomes from the *Vibrionaceae* family showed that different strains of *Vibrio cholerae* share between 70–80% proteins while the similarity to organisms outside the species ranged from 30–45% [27]. From that same study, the internal homology (red squares) ranges from 1.3–5.3%. Other studies, such as a study on *Vibrionaceae* have shown numbers ranging from 1.8–5%. Another study analyzed the similarity between *Enterobacteriaceae* genomes, and found a 76–98.8% similarity between 7 genomes of *Escherichia coli* [28] The same study showed an internal homology of approximately 0.3–3% for the 7 *Escherichia coli*.

The second method looks at the cumulative set of all genes, shared across genomes (pan-genome) and the conserved set of gene families across all genomes (core-genome) [29]. The pan- and core-genomes are theoretical representations of a collective protein pool and a conserved protein pool, respectively. When a protein type is found in all genomes in a collection, it is called a core gene of this collection. Here this is implemented in a pan- and core-genome plot (Figure 7) where sequences are compared using BLAST and the 50/50% cutoff described above. As the clusters grow to more than two members, single linkage clustering is used to assign a new sequence to a group. The program performing this analysis is called pancoreplot and the input is a tab separated text file representing a number of FASTA files containing amino acid sequences (Figure 1, Step 3C). For this analysis, the input files and directories are the same as described for the BLAST matrix.

For the first genome, the pan and core are identical, and the core becomes smaller with the addition of a second genome, as genes in this pool now need to be found in both genomes. If a gene from the core is not found in a new genome it is removed from the core, and is then only part of the pan-genome. The pan-genome is the entire gene pool and as such includes the core genome. The order of the genomes can change the course of the graph, but the final shared gene pool (core and pan-genome) will be the same.

A pan- and core-genome plot analysis was performed for all 31 genomes (Figure 7). The final core genome was found to be 134 gene families and the pan genome contains 17 999 gene families. For an average proteome size of around 1 900 within the *Negativicutes*, a core genome of 134 is relatively small. Using the output data from the pan- and core-genome it was possible to analyze gene overlaps and intersections of the dataset. The core genome of the *Veillonella* genomes is 936 protein families, less than half of the average number of genes in these genomes. Of these families, nly 210 are not found in any of the other genomes (complimentary) while 802 families are not found in the core of the other genomes ("compinter"). The pan-genome of the 31 genomes is 17 999 families, indicating a large diversity and many accessory genes in this class. Compared to similar analyses for genomes of the *Vibrionaceae* family, pan- and core-genome sizes was 20 200 and 1 000 respectively [27]. The *V. cholerae* genomes have a core genome of 2 500, more than 60% of the average size of these genomes, 4 000 genes [27].

Materials and Methods

The CMG-biotools

CMG-biotools is a modified setup of the publicly available Xubuntu 10.10 (www.xubuntu.org/get) operating system. Xubuntu is a community developed operating system that is well-suited for laptops and desktops. It natively contains all applications from word processing and email applications to web server software and programming tools and is part of the Ubuntu project, published under the GPL (GNU General Public License). A number of bioinformatic tools have been added to the system to allow for analysis of microbial genome sequence data and is here called "CMG-biotools". CMG-biotools is an installable operating system (disc image,.iso format). By installing the software, the user accepts the terms of the license and agreements.

The CMG-biotools operating system can be installed on a local computer or on a virtual computer application, such as VirtualBox (www.virtualbox.org). A standard installation should take less than 25 minutes. The functionalities of CMG-biotools consists of a series of compiled executables, Perl, Python and bash scripts contained in a folder on the system (/usr/biotools/). These scripts can be modified according to the individual licenses of the programs (See.*LICENSE* files for this information). The CMG-biotools system is made to run on a local laptop and uses one processor by default. The computationally heavy programs, blastmatrix and pancoreplot, have built-in options (-cpu) that allows the user to increase the number of processors if available.

Download

The installable disk image file (*.iso*) containing CMG-biotools is available from the webpage (www.cbs.dtu.dk/staff/dave/CMGtools/). The tutorials for the courses taught on this platform are available from the same webpage. The system has been tested using VirtualBox, a free virtual computer application, on Windows and Mac operating systems (www.virtualbox.org).

Programs

Data collection. The getgbk.pl script uses the Entrez E-utils programmatic interface made available by the NCBI to fetch sequence data. The script allows searching within the NCBI nuccore or the new bioproject databases using Genbank Accession identifiers or project identifiers respectively. Records identified in *bioprojects* can be filtered to only fetch matches in RefSeq or GenBank. Extraction of DNA from GenBank format is done using saco_convert [12], which locates the DNA sequences in the GenBank data labeled "ORIGIN" and prints the data in FASTA

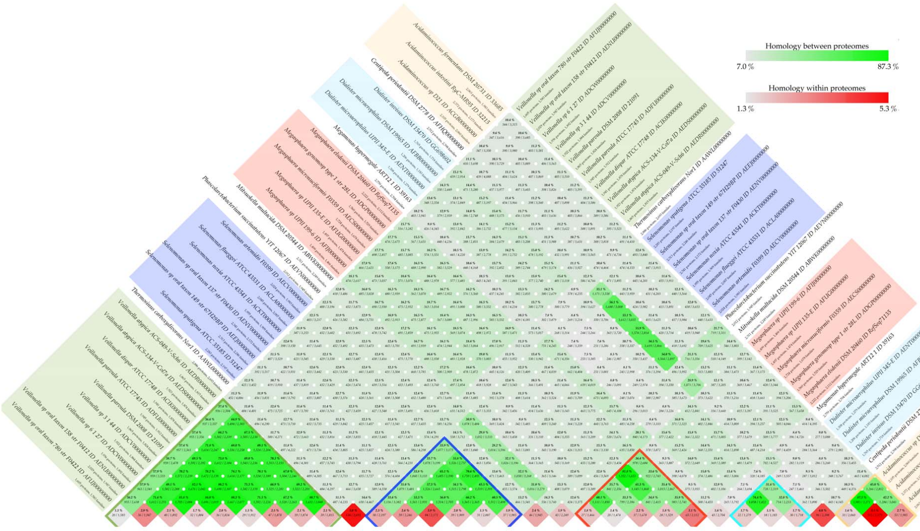


Figure 6. BLAST matrix. An all against all protein comparison was performed using BLAST to define homologs. A BLAST hit is considered significant if 50% of the alignment consists of identical matches and the length of the alignment is 50% of the longest gene. Internal homology (paralogs) is defined as proteins within a genome matching the same 50–50 requirement as for between-proteome comparisons. Self-matches are here ignored. A comparison of 31 *Negativicutes* genomes was performed on the CMG-biotools system (9 hours). A high resolution figure can be found as supplemental Figure S2.
doi:10.1371/journal.pone.0060120.g006

format. Extraction of translated coding sequences from GenBank is done using saco_extract [12]. This program accesses the GenBank data labeled "translation", extracts the sequences and prints them as FASTA format along with the gene identifier, also obtained from the GenBank file. Some GenBank files do not have annotated protein sequences and from these the extraction procedure will not work. In such cases, genefinding should be performed. The input arguments to the saco programs describes input and output file formats, where the first indicates the input file format (for instance GenBank) and the second the output format (for instance FASTA).

Phylogenetic analysis. The RNAmmer [14] program is used for the localization of rRNA sequences in genomic DNA (FASTA format). DNA is extracted from GenBank files using saco_convert and stored in FASTA format (Figure 1, Step 2B). The program uses HMM models to search a DNA sequence for sequences with significant similarity to models of rRNA sequences. Models are included for 5S, 16S and 23S rRNA for bacterial genomes (options TSU, SSU and LSU respectively). For the examples in this paper, each genome sequence was compared to the models for 16S rRNA only. Each sequence is searched and possible rRNA sequences are stored as FASTA formatted DNA sequences. The highest scoring sequence with acceptable length (between 1 400 and 1 800) is extracted from each genome (select16SrRNA) and stored in a FASTA formatted DNA file. It is also possible to use all predicted sequence in stead of selecting the highest scoring one. Some genomes have multiple 16S rRNA sequences and they might yield slightly different phylogenetic relationships. One sequence from each genome is compared in a multiple alignment using ClustalW [15] and the resulting alignment is used to construct a distance tree using 1 000 re-samplings. The tree is visualized using njplot [17].

Genome atlases (structural DNA atlas). The genome atlas presented here is an implementation of the atlas presented earlier by Jensen et al. 1999 [4,6]. Below is a short description of each of the parameters shown in the DNA atlases. Color scales for all parameters follow the same system. The DNA sequence is read and an output file is generated for the various calculated parameters. For each nucleotide in the genome a numerical value is calculated. This file is then read by the GeneWiz program, which calculates the average and standard deviation for each parameter, if the average value of the window is more than 3 standard deviations on either side of the overall average the window is maximally colored. In order to plot the data on a circular map a "window size" is used for longer genomes, which effectively smooths the data for better graphics. For the parameters *Stacking Energy*, *Position Preference* and *Intrinsic Curvature*, the window is 0.002 ×genome length. The window is 0.001 ×genome length for *Percent AT* and *GC skew*. Each of these are calculated separately, wrapped into a pipeline and visualized in a circular plot, called an atlas. The gene annotations are taken directly from a GenBank coding regions; if no such information is found the *CDS*− /+ lanes will be blank. The following lists explanations to each of the lanes in a genome atlas: **Percent AT** is the percent of A's and T's in the genome. **GC skew** is calculated as *((G-C)/(G+C))*, with a window size of 10 000 bp and is useful for determining the origin and terminus of replication [30,31]. **Global Direct Repeats** and **Global Inverted Repeats** refer to a sequence that is present in at least two copies on the same or opposite strands, respectively. **Intrinsic Curvature** is a measure of DNA curvature and is calculated using the CURVATURE program [21,22]. The values are scaled from 0 (e.g. no curvature) to 1, which is the curvature of DNA when wrapped around the nucleosome. **Stacking Energy** is derived from the dinucleotide values provided by Ornstein et al

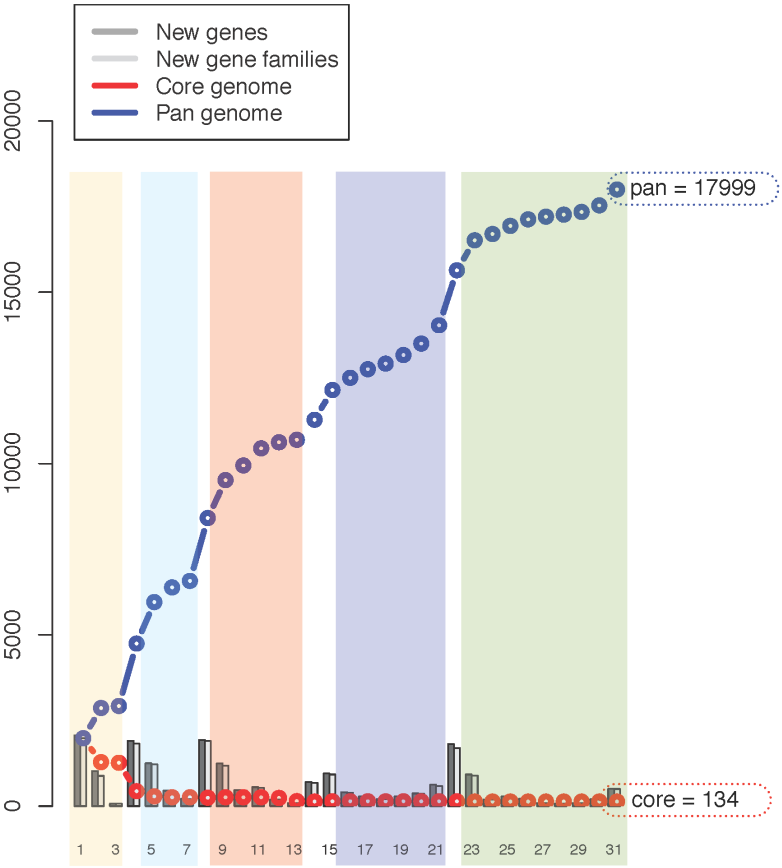


Figure 7. Core and pan genome using BLAST. A pan- and core-genome calculation was performed using BLAST. A BLAST cutoff of 50% identity and 50% coverage of the longest gene was used. If two proteins within a genome matched according to the 50/50% cutoff, they were clustered into one protein family. Protein families were extended via single linkage clustering. If a protein family includes proteins from all genomes in the comparison, the family is a core protein family.
doi:10.1371/journal.pone.0060120.g007

[20]. The scale is in kcal/mol, and the dinucleotide values range from −3.82 kcal/mol (will unstack easily) to −14.59 kcal/mol (difficult to unstack). A positive peak in base-stacking (i.e., numbers closer to zero) reflects regions of the helix which would de-stack or melt more readily. Conversely, minima (larger negative numbers) in this plot would represent more stable regions of the chromosome. **Position Preference** is a measure of preferential location of sequences within nucleosomal core sequences [19]. The trinucleotide values range from essentially zero (0.003, presumably more flexible), to 0.28 (considered rigid). Since very few of the trinucleotide have values close to zero (e.g. little preference for nucleosome positioning), this measure is considered to be more sensitive towards the low ("flexible") end of the scale.

Gene finding. Gene finding is performed using the program Prodigal [13]. The program is wrapped into a formatting program called prodigalrunner. The program reformats the raw output of Prodigal to FASTA formatted open reading frames, DNA and amino acids, along with a draft of a GenBank file and a raw general feature formatted file, a.gff file. The Prodigal program allows for different parameter modifications, including training (prodigalrunner -t <organism>) of the gene finder using given data. This feature increases the computation time of the algorithm, but for less known organisms this feature might improve gene finding. It should be noted that the default behavior when encountering N's is not changed - the program treats runs of N's as masked sequence and does not build genes across them. The CMG-Biotools system also comes with the native Prodigal program, which can be used as published [13].

Amino acid and codon usage. The amino acid and codon usage is calculated using BioPerl modules [32], and is a simple calculation of the fraction of each amino acid or codon count of the total count of amino acids or codons. The bias in third position is found by counting the number of each base on each position in each codon, divided by the total number of codons. The bias in the third position between G/C and A/T was then calculated as *sum(GC)-sum(AT)*, so that 100% GC in third codon position is +1 and −1 for 100% AT. The plots are made using Perl and Gnuplot.

Proteome comparisons using BLAST. The BLAST matrix is a visual presentation of a pairwise proteome comparison using BLAST (Basic Local Alignment Tool) [26]. All sequences are compared to each other and a BLAST hit is significant when 50% of the alignment is identical matches and the length of the alignment is 50% of the longest gene in the comparison. If two sequences are similar according to the cutoff, they are collected in one "protein family". For the comparison of two genomes, protein families are built through single linkage, so that each shared connection must be between sequences from different genomes (shaded green). Paralogs are traditionally defined as a gene which has undergone duplication before speciation; in the BLAST matrix, an internal hit significantly similar to the query protein is grouped into the same gene family. The bottom row of the matrix shows the number of proteins that have homologous hits within the proteome itself (shaded red). The color scales are set automatically from the highest to lowest value observed, but can be changed manually. The procedure is implemented in the program blastmatrix, which takes a XML formatted input file. The input file is created by the program makebmdest.

The pan- and core-genome plot is a different use of BLAST for comparing proteomes (using the 50/50 cutoff as described above). The core-genome consists of protein families with representatives found in all investigated genomes. The pan-genome is the entire set of protein families from all genomes in the comparison. The first genome in the analysis has a core-genome equal to the pan-genome. The addition of an second genome reduces the core-genome of the two genomes and increases the pan-genome. Each sequence of a new genome is compared to a representative from each of the existing gene-families. If the new sequence matches, the family is a core-family, if the sequence does not match a family it becomes a new protein family. When all new sequences have been compared to existing gene-families, core families that did not have a representative in the latest added genome are removed from the core-genome of the genome comparison. The change in the pan- and core-genome is followed as two lines (blue and red, respectively). The number of new proteins, along with how many new protein families that corresponds to, is indicated as gray bars on the plot. The program (pancoreplot), produces a plot and a table which can be used to look up the underlying values of the plot.

The pan- and core-genome calculations can be used to extract subsets of genes for different genome sets. The program that implements this is called specificGenes and works on the BLAST output from the pancoreplot program. The procedure is based on mathematical set theory and works with intersections, unions and complementary genesets. Each genome is treated as a set and the intersection is the gene families that two or more sets have "in common". The intersection of genome A and B, is the set of all gene families which are found in both A and B. The union of two or more sets refers to the gene families which are found in either genome A or B. Calculating the complimentary families of a genome refers to the set of all families which are members of A but not members of B. In the comparative genomic analysis, the sets usually consists of more than one genome, such as the intersection of genome A, B and C while not found (complimentary) in genome D, E and F. This will give families that are found in A, B and C but not found in any of D, E or F. It is also possible to calculate the situation where families are found in A, B and C but not found in the intersection of D, E and F, this is referred to as the "compinter". For more details, see the CMG-biotools manual.

Supporting Information

Figure S1 Genome atlases, DNA structures (Figure 3 at High-Resolution).
(PDF)

Figure S2 BLAST matrix (Figure 6 at High-Resolution).
(PDF)

Acknowledgments

We would like to thank Peter Fisher Hallin for his implementations of good ideas and for contributing his code and work to these studies. Another thanks goes to Hans-Henrik Stæ rfeldt for his help and contribution. We would also like to thank the many people involved in the work on these tools over the years at DTU CBS. A special thanks go to all the students have contributed to the improvements of this tool.

Author Contributions

Conceived and designed the experiments: DU TV. Performed the experiments: TV OA. Analyzed the data: KL TV. Contributed reagents/materials/analysis tools: KL OA TV. Wrote the paper: TV KL DU.

References

1. Casali N, Nikolayevskyy V, Balabanova Y, Ignatyeva O, Kontsevaya I, et al. (2012) Microevolution of extensively drug-resistant tuberculosis in Russia. Genome Research : 735–745.
2. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus inuenza* Rd. Science 269: 496–512.
3. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, et al. (1995) The minimal gene complement of *Mycoplasma genitalium*. Science 270: 397–403.
4. Jensen LJ, Friis C, Ussery DW (1999) Three views of microbial genomes. Research in Microbiology 150: 773–777.
5. Friis C, Jensen LJ, Ussery DW (2000) Visualization of pathogenicity regions in bacteria. Genetica 108: 47–51.
6. Pedersen aG, Jensen LJ, Brunak S, Staerfeldt HH, Ussery DW (2000) A DNA structural atlas for *Escherichia coli*. Journal of Molecular Biology 299: 907930.
7. Snipen L, Almø y T, Ussery DW (2009) Microbial comparative pan-genomics using binomial mixture models. BMC Genomics 10: 385.
8. Pilo P, Frey J (2011) Bacillus anthracis: Molecular taxonomy, population genetics, phylogeny and patho-evolution. Infection, Genetics and Evolution 11: 12181224.
9. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, et al. (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pangenome". Proceedings of the National Academy of Sciences of the United States of America 102: 1395013955
10. DistroWatch (accessed 17/09/2012). http://distrowatch.com/dwres.php?resource = popularity
11. Benson Da, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2011) GenBank. Nucleic Acids Research 39: D32–D37.
12. Jensen L, Knudsen S (1999) Automatic discovery of regulatory patterns in promoter regions based on whole cell expression data and functional annotation. Bioinformatics 16: 326–333.
13. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, et al. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11: 119.
14. Lagesen K, Hallin P, Rø dland EA, Staerfeldt HH, rn Rognes T, et al. (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Research 35: 3100–3108.
15. Larkin Ma, Blackshields G, Brown NP, Chenna R, McGettigan Pa, et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947–2948.
16. Felsenstein J (1985) Confidence limits on phylegenies an approach using bootstrap. Evolution 39: 783–791.
17. Perrière G, Gouy M (1996) WWW-query: an on-line retrieval system for biological sequence banks. Biochimie 78: 364–369.

18. Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proceedings of the National Academy of Sciences of the United States of America 74: 5088–5090.
19. Satchwell SC, Drew HR, Travers AA (1986) Sequence periodicities in chicken nucleosome core DNA. Journal of Molecular Biology 191: 659–675.
20. Ornstein RL, Rein R, Breen DL, Macelroy RD (1978) An optimized potential function for the calculation of nucleic acid interaction energies. I - Base stacking. Biopolymers 17: 2341–2360.
21. Shpigelman ES, Trifonov EN, Bolshoy A (1993) CURVATURE: software for the analysis of curved DNA. Computer Applications in the Biosciences CABIOS 9: 435–440.
22. Bolshoy A, McNamara P, Harrington RE, Trifonov EN (1991) Curved DNA without A-A: experimental estimation of all 16 DNA wedge angles. Proceedings of the National Academy of Sciences of the United States of America 88: 2312–3216.
23. Marín A, Xia X (2008) GC skew in protein-coding genes between the leading and lagging strands in bacterial genomes: new substitution models incorporating strand bias. Journal of Theoretical Biology 253: 508–513.
24. Altschul SS, Gish W, Miller W, Myers EE, Lipman D, et al. (1990) Basic Local Alignment Search Tool. Journal of Molecular Biology 215: 403–410.
25. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25: 3389–3402.
26. Binnewies TT, Hallin PF, Staerfeldt HH, Ussery DW (2005) Genome Update: proteome comparisons. Microbiology (Reading, England) 151: 1–4.
27. Vesth T, Wassenaar TM, Hallin PF, Snipen L, Lagesen K, et al. (2010) On the Origins of a Vibrio Species. Microbial Ecology 59: 1–13.
28. Willenbrock H, Petersen A, Sekse C, Kiil K, Wasteson Y, et al. (2006) Design of a seven-genome *Escherichia coli* microarray for comparative genomic profiling. Journal of Bacteriology 188: 7713–7721.
29. Klockgether J, Würdemann D, Wiehlmann L, Binnewies TT, Ussery DW, et al. (2008) Genome Diversity of *Pseudomonas aeruginosa*. Chapter 2 in Pseudomonas: Genomics and Molecular Biology, (Edited by: Pierre Cornelis, Caister Academic Press).
30. Lobry JR (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. Molecular Biology and Evolution 13: 660–665.
31. Worning P, Jensen LJ, Hallin PF, Staerfeldt HH, Ussery DW (2006) Origin of replication in circular prokaryotic chromosomes. Environmental Microbiology 8: 353–361.
32. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, et al. (2002) The Bioperl toolkit: Perl modules for the life sciences. Genome Research 12: 1611–1618.

5.2 *Veillonella*, between Gram positives and negatives

This paper represents a project illustrates how much variation can be found between genome sequences and how difficult it is to analyze sequences without homology to known and annotated sequences. A set of 24 genomes of the *Negativicutes* class was conducted using the tools provided in the CMG-biotools, to identify unique genetic features for this group. The initial interest in this specific genus arose from the fact that they all stain Gram negative in the Gram cell wall structure test, while in all other aspects are more closely related to Gram positive organism. Although Gram status does not always correlate with taxonomic clustering, this specific genus groups closely with all Gram positives (taxonomy based on 16S rRNA). During the course of the study, it became evident that this group shows very little sequence homology to other organisms. The 24 genomes were compared to a set of diverse genomes using sequence similarity methods such as BLAST and 16S rRNA alignments as well as feature based methods such as Composition Vector Trees and DNA tetramer frequencies. The metabolic potential of each genome analyzed using the Kyoto Encyclopedia of Genes and Genomes (KEGG).

Based on 16S rRNA, complete genomic DNA sequences, and a consensus tree based on conserved proteins, comparisons showed that the *Negativicutes* are only distantly related to *Clostridia*, but are even less related to Gram-negative species. Analyzing genomes of the *Veillonella* genus, under the *Negativicutes* class, showed a total of 1.350 protein were found in all *Veillonella* genomes (core genes), although less than half of these were found in any *Clostridium* genome. Only 27 proteins were found conserved in all analyzed genomes. *Veillonella* has distinct metabolic properties, and significant similarities to other genomes were not detected, with the exception of a shared LPS biosynthesis pathway. The *Negativicutes* exhibits unique properties, most of which are shared with Gram-positives and some with Gram negatives. They are only distantly related to *Clostridia*, but are even less related to Gram-negative species, based on protein and sequence comparisons. Though the *Negativicutes* stain Gram-negative and possess two membranes, the genome and proteome analysis presented here confirm their taxonomic placement. This project highlighted the problem of annotation by sequence homology, but also the need for being able to compare the functional markup of several organisms.

Standards in Genomic Sciences (2013) **9**:431-448 DOI:10.4056/sigs.2981345

Veillonella, Firmicutes: Microbes disguised as Gram negatives

Tammi Vesth[1], Aslı Ozen[1,5], Sandra C. Andersen[1], Rolf Sommer Kaas[1], Oksana Lukjancenko[1], Jon Bohlin[2], Intawat Nookaew[3], Trudy M. Wassenaar[4] and David W. Ussery[6*]

[1]Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Lyngby, Denmark
[2]Norwegian School of Veterinary Science, Department of Food Safety and Infection Biology, Oslo, Norway
[3]Department of Chemical and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden
[4]Molecular Microbiology and Genomics Consultants, Zotzenheim, Germany
[5]The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark
[6]Comparative Genomics Group, Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA

*Correspondence: Dave Ussery (dave@cbs.dtu.dk;)

The *Firmicutes* represent a major component of the intestinal microflora. The intestinal *Firmicutes* are a large, diverse group of organisms, many of which are poorly characterized due to their anaerobic growth requirements. Although most *Firmicutes* are Gram positive, members of the class *Negativicutes*, including the genus *Veillonella*, stain Gram negative. *Veillonella* are among the most abundant organisms of the oral and intestinal microflora of animals and humans, in spite of being strict anaerobes. In this work, the genomes of 24 *Negativicutes*, including eight *Veillonella* spp., are compared to 20 other *Firmicutes* genomes; a further 101 prokaryotic genomes were included, covering 26 phyla. Thus a total of 145 prokaryotic genomes were analyzed by various methods to investigate the apparent conflict of the *Veillonella* Gram stain and their taxonomic position within the *Firmicutes*. Comparison of the genome sequences confirms that the *Negativicutes* are distantly related to *Clostridium* spp., based on 16S rRNA, complete genomic DNA sequences, and a consensus tree based on conserved proteins. The genus *Veillonella* is relatively homogeneous: inter-genus pairwise comparison identifies at least 1,350 shared proteins, although less than half of these are found in any given *Clostridium* genome. Only 27 proteins are found conserved in all analyzed prokaryote genomes. *Veillonella* has distinct metabolic properties, and significant similarities to genomes of *Proteobacteria* are not detected, with the exception of a shared LPS biosynthesis pathway. The clade within the class *Negativicutes* to which the genus *Veillonella* belongs exhibits unique properties, most of which are in common with Gram-positives and some with Gram negatives. They are only distantly related to *Clostridia*, but are even less closely related to Gram-negative species. Though the *Negativicutes* stain Gram-negative and possess two membranes, the genome and proteome analysis presented here confirm their place within the (mainly) Gram positive phylum of the *Firmicutes*. Further studies are required to unveil the evolutionary history of the *Veillonella* and other *Negativicutes*.

Background

The genus *Veillonella*, belonging to *Negativicutes*, consists of anaerobic, non-fermentative, Gram-negative cocci, that are normally observed in pairs or short chains, and are non-sporulating and non-motile [1]. *Veillonella* spp. are abundant in the human microbiome and are found in the oral, respiratory, intestinal and genitourinary flora of humans and animals; they can make up as much as 10% of the bacterial community initially colonizing the enamel [2] and are found throughout the entire oral cavity [3], especially on the tongue dorsum and in saliva [4]. The importance of *Veillonella* spp. in human infections is uncertain, and they are generally considered to be of low virulence. *Veillonella* form biofilms, often with *Streptococcus* spp., and species of these genera have been found to be more abundant in the oral microflora of people with poor oral health [5]. Studies have shown that during formation of early dental plaque, the fraction of *Veillonella* spp. changes in mixed-microbial colonies with streptococci [6]. Thus, *Veillonella* spp. may play a role in caries formation as they utilize the lactic acid produced by the organisms conducive to caries [7]. *Veillonella* are also among the most common anaerobic species reported from pulmonary samples and are frequently recovered from cystic fibrosis cases [8]. The organisms are also abundant in the human gut flora, where their numbers were found to be higher in children with type I diabetes compared to healthy controls [9]. Currently, 12 species of *Veillonella* have been characterized [10,11] including *V. parvula*, *V. atypica* and *V. dispar*, which are found in the human oral cavity.

The *Negativicutes* are the only diderm (literally 'two skins') members of the phylum *Firmicutes* as they possess an inner and an outer membrane. Their placement within the *Firmicutes* has been widely accepted, and has been confirmed by 16S rRNA analysis [12]. However, their genomes have not been analyzed in detail to confirm their taxonomic position. This work presents a broad analysis of the *Negativicutes* with focus on the *Veillonella* spp. using comparative microbial genomics. A total of 24 genomes from the *Negativicutes* were compared to 121 genomes covering most of the taxonomic span of sequenced bacterial genomes. We investigated how the *Negativicutes* genomes compared to other bacterial genomes using three different and complementary approaches: 1) phylogenetic trees to visualize the relative distance of the *Negativicutes* genomes to other genomes; 2) amino acid composition, nucleotide tetramer frequency and metabolism analysis using 2-D clustering and heatmaps to compare genomes; and 3) proteomic comparison across the *Negativicutes* genomes.

Materials and Methods

Genome sequences used for analysis

The set of 145 genomes included in this study (24 *Negativicutes* genomes and 121 other prokaryotic genomes covering 26 phyla) are listed in Table 1.

Table 1. Genomes used in this study

Phylum	Name of organism and strain	Strain designation	Type strain	NCBI Taxon ID	NCBI Project ID
Acidobacteria	*Acidobacterium capsulatum*	ATCC 51196	Yes	240015	28085
Acidobacteria	*"Korebacter versatiles"*	Ellin 345		204669	15771
Acidobacteria	*"Solibacter usitatus"*	Ellin6076		234267	12638
Actinobacteria	*Bifidobacterium bifidum*	317B	No	1681	42863
Actinobacteria	*Catenulispora acidiphila*	ID139908, DSM 44928	Yes	479433	21085
Actinobacteria	*Corynebacterium pseudotuberculosis*	C231	No	681645	40875
Actinobacteria	*Segniliparus rugosus*	ATCC BAA-974	Yes	679197	40685
Actinobacteria	*Streptomyces* bingchenggensis	BCW-1	Name not validly published	749414	46847
Actinobacteria	*Tropheryma whipplei*	Twist	Yes	*203267*	*95*
Aquificae	*Persephonella marina*	EX-H1	Yes	123214	12526
Aquificae	*Sulfurihydrogenibium sp.*	YO3AOP1	No type strain available	436114	18889
Aquificae	*Thermocrinis albus*	HI 11/12, DSM 14484	Yes	638303	37275
Bacteroidetes	*Bacteroides thetaiotaomicron*	VPI-5482	Yes	226186	399
Bacteroidetes	*Candidatus* Sulcia muelleri	DMIN		641892	37785
Bacteroidetes	*Chitinophaga pinensis*	UQM 2034, DSM 2588	Yes	485918	27951
Bacteroidetes	*Paludibacter propionicigenes*	WB4, DSM 17365	Yes	694427	42009
Chlamydiae	*Protochlamydia amoebophila*	UWE25	Yes	264201	10700
Chlamydiae	*Chlamydia trachomatis*	E/Sweden2	No	634464	43167
Chlamydiae	*Chlamydophila pneumoniae*	AR39	No	115711	247
Chlamydiae	*Waddlia chondrophila*	WSU 86-1044	Yes	716544	43761

Table 1. Genomes used in this study (cont.)

Phylum	Name of organism and strain	Strain designation	Type strain	NCBI Taxon ID	NCBI Project ID
Chlorobi	*"Chlorobium chlorochromatii"*	CaD3	Name not validly published	340177	13921
Chlorobi	*Chlorobium tepidum*	TLS	Yes	194439	302
Chloroflexi	*Chloroflexus aggregans*	DSM 9485	Yes	326427	16708
Chloroflexi	*Dehalococcoides sp*	BAV1	No	216389	15770
Chloroflexi	*Herpetosiphon aurantiacus*	ATCC 23779	Yes	316274	16523
Chloroflexi	*Roseiflexus sp.*	RS-1	No type strain available	357808	16190
Cyanobacteria	*Anabaena variabilis* 3	ATCC 2941	No	240292	10642
Cyanobacteria	*Cyanothece sp.*	PCC 7822	No	497965	28535
Cyanobacteria	*Prochlorococcus marinus*	MIT9301	No	167546	15746
Cyanobacteria	*Synechocystis sp.*	PCC6803	No	1148	60
Deferribacteres	*Calditerrivibrio nitroreducens*	Yu37-1, DSM 19672	Yes	768670	49523
Deferribacteres	*Deferribacter desulfuricans*	SSM1, DSM 14783	Yes	197162	37285
Deferribacteres	*Denitrovibrio acetiphilus*	N2460, DSM 12809	Yes	522772	29431
Deinococcus-Thermus	*Oceanithermus profundus*	506, DSM 14977	Yes	670487	40223
Deinococcus-Thermus	*Thermus thermophilus*	HB8	Yes	300852	13202
Deinococcus-Thermus	*Truepera radiovictrix*	RQ-24, DSM 17093	Yes	649638	38371
Dictyoglomi	*Dictyoglomus turgidum*	DSM 6724	Yes	515635	29175
Elusimicrobia	*Elusimicrobium minutum*	Pei 191	Yes	445932	19701
Fibrobacteres	*Fibrobacter succinogenes*	S85	Yes	59374	32617
Firmicutes	*Acetohalobium arabaticum*	Z-7288, DSM 5501	Yes	574087	32769
Firmicutes	*Acidaminococcus fermentans*	VR4, DSM 20731	Yes	591001	33685
Firmicutes	*Acidaminococcus sp.*	D21	No type strain available	563191	34117
Firmicutes	*Alkaliphilus oremlandii*	OhILAs	Yes	350688	16083
Firmicutes	*Bacillus subtilis subsp. subtilis*	168	Yes	224308	76
Firmicutes	*Clostridium botulinum*	F Langeland	No	441772	19519
Firmicutes	*Clostridium cellulolyticum*	H10	Yes	394503	17419
Firmicutes	*Clostridium difficile*	630 (epidemic type X)	No	272563	78
Firmicutes	*"Desulfotomaculum reducens"*	MI-1	Name not validly published	349161	13424
Firmicutes	*Dialister invisus*	DSM 15470	Yes	592028	33143
Firmicutes	*Dialister micraerophilus*	Oral Taxon 843 DSM 19965	Yes	888062	53029
Firmicutes	*Dialister micraerophilus*	UPII-345-E	No	910314	59521
Firmicutes	*Enterococcus faecalis*	V583	No	226185	70
Firmicutes	*Eubacterium cylindroides*	T2-87	No	717960	45917
Firmicutes	*Eubacterium rectale*	A1-86, DSM 17629	No	39491	39159

Table 1. Genomes used in this study (cont.)

Phylum	Name of organism and strain	Strain designation	Type strain	NCBI Taxon ID	NCBI Project ID
Firmicutes	*Exiguobacterium sibiricum*	255-15	Yes	262543	10649
Firmicutes	*Geobacillus kaustophilus*	HTA426	Yes	235909	13233
Firmicutes	*Lactococcus lactis*	cremoris MG1363	No	416870	18797
Firmicutes	*Lysinibacillus sphaericus*	C3-41	No	444177	19619
Firmicutes	*Megamonas hypermegale*	ART12/1	No	158847	39163
Firmicutes	*Megasphaera* genomo sp.	type 128L	No type strain available	699218	42553
Firmicutes	*Megasphaera micronuciformis*	F0359	No	706434	43125
Firmicutes	*Mitsuokella multacida*	A 405-1, DSM 20544	Yes	500635	28653
Firmicutes	*Paenibacillus sp.*	JDR-2	No	324057	20399
Firmicutes	*Phascolarctobacterium sp.*	YIT 12067	No	626939	48505
Firmicutes	*Selenomonas artemidis*	F0399	No	749551	47277
Firmicutes	*Selenomonas flueggei*	ATCC 43531	Yes	638302	37273
Firmicutes	*Selenomonas noxia*	ATCC 43541	Yes	585503	34641
Firmicutes	*Selenomonas sp.*	Oral Taxon 137 F0430	No type strain available	879310	52055
Firmicutes	*Selenomonas sp.*	Oral Taxon 149 67H29BP	No type strain available	864563	50535
Firmicutes	*Selenomonas sputigena*	DSM 20758	Yes	546271	51247
Firmicutes	*Staphylococcus aureus aureus*	ED98	No	681288	39547
Firmicutes	*Streptococcus pneumoniae*	TIGR4	No	170187	277
Firmicutes	*Thermoanaerobacter* sp.	X514	Name not validly published	399726	16394
Firmicutes	*Thermosinus carboxydivorans*	Nor1	Yes	401526	17587
Firmicutes	*Turicibacter sp.*	PC909 702450 42765	No		
Firmicutes	*Veillonella atypica*	ACS-049-V-Sch6	No	866776	51075
Firmicutes	*Veillonella atypica*	ACS-134-V-Col7a	No	866778	51079
Firmicutes	*Veillonella dispar*	ATCC 17748	Yes	546273	30491
Firmicutes	*Veillonella parvula*	ATCC 17745	No	686660	41557
Firmicutes	*Veillonella parvula*	Te3, DSM 2008	Yes	479436	21091
Firmicutes	*Veillonella sp.*	3 1 44	Name not validly published	457416	41975
Firmicutes	*Veillonella sp.*	6 1 27	Name not validly published	450749	41977
Firmicutes	*Veillonella sp.*	Oral Taxon 158 F0412	Name not validly published	879309	52053
Fusobacteria	*Fusobacterium nucleatum nucleatum*	ATCC 25586	Yes	190304	295
Fusobacteria	*Ilyobacter polytropus*	CuHBu1, DSM 2926	Yes	572544	32577
Fusobacteria	*Leptotrichia buccalis*	C-1013-b, DSM 1135	Yes	523794	29445
Fusobacteria	*Sebaldella termitidis*	NCTC 11300	Yes	526218	29539

Table 1. Genomes used in this study (cont.)

