A Biased Urn Model for Taxonomic Identification

Dissertation
zur Erlangung des mathematisch-naturwissenschaftlichen
Doktorgrades "Dr. rerum naturalium"
an der Georg-August-Universität Göttingen

vorgelegt von

Katharina Surovcik

aus Bratislava in der Slowakischen Republik

Göttingen 2008

D7
Referent: Prof. Dr. Stephan Waack
Korreferent: Prof. Dr. Manfred Denker

Tag der mündlichen Prüfung: 26. Juli 2008

Acknowledgements

This work would never have reached the present form without the input of many people. I would like to thank my supervisor Stephan Waack for his support and for the many ideas he was willing to share. I am grateful to Susanne Koch for enlarging and sharpening my view of the delicate science of stochastics as well as for her constant encouragement. Equally important was my delight to learn from her how one can be a woman and a mathematician at the same time. Moreover, I am thankful to Manfred Denker for intensive discussions and his asking the right questions.

I would also like to thank Axel Munk for his continuous interest in the progress of this work and his numerous hints concerning statistical methods, Hajo Holzmann for his contribution to the development of the model and Thorsten Hohage for pointing out how to overcome numerical complications. My work on this project was made possible by the sponsorship from the Graduiertenkolleg Identification in Mathematical Models: Synergy of Stochastic and Numerical Methods, for which I am thankful.

Productive scientific work can only happen within the right environment. In this context, I would like to thank Oliver Keller and Roman Asper who I shared an office with and who helped me out on many occasions.

I am deeply indebted to Aurel Cornea for giving me the feeling that science was offering unparalleled freedom and enriching exchange of thoughts with sharp-minded people. In this world, mathematics was to play a key role, building the tools of imagination. His seeing me taking my place in this world has been and still is an important source of inspiration, motivation and strength.

The mentorship of Jürgen Rohlfs has been very important to me. I thank Studienstiftung des deutschen Volkes for supporting me during my studies in Eichstätt as well as enabling the experience of my time in Oxford.

The ongoing support of my parents was partly edible and delivered by mail. However, it is their love and concern that contribute most to keeping me alive. I am very glad to have a caring little brother. Thousand thanks go to Thorsten Best for all the tea, and so much more.

In memoriam

Prof. Dr. Aurel Cornea

Contents

Introduction

Some time ago, Plato developed his concept of the world of ideas. Wherever two phenomena in the real world were *similar*, these supposedly represented a common idea. This notion of an abstract classification scheme for real world objects has underlied science ever since and stimulated developments such as the periodic table of the elements, the marxist idea of patterns in history or the conception of elementary particles.

Nowadays, the problem to identify entities as members of some class constantly arises in a practical sense. Important examples are image recognition, fingerprint assignment or earthquake prediction.

Manifold instances of this problem occur in the life sciences. An example of current interest is the challenge of taxonomic identification. It is highly desirable to tackle this on the fundamental level that governs the functioning of living cells, i. e. based on the genomic information. In the following, we concentrate on prokaryotic species. These single-cellular organisms are characterised by the absence of cell nuclei and by reproduction through binary fission. They usually possess a single chromosome carrying their genomic information.

This can be seen as a sequence consisting of the nucleotides Adenine, Cytosine, Guanine and Thymine, short A, C, G and T. Genes are subsequences which encode proteins, that is sequences of amino acids. Each of those corresponds to a triple of nucleotides termed the *codon*. This representation is not one-to-one, rather there is a characteristic coding redundancy where several codons lead to the same amino acid (39).

A statistical analysis can exploit such redundancy. This is the basis of sequence measures like codon usage, mono- or oligonucleotide bias.

In the present work, we focus on modelling the dinucleotide bias in prokaryotic genomes. To this end, we construct a biased urn model based on the noncentral hypergeometric distribution by Wallenius (112). Our method is sufficiently general to capture either dinucleotide with respect to the positions in the codon and combinations thereof. However, our emphasis is mostly on the intercodon transition for its complementarity to existing codon usage approaches.

The work is composed of the following parts: The initial chapter 1 defines the biological background and clarifies the terminology. Chapter 2 investigates important stochastic properties for the genomic data under consideration. Subsequently, the model is layed out in chapter 3 with an introduction to the application of the underlying distribution. Afterwards, it is time to feed the model with real world data. This is done in chapter 4, using various test scenarios and comparing to exising literature results.

We also consider the possibility of extending the applicability of our model towards metagenomics, to be defined later. Finally, our conclusions are presented in chapter 5.

1 Biological Background

In this chapter, the biological background of macrobial and bacterial evolution is introduced. It ought to provide enough basic information to understand the problem and its history as well as our idea to solve it which will be presented later on.

1.1 Definition of Species and Phylogeny

Since the beginning of the evolution theory biologists have used the metaphor of the 'tree of life' to describe how the variety of organisms has evolved whose essence is the vertical inheritance of traits through generations from an individuum to its descendants. Charles Darwin proposed this idea and the concept of common descent behind it in his book 'On the Origin of Species' in 1859 (17). Ever since, evolution has been envisioned as a bifurcating process whose imprints on the map of life would thus display a tree.

Classical Phenotypical Approach to Taxonomy

Having the theory, it still has not been easy to set the different stages on the tree, or, putting it in biological terms, to define the taxonomy of life. In fact, there are no stages in nature. Every distinction on the evolutionary steps as well as the separation into species is man-made and as such open to discussion. For higher organisms, this endeavour has led to an agreement upon some taxonomic categories. Although the concept of species is backed up by evolutionary processes, nature has not left niches between variations for us to set the taxonomical boundaries in. Thus there is still space for interpretation, in Darwin's words "in determining whether a form should

be ranked as a species or a variety, the opinion of naturalists having sound judgement and wide experience seems the only guide to follow."(17, Chapter 2) When it comes to unknown grounds, like classifying organisms from new regions, new experience has to be acquired first before one can rely on it. The problem has not more than maybe shifted since Darwin's time.

This is all the more the case at the level of unicellular organisms considering that here the species demarcation according to classical phenotypes is not defined by a theory-based concept and tends to be more arbitrary, anthropocentric or rooted in practical necessity (38). This might have been a grouping by the natural habitat or by the symptoms the organisms caused by pathogenic behaviour in other organisms. A pathogenic and a non-pathogenic individuum of the same species cause very different effects, yet they are closely related. Thus, a mapping on these criteria could have been, if at all, only poor.

Broader Molecular View

Hence, a different, deeper approach was indeed inevitable. That is when scientists started looking at the genetic level of relations and organisms. "It is deep down on the level of molecules and molecular sequences where the evolutionary process demonstrates its workings" (118). Unsurprisingly, the attempt to construct a taxonomical tree using molecular characteristics led to a breakthrough.

To cover a broad spectrum of the species and their interrelation, a molecule common to all of them was required. This necessity comes hand in hand with the problem of balance. The molecule has to be conserved enough such that it can be found in all species, and it also has to have evolved making building of a tree based on these alterations possible. In their pioneering work (115), Woese and Fox made use of the fact, that all self-replication in the cell involves ribosomal RNA (rRNA). The structure of the 16S rRNA molecule had already been characterised in a variety of organisms such that Woese and Fox could utilise it for their purpose.

Their work resulted in the division of unicellular organisms into the three primary domains *Eucaryota* (also referred to as *eucaryotes*), *Archaea* and *Bacteria* (116). Since then this tree has been endetailed and adjusted as

soon as new genetic sequences and their 16S rRNA characterisation became available.

More and more genetic data was published in the 1980s and new research on them stumbled upon irregularities in some genetic information which perforated the fundamental division between the domains. The anomalies did not stop there, instead, they happened to occur at all levels of relatedness. Suddenly, different branches became uninterpretable in the old fashion since the analysed microbial genomes were much more intermixed than anticipated (22; 65).

When Darwin mapped out the theory of evolution, he was inspired by his discoveries in animals and plants. However, these are very complex organisms.

Due to the high degree of organisation in a variation of cells, cell types and organs that they show, only subtle alterations are possible without the loss of functionality in the new organism. It was these little changes which Darwin and other saw and that made the *tree of life* plausible. Now however, a modified metaphor is needed for the new findings (64).

1.2 Horizontal Gene Transfer

Looking at microbial organisms, the variation in their metabolic properties, lifestyles and cellular structures is extraordinary. Point mutation and inheritance, which would go along with Darwin's ideas, causing modification or inactivation of existing genes, cannot be the only explanation for the diversity among even closely related species (65; 92). The exceptional ability of microbes to adapt to new environments calls for a different interpretation.

In the microbial world vertical inheritance of genes is not the only factor of evolution. Genetic information cannot only be exchanged among individuums, but also among different species and even between species belonging to different domains (23; 47; 3; 79).

Lateral Gene Transfer or *Horizontal Gene Transfer*[1], as the propagation of genetic information among contemporaries is called, has had and still has a major impact on microbial evolution (4; 53).

Therefore, some other metaphors instead of the tree of life have been suggested. On the one hand, the *net* or *reticulated tree* metaphor (24) takes the ongoing horizontal gene transfer into account. However, it still assumes some kind of rooting or beginning. This "universal ancestor is not a discrete entity. It is, rather, a diverse community..." (114). The metaphor which makes allowance to the ancient gene transfer of the mentioned diverse community is the *Ring of Life* as proposed by Rivera and Lake in (94) and discussed in (5; 74). It offers a solution to the difficulties encountered when trying to look for a root of the tree. Still, a solution is far from being found due to the fact that the incidence of past events has not been and cannot be fully analysed. These metaphors are "...not an answer to a question, [they are] a picture of a problem." (73) Furthermore, a metaphor taking care of all effective events has not yet been formulated.

