Analysis and Design of Genomic Sequences

A Dissertation Presented
by
Dimitris Papamichail

to
The Graduate School
in partial fulfillment of the requirements
for the degree of
doctor of philosophy
in
computer science
Stony Brook University

August 2007

Stony Brook University

The Graduate School

Dimitris Papamichail

We, the dissertation committee for the above candidate for the degree of Doctor of Philosophy, hereby recommend acceptance of this dissertation.

Professor Steven S. Skiena, Advisor
Computer Science Department

Professor Amanda Stent, Chairman of Defense
Computer Science Department

Professor Radu Grosu
Computer Science Department

Professor Eckard Wimmer
Microbiology Department

Doctor Daniel van der Lelie
Biology Department, Brookhaven National Laboratory

This dissertation is accepted by the Graduate School.

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

Analysis and Design of Genomic Sequences

by

Dimitris Papamichail

Doctor of Philosophy

in

Computer Science

Stony Brook University

2007

Genomic sequences contain genetic information for the development and functioning of living organisms. Sequence variability can be used both to determine organism identity and as a tool to alter function.

Although microorganisms dominate the biosphere, most have not been identified or studied. In this dissertation, we present an oligonucleotide (k-mer classification method based on conditional probabilities, which performs substantially better than other known methods and can be used to identify bacterial species, even from mixed populations, using modest amounts of sample sequence [96].

Here we also deal with the problem of population analysis, leading to determination of diversity and function of members of microbial communities [72]. We develop homology based tools for robust phylotype determination, enhancing closely related sequence associations, and a methodology for achieving more accurate richness estimation, using different clustering criteria [95].

The emerging field of synthetic biology is broadly defined as the intersection of biology and engineering that focuses on the modification or creation of novel biological systems that do not have a counterpart in nature. Working with the group that achieved the first genome-level synthesis of a virus, we have designed, synthesized, and evaluated new variants of poliovirus to serve as vaccines. Specifically, we sought weakened but viable strains that could be used for preparations of a killed poliovirus vaccine. Our designs result in a virus with roughly 100-fold lower specific infectivity than the wildtype virus. Here we detail the theory behind gene design in the context of optimizing a DNA sequence for particular desired properties while simultaneously coding for a given amino acid sequence [87].

We have also explored the problem of designing the provably shortest genomic sequence to encode a given set of genes by exploiting alternate reading frames. We have developed an algorithm for designing the shortest DNA sequence simultaneously encoding two given amino acid sequences. We have shown that the coding sequences of naturally occurring pairs of overlapping genes approach maximum compression, as well as investigated the impact of alternate coding matrices on overlapping sequence design [129].

Contents

I Genomic Sequence Analysis 7

List of Tables

List of Figures

Acknowledgements

I would like to acknowledge many people for helping me during my doctoral work. I would especially like to thank my advisor, Professor Steven Skiena, for his generous time, commitment and help in every aspect of my academic life. Throughout my doctoral work he encouraged me to develop independent thinking and research skills, while paving the path with continuous support. His patience has proven exemplary and his assistance with my scientific writing invaluable. Undoubtfully, I would have not achieved even a fraction of this work without his mentoring and I will always crave for his advice in the years to come.

I am also very grateful to Dr. van der Lelie, an exceptional researcher and mentor, who not only financed a great deal of my work and time, but also led a fruitful and enjoyable collaboration.

Professor Eckard Wimmer, Dr. Steffen Mueller and Rob Coleman have been a great team to work with and deserve special thanks for introducing me to the exciting area of synthetic biology and for all our amazing conversations, where science was explained and created at the same time.

I am extremely grateful for the assistance, generosity, and advice I received from Dr. Sean McCorkle, to whom I owe a big part of my knowledge in bioinformatics, and who is a constant source of interesting problems and discussions. Special thanks to my collaborator Dr. Celine Lesaulnier, for all her help and hard work in analyzing microbial populations, Bei Wang, for involving me in

the gene overlapping project, and Professor Nickolas Delihas, for his excellent advice, tranfer of knowledge and collaboration on a fascinating subject.

Finally, I would like to thank my family, both in Greece and the US. My parents Panayiotis and Katerina were always behind me in every step, with endless love and support in every point in my life. No words here could really describe their contributions. In addition I would like to thank my brother Giorgos, for his support and encouragement all these years, but also giving me a boost in jumpstarting my research. And of course my wife Jaehee, who cared lovingly for me and our two children, Panayiotis and Katerina, for the past five years, enabling the completion of this work.

Chapter 1

Introduction

Bioinformatics is a field that derives knowledge from computer analysis of biological data. This data can consist of information stored in the genetic code, but also experimental results from various sources, patient statistics, and the scientific literature. Research in bioinformatics includes development of methods for the storage, retrieval, and analysis of data. Bioinformatics is a rapidly developing branch of biology and is highly interdisciplinary, using techniques and concepts from informatics, statistics, mathematics, chemistry, biochemistry, physics, and linguistics. It has many practical applications in different areas of biology and medicine.

In this dissertation we will concentrate in two major branches of bioinformatics. The first is sequence analysis, as in classification, phylogenetic analysis and motif finding. The second is sequence design and synthesis.

In the course of my studies I have collaborated with two groups of biologists: One headed by Dr. van der Lelie in the Biology Department of Brookhaven National Laboratory (BNL) and the other headed by Professor Eckard Wimmer in the Microbiology Department of SUNY at Stony Brook. The BNL group seeks analysis of microbial populations and phylogeny determination of microorganisms associated with plants, either endophytically or in their rhizosphere. With the second group we have designed, synthesized and analyzed an array of viral segments/genes, for the purposes of attenuating the translation of these viruses and creating candidate vaccines.

1.1 Sequence Classification and Analysis

Phylogenetic methods can be used for many purposes, including analyzing morphological and molecular data. Here we concentrate on the analysis of DNA and RNA sequences.

In the first part of this dissertation, we will attempt to answer the following question:

> How can we conduct a census of the members of a bacterial community, when the overwhelming number have never been sequenced?

In this problem the input is sequence data from random samples of one or more communities, and the output expected is (1) the determination of the phylogenetic groups present and (2) how these groups change under different environmental conditions. With the motive of global warming, answering this question today is more crucial than ever.

Although there are several techniques and tools for discovering the origin of sequences, they cannot detect genomic fragments not resembling anything registered in databases, especially when these fragments are randomly extracted and are not associated with some known gene or area of a genome. But hope comes from the results of Sandberg et al. [109], who investigated the identification of bacterial genomic sequences with the aid of oligonucleotide distributions instead of sequence matching. Using a naïve Bayesian classifier and the fact that the oligonucleotide distribution of an organism comprises a "genomic signature", capable of distinguishing among different taxa while varying very little intergenomically, random fragments can now be classified to their closest known sequenced relative. Here we enhance this classification method by exploiting the overlapping nature of oligonucleotides, as well as explore the identification possibilities in mixed populations.

There are methods to assay complex microbial communities other than extracting random genomic fragments however. One of the most common is targeting well known, universally distributed, functionally constant and conserved regions, such as the ribosomal RNA genes. Using taxonomically specific primers, meaning conserved oligonucleotides at certain locations of such genes, one can isolate a large number of samples with minimum amount of sequencing and a well defined target in each sample. Approximate sequence matching techniques work accurately to identify the origin of these genes, having tens of thousands of reference sequences available to compare them to. We used such techniques for a large scale phylogenetic analysis project, aiming to observe microbial population changes under conditions of elevated atmospheric CO_2. By parameterized approximate local alignment and significance evaluations of classification assignments, we analyzed the major groups involved in the changes observed under conditions of elevated CO_2. We also determined the richness and diversity of these populations, using provably accurate and robust clustering techniques.

In several occasions a microbiologist acts as an investigator, trying to solve a case using clues and arguments, based on evidence and proof. In the case of Photorhabdus luminescens, an effective killer of insects and potential source of effective insecticide toxins, the case involved the discovery of outer membrane porin proteins and associated regulatory RNA genes. Using motif searching and pattern matching techniques, several “suspects” were identified and analyzed in the search for the ideal candidate fitting the profile. Despite our efforts, the *micF* regulatory antisense RNA gene seems absent from P. luminescens, a fact that is supported by the life cycle and symbiotic relationships of the bacterium.

1.2 Genomic Sequence Synthesis

Genomic synthesis is signaling a new area of research that combines science and engineering in order to design and build novel biological functions and systems. Biologists are interested in learning more about how natural living systems work. One direct (but not always simple) way to test our current understanding of a natural living system is to build an instance (or version) of the system in accordance with our current understanding of its functionality. This way we can test a hypothesis on a complex system, adjusting only a few controls.

From an engineering point of view, biology can be viewed as a technology. Synthetic Biology includes the broad redefinition and expansion of biotechnology, with the ultimate goals of being able to design and build engineered biological systems that process information, manipulate chemicals, fabricate materials and structures, produce energy, provide food, and maintain and enhance human health and our environment. The reverse engineering of nature may often oversimplify the mechanisms governing bio-processes, but discovers major principles building a block at a time, often recovering details in its way.

In this part of the dissertation we will attempt to answer the following question:

> How can we rapidly create a vaccine for a new viral disease?

This problem expects as input the genome of a virus, while it returns a design of a “better” virus, to serve as a vaccine. The need and motives for the existence of such a methodology are obvious, however it is only today that synthesis technologies have advanced to such a degree that low cost and high efficiency permit its development.

RNA viruses is the largest virus group; it contains some of the most dreaded human pathogens, like HIV, Ebola, SARS and Influenza. High mutation rates confer high adaptability to changing conditions and environments and they escape from human intervention using drugs, leaving few effective treatment options. Weakening a pathogenic virus to a degree that is safe for human administration while retaining ability to elicit protective immune response creates live attenuated virus vaccines. But the process of passaging the pathogenic virus through diverse non-human cell cultures and animal hosts in order to acquire mutations and adapt to new host conditions, so that it loses its pathogenic potential in humans, is poorly defined, costly and time consuming.

In this dissertation we demonstrate methods to engineer attenuated stable virus vaccines by introducing synonymous mutations to alter the translation efficiency of the virus. Using the ideas of species-specific codon bias and the effects of altered codon distribution towards underrepresented codons in humans, we synthesized capsid proteins that translate minimally, while encoding for the same capsid proteins and thus eliciting a robust immune response.

Codon pair bias was discovered in prokaryotic cells, but has since been seen in all other examined species, including humans, and has high statistical significance. To investigate the effects of altered codon pair distribution we designed and synthesized optimized poliovirus capsid encodings, using over- and under-represented human codon pairs. The algorithms we developed for designing these novel sequences were used to embed/remove patterns, secondary structures, and altering the codon and codon pair distributions. Since heterogeneous combinations of such preferences lead to computationally intractable (NP-complete) problems, the simulated annealing heuristic was employed to realize our designs.

Another problem we explore here is the design of the provably shortest genomic sequence to encode a given set of genes by exploiting alternate reading frames. We present a dynamic programming algorithm for designing the shortest DNA sequence simultaneously encoding two given amino acid sequences. We show that the coding sequence of naturally occurring pairs of overlapping genes approach maximum compression and we investigate the impact of alternate coding matrices on overlapping sequence design.

Part I

Genomic Sequence Analysis

Chapter 2

Oligonucleotide Classification of Microbial Genomic Sequences

2.1 Introduction

Microorganisms are the largest reservoir of genetic and biochemical diversity on earth. Understanding the structure, functional roles, and diversity of complex communities of microbes is key to using their wide-ranging capabilities. Microorganisms dominate the biosphere, yet most have not been identified or studied. Traditional methods for culturing and characterizing microorganisms limit analysis to those that will grow under laboratory conditions, which represent less than 1% of all microorganisms. The recent surge of research in molecular microbial ecology provided compelling evidence for the existence of many novel types of microorganisms in the environment in numbers and varieties that dwarf those of the comparatively few amenable to laboratory cultivation.

There is currently no effective technology to assay the relative abundance of complex microbial communities. Probe-based methods such as microarrays

This chapter is drawn from our paper [96].

can only hope to detect species which have already been at least partially sequenced; but these represent a vanishingly small fraction of the millions of microbial species. The *genomic sequence tag* (GST) approach, pioneered by Dunn et al. [33], promises to make such analysis possible for the first time. It has important applications in many areas of the life sciences, but particularly in ecological and medical research.

Genomic sequence tags (GSTs) are short (e.g. 21 base) sequence fragments sampled more or less at random from microbial genomes in the given population. Such tags are inexpensive to assay, yet long enough to allow for straightforward species identification against sequence databases. However, such identification techniques cannot hope to identify non-sequenced species, which will constitute the vast majority of microbes into the foreseeable future.

Hope comes from the intriguing results of Sandberg et al. [109], who investigated identifying bacterial genomic sequences using k-mer distributions instead of sequence matching. They found that microbial species could be correctly identified with an accuracy of approximately 85% from k-mer distributions from sequence samples as short as 400 bases. In this chapter, we build on these observations in several directions:

- *Improved Classification Method* – We give a classification method based on conditional probabilities which performs substantially better than the method of Sandberg et al. [109] when using small amounts of sample sequence. In particular, our conditional probability approach improved species identification accuracy by up to 20% for short sequence segments (35bp) over the naive Bayesian classifier. These results are significant, because the cost of an assay increases linearly with the amount of required sequence.

- *Accurate Recognition Using Fragmented Sequence Data* – We demonstrate that k-mer analysis of short sequence tags is *more* effective than analysis of equivalent amounts of contiguous sequence. These results are fortuitous, because they imply that our results can be readily applied to GST and long SAGE [106, 126] assays. They are also surprising, because (1) fragmentation inherently reduces the information available for k-mer analysis, and (2) individual short tags have a low (between 5-8%) sequence-recognition specificity, as shown in Table 1.

- *Signature Analysis for Unsequenced Species* – Recognizing new and unsequenced species is critical to tagging-based population analysis. Success depends upon the extent to which k-mer distribution is preserved among related strains and higher order classifications (order and genus).

  We demonstrate that k-mer distributions are well-preserved among related strains/species, by demonstrating that bacterial genomes can be clustered into natural groups according to k-mer distribution similarities.

  In particular, we demonstrate that we can obtain both coarse phylogenetic relationships [34] and fine information from analyzing genomic signatures.

  We give accurate methods of identifying the order, genus and species of unsequenced bacteria from short tags. In particular, we show that unsequenced bacterial species can be accurately identified with respect to the 16S ribosomal RNA phylogenetic information on the basis of short tags.

- *Frequency Analysis of Mixed Populations* – We demonstrate that it is possible to identify bacterial species from mixed populations via k-mer distributions using modest amounts of sample sequence. Consider sequence tags collected from a mixture of two equally-represented species: our clustering-based approach proves capable of identifying at least one of two species 95% of the time.

> Further, our methods extend beyond species identification to frequency analysis. By careful analysis of modest amounts of sequence data, we can predict the frequency of the most dominant species in a population – even for unsequenced organisms. Further, our predictions grossly match the actual population over wide range of dominant-species frequencies.

This chapter is organized as following. Genomic sequence tag methods and previous work on bacterial population assays are discussed in Section 2.1.1. In Section 2.2, we extend the work of Sandberg et al. [109] on k-mer recognition of contiguous sequence fragments. In Section 2.3, we generalize this work to short sequence tags. We consider the clustering and recognition of unsequenced species with the respect to k-mer distribution and phylogenetic classifications in Section 2.4. Finally, we consider the problem of deconvolving tags from mixed species populations in Section 2.5.

2.1.1 Previous Work

Genomic Sequence Tags (GSTs) are short (21 base) fragments, product of a method for identifying and quantitatively analyzing genomic DNA without a priori knowledge of the genome. The DNA is initially fragmented with a type II restriction enzyme. An oligonucleotide adaptor containing a recognition site for M*me*I, a type IIS restriction enzyme, is then used to release 21-bp tags from fixed positions in the DNA relative to the sites recognized by the fragmenting enzyme. These tags are PCR-amplified, purified, concatenated and sequenced, to create a high-resolution GST sequence profile of the genomic DNA.

The GST approach has proven efficient in providing quantitative information for samples of different microbe sequences, even from non-sequenced genomes. Tags that appear in a sample with significantly different frequencies presumably come from organisms occurring with different frequencies in the population. Difficulty arises when specific organisms appear with similar frequency in the sample, or when tags appear with more than singular multiplicity.

This approach for characterizing prokaryotic or eukaryotic genomes is similar to long serial analysis of gene expression (long SAGE [106, 126]) in that it produces large numbers of positionally defined 21-bp tag sequences that can be used to examine intra-specific genomic variation and, if genome information is available, provide immediate species identity. Other methods of large-scale scanning of microbial genomes on a quantitative and qualitative basis include the *Not*I passporting [134] and the restriction site tagged (RST) microarrays [135], as well as the original SAGE procedure [127, 133, 138], which produces positionally defined short tags of 13 to 14 bp with an increased throughput.

Genomic signatures based on compositions of nucleotides have been proven useful in identifying the origin of small sequences [60, 108, 109]. Frequencies of short sequence motifs – down to the level of dinucleotides – have shown great potential in providing a way of distinguishing different genuses in a coarse level [34], but also differentiate between strains of the same species in eubacterial organisms [62]. Dinucleotide composition was also shown to determine in a great degree the DNA local curvature, which is important in transcription, replication, recombination and chromatin structure [81].

Genomic signatures have been used for identification/detection of pathogenicity islands [62], while differences in the use of mutually symmetric and complementary triplets distinguish between coding and non-coding genomic sequences [90]. Bacterial phage genome signatures are strongly correlated with the nature of the host and the extent to which the phage uses the host-cell machinery [6]. Intragenomically, the dinucleotide relative abundance varies little between 50 kilobase or longer windows on a given genome [30, 44], but is stable even in windows ranging in size from 50 kilobases down to 125 bases [57]. It is difficult though to use genomic signatures in order to differentiate between strands where there are substantial chromosomal rearrangements mediated through homologous recombination or other segment shuffling recombination events which have occurred in nature and are not strongly selected against [14].

Different bacterial genomes have distinct combinations of attributes like characteristic codon usage, G+C content, (ranging from about 75% to 25%), GC strand bias, nearest neighbor frequencies and oligonucleotide frequencies, and although their mechanistic origins are not always entirely clear, these characteristics likely evolve slowly enough to be of use in attempting to decipher evolutionary histories of horizontally transferred DNA regions [14, 64]. The evolutionary implications of microbial genome tetranucleotide frequency biases can produce phylogenetic trees that demonstrate a level of congruence with 16S mRNA trees [99].

Using whole genome signatures and concentrating on a varying number of species one can easily distinguish differences between domains of life and families [20].

Other statistical measures used for characterization and classification of species include: (i) the *linguistic complexity* [122], (ii) the *Chaos Game Representation of Sequences* [30], which provides a unique way of visualizing the frequencies of oligonucleotides in the form of images and construct visual proofs on characteristics observed by image manipulation, and (iii) the *compositional spectrum* [69], which uses a subset of long oligonucleotides ($10 - 25$ bp) and imperfect matching, based on Hamming distance or smallest weighted sum of edit distance. The species specificity of different statistical measures is quantified in [108], where genomic signatures, synonymous codon choice, amino acid usage and G + C content are explored.

	3-mer	4-mer	5-mer	6-mer	7-mer	8-mer
Recognition percentage	5.58%	6.03%	6.51%	6.68%	7.09%	8.16%

Table 1: Average Origin Identification accuracy of 1000 randomly drawn 20-mers for varying k-mer size

2.2 Identifying the Origin of Contiguous Sequences

Sandberg et al. [109] developed a naive Bayesian classifier to investigate the possibility of predicting the genome of origin for a specific genomic sequence. They found that sequences as short as 400 bases could be correctly classified with an accuracy of approximately 85%. The classifier was applied to 25 fully sequenced genomes, all of which came from unrelated species. The samples in all experiments originated from the same set of organisms.

The Sandberg et al. classifier calculates the probability of finding a sequence S of length N in a genome G_i as the product of the $N-(k-1)$ probabilities of finding each of the $N-(k-1)$ k-mers (motifs of length k, $k \leq N$) that constitute S in G_i. This is a valid measure of relating a sequence with a genome which can effectively be used as a rating, although it does not represent a correctly defined probability.

We propose a different method for classifying sequences. Instead of using the absolute probability of a k-mer being drawn from a genome G_i, we calculate the conditional probability of the last character of a k-mer appearing after the $k-1$ preceding characters of the k-mer. This conditional probability takes into consideration the dependence of the overlapping k-mers in a sequence, recognizing that the first $k-1$ characters have already appeared as a suffix of the previous k-mer, so it is the last character of the k-mer that will provide new information. This way we do not have to account for the overlaps independently and do not have to make any further assumptions about the dependence. This modification overcomes the k-mer independence assumptions and does not increase the order of needed computation. Further information can be found in the context of statistical natural language processing [79]. Additionally, the improved classification using this method does not come with any increase in the order of needed computation or ease of implementation. Further information about this method in the context of statistical natural language processing can be found in [79], where in the context of natural language processing, the classification of the previous $n-1$ words, the *history*, is used to predict the next word in an n-gram.

We say that a bacterial genome is identified when the Bayesian/conditional probability, calculated as the product of the individual k-mer statistical probabilities, is the highest among the 104 probabilities calculated for all the genomes.

2.2.1 Experimental Results

In order to compare the two methods with respect to the original study of Sandberg et al., we reproduced the original experiment conditions using 25 eubacteria and archaea species whose completely sequenced genomes were available before September 2001. Random pieces of different sizes were drawn from each of the 25 microbe sequences and k-mer distributions used in calculating the probabilities for varying values of k.

Figure 1 compares the results of the naive Bayesian classifier method and the conditional probability method. We use whole genomic sequences to create the k-mer statistics and also draw random sequences from the same genomes. For each point in the graphs, all 25 microbe sequences are sampled and 10 samples are drawn in random. The classification accuracy is then averaged over the 250 cases.

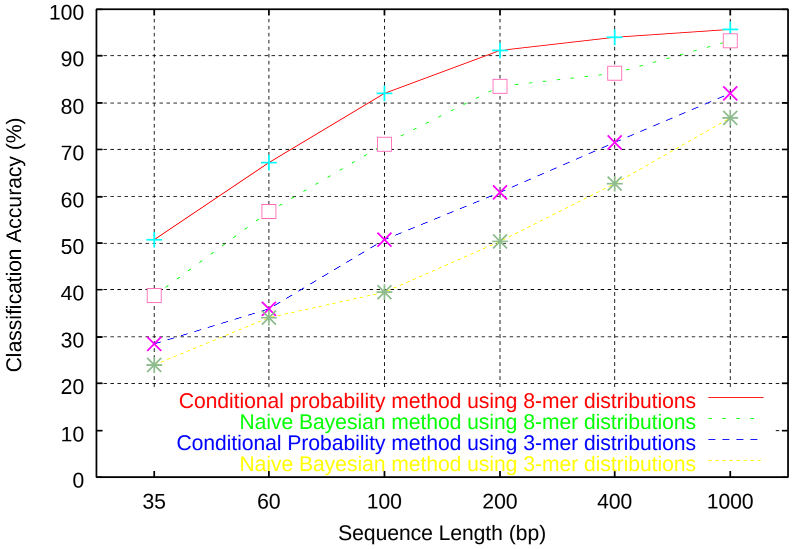


Figure 1: Comparison of Naive Bayesian and Conditional Probability Classifiers.

Figure 1 shows that our conditional probability method performs consistently better, with up to 20% improvement in short sequences of 35 bases. Using the conditional probability method we can now identify short sequences of 400 bases with more than 90% accuracy using 8-mer frequency distributions.

The probabilities in both methods are calculated by multiplying overlapping k-mer probabilities. One must be careful when handling k-mers that do not appear in specific distributions, since the frequency appears as 0. Since we want to be able to classify sequences from unknown bacteria, we must be able to handle k-mers that do not appear in some or all of the available genomes. For that reason, we discount the probabilities of finding a k-mer by assigning a small portion of the probability space to events that have not been encountered. We use Lidstone's Law [79] for discounting:

$$P(w) = \frac{C(w) + \lambda}{N + B\lambda}$$

where P is the assigned probability, w is a training instance, $C(w)$ is the training instance frequency, N is the number of training instances, B is the number of bin training instances are divided into and λ is a constant.

2.2.2 Correcting for Repeated Strains

Sandberg et al. [109] experimented on the 28 different archaea and eubacteria organism genomic sequences available on May 2000. In September 2003, when we started our experiments, 104 full genome sequences were available from NCBI.

Although complete genome sequences are rapidly becoming available, the species diversity of available genomes is increasing at a slower rate because of research biases. Attention is concentrated on human pathogenic microbes, which results in different sequenced strains of similar species.

The frequency profiles of short oligonucleotides (k-mers) of certain length for different microbes, although providing enough specificity for distinguishing different species, becomes less effective for intra-species variation. Sandberg et al. [109] dealt with the problem of reduced specificity by merging multiple strains of the same species in classes, resulting in 25 different classes, out of 28 available microbial sequences.

The 104 available bacterial genomes we studied included several resequenced strains. To eliminate this bias, we grouped bacteria into clusters based on correlation of the k-mer frequency distributions. We grouped the bacteria using agglomerative clustering and the averaging method for merging clusters, resulting in the clustering for 3-mer frequency distributions of Figure 2.