Phylum	Name of organism and strain	Strain designation	Type strain	NCBI Taxon ID	NCBI Project ID
Fusobacteria	*Streptobacillus moniliformis*	9901, DSM 12112	Yes	519441	29309
Planctomycetes	*Pirellula staleyi*	DSM 6068	Yes	530564	29845
Planctomycetes	*Planctomyces limnophilus*	Mu 290, DSM 3776	Yes	521674	29411
Proteobacteria	*Acinetobacter baumannii*	SDF	No	509170	13001
Proteobacteria	*Alkalilimnicola ehrlichii*	MLHE-1	Yes	187272	15763
Proteobacteria	*Arcobacter nitrofigilis*	DSM 7299	Yes	572480	32593
Proteobacteria	*Burkholderia xenovorans*	(fungorum) LB400	Yes	266265	254
Proteobacteria	*Campylobacter jejuni*	doylei 269.97	No	360109	17163
Proteobacteria	*Candidatus Pelagibacter ubique*	SAR11 HTCC1062	Name not validly published	335992	13989
Proteobacteria	*Candidatus Zinderia insecticola*	CARI	Name not validly published	871271	51243
Proteobacteria	*Cellvibrio japonicus*	Ueda107	Yes	498211	28329
Proteobacteria	*Cupriavidus taiwanensis*	LMG19424	Yes	164546	15733
Proteobacteria	*Escherichia coli*	K-12, MG1655	No	511145	225
Proteobacteria	*Geobacter uraniireducens*	Rf4	Yes	351605	15768
Proteobacteria	*Hahella chejuensis*	KCTC 2396	Yes	349521	16064
Proteobacteria	*Haliangium ochraceum*	SMP-2, DSM 14365	Yes	502025	28711
Proteobacteria	*Helicobacter pylori*	908	No	869727	50869
Proteobacteria	*Lawsonia intracellularis*	PHE/MN1-00	No	363253	183
Proteobacteria	*Magnetococcus* sp.	MC-1	Name not validly published	156889	262
Proteobacteria	*Methylobacterium nodulans*	ORS2060	Yes	460265	20477
Proteobacteria	*Neisseria meningitidis*	Z2491	No	122587	252
Proteobacteria	*Neorickettsia sennetsu*	Miyayama	Yes	222891	357
Proteobacteria	*Nitrosomonas eutropha*	C91 (C71)	Yes	335283	13913
Proteobacteria	*Photorhabdus luminescens laumondii*	TT01	Yes	243265	9605
Proteobacteria	*Polynucleobacter necessarius*	STIR1	No	452638	19991
Proteobacteria	*Pseudomonas aeruginosa*	LESB58	No	557722	31101
Proteobacteria	*Pseudomonas fluorescens*	SBW25	No	216595	31229
Proteobacteria	*Pseudomonas stutzeri*	A1501	No	379731	16817
Proteobacteria	*Salmonella enterica enterica*	PT4 P125109	No	550537	30687
Proteobacteria	*Shewanella oneidensis*	MR-1	Yes	211586	335
Proteobacteria	*Sorangium cellulosum*	So ce56	No	448385	28111
Proteobacteria	*Stigmatella aurantiaca*	DW4 /3-1	No	378806	52561
Proteobacteria	*Sulfurospirillum deleyianum*	5175, DSM 6946	No	525898	29529
Proteobacteria	*Vibrio cholerae*	O395	No	345073	32853
Spirochaetes	*Borrelia turicatae*	91E135	Yes	314724	13597
Spirochaetes	*Brachyspira murdochii*	56-150, DSM 12563	Yes	526224	29543
Spirochaetes	*Leptospira interrogans*	lai 56601	No	189518	293
Synergistetes	*Thermanaerovibrio acidaminovorans*	Su883, DSM 6589	Yes	525903	29531

Table 1. Genomes used in this study (cont.)

Phylum	Name of organism and strain	Strain designation	Type strain	NCBI Taxon ID	NCBI Project ID
Tenericutes	*Acholeplasma laidlawii*	PG-8A	No	441768	19259
Tenericutes	*Candidatus* Phytoplasma asteris	yellows witches'-broom AY-WB 322098	Name not validly published	13478	
Tenericutes	*Candidatus* Phytoplasma mali	AT	Name not validly published	37692	25335
Tenericutes	*Mycoplasma genitalium*	G37	Yes	243273	97
Tenericutes	*Mycoplasma pneumoniae*	FH	No	722438	49525
Tenericutes	*Ureaplasma parvum*	sv 3, ATCC 27815	No	505682	19087
Thermotogae	*Fervidobacterium nodosum*	Rt17-B1	Yes	381764	16719
Thermotogae	*Kosmotoga olearia*	TBF 19.5.1	Yes	521045	29419
Thermotogae	*Petrotoga mobilis*	SJ95	Yes	403833	17679
Thermotogae	*Thermotoga naphthophila*	RKU-10	Yes	590168	33663
Verrucomicrobia	*Akkermansia muciniphila*	ATCC BAA-835	Yes	349741	20089
Verrucomicrobia	*Opitutus terrae*		Yes	PB90-1	452637
Crenarchaeota	*Sulfolobus solfataricus*	P2		273057	108
Crenarchaeota	*Thermosphaera aggregans*	M11TL, DSM 11486	Yes	633148	36571
Euryarchaeota	*Halogeometricum borinquense*	PR3, DSM 11551	Yes	469382	20743
Euryarchaeota	*Methanocella sp.*	RC-I	Name not validly published	351160	19641
Euryarchaeota	*Methanothermus fervidus*	V24S, DSM 2088	Yes	523846	33689
Korarchaeota	*Candidatus* Korarchaeum cryptofilum	OPF8	Name not validly published	374847	16525
Nanoarchaeota	*"Nanoarchaeum equitans"*	Kin4-M	Name not validly published	228908	9599

16S rRNA tree

For this analysis, 16S rRNA sequences were predicted from the whole genome sequences of the selected organisms, using the RNAmmer algorithm [13]. These sequences were aligned using the MAFFT program, with the iterative refinement algorithm using maximum iteration (1000) and default parameters for gap penalties [14]. A distance tree was constructed using MEGA5 [15] with the Neighbor-joining algorithm [16] and 1,000 bootstrap resamplings. The taxa in the resulting tree were collapsed to phyla, except for the *Negativicutes*.

Composition Vector Tree (CV)

A Composition Vector Tree was constructed based on protein sequences of the 145 selected genomes using a webserver (available at tlife.fudan.edu.cn/cvtree) with the K parameter set at 6 [17]. The outcome from the program is a distance matrix based on amino acid sequence comparisons, which is then used to generate a phylogenetic tree with the neighbor-joining method. In the shown tree, the outgroup chosen was *Methanothermus fervidus* (an *Archaea*). After tree visualization with MEGA5, branches were collapsed wherever possible with the exception of the *Negativicutes* branch, which remained expanded.

Consensus tree of conserved genes

Using the list of universally conserved core genes, previously identified by Ciccarelli *et al.* [18], and an implementation of BLAST, a set of genes that was shared among all 145 genomes was identified. Proteins that had no match in at least one genome or showed poor E-value were eliminated. The 27 conserved core genes were extracted (Table 1) and a multiple alignment was produced using MUSCLE software [19]. A set of phylogenetic trees was constructed by PAUP [20] and a best-fit consensus tree was generated using Phylogeny Inference package (PHYLIP) as described elsewhere [21]. Bootstrap values were found after 27 re-samplings, which is equal to the number of gene families conserved in all the analyzed genomes.

DNA tetramer analysis and amino acid usage

A tetramer frequency heatmap was constructed from the observed ratios of tetra-nucleotide frequencies divided by estimated tetra-nucleotide frequencies for each genome [22]. The estimated tetra-nucleotides were computed from the genomes' base composition. The ratio of observed over expected frequency was used for hierarchical clustering using complete linkage and Euclidean distance, which was subsequently performed with respect to both strain and tetramer frequencies.

The amino acid heatmap is based on frequencies of deduced proteomic amino acids from each genome normalized with respect to the total number of amino acids in each genome. The amino acid frequencies for each genome were clustered using complete linkage and Euclidean distance with respect to both genomes and amino acids. The heatmap was made using the R package ggplot2 [23].

Comparison of metabolism potential

The protein sequences of Kyoto Encyclopedia of Genes and Genomes (KEGG) orthology categories [24] were downloaded and only the Bacterial sequences were considered. The Hidden Markov model (HMM) of each ortholog was generated using HMMER version 3 [25] based on the multiple alignment of each orthologous set of KEGG proteins, using MUSCLE software [19]. The 145 proteomes were queried against the HMMs to infer their ontology. A cutoff of 1×10^{-30} was used for statistical significance. A heatmap of each pathway and process derived from the database KEGG was illustrated based on normalized abundance of the enzymes present in each pathway. The heatmap and hierarchical clustering were performed in the software R [23].

Construction of BLAST matrix and proteome comparison

Reciprocal BLAST was performed between each genome pair. The program blastall version 2.2.25 was used for BLAST implementation using default settings (BLASTp, E-value set to 1×10^{-5} for non-homologs and 1×10^{-8} for homologs, without filtering). A hit was considered significant at a BLAST cutoff of 95% identity and 95% coverage (of the longest gene in comparison). The number of hits was then given as a percentage of the genes in the column representing the corresponding genome. The diagonal designates internal homologs, computed by blasting each genome with itself. To avoid including identical genes, the second highest scoring hits were used. Furthermore, we also performed homology reduction of the diagonal hits, using an implementation of the Hobohm algorithm [26].

Results

Twenty-four *Negativicutes* genomes were compared to 121 other prokaryotic genomes covering 22 Bacterial and 4 Archaeal phyla. When available, at least two genomes were included for every phylum. The first analysis presented here is based on 16S rRNA alignments. A single 16S rRNA gene was extracted from each of the genomes and an alignment was produced spanning the maximum length of the gene. A phylogenetic tree was constructed based on this alignment, as shown in Figure 1. With the exception of the *Negativicutes*, branches of the tree were collapsed in those cases where the analyzed species within a phylum clustered together. With the exception of some *Firmicutes*, the analyzed genomes cluster according to their phylum, although the *Deferribacteres* phylum is mixed with the *Proteobacteria* phyla, and two members of *Proteobacteria* are not positioned with other members of their phylum (*Lawsonia intracellularis* and *Magnetococcus*). That most phyla could be collapsed is consistent with the weight of 16S rRNA similarities in currently accepted taxonomic descriptions of prokaryotes. The *Firmicutes*, however, show less consistency. Although most of the analyzed *Firmicutes* cluster together, two species are separated from the *Firmicutes* branch (*Eubacterium cylindroides* and *Thermoanaerobacter* sp., both members of *Clostridia*). The *Negativicutes* are positioned within the *Firmicutes* cluster, and this part of the tree is expanded in the figure for clarity. As can be seen, phylogeny of the 16S rRNA gene provides good resolution between the different genera of the analyzed *Negativicutes*. All *Veillonella* spp. are clustered within one branch of the *Negativicutes*. The *Acidaminococcaceae* (to which *Phascolarctobacterium* spp. also belong) are placed within the cluster of the *Veillonellaceae*, in accordance with their current classification [27]. The *Acidaminococcaceae* used to be recognized as a separate family within the *Negativicutes*, just like the *Veillonellaceae*, and during preparation of this contribution these two families were presented as such in the Taxonomy database at NCBI. Of note is the relatively close relationship between *Negativicutes* and two *Clostridium* species (*C. botulinum* and *C. cellulolyticum*), which does not cluster with other members of the *Clostridium* genus (Figure 1). That genus displays a high degree of variation and re-classification of some of the members of this genus is in progress (see for example [27]). That two members of the *Clostridia* are even placed outside the *Firmicutes* phylum is an indication of 16S rRNA gene sequence heterogeneity within this class.

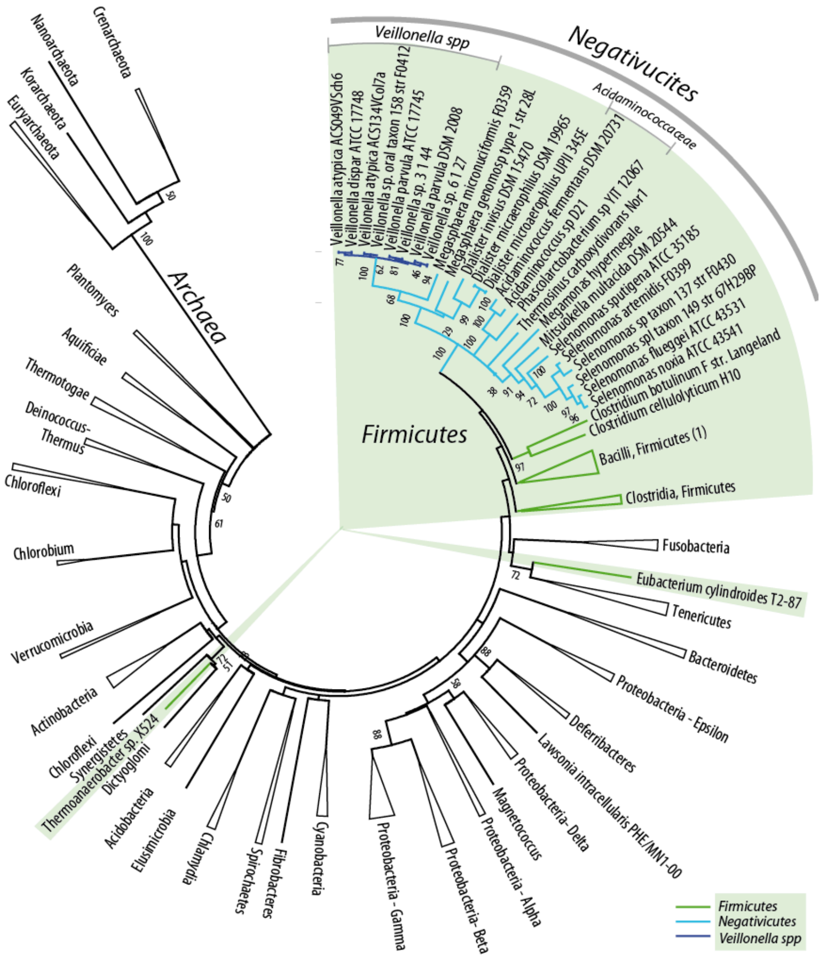

Veillonella spp
Negativucites
Acidaminococcaceae
Veillonella atypica ACS049VSch6
Veillonella dispar ATCC 17748
Veillonella atypica ACS134VCol7a
Veillonella sp. oral taxon 158 str F0412
Veillonella parvula ATCC 17745
Veillonella sp. 3 1 44
Veillonella parvula DSM 2008
Veillonella sp. 6 1 27
Megasphaera micronuciformis F0359
Megasphaera genomosp type 1 str 28L
Dialister invisus DSM 15470
Dialister micraerophilus DSM 19965
Dialister microaerophilus UPII 345E
Acidaminococcus fermentans DSM 20731
Acidaminococcus sp D21
Phascolarctobacterium sp YIT 12067
Thermosinus carboxydivorans Nor1
Megamonas hypermegale
Mitsuokella multacida DSM 20544
Selenomonas sputigena ATCC 35185
Selenomonas artemidis F0399
Selenomonas sp taxon 137 str F0430
Selenomonas spl taxon 149 str 67H29BP
Selenomonas flueggei ATCC 43531
Selenomonas noxia ATCC 43541
Clostridium botulinum F str. Langeland
Clostridium cellulolyticum H10
Bacilli, Firmicutes (1)
Clostridia, Firmicutes
Firmicutes
77
100
62
81
46
94
68
100
99
100
100
29
100
100
38
91
94
72
100
100
97
96
97
Crenarchaeota
Nanoarchaeota
Korarchaeota
Euryarchaeota
50
100
Archaea
Plantomyces
Aquificae
Thermotogae
Deinococcus-Thermus
Chloroflexi
Chlorobium
50
61
Verrucomicrobia
Actinobacteria
Chloroflexi
Synergistetes
Thermoanaerobacter sp. X524
Dictyoglomi
72
57
68
Acidobacteria
Elusimicrobia
Chlamydia
Spirochaetes
Fibrobacteres
Cyanobacteria
88
Proteobacteria - Gamma
Proteobacteria- Beta
58
Proteobacteria - Alpha
Magnetococcus
Proteobacteria- Delta
Lawsonia intracellularis PHE/MN1-00
Deferribacteres
88
Proteobacteria - Epsilon
Bacteroidetes
Tenericutes
72
Eubacterium cylindroides T2-87
Fusobacteria
Firmicutes
Negativicutes
Veillonella spp

Figure 1. Phylogenetic neighbor-joining tree based on 16S rRNA genes extracted from 145 genomes (24 *Negativicutes* and 121 prokaryotic genomes representing 26 phyla). Bootstrap values of 50 and higher are indicated. With the exception of the *Negativicutes*, branches where all organisms belong to the same phyla are collapsed and named by the phyla they represent. The green shading indicates the position of *Firmicutes*. The collapsed branch of the *Bacilli*, marked (1), contains *Turicibacter sanguinis*, a *Firmicutes* member of the *Erysipelotrichales* as well as *Bacilli* members. An uncollapsed tree is included in the supplementary material.

Next, all protein-coding genes of the analyzed genomes were compared and a composition vector tree (CVtree) was produced, based on amino acid sequences (Figure 2). The topology of the resulting tree is generally in accordance with the 16S rRNA tree shown in the previous figure. As indicated by the collapsed branches, the CVtree grouped most genomes according to their known taxonomic phyla, although not all *Spirochaetes* cluster together. In contrast to the 16S rRNA tree, in this protein tree all the *Firmicutes* cluster together, and are distinct from other phyla. The *Negativicutes* genomes, nested within the *Firmicutes*, again have the *Acidaminococcaceae* placed within the *Veillonellaceae*, while all *Veillonella* spp. are found in one cluster. All *Clostridia*, this time divided into two collapsed branches, are positioned as the closest relatives to *Negativicutes*. It is of interest that among the closest relatives to *Firmicutes*, based on this analysis, are the *Fusobacteria* and the *Elusimicrobia*; these are atypical diderm bacteria that produce lipopolysaccharides [28]. However, the spirochete, *Brachyspira murdochii*, does not possess two membranes, but is nevertheless grouped with atypical diderms. On the other hand while the *Synergistetes* are atypical diderm bacteria, they are placed elsewhere in the tree (Figure 2).