Mechanisms of Horizontal Gene Transfer

As the HGT events are not self-explanatory, both their existence as well the mechanisms inducing them were discovered rather late and could bear some description.

Compared to the modification of existing sequences, where no prerequisites other than the replication are needed, the acquisition of new traits through horizontal gene transfer necessitates certain mechanisms. The organism in question needs to be able to acquire DNA, to keep it and to use it.

First of all, the DNA must be transferred from the donor cell to the recipient cell. Once it is there, it needs to be embedded into the recipient's replication element, e. g. the genome (83). These requirements are indifferent to the transferred genes or their properties and are met by *transformation, transduction* or *conjugation*, which, however, do not work with all genes.

[1] In the following and the subsequent chapters, these events will be referred to as HGT events.

Transformation: This is the process whereby naked DNA from the microbe's surroundings is taken up. This event relies on the exposure to extracellular DNA, which is released into the environment from decomposing cells (86; 36), disrupted cells, viral particles or excretion from living cells (68; 76). Extracellular DNA degradation varies considerably in different environments, whereby the exact rates need yet to be determined. In human serum and plasma the present DNases have been found to degrade DNA within between a few minutes (15) and a couple of hours (96), the latter one providing enough time for transformation. In other environments, the rates have been reported to be similarly diverse. Extracellular DNA has been found in soil (67; 84) fresh and sea water (19; 59) bacterial cultures (67) the mammalian intestinal system (27; 98; 14) and as an important component in biofilm formation (113). Active excretion from living cells has been reported for many kind of bacteria (108; 68; 76; 85).

Yet, the exposure to DNA is not enough for transformation to happen. The bacteria also have to develop a state of *competence*, which is the ability to bind and take up extracellular DNA. It involves several proteins and is usually a response to specific conditions in the environment such as growth conditions, nutrient access, cell density and starvation (108). This has been detected in approximately 1% of the described bacterial species (55) so far, forming a wide variety, including pathogenic bacteria (68).

Finally, the DNA needs to be incorporated into the genome. If the sequence of an incoming DNA strand contains regions that are highly similar to a host sequence, the former may be integrated therein by homologous recombination (108).

The occurance of the ability of transformation in such various organisms suggests its value in many different environments (108). It has first been discovered when actually looking at horizontal gene transfer. Some observations indicate that unicellular organisms take up DNA less as a source of information than as an energy-efficient source of nucleotides, to be used as such for DNA repair or components of them (90; 91).

Transduction: In this process, a virus is utilised in the transfer of bacterial DNA from one bacterium to another. Hereby, the transfer deploys a possible flaw in the viral strategy. When bacteriophages, which are the bacterial viruses, infect a bacterial cell, they harness the DNA replication system of

the host by employing it to produce numerous copies of their own genetic information. These are then packaged into new bacteriophage copies (66, Chapter 14), ready to infect new bacterial cells.

At the stage of this packaging process two mistakes occur at a low frequency: either too much or too little of the genetic material is packaged; some bacterial genetic information might be incorporated into a copy of the phage (generalised transduction) or some phage genes might be left behind in the bacteria (specialised transduction) (58; 107).

The first mistake leads to a noninfectious phage. Nevertheless, being a phage it tries to infect a bacteria and thereby inserts the newly acquired DNA into it. The phage's replication might not work anymore, but a recombination of the recipients genetic information integrating the donated DNA may take place. In the second case the viral DNA itself can get incorporated into the bacterial DNA (12; 9). This mechanism enables viruses with a broad range of possible hosts to transfer genetic information across enormous phylogenetic distances (58).

Conjugation: This kind of transfer of genetic material involves the physical contact between two organisms. Therefore, it is also known as bacterial sex. However, this term is a bit misleading since the transfer does not involve any of the mechanisms defining sexual replication (91). It is merely the transfer of genetic information from one cell to another. The donating cell must host a conjugative or mobilisable genetic strain, often a conjugative plasmid some of which can integrate themselves into the bacterial chromosome later on (108). This transfer occurs between distantly related organisms (54; 117; 31), even between domains (47; 6).

Reason for Horizontal Gene Transfer

While all three of these mechanisms appear to have evolved because of advantages they serve to other traits, they repeatedly were and are utilised for the transfer of genes. As the layout suggests, larger strains of DNA can be acquired when a transfer happens. Furthermore, this event might take place more than once and the recombination leads to a mosaic structure, where the genes originate from a variety of organisms (102; 80). Therefore, the term of genetic islands has developed (44; 43; 35).

However, the replication of DNA is costly and thus information that does not provide a selective advantage accounts negatively. Hence, the genome often loses these parts again (82; 72; 46).

The effect of HGT is a fast adaptation to a change in and of environment, since genes and the traits they represent do not have to be developed all over again but are instantaneously transferred.

1.3 Detection of Horizontal Gene Transfer Events

To reveal HGT one can either look at the genome and the information within or take other information into account. Both approaches have their advantages and disadvantages.

Extrinsic information may exist in the form of a database of gene and/or protein families in related species. The occurance or absence of genes is used in a phylogenetic tree to reconstruct gene transfer and gene loss events. If enough related species are sequenced and the gene families known, this method detects the HGT events that contradict the phylogeny. As accurate as this systematic analysis is, the drawbacks are also apparent. The considered taxon has to be well studied and the gene families known. This is usually not the case for the majority of genes in question. Therefore, the result depends highly on the currently available databases and may change as more known information is taken into account (83). Furthermore, as horizontal gene transfer contradicts a pure tree structure, it is always an ambiguous task to build a phylogenetic tree that contains this event, with the outcome depending on the specific algorithm.

A different approach is based on genome intrinsic information only. In this case, taxonomic "identity" is contained in redundancies in the genetic coding. Clearly, this can be any expoitable redundancy in the genetic coding that is, any variability in a certain feature that is not based on underlying biochemical imperatives. HGT events manifest themselves in significant differences of that feature in genomic subsequences.

There are several features that can be considered, i. e.

- G-C content (105; 95; 18)
- codon usage (40; 100; 111)
- nucleotide and various oligonucleotide (88; 32; 25; 10) and
- dinucleotide bias (45; 50; 11; 60).

Comparative studies can be found in (61; 75; 97).

The distinctive characteristic of each feature lies in its complexity which allows for a deeper insight at the price of more computational effort, more parameters to be taken into account and more detail knowledge necessary for the application. Therefore, an optimal trade-off between complexity and versatility has to be chosen depending on a specific application and aim. While a combination of several features might yield insight in specific situations, it would be conceptually desirable to work with one standard feature.

It is intuitively clear, that these statistical methods have their strength in detecting HGT between very remotely connected species while it will be harder for them to deal with HGT in neighbouring species. On the other hand, phylogenetic methods work best for species that are very closely related. Yet, they can only handle HGT within the considered set of genes. Therefore, the choice of methods largely depends on whether broad panoramic view involving a variety of species or a close-up look through a magnifying glass on a small taxonomic unit is desired.

This work concentrates on the statistical broad view approach, specifically the dinucleotide feature which we will demonstrate to have equivalent performance as the codon usage approach while the underlying model is smaller and easier to handle.

2 Stochastic View on the Data

The position of a nucleotide within a codon is the main emphasis of the model that we will introduce later. Therefore, only protein coding sequences are considered as genes and, moreover, genes are seen as sequences of codons.

2.1 Dinucleotides

Given this sequence, pairs of nucleotides, which are called *dinucleotides*, might be formed. Having the underlying structure of the codons, nucleotides are not just grouped, but rather their positions within a codon are taken into account. However, the nucleotides do not have to be neighbours and can be part of one or more codons. While this dismantles most boundaries in the building of pairs we will focus on the three cases where the nucleotides are adjacent. These cases are

1-2 dinucleotide and **2-3 dinucleotide:** Here, the first and second or the second and third nucleotide, respectively, build a tupel. Since the dinucleotides are parts of codons their number in a gene is the same as the number of codons in the gene.

3-1 dinucleotide: The tupel building is across two sequential codons. The third nucleotide of the first codon and the first nucleotide of the second codon are paired. Then, the number of dinucleotides in a gene is the number of transitions between codons and therefore one less than their number.

2.1.1 Dinucleotides as Random Variables

To be able to look at the data with a stochastic eye, some formalisation needs to be faced. As we consider only coding sequences, there are always multiples of 3 nucleotides in a gene and we define a stochastic process

$$X_1, X_2, X_3, \ldots, X_{3n-1}, X_{3n},$$

where $X_i \in \{A, C, G, T\}$ are random variables taking values in the nucleic alphabet and n is the length of the sequence, measured in codons. In this notation the dinucleotides are

$$\begin{array}{ll} Y_i^{(12)} & = (X_{3i-2}, X_{3i-1}), \quad \text{for } i = 1, \ldots, n, \\ Y_i^{(23)} & = (X_{3i-1}, X_{3i}), \quad \text{for } i = 1, \ldots, n, \\ Y_i^{(31)} & = (X_{3i}, X_{3i+1}), \quad \text{for } i = 1, \ldots, n-1, \end{array}$$

for the 1-2, the 2-3 and the 3-1 dinucleotide, respectively. If the type of the dinucleotide is clear or if general properties are discussed, the specifying index will be dropped and $Y_i \in \{A, C, G, T\}^2$ for all i and we speak of the random process $\{Y_i : i \in \mathbb{N}\}$.

Later, we would like to model this process based on certain assumptions. The validity of those assumptions for the given data is investigated in the following. Before we can do this, we need to define a few more random variables that we can work with.