The bacteria sequences can now be grouped according to the height of the cluster difference. into a set of 80 classes. For example, yielding the grouping of Table 2. Agglomerative clustering allows us to divide in any number of classes desired. We found that partitioning into 80 classes satisfied both a close proximity in distribution correlation difference while retaining biological

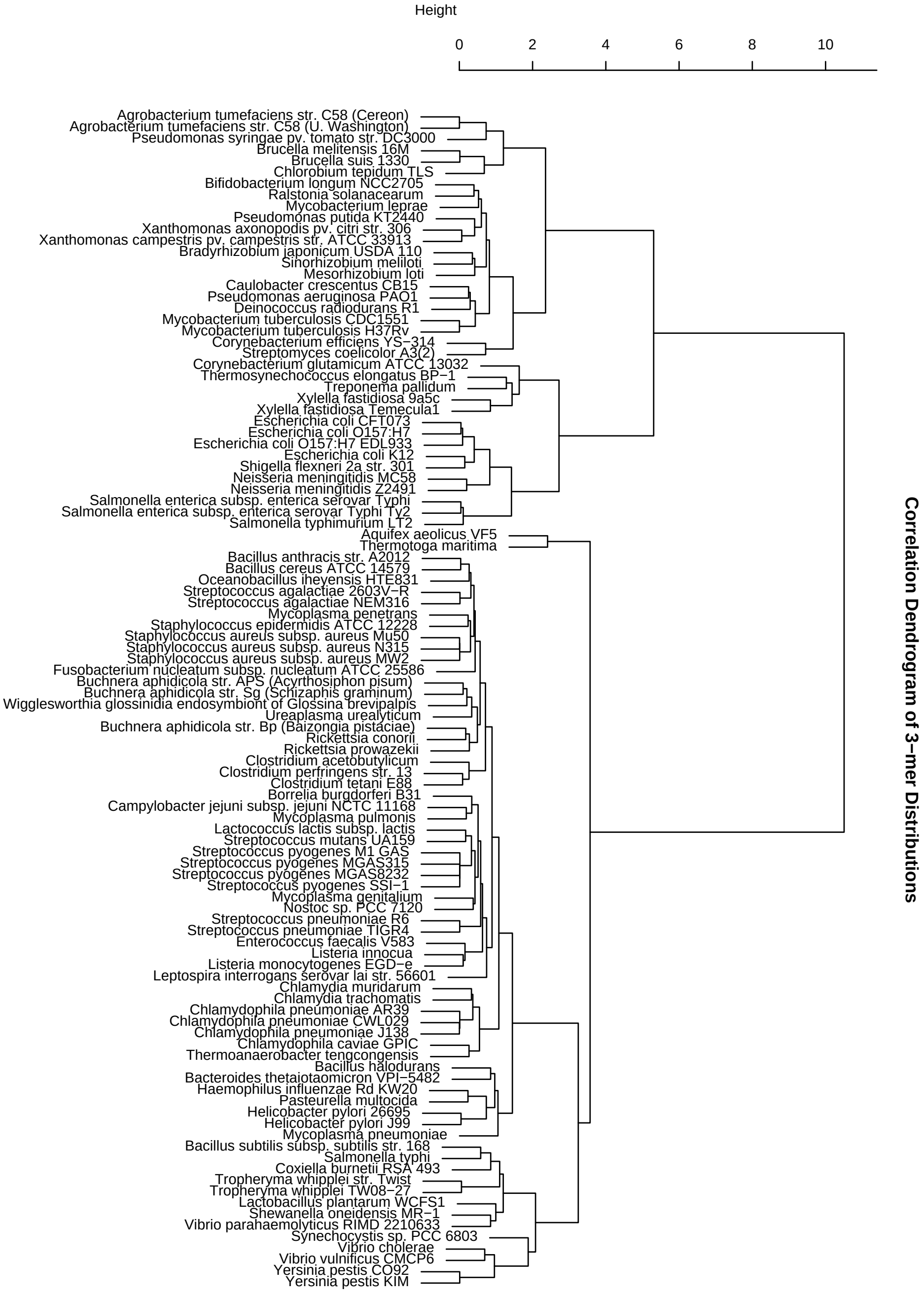


Figure 2: 3-mer frequency distribution clustering based on the averaging (cluster merge) agglomerative method

A. tumefaciens (Cereon) A. tumefaciens (U. Washington)
P. syringae
B. melitensis 16M B. suis 1330
C. tepidum TLS
B. longum NCC2705
R. solanacearum
M. leprae
P. putida KT2440
X. axonopodis X. campestris
B. japonicum USDA 110
S. meliloti
M. loti
C. crescentus CB15
P. aeruginosa PAO1
D. radiodurans R1
M. tuberculosis CDC1551 M. tuberculosis H37Rv
C. efficiens YS-314
S. coelicolor A3(2)
C. glutamicum ATCC 13032
T. elongatus BP-1
T. pallidum
X. fastidiosa 9a5c
X. fastidiosa Temecula1
E. coli CFT073 E. coli O157:H7 E. coli O157:H7 EDL933
E. coli K12
S. flexneri 2a str. 301
N. meningitidis MC58
N. meningitidis Z2491
S. enterica S. enterica Ty2 S. typhimurium LT2
A. aeolicus VF5
T. maritima
B. anthracis str. A2012 B. cereus ATCC 14579
O. iheyensis HTE831
S. agalactiae 2603V-R S. agalactiae NEM316
M. penetrans
S. epidermidis ATCC 12228
S. aureus subsp. aureus Mu50 S. aureus subsp. aureus MW2 S. aureus subsp. aureus N315
F. nucleatum subsp. nucleatum
B. aphidicola str. APS B. aphidicola str. Sg
W. glossinidia
U. urealyticum
B. aphidicola str. Bp
R. conorii
R. prowazekii
C. acetobutylicum
C. perfringens str. 13 C. tetani E88
B. burgdorferi B31
C. jejuni subsp. jejuni
M. pulmonis
L. lactis subsp. lactis
S. mutans UA159
S. pyogenes M1 GAS S. pyogenes MGAS315 S. pyogenes MGAS8232 S. pyogenes SSI-1
M. genitalium
Nostoc sp. PCC 7120
S. pneumoniae R6 S. pneumoniae TIGR4
E. faecalis V583
L. innocua L. monocytogenes EGD-e
L. interrogans serovar lai
C. muridarum
C. trachomatis
C. pneumoniae AR39 C. pneumoniae CWL029 C. pneumoniae J138
C. caviae GPIC
T. tengcongensis
B. halodurans
B. thetaiotaomicron VPI-5482
H. influenzae Rd KW20
P. multocida
H. pylori 26695 H. pylori J99
M. pneumoniae
B. subtilis subsp. subtilis
S. typhi
C. burnetii RSA 493
T. whipplei str. Twist T. whipplei TW08-27
L. plantarum WCFS1
S. oneidensis MR-1
V. parahaemolyticus
S. sp. PCC 6803
V. cholerae
V. vulnificus CMCP6
Y. pestis CO92 Y. pestis KIM

Table 2: Bacterial species clustered into 80 groups

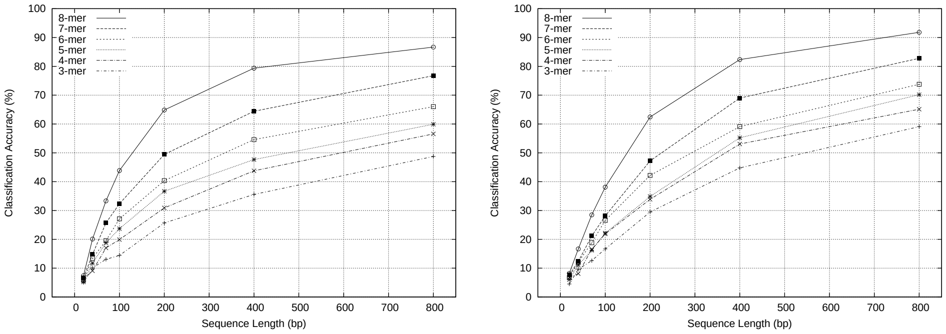
significance, and so will use these classes in subsequent sections of this chapter.

2.3 Dealing with Fragmented Sequences

The genomic sequence tag (GST) method results in fragments of approximately 20 bases extracted from specific locations in a genome, relative to restriction sites. Using short tags has the advantage of avoiding oversampling from repetitive or non-representative (in a genomic signature sense) regions, but individually have low specificity, inadequate of discriminating species, as seen in Table 1.

For a fixed size sample of sequence, fragmented sequences give a reduced amount of k-mers over unfragmented sequences. For example, a sequence of 400 bases can yield 396 5-mers if in one contiguous piece, but only 320 5-mers if the sequence is fragmented into 20 pieces of size 20. Still, for the same sequence length, our methods prove better at identifying fragmented sequences than contiguous sequences. Our results appear in Figure 4(a). Here the contiguous and fragmented sequence experiment results are presented for 3-mer, 6-mer and 8-mer distributions. Graphs of all the results are presented in Figure 3.

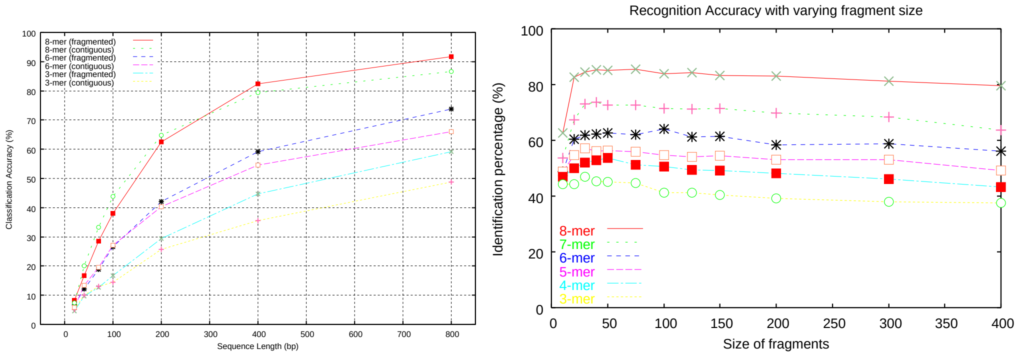
To see how the tag size affects the recognition accuracy, we conducted an experiment where we kept the amount of available sequence constant at 400bp and varied the tag size. The results are shown in Figure 4(b). We observe that the optimal tag size varies with the size of the k-mers used to analyze the data. For distributions of trinucleotide frequencies, the tag length where identification accuracy is maximized is around 30bp, where the optimal tag size is around 75bp for 8-mer frequency distributions. These experiments were performed on all 104 bacteria, with random sampling of 400bp in tags of varying size, where each data point represents 20 averaged repeats.



(a) Classification for contiguous sequences

(b) Classification for fragmented sequences

Figure 3: Sequence identification accuracy as a function of sample length.



(a) Comparison of classification accuracy of contiguous and fragmented sequences

(b) Classification accuracy as a function of fragment/tag size

Figure 4: Classification accuracy for single bacterial targets.

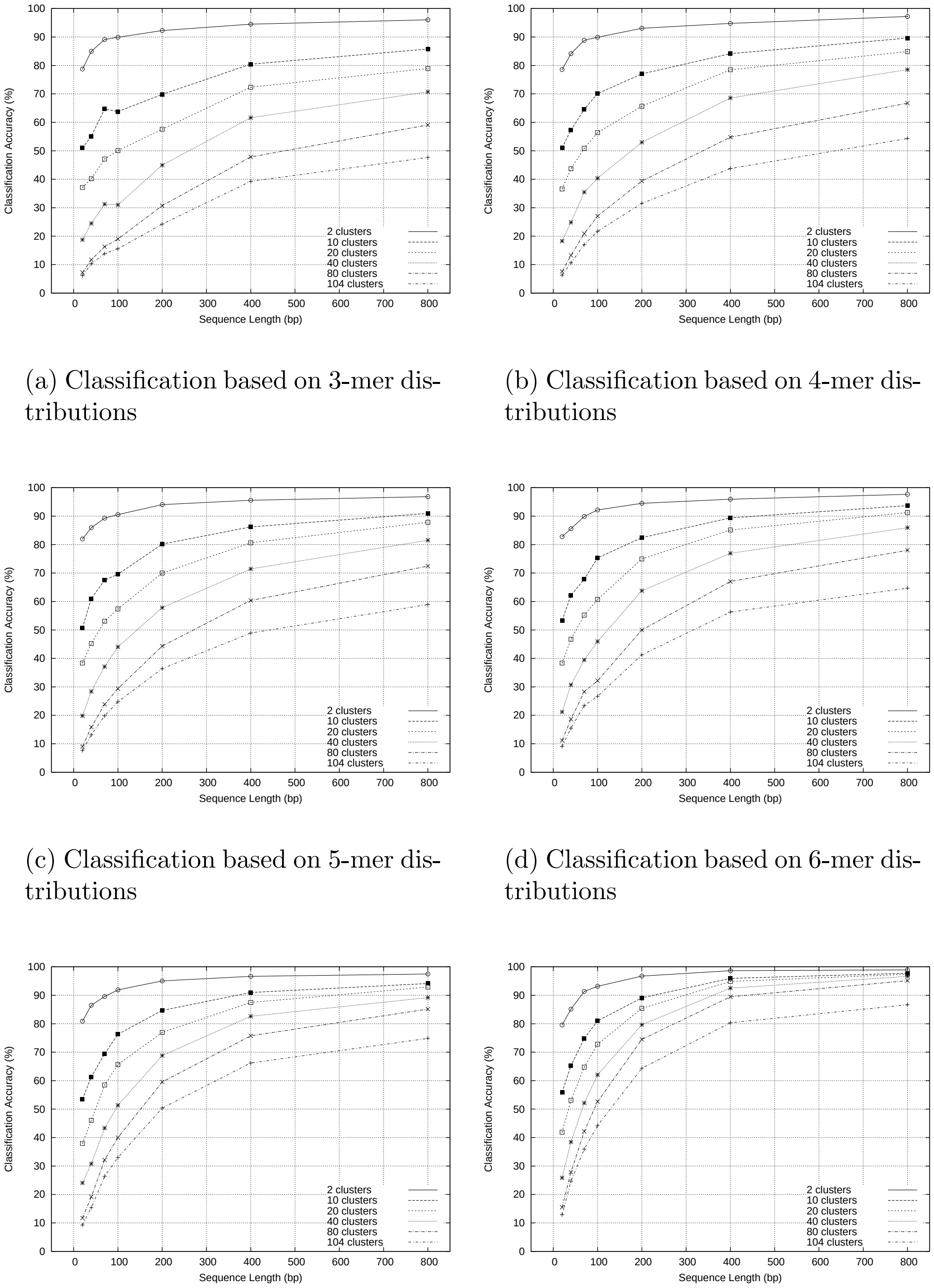
There are two reasons behind this surprising result. First, although the number of k-mers is reduced when using fragmented pieces, the size of the largest independent set of non-overlapping k-mers is not significantly smaller. With fragmented pieces we get at least one new non-overlapping k-mer every time we have a new piece. Second, by sampling from different locations of the genome we decrease the chance that the samples were drawn from an area not representative of the frequency distribution for the specific bacteria. Intergenomic differences are generally higher than intragenomic differences [108, 109].

2.4 Phylogenetic Classification from k-mer Distributions

Estimates of the number of distinct bacterial species go into the millions, which makes it unlikely an observed species will correspond to a sequenced organism. In general, we are interested in obtaining coarser identification than distinct species. Thus we seek to identify which general class of bacteria our prediction indicates as the origin of a sequence. These classes can be formed from the clustering of the bacteria according to their k-mer distributions. We can group bacteria that reside in clusters with a specific distance, represented as the height attribute of the dendogram of Figure 2.

Using the dendrogram derived by the 3-mer distribution correlation and classifying sequences from the 104 bacteria according to whether their class was correctly identified, we get the results shown in Figure 5.

Identifying bacteria in groups instead of single entities makes sense in several ways. First, a number of bacteria have great k-mer similarities to each other, where others have very distant k-mer frequency distributions. Microbial groups which are more important to humans are preferentially sequenced, and



(a) Classification based on 3-mer distributions

(b) Classification based on 4-mer distributions

(c) Classification based on 5-mer distributions

(d) Classification based on 6-mer distributions

(e) Classification based on 7-mer distributions

(f) Classification based on 8-mer distributions

Figure 5: Classification accuracy within varying classes

bacteria from related strains share enough material to have very similar k-mer frequency distributions. Grouping related strains into classes enables us to give a more stable characterization which yields more significant comparative results. Second, our groups of bacteria are generally consistent with the phylogenetic tree constructed by $16S$ ribosomal gene differences and other criteria discussed in Section 2.4.

In this section, we will show that both known and unknown (non fully sequenced) bacteria can be identified with even greater accuracy with respect to phylogenetic categorization.

We use a procaryotic phylogenetic listing of small subunit 16S rRNA found at the Ribosomal Database Project Website [83]. Each bacterial species has a unique index number, consisting of a series of numbers separated by '.', each indicating a different genealogic attribute (kingdom, order, genus, species). We consider each of these numbers as branching points in our inferred tree.

All experiments in this section, involving identifying bacteria based on 16S ribosomal criteria, were averaged over 100 repeats, where fifty 20-mer tags (1000 bp) were randomly selected from each sampling bacteria.

2.4.1 Identifying Bacteria with Known k-mer Statistical Distributions

In this section, we analyze how often the top-scoring bacterium of our classifier happens to match the order, genus, and species of the closest appropriate species in the 16S rRNA database. We measure distance to our sampling bacteria using the inferred subtree (which now contains only our 104 fully sequenced genomes). Closest to our sampling bacteria is considered the species of the inferred subtree with the minimum distance in the number of node traversals (hops) needed to reach the former in the 16S rRNA phylogenetic tree.

Order match	Genus match	Species match	Closest match
99.98%	99.95%	99.83%	99.42%

Table 3: Bacterial classification by ribosomal rRNA similarity (1000bp samples)

The bacterial samples were identified in the correct order with 99.98% accuracy, in the correct genus with 99.95% accuracy and in the correct species category with 99.83% accuracy. The exact strain of origin was identified correctly 99.42% of the time.

The higher than 99% positive identification exceeds even the classification accuracy using the statistically derived clustering tree by approximately 3%, for similar group sizes. For classification in the corresponding order, 98% of all bacteria were correctly classified 100% of the time, where the percentages for perfect identification in the genus and species categories were 97% and 87% respectively.

2.4.2 Identifying Bacteria in the Absence of Statistical Information

Thirty-five additional bacterial genomes were published in the six months after we down-loaded the 104 bacterial genomes known at the beginning of our study. This new data gave us the opportunity to try to identify unknown bacterial sequences, some of which have related strains in our distribution database and others that are distant to any existing entry. To determine the accuracy of our classification we use the 16S ribosomal phylogenetic tree as reference.

Table 2.4.2 shows that bacteria which are closely related to others in our frequency distribution database have a significantly better chance of being

Bacterium Name	Index Number	Order match (%)	Genus match (%)	Species match (%)	Closest match (%)	Closest match index
M. bovis	2.30.1.13.1.1	100	100	100	100	2.30.1.13.1.1
S. flexneri	2.28.3.27.2	100	100	100	100	2.28.3.27.2
C. pneumoniae	2.20.6.2.3	100	100	100	100	2.20.6.2.3
B. anthracis	2.30.7.12.4	100	100	98	98	2.30.7.12.4
B. cereus	2.30.7.12.4	100	99	95	95	2.30.7.12.4
V. vulnificus	2.28.3.23.9	100	100	99	93	2.28.3.23.9
B. bronchiseptica	2.28.2.8.4	100	77	–	77	2.28.2.14
B. parapertussis	2.28.2.8.4	100	73	–	73	2.28.2.14
L. johnsonii	2.30.7.17.3	99	99	0	0	2.30.7.17.6
R. palustris	2.28.1.6.12.5	99	92	91	78	2.28.1.6.12
B. pertussis	2.28.2.8.4	99	79	–	79	2.28.2.14
C. violaceum	2.28.2.1.4	99	39	0	0	2.28.2.1.9
M. gallisepticum	2.30.8.4.4	99	0	0	0	2.30.8.4.3
S. avermitilis	2.30.1.8.1.2	92	92	84	84	2.30.1.8.1.12
O. yellows	2.30.8.2.3	92	0	–	0	2.30.8.4.1
C. diphtheriae	2.30.1.13.2.10	91	91	91	82	2.30.1.13.2.8
N. europaea	2.28.2.4.6	85	0	–	0	2.28.2.1.9
M. avium	2.30.1.13.1.1	83	83	78	78	2.30.1.13.1.1
M. mycoides	2.30.8.3.1	83	5	–	1	2.30.8.4.1
P. luminescens	2.28.3.27.13.5	82	82	71	66	2.28.3.27.14.5
P. gingivalis	2.15.1.2.7	71	71	71	71	2.15.1.2.8
H. ducreyi	2.28.3.26.13	67	67	51	8	2.28.3.26.10
G. sulfurreducens	2.28.4.7.4	66	–	–	0	2.28.3.27.14.5
B. bacteriovorus	2.28.4.8	56	–	–	6	2.28.3.27.14.5
W. succinogenes	2.28.5.1.2	49	0	0	0	2.28.5.1.1
H. hepaticus	2.28.5.1.1	17	4	0	0	2.28.5.1.1
T. denticola	2.27.3.2.3	6	0	0	0	2.27.3.2.3
P. marinus	2.21.1.9	1	1	–	0	2.21.1.3
Synechococcus	2.21.1.9	1	1	–	0	2.21.1.3
Wolbachia	2.28.1.8.5.10	0	0	0	0	2.28.1.8.5.5
P. marinus	2.21.1.9	0	0	–	0	2.21.1.3
Pirellula	2.20.1.1	0	–	–	0	2.20.6.2.3
C. Blochmannia	2.15.4.3	0	–	–	0	2.15.1.2.8
G. violaceus	2.21.4	0	–	–	0	2.21.1.3
Average		64.97	53.45	58.33	37.27	

Table 4: Identifying unknown bacterial species according to ribosomal rRNA similarity, selected results (1000bp samples)

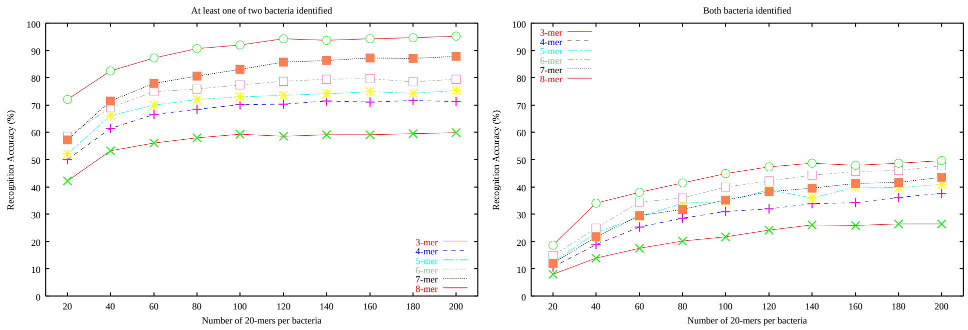
identified than distant ones. One interesting observation is the fact that specific microbial species are identified with high percentages as other ‘unrelated’ species, where a closer relative from the same class or even subclass may exist in the k-mer frequency distribution database. This would indicate that there are differences in the classification based on the 16S rRNA phylogenetic tree and one based on the genomic signatures alone.

2.5 Identifying Bacteria from Mixed Samples

More than just identify the members of a complex microbial community, we seek to assay their relative population frequency. We have shown that individual 20-mers identify the correct species only 8% of the time, using 8-mer frequencies, thus identifying the relative frequencies of bacteria in a mixed sample is a difficult task. An easier problem is the identification of a subset of species in the sample, especially the single most populous member of the sample.

For this purpose we constructed 20-mer tag data sets where half were derived from one bacteria and half from another. To identify the appropriate species, we cluster the 20-mers according to k-mer similarity, as follows: First we create for each 20-mer a vector of size 104, each position containing the conditional probability of the 20-mer being originated from the corresponding known bacteria genome. Then we cluster the 20-mers using *k-means* clustering into two clusters, according to the Euclidean distance of their corresponding vectors. We then classify the 20-mers of the two clusters separately, which gives us two candidate bacteria.

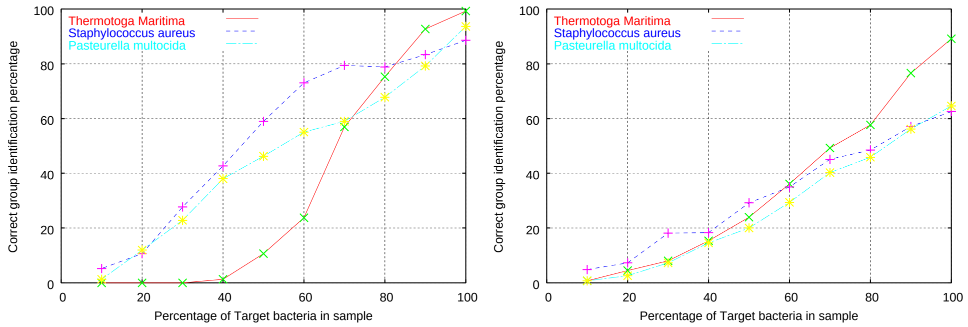
Figure 6 shows that we can identify both bacteria 50% of the time and one of the two 95% of the time, provided we sample a sufficient number of 20-mers from each bacteria.



(a) Recognition rates of ≥ 1 bacteria

(b) Recognition rates of both bacteria

Figure 6: Recognition accuracy of pairs of equi-probable bacteria, averaged over 500 different bacteria genome pairs.



(a) Using 2 k-means cluster

(b) Using 8 k-means clusters

Figure 7: Identifying bacteria from mixed sample containing percentage p of target bacteria, using 8-mer frequency distributions and variable cluster numbers

As a second experiment, we created samples where a specific percentage p is taken from a primary bacteria that we want to identify, where the rest of the sample is populated with 20-mers from randomly selected bacteria genomes. Then we try to identify the specified bacteria by creating a number of clusters and counting the total percentage of the identified clusters that matches the primary sampled bacteria.

Results for three different bacterial strains (*Thermotoga maritima, Pasteurella multocida* and *Staphylococcus aureus subsp. aureus Mu50*) are provided in Figure 7. These three bacteria were selected as random choices of a hard, medium and easy-to- recognize bacteria strains by their k-mer distribution frequencies. *T. maritima* is pretty distant to other bacteria found in our database, *P. multocida* frequency distribution resembles few other in our database and *S. aureus* has another three relative strains present, which are divided in two groups according to an 80-group clustering of the available genomic sequences.

In Figure 7, we can observe that recognition accuracy when a specific bacteria is comprising more than half of the sequence material in our sample is significant, especially when compared with an expected recognition percentage of 1.25% of a totally random sample. As expected, the recognition rates for *T. maritima* drop significantly faster than of the other representative samples, since having related strains in the database gives a larger space for recognition and *T. maritima* has a pretty distant k-mer frequency distribution. All three bacteria have a higher than 50% recognition percentage when they comprise more than 70% of the sample.

Comparing the results of clustering in 2 or 8 groups, we can see that 2-group clustering performs generally better, which is expected considering we are seeking to identify only one bacterial strain. The difference, though, diminishes (or even reverses, in the case of distant bacteria like *T. maritima*) when the bacteria comprises a smaller percentage of our sample. This can be explained by the fact that the specificity of the existing 20-mers of our target bacteria in the sample is absorbed by the noise of the other 20-mers in a larger group, where the target 20-mers could actually form smaller easier to identify groups (given enough 20-mers).

All majority-identifying experiments were performed 100 times for each bacteria to create data points in our graph, for 100 20-mers drawn randomly from the target and random other bacteria in our frequency distribution database, and averaged.

2.6 Methods

All experiments were performed on a set of 104 eubacterial genomes, except the recreation of the experiment conditions of [109], where 25 archaea eubacterial genomes were used.

2.6.1 Conditional Probability Classifier

For each classifying experiment, random sequences/fragments were drawn from each bacterial genome in order to identify its origin. As stated in Section 2, we say that a bacterial genome is identified when the Bayesian/conditional probability, calculated as the product of the individual k-mer statistical probabilities, is the top one (highest) among the 104 probabilities calculated for all the genomes. In the case of considering the probability being in the top m, we compare the origin genome of the samples with the top m ranking genomes from the probability calculations.

Every classification experiment, for every bacteria, is repeated ten times and the results are averaged.

For the experiments with mixed samples from two bacteria with equal

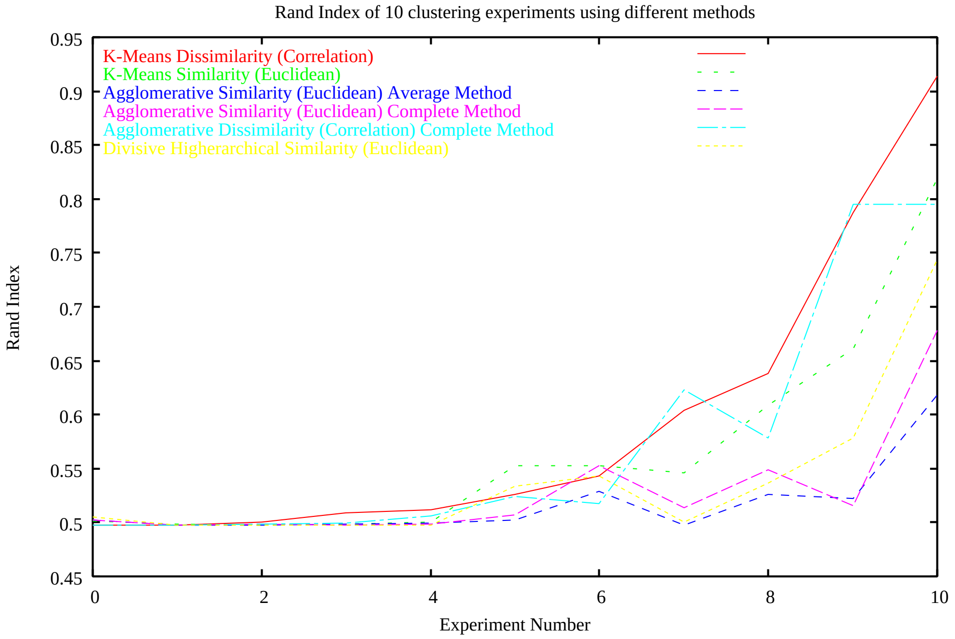


Figure 8: Six clustering methods comparison on 20-mer vector datasets based on 8-mer statistics

percentages, we used 500 different bacteria genome pairs and averaged the results. For the experiments where a primary bacteria was sampled comprising p percentage of the sample and the rest was populated by randomly selected pieces from other bacteria, each point represents the average of 100 repeats, each applied on random 20-mer samples of the bacterial pair.

2.6.2 Clustering and Grouping

The different clustering methods used are the agglomerative, divisive hierarchical and k-means, as implemented in the R-Project statistical package. Further information can be found in the *R-Project documentation* [100]. We should note that the k-means clustering method in R uses “methods” instead of “centroids”, which allows clustering experiments even when the distance between elements is not a metric, which is the case when comparing the correlation of vectors.

Since many of our statistical experiments depended on how well vectors can be clustered based on similarity/dissimilarity, we performed a number of comparisons of the different methods. In general, both k-means methods, dissimilarity with correlation distances and similarity with Euclidean distances, perform best, but knowing the number of clusters that we are seeking. Surprisingly, the divisive hierarchical method, using similarity vectors and Euclidean distances performs equally well with the k-means methods, without any information about the number of clusters. Finally, the agglomerative methods perform a bit weaker, especially when not using the complete method. A sample result of ten different pair instances of 20-mer vectors clustered with six different clustering methods can be seen in Figure 8. Here we are using 8-mer distribution frequencies to cluster 20-mers originated from two different bacteria, each of which comprises half or our sample of 100 total 20-mers. The comparatively measuring how well each method performs, we are using the Rand Index [101]. We should note that the minimum and average values of the Rand Index for the number of elements used is approximately 0.49 and 0.67 respectively.