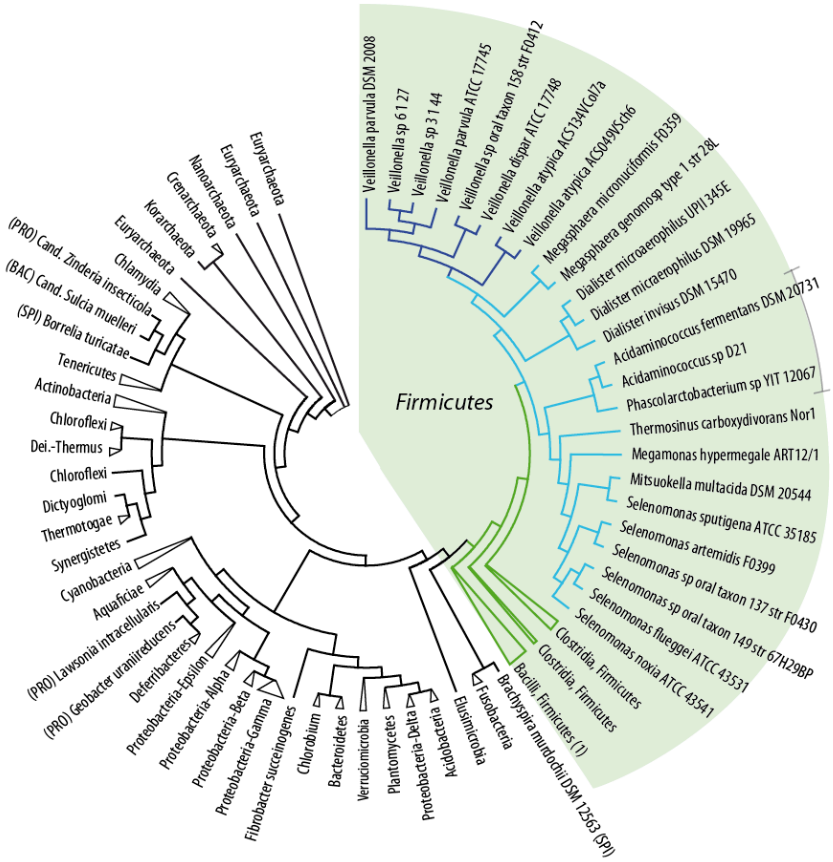

Firmicutes
Veillonella parvula DSM 2008
Veillonella sp 6 1 27
Veillonella sp 3 1 44
Veillonella parvula ATCC 17745
Veillonella sp oral taxon 158 str F0412
Veillonella dispar ATCC 17748
Veillonella atypica ACS134VCol7a
Veillonella atypica ACS049VSch6
Megasphaera micronuciformis F0359
Megasphaera genomosp type 1 str 28L
Dialister microaerophilus UPII 345E
Dialister micraerophilus DSM 19965
Dialister invisus DSM 15470
Acidaminococcus fermentans DSM 20731
Acidaminococcus sp D21
Phascolarctobacterium sp YIT 12067
Thermosinus carboxydivorans Nor1
Megamonas hypermegale ART12/1
Mitsuokella multacida DSM 20544
Selenomonas sputigena ATCC 35185
Selenomonas artemidis F0399
Selenomonas sp oral taxon 137 str F0430
Selenomonas sp oral taxon 149 str 67H29BP
Selenomonas flueggei ATCC 43531
Selenomonas noxia ATCC 43541
Clostridia, Firmicutes
Clostridia, Firmicutes
Bacilli, Firmicutes (1)
Brachyspira murdochii DSM 12563 (SPI)
Fusobacteria
Elusimicrobia
Acidobacteria
Proteobacteria-Delta
Plantomycetes
Verruciomicrobia
Bacteroidetes
Chlorobium
Fibrobacter succinogenes
Proteobacteria-Gamma
Proteobacteria-Beta
Proteobacteria-Alpha
Proteobacteria-Epsilon
Deferribacteres
(PRO) Geobacter uraniireducens
(PRO) Lawsonia intracellularis
Aquificae
Cyanobacteria
Synergistetes
Thermotogae
Dictyoglomi
Chloroflexi
Dei.-Thermus
Chloroflexi
Actinobacteria
Tenericutes
(SPI) Borrelia turicatae
(BAC) Cand. Sulcia muelleri
(PRO) Cand. Zinderia insecticola
Chlamydia
Euryarchaeota
Korarchaeota
Crenarchaeota
Nanoarchaeota
Euryarchaeota
Euryarchaeota

Figure 2. Phylogenetic tree based on composition vector analysis (CVtree) of all protein coding genes (amino acid sequences) derived from the analyzed genomes. Note that the branch lengths in this plot are artificial. The coloring is the same as in Figure 1 and branches have been collapsed. The *Firmicutes* branch *Bacilli*, marked (1), contains *Turicibacter sanguinis*. An uncollapsed tree is included in the supplementary material.

A third analysis was based on a subset of proteins found conserved amongst all analyzed genomes. These conserved proteins were selected based on a protein BLAST (a cutoff of 50% identity and 50% coverage of the query length was used) and single linkage clustering. The analysis identified 29 genes that are shared among all 145 genomes [Table 2]. A consensus tree was constructed based on these 29 conserved proteins (Figure 3). The results confirm the global observations of the other two phylogenetic analyses: the *Negativicutes* cluster together and are most closely related to *Clostridia* (in this case the most closely related species are *Desulfotomaculum reducens* and *Acetohalobium arabaticum*). As before, the *Acidaminococcaceae* cluster together but within the *Veillonellaceae*. The position of *Turicibacter sanguinis* within the *Bacilli* group of *Firmicutes* is consistent with the other two trees but contrasts with its taxonomic description at NCBI as a member of the *Erysipelotrichia*.

Table 2. Universally conserved COGs

Group	Average length (aa)	Annotation
COG0012	380	Predicted GTPase, probable translation factor
COG0016	423	Phenylalanine-tRNA synthethase alpha subunit
COG0048	137	Ribosomal protein S12
COG0049	182	Ribosomal protein S7
COG0052	240	Ribosomal protein S2
COG0080	154	Ribosomal protein L11
COG0081	230	Ribosomal protein L1
COG0087	288	Ribosomal protein L3
COG0091	157	Ribosomal protein L22
COG0092	240	Ribosomal protein S3
COG0093	130	Ribosomal protein L14
COG0094	182	Ribosomal protein L5
COG0096	131	Ribosomal protein S8
COG0097	177	Ribosomal protein L6P/L9E
COG0098	220	Ribosomal protein S5
COG0100	145	Ribosomal protein S11
COG0102	167	Ribosomal protein L13
COG0103	172	Ribosomal protein S9
COG0172	442	Seryl-tRNA synthetase
COG0184	154	Ribosomal protein S15P/S13E
COG0186	122	Ribosomal protein S17
COG0197	175	Ribosomal protein L16/L10E
COG0200	166	Ribosomal protein L15
COG0201	445	Preprotein translocase subunit SecY
COG0202	323	DNA-directed RNA polymerase, alpha subunit
COG0256	178	Ribosomal protein L18
COG0495	854	Leucyl-tRNA synthetase
COG0522	199	Ribosomal protein S4 and related proteins
COG0533	375	Metal-dependent proteases with chaperone activity

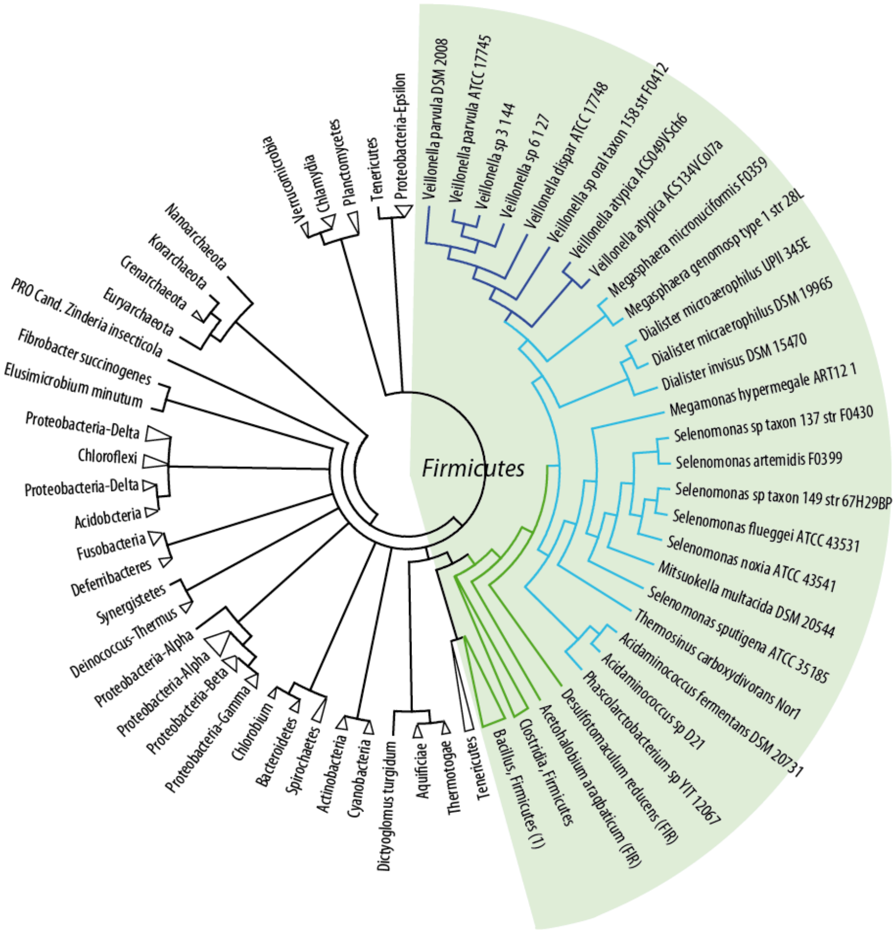

Nanoarchaeota
Korarchaeota
Crenarchaeota
Euryarchaeota
PRO Cand. Zinderia insecticola
Fibrobacter succinogenes
Elusimicrobium minutum
Proteobacteria-Delta
Chloroflexi
Proteobacteria-Delta
Acidobcteria
Fusobacteria
Deferribacteres
Synergistetes
Deinococcus-Thermus
Proteobacteria-Alpha
Proteobacteria-Alpha
Proteobacteria-Beta
Proteobacteria-Gamma
Chlorobium
Bacteroidetes
Spirochaetes
Actinobacteria
Cyanobacteria
Dictyoglomus turgidum
Aquificiae
Thermotogae
Tenericutes
Bacillus, Firmicutes (1)
Clostridia, Firmicutes
Acetohalobium arabaticum (FIR)
Desulfotomaculum reducens (FIR)
Phascolarctobacterium sp YIT 12067
Acidaminococcus sp D21
Acidaminococcus fermentans DSM 20731
Thermosinus carboxydivorans Nor1
Selenomonas sputigena ATCC 35185
Mitsuokella multacida DSM 20544
Selenomonas noxia ATCC 43541
Selenomonas flueggei ATCC 43531
Selenomonas sp taxon 149 str 67H29BP
Selenomonas artemidis F0399
Selenomonas sp taxon 137 str F0430
Megamonas hypermegale ART12 1
Dialister invisus DSM 15470
Dialister micraerophilus DSM 19965
Dialister microaerophilus UPII 345E
Megasphaera genomosp type 1 str 28L
Megasphaera micronuciformis F0359
Veillonella atypica ACS134VCol7a
Veillonella atypica ACS049VSch6
Veillonella sp oral taxon 158 str F0412
Veillonella dispar ATCC 17748
Veillonella sp 6 1 27
Veillonella sp 3 1 44
Veillonella parvula ATCC 17745
Veillonella parvula DSM 2008
Proteobacteria-Epsilon
Tenericutes
Planctomycetes
Chlamydia
Verrucomicrobia
Firmicutes

Figure 3. Consensus tree based on the phylogenetic trees of 27 genes conserved in all 145 genomes. The collapsed branch of the *Bacilli,* marked (1), contains *Turicibacter sanguinis*. An uncollapsed tree is available as a supplemental figure.

In conclusion, based on three independent phylogenetic analyses, the closest relatives to the *Negativicutes* seem to be the *Clostridiaceae*. The observed clustering of species within the *Negativicutes* is consistent with their assigned taxonomy. Furthermore, these analyses show that *Veillonella spp.* form a distinct branch, most different from the other *Negativicutes*, while the recent change of status of the *Acidaminococcaceae* (they are no longer a separate family) is confirmed by these analyses.

Apart from comparing proteins and genes, genomes can also be compared based on nucleotide composition irrespective of their coding capacity. For instance, the frequency of nucleotide combinations can reveal similarities between genomes that are independent of protein-coding information. We compared the frequency of tetranucleotides for all 145 genomes. The observed frequency of all 64 tetranucleotide combinations was extracted for each genome and these frequencies were divided by the theoretically calculated, expected frequencies (corrected for differences in base composition). This ratio, which could be interpreted as a genomic signature, was expected to reflect taxonomic divisions [29]. However, although the analysis identified a high similarity in tetranucleotide frequency for all of the analyzed *Veillonella* genomes, most of the clustering observed was not in accordance with known taxonomic relationships. Not only were *Negativicutes* other than *Veillonella* separated from each other and strewn across the phyla, but also several other *Firmicutes* were distributed over various branches (data shown as supplementary material). In fact, for most of the analyzed genomes, members of identical phyla did not cluster together and even the *Archaea* were mixed with *Bacteria*, although some closely related species were indeed clustered. This may explain why all *Veillonella* genomes grouped together. Several organisms with similar tetranucleotide frequencies did not share a common ecological niche, in contrast to previously reported observations (reviewed in [30]). Neither was the obtained clustering dictated by GC-content. The conclusion from this analysis was that tetranucleotide analysis is only taxonomically informative for closely related genomes.

We also compared whole-genome amino acid frequencies in each of the deduced proteomes. Although the results are slightly more in agreement with known taxonomy as compared with the genomic signatures discussed above, this analysis does not cluster organisms according to their phyla, and again some *Archaea* are mixed with *Bacteria*. The relevant part of the heatmap based on amino acid frequency is shown in Figure 4. All *Veillonella* genomes cluster together within the *Negativicutes*, with the exception of two of the three *Dialister* genomes, which are found most closely related to *Clostridium* species (See supplemental information for a version of this figure showing all the genomes). The major *Negativicutes* cluster also contains a *Geobacillus* (which is a Gram-positive *Firmicutes*) and a methanogenic Archaean. Interestingly, the closest relatives to this cluster are not *Clostridia*, as the previous phylogenetic trees suggest, but a number of *Proteobacteria*. It is striking that the amino acid frequency analysis detects similarities to *Proteobacteria*, with which the *Negativicutes* have their two membranes in common.

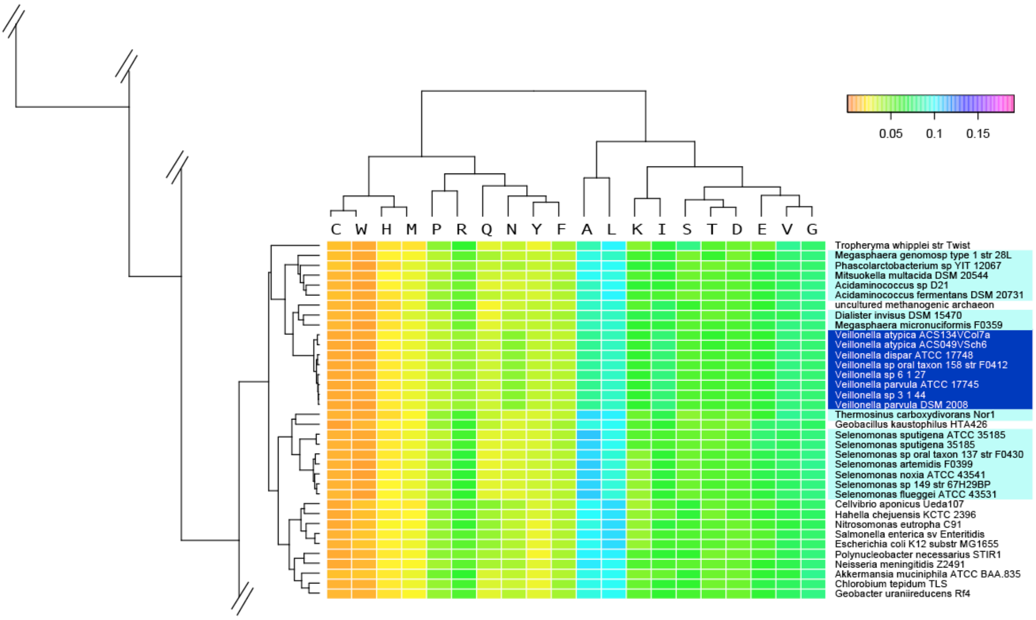


Figure 4. A zoomed heatmap of the amino acid frequency found in the deduced proteomes of all 145 genomes. A fragment of the heatmap is shown, presenting the cluster in which all but two *Negativicutes* are found. The remaining two, both *Dialister microaerophilus* genomes, are positioned elsewhere in the tree, closest to *Clostridium cellulolyticum* (not shown in this zoom). The color scale indicates highly underrepresented (orange) to highly overrepresented amino acid frequency (magentum). The full figure is available as supplementary information.

The metabolic properties encoded by the genomes were analyzed next, based on KEGG comparisons [24]. The results are again visualized in a heatmap (Figure 5). We hypothesized that this analysis could identify similarities based on niche adaptation. For simplicity, only a selected number of phyla are shown: apart from the *Firmicutes*, genomes are included that represent *Bacteroidetes* and *Proteobacteria* (both of which contain members frequently found in the oral or gut microbiome), while *Cyanobacteria* are included as representatives of a phylum that occupy an environmental niche. Since the genomes are compared based on predicted proteomes, their annotation was standardized in order to reduce artificial variation caused by gene annotation differences. As can be seen in Figure 5, the *Veillonella* genomes all cluster together at the right-hand side of the plot, within a larger cluster containing most of the other *Negativicutes* and some *Firmicutes*. The three *Dialister* species are placed outside the *Negativicutes* cluster. The other *Firmicutes* that are found combined with the *Negativicutes*, based on their metabolic potential, are *Clostridium cellulolyticum, Eubacterium rectale, Lactococcus lactis, Streptococcus pneumoniae* and *Turicibacter sanguinis*. These are all common members of the oral or intestine microbiome. As expected, the metabolic pathway for lipopolysaccharide biosynthesis is shared between the *Negativicutes* and other Gram-negative species, as indicated by the arrows in Figure 5. Interestingly, the *Cyanobacteria* form a small cluster within, not outside the tree, together with a *Haliangium* and a *Sorangium* species as their closest neighbors (both are social *Myxococcales* belonging to the *Deltaproteobacteria*). The exclusive ability of carbon fixation by *Cyanobacteria* is apparent from the dark red square in the block 'energy'. The lanes of *Veillonella* in Figure 5 are dominated by light colors, indicative of medium metabolic potential; that is, in contrast to some genomes where most of the pathways are present (dark red for Proteobacteria for example) or missing (dark green for other *Negativicutes*), the *Veillonella* genomes have partial pathways (based on knowledge primarily from aerobic genomes). There is no reason to believe that the *Veillonella* genomes should have less metabolic potential than other *Negativicutes*. Indeed, it is likely that the differences in metabolic potential of *Veillonella* are truly reflective of alternative capabilities for these bacteria.

It was further investigated how conserved the predicted proteomes are within the *Negativicutes*. As a quantitative measure for homology, shared protein-coding genes were identified by pairwise BLASTP comparison and expressed as a percentage of the combined proteomes. The results are shown in a matrix (Figure 6). In addition to the proteomes of the 24 *Negativicutes*, the comparison includes *Clostridium botulinum*, *Cl. cellulolyticum* and *Desulfotomaculum **reducens***, as these *Firmicutes* were shown to share characteristics with *Negativicutes* in previous analyses (*cf.* Figures 1 and 3). The proteome of *E. coli* K12 is included as an example of a Gram-negative intestinal bacterium. The BLAST matrix was constructed using reciprocal best BLAST hits to determine the presence of shared protein family between two genomes. Inspection of Figure 6 shows that the genus *Veillonella* is relatively homogeneous; any two members of this genus share between 67% and 90% homology (1,357 to 1,682 protein families), irrespective of the species. The genus *Selenomonas* is more heterogeneous, with pairwise homology varying from 42% to 82% between any two species (980 to 1659 protein families). The three proteomes of *Dialister spp.*, covering two species, share between 40% and 84% homology. The highest homologous fraction identified between two members of different genera within the *Negativicutes* is 43% (*Mitsuokella multacida* compared to *Selenomonas sputigena*, whereas the lowest homology is 15% (*Dialister* spp. compared to *Thermosinus carboxydivorans*). *Negativicutes* share between 9% and 33% homology with the analyzed *Firmicutes*, whereas slightly lower homology is detected with *E. coli* (between 7% and 24%).

Finally, we assessed the gene pool conserved within all analyzed *Negativicutes*. Using the same cutoff for protein BLAST comparison as before, a core-genome is identified that contains about 300 conserved protein families (data not shown). This is a relatively low number of conserved proteins, reflective of the extensive genetic heterogeneity within this bacterial class.