Definition of Sums

Let h be a bijection between $\{1, 2, \ldots, 16\}$ and $\{A, C, G, T\}^2$. First, we define a vector of the partial sums of the sequence as

$$S_m := \left(\sum_{i=1}^{m} \mathbb{1}_{\{Y_i = h(k)\}} \right)_{(k=1,\ldots,16)}, \tag{2.1}$$

this way S_m takes values in $\mathbb{N}^{16}$. Dividing by m we get

$$Z_m := \frac{1}{m} \cdot S_m = \frac{1}{m} \left(\sum_{i=1}^{m} \mathbb{1}_{\{Y_i = h(k)\}} \right)_{(k=1,\ldots,16)} \tag{2.2}$$

the normalised partial sums. Additionally, we want to consider blocks of the sequence. To this means, let

$$S_{s,l} := \left(\sum_{i=s}^{s+l-1} \mathbb{1}_{\{Y_i = h(k)\}} \right)_{(k=1,\ldots,16)} \tag{2.3}$$

be the sum of a block of the sequence of length l starting at s. Again, we can normalise by the block length and obtain

$$Z_{s,l} := \frac{1}{l} \cdot S_{s,l}\,, \tag{2.4}$$

analogously. In this notation, we have the relation to the previous two equations 2.1 and 2.2 by $S_m = S_{1,m}$ and correspondingly similar with Z_m.

2.1.2 Genes as Paths

Looking at the data, we come across genes in genomes that can be understood as realisations or sample paths of the process $\{Y_i : i \in \mathbb{N}\}$ of a given length, which can be either the length of the gene or the length of the genome.

As we assume the genome G being the sequence of the genes g of the length n_g, we can further define

$$S_G := \sum_{g \in G} S_{n_g} \tag{2.5}$$

and

$$Z_G := \frac{1}{\sum_g n_g} \sum_{g \in G} S_{n_g}\,, \tag{2.6}$$

where g runs over all genes in the genome. [1]

[1] If a gene g is from a genome G, we use the notation $g \in G$. Stricly speaking, we could argue that $g \subset G$ as it is a subsequence. However, the former sloppy notation is more intuitive in the context of summation where it will be mostly used.

2.2 Properties of the Data

In the following, we turn towards the genomic data and study their stochastic properties in more detail.

The considered data are 420 bacterial genomes from the EMBL Nucleotide Sequence Database (28), consisting mostly of 1000 to 6000 genes each, though some shorter genomes can be found. Thereby, the length of a gene varies between about 50 up to 1500 and more codons, the genome mean of the gene length being between 250 and 350.

In a fixed genome G, we expect the Z_{n_g} to be identically distributed for all genes $g \in G$. We further assume that Z_G follows the same distribution.

The given showcases are randomly picked out of the dataset. Similar results can be obtained for all genomes.

Convergence

First, the convergence of the Y_i is examined. By the Law of Large Numbers, Z_m should converge to a random variable Z.

However, it is not a priori clear, if the individual genes are long enough for this to happen. Therefore, the behaviour of Z_m for increasing m is studied. If our assumption about the distributions of Z_G and Z_{n_g} for all $g \in G$ being identical is true, $\mathbb{E}(Z)$ can be approximated by Z_G.

For each coordinate Z_m^k of Z_m we have, as $Z_m^k \overset{m\to\infty}{\longrightarrow} \mathbb{E}(Z^k)$ and $\mathbb{E}(Z^k) \neq 0$, that

$$\frac{Z_m^k}{\mathbb{E}(Z^k)} \approx \frac{Z_m^k}{Z_G^k} \overset{m\to\infty}{\longrightarrow} 1$$

which we examine more closely.

As an example, figure 2.1 on the facing page shows Z_m^k/Z_G^k for coordinates k_1 and k_2 with $h(k_1) = AC$ and $h(k_2) = GT$ as well as $\max_k Z_m^k/Z_G^k$, $\min_k Z_m^k/Z_G^k$, whereby g is the gene tagged *SSON_1747 ydbK* in the organism *Shigella*

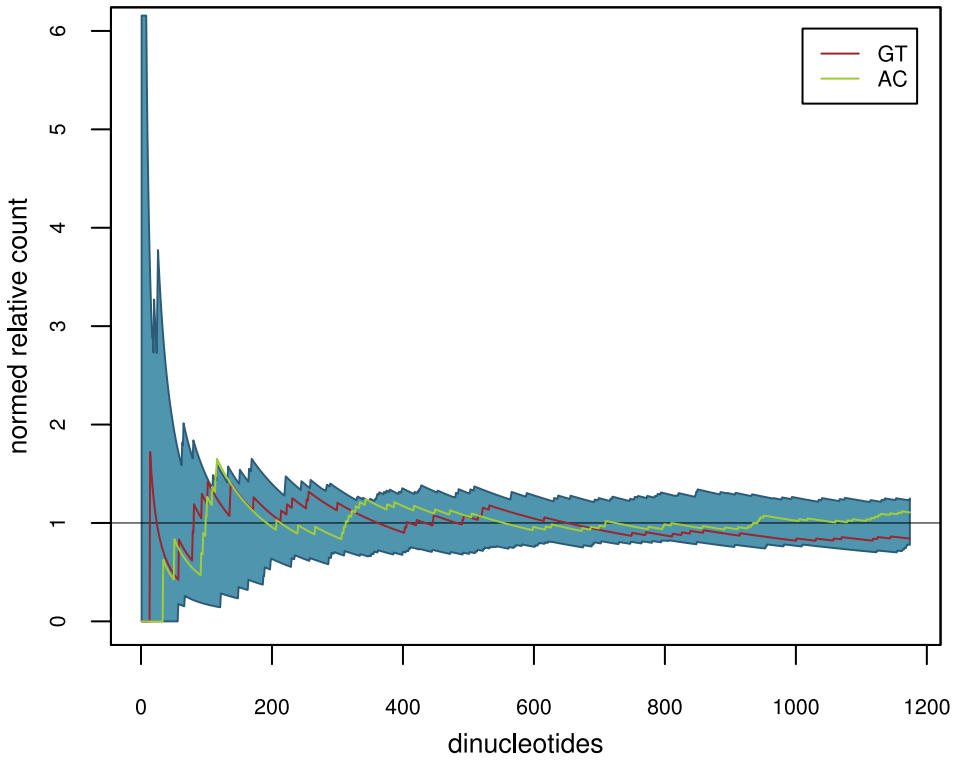

Figure 2.1: Corridor of the dinucleotide frequencies relative to their expected value, given by $\max_k \frac{Z_m^k}{Z_G^k}$ *and* $\min_k \frac{Z_m^k}{Z_G^k}$. *(The values are taken from the 3-1 dinucleotide of the gene tagged* SSON_1747 ydbK *of* Shigella sonnei, strain Ss046.*)*

sonnei (strain Ss046). For practical reasoning, convergence can be assumed as fast as $n \approx 200$.

The values to which the Z_m^k converge can differ significantly for individual k. This can be seen in figure 2.2 on the next page in the gene tagged *TBFG_01212* of *Mycobacterium tuberculosis (strain F11)* looking at the 2-3 dinucleotide. These values also vary in different genomes. The figure also shows that for some k the partial sums S_m^k are rather small, which results in discrete jumps in Z_m^k. These need to be taken care of in the modelling later.

Next, we consider the normality character of the partial sums. By the central limit theorem, we expect their distribution to approach the normal distribu-

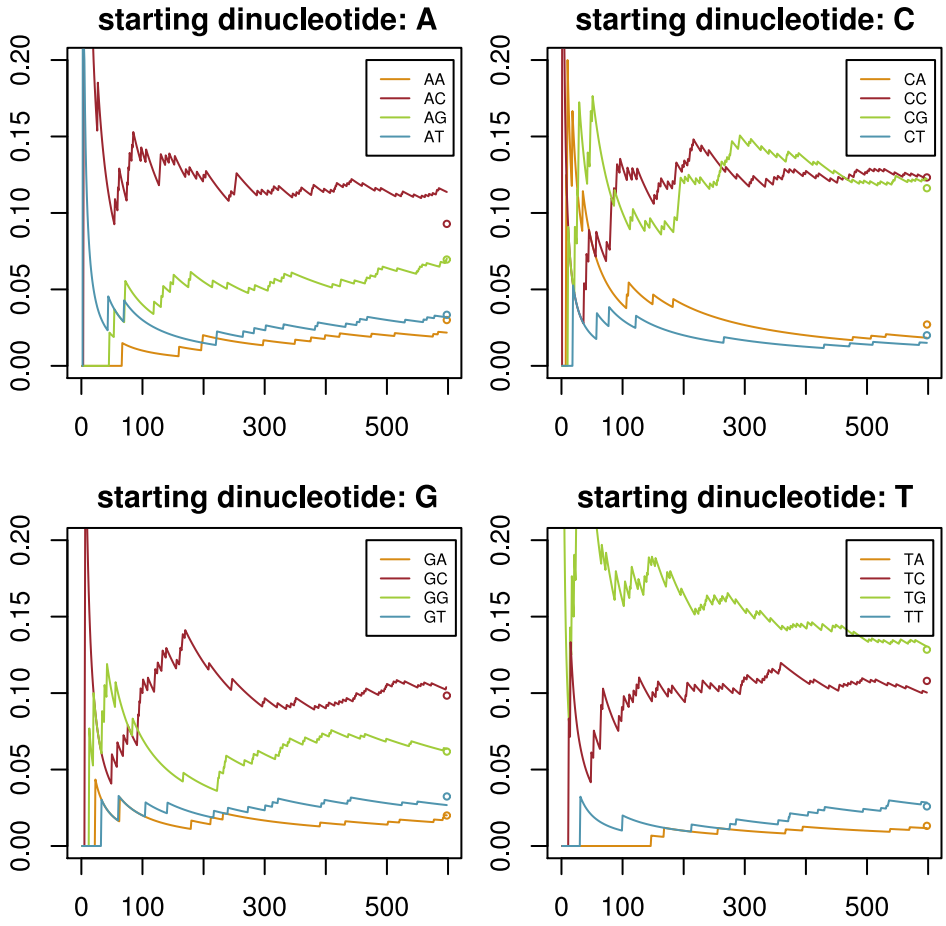

Figure 2.2: Individual dinucleotide frequency curves for all the dinucleotides, grouped by four. Additionally, the genome wide frequencies are shown as the circles on the right hand side for each plot. (The values are taken from the 2-3 dinucleotide of the gene tagged TBFG_01212 *of* Mycobacterium tuberculosis, strain F11.*)*

tion with mean μ and variance σ^2 as $n \to \infty$. The mean and variance depend on the data under consideration.