2.7 Conclusions

Through computational experiments, we have demonstrated that the analysis of short DNA sequence reads or tags can be used to determine the composition of complex microbial communities. Such methods hold particular promise as inexpensive, high-throughput methods of producing short sequence reads become available. Unlike microarray-based techniques for population analysis, our approach appears capable of recognizing previously unsequenced species. We are now applying these techniques to the analysis of actual sequence data from samples of the poplar rhizosphere grown under different environmental conditions.

Chapter 3

Local Alignment Classification for Analyzing the Poplar Rhizosphere

One of the major contributing factors associated with climate change and global warming is the ever increasing concentration of atmospheric CO_2. Forests account for a large proportion of global net primary productivity (NPP), the rate at which new biomass accrues in an ecosystem [68]. Much research has focused on these ecosystems as a component of the terrestrial carbon sink and their potential to mitigate the effects of this greenhouse gas. Though no widely accepted model exists accounting for subsequent plant responses, elevated atmospheric CO_2 has been documented to increase the carboxylation efficiency of Rubisco [17] resulting in enhanced plant growth [27], greater fine root production [56] and augmentation of soil carbon allocation via secretion of root exudates from the root tips and increased turnover of fine roots [136, 54].

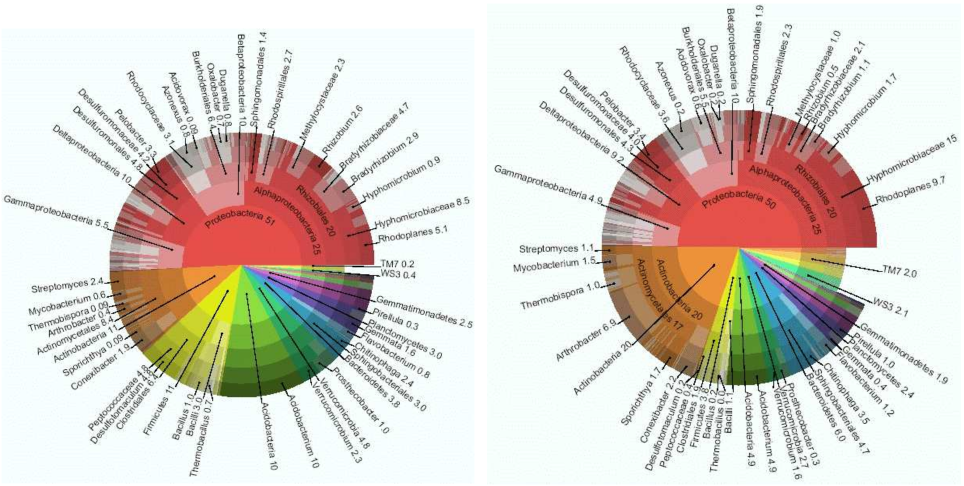
This chapter is drawn from our papers [72] and [95]. My contributions in this work are limited to the development and implementation of the sequence classification and analysis methods.

These processes result in a concomitant increase in soil microbial respiration and carbon turnover [50]. There is no consensus on many of the secondary effects associated with these plant responses. Therefore their importance in regulating the terrestrial carbon sink still remains to be determined.

Here, we present the tools facilitating an in-depth analysis of the microbial community composition of trembling aspen (*Populus tremuloides*), grown at the Rhinelander WI free-air CO_2 and O_3 enrichment (FACE) experiment, and how it is affected by plant responses to elevated CO_2. Initial studies on our soil core samples for both ambient or elevated atmospheric CO_2 showed that total fungal biomass, community composition and metabolism did not significantly change between each of the triplicate FACE plots for each experiment belonging to either ambient or elevated CO_2 [23]. However, significant changes in enzymatic activities were noted between ambient and elevated CO_2 treatments, indicating changes in microbial community composition or increased activity under conditions of elevated CO_2. In order to obtain the most complete global profile of the microbial community, we conducted an in-depth community analysis on composites of these previously characterized soils.

Previous soil population studies on *Bacteria* and *Eukarya* at the domain and phylum levels showed that total microbial abundance does not significantly change under elevated CO_2 at the Rhinelander FACE site [137, 23], which was confirmed with q-PCR. In order to address changes in microbial diversity (detection and frequency of operational taxonomic units (OTUs) and microbial richness (total number of different OTUs), a total of 5061 16S (prokaryotic and archaeal) and 1935 18S (eukaryotic) ribosomal rDNA clones (Table 5) were generated from total soil DNA extractions obtained from trembling aspen under ambient and elevated (560 ppm) CO_2 concentrations.

A general overview of the bacterial community compositions is outlined



(a) Classification assignment of prokaryotic sequences under ambient CO_2 conditions

(b) Classification assignment of prokaryotic sequences under elevated CO_2 conditions

Figure 9: Taxonomic breakdown of classified 16S rDNA sequences in prokaryotic populations under ambient (a) and elevated (b) atmospheric CO_2. Central pie shows percentages by phyla; each outer annulus progressively breaks these down by finer taxonomic levels: class, order, family and genus in the outermost annulus. Numbers indicate the relative abundance, expressed as a percentage, of the different taxonomic groups.

in Figures 9a-b. Complete comparisons of community composition for all domains can be found at http://genome.bnl.gov/FACE/. The microbial abundance of many taxonomic groups remained unchanged. These were predominantly affiliated with γ- and δ- Proteobacteria, and provide an additional internal standard further validating the comparison of these composite samples to determine microbial community composition.

In the rest of this chapter we will detail the bioinformatic methods enabling a robust analysis of the diversity and identification of population changes in the microbial community composition under conditions of ambient/elevated CO_2.

3.1 Genomic Sequence Classification

Metagenome shotgun sequencing and analysis is revolutionizing the field of molecular ecology, revealing a more complete vision of the biodiversity and functions within ecosystems [4, 8, 9, 10, 13, 49, 121, 123, 128]. Preliminary to constructing and sequencing metagenome libraries, it is necessary to estimate the complexity and richness of the microbial community to be examined. This can be adequately done by constructing and sequencing ribosomal RNA (rRNA) gene libraries [21]. Thus, the means to accurately identify and classify large datasets of rRNA genes and to determine community richness is an essential step in metagenome sequencing projects. The increasing library sizes to cover the metagenomes of complex microbial communities and the subsequent amount of screening and analysis needed necessitates the development and application of adequate and robust bioinformatics tools for community composition analysis that can handle large rRNA gene data sets.

The following tools are available for analyzing microbial community composition:

Phylogenetic trees are useful in determining novel groups of organisms, but are limited in their reliability for large datasets. Bootstrap values become expensive to calculate and tree topologies become unreliable as tree size increases [88].

Oligonucleotide (k-mer) based classifiers provide fast and reliable techniques which exploit the fact that closely related sequences share common small subsequences and that organisms and regions have their own distinct signatures, a term to describe distributions of oligonucleotides [30, 61, 96, 109]. A disadvantage is the lack of positional information of oligonucleotides, which becomes more problematic with decreasing k (oligonucleotide size), in which case the probability of finding oligos by chance increases exponentially. Adding location information to oligonucleotides leads to computation time increases, while limiting its generality. In addition, altering a single base in a sequence results in k different k-mers, and random mutations can therefore have devastating results on the oligonucleotide distribution. Correcting this problem by approximate matching for k-mers leads to time increases exponential in the number of errors allowed.

At present, the most widely used tool for species identification based on ribosomal RNA gene sequences is the ribosomal database project (RDP) Naïve Bayesian classifier [25]. This tool uses 8-mers to cluster and classify 16S rRNA sequences based upon vetted sequences with well assigned taxonomy (http://rdp.cme.msu.edu/index.jsp, James Cole, private communication, [43]).

Sequence alignment classifiers extract differences and calculate distances between DNA sequences. They can be used to identify the closest match of a sequence to a vetted reference data set. The computation time required for aligning sequences increases quadratically to the length of the sequences, which significantly delays the analysis of large datasets, e.g. rRNA gene sequences representing a complex microbial community.

The program BLAST [1] performs approximate sequence alignments, finding locally very similar pieces as opposed to globally calculating the best way to convert one sequence to the other. BLAST can be calibrated to achieve a desired speed/sensitivity ratio. Local alignment provides flexibility in handling sequencing errors, incorrectly inserted or omitted prefixes, suffixes and subsequences, as well as ambiguous characters. A further advantage of BLAST is the ability to compare sequences, one at a time, against a database of vetted sequences. Classification using alignment is highly parallelizable, with great promise with multi-processor and future multi-core systems.

In the next section we describe the use of local alignment classification as a robust tool in grouping closely related sequences, while maintaining equivalent or slightly better classification accuracy as compared to the RDP naïve Bayesian classifier.

3.1.1 RDP Naïve Bayesian Classification and Initial Drawbacks

We initially tried to use the Ribosomal Database Project (RDP) Naïve Bayesian classifier to analyze the 16S rRNA gene sequence data from a large scale sequencing project, which aimed at determining changes in the soil microbial community composition of trembling aspen when three were exposed to ambient (360 ppm) or elevated (560 ppm) atmospheric CO_2. Upon initially analyzing 2774 16S rRNA gene sequences it became apparent that a large number of them classified in the same genus had edit distances accounting for >20% of the total sequence length, and sometimes up to 50%! Many of these occurred between sequences classified with low confidence estimates (<50%), creating uncertainty for a large number of classification groupings. This provided the first clue that many of the sequences we were trying to classify were distant from all the vetted sequences available, possibly representing new phylogenetic groups. We also noticed that some sequences with >99% edit distance similarity were found classified in different taxa, with a couple of occurrences even at the phylum level. These findings bring to light the fact that a perfect classification of rRNA gene sequences can currently not be achieved, and that errors will be found even when classifying unknown sequences to closely related characterized ones. These results prompted us to explore sequence identification methods alternative to the RDP Naïve Bayesian classifier.

3.1.2 Refinement of Vetted Sequence Classification using BLAST

The BLAST utility bl2seq [118], which performs pairwise sequence alignment, was used as an alternative to maximize 16S rRNA gene classification accuracy. Several key parameters (match, mismatch, gap_start, gap_extend) were adjusted to minimize misclassifications. To do so, a genus level leave-one-out test (see methods) for each of the 5574 vetted sequences (RDP data set) was performed to determine the parameter set that optimally separates the scores of closely related species from distant ones.

After examining the different parameter sets we determined the optimal set of key parameters (*match* = 1*, mismatch* = 5*, gap_open* = 3, and *gap_extend* = 2.5), where the *match* is positive and considered a reward, and *mismatch*, *gap_open* and *gap_extend* are negative and are considered penalties. Compared to the default bl2seq parameter values, this set of parameters reduced the number of misclassified vetted sequences from 284 to 268 out of a total of 5246 (328 sequences are unique in their genus in the vetted set).

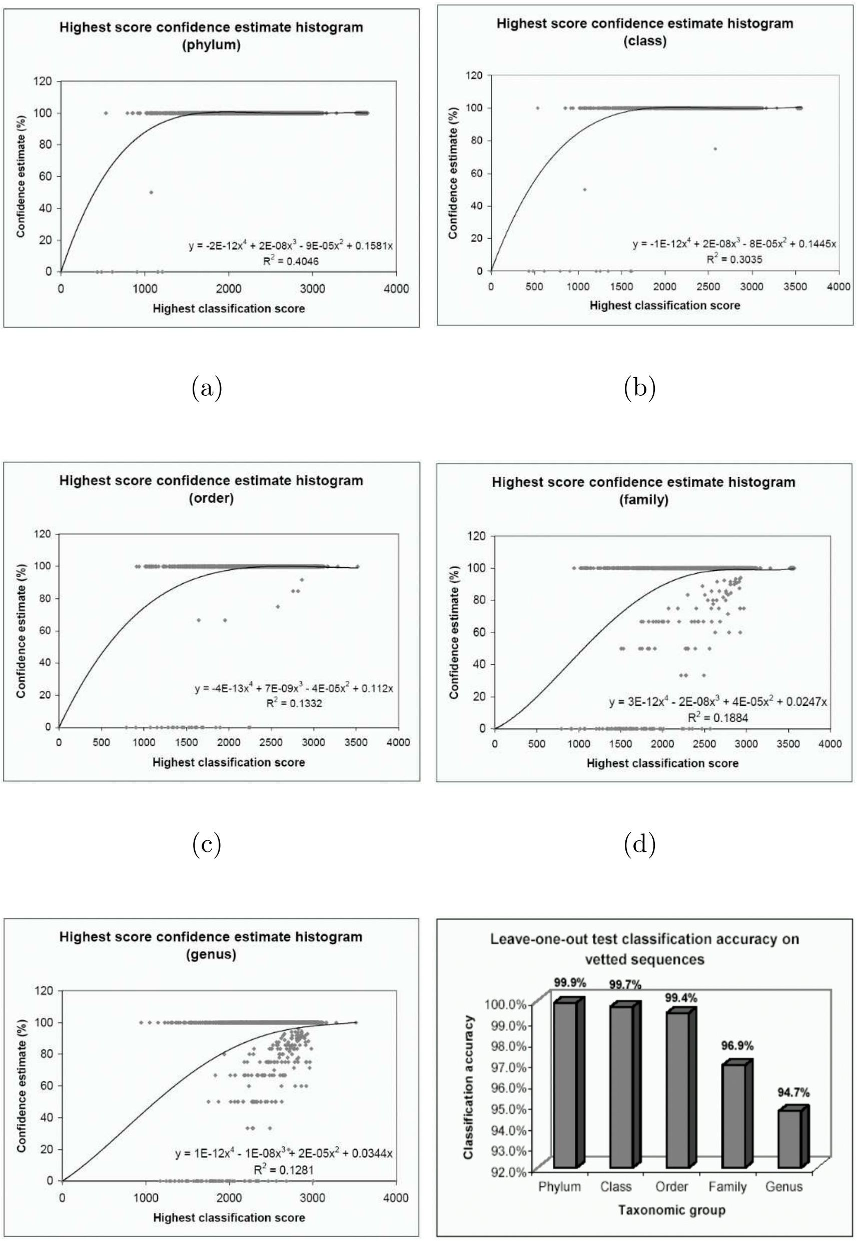
3.2 Classifier Confidence Levels

We created a confidence estimate for each bl2seq alignment score using the value of the highest scoring pair-wise alignment for each sequence, setting the Boolean value for correct (1) or incorrect (0) classification at the genus, family, order, class or phylum level. This gives us the ability to assign a confidence estimate to each specific score value, according to the number of times an alignment with such a value resulted in the correct phylogenetic classification. A rating for this score was then based on these results. Fourth degree polynomial regression curves were used to determine the relation between classification scores and confidence estimate values for the different phylogenetic levels.

These curves (Figures 10a-e) were, smoothed by the addition of extra points, represent high confidence estimates at very high scores and zero confidence estimates at very low scores. The confidence estimate of each score is calculated from the value of the polynomial for this specific score. From this analysis we can conclude that confidence estimates decrease when classifying the species at lower phylogenetic levels (from phylum to genus). Figure 10f demonstrates that we still obtain a 94% classification accuracy at the genus level, when using the optimal parameter set and the bl2seq utility, and increasing accuracies for higher phylogenetic levels.

3.2.1 Exploring the Alignment Parameter Set

We subsequently constructed a vetted sequences database in BLAST format and used the *blastn* utility, like previously done for bl2seq, in a leave-one-out test on this database. The faster *blastn* processing of sequences, compared to bl2seq, was used to explore all possible combinations of key parameters. In order to identify the best set of parameters, we performed a full coverage scan at value increments of 5, fixing the reward value of match to 10, and allowing for all possible combinations of the three other parameters (*mismatch, gap_open and gap_extend*). Restrictions in the *blastn* utility did not allow ratios of mismatch/match lower than 1 (except the ratio 8/10) and higher than 5. Also the mismatch/match ratio of 9/2 was not permitted. We observed that when using *blastn*, the values for *gap_start* and *gap_extend* did not alter our results, this in contrast to their influence on scores using the *bl2seq* utility. At the end, the (*match = 10, mismatch = 50*) assignment gave the best classification accuracy, with percentages for the different phylogenetic levels shown in Table 5, together with the published accuracy results of the RDP naïve Bayesian classifier [130].



(a)

(b)

(c)

(d)

(e)

(f)

Figure 10: Fourth degree smoothed polynomial regression curves and classification score confidence estimates for different phylogenetic levels (phylum, class, order, family, genus). The confidence estimate of each score is calculated from the value of the polynomial for this specific score.

	Phylum	Class	Order	Family	Genus
RDP naïve Bayesian classifier	**99.9**	**99.9**	99.3	**97.1**	94.3
blastn classifier	**99.9**	99.7	**99.4**	**97.1**	**94.9**
bl2seq classifier	**99.9**	99.7	**99.4**	96.9	94.7

Table 5: Classification accuracy of RDP, blastn and bl2seq classifiers. The numbers represent percentages of sequences correctly classified in their known phylogenetic levels in leave-one-out tests.

3.2.2 Improved Grouping of Closely Related Sequences using the Blastn Classifier

Although the accuracy results presented in Table 5 do not differ significantly for the different methods tested, our *blastn* classifier groups closely related sequences with increased accuracy. To demonstrate this, Levenshtein pair wise edit distances [73] were calculated for our set of 2774 16S rRNA gene sequences. Using the complete linkage method these were subsequently clustered into groups, in which the percentage difference among sequences is below a cut-off value.

Taking as an example a 1% cut-off sequence difference (this percentage is calculated proportional to the sequence length, which for the 16S bacterial rRNA gene is approximately 15), we would expect all group members to belong to the same phylogenetic group, ranging from genus to phylum, since at the 1% dissimilarity level even species are expected to cluster together. Considering the high identification percentages for both the RDP classifier and the *blastn* classifier, we counted the number of groups that were heterologous at a given phylogenetic level (e.g. contained members of more than one phylotype). A few indicative results are the following: (i) For the ambient CO_2 community 16S rRNA sequence set, which comprised of 132 groups with more than one element, 23 groups had elements classified in different genera by the

RDP classifier, breaking them into 52 subgroups. The blastn classifier divided only 15 groups into 30 subgroups. In 14 out of the 15 groups this was due to the presence of a single misclassified sequence. The biggest group of 32 elements was identified as 32 *Desulfotomaculum* by our classifier, where the RDP broke it into 6 distinct groups, partitioned as: 4 *Thermodesulfovibrio*, 1 *Succiniclasticum*, 15 *Gelria*, 1 *Propionispora*, 3 *Thermovenabulum* and 8 *Pelotomaculum.* (ii) For the elevated CO_2 community 16S rRNA sequence set, 28 groups out of a total of 87 groups were divided into 67 subgroups by the RDP classifier, where the blastn classifier only divided 10 groups into 20 subgroups. These examples show that the blastn classifier reduces classification ambiguities compared to the RDP classifier.

3.3 Significance of Population Differences

Probability values measuring significance in phylotype population changes are calculated as follows: Considering the null hypothesis that there is no difference between population proportions, we test it by calculating the p-values derived from the standardized variable $z = \frac{P_1 - P_2}{\sigma_{P_1 - P_2}}$, where $\sigma_{P_1 - P_2} = \sqrt{pq(\frac{1}{N_1} + \frac{1}{N_2})}$ is the standard deviation of the sampling distribution of differences in proportions, which is approximately normally distributed. P_1 and P_2 are the phylotype proportions in the ambient (N_1) and elevated (N_2) CO_2 sample sizes respectively. $p = \frac{N_1 P_1 + N_2 P_2}{N_1 + N_2}$ is an estimate of the common population proportion under the null hypothesis and $q = 1 - p$ [49]. To derive one-tail test p-values from the standardized variable z values, we use the Ibbetson polynomial approximation [50], as adapted by John Walker from C implementations written by Gary Perlman of Wang Institute, Tyngsboro, MA, found at http://www.fourmilab.ch/rpkp/experiments/analysis/zCalc.html and available in the public domain.

Probability values measuring significance in phylotype population changes can also be calculated using the one side Fisher exact test. We test against the null hypothesis that there is no difference between population proportions. Differences in proportions follow the multinomial distribution and the Fisher test combinatorially calculates exactly the probability of a difference with no approximations. This test, although not an approximation, such as the z test mentioned above, is computationally expensive and involves large integers, even for small group sizes.

To account for the number of simultaneous statistical tests being performed to calculate p-values of the differences in the populations, we adjust the alpha value (significance threshold) by applying the Bonferroni correction. Thus we divide the alpha value by the number of tests at each phylogenetic level, which equals the number of groups compared. The divisors for the 16S populations are: 20 for the phylum level, 27 for the class level, 52 for the order level, 120 for the family level and 273 for the genus level. The seven taxonomic levels for the 18S sequences have the following divisors: 11, 25, 39, 40, 45, 47 and 42. As an example, there were 35 sequences identified as members of the Bacilli class in the ambient CO_2 population samples, where only 13 in the elevated CO_2 samples. The one-tail Fisher exact test gives a probability value of 0.001463 against the hypothesis that the Bacilli populations are the same. Since there are 27 classes for which we calculate p-values, adjusting an alpha value of 0.01 with the Bonferroni correction will result in a new alpha value of $0.01/27 = 0.000370$, according to which the Bacilli members did not change significantly as a proportion of the sampled populations. Against an initial alpha value of 0.05, the adjusted value of $0.05/27 = 0.001851$ suggests there is a significant change of the Bacilli population.

3.4 Population Richness

Rather than using the phylogenetic grouping determined from the RDP vetted sequences, the Chao1 non-parametric estimator was used to determine phylotype richness [18, 19] and calculated on groups clustered according to specific Levenshtein edit distance values [73]. In addition to single linkage clustering commonly used for determining phylotypes [111], we also used average and complete linkage methods which are more appropriate to larger groupings (Table 6).

Population equitability was calculated using the Simpson evenness index E_D [112, 3, 55], defined as the reciprocal Simpson index D over the maximum number of phylotypes observed D_{max}:

$E_D = \frac{D}{D_{max}}$, where $D = \frac{1}{\sum_{i=1}^{S} p_i^2}$

Here, p_i is the proportion of the population constructed from the ith phylotype and S the total number of phylotypes.

Kemp and Aller [67] argue that the amount of sampling required to detect all phylotypes and reach asymptotic values of the Chao1 index (therefore significantly reducing the probability that further sampling will discover novel phylotypes) correlates well with evenness.

Low evenness values (<0.4) are indicative of under-sampling (by a factor of 8 or more), relative to the Chao1 index.

3.5 Methods

All computational experiments were performed on either a hyper threaded Pentium 4 at 3.2GHz desktop with 2GB of memory or dual Xeon at 2.8GHz server with 4GB of memory. Since most classification computations are parallelizable, a cluster of 4-10 desktops has been used for reducing computational time. All time references referring to computation will assume use of one CPU

Data Set		Ambient									Elevated								
Clustering Method		Single			Average			Complete			Single			Average			Complete		
Similarity		99%	97%	95%	99%	97%	95%	99%	97%	95%	99%	97%	95%	99%	97%	95%	99%	97%	95%
Archaea	Observed Phylotypes	448	208	119	518	274	175	565	330	220	257	100	53	334	135	81	391	176	109
	Chao1 Index	4069	1664	2099	3221	1430	845	2930	1146	758	1847	342	284	1635	350	211	1365	364	201
	Chao1 Index SD	873	491	1205	527	305	221	427	180	155	480	94	136	309	70	56	200	55	35
	Simpson Index	0.05	0.03	0.02	0.08	0.06	0.04	0.13	0.09	0.09	0.03	0.02	0.03	0.05	0.05	0.05	0.10	0.08	0.08
Bacteria	Observed Phylotypes	882	811	696	884	819	730	887	819	745	839	803	738	841	805	752	845	807	762
	Chao1 Index	3,989	3,042	2,281	3,966	3,001	2,197	3,985	2,927	2,163	5,888	4,854	3,858	5,917	4,867	3,745	5,975	4,893	3,585
	Chao1 Index SD	400.7	292.6	217.9	395.9	282.1	193.6	397.8	268.8	184.0	771.9	605.3	468.8	775.8	606.9	436.9	783.6	610.2	402.1
	Simpson Index	0.39	0.33	0.25	0.40	0.39	0.31	0.43	0.40	0.38	0.24	0.21	0.20	0.24	0.24	0.21	0.24	0.24	0.24
Eukaryota	Observed Phylotypes	379	302	206	387	322	243	392	330	267	358	275	216	365	299	235	373	307	259
	Chao1 Index	774	576	390	771	579	389	786	583	414	914	685	445	911	686	480	902	684	496
	Chao1 Index SD	73.7	58.4	49.0	71.0	53.2	35.8	72.6	51.8	34.9	111.1	96.9	60.0	107.0	86.8	61.6	101.9	83.4	57.7
	Simpson Index	0.31	0.21	0.14	0.34	0.31	0.20	0.35	0.32	0.32	0.19	0.13	0.12	0.20	0.18	0.13	0.24	0.20	0.20
Fungi	Observed Phylotypes	88	60	35	89	68	47	91	70	54	54	39	31	57	46	33	59	47	41
	Chao1 Index	345	187	323	346	148	112	366	139	110	142	100	97	138	106	86	140	111	86
	Chao1 Index SD	118.8	66.7	311.5	118.8	36.9	37.2	126.1	31.5	30.9	45.3	35.0	43.8	40.4	32.8	33.2	40.4	34.7	25.5
	Simpson Index	0.41	0.25	0.14	0.41	0.36	0.19	0.41	0.39	0.36	0.18	0.11	0.10	0.18	0.16	0.12	0.26	0.16	0.14

Table 6: Phylotype richness calculated for the archaeal, bacterial, eukaryotic and fungal populations.

desktop with at least 2GB of memory.

All programs/scripts performing computations and statistical evaluations were written in perl, except the edit distance calculations, which were written in C, in order to decrease computational time.

3.5.1 16S rRNA Sequence Classification

Classification accuracy was measured by performing leave-one-out tests of the 5574 vetted sequences against themselves. A subset of vetted sequences is created, excluding the singletons, which are sequences that belong to a phylogenetic group with only one member. Each sequence of this new dataset is then separated from the dataset and classified against it. The number of correctly classified sequences is then divided by the total number of sequences present in this subset (with singletons being excluded) to produce the final accuracy percentages.

Classification with the RDP naïve Bayesian classifier takes approximately 1.5 minutes, when submitted through the web, to produce a complete 1000 sequence assignment with confidence estimates, as calculated by bootstrap trials. The *blastn* classifier requires, for the same number of sequences and against the same database of vetted sequences, approximately 35 minutes. Using *bl2seq*, for performing pair wise comparisons, requires significantly more time, in the range of days. Increasing the accuracy of BLAST by lowering the values of the word size W and the X *drop-off* parameters results in a significant increase in computation, while increasing the classification accuracy at the genus level by no more than 0.1%.

The naïve Bayesian classifier principles are described in [109]. The RDP classifier uses oligonucleotides of size 8 and randomizes the selection of oligonucleotides to be used for the confidence calculation. In each of the 100 bootstrap trials, 1/8 of all possible 8-mers of a sequence are selected randomly.

The BLAST parameters adjusted for *bl2seq* and *blastn* in order to minimize misclassifications were:

1. *match*: The reward for a matched character, common to both sequences compared.
2. *mismatch*: The penalty for a character substitution
3. *gap_start*: The penalty for initiating a gap.
4. *gap_extend*: The penalty for extending an initiated gap by one character.

These values can be scaled proportionately without affecting the alignment, but only the score, although the relative scores under the same parameter set remain proportional.

The score for each BLAST alignment, used to determine confidence values, was calculated by summing up the individual scores of the locally aligned pieces, which is already normalized against the length of the sequences being compared. The two tests to measure misclassifications and calibrate local alignment parameter space were (i) the number of rRNA sequences that score better against a sequence from another genus than against all of the sequences in their genus and (ii) the number of ribosomal sequences that score better against a sequence from another genus than at least one sequence from their own genus. The first test was used predominantly, since the final classification decision is based on the top scoring vetted sequence and can lead to a misclassification only if the test sequence aligns with a higher score against a foreign-genus vetted sequence.

For scoring an alignment, we used the sum of scores of individual local aligned pieces, which are not overlapping. A query sequence is assigned the phylogenetic lineage of the highest matching score.