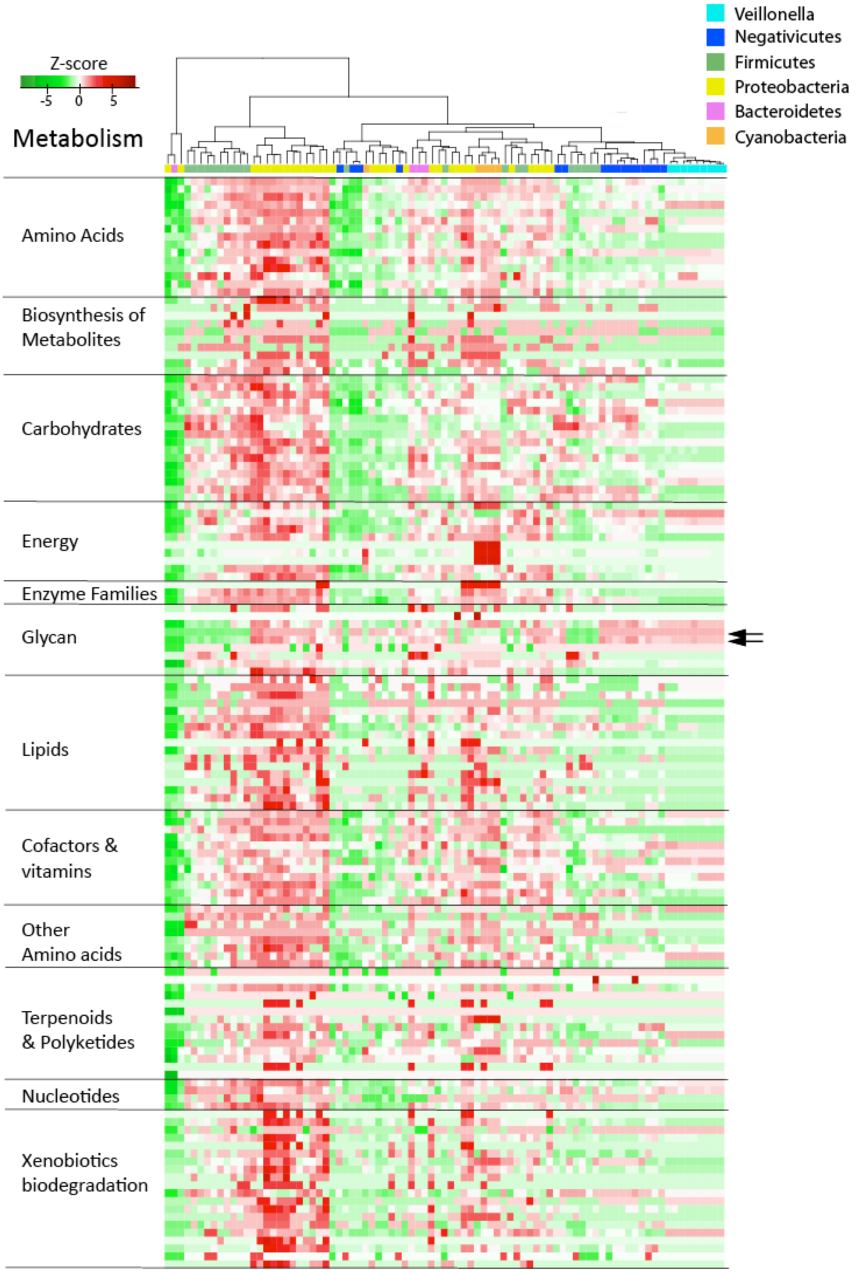


Figure 5. Heatmap of metabolism potential, based on Kyoto Encyclopedia of Genes and Genomes ontology (KEGG). The green color in the heatmap indicates weak metabolic potential, while red signals strong potential. The arrows to the right indicate the scores for lipopolysaccharide biosynthesis. A version summarizing the metabolism pathways and showing the species legend is available as supplementary material.

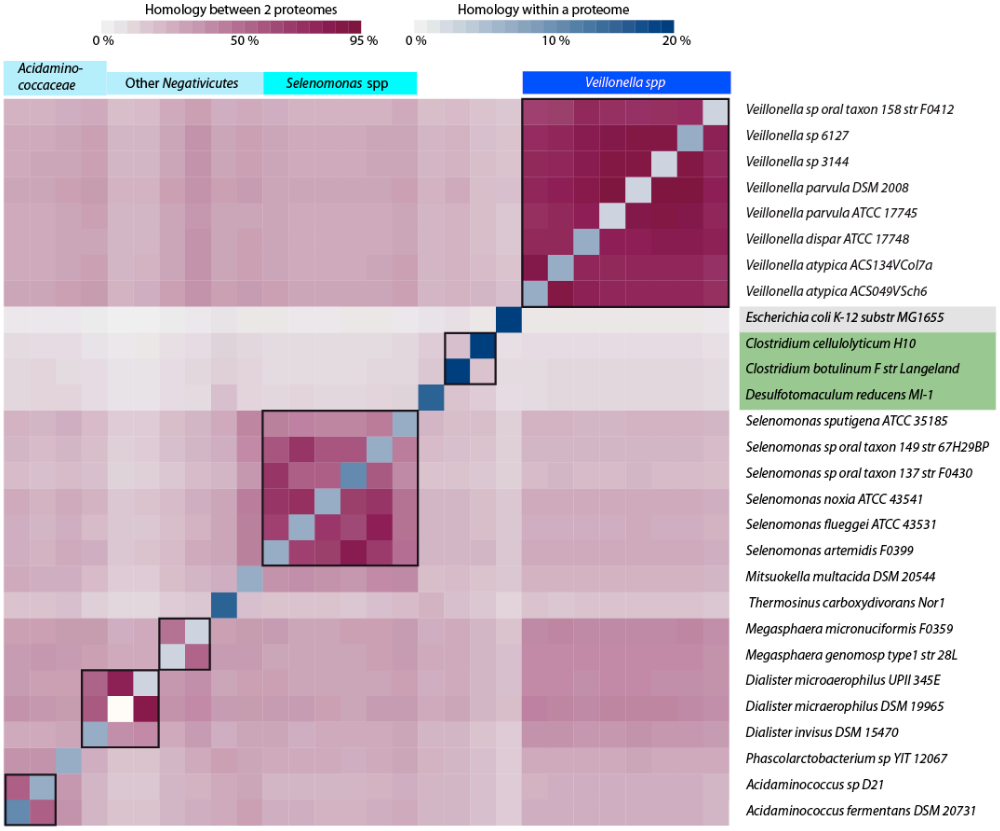


Figure 6. Proteome comparison represented by a BLAST matrix, based on 24 *Negativicutes* genomes with reciprocal best hits. The genomes of *Clostridium botulinum*, *Cl. cellulolyticum*, *Desulfotomaculum reducens* and *E. coli* are added for comparison. Inter-genus comparisons are indicated by black squares. A version reporting the numerical values of homology percentages is available as supplementary information.

Discussion

The availability of complete sequences for a large and diverse set of Bacterial genomes has helped in exploring the conundrum of the genus *Veillonella*, a genus within the *Negativicutes* class, all of which are Gram negative *Firmicutes*. The 16S rRNA tree shown as Figure 1 illustrates how "close" the *Negativicutes* are to other *Firmicutes*. The closest Gram positive *Clostridium* species are actually quite distant to *Veillonella* and other *Negativicutes* genomes, as can be seen in the low fraction of shared protein families in Figure 6. The Gram-negative *Firmicutes* are even more distant to other Gram negatives, such as *Proteobacteria* (e.g., *E. coli*). It should be noted that the family *Clostridiaceae* is a largely diverse group with many members being re-classified [27]. It is therefore possible that the taxonomic description of some *Clostridium* genomes may change in future. However, our analyses did not identify one single Gram-positive *Firmicutes* (*Clostrida* or others) that consistently was identified as most closely related to *Veillonella*. As seen from three types of phylogenetic analysis, the *Negativicutes* class genomes form a distinct cluster within the *Firmicutes*, and the *Veillonella* genus forms a relatively homogeneous group of species within the *Negativicutes*, with relatively conserved metabolic properties (Figure 5). In comparison, the *Selenomonas* genus is more heterogeneous, at least based on their total gene comparison, as illustrated in Figure 6.

In contrast to expectations, relatively little homology between *Negativicutes* and other Gram-negative genomes was detected in our analyses. Neither gene-dependent phylogenetic analysis, nor gene-independent DNA tetramer analysis identified a significant commonness between *Negativicutes* and, say, *Proteobacteria*. Only whole-genome frequency analysis of amino acid usage identified some similarity to a few *Proteobacteria*, and this might be more reflective of environment the organism is adapted to, and not phylogeny. Using KEGG pathways for metabolic comparison of the proteomes we found few pathways in common, with the exception of a shared lipopolysaccharide biosynthesis pathway. From all analyses combined, it is clear that the taxonomic placement of *Negativicutes* within the *Firmicutes* reflects their genetic and genomic characteristics, although the proteins encoded by the *Negativicutes* genomes are quite distinct from their Gram-positive cousins. It could be speculated that the double membrane of the *Negativicutes* evolved in a lineage that used to be a single-membrane (Gram-positive) Firmicute. Whether this event co-evolved independently of the formation of other Gram-negative phyla, or was the result of lateral gene transfer, cannot be stated for certain at present; estimations of horizontally transferred regions in *Veillonella parvula* DSM 2008, the only fully assembled *Veillonella* genome available, using the least conservative method on the Islandviewer web-site [31], revealed that only 2% of the genome is of foreign origin. In comparison, 9% of the *E. coli* K-12 subsp. MG1655 genome was predicted as horizontally transferred. Further analyses are therefore needed to assess this in more detail.

Author's contributions

Tammi Vesth was a main contributor to the writing of the manuscript and to the organization of the work. Trudy Wassenaar helped considerably in editing and improving the manuscript. Individual contributions: Asli Ozen (16s rRNA and CV tree), Oksana Lukjancenko (consensus tree), Sandra Andersen (initial investigations and background research, early version of the manuscript), Rolf Sommer Kaas (BLAST matrix), Jon Bohlin (tetramer and amino acid usage heatmaps), Intawat Nookaew (metabolism heatmaps). David Ussery provided the original idea for this manuscript, suggested the figures, helped in early drafts of the manuscript, and supervised the project.

Acknowledgements

This research was supported by grants from the Danish Research Council, and in part by a grant 09-067103/DSF from the Danish Council for Strategic Research.

References

1. Delwiche EA, Pestka JJ, Tortorello ML. The *veillonellae*: gram-negative cocci with a unique physiology. *Annu Rev Microbiol* 1985; **39**:175-193. PubMed http://dx.doi.org/10.1146/annurev.mi.39.100185.001135
2. Diaz PI, Chalmers NI, Rickard AH, Kong C, Milburn CL, Palmer RJ, Kolenbrander PE. Molecular Characterization of Subject-Specific Oral Microflora during Initial Colonization of Enamel. *Appl Environ Microbiol* 2006; **72**:2837-2848. PubMed http://dx.doi.org/10.1128/AEM.72.4.2837-2848.2006
3. Aas JA, Paster BJ, Stokes LN, Olsen I, Dewhirst FE. Defining the Normal *Bacteria*l Flora of the Oral Cavity. *J Clin Microbiol* 2005; **43**:5721-5732. PubMed http://dx.doi.org/10.1128/JCM.43.11.5721-5732.2005
4. Mager DL, Ximenez-Fyvie LA, Haffajee AD, Socransky SS. Distribution of selected bacterial species on intraoral surfaces. *J Clin Periodontol* 2003; **30**:644-654. PubMed http://dx.doi.org/10.1034/j.1600-051X.2003.00376.x
5. Olson JC, Cuff CF, Lukomski S, Lukomska E, Canizales Y, Wu B, Crout RJ, Thomas JG, McNeil DW, Weyant RJ, *et al.* Use of 16S ribosomal RNA gene analyses to characterize the bacterial signature associated with poor oral health in West Virginia. *BMC Oral Health* 2011; **11**:7. PubMed http://dx.doi.org/10.1186/1472-6831-11-7

6. Chalmers NI, Palmer RJ, Cisar JO, Kolenbrander PE. Characterization of a *Streptococcus sp.*-*Veillonella sp.* Community Micromanipulated from Dental Plaque. *J Bacteriol* 2008; **190**:8145-8154. PubMed http://dx.doi.org/10.1128/JB.00983-08

7. Leuckfeld I, Paster BJ, Kristoffersen AK, Olsen I. Diversity of *Veillonella* spp. from subgingival plaque by polyphasic approach. *APMIS* 2010; **118**:230-242. PubMed http://dx.doi.org/10.1111/j.1600-0463.2009.02584.x

8. Tunney MM, Field TR, Moriarty TF, Patrick S, Doering G, Muhlebach MS, Wolfgang MC, Boucher R, Gilpin DF, McDowell A, Elborn JS. Detection of anaerobic bacteria in high numbers in sputum from patients with cystic fibrosis. *Am J Respir Crit Care Med* 2008; **177**:995-1001. PubMed http://dx.doi.org/10.1164/rccm.200708-1151OC

9. Murri M, Leiva I, Gomez-Zumaquero JM, Tinahones FJ, Cardona F, Soriguer F, Quepo-Ortuño MI. Gut microbiota in children with type I diabetes differs from that in healthy children: a case-control study. *BMC Med* 2013; **11**:46. PubMed http://dx.doi.org/10.1186/1741-7015-11-46

10. Kolenbrander PE, Moore LVH. The genus *Veillonella*. in:H.G. Balows, M. Trüper, W. Dworkin, W. Harder, K.H. Schleifer (Eds.), The prokaryotes (2nd ed.), Springer, New York (1992), pp. 2034–2047.

11. Mashima I, Nakazawa F. Identification of *Veillonella tobetsuensis* in tongue biofilm by using a species-specific primer pair. *Anaerobe* 2013; **22**:77_81.

12. De Vos P, Garrity GM, Jones D, Krieg NR, Ludwig W, Rainey FA, Schleifer KH, Whitman WB. Volume 3: The *Firmicutes*. In Bergey's Manual of Systematic Bacteriology, Springer 2009.

13. Lagesen K, Hallin P, Rødland EA, Staerfeldt HH, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* 2007; **35**:3100-3108. PubMed http://dx.doi.org/10.1093/nar/gkm160

14. Katoh K, Toh H. Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics* 2010; **26**:1899-1900. PubMed http://dx.doi.org/10.1093/bioinformatics/btq224

15. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Mol Biol Evol* 2011; **28**:2731-2739. PubMed http://dx.doi.org/10.1093/molbev/msr121

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

17. Xu Z, Hao B. CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes. *Nucleic Acids Res* 2009; **37**(suppl 2):W174-W178. PubMed http://dx.doi.org/10.1093/nar/gkp278

18. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. Toward automatic reconstruction of a highly resolved tree of life. *Science* 2006; **311**:1283-1287. PubMed http://dx.doi.org/10.1126/science.1123061

19. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004; **32**:1792-1797. PubMed http://dx.doi.org/10.1093/nar/gkh340

20. Fink WL. Microcomputers and phylogenetic analysis. *Science* 1986; **234**:1135-1139. PubMed http://dx.doi.org/10.1126/science.234.4780.1135

21. Retief JD. Phylogenetic analysis using PHYLIP. *Methods Mol Biol* 2000; **132**:243-258. PubMed

22. Karlin S, Burge C. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet* 1995; **11**:283-290. PubMed http://dx.doi.org/10.1016/S0168-9525(00)89076-9

23. Wickham H: ggplot2: Elegant Graphics for Data Analysis (Use R!). Springer New York, 2009. ISBN-10: 0387981403 | ISBN-13: 978-0387981406

24. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. *Nucleic Acids Res* 2004; **32**:277-280. PubMed http://dx.doi.org/10.1093/nar/gkh063

25. Eddy SR. BIOINFORMATICS REVIEW Profile hidden Markov models. [PubMed]. *Bioinformatics* 1998; **14**:755-763. PubMed http://dx.doi.org/10.1093/bioinformatics/14.9.755

26. Hobohm U, Sander C. Enlarged representative set of protein structures. *Protein Sci* 1994; **3**:522-524. PubMed http://dx.doi.org/10.1002/pro.5560030317

27. Ludwig W, Schleifer KH, Whitman W. Revised road map to the phylum *Firmicutes*. In Bergey's Manual of Systematic Bacteriology, Springer 2009:1-13.

28. Gupta RS. Origin of diderm (Gram-negative) bacteria: antibiotic selection pressure rather than endosymbiosis likely led to the evolution of bacterial cells with two membranes. *Antonie van Leeuwenhoek* 2011; **100**:171-182. PubMed http://dx.doi.org/10.1007/s10482-011-9616-8

29. Pride DT, Meinersmann RJ, Wassenaar TM, Blaser MJ. Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res* 2003; **13**:145-158. PubMed http://dx.doi.org/10.1101/gr.335003

30. Dutta C, Paul S. Microbial lifestyle and genome signatures. *Curr Genomics* 2012; **13**:153-162. PubMed http://dx.doi.org/10.2174/138920212799860698

31. Langille MG, Brinkman FS. IslandViewer: an integrated interface for computational identification and visualization of genomic islands. *Bioinformatics* 2009; **25**:664-665. PubMed http://dx.doi.org/10.1093/bioinformatics/btp030

5.3 CMGfunc: Comparative functional annotation of bacterial proteins

The paper described in this section represents the main work of this project, a stand alone functional annotation system designed for comparative analysis of bacterial genomes. The system consists of 1.216 models of different functions contained in a stand alone virtual computer which can be installed on any platform. Each model is the result of a Artificial Neural Network training procedure and each model represents a function trained on protein clusters of between 75 and 35.570 sequences (mean 1.471). Each function is described by Pfam-A domains, Pfam clans and GO terms and can be connected to Interpro. The pipeline consists of three steps and takes a protein FASTA file as input. A total of 75 sequence features are calculated for each protein and each protein is compared to the functional models. The model which gives the sequence the highest value (a 100% match equals a value of 1) is recorded and the function of the models is assigned to the sequence. Frequencies of each function is calculated for each genome or input set and visualized in a set of heat maps. When multiple genomes are analyzed, the analysis performs a clustering procedure of genomes based on shared functional frequencies. The analysis divides functions into the three GO ontologies, molecular function, biological process and cellular component, allowing for evaluations of similarity based on different levels of functional annotation.

The performance of CMGfunc was evaluated using the CAFA 1 (Critical Assessment of Functional Annotation) data. Each protein was classified using the ANN models, as described above, 98% of the sequences were assigned a function. Comparing the GO terms output from CMGfunc with the GO terms from Uniprot entries for the dataset revealed a 64% correct prediction rate on GO level 3 and 55% on GO level 4. The coverage of the CMGfunc models was further investigated using proteins that did not have matches to Pfam-A. A set of 47.050 proteins which could not be annotated using Pfam-A was compared to each of the CMGfunc models and functions for 98% of the dataset were predicted. Among these functions, the helix-turn-helix(HTH) Pfam clan (1.139 proteins) was the most common. Although not a direct function, such a pattern does suggest evidence for an HTH structure being present in these proteins.

This project presented a set of models and tools for comparative functional annotation and is available as a stand alone virtual computer as well as individual scripts from GitHub, where wiki documentation is also found. The offline use and cross platform installation ensures that the pipeline can be used for confidential data analysis and does not require expensive computational infrastructure.