To this end, we examine $S_{s_b,l}$ as defined in 2.3 with $s_b = b(l-1)+1$ and b the number of the block, where the length l is chosen such that convergence of Y_i within the blocks $S_{s_b,l}$ can be assumed.

The distribution of the $S_{s_b,l}$ is investigated in a Quantile-Quantile-Normality Plot in 2.3 on the following page. Additionally, we perform a Shapiro-Wilk Test (99) and present the obtained p-values in the same graph. In general, very good agreement is found for all k. The data in figure 2.3 are taken from the organism *Pseudoalteromonas haloplanktis (strain TAC125, chromosome I)* from the 1-2 dinucleotide of gene tagged *PSHAa2216 hrpB*. Though the results here are typical, some rare genes might behave pathologically as seen in 2.4 on page 19.

Stationarity

In the following, we want to see if our data shows any stationary behaviour.

As defined in (42), a stochastic process $X = \{X(t) : t \in T\}$ is *strongly stationary* if the families $\{X(t_1), X(t_1), \ldots, X(t_m)\}$ and $\{X(t_1 + l), X(t_1 + l), \ldots, X(t_m + l)\}$ have the same joint distribution for all $t_1, t_2, \ldots, t_m \in T$ and $l > 0$.

A *weakly stationary* process has the property, that $\mathbb{E}(X(t_1)) = \mathbb{E}(X(t_2))$ and $\operatorname{cov}(X(t_1), X(t_2)) = \operatorname{cov}(X(t_1 + h), X(t_2 + h))$ for all $t_1, t_2 \in T$ and $h > 0$.

However, both these definitions are not very well applicable in our data. The best criterion is a look at the behaviour of the sample means.

We sampled m timepoints $t_1, \ldots, t_m$ from the first half of a gene to avoid any kind of implication of the property in the setup. We then defined a random variable

$$N(a) := \frac{1}{m}\left(\sum_{i=1}^{m} \mathbb{1}_{\{Y_{t_i+a}=h(k)\}}\right)_{(k=1,\ldots,16)},$$

that sums up the values of the random process at the timepoints for $a = 0$ and shifts the timepoints by a for $a > 0$, respectively. In order to investigate the gene g we evaluate $N(a)$ on g. If the process is stationary, a horizontal line with some noise is expected for the plot of each of the coordinates of $N(a)(g)$. These plots can be seen for *Nitrosococcus oceani (strain ATCC 19707)* and the gene *Noc_0602* in figure 2.5 on page 20. Here, m was chosen to be 160. Plots with $m = 70$ and $m = 100$, that are not presented, were also generated and showed similar trends, with higher noise as the only difference.

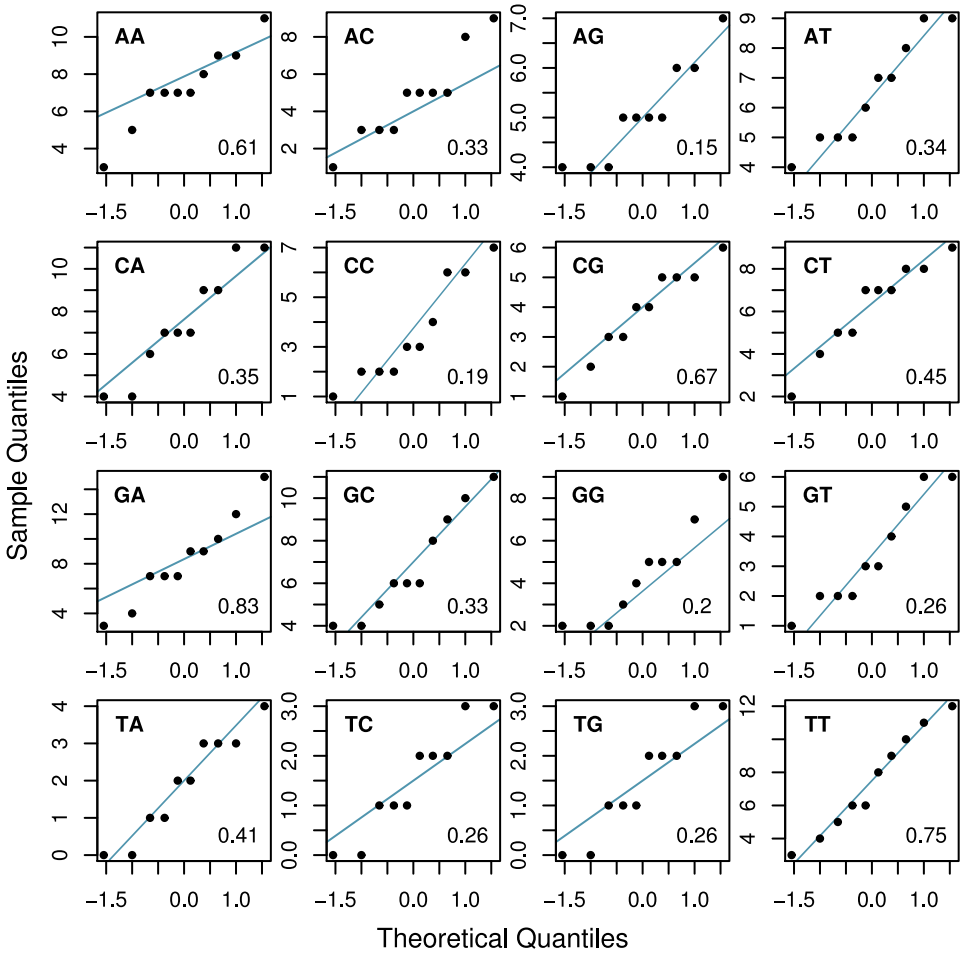

Figure 2.3: Quantile-Quantile Normality Plots of all 16 dinucleotides. In the lower right corner of each plot the p-value of the corresponding Shapiro-Wilk test is given. (The values are taken from the 1-2 dinucleotide of the gene tagged PSHAa2216 hrpB *of* Pseudoalteromonas haloplanktis (strain TAC125, chromosome I)*.)*

Encouraged by the plots, we now try a closer look at the distribution of the data. To this means, we compute $S_{s_b,l}$ as defined in 2.3 with $s_b = (b-1)\,l+1$. If the distribution of the $S_{s_b,l}$ is the same for different b we can assume that the distribution of our sequence is independent of the starting index s_b. This can now be verified using Pearson's χ^2-test (87).

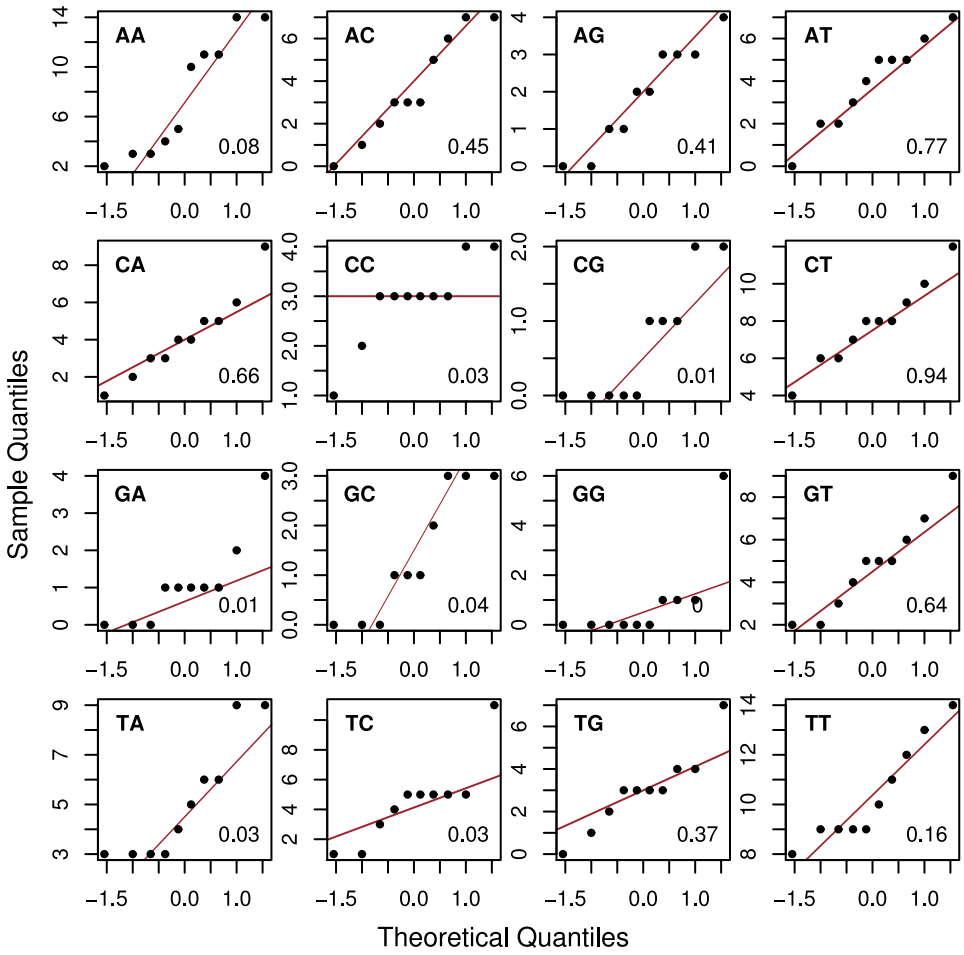

Figure 2.4: Quantile-Quantile Normality Plots with the p-values of the Shapiro-Wilk test of all 16 dinucleotides as an example of a pathologically behaving gene. (The values are taken from the 2-3 dinucleotide of the gene tagged LBA1966 copA *of* Lactobacillus acidophilus, strain NCFM.*)*