3.6 Conclusions

The long-term sustainability of ecosystem productivity requires detailed knowledge of its biodiversity coupled to profound understanding in its function. Furthermore, to better understand the implications that elevated atmospheric CO_2 has on microbial communities, we provide the first detailed analysis profiling changes in specific groups of microbes to specific soil processes.

Our results show that microbial communities appear to be altered by elevated atmospheric CO_2 and that these changes may have implications for ecosystem function, especially via effects on the cycling of essential elements. Future investigations should shed more light on how elevated atmospheric CO_2 affects the diversity of life, the complexity and functioning of microbial communities in soil, the cycling of essential elements, and may further facilitate the prediction of such environmental impacts providing the key for their future correction.

We developed a *blastn* classifier with optimal key parameter set that performs better than the RDP II classifier for 16S rRNA based identification, especially when it comes to grouping of related sequences, thus reducing classification ambiguities. However, every classifier has a closed architecture and will assign every sequence to one in its dataset. The view of the biodiversity contained within a sample is therefore subject to the biases incurred by the limited number of sequences contained within the vetted sequence database, against which we classify.

In conclusion, our classifier has been proven to provide consistent and robust analysis. Further improvements could be realized in both accuracy and speed, especially through the contributions of advances in parallel and core architectures. These developments should enhance significantly the utility of database search and taxonomic annotation methods to the molecular biologist.

Chapter 4

Richness Estimation

4.1 Introduction

Two important factors that describe a microbial community are richness, meaning the number of species present, and diversity, which is their relative abundance [93]. The latter can be estimated from the classification efforts and/or phylogenetic analysis of community samples. Richness estimation requires information on the number of distinct subpopulations present in the community, according to a threshold set to determine them (e.g. genus level), as well as the evenness information, meaning how different the sizes of the subpopulations in the community are. Several richness estimator methodologies have been developed including extrapolation from accumulation curves, parametric estimators and non-parametric estimators, the latter being the most promising for microbial studies [55]. Among this last class of estimators, Chao1 [18] seems to be the most suited method for estimating phylotype richness from prokaryotic 16S rRNA libraries [67].

Here we detail our bioinformatics methods for analyzing population distribution and richness in large and diverse microbial communities. This was

This chapter is drawn from our paper [95].

achieved via a comparison of different clustering methods for achieving more accurate richness estimations. Our methodology, which we developed using the RDP vetted 16S rRNA gene sequence dataset, was validated against a large 16S rRNA gene dataset of approximately 2300 sequences, obtained from a soil microbial community study. We concluded that the best approach to group closely related sequences is by using complete linkage clustering, in order to calculate richness and evenness indices for the communities.

In the rest of this chapter we will argue against the single linkage clustering methodology for richness estimation calculations, in favor of the equally computationally-attractive complete linkage method.

4.2 Clustering Methodology for Richness Estimation

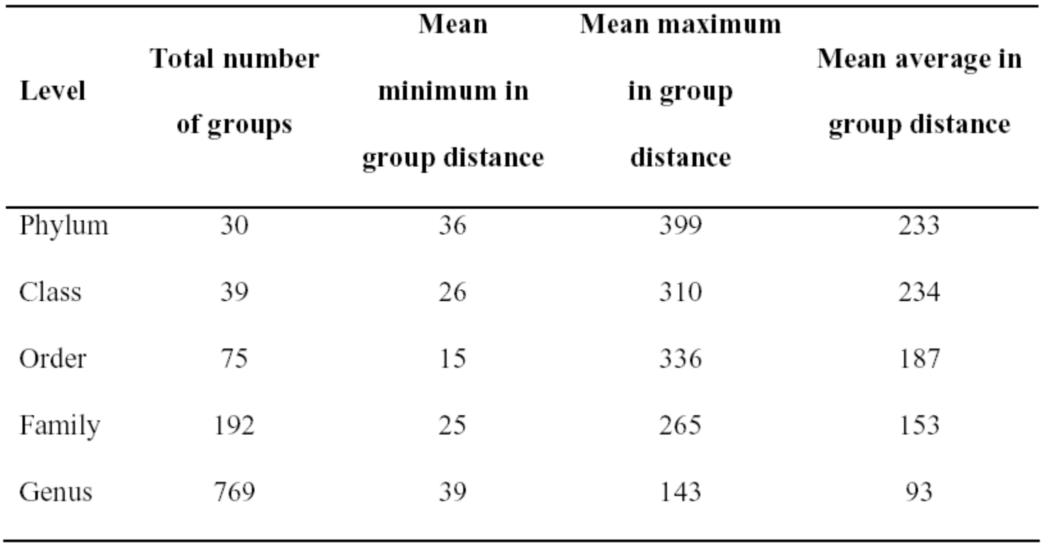
To deconvolute community composition, it is necessary to calculate the richness of a microbial population, meaning the number of phylotypes present. This requires partitioning the sampled sequences into sets according to their similarity. This can theoretically be achieved by using the output of our classifier, as presented in the previous chapter, where information is known for the identification of all sequences at different phylogenetic levels. However, this would require that all sequences are identified with the same confidence level, which is not always the case. In addition, highly dissimilar sequences can sometimes be classified in the same phylogenetic group when they have the same vetted sequence as their closest neighbor.

We examined three traditionally used clustering methods, the single linkage, average linkage and complete linkage methods, all which fall under the agglomerative hierarchical (bottom-up) approach and produce clustering trees [65]. All hierarchical clustering methods treat each data point as a singleton cluster, and then successively merge clusters until all points have been merged into a single remaining cluster. In single linkage hierarchical clustering, two clusters are merged in each step, whose two closest members have the smallest distance. In complete linkage clustering, we merge in each step the two clusters whose merger has the smallest diameter. In average linkage clustering, the two clusters merged in each step have a minimum average distance between their members.

We compared the three methods using the set of 5574 vetted sequences, for which phylogenetic information for all phylotypes is well assigned. Before applying the clustering methods, we used the known phylogenetic partitioning of the vetted sequences to calculate statistics about the number of groups they form at each phylogenetic level, as well as the minimum, average and maximum Levenshtein edit distances between sequences in these phylogenetic groups. The means of all these values are shown in Table 7.

The known phylogenetic partitioning of 5574 vetted sequences was used to calculate statistics about the number of groups they form at each phylogenetic level, as well as the minimum, average and maximum Levenshtein edit distances between sequences in these phylogenetic groups.

According to the number of groups in each phylotype (see Table 7), we determined the necessary cut off edit distance value, which, when applied to the inferred clustering tree, would produce the same number of phylogenetic groups. This is demonstrated in Figure 11 on a random subset of 100 vetted 16S rRNA sequences. In this figure, for example, we can observe that a cut-off edit distance value of 300 will result in the formation of 23 groups, for the given 100 sequences. Inversely, if we want to acquire 15 groups, a cutoff edit distance of 380 is required. Knowing the number of distinct groups for all taxa for our vetted sequence set allows us to determine cutoff levels that will generate the same number of groups, when clustering these sequences with



Level	Total number of groups	Mean minimum in group distance	Mean maximum in group distance	Mean average in group distance
Phylum	30	36	399	233
Class	39	26	310	234
Order	75	15	336	187
Family	192	25	265	153
Genus	769	39	143	93

Table 7: Phylogenetic partitioning of the vetted sequences in groups and their statistics. The known phylogenetic partitioning of 5574 vetted sequences was used to calculate statistics about the number of groups they form at each phylogenetic level, as well as the minimum, average and maximum Levenshtein edit distances between sequences in these phylogenetic groups.

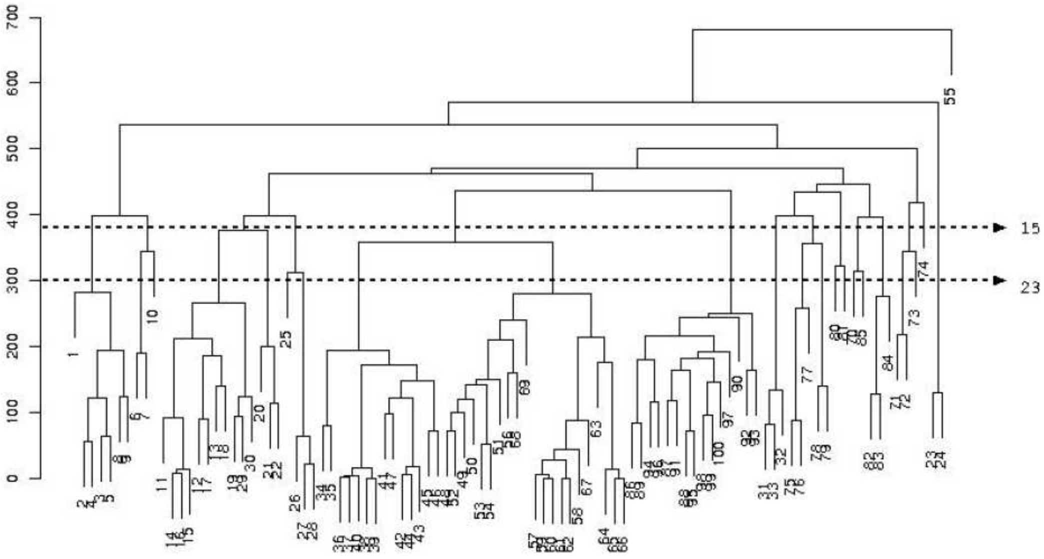


Figure 11: Complete linkage clustering of a subset of 100 vetted 16S rRNA sequences (for demonstration purposes). For two different cut off edit distance values of 300 and 380, the set is partitioned into 23 and 15 groups, respectively.

our three hierarchical clustering methods. This allows the evaluation of the clustering methods independently of the error in the cut-off estimation, which is actually a separate problem for all clustering and partitioning methods, and usually is calculated based on observations [55].

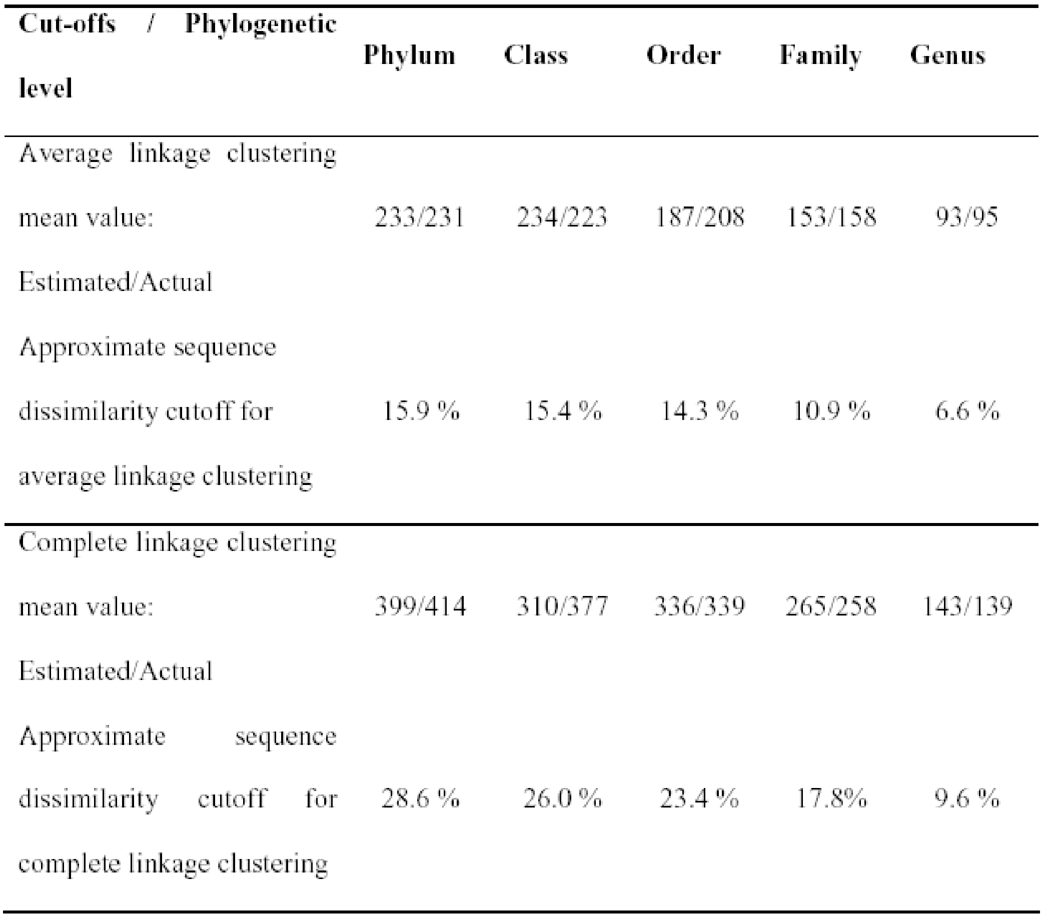
Correct cutoff values, as shown in Figure 11, cannot be calculated directly from vetted sequence statistics. To illustrate this point, one would expect that for the complete linkage clustering method, the correct threshold could be determined by calculating the maximum in-group distance, when the number of groups formed is the same as in the vetted sequence set at some phylogenetic level. It happens though that, even at the genus level, a group exists (Clostridia) in the vetted sequence set with a maximum in-group distance of 626, which indicates approximately 43% sequence dissimilarity.

More appropriate thresholds can be determined by considering the mean of the average distances inside the groups at every merging step of the average linkage clustering hierarchical algorithm, and then comparing this value to the known mean for the corresponding groups of the vetted sequences. These values are quite similar, as can be seen in Table 8. The same effect is observed for the complete linkage clustering method (Table 8), where the threshold value for partitioning is determined based on the maximum distance inside each group. Single linkage clustering does not offer such a measure for estimating a cut-off, since there is no averaging process in the algorithm.

By sorting the groups, formed at the different phylogenetic levels by using the different clustering methods, according to their cardinality, and by comparing this to the known phylogeny, we created graphs showing the trends in group sizes. The known phylogeny group cardinalities were the best approximated by the complete, followed by the average clustering method. This is shown in Figure 12 for the 75 groups at the order level, where similar figures were produced for all taxonomic levels.

To quantify the better performance of the complete clustering method, as observed in Figure 12, we calculated the Pearson correlation (difference in variance) and square difference (distance of each individual group of the same index, according to the corresponding sorted position). The results are presented in Table 9 and confirm that the complete linkage clustering method provides a better correlation to the known classification at all five phylogenetic levels.

As a case study for the richness estimation based on different clustering methods, we calculated the Chao1 index at different phylogenetic levels. Since the Chao1 estimate is based on the ratio of singletons and doubletons of a sequence grouping, it can vary significantly with changes to these small integers. For that reason, we tested the Chao1 richness estimations on 1000 random selected sequences from the vetted sequence set, repeating the test a total of 1000 times. In Table 10 we present the average richness estimations of these



Cut-offs / Phylogenetic level	Phylum	Class	Order	Family	Genus
Average linkage clustering mean value: Estimated/Actual	233/231	234/223	187/208	153/158	93/95
Approximate sequence dissimilarity cutoff for average linkage clustering	15.9 %	15.4 %	14.3 %	10.9 %	6.6 %
Complete linkage clustering mean value: Estimated/Actual	399/414	310/377	336/339	265/258	143/139
Approximate sequence dissimilarity cutoff for complete linkage clustering	28.6 %	26.0 %	23.4 %	17.8%	9.6 %

Table 8: Estimated cut-off values based on the vetted sequence statistics and actual cut-off values for different phylogenetic levels determined using the average and complete linkage clustering methods. The estimated cut-off values are the means of the vetted sequence statistics for in- group average and maximum sequence distances (See Table 7). Sequence dissimilarity cutoffs are presented as edit distance over average 16S sequence length percentages.

experiments. As seen in Table 10, richness estimations based on groups clustered with the complete linkage method are the most accurate. In conjunction with the consistently better correlation of the group size histograms, complete linkage clustering is preferable for use in richness estimation analysis.

We should note here that computation times for the single, average and

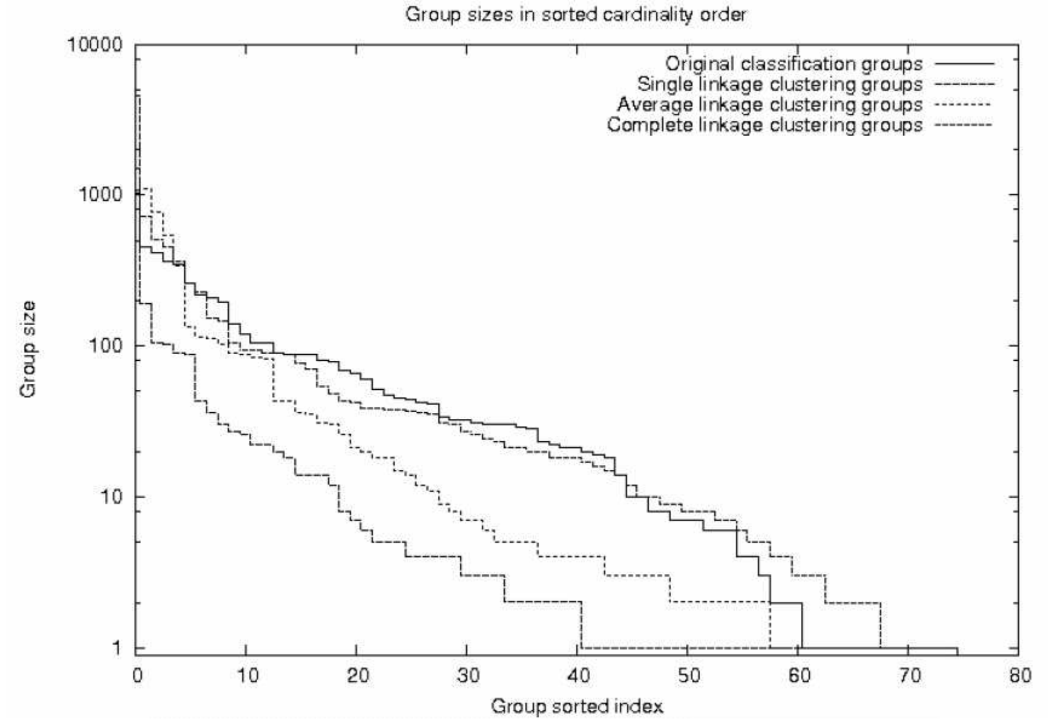
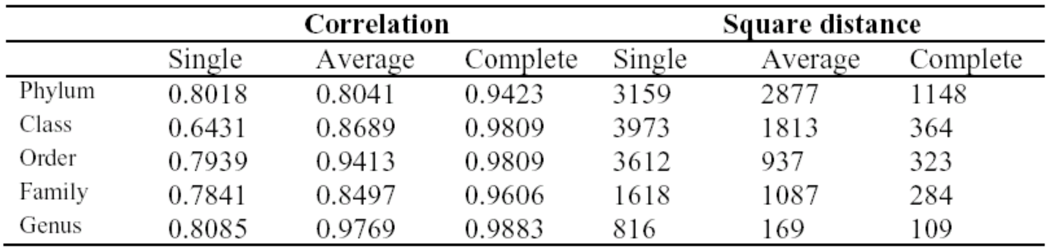


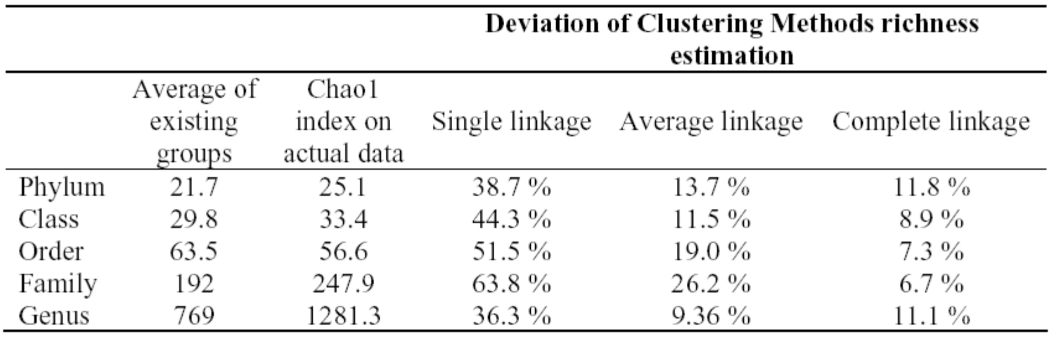
Figure 12: Cardinalities of the 75 groups formed by the single, average and complete clustering methods, compared to the original 75-group partitioning of all 16S vetted sequences, at the order level. The logarithmic values of the group sizes are presented in reverse sorted order. The two rightmost steps of each curve show the number of the doubletons and singletons, representing groups with two and one members, respectively.

complete methods differ insignificantly and have a quadratic time dependence to the number of sequences (or linear to the number of pair-wise sequence distances). For our clustering analysis. we used the statistical analysis program R [119]. Evaluation of the average and maximum in-group distances of a clustering were performed using the height and order arrays provided by R, in conjunction with the calculated pair-wise edit distances of the vetted sequences.



	Correlation			Square distance		
	Single	Average	Complete	Single	Average	Complete
Phylum	0.8018	0.8041	0.9423	3159	2877	1148
Class	0.6431	0.8689	0.9809	3973	1813	364
Order	0.7939	0.9413	0.9809	3612	937	323
Family	0.7841	0.8497	0.9606	1618	1087	284
Genus	0.8085	0.9769	0.9883	816	169	109

Table 9: Pearson correlation and square differences of the sorted cardinality lists of partition groups, created by the three clustering methods, against the original partitions of the vetted sequences. The Pearson correlation and square differences were calculated for different phylogenetic levels.



			Deviation of Clustering Methods richness estimation		
	Average of existing groups	Chao1 index on actual data	Single linkage	Average linkage	Complete linkage
Phylum	21.7	25.1	38.7 %	13.7 %	11.8 %
Class	29.8	33.4	44.3 %	11.5 %	8.9 %
Order	63.5	56.6	51.5 %	19.0 %	7.3 %
Family	192	247.9	63.8 %	26.2 %	6.7 %
Genus	769	1281.3	36.3 %	9.36 %	11.1 %

Table 10: Average Chao1 richness estimation index calculated for random 1000 sequence subsets from the RDP vetted sequence dataset and for the groupings from different clustering methods. The first two columns present the average number of phylotypes in the 1000 randomly selected sequences and the estimated Chao1 richness for each phylogenetic level, based on the known taxonomical grouping. The last three columns present the average deviation of the Chao1 richness estimate, based on the groupings acquired from different clustering methods, as a percentage of the Chao1 richness estimate of the known taxonomy. Here complete linkage clustering is outperforming the other clustering methods in all but one level (genus).

4.3 Conclusions

The 16S rRNA gene sequence enables the association of phylogeny [76, 116] and remains the most reliable method to determine completely new or divergent organisms. Aside from the availability of a curated dataset (i.e. the vetted sequences), the analysis of 16S rRNA gene sequences serves as a choice model, as it also permits a direct comparison of the composition of different communities.

Although tree manipulation and visualization utilities like *arb* [77], which use multiple alignment to construct phylogenetic trees, have the capability of handling large datasets, editing their input becomes a laborious and tedious task. Therefore, the need exists to develop classification tools to overcome both the computational limitations in accurately identifying taxonomical relationships, and reconstructing phylogenetic trees for the purpose of better extrapolating ecological roles.

Because of its simplicity and efficiency, single linkage clustering has often been used for clustering sequences [111]. Other tools, such as DOTUR [110], give the user the option to select different clustering methods, but no information is provided on which method actually performs better or what dissimilarity cutoff should be used to differentiate groups at a given phylogenetic level. We demonstrate that the complete linkage clustering method seems to be the preferential approach to create clusters of closely related sequences, taking into account that it is less computational intense than full phylogenetic tree analysis. The output of this clustering method can subsequently be used for richness estimation of the microbial community, using e.g. the Chao1 index, as we did for the different microbial communities associated with trembling aspen under conditions of ambient and elevated CO_2, presented in the previous chapter.

Chapter 5

Identifying Outer Membrane Genes in Photorhabdus Luminescens

5.1 Introduction

Photorhabdus luminescens (P. luminescens) is phylogenetically a member of the γ-proteobacteria based on analyses of 50 γ-proteobacterial 16S rRNA genes [86]. In a phylogenetic tree based on the sctV gene (which encodes a highly conserved inner membrane protein), P. luminescens falls into the Yersinia family [11].

P. luminescens has a complex life cycle and proliferates in two distinctly different environments [37, 132]. P. luminescens lives symbiotically in the nematode gut, but also has a pathogenic phase when the worm, which normally resides in the soil, infects an insect. In this stage, P. luminescens cells are

This chapter is drawn from our paper [94]. My contributions in this work are limited to the design and development of the motif and promoter locator utilities and obtaining data for micF.

released into the circulatory system (hemocoel) of the insect by the nematode. Here the bacteria grow and commence with the rapid killing of the insect and both the nematode and the bacteria feed from the dead insect [37, 132]. After nutrients derived from the insect carcass are depleted, the bacteria re-associate with the nematode and the symbiotic relationship is re-established. P. luminescens has not been found as a free living organism and thus differs significantly from E. coli and most other closely related γ-proteobacteria.

During the evolutionary period when P. luminescens evolved into a symbiont and a pathogen, its genome expanded such that it has one of the largest chromosomes of the γ-proteobacteria (5.7 Mb) [32]. This expansion is related to its pathogenic phase [37, 132, 32]. However certain genetic elements that contribute towards survival in a harsh environment but are no longer needed may have been lost from the genome during evolution of the organism.

Using a bioinformatics approach, the P. luminescens genome was analyzed for outer membrane porin protein and associated regulatory RNA genes. We find a limited presence of the porin genes and their RNA regulators.

5.1.1 Outer Membrane Proteins

Outer membrane porin proteins allow for the passive diffusion of small solutes into the bacterial cell. Passage of molecules through the cell envelope and control of this process are crucial to cell survival when nutrients are scarce or when the cell is exposure to toxins or other adverse conditions. In E. coli and related γ-proteobacteria, the major outer membrane porin proteins are OmpF and OmpC [89]. ompF and ompC genes are regulated transcriptionally by transcription factor OmpR in response to osmolarity change in the environment [38]. ompF is also regulated post-transcriptionally at the level of messenger RNA stability by the trans-encoded antisense RNA micF in response to various environmental factors such as temperature increase, oxidative stress and exposure to toxic compounds [29]. Regulatory non-coding RNAs (ncRNAs) in prokaryotes are also referred to as trans-encoded antisense RNAs. ompC in E. coli is regulated post-transcriptionally by the regulatory ncRNA micC [22]. OmpA, another major outer membrane protein, has multiple and more complex functions.[131] For example, OmpA adds to the stability of the cell envelope by linking the outer membrane to the peptidoglycan. It is involved in bacterial conjugation [102] and functions as a porin protein as well [39]. The stability of ompA mRNA varies with bacterial growth rate [117] and ompA mRNA is degraded at a fast rate when cells enter stationary phase [45]. Udekwu et al [124] recently showed that the regulatory micA RNA post-transcriptionally regulates ompA mRNA. In addition, micA is induced at stationary phase, a stress condition [124]. Thus in E. coli, three major outer membrane proteins, OmpF, OmpC, and OmpA are all regulated by specific small RNAs in response to stress factors.

5.2 Searching for RNA Primary and Secondary Structure Motifs in P. luminescens

The strategy used to search for a putative *micF* RNA in *P. luminescens* was to scan the genome using the conserved 13 nt 5' end *micF* sequence, i.e., 5'G_1CTATCATCATTA_{13}3' as well as variations of this sequence. Variations included T at position 2, T at positions 6 and 9, and in addition, a total of 4 random substitutions. A different first pattern that provides perfect complementarily to the *ompF* mRNA 5' UTR was also employed: 5'G_1TTTCATCATTATT_{14}3'. Variations included a total of four random substitutions and also allowing for the insertion of an A residue randomly between the 3rd and 10th base of the pattern. Additional constraints consisted of a rho- independent termination pattern situated $35-85$ bp downstream the two basic 5' end patterns shown above. The parameters used for the terminal rho-independent structure were a stem-loop followed by at least four T residues. The stem was $4-15$ bp with a minimum of three G-C pairs, the loop $3-8$ bases, and the maximum folding energy of loop was -9 Kcal/mol. Scans for the termination motif were performed after the initial identification of the two patterns shown above.

For scanning the *P. luminescens* genome, the perl programing language was used.