CMGfunc: Comparative functional annotation of bacterial proteins using artificial neural networks and proteins domains

Tammi Vesth[1], David W. Ussery[1,2], and Karin Lagesen[3,4]

[1]**Center for Biological Sequence Analysis, Department of Systems Biology, The Technical University of Denmark, Building 208, DK-2800 Kgs. Lyngby, Denmark. E-mail: tammi@cbs.dtu.dk**
[2]**Comparative Genomics Group, Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA. Email: dave@cbs.dtu.dk**
[3]**Norwegian Sequencing Centre, Department of Medical Genetics, Oslo University Hospital, 0407 Oslo, Norway**
[4]**Biomedical Informatics, Department of Computer Science, University of Oslo, Oslo, Norway. E-mail: karin.lagesen@medisin.uio.no**

Abstract **Background:** Assignment of protein function, particularly in the absence of strong matches to proteins of known and well characterized function, is a difficult task. Many programs and pipelines have been developed as an effort to improve the automated functional annotation of protein sequences. These vary greatly in ease of use, level of detail in annotation, manual curation and modelling approaches. One shortcoming of such systems is that they do not readily allow for the comparison of annotations across genomes. Here we present CMGfunc (Comparative Microbial Genome functions), a bioinformatics pipeline and model collection for comparative functional annotation of bacterial genomes. The models are based on protein clusters created on the basis of shared Pfam-A domain and modeled using artificial neural networks, using 75 sequence features as input values.

Results: The performance of the CMGfunc method was assessed using the dataset from a previous contest for prediction of protein function, CAFA 1. CMGfunc consists of 1216 functional models based on artificial neural networks, were shown to cover 98% of 10,019 protein sequences used in the CAFA 1 challenge and correctly annotate 60% proteins at the GO term level 3. Furthermore, the methods, although based on protein clustering using Pfam, assigned functions to 98% of proteins with no match to Pfam (example set of 46.389 sequences).

Background

Advances in DNA sequencing technologies over the past 10-20 years, both in speed, precision and price, has increased the amount of sequence data tremendously. This development has moved bioinformatics from the periphery of life sciences, to a more central role, as analysis of massive amounts of sequence data has become crucial. An important area of bioinformatics is the functional annotation of genes and proteins based on sequence data. Although protein functions might at first appear to be a well defined problem, the definition of a function varies based on the context in which it is used[1]. As such the ambiguity of the protein function concept can be described as a matter of defining the word ”understand”[2]. As the aim of sequencing DNA is to get greater understanding of biology, the process of assigning a function to a gene or protein becomes a question of what is meant by the word ”function”. The number of genome projects has not only increased the available data to resolve the questions about DNA and the functions of genes but has also greatly increased the amount of software to aid in finding the answers.

Standardizing functional annotation

Though functional annotation has become increasingly more common and new systems are emerging rapidly, evaluating the performance of these systems has been standardized. Due to differences in functional standards and the standardized test sets, it has been difficult to compare one system to another. The Gene Ontology (GO), from the gene Ontology Consortium, has emerged as a possible solution to some of the problems with evaluation of functional annotation. The consortium was originally launched as a collaborative project between three eukaryotic model organism databases, but has since expanded to include many microbial data sources as well[3, 4]. The ontology consists of three structured controlled vocabularies (ontologies) of functional descriptions (GO terms) and unique identifiers constructed through manual annotation and combines data from several databases and scientific literature. The different vocabularies cover three aspects of gene product function: molecular function, biological process and cellular component. The descriptions in GO are organized in a relational manner with ”child-parent” relationships between different terms. The GO has proved useful in making annotation comparable and standardized and is now used by many annotation pipelines [5, 6, 7].

With an established system like GO other initiatives have been set up to promote advances in functional annotation. The Critical Assessment of Functional Annotation experiment (CAFA) aims at improving the performance and evaluation of functional annotation of proteins. The project constructs a functionally unknown dataset from public data (Swiss-Prot and the Enzyme Function Initiative[8]) and research groups sign up to attempt to assign functions to the data. After a year, predicted functions are compared to accumulated experimental functions and performance is evaluated. The project ran for the first time in 2006 and has collected many useful approaches to functional annotation. Furthermore, the experiment has highlighted the slow progress in experimental verification, further supporting the need for computational methods.

Sequences domains

Proteins domains have long been of interest in the field of functional annotation, because many of the functions in the cell are done by proteins. The ”sequence hypothesis”, upon which molecular biology is built, assumes that the amino acid sequence of a protein determines its structure, and the structure determines its function(s). Based on the observations that many known structures contain conserved domains which form specific functions, a modular approach seems reasonable - where a protein can be divided into sets of functional domains. One such approach to modelling protein functional domains is the Pfam database, first published in 1997 [9]. A functional domain might serve as a specific binding site, or create a specific secondary or tertiary structure essential for the proteins function. These characteristic regions have been used to build models of functional or structural domains of proteins. Different approaches in defining protein domains are being used with Pfam, SUPERFAM, TIGRFAM and PANTHER being among the best known databases [10, 11, 12, 13]. The approach used here is based on the Pfam database of Hidden Markov Models for domain identification. One part of Pfam is a manually curated set of domain models (Pfam-A) while another part is automatically generated from common sequence patterns (Pfam-B). Additionally, Pfam includes a structure called ”clans” consisting of manually curated domains sets with related structure or function or with similarities between sequence profiles. Not all domains belong to a clan. Through the InterPro database[14] it is possible to connect some (but not all) Pfam domains to GO terms.

Comparative Microbial Genomics functions, CMGfunc

The method presented here, CMGfunc (Comparative Microbial Genome functions), uses Pfam-A and Pfam clans to create functionally related protein clusters. The aim of the method is to create functional models with a degree of generalization that allows comparison of large sets of genomes. CMGfunc does not give a detailed prediction of each protein function. If the function of a protein is very specific, the chance of finding the same exact function in another genome is very low. However, when comparing genomes it is usually more interesting to discover which processes are found in both genomes. For proteins with no strong match to a reference with known structure and function, prediction of general properties is better than nothing.

Proteins from NCBI GenBank are clustered using Pfam-A domains and clans. Additionally, 75 sequence features are calculated for each protein. These features describe biochemical, structural or functional signatures of the individual protein and include amino acid counts, molecular weight, estimates of subcellular location and signal peptides. Each cluster is modeled using artificial neural networks, where the input is the features of a protein sequence encoded numerically and the response is a score describing how well a protein matches to that specific function. Each network is a model describing a specific protein function. A comparative pipeline allows for comparison of new proteins to all the models and returns both text and graphical output of the genome annotation. The pipeline also allows for comparison of functional annotation across a number of genomes. Figure 1 shows the analysis flow of the pipeline.

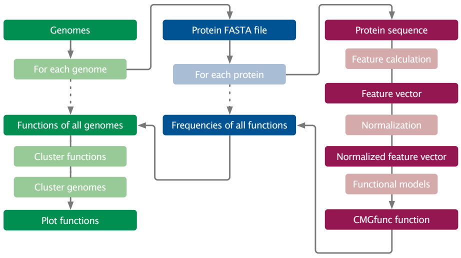


Figure 1. Flow of analysis in CMGfunc.

Materials and Methods

Genome and proteome data

A set of 1632 bacterial genomes was obtained from the NCBI GenBank FTP database as of November 2010 (*ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/*, see Supplementary Table S1 for full list of genomes). Complete DNA sequences were extracted from the GenBank files and predicted proteomes were constructed using the gene-finding algorithm implemented in Prodigal version 2.0 (March 2010)[15]. Prodigal was run with the option of not constructing genes over DNA sequence containing unknown bases (option $-m$). A total of 5.317.141 proteins were predicted from the 1.623 genomes, and stored in a MySQL database designed for the purpose.

Sequence clustering

Protein clusters with functionally related sequences were created based on shared Pfam domains. The clustering was done using three consecutive criteria:

- Criteria 1: Does the protein contain a Pfam-A domain?
- Criteria 2: Does the architecture match a Pfam-A clan?
- Criteria 3: Is the protein group too large? Are too few proteins in the cluster?

Criteria 1: Does the protein match a Pfam-A domain? Each protein was compared to the Pfam-A (version 26.0 November 2011, 13.672 families) database[16] using pfam_scan.pl[17]. Matches between sequences and Pfam-A models were recorded and the presence/absence of domains was used to create an architecture for each protein. Multiple matches of one domain in one sequence were ignored as well as the relative position of the domain in the sequence (e.g., AB = BAA). A protein sequence was allowed to match multiple unique domains as long as they did not overlap - if overlaps were detected, the highest scoring domain was used. The combination of domains in a sequence is called the protein "architecture" and architectures can consist of just a single domain. Each of the 5,3 million proteins was compared to the Pfam-A database. About one million (1.130.097) proteins did not match Pfam-A while more than 4 million (4.177.021 or 78%) matched at least one domain. The 4,1 million proteins containing Pfam-A domains were then clustered based on shared Pfam domain architectures, yielding 26.179 architectures clusters.

Criteria 2: Does the architecture match a Pfam-A clan? Architecture clusters were connected to Pfam clans by single linkage, if a domain in the architecture could be connected to the clan the architecture was connected to that clan. When an architecture could be connected to several clans, the architecture was assigned to both clans. Of the 26.179 architectures, 76% (19.988) could be assigned to a clan and are referred to as clan clusters. Architectures consisting of one domain were not included, as these did possibly not hold more information than that Pfam domain model itself.

Criteria 3: Is the protein group too large or small? Some architecture clusters were very small (16.041 groups had less than 10 proteins, 13.749 have less than 5 and 21.506 have less than 100) and 13 clan clusters contained more than 50.000 proteins. If a clan group contained more than 50.000 sequences, it was split into architecture groups while retaining its clan description (architecture clans). After this step, any cluster with more than 50.000 or less than 100 sequences was discarded. The result of this clustering is three types of functional clusters: architectures with no clan association (2.570), architectures with a clan association (914) and clans (381).

Gene Ontology terms

Each of the clan models were connected to the Gene Ontology terms using the Pfam to GO mapping provided by Gene Ontology[3, 14]. The mapping contained a list of Pfam domains and the GO terms they belong to, thus there could be more GO terms per Pfam as well as more Pfams with the same GO term. The terms included both Molecular Function (MF), Biological Process (BP) and Cellular Component (CC). GO terms were associated with Pfam-A domains through a Pfam to GO mapping constructed by the InterPro [14] project and was obtained March 2012 from *http://www.geneontology.org/external2go/pfam2go*. This

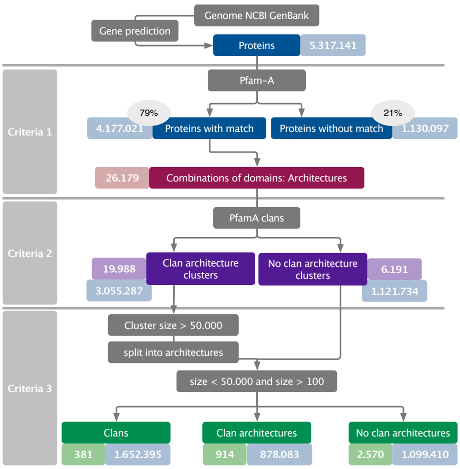


Figure 2. Protein clustering and functional cluster selection.

mapping does not cover all Pfam domains and 4.628 different domains were mapped to 2.176 GO terms. Terms were connected to architectures and clans via single linkage - if one domain in an architecture was connected to the term, that term was assigned to the entire architecture or clan. All clans were connected to at least one GO term and 67% (2.317 of 3.484) of the architectures could be connected to a term.

Sequences features

A set of 75 sequence features were calculated for each protein, based on a number of external programs and pipelines. Using the python module ProtParam (*http://biopython.org/w/index.php?title=ProtParam*) six values, or features, were calculated based on the ExPaZy Protparam pipeline. Another 11 features came from Psort [18] running the program with Gram positive and negative settings for each protein. High and low complexity regions were calculated using SEG[19] and the output was six values for each protein. SignalP was used, again using both Gram positive and negative settings for each protein, and produced 12 features. Normalization is the process of adjusting all parameters to the same scale and here each feature was normalized to a scale of 0 to 1. If the measure had a fixed scale already, like PSORT giving values between 0 and 10, this was used to normalize. If no fixed scale existed, like the molecular weight, the highest value for that feature, observed in the dataset was used to normalize (See Supplementary Table S2). The 75 features are combined into a single vector which is used as the input to the functional group modeling engine for the artificial neural network.

Artificial Neural Networks

Artificial neural networks were used to model the sequence groups formed by architectures and clans. One network model was created per group, so that the model could be used to evaluate whether a protein belongs to that group or not. Training is done by presenting sequence feature vectors to the network engine, which then uses that specific network model to calculate a value based on the input. The model is then gradually adjusted so that it will approximately give the desired output for the specified input. In this case, the desired output was set to be 1 for membership and 0 for non-membership.

Training was performed using 75% of positive data and tested using the remaining 25%. Negative data used for training and testing was selected from each group excluding the group currently being trained on. The number of examples (proteins) selected from each group was set to 30 unless the total number of positive examples divided by the number of negative clans was larger than 30. For 381 clans, this means that at least 11,430 negative examples were used. Preliminary data indicated that the networks showed a tendency to predict many false positives if they were not trained with large amounts of negative examples. The same number of sequences were taken from each group, thus ensuring that negative data was selected from all groups not in the positive set. Three sets were generated for each functional group (protein cluster), randomly creating the positive and negative sets to address problems with bias in data. A fully connected feed-forward neural network architecture was constructed, with two hidden layers and 30 neurons in each hidden layer. Each layer also contained a bias node. A sigmoid activation function was used on all connections. The networks were constructed using the Python library, pybrain[20] for architecture and algorithm implementations.

As the network trains, the performance is monitored using the mean squared error (MSE) of the desired and predicted output. The more frequently the network predicts a value close to the desired output the lower the MSE becomes. As the MSE is a mean of all predicted/target difference, it is sensitive to large numbers of examples. If the number of negative examples is very high compared to the number of positive examples, the network can get a low MSE even by predicting all outputs to 0, the negative data simple overpowers the effect of the false negatives. For this reason, the networks are not evaluated solely on MSE values. The MSE is calculated at each training round (iteration) and can therefore be used to stop the training when the network seems to perform good enough. However, for the reason described above, a low MSE can be misleading. Therefore, the desired MSE was set very low for training (0.0001 for datasets with less than 1.000 positive examples and 0.001 for datasets with more positive examples). The training stops when the desired MSE value is met or when 1.000 iterations have been run, whichever comes first. Networks were evaluated using Matthew's Correlation Coefficients (MCC) and MSE. The MCC was selected for its ability to measure the performance of a classifying program. The three different models created for each cluster were compared based on the MCC value for the testing and the best performing version was selected. Each of the functional cluster groups were tested for MCC values above 0.85 and of the initial set, good performance was found for 756 of 2.570 architectures with no clan association, 296 of 914 architectures with a clans association and 164 of 381 clans.

CMGfunc pipline

The models constructed above were then built into a pipeline for use on unknown FASTA sequences. The system is called CMGfunc and is built into a virtual computer available for download and local installation. A virtual computer can be run on any platform using virtualization software and guidelines for installation are available on GitHub under the repository name ”cmgfunc”. The pipeline consists of three parts:

1) CMGfunc.pl - this part takes a protein FASTA file as an input and a directory of functional models (neural networks). Sequence features are calculated and normalized for each protein and each protein is compared to a set of functional models. The output is a result file containing the best scoring function for each sequence ($*res$) and another list ($*res.all$). The score indicates how well a protein match the model for a specific function. A perfect match would be 1 while a reasonable match would score around 0.8.

2) CMGfunc_analyzeGenome.pl - the purpose of this step is to summarize the functions identified in the first step. The frequency of each function is recorded and combined with functional descriptions for Pfam-A clans and domains as well as GO terms. The frequency of the clans is recorded and associated GO terms are listed. No clan architectures are described by domain descriptions and possible GO description while clans and architecture clans are listed with clan descriptions and GO terms. The output is a table in raw text format.

3) CMGfunc_plot_analyzeGenome.pl - the table from above is used as input and several plots and tables are generated. Three plots are created for function frequencies, one for molecular function, one for biological process and one for cellular component. Percentages plots are created for the same three GO ontologies. A table containing data used for the plots is created and also contains additional functional information about each functional model.

0.1 Comparison to CAFA 1

The CAFA 1[21] dataset was downloaded from the webpage: http://biofunctionprediction.org/content/previous-cafa-data. The data itself was not part of the data used to train the CMGfunc models. The set consists of 11,532 FASTA sequences with sequence names and UniProt identifiers. Accessing a local version of UniProt (Release October 2013) 11,039 sequences were connected to a UniProt entry. Furthermore, not all entries contained GO terms and 92% (10,111 of 11,039) entries were found to be connected with one or more GO terms. Each of the 11,532 protein were compared to the CMGfunc functional models and 98% (11,293) were found to match a function with a score higher than 0.9. To compare the results with the UniProt functions, only matches with associated GO terms can be used. Some architectures do not have GO terms. The set of predicted functions with score above 0.9 and associated GO terms include 9814 sequences. Three different comparisons were done between the CMGfunc predictions and the Uniprot annotations. The first was an exact term similarity comparison, testing how many of the exact same GO terms were found in the CMGfunc predictions and the UniProt data. For the two next comparisons, BLAST2GO was used to identify level 3 and 4 GO terms for Each Uniprot and CMGfunc GO term set[22]. The upper level terms were identified using the ”Combined Graph” analysis, exporting the graph text data and extracting terms for the desired level. The Overlap in GO terms was then calculated again.

Results and Discussion

A set of 1632 genomes were obtained from NCBI GenBank as of November 2012. From these, more than 5 million proteins were identified using the genefinder, Prodigal. Each protein was compared to the Pfam-A database and 78% were found to match one or more domains. Proteins were clustered based on shared presence/absence of domains as well as connections to Pfam clans. Clusters with more than 50,000 and less than 100 proteins were discarded. The remaining set of 3865 protein clusters were modeled using 75 sequence features and a feed-forward artificial neural network setup. The models showed varying performance (See Supplementary Table S3) and 1216 were selected as acceptably performing networks based on Matthew's Correlation Coefficient (above 0.85). These functional models are described by GO terms as well as Pfam domain or clan descriptions and can be connected to InterPro entries if desired. The models are used as the backbone in the CMGfunc pipeline. A set of input proteins are compared to each model and the best comparison is recorded. If the best score is close to 1, the protein will likely have the same function as the model it resembles, and is connected with the GO terms and descriptions of that model. The output of the pipeline is a TAB delimited table file and six heat-map plots. The heat-maps represent the frequency and fraction of each function in each genome. The functions are further split into GO ontologies, with separate plots of Molecular Function, Biological Process and Cellular Components. The plots are limited by user defined thresholds, showing only fractions or frequencies above the threshold. Figure 3 shows the comparison of functions for 15 *Escherichia coli* genomes using a frequency threshold of 10 molecular function (See Supplementary Figure S1 for biological process and cellular component and percentage). Clustering is automatically performed on both function and genomes (For single genome annotation, see Supplementary Figure S2). The genome set includes three genome sequences of *E. coli* BL21 DE3 (NCBI project ID

20713, 28965 and 30681) and another 12 genomes. One genome *E. coli* SMS 3 5 (NCBI project ID 19469) was included twice, the exact same sequence, to verify the consistency of the annotation. The genomes form distinct clusters on all three plots with genomes but the clusters are not the same when comparing molecular functions, biological processes and cellular components.