We have chosen the block length $l = \frac{n}{3}$, where n is the gene length and s_b for $b = 1, 2, 3$ as the start indices of the three thirds of the gene sequence. Next, we test for pairwise independence of $S_{s_b,l}$ for all three dinucleotide sequences. A histogram of the resulting p-values in all genes of all organisms is presented in 2.6 on page 21 and shows that 92% of all 10636011 tests return a p-value below $\alpha = 5\%$. As we did not exclude any genes, especially not

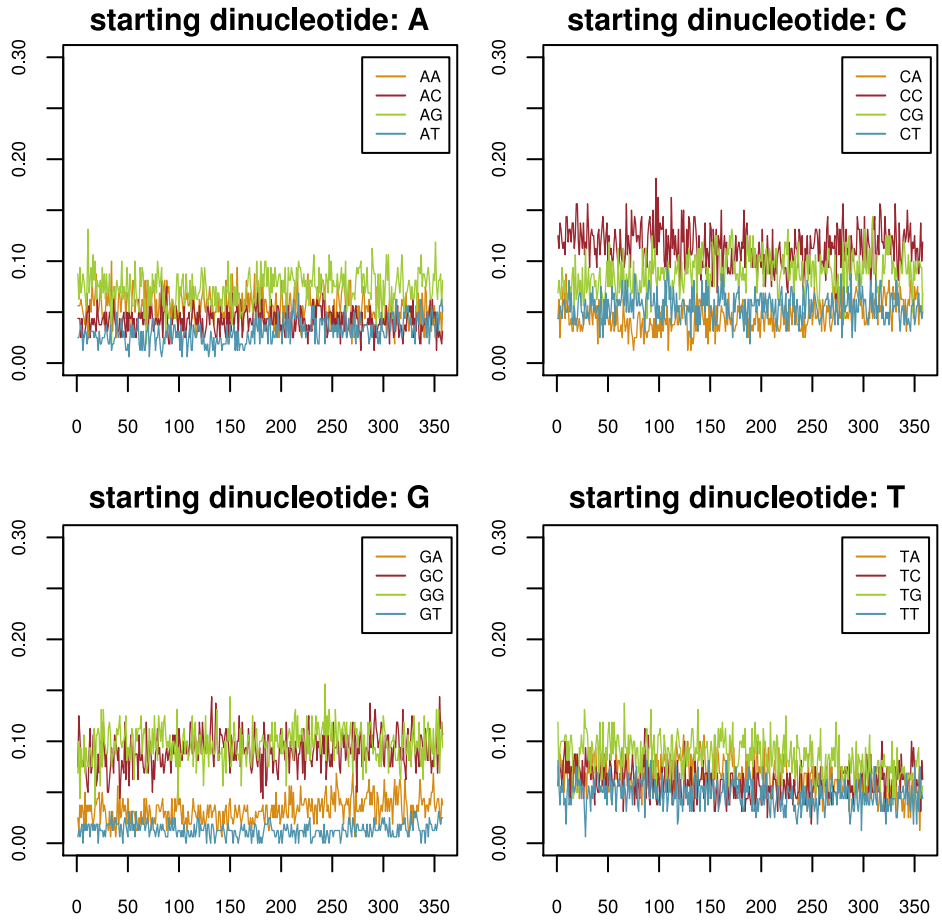

Figure 2.5: Evaluation of $N(a)$ presenting the behaviour of the sample means of the dinucleotides when shifting along the gene. (The values are taken from the 2-3 dinucleotide of the gene tagged Noc_0602 *of* Nitrosococcus oceani (strain ATCC 19707)*.)*

short ones which constitute about one third of all genes, this result is better than anticipated.

Ergodicity

A weakly ergodic process is defined as a weakly stationary process $\{X_i : i \in \mathbb{N}\}$ and a random variable Y with $\mathbb{E}(Y) = \mathbb{E}(X_1)$ and $\frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow Y$ in mean square.

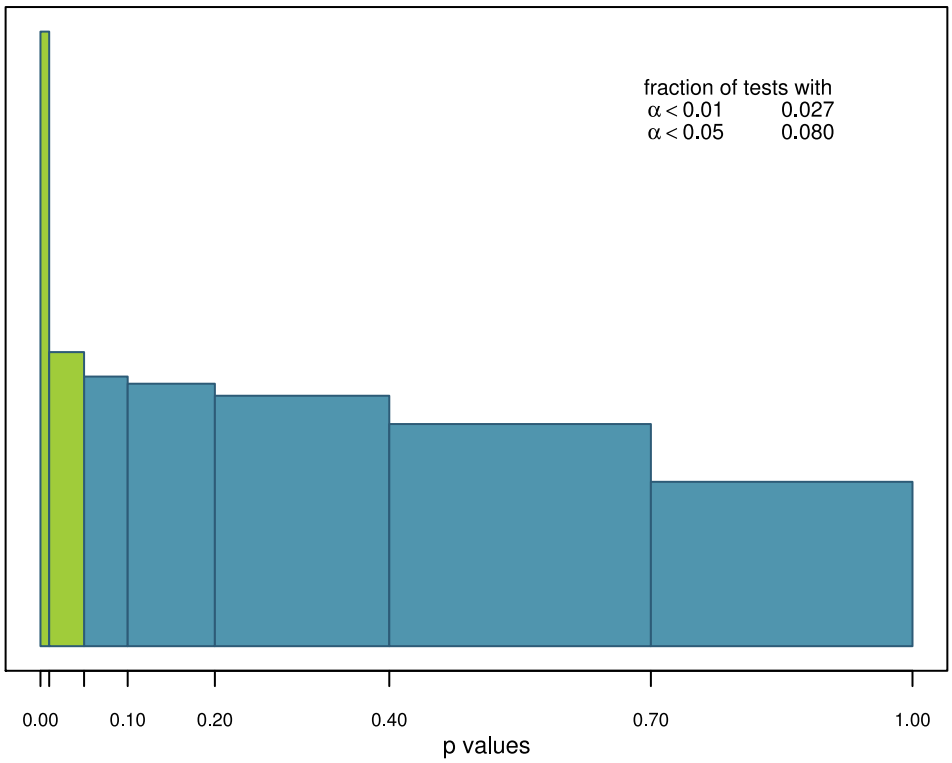

Figure 2.6: Histogram of p-values of the stationarity χ^2 test as described in detail in the main text. The data contain 10636011 *individual test runs.*

This property is difficult to show in individual genetic sequences as these represent paths and which in general are too short for any subdivisions. However, what is relevant in our case, is the ergodicity across the genome.

To test this, we combined the genes to a long sequence which we cut into n pieces of length n, not taking care of gene ends and beginnings. This way, we avoid any effects of a potential gene starting distribution, that might show in the first two dinucleotides (110; 89). We then again computed the contingency tables of the dinucleotides for each of this blocks as $S_{s_b,n}$ with $s_b = (b-1)\,n+1$. Additionally, we computed the sum of all block beginnings Y_{s_b} as

$$S_{beg} := \left(\sum_{i=s_b}^{b\cdot n} \mathbb{1}_{\{Y_i = h(k)\}} \right)_{(i=1,\ldots,16)}$$

and compare it to all $S_{s_b,n}$. Hereby, we can again perform Pearson's χ^2-test of independence.

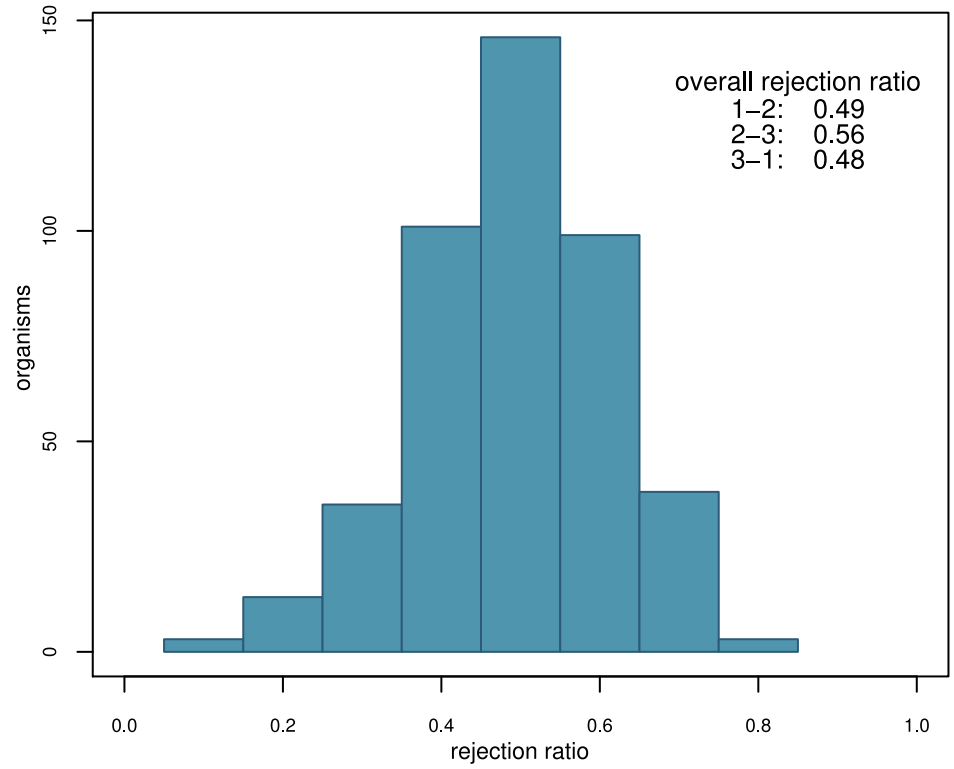

Figure 2.7: Histogram of the rejection ratio for all organisms. The rejection ratio is the fraction of tests on an individual organism which reject the hypothesis of ergodicity at a confidence level of $\alpha = 5\%$. The results for all three dinucleotides have been added up. The overall rejection ratio considers the individual dinucleotides summarised over all organisms.