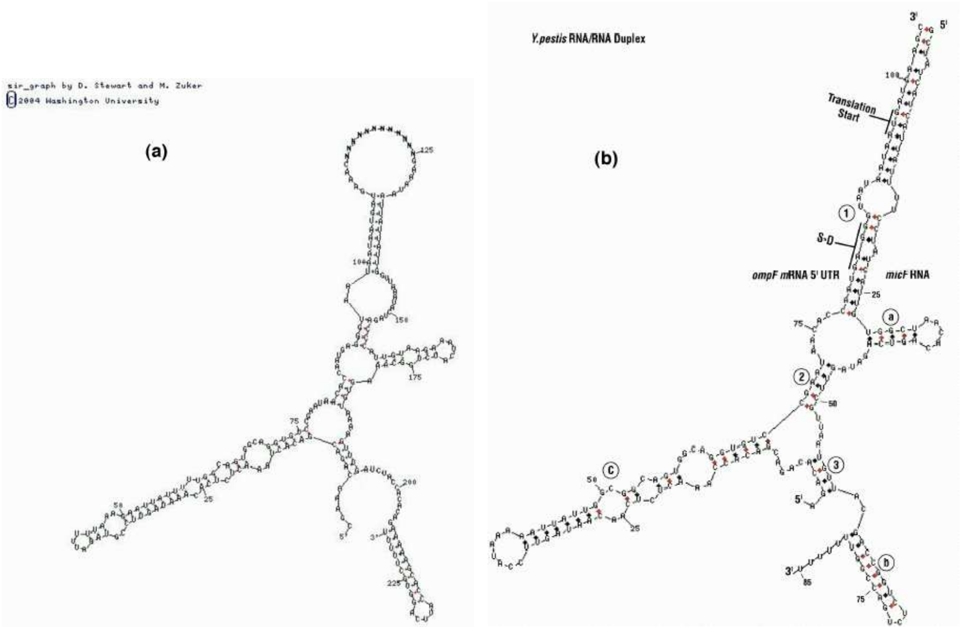
Additional scans were performed for –10 and –35 promoter sequences. To avoid 0 values, discounting for the probabilities in the consensus sequences were applied. Jeffrey Perk's law was used [79]:

Jeffrey Perk's law: $P(w) = (C(w)+1/2)/(N+B/2)$, where P is the assigned probability, w is a DNA character assignment, C(w) is the frequency of the character in the consensus table for the specific position, N is the number of training sequences used for the creation of the consensus table and B is the number of possible values for our character i.e., 4.

5.2.1 Searching for a Putative P. luminescens micF RNA

In the search for a putative *P. luminescens micF* RNA, the "fifth positive" sequence (described in the main text under the section **a. *ompC* and *micF***) is found in an intergenic region. It is 414 bp from the 5' side of the *ptsH* gene and 35 bp from the 3' side of the *cysK* gene and is located at positions 1674515-1674620 of the *P. luminescens* genome.

A promoter search was also performed for the *P.luminescens* "fifth positive" sequence. The results are described in terms of a p-value (the probability that the examined sequence appears by chance). The most prominent



(a) RNA/RNA secondary structure model of the fifth positive sequences

(b) Y. pestis micF RNA/ompF mRNA 5' UTR duplex model, reproduced from [28]

Figure 13: Examining micF candidates

promoter candidate for the "fifth positive sequence" had a p-value of 2.0x10^{-3} and was calculated for the P-35 at -37 and a spacer of 17 bp between the –35 and –10 sequences. For the *E.coli micF* promoter [26] the calculated p-value is 3.3x10^{-3} (P-35 at –36 and spacer of 17 bp). Although the "fifth positive" sequence provides a promoter probability in the range of that for *E. coli micF*, the *micF* promoter produces a weak signal. The statistical methods used are described below.

Figure 13a shows a model of the RNA/RNA duplex structure of the "fifth positive" sequence with *P. luminescens ompF* mRNA 5' UTR. Figure 13b (reproduced from [28]) shows the *Y. pestis micF* RNA/*ompF* mRNA 5' UTR duplex as a comparison. This structure has the characteristics of *micF/ompF*

RNA/RNA duplexes with the exception of "blunt ends" which is characreristic of duplexes from all bacteria known to have a *micF* gene. Thus although we have discounted this sequence as a potential *micF* in *P. luminescens*, we cannot rule out that the "fifth positive sequence" encodes a small RNA as it has possible promoter and termination signals.

5.3 Promoter Search Methods

In order to locate possible P-35 and P-10 promoter sites for candidate sequences, a search was performed in the region upstream of the sequence, starting at -50 and ending at -30 for the P-35 promoter. The spacer between the P-35 and P-10 promoter allowed for a length of 15-19 bp.

To calculate p-values (probability that the examined sequence appears by chance) and to examine the significance of the promoter findings, the base frequency distributions in the consensus hexamers as compiled in reference [75] was used. Initially, the probability of the two hexamers appearing in a specific location (under the constraints previously specified) is calculated as the geometric mean of the product of the individual frequencies of all bases of the candidate promoters at this location. In order to calculate a p-value for this probability, a pre-compiled sorted table of all possible 4^{12} values of probabilities for each hexamer pair that could comprise our two promoters is consulted, in order to locate the value of the probability for the location we are examining. The p-value is then calculated as the number of values equal or higher than the one we calculated, over the total number of values (4^{12}).

Although the p-values are significant, they become less significant if corrected for the number of samples, e.g. Bonferroni correction [7], since a variety of possible positions for the P-35 promoter and five values, 15-19 for the spacer between the two promoters were considered. This is true for the known promoter site for the *E. coli micF*. Thus this accounts for attributing the *E.coli micF* as a weak signal.

5.4 Conclusion

By bioinformatics analysis of conserved genetic loci, mRNA 5' UTR sequences, RNA secondary structure motifs, upstream promoter regions and protein sequence homologies, an ompF like porin gene in P. luminescens as well as a duplication of this gene have been predicted. Gene loci for micF RNA, as well as OmpC protein and its associated regulatory micC RNA, were not found. Significantly, a sequence bearing the appropriate signatures of the E. coli micA RNA was located. The ompA homolog was previously annotated in P. luminescens.

Presence of an ompF-like porin in P. luminescens is in keeping with the necessity to allow for passage of small molecules into the cell. The apparent lack of ompC, micC and micF suggests that these genes are not essential to P. luminescens and ompC and micF in particular may have been lost when the organism entered its defined life cycle and partially protected habitat. Control of porin gene expression by RNA may be more prevalent in free-living cells where survival is dependent on the ability to make rapid adjustments in response to environmental stress. Regulation of ompA by micA may have been retained due to a necessity for ompA control during one or both stages of the P. luminescens life cycle.

Part II

Genomic Sequence Design

Chapter 6

Design of Overlapping Genes

6.1 Introduction

The emerging field of *synthetic biology* moves beyond conventional genetic manipulation to construct novel life forms that do not originate in nature. The synthesis of poliovirus from off-to-shelf components [15] attracted worldwide attention when announced in July 2002. Subsequently, the bacteriophage PhiX174 was synthesized using different techniques in only three weeks [115], and Kodumal, et.al [70] recently set a new record for the longest synthesized sequence, at 31.7 kilobases. The ethics and risks associated with synthetic biology continue to be debated [2], but the pace of developments is quickening. Indeed, Tian, et.al. [120] have just proposed a method for DNA synthesis based on microarrays and multiplex PCR that promises a substantial reduction in cost.

Once you can synthesize an existing genome from scratch, you can do the same for new and better designs as well. In this chapter, we explore an interesting problem in genome design, namely designing the provably shortest

This chapter is drawn from our paper [129]. My contributions in this work are limited to the development and implementation of the algorithm for gene overlapping and observations on natural and random gene overlap dynamics.

genomic sequence to encode a given set of genes, by exploiting alternate reading frames and the redundancy of the genetic code. Theoretically, up to six proteins can be encoded on the same genomic sequence using three alternate reading frames on both strands. Indeed, long gene overlaps occur frequently in nature.

Our contributions are:

- *Finding Shortest Encodings for Given Protein Pairs* – We present an algorithm for designing the shortest DNA sequence simultaneously encoding two given amino acid sequences. Our algorithm runs in worst-case quadratic time, but we provide an expected-case analysis explaining its observed linear running time when employing the standard DNA triplet code.

- *Comparing Natural and Synthetic Coding-Pair Sequences* – We compare the overlapping gene designs constructed by our algorithm to those occurring in natural viral sequences. We show that the coding sequence of naturally occurring pairs of overlapping genes in general approach maximum compression, meaning that it is impossible to design overlapping shorter coding sequences for them which save more than 1-2% over independent genes. This counterintuitive result has natural explanation in terms of the evolutionary mechanics of overlapping gene sequences.

  Further, we show interesting differences between the preferred phase (reading frame), strand, and orientation of natural and optimized overlapping sequences.

- *Impact of Alternate Coding Matrices on Overlapping Sequence Design* – Protein designs are not immutable; indeed, certain pairs of amino acids share such similar physical/chemical properties they can be fairly freely substituted without altering protein function. This freedom can be exploited to design substantially shorter encodings for a given pair of proteins.

We investigate the impact of increasingly permissive amino acid substitution matrices (derived from the well-known PAM250 matrix) on the potential for constructing tight encodings. Extremely tight encodings are often possible while largely preserving the hydrophobicity of the associated residues. Further, the encodings designed under each of these matrices shows interesting differences between the preferred phase (reading frame), strand, and orientation.

- *Biotechnology Applications of Nested Encodings* – We propose an interesting application for overlapping gene design, namely the interleaving of an antibiotic resistance gene into a target gene inserted into a virus or plasmid for amplification. Selective pressures tend to quickly remove such target genes as disadvantageous to the host. However, coupling such a target with a resistance gene provides a means to select for individuals *containing* the arbitrarily selected target gene.

  To demonstrate the feasibility of this technique, we apply our algorithm to encode each of five important antibiotic resistance genes within the body of the Hepatitis C virus. In fact, we demonstrate there are several possible places to encode each resistance gene within the virus, assuming a sufficiently (but not excessively) permissive codon replacement matrix.

These sequence design problems naturally arise in our project, currently underway, to design and synthesize weakened viral strains to serve as candidate vaccines. This work also follows our previous efforts to design encoding sequences for proteins which minimize or maximize RNA secondary structure [24] and avoid restriction sites [114].

This chapter is organized as follows. In Section 6.2, we survey the literature concerning why gene overlaps occur in nature and how they evolve. We present our algorithm for constructing optimal encodings in Section 6.3, with associated analysis, and compare our synthesized designs with wildtype viral encodings. In Section 6.4, we study the impact of alternate codon substitution matrices on the size and parity of minimal pairwise gene encodings. Finally, in Section 6.5, we present our results on encoding antibiotic resistance genes within viral coding sequences.

6.2 Overlapping Genes in Nature

Overlapping genes are adjacent genes whose coding regions are at least partly overlapping. They occur most frequently in prokaryotes, bacteriophages, animal viruses and mitochondria, but are seen in higher organisms as well. Gene overlapping presumably results from evolutionary pressure to minimize genome size and maximize encoding capacity. For viruses, this is manifested in two ways; first when genome size substantially affects the speed of replication, and second when an upper bound on the genome size is imposed by packaging.

Overlapping genes are common for viruses with prokaryotic hosts because they must be able to replicate sufficiently fast to keep up with their host cells [12]. As an example, many bacteriophages have compact genomes which maximize coding information into the minimum genome size [12]. In term of evolutionary pressure to minimize genome size, packaging size pressure (the packaging size of the virus particle as the amount of nucleic acid which can be incorporated into the virion) sets the genome size upper bound for viruses with eukaryotic hosts [12].

Overlap between genes is very common in genomes mutating at high rates,

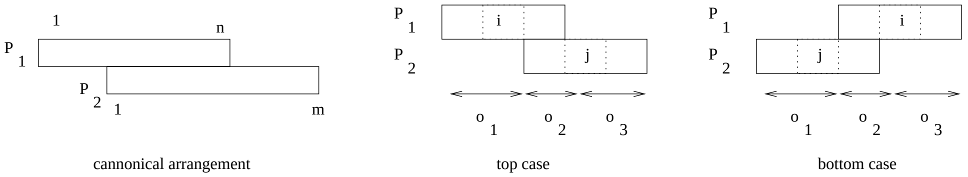


Figure 14: Notation for the gene encoding algorithm: the canonical encoding (left), with the top (center) and bottom (right) overhang cases.

such as bacteria and mitochondria, but especially viruses. Although a mutation in an overlapping region can impair more than one protein and would be naturally selected against, there are several reasons overlapping genes can benefit an organism:

- By reducing the size of the genome, without affecting the number of genes encoded.
- By generating new (or sometimes more complex) proteins without increasing the size of the genome.
- By coordinating the expression levels of functionally related genes.
- By coordinating the expression levels of genes, where the expression of one gene requires the deactivation of the other.

The first two functions are supported by the theory of "overprinting", which attempts to describe the origin of new genes from an existing genome with minimal mutational change [66]. Size reduction is considered important under the assumption that replication rate is inversely related to genome length, since it has an obvious effect in increased rates of replication and minimization of mutation load.

Overlapping reading frames can serve to expedite efficient translation. Overlaps can bring translation machinery close to both overlapping genes, which can co-ordinate or co-regulate their expression. In other cases an overlap can bring the termination site of one gene into the same region as the translation initiation site for the next gene [92].

The rate of evolution can be expected to be slower in overlapping genes [84]. Since point mutations in overlapping regions can affect two genes simultaneously, a mutant variant produced with a mutation in an overlapping region will have a lower growth rate and in most cases cannot compete with the wild type variant.

Although high mutation rates and selection towards a compacted genome would indicate that overlapping genes should occur mostly in viral and cellular prokaryotic genomes and mitochondria, recent studies show that mammalian genomes have relatively frequent occurrences of overlapping genes too. The observed 774 overlapping genes in the human and 542 overlapping genes in the mouse genome [125] do not compare favorably with the 806 overlapping gene pairs in the genome of E.Coli, since the latter genomes is three orders of magnitude smaller. Nevertheless, the same mechanisms of evolution, like rearrangements or loss of parts and utilization of neighboring gene signals, provide explanation for the origin of these overlaps.

Overlapping genes offer an efficient way to study how coding and control sequences have evolved. With direct comparison of the overlapping genes for related species, one can determine how the overlaps evolved and under which conditions, like neighboring gene distance (for example, in closely related bacterial species it has been observed that most of the overlapping genes were generated or degraded in gene pairs that have a short intergenic region [41]). By comparing gene overlaps that are not conserved between related species, the mutational changes that caused the diversion can often be identified. In other cases further species sequencing are necessary to decipher the evolutionary mechanisms and tendencies [41, 125].

There is evidence that in certain gene overlaps the overlapping region is younger than the coding sequences [91]. Other cases clearly indicate that overlap occurred after the loss of a stop codon, the start codon, or both simultaneously [41].

The stability of an overlap greatly depends on the direction and phase of the overlap. Theoretically we expect antiparallel (head-on or end-on) +1-phase to be most common, followed by parallel +1-phase and +2-phase and then antiparallel 0-phase and +2-phase overlap [71]. This order is not observed in our experiments on sets of viral genomes, which would indicate that specific evolutionary changes like the loss of end and start codons occur more frequently than larger piece insertions and deletions, the former being the cause of the more popular parallel and head-on overlaps. To support the last argument, natural gene overlaps in many bacterial organisms match the unidirectional, convergent (head-on) and divergent (end-on) relative occurrences, with the unidirectional type appearing most, followed by the convergent and divergent types.

In bacterial species it has been observed that the total number of overlapping genes depends on the genome size or the total number of genes, which could imply that the rates for the accumulation and degradation of overlapping genes are universal among bacterial species [41].

Overlapping gene regions can also provide information for evolution patterns among classes of organisms and seem to converge with ribosomal RNA phylogenetic methods' results [103]. For certain bacterial species, the extent of conservation of unidirectional overlaps correlates with the evolutionary distances between pairs of species. Gene overlaps have even been correlated with certain human disease genes; further genomic rearrangements are likely to occur within overlapping regions, possibly as a consequence of anomalous sequence features prevalent in these regions [63].

6.3 Finding Maximally Compressed Gene-Pair Encodings

Our algorithm for constructing the maximally compressed encoding for a given pair of amino acid sequences P_1 and P_2 can be most succinctly described via a dynamic program. We consider the canonical case where the encoding of P_1 starts to the left (5' end) of P_2 as shown in Figure 14(left); the reverse case follows by simply relabeling the proteins. We present only the algorithm for the case of same-strand encodings; the case of alternate strand encodings follows analogously.

Let P_1 contain n residues and P_2 m residues, respectively. Let o_1, o_2, and o_3 denote possible DNA sequences of 0 to 3 bases in length. There are two general cases:

- We say that $C[i, j, o_1, top]$ is *realizable* iff there exists a pair of sequences o_2, o_3 such that o_1o_2 codes for residue $P_1[i]$, o_2o_3 codes for residue $P_2[j]$, and $C[i+1, j, o_3, bottom]$ is realizable or $i = n$.
- We say that $C[i, j, o_1, bottom]$ is *realizable* iff there exists a pair of sequences o_2, o_3 such that o_1o_2 codes for residue $P_2[j]$, o_2o_3 codes for residue $P_1[i]$, and $C[i, j+1, o_3, top]$ is realizable or $j = m$.

An exception occurs only when the residues are aligned, where only one case is needed, in which we advance both indices i and j and we check for reaching both ends of the proteins.

The basis cases for the canonical labeling assert that an overlap is attainable ($C[n, j, o_1, top]$ or $C[i, m, o_1, bottom]$ is realizable) iff $C[j, 1, o_1, top]$ is realizable for some $1 < j < n$. Since there are only a constant number of possible short strings o_1, o_2, and o_3, it takes constant time to evaluate a given value of $C[i, j, o, b]$ given the solution of all smaller cases. With $\Theta(mn)$ values to evaluate, the algorithm runs in worst case $\Theta(mn)$ time.

By ceasing evaluation once no realizable values remain, the longest overlap can be computed in $O((n + m)l)$, where l is the length of the longest overlap between the protein sequences. Below, we argue that l should in general be of constant length on non-degenerate substitution matrices; this states that on average this algorithm should run in linear time on such matrices.

We say that two overlapping proteins are *in-phase* if the overlap length is congruent to 0 mod 3, i.e. they align along codon boundaries. Non-trivial in-phase, same-strand overlapping designs are in principle forbidden by the fact that proteins must end with stop codons. However, we consider an abstraction of this case to simplify the analysis.

Here we consider the expected length of the maximal overlap as a function of the *residue equivalence probability*, defined as the probability that two randomly selected amino acids have an equivalent codon between them. This residue equivalence probability p is a function both of the codon substitution matrix and the distribution of amino acids in the proteins.

Assuming independence of the protein sequences, the expected length of the longest left-right overlap $E(O)$ of two random sequences P_1 and P_2 is given by

$$E_1(O) = \sum_{l=0}^{\infty} l p^l \prod_{i=l+1}^{\infty} (1 - p^i)$$

For the case of two-sided overlaps (i.e. either P_1 or P_2 may occur on the left side of the alignment),

$$E_2(O) = \sum_{l=0}^{\infty} l(2p^l - p^{2l}) \prod_{i=l+1}^{\infty} (1 - (2p^i - p^{2i}))$$

The above analysis demonstrates that the expected maximum overlap length remains quite small until the residue equivalence probability approaches

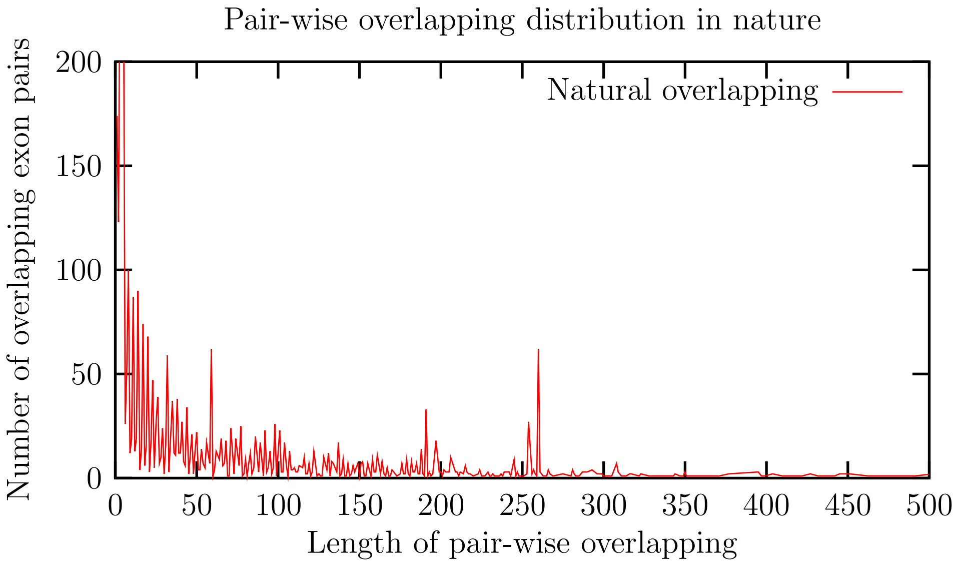


Figure 15: Length distribution of pairwise-overlapping genes in viral genomes.

	All Overlaps, parity mod 3					Length > 4, parity mod 3			
Pattern	0	1	2	All		0	1	2	All
SAME	0.0%	23.1%	39.9%	63.0%		0.0%	12.9%	53.3%	66.2%
HH	4.4%	1.9%	4.8%	21.1%		5.9%	4.9%	6.5%	17.3%
TT	3.0%	1.9%	11.0%	15.9%		4.0%	2.6%	9.9%	16.5%
total	7.4%	36.9%	55.7%	100%		9.9%	20.4%	69.7%	100%

Table 11: Parities of natural gene overlaps, ties discarded. All 3232 gene pairs (left). The 2407 gene pairs with overlap > 4.

1. This suggests that two arbitrary proteins are unlikely to permit substantially compressed in-phase encodings except under a forgiving (degenerate) coding matrix.

Still, all is not lost. Our analysis of both wildtype and synthetic overlaps demonstrates that out-of-phase encodings are likely to be substantially longer than in-phase encodings. This phenomenon appears to be difficult to analyze in general because it strongly depends upon the properties of the codon

	All Overlaps, parity mod 3					Length > 4, parity mod 3			
Pattern	0	1	2	All		0	1	2	All
SAME	0.01%	31.92%	1.47%	33.40%		0.05%	3.04%	13.16%	16.25%
HH	5.09%	35.51%	2.31%	42.91%		45.55%	7.77%	21.07%	74.39%
TT	0.08%	0.91%	22.70%	23.69%		0.62%	8.28%	0.46%	9.36%
total	5.18%	68.34%	26.48%	100.00%		46.22%	19.09%	34.69%	100.00%

Table 12: Longest optimized overlap using codon matrix, ties discarded. All 135,869 overlapping gene pairs (left), the 14,925 overlapping gene pairs of length > 4 (right).

equivalence matrix.

Each amino acid is encoded by a minimum of one and a maximum of six different codons. In total, 61 of the 64 codons encode 20 amino acids while the other three are stop codons, a termination point for protein-synthesizing machinery. Thus there is an approximate 1-to-3 correspondence between amino acids and their codon encodings. It is this redundancy that offers the flexibility in amino acid sequence encoding.

To study the extent of gene overlapping in viruses, we analyzed all 1058 completely sequenced viral genomes available in Genbank as of February 22, 2004. After excluding 273 genomes containing a single annotated gene (and hence not a candidate for gene overlapping) and 108 genomes with sequence ambiguity or obvious annotation errors, we were left with 677 viruses of interest. In total, these viruses contained 3,232 pairs of overlapping genes, 2,407 of which had overlaps of length greater than four bases.[1]

Figure 15 presents the frequency distribution of gene overlaps by length, the tail of which demonstrates that long overlaps occur with surprisingly high frequency. Table 11 shows a partition of these overlaps into disjoint cases, distinguished by whether the genes occur on the same strand, or are head-to-head or tail-to-tail on opposing strands. Same strand overlaps dominate in the sample. Table 11 also shows partitions of these overlaps by the length mod 3.

[1] Overlaps of less than four bases are not particularly interesting, since the possible overlaps are restricted to the start and stop codons possessed by every gene.

In-phase overlaps are understandably rare (any stop codon breaks both same strand sequences), but there is also a clear preference for 2 mod 3 parity over 1 mod 3.

Using our gene pair encoder, we attempted to find more compressed representations of the wildtype gene pairs. In general we failed badly, with the vast majority of cases having zero or insignificant improvement (recall that approximately one third of all natural overlaps were of length 4 or less). In no case were we able to increase the overlap length of such an overlapping gene pair by more than 20 bases.

The lesson here is that gene overlaps occur because the proteins evolved together – significant potential overlaps are extremely unlikely to arise in unrelated sequence pairs because the genetic code does not provide sufficient flexibility. Figure 16 presents the results of optimally encoding 135,869 pairs of unrelated proteins. In no case were we able to reduce the length of an overlapping gene pair by more than 30 bases.

More interesting is the breakdown of our optimized encodings by strand and parity, reported in Table 12. The optimized encodings show sharply different preferences than the wildtype encodings. Functional demands likely constrict the choice of same strand encodings, although it is less obvious why there is such dramatic difference in head-to-head and tail-to-tail preferences. The difference in preferred parity is largely explained by the change in strand encoding distribution.

6.4 Experiments in Synthetic Gene Encoding

Recent studies [40] have demonstrated that the genetic code maximizes the likelihood that a gene mutation will not harm and may even improve the

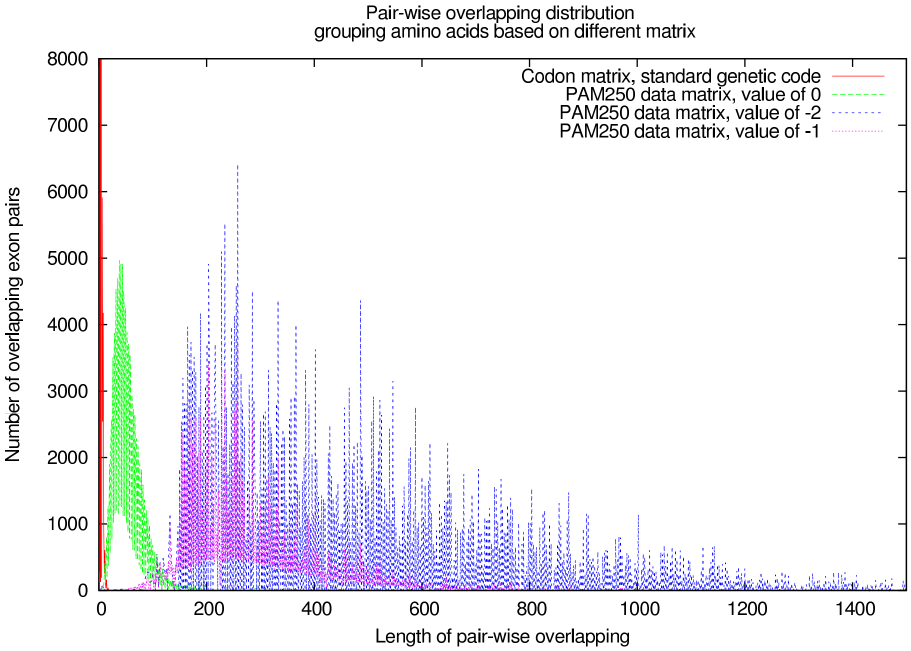


Figure 16: Distribution of maximum overlaps under four different codon substitution matrices.

protein. In general, the code is resilient to random mutations leading to significant changes of the affected amino-acid properties, so that a misread codon often codes for the same amino acid or one with similar biochemical properties. Furthermore, simulations by Gilis et al. [46] have shown that taking the amino-acid frequency into account further increases the resilience of the code compared to random codings. It is also known that proteins with a limited number of point mutations which lead to non-synonymous substitutions fold in similar ways, in a degree that homology database search can detect function similarity in proteins differing in up to 50% of their amino-acid compositions [80].

Based on these results, we decided to further investigate the pairwise gene overlapping possibilities using non-synonymous amino-acid substitution matrices, which increase the combinatorial possibilities of compressed overlapping representations at the cost of minor changes in the residues in the underlying proteins.

Our substitution matrices are derived from the well-known PAM 250 amino acid substitution scoring matrix. The value of each entry describes the reward or penalty in replacing an instance of the first amino acid with the second in aligned sequences. Positive values contribute favorably to an alignment, and negative values unfavorably. We may derive a permissive codon equivalence matrix from PAM 250 as a function of a threshold t by permitting replacement of amino acid x with y if the score is $\geq t$. By decreasing t, we can define a sequence of increasingly permissive substitution matrices for our experiments.