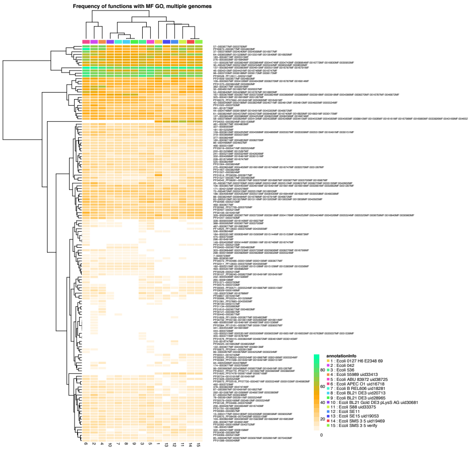


Figure 3. CMGfunc results for set of genomes, molecular function GO terms, frequencies over 10.

The method was tried on the CAFA 1[21] dataset comparing the GO terms predicted by CMGfunc and recorded UniProt annotations for each protein. UniProt entries were obtained from the CAFA dataset and 10,019 proteins were found to be connected to UniProt as well as GO terms. CMGfunc predicted functions for 98% (9814) of the proteins with a score of 0.9 or higher (a perfect match would be 1). Comparing the GO terms predicted by CMGfunc with the UniProt terms, using exact term matching, the predicted terms overlapped with the real terms 29% of the time (2,902 of 10,019). Most of the functional agreement was on molecular function level, with 1,750 entries. Using BLAST2GO, the GO terms of both CMGfunc and CAFA UniProt were normalized to the third and fourth GO graph level. Calculating the overlap again, with exact term matching, CMGfunc predicted the same level 3 GO terms in 64% and 55% on level 4.

The CMGfunc functional models are based on more than 4 million proteins with matches to Pfam-A; however, 22% of the proteins acquired could not be matched to Pfam-A and could as such not be annotated using the information already in the database. A random selection of 47,050 proteins from this unmatched set was compared to the CMGfunc models and 98.6% (46,389) were found to match on of the models with a score above 0.9. The most common models matched to these proteins is the helix-turn-helix clan. This structure accounts for 1,139 of the proteins. Although not a direct function, this pattern does suggest that these proteins do have the structure despite not matching the Pfam domains associated with it. The second most common function is the ribonuclease H-like clan (908 proteins) which includes ”Any process that initiates the activity of the inactive enzyme MAP kinase kinase kinase in the context of cell wall biogenesis, the assembly and arrangement of the cell wall, the rigid or semi-rigid envelope lying outside the cell membrane of plant, fungal and most prokaryotic cells”. Other common functions include clans for lysozyme-like proteins, MetJ/Arc repressors and periplasmic binding proteins (See Supplementary Figure S3 for CMGfunc heatmaps of the protein annotation).

Conclusion

The method presented here, CMGfunc, consists of 1,216 functional models based on artificial neural networks. The models were shown to covered 98% of 10,019 protein sequences used in the CAFA 1 challenge and correctly annotate 64% using on GO term level 3. Furthermore, the method, although based on protein clustering using Pfam, assigned functions to 98% of proteins with no match to Pfam (example set of 46,389 sequences).

Author contributions

DWU, idea and initial project description and editing of the manuscript. KL, method development and writing. TV, method development, coding, pipeline design and writing.

Competing interests

No competing interests were disclosed.

Acknowledgements

We would like to acknowledge Steven J. Otto for his input on artificial neural network implementations.

References

[1] Iddo Friedberg. Automated protein function prediction–the genomic challenge. *Briefings in bioinformatics*, 7(3):225–42, September 2006.

[2] Michael Y. Galperin and Eugene V. Koonin. From complete genome sequence to ‘complete’ understanding? *Trends in Biotechnology*, 28(8):398–406, August 2010.

[3] Michael Ashburner, CA A Ball, JA A Blake, David Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, G Sherlock, The Gene, and Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, May 2000.

[4] M A Harris, J Clark, A Ireland, J Lomax, M Ashburner, R Foulger, K Eilbeck, S Lewis, B Marshall, C Mungall, J Richter, G M Rubin, J A Blake, C Bult, M Dolan,

H Drabkin, J T Eppig, D P Hill, L Ni, M Ringwald, R Balakrishnan, J M Cherry, K R Christie, M C Costanzo, S S Dwight, S Engel, D G Fisk, J E Hirschman, E L Hong, R S Nash, A Sethuraman, C L Theesfeld, D Botstein, K Dolinski, B Feierbach, T Berardini, S Mundodi, S Y Rhee, R Apweiler, D Barrell, E Camon, E Dimmer, V Lee, R Chisholm, P Gaudet, W Kibbe, R Kishore, E M Schwarz, P Sternberg, M Gwinn, L Hannick, J Wortman, M Berriman, V Wood, N De La Cruz, P Tonellato, P Jaiswal, T Seigfried, and R White. The Gene Ontology (GO) database and informatics resource. *Nucleic acids research*, 32(Database issue):D258–D261, 2004.

[5] Kevin Galens, Joshua Orvis, Sean Daugherty, Heather H. Creasy, Sam Angiuoli, Owen White, Jennifer Wortman, Anup Mahurkar, and Michelle Gwinn Giglio. The IGS Standard Operating Procedure for Automated Prokaryotic Annotation. *Standards in Genomic Sciences*, 4(2):244–251, April 2011.

[6] Konstantinos Mavromatis, Natalia N. Ivanova, I-Min a. Chen, Ernest Szeto, Victor M. Markowitz, and Nikos C. Kyrpides. The DOE-JGI Standard Operating Procedure for the Annotations of Microbial Genomes. *Standards in Genomic Sciences*, 1(1):63–67, July 2009.

[7] David M. Tanenbaum, Johannes Goll, Sean Murphy, Prateek Kumar, Nikhat Zafar, Mathangi Thiagarajan, Ramana Madupu, Tanja Davidsen, Leonid Kagan, Saul Kravitz, Douglas B. Rusch, and Shibu Yooseph. The JCVI standard operating procedure for annotating prokaryotic metagenomic shotgun sequencing data. *Standards in Genomic Sciences*, 2(2):229–237, March 2010.

[8] JA Gerlt, KN Allen, and SC Almo. The Enzyme Function Initiative. *Biochemistry*, 50(46):9950–9962, 2011.

[9] E L Sonnhammer, S R Eddy, and R Durbin. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, 28(3):405–20, July 1997.

[10] Robert D Finn, Jaina Mistry, John Tate, Penny Coggill, Andreas Heger, Joanne E Pollington, O Luke Gavin, Prasad Gunasekaran, Goran Ceric, Kristoffer Forslund, Liisa Holm, Erik L L Sonnhammer, Sean R Eddy, and Alex Bateman. The Pfam protein families database. *Nucleic acids research*, 38(Database issue):D211–22, January 2010.

[11] J Gough, K Karplus, R Hughey, and C Chothia. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *Journal of molecular biology*, 313(4):903–19, November 2001.

[12] Jeremy D Selengut, Daniel H Haft, Tanja Davidsen, Anurhada Ganapathy, Michelle Gwinn-Giglio, William C Nelson, Alexander R Richter, and Owen White. TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Research*, 35(Database issue):D260–D264, 2007.

[13] Huaiyu Mi, Anushya Muruganujan, and Paul D Thomas. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic acids research*, 41(Database issue):D377–86, January 2013.

[14] Sarah Hunter, Rolf Apweiler, Teresa K Attwood, Amos Bairoch, Alex Bateman, David Binns, Peer Bork, Ujjwal Das, Louise Daugherty, Lauranne Duquenne, Robert D Finn, Julian Gough, Daniel Haft, Nicolas Hulo, Daniel Kahn, Elizabeth Kelly, Aurélie Laugraud, Ivica Letunic, David Lonsdale, Rodrigo Lopez, Martin Madera, John Maslen, Craig McAnulla, Jennifer McDowall, Jaina Mistry, Alex Mitchell, Nicola Mulder, Darren Natale, Christine Orengo, Antony F Quinn, Jeremy D Selengut, Christian J a Sigrist, Manjula Thimma, Paul D Thomas, Franck Valentin, Derek Wilson, Cathy H Wu, and Corin Yeats. InterPro: the integrative protein signature database. *Nucleic acids research*, 37(Database issue):D211–5, January 2009.

[15] Doug Hyatt, Gwo-Liang Chen, Philip F Locascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, 11:119, January 2010.

[16] Marco Punta, Penny C Coggill, Ruth Y Eberhardt, Jaina Mistry, John Tate, Chris Boursnell, Ningze Pang, Kristoffer Forslund, Goran Ceric, Jody Clements, Andreas Heger, Liisa Holm, Erik L L Sonnhammer, Sean R Eddy, Alex Bateman, and Robert D Finn. The Pfam protein families database. *Nucleic acids research*, 40(November 2011):290–301, November 2011.

[17] Jaina Mistry, Alex Bateman, and Robert D Finn. Predicting active site residue annotations in the Pfam database. *BMC bioinformatics*, 8:298, January 2007.

[18] Nancy Y Yu, James R Wagner, Matthew R Laird, Gabor Melli, Sébastien Rey, Raymond Lo, Phuong Dao, S Cenk Sahinalp, Martin Ester, Leonard J Foster, and Fiona S L Brinkman. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics (Oxford, England)*, 26(13):1608–15, July 2010.

[19] JC Wootton and S Federhen. Statistics of local complexity in amino acid sequence databases. *Computers & chemistry*, 17(2):149–163, 1993.

[20] Tom Schaul, J Bayer, D Wierstra, and Y Sun. PyBrain. *Journal ofMachine Learning Research*, 11:743–746, 2010.

[21] Predrag Radivojac, Wyatt T Clark, Tal Ronnen Oron, Alexandra M Schnoes, Tobias Wittkop, Artem Sokolov, Kiley Graim, Christopher Funk, Karin Verspoor, Asa Ben-Hur, Gaurav Pandey, Jeffrey M Yunes, Ameet S Talwalkar, Susanna Repo, Michael L Souza, Damiano Piovesan, Rita Casadio, Zheng Wang, Jianlin Cheng, Hai Fang, Julian Gough, Patrik Koskinen, Petri Törönen, Jussi Nokso-Koivisto, Liisa Holm, Domenico Cozzetto, Daniel W a Buchan, Kevin Bryson, David T Jones, Bhakti Limaye, Harshal Inamdar, Avik Datta, Sunitha K Manjari, Rajendra Joshi, Meghana Chitale, Daisuke Kihara, Andreas M Lisewski, Serkan Erdin, Eric Venner, Olivier Lichtarge, Robert Rentzsch, Haixuan Yang, Alfonso E Romero, Prajwal Bhat, Alberto Paccanaro, Tobias Hamp, Rebecca Kaß ner, Stefan Seemayer, Esmeralda Vicedo, Christian Schaefer, Dominik Achten, Florian Auer, Ariane Boehm, Tatjana Braun, Maximilian Hecht, Mark Heron, Peter Hönigschmid, Thomas a Hopf, Stefanie Kaufmann, Michael Kiening, Denis Krompass, Cedric Landerer, Yannick Mahlich, Manfred Roos, Jari Björne, Tapio Salakoski, Andrew Wong, Hagit Shatkay, Fanny Gatzmann, Ingolf Sommer, Mark N Wass, Michael J E Sternberg, Nives Škunca, Fran Supek, Matko Bošnjak, Panče Panov, Sašo Džeroski, Tomislav Šmuc, Yiannis a I Kourmpetis, Aalt D J van Dijk, Cajo J F ter Braak, Yuanpeng Zhou, Qingtian

Gong, Xinran Dong, Weidong Tian, Marco Falda, Paolo Fontana, Enrico Lavezzo, Barbara Di Camillo, Stefano Toppo, Liang Lan, Nemanja Djuric, Yuhong Guo, Slobodan Vucetic, Amos Bairoch, Michal Linial, Patricia C Babbitt, Steven E Brenner, Christine Orengo, Burkhard Rost, Sean D Mooney, and Iddo Friedberg. A large-scale evaluation of computational protein function prediction. *Nature methods*, 10(3):221–7, March 2013.

[22] Ana Conesa and Stefan Götz. Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics. *International journal of plant genomics*, 2008:619832, January 2008.

6 Concluding remarks

This thesis has presented the challenges and current status of large scale and functional annotation. The primary goal of this project is to develop methods for analyzing sequences without homology to known and annotated sequences. The aim has been to overcome the obstacles in the coverage of functional annotation as well as creating an environment for comparison of these annotations. One of the requirements of the method is high coverage that includes many of the proteins that cannot be annotated using sequence similarity methods. The method should ensure comparability by allowing for functional profiles to be compared across genomes. Lastly, access to data for verification, reproducibility and usage in local research are desired.

This work has included studies of existing methods and databases which have highlighted a number of strengths, initiatives, as well as problems of these methods. Several of the described methods involve the curation of database information and sequences to create new databases and models, such as InterPro. In the process, these projects are accumulating information about individual sequence models, making them increasingly difficult to compare and automatically process. Although existing resources serve important functions, such as systematically storing and connecting experimental and published data, they do not work well for comparative annotation. For such purposes systems like the Gene Ontologies (GO) are better suited but still cause other problems. The gene ontologies reflect the manual curation of the system and the graph structure reflects experimental results, with one function having multiple upper level functions and several functions per protein. When assigning a GO term to a new protein or protein family, there are no requirements as to which level should be assigned, and since GO is a graph system, finding upper level terms for such a group is not straightforward. Furthermore, GO was developed for eukaryotes, and although effort are being made to cover bacterial proteins as well, it is likely that prokaryotic specific initiatives might be required as more metagenomes are sequenced and as such more of the vast bacterial diversity is covered. Standards in the assignment of GO terms, as well as standard approaches for the comparison of terms should be established. As computational biology becomes a bigger part of medical science and society in general, standards must be created to ensure that projects are conducted properly and can be validated against other projects. In the field of functional annotation, the Critical Assessment of Functional Annotation (CAFA) is a step in the right direction and in time it might take on a role like the one seen for CASP.

The functional models presented in this project were designed for comparative analysis, and the 1.216 different functions are not proposed to describe the exact function of every protein in bacteria. Instead they offer a level of annotation which can be compared across very different genome, as the coverage of these models has been shown to be very high (98% of proteins without match to Pfam-A). Although the coverage of the functional models is high, the speed and precision could be improved. The process of calculating features for each new protein is the most limiting factor, although the process currently takes less than 10 seconds per sequence. Improvements could include reprogramming of feature programs or better parallel processing. The predictive performance of the models might be improved by including more features, including more complex structural models, codon usage or tetramer counts. The options of features is almost unlimited but as more features are included, the speed of comparing each new sequence goes down, so this is a cost/benefit problem. The CMGfunc models have been shown to assign functions to proteins without matches to Pfam-A and showed a large coverage of the sequences selected for CAFA 1. The models have a low precision but since the coverage is so high, these models might be used in combination with other more specific annotation tools. Combining Pfam-A or InterPro annotations with CMGfunc might be the way to assign functions to the large number of functionally unknown proteins. Finally, the method includes a setup for creating network models using the CMGfunc features and network architecture for modeling a local set of proteins. This procedure makes it possible for biologists working with a specific protein to create models and descriptions for this specific function based on sequences. This approach might also add to the future coverage of the method as is allows the specialist to make their own networks, since they have the highest expertise in their field.

The CMGfunc and CMG-biotools presented here are methods for creating stand alone tools. The systems offer several advantages, including no need for internet access which is useful for confidential data analysis. Furthermore, the virtual computer setup makes the systems installable on a wide range of platforms and makes it easy to receive support on the method, as every user is using the method on the exact same system. Using this type of setup for distributing a system or method also allows for through reproduction of results as any user has access to the same system on which the method was developed. Lastly, both of these methods have been documented on wiki type web pages, giving users and peers access to up to date documentation and changes on the system.

As more genomes, and metagenomes, become available, it is hoped that speed and coverage of functional annotation will improve. To some extent, it might be desired to include a larger fraction of false positives in order to insure high coverage, and then afterwards, add additional tests to filter out the false results. Such approaches might be needed to discover new functions of genes. In the context of new discoveries, it is also important to bring the sequence analysis to the biologists, allowing them to combine their biological knowledge with bioinformatics, and not wait until biological and sequence results have been published separately. Although much work is still left to be done resources are flowing into the area of sequence analysis and progress is being made every day. As such, many different approach are being tried out and tested which will, in time, improve the knowledge gained from sequencing genomes.

Bibliography

[1] R D Fleischmann et al. “Whole-genome random sequencing and assembly of Haemophilus influenzae Rd.” In: *Science* 269.5223 (July 1995), pp. 496–512. ISSN: 0036-8075.

[2] C M Fraser et al. “The minimal gene complement of Mycoplasma genitalium.” In: *Science* 270.5235 (Oct. 1995), pp. 397–403. ISSN: 0036-8075.

[3] E S Lander et al. “Initial sequencing and analysis of the human genome.” In: *Nature* 409.6822 (Feb. 2001), pp. 860–921. ISSN: 0028-0836.

[4] Alban Mancheron, Raluca Uricaru, and Eric Rivals. “An alternative approach to multiple genome comparison.” In: *Nucleic acids research* (June 2011), pp. 1–11. ISSN: 1362-4962.

[5] Dennis a Benson et al. “GenBank.” In: *Nucleic acids research* 41.Database issue (Jan. 2013), pp. D36–42. ISSN: 1362-4962.

[6] The UniProt Consortium. “Update on activities at the Universal Protein Resource (UniProt) in 2013.” In: *Nucleic acids research* 41.Database issue (Jan. 2013), pp. D43–7. ISSN: 1362-4962.

[7] Michele Magrane and Uniprot Consortium. “UniProt Knowledgebase: a hub of integrated protein data.” In: *Database : the journal of biological databases and curation* 2011 (Jan. 2011), bar009. ISSN: 1758-0463.

[8] Konstantinos Liolios et al. “The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata.” In: *Nucleic acids research* 38.Database issue (Jan. 2010), pp. D346–54. ISSN: 1362-4962.

[9] Richard J Roberts et al. “COMBREX: a project to accelerate the functional annotation of prokaryotic genomes.” In: *Nucleic acids research* 39.Database issue (Jan. 2011), pp. D11–4. ISSN: 1362-4962.

[10] Alfonso Benítez-Páez. “Considerations to improve functional annotations in biological databases.” In: *Omics : a journal of integrative biology* 13.6 (Dec. 2009), pp. 527–35. ISSN: 1557-8100.

[11] Selvarajan Sivashankari and Piramanayagam Shanmughavel. “Functional annotation of hypothetical proteins - A review.” In: *Bioinformation* 1.8 (Jan. 2006), pp. 335–8. ISSN: 0973-2063.

[12] Iddo Friedberg. “Automated protein function prediction–the genomic challenge.” In: *Briefings in bioinformatics* 7.3 (Sept. 2006), pp. 225–42. ISSN: 1467-5463.

[13] Michael Y. Galperin and Eugene V. Koonin. “From complete genome sequence to, complete, understanding?” In: *Trends in Biotechnology* 28.8 (Aug. 2010), pp. 398–406. ISSN: 01677799.

[14] M Riley. “Functions of the gene products of Escherichia coli.” In: *Microbiological reviews* 57.4 (Dec. 1993), pp. 862–952. ISSN: 0146-0749.

[15] M a Andrade, C Ouzounis, C Sander, J Tamames, and a Valencia. “Functional classes in the three domains of life.” In: *Journal of molecular evolution* 49.5 (Nov. 1999), pp. 551–7. ISSN: 0022-2844.

[16] F. R. Blattner. “The Complete Genome Sequence of Escherichia coli K-12”. In: *Science* 277.5331 (Sept. 1997), pp. 1453–1462. ISSN: 00368075.