The test results are given in 2.7. The average rejection ratio over all organisms and dinucleotides at a confidence level of $\alpha = 5\%$ is 51%, where the variation for the different dinucleotides is small compared to the variation over the different organisms. The rejection rate for individual organisms ranges between 13% for *Sinorhizobium meliloti (strain 1021)* and 83% for *Xylella fastidiosa (strain 9a5c)*.

Markov Property

A stochastic process X has the Markov property, if

$$\mathbb{P}\left(X_n = s | X_0 = s_0, X_1 = s_1, \ldots, X_{n-1} = s_{n-1}\right) = \mathbb{P}\left(X_n = s | X_{n-r} = s_{n-r}, \ldots, X_{n-1} = s_{n-1}\right)$$

and is homogeneous if

$$\mathbb{P}\left(X_n = s | X_{n-r} = s_{n-r}, \ldots, X_{n-1} = s_{n-1}\right) = \mathbb{P}\left(X_r = s | X_0 = s_0, \ldots, X_{r-1} = s_{r-1}\right)$$

and is then called a homogeneous Markov chain of order k (42; 30). These processes are well-studied and a lot of standard methodology is at hand for them (93; 81). Therefore, it would be interesting to demonstrate Markov property for our genome data.

We assume Markov property and test for the order of the chain as introduced in (1) and implemented in (106) with the hypotheses

$$H_0: \text{ Markov chain of order } k \quad \text{vs.} \quad H_1: \text{ Markov chain of order } k+1.$$

Starting at $k = 0$, we test independence against first order Markov chain, whereby the test results in Pearson's χ^2-test for independence.

Upon rejection of the null hypothesis, we repeat the test for $k = 1$ and so forth until either the order k of the chain is found or $k+1$ becomes too large for the test to be performed in practice.

In a genome, we summarise the transition counts of the dinucleotides in a gene over all genes in order to get numbers that enable us to test for orders up to $k = 3$. As an average gene length is ≈ 270, already the transition count table for independence against first order with its $16^2 = 256$ entries would not be sufficient for the estimation of transition probabilities if only one gene was considered.

After the data preparation, we perform the test up to the order $k = 3$. If our data happen to follow the Markov property, but for a higher order than k, the property is still not useful for the purpose of data analysis because of the typical data length.

dinucleotide	1-2	2-3	3-1
order 0	$0.00 \cdot 10^{0}$	$0.00 \cdot 10^{0}$	$0.00 \cdot 10^{+0}$
order 1	$1.08 \cdot 10^{-300}$	$7.39 \cdot 10^{-171}$	$8.28 \cdot 10^{-75}$
order 2	$9.71 \cdot 10^{-267}$	$1.10 \cdot 10^{-107}$	$1.49 \cdot 10^{-67}$
order 3	$0.00 \cdot 10^{0}$	$0.00 \cdot 10^{0}$	$0.00 \cdot 10^{0}$

Table 2.1: The p-values of the Markov test for orders 0 *to* 3 *for all three dinucleotides. All gene sequences in a organism are used to perform this test. (The values are taken from the organism* Bertonella basilliformis (strain KC583)*.)*

As can be seen in the examplary table 2.1, the hypothesis of independence is rejected in favour of more dependence, here order 1. Next, the hypothesis of order 1 is rejected in favour of even more dependence, order 2. But the same happens to orders 2 and 3. This is shown for all three dinucleotides. In fact, the rejection happens on any confidence level as the p-values are (close to) vanishing.

Therefore, we conclude that either the Markov chain is of higher order than 3 or our assumption of the Markov property in the genetic sequence was wrong in the first place. In both cases, we cannot make use of the Markov property in our data.

Conclusion

The sequences of genomic data that we would like to analyse display convergence. Furthermore, we can assume stationarity within them with confidence. The assumption of ergodicity is somewhat more problematic with some variation over the different organisms. Still, a reasonable fraction of the tests do not reject ergodicity. Therefore, we can safely make use of this assumption for practical purpose, especially, since we do not heavily rely on it.

On the other hand, the data do definitely not satisfy the Markov property. Although many models assume it nonetheless, we decided against it and build a model with a higher dependence structure. This model is introduced in the next chapter.

3 Model

This chapter is meant to introduce the developed model. First, the distribution of interest is presented. In the second part, our model is developed and the application of the distribution is demonstrated.

3.1 Noncentral Hypergeometric Distribution

Before describing the noncentral hypergeometric distribution, some basic properties of the central hypergeometric distribution[1] ought to be recalled.

3.1.1 Urn Model

Assume an urn with balls of k different colours $C_1, \ldots, C_k$, whereby there urn contains m_j balls of colour C_j with a total of $m = \sum_{j=1}^{k} m_j$ and written as $\boldsymbol{m} = (m_1, \ldots, m_k)$. Then, n balls are drawn without replacement. The arising questions are

1. How many balls of each colour have been drawn after n steps?
2. What is the probability of a given set of ball colours to be drawn?

The answers are provided by the hypergeometric distribution which represents this urn model.

[1] This distribution is usually called the *hypergeometric distribution*. The adjective *central* is only used to emphasise the difference.

Hypergeometric Distribution

The known standard hypergeometrical distribution shows no dependance between the colour of a ball in an urn and its probability to be drawn. The only influencing paramater is the number of balls of the different colours in the urn.

The probability to draw a ball of colour C_i is its relative frequency in the urn. Is x_j the number of balls of colour C_j for $j = 1, \ldots, k$ in the urn at a certain time, the relative frequency for the next step is

$$\mathbb{P}\left(\text{ball of colour } C_i \text{ drawn}\right) = \frac{x_i}{\sum_{j=1}^{k} x_j}$$

for $i = 1, \ldots, k$.

Noncentral Hypergeometric Distribution

Does one want to model the preferences in drawing balls of different colours, weight parameters are introduced. The resulting distribution is then called the *noncentral hypergeometric distribution* and was developed by Wallenius in (112) for the bivariate case and extended to a multivariate distribution by Chesson in (13), an alternative formulation was given in (69).

The set-up remains the same. What changes, is the probability to draw a ball in a single step. Instead of the relative frequency we have the weighted relative frequency for $i = 1, \ldots, k$,

$$\mathbb{P}\left(\text{ball of colour } C_i \text{ drawn}\right) = \frac{\beta_i x_i}{\sum_{j=1}^{k} \beta_j x_j},$$

where the parameters $\beta_j \geq 0$ are normed such that $\sum_{j=1}^{k} \beta_j = 1$.

After n steps, the closed-form expression for the outcome $\boldsymbol{n} = (n_1, \ldots, n_k)$ is

$$\mathbb{P}\left(\boldsymbol{n}\right) = A \cdot \int_0^1 \left(\prod_{j=1}^{k} (1 - t^{\beta_j c})^{n_j} \right) dt \tag{3.1}$$

$$= A \cdot \sum_{l_1=0}^{n_1} \cdots \sum_{l_k=0}^{n_k} \left(\prod_{j=1}^{k} \binom{n_j}{l_j} \right) \frac{(-1)^{\sum_{j=1}^{k} l_j}}{c \sum_{j=1}^{k} \beta_j \, l_j + 1} \tag{3.2}$$

where $A = \left(\prod_{j=1}^{k} \binom{m_j}{n_j}\right)$ and $c = (\sum_{j=1}^{k} \beta_j(m_j - n_j))^{-1}$. With $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$ the distribution is noted as

$$\text{hyp}\,(\boldsymbol{m}, \boldsymbol{\beta}, n) \tag{3.3}$$

with $k-1$ parameters $\beta_1, \ldots, \beta_{k-1}$ and $\beta_k = 1 - \sum_{j=1}^{k-1} \beta_j$.

3.1.2 Computation of the Probabilities

As satisfying as it is to have expression 3.2, one can see that its usage might be difficult, or at least limited, if the data $\boldsymbol{m}$ and $\boldsymbol{n}$ take large values.

The R-package `BiasedUrn` by Agner Fog provides a computational solution to this problem using various numerical concepts (33; 34). However, it was not available when we started working with the distribution. Therefore, we developed our own solution that is presented below. It provides identical performance in terms of accuracy and run time.

The numbers that are dealt with are tiny, particularly for larger n, but already for smaller ones too, we had to make use of a special precision arithmetic library. Here, both multiplication and division are not computable in linear time but in $\mathcal{O}(n^2)$.

Exact Sums

The use of the summation formula 3.2 on the facing page is not recommended for larger data. Its running time can be bounded by the following argument. The number of summands is largest if all n_i for $i = 1, \ldots, k$ are of similar size, which would be $\frac{n}{k}$. Then, we have approximately $\left(\frac{n}{k}\right)^k$ summands. For each summand we need to compute the binomial coefficient and the fraction. These are $\mathcal{O}(n^2)$ operations each in the precision arithmetic library. Multiplying by the number of summands, this yields to $\mathcal{O}(n^{k+2})$.

However, for smaller data this formula is perfectly fine as it is exact and not slower than any approximation.