Clearly other substitution matrices are possible (e.g. Levitt's hydrophobicity scoring matrix [74]), and perhaps even preferable. Our primary interest is establishing the flexibility for compressed sequences as a function of more tolerant substitution matrices.

For this purpose we used three substitution matrices, one to indicate amino-acid compatibility based on mutational changes and two matrices which indicate hydrophobicity compatibility, each one allowing a different level of acceptable substitution distances. In the substitution matrix based on mutations, the amino-acids are separated in disjoint equivalence groups, where the hydrophobicity matrix limits substitutions in overlapping groups for each amino-acid, based on the hydrophobicity "neighborhood" of each. In the two cases we considered, the neighborhoods are limited by a hydrophobicity of 10 and 9 respectively. The substitution possibilities are presented in Table 6.4.

The results of our overlapping experiments with the use of the alternate substitution matrices are shown in Figure 16. One can observe the significant increase in both the number and frequency of long overlaps with increasing

AA	Codons	Mut	Count	Hydro 10	Count	Hydro 9	Count
R	6	HRK	10	RK	8	RKDE	12
K	2	HRK	10	RK	8	RKDE	12
D	2	NDEQ	8	DE	4	RKDE	12
E	2	NDEQ	8	DE	4	RKDE	12
S	6	STPAG	22	SNQG	14	SNQGTHA	24
N	2	NDEQ	8	SNQG	14	SNQGTHA	24
Q	2	NDEQ	8	SNQG	14	SNQGTHA	24
G	4	STPAG	22	SNQG	14	SNQGTHA	24
T	4	STPAG	22	THA	10	SNQGTHACM	27
H	2	HRK	10	THA	10	SNQGTHACMP	31
A	4	STPAG	22	THA	10	SNQGTHACMP	31
C	2	C	2	CM	3	THACMPVLI	30
M	1	MILV	14	CMPV	11	THACMPVLI	30
P	4	STPAG	22	MPV	9	HACMPVLIY	28
V	4	MILV	14	MPVLI	18	CMPVLIY	22
L	6	MILV	14	VLI	13	CMPVLIYF	24
I	3	MILV	14	VLI	13	CMPVLIYF	24
Y	2	FYW	5	YF	4	PVLIYF	21
F	2	FYW	5	YF	4	LIYFW	14
W	1	FYW	5	W	1	FW	3
Z	3	Z	3	Z	3	Z	3

Table 13: Matrix Amino Acid Equivalence, based on codon matrix, hydro 9 matrix, hydro 10 matrix and mutation matrix.

length as the matrices become more permissive. In particular, almost arbitrarily long overlaps appear possible under $t \geq -3$ substitution.

6.5 Hiding Short Genes in Long Genes

Here we report on proof-of-concept simulations of two related biotechnology applications for carefully designed overlapping of synthetic gene sequences:

- *Plasmid incorporation into mammalian cells* – A common technique for incorporating target gene expression into mammalian cell involved plasmid incorporation and mammalian cell transfection. Initially, the plasmid containing the target gene is propagated in bacteria. The naked plasmid DNA is extracted and then introduced into the mammalian cell by transfection.

  Typically the target gene is paired with an antibiotic resistance gene, so as to create a marker for selection in the eukaryotic cell. All cells not expressing this marker can be eliminated with the corresponding antibiotic drug (ex. geneticin or G418), to isolate cells expressing the target protein. Sometimes, however only one gene is expressed, such as when the cell fails to incorporate the entire plasmid. Because the plasmid is linearized to be incorporated in a chromosome, the cut may also occur in the target gene location.

  By overlapping the target and marker genes, we reduce the probability that either the cut will eliminate the target gene but the not the marker, as well as the probability that the two genes will be separated.

- *Foreign gene incorporation into viruses* – RNA viruses are very prone to recombination, so an added sequence has a high probability to be deleted.

> Since RNA viruses are streamlined to perform a limited number of specific tasks, the addition of a gene slows down the virus processes, merely by extending slightly its length. Since the foreign gene is undesirable, its deletion will result in a faster produced replicon that will eventually outgrow the engineered virus we implanted.
>
> Interleaving the target gene into a gene that the virus needs can prohibit its deletion through reversion.

Positive indications in the direction of gene overlap engineering are the recent results of [40], which show that the amino acid code minimizes the effects of mutations and maximizes the likelihood that a gene mutation will improve the resulting protein. Additionally, methods of local sampling [35, 52] can help us simulate the behavior of slightly altered proteins in respect to folding and docking, so that we can test the codon substitution effects without lab experimentation.

To evaluate the potential for such synthetic overlap encodings, we attempted to find maximal encodings of five important antibiotic genes (whose length ranges from 375 to 1026 nucleotides) within the coding region of the Hepatitis C virus (HCV). Consistent with the results of the previous section, only trivial overlaps can be obtained using synonymous substitutions. Partial overlaps, but not complete nesting can be obtained with the hydro-10 and mutation substitution matrices.

However, multiple complete encodings are possible under $t \leq -2$ and $t \leq -3$ substitution for each of the five antibiotic resistance genes, as reported in Table 14. There is a strong bias for alternate strand encodings, although all five antibiotic resistance genes offer same strand encodings for $t \leq -3$. In fact, the preferable target for the inserted gene encoding (and promoter region) in the virus application is the minus strand, so this bias appears fortunate.

	accession		number of $t < -2$ encodings		number of $t < -3$ encodings	
Resistance gene	number	length	same strand	alternate strand	same strand	alternate strand
Hygromycin	X03615	1026	0	1	4	1
Neomycin	M55520	795	0	1	2	3
Puromycin	X92429	600	0	11	16	25
Blasticidin	AYI96214	423	56	250	217	442
Zeocin	A31902	375	35	132	163	175

Table 14: Number of fully-enclosed $t \leq -2$ and $t \leq -3$ encodings of antibiotic resistance genes within the Hepatitis C virus.

Based on these results, we are pursing more rigorous designs for intended synthesis and implementation.

Chapter 7

Codon Bias Designs

7.1 Introduction

The rapidly developing technologies in the field of synthetic biology allow for the cost-efficient de novo synthesis of DNA sequences without the need for a natural template. This allows for the generation of entirely novel coding sequences or the modulation of existing sequences to a degree practically impossible with traditional cloning methods. Inspired by a previous work on the chemical synthesis of poliovirus (PV) in the absence of natural template [15], we are now actively exploring the utility of de novo gene synthesis for the customization of virus properties.

As a result of the degeneracy of the genetic code, all but two amino acids in the protein coding sequence can be encoded by more than one synonymous codon. The frequencies of synonymous codon use for each amino acid are unequal and have coevolved with the cell s translation machinery to avoid excessive use of suboptimal codons which often correspond to rare or otherwise disadvantaged tRNAs [47]. This results in a phenomenon termed *synonymous*

This chapter is drawn from our paper [87]. My contributions in this work are limited to the design of the synthetic poliovirus capsid regions.

codon bias, which varies greatly between evolutionarily distant species and possibly even between different tissues in the same species [98].

While codon optimization by recombinant methods (that is, to bring a gene's synonymous codon use into correspondence with the host cell's codon bias) has been widely used to improve cross-species expression [47], the opposite direction of reducing expression by intentional introduction of suboptimal synonymous codons has seldom been chosen.

7.2 The Effects of Altered Codon Distribution

In the present work, we have reengineered the capsid coding region of poliovirus type 1 Mahoney [PV(M)] by introducing through de novo gene synthesis the largest possible number of rarely used synonymous codons (PV-AB) or the largest possible codon position changes while maintaining the original codon bias (PV-SD). We found viruses arising from PV-AB-type designs to be attenuated by a previously underappreciated mechanism. While the primary defect of these genomes was at the level of genome translation, codon-deoptimized viruses were marked by a reduction in virus-particle-specific infectivity up to 1,000-fold. Thus, while producing similar amounts of virus particles per cell, production of infectious units (measured by functional assays) is greatly reduced. Due to the high degree of genetic stability as a result of the large number of mutations contributing to the phenotype, we propose that codon-deoptimized virus may present a useful and safer alternative for the production of poliovirus vaccines, especially inactivated vaccines.

In order to design of codon-deoptimized polioviruses, we produced two different synonymous encodings of the poliovirus P1 capsid region, each governed by different design criteria. We limited our designs to the capsid, as it has been conclusively shown that the entire capsid coding sequence can be deleted from the poliovirus genome or replaced with exogenous sequences without affecting replication of the resulting subgenomic replicon [58, 59]. It is therefore quite certain that no unidentified crucial regulatory RNA elements are located in the capsid region, which might be affected inadvertently by modulation of the RNA sequence.

In our first design (PV-SD), we sought to maximize the number of RNA base changes while preserving the exact codon usage distribution of the wild-type P1 region (Figure 17). To achieve this result, we exchanged synonymous codon positions for each amino acid by finding a maximum weight bipartite match [42] between the positions and the codons, where the weight of each position-codon pair is the number of base changes between the original codon and the synonymous candidate codon to replace it. To avoid any positional bias from the matching algorithm, the synonymous codon locations were randomly permuted before creating the input graph and the locations were subsequently restored. We used Rothberg's maximum bipartite matching program [104] to compute the matching. A total of 11 useful restriction enzyme sites, each 6 nucleotides, were locked in the viral genome sequence so as to not participate in the codon location exchange. The codon shuffling technique potentially creates additional restriction sites that we prefer to remain unique in the resulting reconstituted full-length genome. For this reason, we further processed the sequence by substituting codons to eliminate the undesired sites. This resulted in an additional nine synonymous codon changes that slightly altered the codon frequency distribution. However, no codon had its frequency changed by more than 1 over the wild-type sequence. In total, there were 934 out of 2,643 nucleotides changed in the PV-SD capsid design, when compared to the wild-type P1 sequence, while maintaining the identical protein sequence of the capsid coding domain (Figure 17). Since the codon usage in this design was not changed, the GC content in the PV-SD capsid

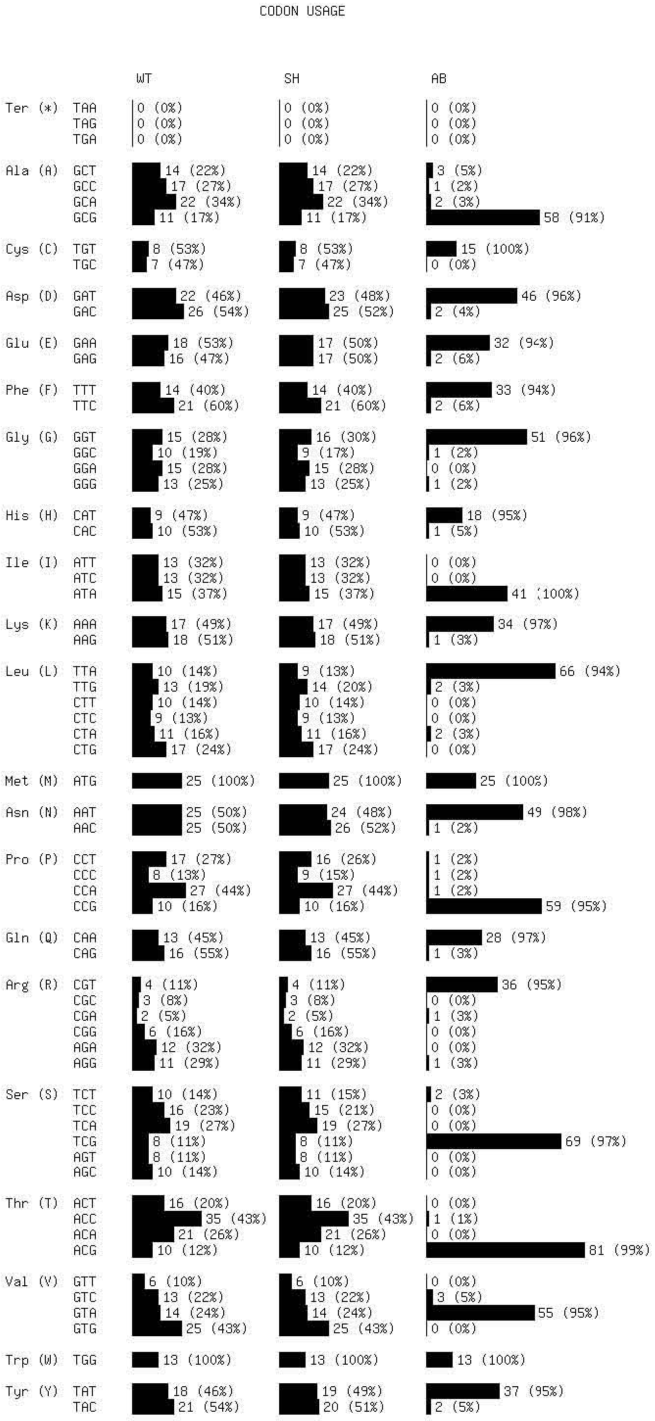


Figure 17: Codon use statistics in synthetic P1 capsid designs.

coding sequence remained identical to that in the wildtype at 49%.

The second design, PV-AB, sought to drastically change the codon usage distribution over the wild-type P1 region. We were influenced by recent work suggesting codon bias may impact tissue-specific expression [98]. Our desired codon usage distribution was derived from the most unfavorable codons observed in a previously described set of brain-specific genes [53, 98]. We synthesized a capsid coding region maximizing the usage of the rarest synonymous codon for each particular amino acid as observed in this set of genes (Figure 17). Since for all amino acids but one (leucine) the rarest codon in brain corresponds to the rarest codons among all human genes at large, in effect this design would be expected to discriminate against expression in other mammalian tissues as well. Altogether the PV-AB capsid differs from the wildtype capsid in 680 nucleotide positions. The GC content in the PV-AB capsid region was reduced to 43% compared to 49% in the wildtype.

Codon-deoptimized polioviruses display severe growth phenotypes. Of the two initial capsid ORF replacement designs (Figure 18A), only PV-SD produced viable virus. In contrast, no viable virus was recovered from four independent transfections with PV-AB RNA, even after three rounds of passaging (Figure 18E). It appeared that the codon bias we introduced into the PV-AB genome was too severe. Thus, we subcloned smaller portions of the PV-AB capsid coding sequence into the PV(M) background to reduce the detrimental effects of the nonpreferred codons. Of these subclones, PV-AB$^{2954-3386}$ produced cytopathic effect 40 h after RNA transfection, while PV-AB$^{755-1513}$ and PV-AB$^{2470-2954}$ required one or two additional passages following transfection, respectively (compared to 24 h for the wildtype virus). Interestingly, they represent the three subclones with the smallest portions of the original AB sequence, an observation suggesting a direct correlation between the number of nonpreferred codons and the fitness of the virus.

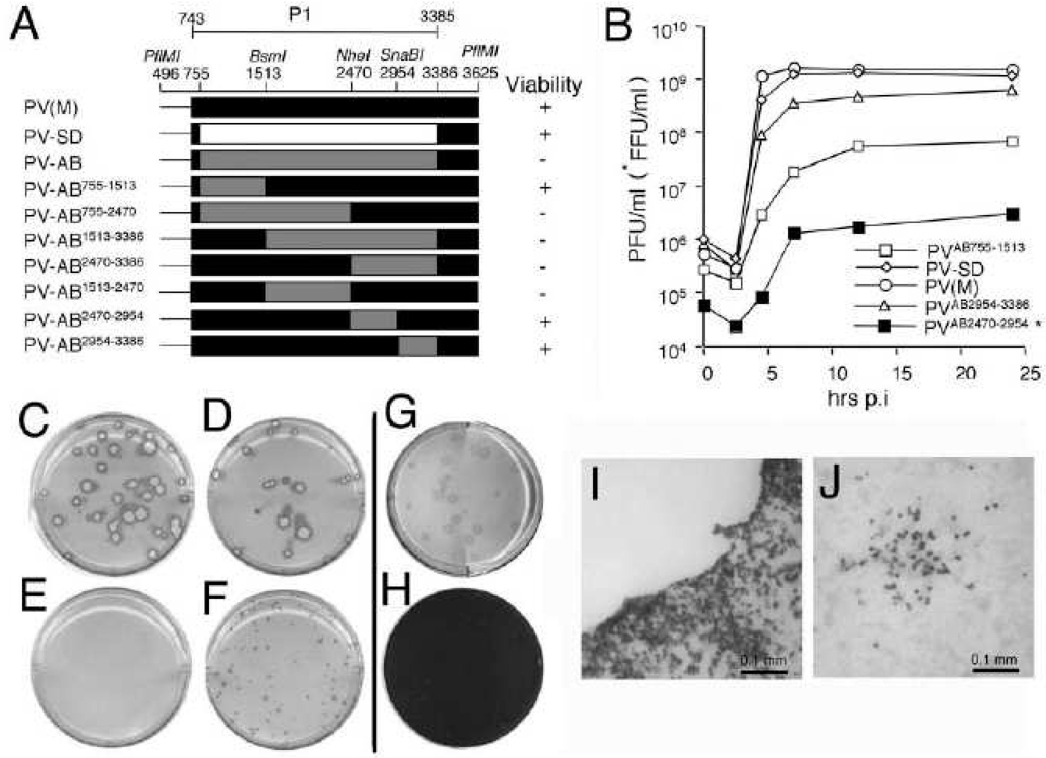


Figure 18: Codon-deoptimized virus phenotypes. (A) Overview of virus constructs used in this study. (B) One-step growth kinetics in HeLa cell monolayers. (C to H) Plaque phenotypes of codon-deoptimized viruses after 48 h (C to F) or 72 h (G and H) of incubation; stained with anti 3Dpol antibody to visualize infected cells. (C) PV(M), (D) PV-SD, (E) PV-AB, (F) PV-AB$^{755-1513}$, (G and H) PV-AB$^{2470-2954}$. Cleared plaque areas are outlined by a rim of infected cells (C and D). (H) No plaques are apparent with PV-AB$^{2470-2954}$ after subsequent crystal violet staining of the well shown in panel G. (I and J) Microphotographs of the edge of an immunostained plaque produced by PV(M) (I) or an infected focus produced by PV-AB$^{2470-2954}$ (J) after 48 h of infection.

Despite 934 single-point mutations in its capsid region, PV-SD replicated at wild-type capacity (Figure 18B) and produced similarly sized plaques as the wild type (Figure 18D). While PV-AB$^{2954-3386}$ grew with near-wildtype kinetics (Figure 18B), PV-AB$^{755-1513}$ produced minute plaques and approximately 22-fold less infectious virus (Figure 18B and F, respectively).

In order to quantify the cumulative effect of a particular codon bias in a protein coding sequence, we calculated a relative codon deoptimization index ($RCDI$), which is a comparative measure against the general codon distribution in the human genome. An $RCDI = 1/codon$ indicates that a gene follows the normal human codon frequencies, while any deviation from the normal human codon bias results in an $RCDI$ higher than 1. We derived the $RCDI$ by the formula $RCDI = [\sum (C_iF_a/C_iF_h) \cdot N_{C_i}]/Ni$ ($i = 1$ through 64). C_iF_a is the observed relative frequency in the test sequence of each codon i out of all synonymous codons for the same amino acid (0 to 1), C_iF_h is the normal relative frequency observed in the human genome of each codon i out of all synonymous codons for that amino acid (0.06 to 1), N_{C_i} is the number of occurrences of that codon i in the sequence, and N is the total number of codons (amino acids) in the sequence. Thus, a high number of rare codons in a sequence results in a higher index. According to this formula, we calculated $RCDI$ values of the various capsid coding sequences of 1.14 for PV(M) and PV-SD which is very close to a normal human distribution. The $RCDI$ values for the AB constructs are 1.73 for PV-AB$^{755-1513}$, 1.45 for PV-AB$^{2470-2954}$, and 6.51 for the parental PV-AB.

For comparison, the $RCDI$ for probably the best known codon-optimized protein, humanized green fluorescent protein (GFP), was 1.31 compared to an $RCDI$ of 1.68 for the original Aequora victoria gfp gene [139]. According to these calculations, a capsid coding sequence with an $RCDI$ of < 2 is associated with a viable virus phenotype, while an $RCDI$ of < 2 (PV-AB 6.51, PV-AB$^{1513-3386}$ 4.04, PV-AB$^{755-2470}$ 3.61) would result in a lethal phenotype.

Codon-deoptimized viruses are deficient at the level of genome translation. Since our synthetic viruses and the wildtype PV(M) are indistinguishable in their protein makeup and no known RNA-based regulatory elements were altered in the modified RNA genomes, these designs allowed us to study the effect of reduced genome translation/replication on attenuation without affecting cell and tissue tropism or immunological properties of the virus. The PV-AB genome was designed under the hypothesis that introduction of many suboptimal codons into the capsid coding sequence should lead to a reduction of genome translation. Since the P1 region is at the N terminus of the polyprotein, synthesis of all downstream nonstructural proteins is determined by the rate of translation through the P1 region. To test whether in fact translation is affected, in vitro translations were performed. Unexpectedly, our initial translations in a standard HeLa-cell based cytoplasmic S10 extract [85] showed no difference in translation capacities for any of the genomes tested. However, as this translation system is optimized for maximal translation, it includes the exogenous addition of excess amino acids and tRNAs, which could conceivably compensate for the genetically engineered codon bias. Therefore, we repeated in vitro translations with a modified HeLa cell extract, which was not dialyzed and in which cellular mRNAs were not removed by micrococcal nuclease treatment. Translations in this extract were performed without the addition of exogenous tRNAs or amino acids. Thus, an environment was created that more closely resembles that in the infected cell, where translation of the PV genomes relies only on cellular supplies while competing for resources with cellular mRNAs. Due to the high background translation from cellular mRNA and the low [35S] Met incorporation rate in nondialyzed extract, a set of virus-specific translation products were detected by Western blotting with anti-2C antibodies [97]. These modified conditions resulted in dramatic reduction of translation efficiencies of the modified genomes which correlated with the extent of the deoptimized sequence. Whereas translation of PV-SD was comparable to that of the wildtype, translation of three noninfectious genomes, PV-AB, PV-AB$^{1513-3386}$, and PV-AB$^{755-2470}$, was reduced by approximately 90%.

7.3 Conclusions

In this chapter, we have demonstrated the utility of large-scale codon deoptimization of poliovirus capsid coding sequences by de novo gene synthesis for the generation of attenuated viruses. It was our initial goal to explore the potential of this technology as a tool for generating live attenuated virus vaccines. However, we found codon-deoptimized viruses to be marked by a very low specific infectivity. In addition, our intention to design PV capsid encoding with a synonymous codon bias that specifically discriminated against expression in the central nervous system did not bear fruit, as the tissue-specific differences in codon bias described by others [98], if at all significant, are too small to bring about a tissue-restrictive virus phenotype. In a larger set of brain-specific genes than the one used by Plotkin and colleagues [98] in their calculations, we could not detect any appreciable tissue-specific codon bias (data not shown). These observations may pose hurdles to using this new technology to develop codon-deoptimized viruses as candidates for live attenuated vaccines, although much more work has to be carried out. Fine-tuning of codon deoptimization may still allow the alteration of tissue tropism and virulence required for attenuation. On the other hand, codon de-optimized viruses produced similar amounts of progeny per cell, while being 2 to 3 orders of magnitude less infectious. Such viruses may prove very useful as safer alternatives in the production of inactivated poliovirus vaccine. Since they are 100% identical to the wildtype virus at the protein level, an identical immune response in hosts who received inactivated virus is guaranteed. This may be of great advantage at the stage of licensing any potential vaccine based on this strategy. Due to the distribution effect of many silent mutations over large genome segments, codondeoptimized viruses should prove extremely genetically stable. Although long-term passaging experiments are still under way, no faster-replicating escape variants have been isolated from either PV-AB$^{2470-2954}$ or PV-AB$^{755-1513}$ after five passages, as assessed by the absence of faster-growing or large-plaque revertants (data not shown).

In the infectious cycle of AB-type viruses described here, steps 1 and 2 should be identical to a PV(M) infection as their capsids are identical. Likewise, identical 5' nontranslated regions should perform equally well in assembly of a translation complex (step 3). Viral polyprotein translation, on the other hand (step 4), is severely debilitated due to the introduction of a great number of suboptimal synonymous codons in the capsid region. It is thought that the repeated encounter of rare codons by the translational machinery causes stalling of the ribosome as by the laws of mass action rare aminoacyl-tRNA will take longer to diffuse into the A site on the ribosome. As peptide elongation to a large extent is driven by the concentration of available aminoacyl-tRNA, dependence of an mRNA on many rare tRNAs, consequently, lengthens the time of translation [47]. Alternatively, excessive stalling of the ribosome may cause premature dissociation of the translation complex from the RNA and result in a truncated protein destined for degradation. Both processes lead to a lower protein synthesis rate per mRNA molecule per unit of time. While our data presented here suggest that the phenotypes of codon-deoptimized viruses are determined by the rate of genome translation, other mechanistic explanations may be possible. It has been suggested that the conserved positions of rare synonymous codons throughout the viral capsid sequence in the hepatitis A virus are of functional importance for the proper folding of the nascent polypeptide by introducing necessary translation pauses [107]. Large-scale alteration of the codon composition may therefore conceivably change some of these pause sites to result in an increase of misfolded capsid proteins. Whether these considerations also apply to the PV capsid is not clear. If so, we would have expected a phenotype with our PV-SD design, in which the wildtype codons were preserved but their positions throughout the capsid were completely changed: that is, none of the purported pause sites would be at the appropriate position with respect to the protein sequence. No phenotype, however, was observed and PV-SD translated and replicated at wildtype levels. Another possibility is that the large-scale codon alterations in our designs may create fortuitous dominant-negative RNA elements, such as stable secondary structures, or sequences that may undergo disruptive long-range interactions with other regions of the genome.

The near-identical production of particles per cell by codon de-optimized viruses indicates that the total of protein produced after extended period of times is not severely affected, whereas the rate of protein production has been drastically reduced. This is reflected in the delayed appearance of CPE, which may be a sign that the virus has to go through more RNA replication cycles to build up similar intracellular virus protein concentrations. It appears that codon-deoptimized viruses are severely handicapped in establishing a productive infection because the early translation rate of the incoming infecting genome is reduced. As a result of this lower translation rate, poliovirus proteins essential for disabling the cell s antiviral responses (most likely proteinases 2Apro and 3Cpro) are not synthesized at sufficient amounts to pass this crucial hurdle in the life cycle quickly enough. Consequently, there is a better chance for the cell to eliminate the infection before viral replication could unfold and take over the cell. Thus, the chance for productive infection events is reduced and the rate of abortive infection is increased. However, in the case where a codon-deoptimized virus does succeed in disabling the cell, this virus will produce nearly identical amounts of progeny to the wild type. Our data suggest that a fundamental difference may exist between early translation (from the incoming RNA genome) and late translation during the replicative phase, when the cell s own translation is largely shut down. Although this may be a general phenomenon, it might be especially important in the case of codon-deoptimized genomes. Host cell shutoff very likely results in an overabundance of free aminoacyltRNAs which may overcome the imposed effect of the suboptimal codon usage as the PV genomes no longer have to compete with cellular RNAs for translation resources. This, in fact, may be analogous to our observations with the modified in vitro translation system described above. Here, in the translation extract, which was not nuclease treated (and thus contained the cellular mRNAs) and was not supplemented with exogenous amino acids or tRNA, clear differences were observed in the translation capacity of different capsid design mutants. Under these conditions, viral genomes have to compete with cellular mRNAs in an environment where supplies are limited. On the other hand, in the traditional translation extract, in which endogenous mRNAs were removed and tRNAs and amino acids were supplemented in excess, all PV RNAs translated equally well regardless of codon bias. These two different in vitro conditions may be analogous to in vivo translation during the early and late phases in the PV-infected cell.

Since the dramatic effect of codon bias on poliovirus fitness could not be predicted, it should be possible in future designs to make less severe codon changes distributed over a larger number of codon sequences. This should continue to improve the genetic stability of the individual phenotypes and improve their potential as vaccine candidates.

Chapter 8

Codon Pair Bias Designs

8.1 Introduction

Genes are DNA sequences, meaning chains of desoxyribonucleic acids, which can be represented as strings on a four-letter alphabet. These strings describe protein sequences, which are chains of amino acids, and can in turn be represented as strings on a 20-letter alphabet. Every triplet of DNA sequence characters maps to a single amino-acid in the protein sequence, creating the *triplet code*. In this code, all $4^3 = 64$ possible DNA strings of length three (called *codons*) map to elements of the 20-letter protein alphabet, with most of the amino acid letters being encoded by more than one triplet, up to a maximum of six.