[17] R L Tatusov, M Y Galperin, D a Natale, and E V Koonin. “The COG database: a tool for genome-scale analysis of protein functions and evolution.” In: *Nucleic acids research* 28.1 (Jan. 2000), pp. 33–6. ISSN: 0305-1048.

[18] J.A. Eisen. “Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis”. In: *Genome research* 8.3 (1998), p. 163.

[19] David Vallenet et al. “MaGe: a microbial genome annotation system supported by synteny results.” In: *Nucleic acids research* 34.1 (Jan. 2006), pp. 53–65. ISSN: 1362-4962.

[20] Michael Ashburner et al. “Gene Ontology: tool for the unification of biology”. In: *Nature genetics* 25.1 (May 2000), pp. 25–29. ISSN: 1061-4036.

[21] M A Harris et al. “The Gene Ontology (GO) database and informatics resource.” In: *Nucleic acids research* 32.Database issue (2004), pp. D258–D261.

[22] Kevin Galens et al. “The IGS Standard Operating Procedure for Automated Prokaryotic Annotation”. In: *Standards in Genomic Sciences* 4.2 (Apr. 2011), pp. 244–251. ISSN: 1944-3277.

[23] Konstantinos Mavromatis, Natalia N. Ivanova, I-Min a. Chen, Ernest Szeto, Victor M. Markowitz, and Nikos C. Kyrpides. “The DOE-JGI Standard Operating Procedure for the Annotations of Microbial Genomes”. In: *Standards in Genomic Sciences* 1.1 (July 2009), pp. 63–67. ISSN: 1944-3277.

[24] David M. Tanenbaum et al. “The JCVI standard operating procedure for annotating prokaryotic metagenomic shotgun sequencing data”. In: *Standards in Genomic Sciences* 2.2 (Mar. 2010), pp. 229–237. ISSN: 1944-3277.

[25] John E Beaver, Murat Tasan, Francis D Gibbons, Weidong Tian, Timothy R Hughes, and Frederick P Roth. “FuncBase: a resource for quantitative gene function annotation.” In: *Bioinformatics (Oxford, England)* 26.14 (July 2010), pp. 1806–7. ISSN: 1367-4811.

[26] Jesse Gillis and Paul Pavlidis. “The impact of multifunctional genes on ”guilt by association” analysis.” In: *PloS one* 6.2 (Jan. 2011), e17258. ISSN: 1932-6203.

[27] Ruben E Valas and Philip E Bourne. “Save the tree of life or get lost in the woods.” In: *Biology direct* 5.1 (Jan. 2010), p. 44. ISSN: 1745-6150.

[28] David A Lee, Robert Rentzsch, and Christine Orengo. "GeMMA: functional subfamily classification within superfamilies of predicted protein structural domains." In: *Nucleic acids research* 38.3 (Jan. 2010), pp. 720–37. ISSN: 1362-4962.

[29] François Jacob and Jacques Monod. "Genetic regulatory mechanisms in the synthesis of proteins". In: *Journal of Molecular Biology* 3.3 (June 1961), pp. 318–356. ISSN: 00222836.

[30] Blanca Taboada, Cristina Verde, and Enrique Merino. "High accuracy operon prediction method based on STRING database scores." In: *Nucleic acids research* 38.12 (Apr. 2010). ISSN: 1362-4962.

[31] Ramy K Aziz, Mya Breitbart, and Robert a Edwards. "Transposases are the most abundant, most ubiquitous genes in nature." In: *Nucleic acids research* 38.13 (July 2010), pp. 4207–17. ISSN: 1362-4962.

[32] Amit Dhingra, Archie R Portis, and Henry Daniell. "Enhanced translation of a chloroplast-expressed RbcS gene restores small subunit levels and photosynthesis in nuclear RbcS antisense plants." In: *Proceedings of the National Academy of Sciences of the United States of America* 101.16 (Apr. 2004), pp. 6315–20. ISSN: 0027-8424.

[33] Murray Ronald Badger and Emily Jane Bek. "Multiple Rubisco forms in proteobacteria: their functional significance in relation to CO2 acquisition by the CBB cycle." In: *Journal of experimental botany* 59.7 (Jan. 2008), pp. 1525–41. ISSN: 1460-2431.

[34] Agnieszka S Juncker et al. "Sequence-based feature prediction and annotation of proteins." In: *Genome biology* 10.2 (Jan. 2009), p. 206. ISSN: 1465-6914.

[35] James C Whisstock and Arthur M Lesk. "Prediction of protein function from protein sequence and structure." In: *Quarterly reviews of biophysics* 36.3 (Aug. 2003), pp. 307–40. ISSN: 0033-5835.

[36] Michal Brylinski and Jeffrey Skolnick. "Comparison of structure-based and threading-based approaches to protein functional annotation." In: *Proteins* 78.1 (Jan. 2010), pp. 118–34. ISSN: 1097-0134.

[37] D T Jones. "Protein secondary structure prediction based on position-specific scoring matrices." In: *Journal of molecular biology* 292.2 (Sept. 1999), pp. 195–202. ISSN: 0022-2836.

[38] Morten Nielsen, Claus Lundegaard, Ole Lund, and Thomas Nordahl Petersen. "CPHmodels-3.0–remote homology modeling using structure-guided sequence profiles." In: *Nucleic acids research* 38.Web Server issue (July 2010), W576–81. ISSN: 1362-4962.

[39] a Krogh, B Larsson, G von Heijne, and E L Sonnhammer. "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes." In: *Journal of molecular biology* 305.3 (Jan. 2001), pp. 567–80. ISSN: 0022-2836.

[40] a Lupas, M Van Dyke, and J Stock. “Predicting coiled coils from protein sequences.” In: *Science (New York, N.Y.)* 252.5009 (May 1991), pp. 1162–4. ISSN: 0036-8075.

[41] Aron Marchler-Bauer et al. “CDD: a Conserved Domain Database for the functional annotation of proteins.” In: *Nucleic acids research* 39.Database issue (Nov. 2010), pp. D225–9. ISSN: 1362-4962.

[42] Ivica Letunic, Tobias Doerks, and Peer Bork. “SMART 7: recent updates to the protein domain annotation resource.” In: *Nucleic acids research* 40.Database issue (Jan. 2012), pp. D302–5. ISSN: 1362-4962.

[43] Tania Lima et al. “HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot.” In: *Nucleic Acids Research* 37.Database issue (2009), pp. D471–8. ISSN: 13624962.

[44] A R Mushegian and E V Koonin. “A minimal gene set for cellular life derived by comparison of complete bacterial genomes.” In: *Proceedings of the National Academy of Sciences of the United States of America* 93.19 (Sept. 1996), pp. 10268–73. ISSN: 0027-8424.

[45] a Bairoch and R Apweiler. “The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.” In: *Nucleic acids research* 28.1 (Jan. 2000), pp. 45–8. ISSN: 0305-1048.

[46] Kim D Pruitt, Tatiana Tatusova, and Donna R Maglott. “NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.” In: *Nucleic acids research* 33.Database issue (Jan. 2005), pp. D501–4. ISSN: 1362-4962.

[47] Y Tateno et al. “DNA Data Bank of Japan (DDBJ) for genome scale research in life science.” In: *Nucleic acids research* 30.1 (Jan. 2002), pp. 27–30. ISSN: 1362-4962.

[48] Marco Punta et al. “The Pfam protein families database.” In: *Nucleic acids research* 40.November 2011 (Nov. 2011), pp. 290–301. ISSN: 1362-4962.

[49] Sarah Hunter et al. “InterPro: the integrative protein signature database.” In: *Nucleic acids research* 37.Database issue (Jan. 2009), pp. D211–5. ISSN: 1362-4962.

[50] Christian J a Sigrist et al. “New and continuing developments at PROSITE.” In: *Nucleic acids research* 41.Database issue (Jan. 2013), pp. D344–7. ISSN: 1362-4962.

[51] D H Haft et al. “TIGRFAMs: a protein family resource for the functional identification of proteins.” In: *Nucleic acids research* 29.1 (Jan. 2001), pp. 41–3. ISSN: 1362-4962.

[52] Huaiyu Mi, Anushya Muruganujan, and Paul D Thomas. “PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees.” In: *Nucleic acids research* 41.Database issue (Jan. 2013), pp. D377–86. ISSN: 1362-4962.

[53] a G Murzin, S E Brenner, T Hubbard, and C Chothia. “SCOP: a structural classification of proteins database for the investigation of sequences and structures.” In: *Journal of molecular biology* 247.4 (Apr. 1995), pp. 536–40. ISSN: 0022-2836.

[54] Katerina Michalickova et al. “SeqHound: biological sequence and structure database as a platform for bioinformatics research.” In: *BMC bioinformatics* 3 (Oct. 2002), p. 32. ISSN: 1471-2105.

[55] Herbert Schmidt and Michael Hensel. “Pathogenicity islands in bacterial pathogenesis”. In: *Clinical Microbiology Reviews* 17.1 (2004), p. 14. ISSN: 0893-8512.

[56] Nicki Tiffin, Miguel A Andrade-Navarro, and Carolina Perez-Iratxeta. “Linking genes to diseases: it’s all in the data.” In: *Genome medicine* 1.8 (Jan. 2009), p. 77. ISSN: 1756-994X.

[57] Ea Zankari et al. “Identification of acquired antimicrobial resistance genes.” In: *The Journal of antimicrobial chemotherapy* 67.11 (Nov. 2012), pp. 2640–4. ISSN: 1460-2091.

[58] Euan A Adie, Richard R Adams, Kathryn L Evans, David J Porteous, and Ben S Pickard. “Speeding disease gene discovery by sequence based candidate prioritization.” In: *BMC bioinformatics* 6 (Jan. 2005), p. 55. ISSN: 1471-2105.

[59] Pimlapas Leekitcharoenphon, Oksana Lukjancenko, Carsten Friis, Frank M Aarestrup, and David W Ussery. “Genomic variation in Salmonella enterica core genes for epidemiological typing.” In: *BMC genomics* 13.1 (Jan. 2012), p. 88. ISSN: 1471-2164.

[60] Arthur L Delcher, Kirsten a Bratke, Edwin C Powers, and Steven L Salzberg. “Identifying bacterial genes and endosymbiont DNA with Glimmer.” In: *Bioinformatics (Oxford, England)* 23.6 (Mar. 2007), pp. 673–9. ISSN: 1367-4811.

[61] a V Lukashin and M Borodovsky. “GeneMark.hmm: new solutions for gene finding.” In: *Nucleic acids research* 26.4 (Feb. 1998), pp. 1107–15. ISSN: 0305-1048.

[62] Doug Hyatt, Gwo-Liang Chen, Philip F Locascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. “Prodigal: prokaryotic gene recognition and translation initiation site identification.” In: *BMC bioinformatics* 11 (Jan. 2010), p. 119. ISSN: 1471-2105.

[63] T.S. Larsen and Anders Krogh. “EasyGene - a prokaryotic gene finder that ranks ORFs by statistical significance”. In: *BMC bioinformatics* 4.1 (2003), p. 21. ISSN: 1471-2105.

[64] Huaiqiu Zhu, Gang-Qing Hu, Yi-Fan Yang, Jin Wang, and Zhen-Su She. “MED: a new non-supervised gene prediction algorithm for bacterial and archaeal genomes.” In: *BMC bioinformatics* 8 (Jan. 2007), p. 97. ISSN: 1471-2105.

[65] A. Al-Shahib, Rainer Breitling, David Gilbert, and Others. “FRANKSUM: New feature selection method for protein function prediction”. In: *International journal of neural systems* 15.4 (2005), pp. 259–276.

[66] L.J. Jensen et al. “Prediction of Human Protein Function from Post-translational Modifications and Localization Features”. In: *Journal of Molecular Biology* 319.5 (June 2002), pp. 1257–1265. ISSN: 00222836.

[67] Tina Koestler, Arndt von Haeseler, and Ingo Ebersberger. “FACT: functional annotation transfer between proteins with similar feature architectures.” In: *BMC bioinformatics* 11 (Jan. 2010), p. 417. ISSN: 1471-2105.

[68] Bum Ju Lee, Moon Sun Shin, Young Joon Oh, Hae Seok Oh, and Keun Ho Ryu. “Identification of protein functions using a machine-learning approach based on sequence-derived properties.” In: *Proteome science* 7 (Jan. 2009), p. 27. ISSN: 1477-5956.

[69] E. Gasteiger. “ExPASy: the proteomics server for in-depth protein knowledge and analysis”. In: *Nucleic Acids Research* 31.13 (July 2003), pp. 3784–3788. ISSN: 1362-4962.

[70] Nancy Y Yu et al. “PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes.” In: *Bioinformatics (Oxford, England)* 26.13 (July 2010), pp. 1608–15. ISSN: 1367-4811.

[71] Jannick Dyrlø v Bendtsen, Henrik Nielsen, Gunnar von Heijne, and Sø ren Brunak. “Improved prediction of signal peptides: SignalP 3.0.” In: *Journal of molecular biology* 340.4 (July 2004), pp. 783–95. ISSN: 0022-2836.

[72] JC Wootton and S Federhen. “Statistics of local complexity in amino acid sequence databases”. In: *Computers and chemistry* 17.2 (1993), pp. 149–163.

[73] Robert D Finn et al. “The Pfam protein families database.” In: *Nucleic acids research* 38.Database issue (Jan. 2010), pp. D211–22. ISSN: 1362-4962.

[74] Derek Wilson, Martin Madera, Christine Vogel, Cyrus Chothia, and Julian Gough. “The SUPERFAMILY database in 2007: families and functions.” In: *Nucleic acids research* 35.Database issue (Jan. 2007), pp. D308–13. ISSN: 1362-4962.

[75] Jeremy D Selengut et al. “TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes”. In: *Nucleic Acids Research* 35.Database issue (2007), pp. D260–D264.

[76] BS Saritha and S Hemanth. “An Efficient Hidden Markov Model for Offline Handwritten Numeral Recognition”. In: *arXiv preprint arXiv:1001.5334* (2010).

[77] Sean R Eddy. “What is a hidden Markov model?” In: *Nature biotechnology* 22.10 (Oct. 2004), pp. 1315–6. ISSN: 1087-0156.

[78] E L Sonnhammer, S R Eddy, and R Durbin. “Pfam: a comprehensive database of protein domain families based on seed alignments.” In: *Proteins* 28.3 (July 1997), pp. 405–20. ISSN: 0887-3585.

[79] Barbara E Engelhardt, Michael I Jordan, John R Srouji, and Steven E Brenner. “Genome-scale phylogenetic function annotation of large and diverse protein families.” In: *Genome research* (July 2011), pp. 1969–1980. ISSN: 1549-5469.

[80] Ana Conesa and Stefan Götz. “Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics.” In: *International journal of plant genomics* 2008 (Jan. 2008), p. 619832. ISSN: 1687-5370.

[81] Theodoros G Soldatos et al. “Martini: using literature keywords to compare gene sets.” In: *Nucleic acids research* 38.1 (Jan. 2010), pp. 26–38. ISSN: 1362-4962.

[82] Karin Lagesen, Peter Hallin, Einar Andreas Rø dland, Hans-Henrik Staerfeldt, Torbjø rn Rognes, and David W Ussery. “RNAmmer: consistent and rapid annotation of ribosomal RNA genes.” In: *Nucleic acids research* 35.9 (2007), pp. 3100–3108. ISSN: 1362-4962.

[83] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, and Others. “Basic Local Alignment Search Tool”. In: *Journal of molecular biology* 215.3 (1990), pp. 403–410.

[84] AS Juncker and Hanni Willenbrock. “Prediction of lipoprotein signal peptides in Gram‚Äênegative bacteria”. In: *Protein ...* 1994 (2003), pp. 1652–1662.

[85] Hideki Noguchi, Jungho Park, and Toshihisa Takagi. “MetaGene: prokaryotic gene finding from environmental genome shotgun sequences.” In: *Nucleic acids research* 34.19 (Jan. 2006), pp. 5623–30. ISSN: 1362-4962.

[86] Minoru Kanehisa, Susumu Goto, Shuichi Kawashima, Yasushi Okuno, and Masahiro Hattori. “The KEGG resource for deciphering the genome”. In: *Nucleic Acids Research* 32.Database issue (2004), pp. D277–D280.

[87] Roman L Tatusov et al. “The COG database: an updated version includes eukaryotes.” In: *BMC bioinformatics* 4 (Sept. 2003), p. 41. ISSN: 1471-2105.

[88] C. Claudel-Renard. “Enzyme-specific profiles for genome annotation: PRIAM”. In: *Nucleic Acids Research* 31.22 (Nov. 2003), pp. 6633–6639. ISSN: 1362-4962.

[89] E Quevillon et al. “InterProScan: protein domains identifier.” In: *Nucleic acids research* 33.Web Server issue (July 2005), W116–20. ISSN: 1362-4962.

[90] John Moult. “A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction.” In: *Current opinion in structural biology* 15.3 (June 2005), pp. 285–9. ISSN: 0959-440X.

[91] Jesse Gillis and Paul Pavlidis. “Characterizing the state of the art in the computational assignment of gene function: lessons from the first critical assessment of functional annotation (CAFA)”. In: *BMC Bioinformatics* 14.Suppl 3 (2013), S15. ISSN: 1471-2105.

[92] Predrag Radivojac et al. “A large-scale evaluation of computational protein function prediction.” In: *Nature methods* 10.3 (Mar. 2013), pp. 221–7. ISSN: 1548-7105.

[93] JA Gerlt, KN Allen, and SC Almo. “The Enzyme Function Initiative”. In: *Biochemistry* 50.46 (2011), pp. 9950–9962.

[94] CHQ Ding and Inna Dubchak. “Multi-class protein fold recognition using support vector machines and neural networks”. In: *Bioinformatics* 17.4 (2001), pp. 349–358.

[95] C Pasquier, V J Promponas, and S J Hamodrakas. "PRED-CLASS: cascading neural networks for generalized protein classification and genome-wide applications." In: *Proteins* 44.3 (Aug. 2001), pp. 361–9. ISSN: 0887-3585.

[96] Pierre. Baldi and Sø ren. Brunak. *Bioinformatics : the machine learning approach.* MIT Press, 2001, xxi, 452 s. ISBN: 026202506x, 9780262025065.

[97] Ole. Lund. *Immunological bioinformatics.* MIT Press, 2005, 296 s. ISBN: 0262122804, 9780262122801.

[98] Andrey O Kislyuk et al. "A computational genomics pipeline for prokaryotic sequencing projects." In: *Bioinformatics (Oxford, England)* 26.15 (Aug. 2010), pp. 1819–26. ISSN: 1367-4811.

[99] John P a Ioannidis et al. "Repeatability of published microarray gene expression analyses." In: *Nature genetics* 41.2 (Feb. 2009), pp. 149–55. ISSN: 1546-1718.

[100] Richard Van Noorden. "Science publishing: The trouble with retractions". In: 21.4 (Nov. 2011), pp. 355–367. ISSN: 09545395.

[101] Johan Rung and Alvis Brazma. "Reuse of public genome-wide gene expression data." In: *Nature reviews. Genetics* 14.2 (Feb. 2013), pp. 89–99. ISSN: 1471-0064.

[102] Iain Hrynaszkiewicz and Matthew J Cockerill. "Open by default: a proposed copyright license and waiver agreement for open access research and data in peer-reviewed journals." In: *BMC research notes* 5.1 (Jan. 2012), p. 494. ISSN: 1756-0500.

[103] David L Donoho. "An invitation to reproducible computational research." In: *Biostatistics (Oxford, England)* 11.3 (July 2010), pp. 385–8. ISSN: 1468-4357.