Larger data need to be considered differently. Here, we use the integral formula 3.1 and approach it numerically.

Numerical Integration

To this means, we have a closer look at the functions

$$f(t) = \prod_{j=1}^{k} \underbrace{(1 - t^{\beta_j c})^{n_j}}_{=:f_j(t)}$$

to be integrated. Examples can be seen in figures 3.2 on page 30 through 3.4 on page 32 as the left upper functions for different data sizes.

Independently of the values of $\boldsymbol{m}$ and $\boldsymbol{n}$, we have $f(0) = 1$ and $f(1) = 0$ for all parameters β. For most parameter settings, the function f is stricly monotonic, convex and very close to the axes. As we exploit the monotony later, we now look at the first derivative

$$\begin{aligned} f'(t) &= \sum_{j=1}^{k} f_j'(t) \prod_{\substack{i=1 \\ i \neq j}}^{k} f_i(t) \\ &= \sum_{j=1}^{k} \left[-\underbrace{\beta_j c n_j}_{>0} \underbrace{t^{\beta_j c - 1}(1 - t^{\beta_j c})^{n_j - 1}}_{>0} \prod_{\substack{i=1 \\ i \neq j}}^{k} \underbrace{(1 - t^{\beta_i c})^{n_i}}_{>0} \right] < 0 \end{aligned}$$

for $t \in (0, 1)$ which shows the claim.

Integration Approach

Hereby, a direct quadrature of f is impractical for $t \to 0$. Therefore, we divide the area into three parts and make use of the inverse function f^{-1}.

Let t_x be the fix point of f with $f(t_x) = t_x$. Then, we have

$$\int_0^1 f(t)\, dt = A + B + t_x^2$$

with

$$A := \int_{t_x}^{1} f(t)\,dt \qquad \text{and} \qquad B := \int_{t_x}^{1} f^{-1}(\xi)\,d\xi$$

as shown in figure 3.1.

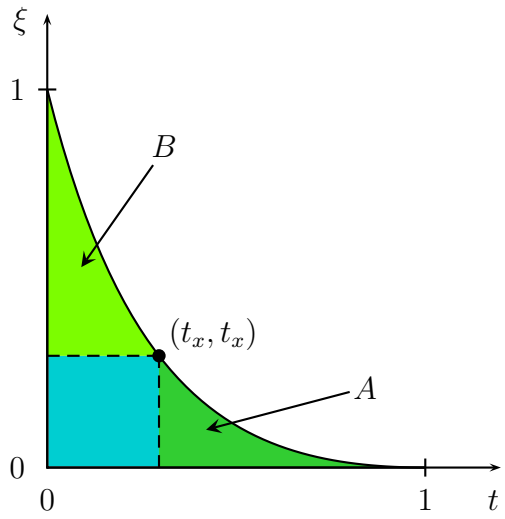

Figure 3.1: Division of the integral area into three parts.

Additionally, using a doubly logarithmic transform we reduce the range of the values. We define $\tau := \ln(t)$ and $F(\ln(t)) := \ln(f(t))$, which in combination yield to

$$F(\tau) = \ln f(\exp(\tau)) = \sum_{j=1}^{k} n_j \ln(1 - \exp(\beta_j c\tau)).$$

Because of the Taylor approximation, we have $\ln(1-t) \approx -t$ for t close to 0 and $0 < t < \delta$, and thus obtain for $\exp(\beta_j c\tau) \ll 1$ the approximation

$$F(\tau) \approx -\sum_{j=1}^{k} n_j \exp(\beta_j c\tau),$$

which we use for computation.

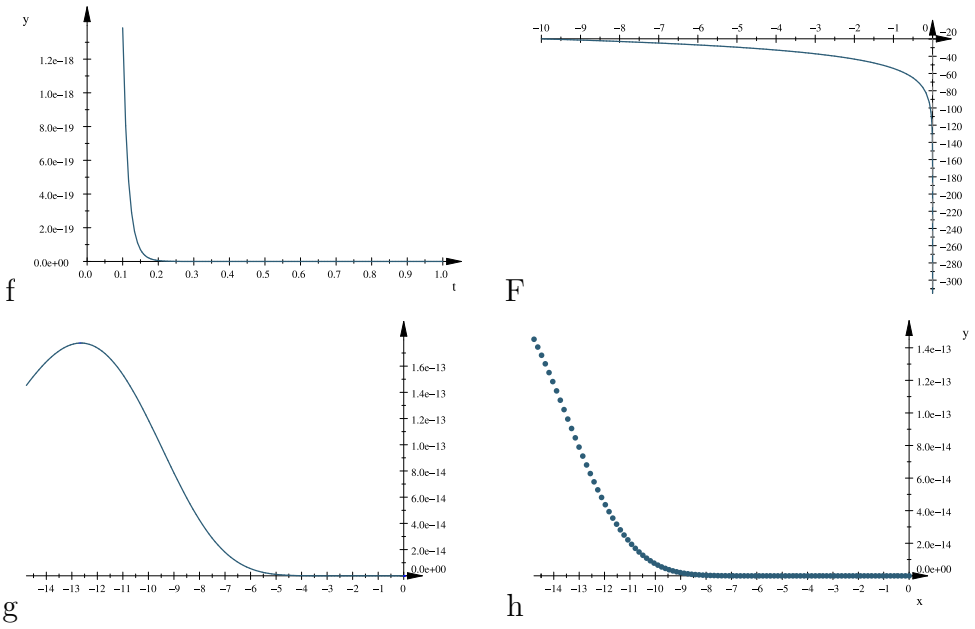

Figure 3.2: This and the following two figures show graphs of the functions f, F, g and h that play a role in the integration. Here, the functions are shown for small data sizes $\boldsymbol{m} = (10, 5, 15, 10)$, $\boldsymbol{n} = (7, 1, 3, 5)$. We obtain $\tau_x = -14.78$.

Now, we approach the integral. Because of $t = \exp(\tau)$ we have $dt = \exp(\tau)d\tau$. For the point t_x we define $\tau_x = \ln(t_x)$ and get $F(\tau_x) = \tau_x$. By substitution of these transforms into A we have

$$\begin{aligned} A &= \int_{t_x}^{1} f(t)\,dt \\ &= \int_{\tau_x}^{0} \underbrace{f(\exp(\tau))\exp(\tau)}_{=:g}\,d\tau \\ &= \int_{\tau_x}^{0} \exp\left[\sum_{j=1}^{k} n_j \ln(1-\exp(\beta_j c\tau))\right]\exp(\tau)\,d\tau \end{aligned}$$

where $g(\tau) = \exp[F(\tau)]\exp(\tau)$. Analogously, we proceed with the inverse

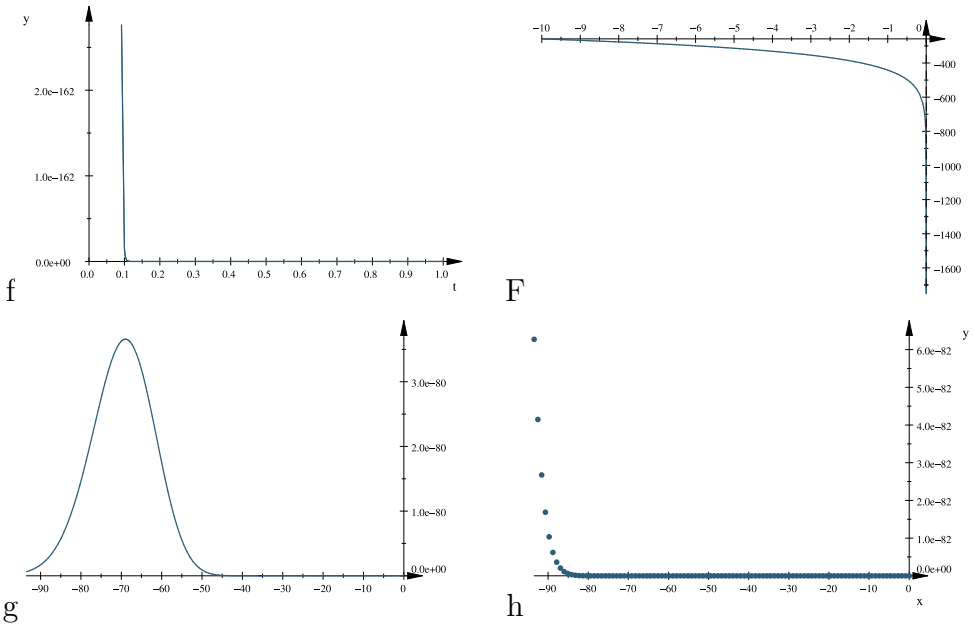

Figure 3.3: Integration with intermediate data sizes $\boldsymbol{m} = (75, 56, 90, 68)$, $\boldsymbol{n} = (24, 10, 35, 11)$, *result in* $\tau_x = -98.49$.

function f^{-1}. Here, we substitute $\xi = \exp(\zeta)$. As $t_x = \xi_x$, we have

$$\begin{aligned} B &= \int_{\xi_x}^{1} f^{-1}(\xi)\, d\xi \\ &= \int_{\zeta_x}^{0} \underbrace{f^{-1}(\exp(\zeta))\exp(\zeta)}_{=:h}\, d\zeta\,, \end{aligned}$$

Since f^{-1} is unknown, values needed for the computation of the integral cannot be obtained directly. To this end, we rewrite h in terms of F as $h(\zeta) = \exp(F^{-1}(\zeta))\exp(\zeta)$ and compute the values using Newton's algorithm.

We evaluate both integrals using the trapezoidal rule. However, for smaller data sizes, special care needs to be taken here when approaching 0. Hereby, the Romberg-method could be applied (20).

Last, we compute the square t_x^2 as $t_x^2 = \exp(2\tau_x)$.