In most organisms, there exists a distinct *codon bias*, which describes the preferences of amino acids being encoded by particular codons more often than others. The genetic information of every organism has a preference in codon usage for amino acids that are encoded by multiple codons. The codon frequency distribution of the codons used to encode proteins is characteristic and often can be used to distinguish between organisms, groups of organisms, but also different tissues, when analyzed specifically for RNA preferencially overexpressed in different cell types. It is widely believed that codon bias is connected to protein translation rates, as been described and demonstrated in the previous chapter.

In addition to codon bias, each species has specific preferences as to whether given pairs of codons appear adjacent to each other in coding sequences, in an ordered manner, something that is called *codon-pair bias*.

8.2 Codon Pair Score

To quantify codon pair bias, we define a *codon pair distance* as the log ratio of the observed over the expected number of occurences (frequency) of codon pairs in the known coding sequences of an organism. Although the calculation of the observed frequency of codon pairs in a set of genes is straightforward, the expected frequency of a codon pair is calculated to as to be independent of amino acid and codon bias, following the paradigm of Fedorov et al. [36] and enhancing Gutman and Hatfield's approach [48]. To achieve independence from amino acid and codon bias, the expected frequency is calculated based on the relative proportion of the number of times an amino acid is encoded by a specific codon. In short:

$$\textit{codon pair score} \ = \log\left(\frac{F(AB)}{\frac{F(A)\times F(B)}{F(X)\times F(Y)} \times F(XY)}\right), \qquad (1)$$

where the codon pair AB encodes for amino acid pair XY and F denotes frequency (number of occurences).

The logarithm provides specific properties to the score, such as providing a positive/negative attribute to over- and under-represented codon pairs respectively, as well as equalizing the distance from the mean of two scores resulting from equal percentage deviations of the observed from the expected frequencies. For example, two codon pairs, one having an observed/expected ratio of 2 and another having a ratio of 1/2, will be equidistant from 0, having opposite values. In addition, this scoring scheme results in an expected arithmetic mean of 0 for all the codon pair scores, assuming the values fluctuate randomly.

A drawback in our score definition is encountered in situations where the observed frequencies do not deviate significantly from their expected values. In such case, an increase/decrease of the observed value of a codon pair over the expected value results in dramatically greater changes for small expected values than for large ones. By this effect, codon pairs expected to be encountered a few times influence the score significantly more than frequent ones.

In practice, we have found that the zero expected average (which was verified through simple computational experiment on random amino-acid and codon generated sets) of our scoring scheme is not encountered in the species examined, all of which seem have negative score averages for the unweighted set of codon pairs. This result could indicate that small expected values in codon pair frequencies are paired with even smaller observed values for the same codon pairs.

For every organism we examine, we create a codon pair score table of all possible codon pairs, excluding the stop codon (for statistical reasons for which most other studies have excluded it as well). This results in vectors of 3721 (61^2) codon pair scores. In order to calculate the score of a random codon pair, we use the values in the vector of a specific organism, which we will consider our reference organism. For most of our experiments, our reference organism is human, and is implied if not explicitely mentioned. Any m-residue protein or coding region can be rated by the arithmetic mean of the scores of all codon pairs that comprise its encoding.

		Random Runs		Wildtype Sequence		Gradient Descent	
Organism/ Gene	length (codons)	Mean Score per base (x10^3)	Mean StdDev	Score per base (x10^3)	SD dist.	Max SD dist.	Min SD dist.
Human hemoglobin	148	-68.24	5.03	85.41	4.52	8.00	-8.67
Human HV1 minor capsid region	677	-18.83	10.40	-33.56	-0.96	22.77	-22.88
Poliovirus WT1 capsid region	881	-81.77	12.81	-34.26	3.27	25.86	-28.97
Encephalomyocarditis virus	2293	-78.69	20.77	-3.86	8.26	41.95	-48.63
Human Enterovirus A	2194	-83.12	20.91	-6.00	8.09	40.36	-47.52
Human Rhinovirus A	2165	-3.19	16.99	49.89	6.76	37.18	-45.26
Human Rhinovirus B	2180	-26.69	18.17	44.55	8.55	37.08	-45.99
SARS coronavirus	7074	-39.24	33.57	23.76	13.28	65.88	-79.61
Hepatitis B *	844	-81.18	12.03	8.07	6.26	27.50	-29.95
Hepatitis C	3012	-86.34	24.27	-28.85	7.13	51.79	-54.61
HIV1 *	5444	19.46	61.41	81.61	5.51	10.32	-3.42
Poliovirus WT1	2210	-75.58	19.66	-17.68	6.51	40.56	-47.24
Cactus virus X	2064	-80.49	19.14	-6.58	7.97	41.73	-43.22
Potato virus A	3060	-58.93	22.04	5.79	8.99	46.06	-55.30
Salmon pancreas disease virus	3923	-98.99	28.61	-72.08	3.69	57.97	-61.91
Bacteriophage AP205 *	1294	-76.96	15.70	-60.91	1.32	30.95	-36.07
Chlamydia phage 2 *	1369	-26.67	13.05	-9.96	1.75	30.91	-36.09
Coliphage phiK *	1693	-56.47	16.89	-20.97	3.56	32.99	-38.50

Table 15: Codon pair bias statistics for selected genes and organisms
* Including only a non-overlapping subset of genes

In examining the extend and characteristics of codon pair bias, we conducted computational experiments with a collection of human coding sequences from the Consensus CDS (CCDS) database, which consists of the core set of human protein coding regions that are consistently annotated and of high quality. We downloaded the March 2005 dataset, up to date as of August 18, 2005, containing a total of 14,795 coding sequences and 13,142 genes, representing more than half of the currently believed human gene number.

8.3 Gene and Organism Codon Pair Bias

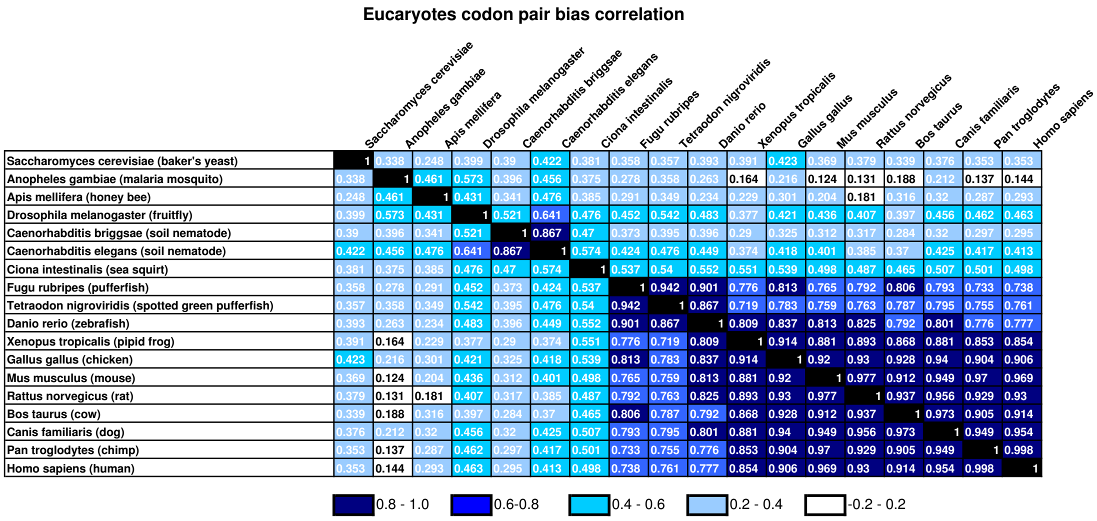
In order to examine the significance of codon pair bias, we performed a series of computational experiments, over a number of different randomly selected genes and organisms. The initial set is composed of human, human viral, other viral, and phage genes.

Since the absolute scoring value of a gene or set of genes of an organism does not provide enough information on the use of favored or unfavored codon pairs in the human genome, for each gene/organism we examined we generated a large number of random permutations of the codons comprising the gene or gene set (when considering a genome). Random genes were generated by permuting the locations of all codons of each corresponding amino-acid, thus creating an equivalent gene (translated to the same protein), having exactly the same codon distribution. The only alterations are the relative locations of the codons, which result in random codon pair associations.

For each of the "scrambled" genes (or gene sets) with altered codon pairs, we calculated the gene scores and for all of the random generations we computed the mean and standard deviation of their scores. This provides with an indication of the expected codon pair score and extend of its deviation. We then calculated the codon pair score of the original coding sequence of each gene and the number of standard deviations from the expected mean, calculated from the random permutations. The latter can also be used to calculate the probability of the codon pairs being selected with bias towards over- or under-represented codon pairs in our reference organism. For example, a sequence with a positive standard deviation value of 3 from the mean would indicate a probability of 0.0013 that the use of over-represented codon pairs occured by chance. Similarly, a positive value of 8 standard deviations from the mean makes the probability that codon pair bias is coincidental a mere 6.11×10^{-16}.

In Table 8.3 we can observe that most genes in human viruses as well as other viruses are predominantly using codon pairs over-represented in human coding regions. This bias is less pronounced in other viruses and especially in phages, as one would expect. Still, the tendency towards the usage of over-represented human codon pairs may indicate the existence of patterns that are generally selected for or against in a wide variety of unrelated organisms.

In most of our computational experiments we observed standard deviations from the mean which indicate strong biases. But what is the maximum and minimum possible deviations that a scrambled gene could achieve? To answer this question, we performed gradient descent experiments, by randomly exchanging equivalent codons in our sequences for a large number of iterations (in the range of 10^6), accepting transitions towards the optimizing direction we select (either maximizing or minimizing the total score). We also performed simulated annealing simulations, which improved the lower and upper bounds only slightly in the expense of being computationally expensive and difficult to parameterize. This verified the accuracy of the limits achieved by gradient descent. These values can be found in Table 8.3.



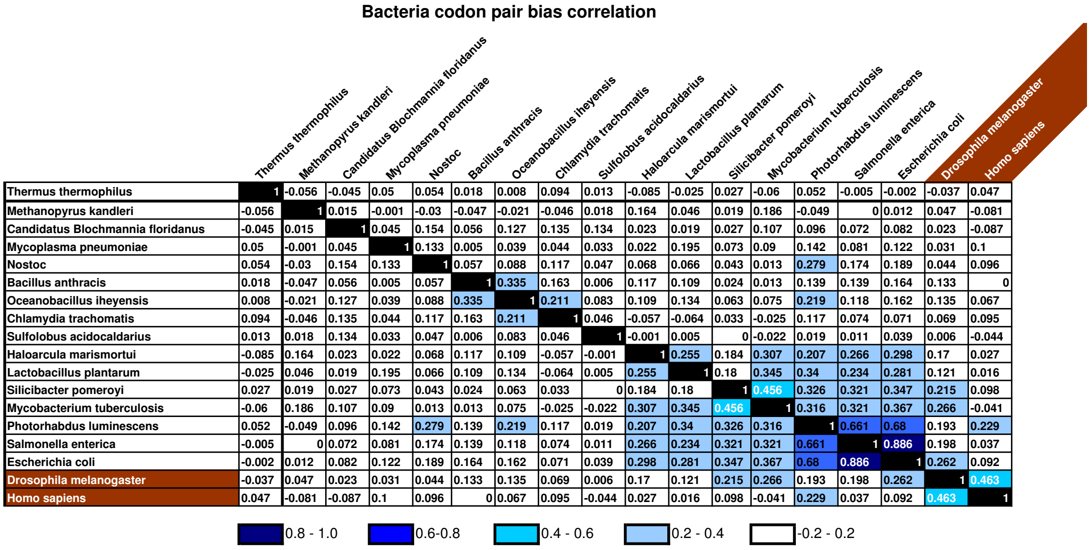
Eucaryotes codon pair bias correlation

	Saccharomyces cerevisiae	Anopheles gambiae	Apis mellifera	Drosophila melanogaster	Caenorhabditis briggsae	Caenorhabditis elegans	Ciona intestinalis	Fugu rubripes	Tetraodon nigroviridis	Danio rerio	Xenopus tropicalis	Gallus gallus	Mus musculus	Rattus norvegicus	Bos taurus	Canis familiaris	Pan troglodytes	Homo sapiens
Saccharomyces cerevisiae (baker's yeast)	1	0.338	0.248	0.399	0.39	0.422	0.381	0.358	0.357	0.393	0.391	0.423	0.369	0.379	0.339	0.376	0.353	0.353
Anopheles gambiae (malaria mosquito)	0.338	1	0.461	0.573	0.396	0.456	0.375	0.278	0.358	0.263	0.164	0.216	0.124	0.131	0.188	0.212	0.137	0.144
Apis mellifera (honey bee)	0.248	0.461	1	0.431	0.341	0.476	0.385	0.291	0.349	0.234	0.229	0.301	0.204	0.181	0.316	0.32	0.287	0.293
Drosophila melanogaster (fruitfly)	0.399	0.573	0.431	1	0.521	0.641	0.476	0.452	0.542	0.483	0.377	0.421	0.436	0.407	0.397	0.456	0.462	0.463
Caenorhabditis briggsae (soil nematode)	0.39	0.396	0.341	0.521	1	0.867	0.47	0.373	0.395	0.396	0.29	0.325	0.312	0.317	0.284	0.32	0.297	0.295
Caenorhabditis elegans (soil nematode)	0.422	0.456	0.476	0.641	0.867	1	0.574	0.424	0.476	0.449	0.374	0.418	0.401	0.385	0.37	0.425	0.417	0.413
Ciona intestinalis (sea squirt)	0.381	0.375	0.385	0.476	0.47	0.574	1	0.537	0.54	0.552	0.551	0.539	0.498	0.487	0.465	0.507	0.501	0.498
Fugu rubripes (pufferfish)	0.358	0.278	0.291	0.452	0.373	0.424	0.537	1	0.942	0.901	0.776	0.813	0.765	0.792	0.806	0.793	0.733	0.738
Tetraodon nigroviridis (spotted green pufferfish)	0.357	0.358	0.349	0.542	0.395	0.476	0.54	0.942	1	0.867	0.719	0.783	0.759	0.763	0.787	0.795	0.755	0.761
Danio rerio (zebrafish)	0.393	0.263	0.234	0.483	0.396	0.449	0.552	0.901	0.867	1	0.809	0.837	0.813	0.825	0.792	0.801	0.776	0.777
Xenopus tropicalis (pipid frog)	0.391	0.164	0.229	0.377	0.29	0.374	0.551	0.776	0.719	0.809	1	0.914	0.881	0.893	0.868	0.881	0.853	0.854
Gallus gallus (chicken)	0.423	0.216	0.301	0.421	0.325	0.418	0.539	0.813	0.783	0.837	0.914	1	0.92	0.93	0.928	0.94	0.904	0.906
Mus musculus (mouse)	0.369	0.124	0.204	0.436	0.312	0.401	0.498	0.765	0.759	0.813	0.881	0.92	1	0.977	0.912	0.949	0.97	0.969
Rattus norvegicus (rat)	0.379	0.131	0.181	0.407	0.317	0.385	0.487	0.792	0.763	0.825	0.893	0.93	0.977	1	0.937	0.956	0.929	0.93
Bos taurus (cow)	0.339	0.188	0.316	0.397	0.284	0.37	0.465	0.806	0.787	0.792	0.868	0.928	0.912	0.937	1	0.973	0.905	0.914
Canis familiaris (dog)	0.376	0.212	0.32	0.456	0.32	0.425	0.507	0.793	0.795	0.801	0.881	0.94	0.949	0.956	0.973	1	0.949	0.954
Pan troglodytes (chimp)	0.353	0.137	0.287	0.462	0.297	0.417	0.501	0.733	0.755	0.776	0.853	0.904	0.97	0.929	0.905	0.949	1	0.998
Homo sapiens (human)	0.353	0.144	0.293	0.463	0.295	0.413	0.498	0.738	0.761	0.777	0.854	0.906	0.969	0.93	0.914	0.954	0.998	1

■ 0.8 - 1.0 ■ 0.6-0.8 ■ 0.4 - 0.6 ■ 0.2 - 0.4 □ -0.2 - 0.2

Figure 19: Selected eukaryote codon pair bias score correlation.

In another effort to investigate the extend and significance of codon pair bias, we created codon pair score vectors for a variety of organisms, from both



Bacteria codon pair bias correlation

	Thermus thermophilus	Methanopyrus kandleri	Candidatus Blochmannia floridanus	Mycoplasma pneumoniae	Nostoc	Bacillus anthracis	Oceanobacillus iheyensis	Chlamydia trachomatis	Sulfolobus acidocaldarius	Haloarcula marismortui	Lactobacillus plantarum	Silicibacter pomeroyi	Mycobacterium tuberculosis	Photorhabdus luminescens	Salmonella enterica	Escherichia coli	Drosophila melanogaster	Homo sapiens
Thermus thermophilus	1	-0.056	-0.045	0.05	0.054	0.018	0.008	0.094	0.013	-0.085	-0.025	0.027	-0.06	0.052	-0.005	-0.002	-0.037	0.047
Methanopyrus kandleri	-0.056	1	0.015	-0.001	-0.03	-0.047	-0.021	-0.046	0.018	0.164	0.046	0.019	0.186	-0.049	0	0.012	0.047	-0.081
Candidatus Blochmannia floridanus	-0.045	0.015	1	0.045	0.154	0.056	0.127	0.135	0.134	0.023	0.019	0.027	0.107	0.096	0.072	0.082	0.023	-0.087
Mycoplasma pneumoniae	0.05	-0.001	0.045	1	0.133	0.005	0.039	0.044	0.033	0.022	0.195	0.073	0.09	0.142	0.081	0.122	0.031	0.1
Nostoc	0.054	-0.03	0.154	0.133	1	0.057	0.088	0.117	0.047	0.068	0.066	0.043	0.013	0.279	0.174	0.189	0.044	0.096
Bacillus anthracis	0.018	-0.047	0.056	0.005	0.057	1	0.335	0.163	0.006	0.117	0.109	0.024	0.013	0.139	0.139	0.164	0.133	0
Oceanobacillus iheyensis	0.008	-0.021	0.127	0.039	0.088	0.335	1	0.211	0.083	0.109	0.134	0.063	0.075	0.219	0.118	0.162	0.135	0.067
Chlamydia trachomatis	0.094	-0.046	0.135	0.044	0.117	0.163	0.211	1	0.046	-0.057	-0.064	0.033	-0.025	0.117	0.074	0.071	0.069	0.095
Sulfolobus acidocaldarius	0.013	0.018	0.134	0.033	0.047	0.006	0.083	0.046	1	-0.001	0.005	0	-0.022	0.019	0.011	0.039	0.006	-0.044
Haloarcula marismortui	-0.085	0.164	0.023	0.022	0.068	0.117	0.109	-0.057	-0.001	1	0.255	0.184	0.307	0.207	0.266	0.298	0.17	0.027
Lactobacillus plantarum	-0.025	0.046	0.019	0.195	0.066	0.109	0.134	-0.064	0.005	0.255	1	0.18	0.345	0.34	0.234	0.281	0.121	0.016
Silicibacter pomeroyi	0.027	0.019	0.027	0.073	0.043	0.024	0.063	0.033	0	0.184	0.18	1	0.456	0.326	0.321	0.347	0.215	0.098
Mycobacterium tuberculosis	-0.06	0.186	0.107	0.09	0.013	0.013	0.075	-0.025	-0.022	0.307	0.345	0.456	1	0.316	0.321	0.367	0.266	-0.041
Photorhabdus luminescens	0.052	-0.049	0.096	0.142	0.279	0.139	0.219	0.117	0.019	0.207	0.34	0.326	0.316	1	0.661	0.68	0.193	0.229
Salmonella enterica	-0.005	0	0.072	0.081	0.174	0.139	0.118	0.074	0.011	0.266	0.234	0.321	0.321	0.661	1	0.886	0.198	0.037
Escherichia coli	-0.002	0.012	0.082	0.122	0.189	0.164	0.162	0.071	0.039	0.298	0.281	0.347	0.367	0.68	0.886	1	0.262	0.092
Drosophila melanogaster	-0.037	0.047	0.023	0.031	0.044	0.133	0.135	0.069	0.006	0.17	0.121	0.215	0.266	0.193	0.198	0.262	1	0.463
Homo sapiens	0.047	-0.081	-0.087	0.1	0.096	0	0.067	0.095	-0.044	0.027	0.016	0.098	-0.041	0.229	0.037	0.092	0.463	1

■ 0.8 - 1.0 ■ 0.6-0.8 ■ 0.4 - 0.6 ■ 0.2 - 0.4 □ -0.2 - 0.2

Figure 20: Selected bacteria codon pair bias score correlation.

eykaryotes and prokaryotes. Then we correlated the vectors using Pearson correlation. As seen in Figures 19 and 20, based only on codon pair score correlation, the phylogenetic tree can be reconstructed with relative accuracy. Codon pair bias seem to be preserved among closely related species and the wider spread in the range of prokaryote similarity as compared to the eukaryotes is also in accord with the timeframe of species divergence in these two kingdoms.

8.4 Codon Pair Bias in Human Viruses

Our specific interests in modifying viruses, and specifically human pathogens, in order to attenuate them, led to a series of computational experiments on the codon pair bias of these viruses. We analyzed coding regions of all human viruses whose full genome has been sequenced before May 2005, grouping them under a number of categories, as seen in table 8.4. The “other viruses” group

Group	Number of genes	Average gene Standard Deviation distance from mean
All Human virus genes	1562	1.462
Herpes virus genes	1016	0.831
Non-Herpes virus genes	546	2.640
Adeno virus genes	201	2.327
Papilloma virus genes	197	2.319
Genes from other viruses	148	3.493

Table 16: Average standard deviation distances from mean of human viruses groups.

includes genes from entero-, lymphotropic-, rhino-, HIV, foamy, respiratory syncytial, spumaretro, parainfluenza, parecho-, astro-, corono-, metapneumo-, erythro-, parvo-, and picobirna-viruses.

The set of all virus genes in aggregate does not have an excessively strong bias towards human over-represented codon pairs, although most of this effect can be attributed to the large number of herpes virus genes. Herpes is a special case of superinfecting viruses, infecting almost every eukaryote, including fungi. In Figure 8.4 we can see a histogram of all CCDS genes, as well as genes from herpes- and adeno-viruses.

8.5 Optimization of a Gene Encoding based on Codon Pair Bias

To examine the effects of codon pair bias on mRNA translation of specific proteins, we decided to alter the codon positions, exchanging synonymous codons, while keeping the same codon distribution and of course the same amino acid chain. So we define the following problem: Given an amino acid sequence and a set of codon frequencies (codon distribution), change the DNA encoding of the sequence such that the codon pair score is optimized (usually

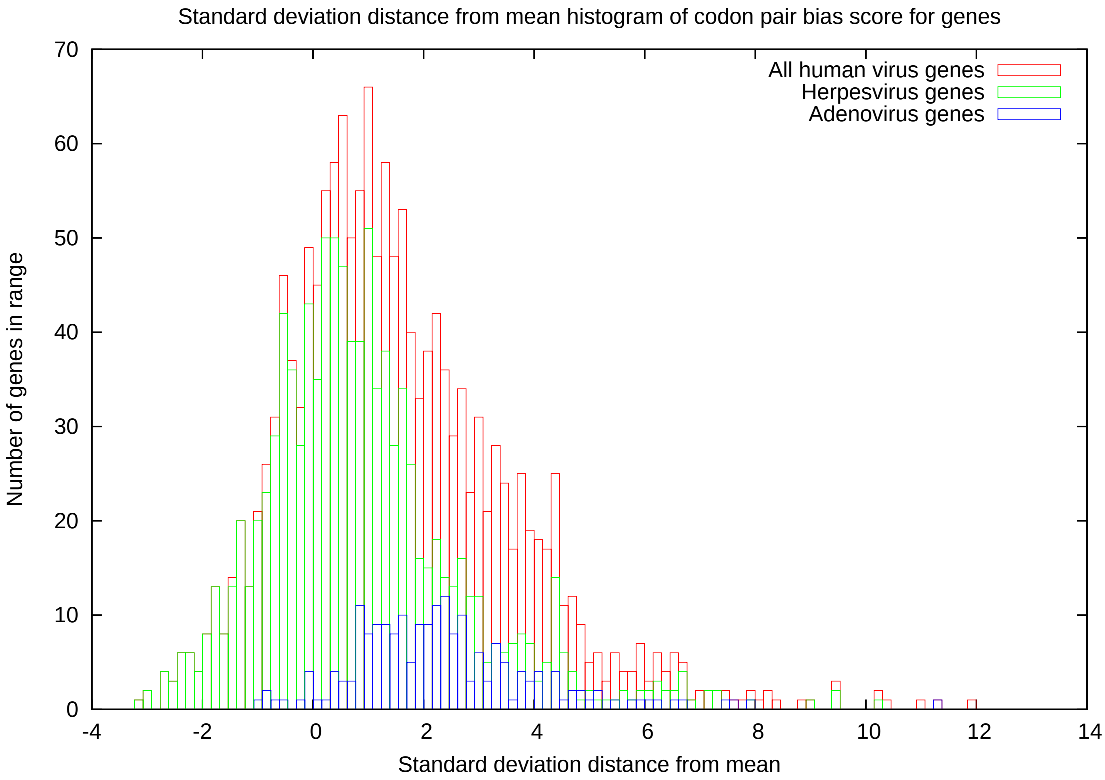


Figure 21: Human viruses gene standard deviations from mean histogram.

minimized or maximized).

Our problem, as defined above, can be associated with the *Traveling Salesman Problem* (TSP), where a salesman has to visit each of a given set of cities, driving the minimum distance possible while visiting each city only once. The traveling salesman problem is one of the most notorious NP-complete problems, which is a function of its general usefulness and ease of description.

Almost any flavor of TSP is going to be NP-complete, so the right way to proceed is with heuristics. These are often quite successful, typically coming within a few percent of the optimal solution, which is close enough for most applications and in particular for our optimized encoding.

Our problem is associated with the problem of finding a traveling salesman path (not tour) under a 64-country metric. In this formulation, each of the 64 possible codons is analogous to a country, and the codon multiplicity modeled as the number of cities in the country. The codon-pair bias measure is reflected as the country distance matrix.

The real biological problem of the design of genes encoding specific proteins using a given set of codon multiplicities so as to optimize the gene/DNA sequence under a codon-pair bias measure is slightly differerent. What is missing in our model in the country TSP model is the need to encode *specific protein sequences.* The DNA triplet code partitions the 64 codons into 21 equivalence classes (coding for each of the 20 possible amino acids and a stop symbol). Any given protein/amino acid sequence can be specified by picking an arbitrary representative of the associated codon equivalence class to encode it.

Since the number of amino acids and codons is fixed in our problem, there actually exists a dynamic programming algorithm that can solve it in $O(n^{65})$ time and space, where n is the protein length (in amino-acid residues). To achieve these bounds, the algorithm progresses iteratively through each residue of the protein, keeping the best possible score for each ending residue, based on all possible codon distributions for the positions encountered so far. This time and space complexity can be slightly improved by restricting the degrees of freedom of the residues, keeping in mind that the sum of the codons in the distributions kept are constant as a total, as well as grouped under their corresponding amino-acids. Even by these reductions, the space and time complexities are prohibitive for any real protein length design.

Because of the special restrictions and the nature of our problem, as well as its adaptability to application of additional criteria in the optimization, we selected the *simulated annealing* heuristic to optimize sequences. A general description of the technique can be found in [113].

Our simulated annealing algorithm works as follows:

1. We initially create a random assignment of the codons in their respective

amino acid allowed positions.

2. We calculate the codon pair score of the coding sequence from the initial assignment.

3. We randomly exchange pairs of codons encoding for the same amino acid, according to the simulated annealing optimization function.

4. We repeat the previous step until no change is observed for a specific number of steps.

For best results, we adjusted the simulated annealing parameters, including the temperature, the number of repetitions at each energy level and the constant k. The performance of simulated annealing was compared to the *gradient descent* method results (which is similar to simulated annealing, but taking no backward steps), which verified the improvement.

8.6 Splice Sites and Secondary Structure

In all codon pair bias designs, we attempted an elimination of donor (3') splice sites. This was achieved by specifically targeting the consensus sequence $CAG|G$, appearing in the vicinity of the donor splice site, with synonymous changes, depending on the position of the wobble base. In the few cases where such changes were not possible, other upstream synonymous changes that reduce the probability of occurence of a donor splice site were applied. These changes were not compensated by complementary changes in other places to keep the same codon distribution, since they would most probably alter negatively the codon pair objective score. The locations of donor splice sites and the confirmation of their elimination were predicted with two neural network splice site prediction services, NetGene2 Server [51, 105, 16] and NNSPLICE

[82, 5]. We consider a splice site eliminated if both services would not predict its appearance with probability higher than 20%.

To ensure that strong secondary structures do not affect translation efficiency, we scanned the capsid region of our designs using the program mfold [31, 78]. We concentrated our search on 100 bases long segments, overlapping with each other every 20 bases. Any segments with lower binding energy than a threshold of $-30Kcal/mol$ would incur random synonymous substitutions at $C-G$ binding locations, such that the binding energy of the segment could be elevated. The synonymous changes would be selected in such a way that the codon pair bias objective would be satisfied as well. Nevertheless, only a few changes resulted in an optimized score over the original selections, since favorable codon pairs had been already selected by the simulated annealing optimization algorithm. The codon distribution was affected as well, but only minimally.

Bibliography

[1] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–10, 1990.

[2] P. Ball. Starting from scratch. *Nature*, 431:624–626, 2004.

[3] M. Begon, J. L. Harper, and C. R. Townsend. *Ecology: Individuals, populations, and communities cambridge.* MA: Blackwell Science Ltd, 1996.

[4] O. Beja, M. T. Suzuki, E. V. Koonin, L. Aravind, A. Hadd, L. P. Nguyen, R. Villacorta, M. Amjadi, C. Garrigues, S. B. Jovanovich, R. A. Feldman, and E. F. DeLong. Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ Microbiol*, 2(5):516–29, 2000.

[5] Berkeley Drosophila Genome Project. NNSPLICE 0.9. http://www.fruitfly.org/seq_tools/splice.html.

[6] B. Blaisdell, A. Campbell, and S. Karlin. Similarities and dissimilarities of phage genomes. *Proc. Natl. Acad. Sci.*, 93:5854–5859, 1996.

[7] C. E. Bonferroni. Il calcolo delle assicurazioni su gruppi di teste. *Studi in Onore del Professore Salvatore Ortu Carboni*, pages 13–60, 1935.

[8] M. Breitbart, B. Felts, S. Kelley, J. M. Mahaffy, J. Nulton, P. Salamon, and F. Rohwer. Diversity and population structure of a near-shore marine-sediment viral community. *Proc Biol Sci*, 271(1539):565–74, 2004.

[9] M. Breitbart, I. Hewson, B. Felts, J. M. Mahaffy, J. Nulton, P. Salamon, and F. Rohwer. Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol*, 185(20):6220–3, 2003.

[10] M. Breitbart, P. Salamon, B. Andresen, J. M. Mahaffy, A. M. Segall, D. Mead, F. Azam, and F. Rohwer. Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A*, 99(22):14250–5, 2002.

[11] K. Brugirard-Ricaud, A. Givaudan, J. Parkhill, N. Boemare, F. Kunst, R. Zumbihl, and E. Duchaud. Variation in the effectors of the type III secretion system among Photorhabdus species as revealed by genomic analysis. *J Bacteriol*, 186:4376–4381, 2004.

[12] A. J. Cann. *principles of molecular virology.* Academic Press, 1993.

[13] A. J. Cann, S. E. Fandrich, and S. Heaphy. Analysis of the virus population present in equine faeces indicates the presence of hundreds of uncharacterized virus genomes. *Virus Genes*, 30(2):151–6, 2005.

[14] S. Casjens. The diverse and dynamic structure of bacterial genomes. *Annu. Rev. Genet.*, 32:339–77, 1998.

[15] J. Cello, A. V. Paul, and E. Wimmer. Chemical synthesis of poliovirus cDNA: generation of infectious virus in the absence of natural template. *Science*, 297:1016–1018, 2002.

[16] Center for Biological Sequence Analysis (CBS). NetGene2 server. http://www.cbs.dtu.dk/services/NetGene2/.

[17] R. Ceulemans and M. Mousseau. Tansley review no 71: Effects of elevated CO_2 on woody plants. *New Phyt*, 127:425–446, 1994.

[18] A. Chao. Nonparametric estimation of the number of classes in a population. *Scand. J. Stat.*, 11, 1984.

[19] A. Chao. Estimating the population size for capture-recapture data with unequal catchability. *Global Change Biol*, 43:783–791, 1987.

[20] C. Chapus, C. Dufraigne, A. Giron, S. Edwards, and B. Fertil. Exploration of phylogenetic relationships using DNA style. *Recomb2003*, 2003.

[21] K. Chen and L. Pachter. Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Comput Biol*, 1(2):106–12, 2005.

[22] S. Chen, A. Zhang, L. B. Blyn, and G. Storz. MicC, a second small-RNA regulator of Omp protein expression in Escherichia coli. *J Bacteriol*, 186:6689–6697, 2004.

[23] H. Chung, D. R. Zak, and E. A. Lilleskov. Fungal community composition and metabolism under elevated CO_2 and O_3. *Oecologia*, 147:143–154, 2006.

[24] B. Cohen and S. Skiena. Natural selection and algorithmic design of mRNA. *J. Computational Biology*, 10:419–432, 2003.

[25] J. R. Cole, B. Chai, T. L. Marsh, R. J. Farris, Q. Wang, S. A. Kulam, S. Chandra, D. M. McGarrell, T. M. Schmidt, G. M. Garrity, and J. M.

Tiedje. The ribosomal database project (rdp-ii): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic Acids Res*, 31(1):442–3, 2003.

[26] J. Coyer, J. Andersen, S. A. Forst, M. Inouye, and N. Delihas. micF RNA in ompB mutants of Escherichia coli: different pathways regulate micF RNA levels in response to osmolarity and temperature change. *J Bacteriol*, 172:4143–4150, 1990.

[27] P. S. Curtis and X. Wang. A meta-analysis of elevated CO2 effects on woody plant mass, form and physiology. *Oecologia*, 113:299–313, 1998.

[28] N. Delihas. Annotation and evolutionary relationships of a small regulatory RNA gene micF and its target ompF in Yersinia species. *BMC Microbiol*, 3:13, 2003.

[29] N. Delihas and S. Forst. micF : an antisense RNA gene involved in response of Escherichia coli to global stress factors. *J Mol Biol*, 313:1–12, 2001.

[30] P. Deschavanne, A. Giron, J. Vilain, G. Fagot, and B. Fertil. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol Biol Evol.*, 16(10):1391–9, 1999.

[31] D.H. Mathews and J. Sabina and M. Zuker and D.H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology*, 288:911–940, 1999.

[32] E. Duchaud, C. Rusniok, L. Frangeul, C. Buchrieser, A. Givaudan, S. Taourit, S. Bocs, C. Boursaux-Eude, M. Chandler, J. F. Charles, E. Dassa, R. Derose, S. Derzelle, G. Freyssinet, S. Gaudriault,

C. Medigue, A. Lanois, K. Powell, P. Siguier, R. Vincent, V. Wingate, M. Zouine, P. Glaser, A. Danchin, and F. Kunst. The genome sequence of the entomopathogenic bacterium Photorhabdus luminescens. *Nat Biotechnol*, 21:1307–1313, 2003.

[33] J. Dunn, S. McCorkle, L. Praissman, G. Hind, D. van der Lelie, W. Bahou, D. Gnatenko, and M. Krause. Genomic signature tags (GSTs): A system for profiling genomic DNA. *Genome Research*, 12:1756–1765, 2002.

[34] S. Edwards, B. Fertil, A. Giron, and P. Deschavanne. A genomic schism in birds revealed by phylogenetic analysis of DNA strings. *Syst Biol.*, 51(4):599–613, 2002.

[35] R. Elber and M. Karplus. Enhanced sampling in molecular dynamics: use of the time-dependent hartree approximation for a simulation of carbon monoxide diffusion through myoglobin. *J. Am. Chem. Soc.*, 112:9161–9175, 1990.

[36] A. Fedorov, S. Saxonov, and W. Gilbert. Regularities of context-dependent codon bias in eukaryotic genes. *Nucleic Acids Research*, 30:1192–1197, 2002.

[37] R. ffrench Constant, N. Waterfield, P. Daborn, S. Joyce, H. Bennett, C. Au, A. Dowling, S. Boundy, S. Reynolds, and D. Clarke. Photorhabdus: towards a functional genomic analysis of a symbiont and pathogen. *FEMS Microbiol Rev.*, 26:433–456, 2003.

[38] S. Forst and M. Inouye. Environmentally regulated gene expression for membrane proteins in Escherichia coli. *Annu Rev Cell Biol*, 4:21–42, 1988.

[39] S. Forst, J. Waukau, G. Leisman, M. Exner, and R. Hancock. Functional and regulatory analysis of the OmpF-like porin, OpnP, of the symbiotic bacterium Xenorhabdus nematophilus. *Mol Microbiol*, 18:779–789, 1995.

[40] S. Freeland and L. Hurst. Evolution encoded. *Sci Am.*, 290(4):84–91, 2004.

[41] Y. Fukuda, Y. N. Y, and M. Tomita. On dynamics of overlapping genes in bacterial genomes. *Gene*, 323:181–187, 2003.

[42] H. Gabow. *Implementation of algorithms for maximum matching on nonbipartite graphs.* PhD thesis, Stanford University, Stanford, California, 1973.

[43] G. M. Garrity, M. Winters, K. A. W., and D. B. Searles. Taxonomic outline of the prokaryotes. In *Bergey's Manual of Systematic Bacteriology.* Springer-Verlag, New York, 2nd edition, 2002.

[44] A. Gentles and S. Karlin. Genome-scale compositional comparisons in eukaryotes. *Genome Research*, 11(4):540–546, 2001.

[45] D. Georgellis, S. Arvidson, and A. von Gabain. Decay of ompA mRNA and processing of 9S RNA are immediately affected by shifts in growth rate, but in opposite manners. *J Bacteriol*, 174:5382–5390, 1992.

[46] D. Gilis, S. Massar, N. Cerf, and M. Rooman. Optimality of the genetic code with respect to protein stability and amino-acid frequencies. *Genome Biol.*, 2(11), 2001.

[47] C. Gustafsson, S. Govindarajan, and J. Minshull. Codon bias and heterologous protein expression. *Trends Biotechnol.*, 22:346–353, 2004.

[48] G. A. Gutman and G. W. Hatfield. Nonrandom utilization of codon pairs in Escherichia coli. *Proc. Natl. Acad. Sci. USA*, 86:3699–3703, 1989.

[49] S. J. Hallam, N. Putnam, C. M. Preston, J. C. Detter, D. Rokhsar, P. M. Richardson, and E. F. DeLong. Reverse methanogenesis: testing the hypothesis with environmental genomics. *Science*, 305(5689):1457–62, 2004.

[50] J. Heath, E. Ayres, M. Possell, R. D. Bardgett, I. J. Black, and H. G. et al. Rising atmospheric CO_2 reduces sequestration of root-derived soil carbon. *Science*, 309:1711–1713, 2005.

[51] S. M. Hebsgaard, P. G. Korning, N. Tolstrup, J. Engelbrecht, P. Rouze, and S. Brunak. Splice site prediction in Arabidopsis thaliana DNA by combining local and global sequence information. *Nucleic Acids Research*, 24(17):3439–3452, 1996.

[52] V. Hornak and C. Simmerling. Generation of accurate protein loop conformations through low-barrier molecular dynamics. *Proteins*, 51:577–590, 2003.

[53] L. L. Hsiao, F. Dangond, T. Yoshida, R. Hong, R. V. Jensen, J. Misra, W. Dillon, K. F. Lee, K. E. Clark, P. Haverty, Z. Weng, G. L. Mutter, M. P. Frosch, M. E. Macdonald, E. L. Milford, C. P. Crum, R. Bueno, R. E. Pratt, M. Mahadevappa, J. A. Warrington, G. Stephanopoulos, and S. R. Gullans. A compendium of gene expression in normal human tissues. *Physiol. Genomics*, 7:97–104, 2001.

[54] S. Hu, F. S. C. 3rd, M. K. Firestone, C. B. Field, and N. R. Chiariello. Nitrogen limitation of microbial decomposition in a grassland under elevated CO_2. *Nature*, 409:188–191, 2001.

[55] J. B. Hughes, J. J. Hellmann, T. H. Ricketts, and B. J. Bohannan. Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl Environ Microbiol*, 67(10):4399–406, 2001.

[56] B. A. Hungate, C. H. J. III, G. Gamara, F. S. C. III, and C. B. Field. Soil microbiota in two annual grasslands: responses to elevated atmospheric CO_2. *Oecologia*, 124:589–598, 2000.

[57] R. Jernigan and R. Baran. Pervasive properties of the genomic signature. *BMC Genomics*, 3:23, 2002.

[58] L. K. Johansen and C. D. Morrow. The RNA encompassing the internal ribosome entry site in the poliovirus 5' nontranslated region enhances the encapsidation of genomic RNA. *Virology*, 273:391–399, 2000.

[59] G. Kaplan and V. R. Racaniello. Construction and characterization of poliovirus subgenomic replicons. *J. Virol.*, 62:1687–1696, 1988.

[60] S. Karlin. Global dinucleotide signatures and analysis of genomic heterogeneity. *Current Opinion in Microbiology*, 1:598–610, 1998.

[61] S. Karlin. Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr Opin Microbiol*, 1(5):598–610, 1998.

[62] S. Karlin, A. Campbell, and J. Mrazek. Comparative DNA analysis across diverse genomes. *Annu. Rev. Genet.*, 32:185–225, 1998.

[63] S. Karlin, C. Chen, A. Gentles, and M. Cleary. Associations between human disease genes and overlapping gene groups and multiple amino acid runs. *Proc. Natl. Acad. Sci.*, 99(26):17008–17013, 2002.

[64] S. Karlin, J. Mrazek, and A. Campbell. Compositional biases of bacterial genomes and evolutionary implications. *J Bacteriol.*, 179(12):3899–913, 1996.

[65] L. Kaufman and P. J. Rousseeuw. *Finding groups in data : an introduction to cluster analysis.* Wiley, New York, 1990.

[66] P. Keese and A. Gibbs. Origins of genes: “Big bang” or continuous creation? *Proc. Natl. Acad. Sci.*, 89:9489–9493, 1992.

[67] P. F. Kemp and J. Y. Aller. Estimating prokaryotic diversity: When are 16s rdna libraries large enough? *Limnol. Oceanogr. Methods*, 2:114–125, 2004.

[68] J. S. King, K. S. Pregitzer, D. R. Zak, W. E. Holmes, and K. Schmidt. Fine root chemistry and decomposition in model communities of north-temperate tree species show little response to elevated atmospheric CO_2 and varying soil resource availability. *Oecologia*, 146:318–328, 2005.

[69] V. Kirzhner, A. Korol, A. Bolshoy, and E. Nevo. Compositional spectrum revealing patterns for genomic sequence characterization and comparison. *Physica A*, 312:447–458, 2002.

[70] S. Kodumal, K. Pael, R. Reid, H. Menzella, M. Welch, and D. Santi. Total synthesis of long DNA sequences: Synthesis of a contiguous 32-kb polyketide synthase gene cluster. *Proc. Nat. Acad. Sci.*, 44:15573–15578, 2004.

[71] D. C. Krakauer. Stability and evolution of overlapping genes. *Evolution*, 54(3):731–739, 2000.

[72] C. C. Lesaulnier, D. Papamichail, S. McCorkle, B. Ollivier, S. Skiena, S. Taghavi, D. Zak, and D. van der Lelie. Elevated CO_2 affects soil microbial diversity associated with trembling aspen. *submitted to Environmental Microbiology.*

[73] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk*, 163(4):845–848, 1965.

[74] M. Levitt. A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.*, 104:59–107, 1976.

[75] S. Lisser and H. Margalit. Compilation of E.coli mRNA promoter sequences. *Nucleic Acids Research*, 21:1507–1516, 1993.

[76] W. Ludwig and K. H. Schleifer. Bacterial phylogeny based on 16s and 23s rrna sequence analysis. *FEMS Microbiol Rev*, 15(2-3):155–73, 1994.

[77] W. Ludwig, O. Strunk, R. Westram, L. Richter, H. Meier, Yadhukumar, A. Buchner, T. Lai, S. Steppi, G. Jobb, W. Forster, I. Brettske, S. Gerber, A. W. Ginhart, O. Gross, S. Grumann, S. Hermann, R. Jost, A. Konig, T. Liss, R. Lussmann, M. May, B. Nonhoff, B. Reichel, R. Strehlow, A. Stamatakis, N. Stuckmann, A. Vilbig, M. Lenke, T. Ludwig, A. Bode, and K. H. Schleifer. Arb: a software environment for sequence data. *Nucleic Acids Res*, 32(4):1363–71, 2004.

[78] M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31 (13):3406–15, 2003.

[79] C. Manning and H. Schutze. *Foundations of statistical natural language processing.* MIT Press. Cambridge, MA, 2003.

[80] M. Marti-Renom, A. Stuart, A. Fiser, R. Sanchez, F. Melo, and A. Sali. Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct.*, 29:291–325, 2000.

[81] E. Merino and A. Garciarrubio. The global intrinsic curvature of archaeal and eubacterial genomes is mostly contained in their dinucleotide composition and is probably not an adaptation. *Nucleic Acids Res.*, 28(12):2431–8, 2000.

[82] M.G. Reese and F.H. Eeckman and D. Kulp and D. Haussler. Improved splice site detection in Genie. *Journal of Computational Biology*, 4:311–323, 1997.

[83] Michigan State University. Ribosomal Database Project (RDP). http://rdp.cme.msu.edu/download/SSU_rRNA/SSU_Prok.phylo.

[84] T. Miyata and T. Yasunaga. Evolution of overlapping genes. *Nature*, 272:532–535, 1978.

[85] A. Molla, A. V. Paul, and E. Wimmer. Cell-free, de novo synthesis of poliovirus. *Science*, 254:1647–1651, 1991.

[86] N. A. Moran, J. A. Russell, R. Koga, and T. Fukatsu. Evolutionary relationships of three new species of Enterobacteriaceae living as symbionts of aphids and other insects. *Appl Environ Microbiol*, 71:3302–3310, 2005.

[87] S. Mueller, D. Papamichail, J. Coleman, S. Skiena, and E. Wimmer. Reduction of the rate of poliovirus protein synthesis through large scale codon deoptimization causes virus attenuation of viral virulence. *Journal of Virology*, 80(19):9687–9696, 2006.

[88] M. Nei and S. Kumar. *Molecular evolution and phylogenetics.* Oxford University Press, Oxford ; New York, 2000.

[89] H. Nikaido and M. Vaara. Molecular basis of bacterial outer membrane permeability. *Nat Biotechnol*, 49:1–32, 1985.

[90] C. Nikolaou and Y. Almirantis. Mutually symmetric and complementary triplets: differences in their use distinguish systematically between coding and non-coding genomic sequences. *Journal of Theoretical Biology*, 223(4):477–487, 2003.

[91] S. Normark, S. Bergstrom, T. Edlund, T. Grundstrom, B. Jaurin, F. Lindberg, and O. Olassan. Overlapping genes. *Annu. Rev. Genet.*, 17:499–525, 1983.

[92] D. Oppenheim and C. Yahofsky. Translational coupling during expression of the tryptophan operon of E. coli. *Genetics*, 95:785–795, 1980.

[93] L. Ovreas. Population and community level approaches for analysing microbial diversity in natural environments, 2000.

[94] D. Papamichail and N. Delihas. Outer membrane protein genes and their small non-coding RNA regulator genes in Photorhabdus luminescens. *Biology Direct*, 1(12), 2006.

[95] D. Papamichail, C. C. Lesaulnier, S. Skiena, S. McCorkle, B. Ollivier, S. Taghavi, and D. van der Lelie. Towards a taxonomical consensus: Diversity and richness inference from large scale rDNA analysis. *submitted to Applied and Environmental Microbiology.*

[96] D. Papamichail, D. van der Lelie, S. R. McCorkle, and S. Skiena. Bacterial population assay via k-mer analysis. *Asian Pacific Bioinformatics Conference (APBC)*, pages 299–308, 2005.

[97] T. Pfister and E. Wimmer. Characterization of the nucleoside triphosphatase activity of poliovirus protein 2C reveals a mechanism by which guanidine inhibits poliovirus replication. *J. Biol. Chem.*, 274:6992–7001, 1999.

[98] J. B. Plotkin, H. Robins, and A. J. Levine. Tissue-specific codon usage and the expression of human genes. 101:12588 12591. *Proc. Natl. Acad. Sci. USA*, 101:12588–12591, 2004.

[99] D. Pride, R. Meinersmann, T. Wassenaar, and M. Blaser. Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Research*, 13:145–158, 2003.

[100] R Development Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2003. ISBN 3-900051-00-3.

[101] W. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66:846–850, 1971.

[102] G. Ried and U. Henning. A unique amino acid substitution in the outer membrane protein OmpA causes conjugation deficiency in Escherichia coli K-12. *FEBS Lett*, 223:387–390, 1987.

[103] I. Rogozin, A. Spiridonov, A. Sorokin, Y. Wolf, J. King, R. Tatusov, and E. Koonin. Purifying and directional selection in overlapping prokaryotic genes. *Trends Genet.*, 18(5):228–232, 2002.

[104] E. Rothberg. wmatch: a C program to solve maximum-weight matching.

[105] S. Brunak and J. Engelbrecht and S. Knudsen. Prediction of human mRNA donor and acceptor sites from the DNA sequence. *Journal of Molecular Biology*, 220:49–65, 1991.

[106] S. Saha, A. Sparks, C. Rago, V. Akmaev, C. Wang, B. Vogelstein, K. Kinzler, and V. Velculescu. Using the transcriptome to annotate the genome. *Nat. Biotechnol.*, 20:508–512, 2002.

[107] G. Sanchez, A. Bosch, and R. M. Pinto. Genome variability and capsid structural constraints of hepatitis A virus. *J. Virol.*, 77:452–459, 2003.

[108] R. Sandberg, C. Branden, I. Ernberg, and J. Coster. Quantifying the species-specificity in genomics signatures, synonymous codon choice, amino acid usage and G+C content. *GENE*, 301:35–42, 2003.

[109] R. Sandberg, C. Winberg, C. Branden, A. Kaske, I. Ernberg, and J. Coster. Capturing whole-genome characteristics in short sequences using a naive bayesian classifier. *Genome Research*, 11:1404–1409, 2001.

[110] P. D. Schloss and J. Handelsman. Introducing dotur, a computer program for defining operational taxonomic units and estimating species richness. *Appl Environ Microbiol*, 71(3):1501–6, 2005.

[111] V. Seguritan and F. Rohwer. Fastgroup: a program to dereplicate libraries of 16s rdna sequences. *BMC Bioinformatics*, 2:9, 2001.

[112] E. Simpson. Measurement of diversity. *Nature*, 163:688, 1949.

[113] S. Skiena. *The algorithm design manual.* Springer Verlag, 1998.

[114] S. Skiena. Designing better phages. *Bioinformatics*, 17:253–261, 2001.

[115] H. Smith, C. Hutchison, C. Pfannkoch, and J. C. Venter. Generating a synthetic genome by whole genome assembly: phiX174 bacteriophage from synthetic oligonucleotides. *Proc. Nat. Acad. Sci.*, 100:15440–15445, 2003.

[116] E. Stackebrandt, W. Frederiksen, G. M. Garrity, P. A. Grimont, P. Kampfer, M. C. Maiden, X. Nesme, R. Rossello-Mora, J. Swings, H. G. Truper, L. Vauterin, A. C. Ward, and W. B. Whitman. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. *Int J Syst Evol Microbiol*, 52(Pt 3):1043–7, 2002.

[117] E. Sugawara and H. Nikaido. Pore-forming activity of OmpA protein of Escherichia coli. *Proc Natl Acad Sci USA*, 76:4350–4354, 1979.

[118] T. A. Tatusova and T. L. Madden. Blast 2 sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett*, 174(2):247–50, 1999.

[119] R. D. C. Team. R: A language and environment for statistical computing, 2006.

[120] J. Tian, H. gong, N. Sheng, Z. Zhou, E. Gulari, X. Gao, and G. Church. Accurate multiplex gene synthesis from programmable DNA microchips. *Nature*, 432:1050–1054, 2004.

[121] S. G. Tringe, C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenholtz, and E. M. Rubin. Comparative metagenomics of microbial communities. *Science*, 308(5721):554–7, 2005.

[122] O. Troyanskaya, O. Arbell, Y. Koren, G. Landau, and A. Bolshoy. Sequence complexity profiles of prokaryotic genomic sequences: A fast algorithm for calculating linguistic complexity. *Bioinformatics*, 18(5):679–688, 2002.

[123] G. W. Tyson, J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar, and J. F. Banfield. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37–43, 2004.

[124] K. I. Udekwu, F. Darfeuille, J. Vogel, J. Reimegard, E. Holmqvist, and E. G. Wagner. Hfq-dependent regulation of OmpA synthesis is mediated by an antisense RNA. *Genes Dev*, 19:2355–2366, 2005.

[125] V. Veeramachaneni, W. Makalowski, M. Galdzicki, R. Sood, and I. Makalowska. Mammalian overlapping genes: The comparative method. *Genome Research*, 14:280–286, 2004.

[126] V. Velculescu. Using SAGE to explore the genome. *Proceedings from SAGE 2001: Frontiers in transcritome exploration*, page 15, 2001.

[127] V. Velculescu, L. Zhang, B. Vogelstein, and K. Kinzler. Serial analysis of gene exression. *Science*, 270:484–487, 1995.

[128] J. C. Venter, K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. H. Rogers, and H. O. Smith. Environmental genome shotgun sequencing of the sargasso sea. *Science*, 304(5667):66–74, 2004.

[129] B. Wang, D. Papamichail, S. Mueller, and S. Skiena. Two proteins for the price of one: The design of maximally compressed coding sequences. *11th International Meeting on DNA Computing (DNA11)*, 2005.

[130] Q. Wang, B. Chai, R. Farris, S. Kulam, D. McGarrell, G. Garrity, J. Tiedje, and J. Cole. The rdp-ii (ribosomal database project): The rdp sequence classifier., 2004.

[131] Y. Wang. The function of OmpA in Escherichia coli. *Biochem Biophys Res Commun*, 292:396–401, 2002.

[132] V. M. Williamson and H. K. Kaya. Sequence of a symbiont. *Nat Biotechnol*, 21:1294–1295, 2003.

[133] J. Yu, L. Zhang, P. Hwang, C. Rago, K. Kinzler, and B. Vogelstein. Identification and classification of p53-regulated genes. *Proc. Natl. Acad. Sci.*, 96:14517–14522, 1996.

[134] V. Zabarovska, A. Kutsenko, L. Petrenko, G. Kilosanidze, O. Ljungqvist, E. Norin, T. Midtvedt, G. Winberg, R. Mollby, V. Kashuba, I. Ernberg, and E. Zabarovsky. NotI passporting to identify species composition of complex microbial systems. *Nucleic Acids Res.*, 31(2):E5–5, 2002.

[135] E. Zabarovsky, L. Petrenko, A. Protopopov, O. Vorontsova, A. Kutsenko, Y. Zhao, G. Kilosanidze, V. Zabarovska, E. Rakhmanaliev, B. Pettersson, V. Kashuba, O. Ljungqvist, E. Norin, T. Midtvedt, R. Mollby, G. Winberg, and I. Ernberg. Restriction site tagged (RST) microarrays: a novel technique to study the species composition of complex microbial systems. *Nucleic Acids Res.*, 31(16):e95, 2003.

[136] D. R. Zak, K. S. Pregitzer, P. S. Curtis, J. A. Teeri, R. Fogel, D. L. Randlett, and A. Friend. Elevated atmospheric CO_2 and feedback between carbon and nitrogen cycles. *Plant and Soil*, 151:105–117, 1993.

[137] D. R. Zak, K. S. Pregitzer, P. S. Curtis, C. S. Vogel, W. E. Holmes, and J. Lussenhop. Atmospheric CO_2, soil-N availability, and allocation of biomass and nitrogen by Populus tremuloides. *Ecological Applications*, 10:34–46, 2000.

[138] L. Zhang, W. Zhou, V. Velculescu, S. Kern, R. Hruban, S. Hamilton, B. Vogelstein, and K. Kinzler. Gene expression profiles in normal and cancer cells. *Science*, 276:1268–1272, 1997.

[139] S. Zolotukhin, M. Potter, W. W. Hauswirth, J. Guy, and N. Muzyczka. A humanized green fluorescent protein cDNA adapted for high-level expression in mammalian cells. *J. Virol.*, 70:4646–4654, 1996.