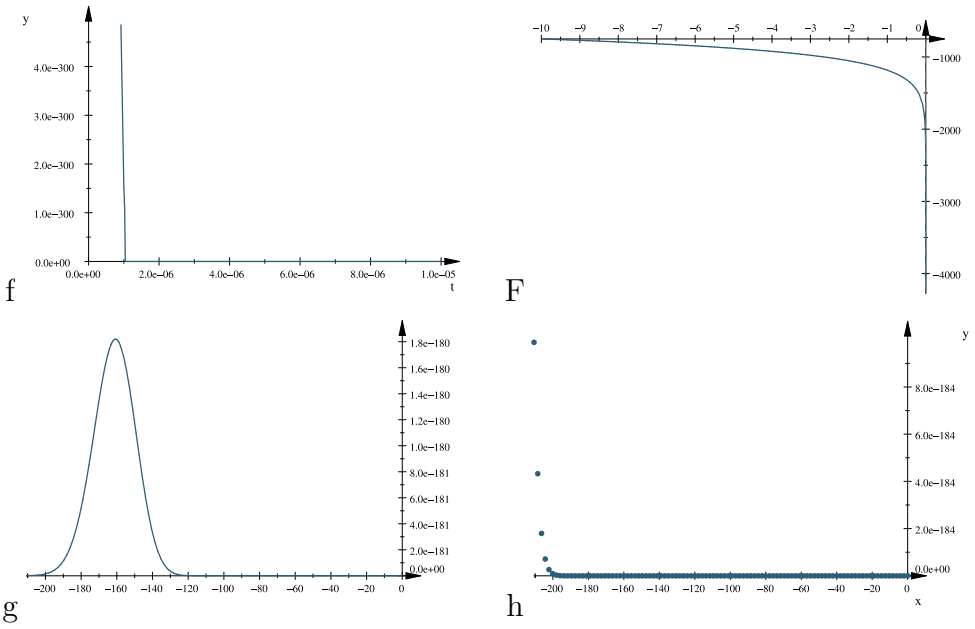

Figure 3.4: Integration with large data sizes $\boldsymbol{m} = (195, 141, 237, 116)$, $\boldsymbol{n} = (57, 26, 76, 30)$, *leads to* $\tau_x = -210.69$.

The functions f, F, g and h are shown in figures 3.2 through 3.4 for different sizes of $\boldsymbol{m}$ and $\boldsymbol{n}$. Although the general behaviour of the functions remains similar, the obtained values differ significantly as can be seen on the axis. The lower two graphs also show, that the proportion of the integral contributed by h decreases as the data sizes increase. The parameter k is always 4 as this is the value which shall be used later for our model.

Implementation

As mentioned above, we use both computational methods, the summation formula as well as the numerical integration, in our implementation. The summation formula is applied to smaller data, an approximate urn size of

$$\sum_{i=i}^{k} m_i \leq 15 \cdot k\,,$$

whereas the numerical integration finds its use for larger urns.

This way, we have one's cake and eat it too. We get the exact values if they

are fast to compute, we avoid the use of the Romberg-method for smaller data sizes which would slow down our computation and we have a fast method and good approximation for larger data.

3.1.3 Estimation of the Parameters

Because of the complexity of the distribution, it is not known how the maximum likelihood estimators could be computed. Nevertheless, a suitable approximation using the Newton-Raphson method was presented by Manly in (70). The estimators are given as

$$\hat{\beta}_i = \frac{\log(\frac{m_i - n_i}{m_i})}{\sum_{j=1}^{k} \log(\frac{m_j - n_j}{m_j})} \tag{3.4}$$

for all $i = 1, \ldots, k$.

As Manly mentioned, these estimators are almost accurate with vanishing biases given that the numbers of balls n_j as well as $m_j - n_j$ are larger than five.

However, for our purpose it is important to get a quantitative idea of the behaviour of the estimators depending on the urn and sample size.

Convergence of the Estimators

In our model, we will be using $k = 4$ and therefore this is the case that we constrain our analysis to. We choose an arbitrary but realistic urn composition and an arbitrary set of parameters, namely $\boldsymbol{m} = (5, 8, 5, 2)$ and $\boldsymbol{\beta} = (0.2, 0.15, 0.2, 0.45)$. This urn composition is then gradually scaled by a common factor. The number n of balls drawn from the urn is chosen as a constant fraction of the total urn size and scaled accordingly.

Explicitely, we look at the fractions $1/4$, $1/3$ and $1/2$ as these naturally occur when our model is defined in part 3.2.

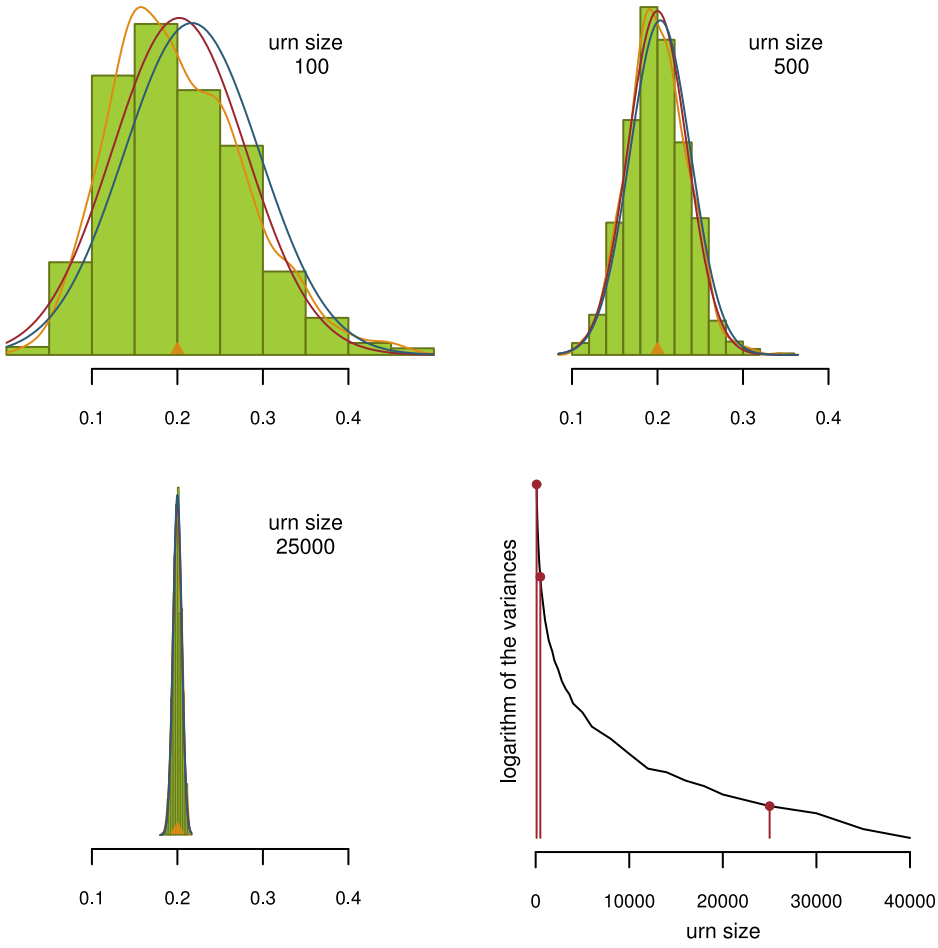

Figure 3.5: Histograms of an estimated parameter in urns of three given sizes. The graph in the lower right corner shows the sample variance of the estimator as a function of the urn sizes. Those sizes used for the histograms are highlighted. The shrinking variance of the estimators is clearly visible and manifests itself also in the width of the histograms. The displayed curves are (1: orange) empirical density estimation, (2: red) normal density with empirical mean and variance and (3: blue) normal density with theoretical mean and variance as given in (70, Manly). The theoretical normal density is systematically shifted towards higher values. All graphs refer to $\beta_i = 0.2$ as marked and $m_i/\text{urn size} = 1/4$. The fraction of balls drawn from the urn is $1/4$.

For a given urn we simulate 1000 times the scenario of drawing n balls and estimate the parameters $\boldsymbol{\beta}$ using the formula 3.4 defined above. The key quantity to judge their variability is the sample variance. In figure 3.5 on the facing page the histograms of one estimated parameter for different urn sizes are displayed. Additionally, its sample variance is shown as a function of all evaluated urn sizes. These simulations allow in principle for a rough estimate of the expected variance in the application to real world data which we shall come back to in chapter 4.

With respect to applications, it is interesting to consider the effect of more extreme urn and weight parameter configurations. An example of such a situation is given in figure 3.6.

Extreme Urn Events

In extreme cases, estimating the β_i by formula 3.4 is not possible. These cases are, in general, those where one of the involved numbers n_j or $m_j - n_j$ equals 0 for at least one j. Hereby, the estimator is 0 or $\log(0)$ appears, respectively. In the following, we are going to look at these cases more closely and propose some alterations. Though it is clear, that they do not approximate the real parameters too well, their advantage is visible on the second glance, the practical one. Since these cases are rare but real events in our data, they need to be taken care of in one way or the other. Without doing so, a program would just stop at such an incident which definitely is undesired behaviour. Excluding these data from the data set also constitutes an objectionable distortion.

We have considered the following possibilities of dealing with the $\log(0)$ issue:

- Adding an integer pseudo count to all cases. This brute force method circumvents the problem by increasing all m_i to $m_i + 1$. The price to pay is a degradation of the estimators even in non-critical cases.
- Adding an integer pseudo count to critical values. In case the urn is exhausted in at least one colour, say i, one can either increase m_i to $m_i + 1$ or decrease n_i to $n_i - 1$. In practice, the difference between these two choices is negligible. The ad hoc character of this method might cause some concern.

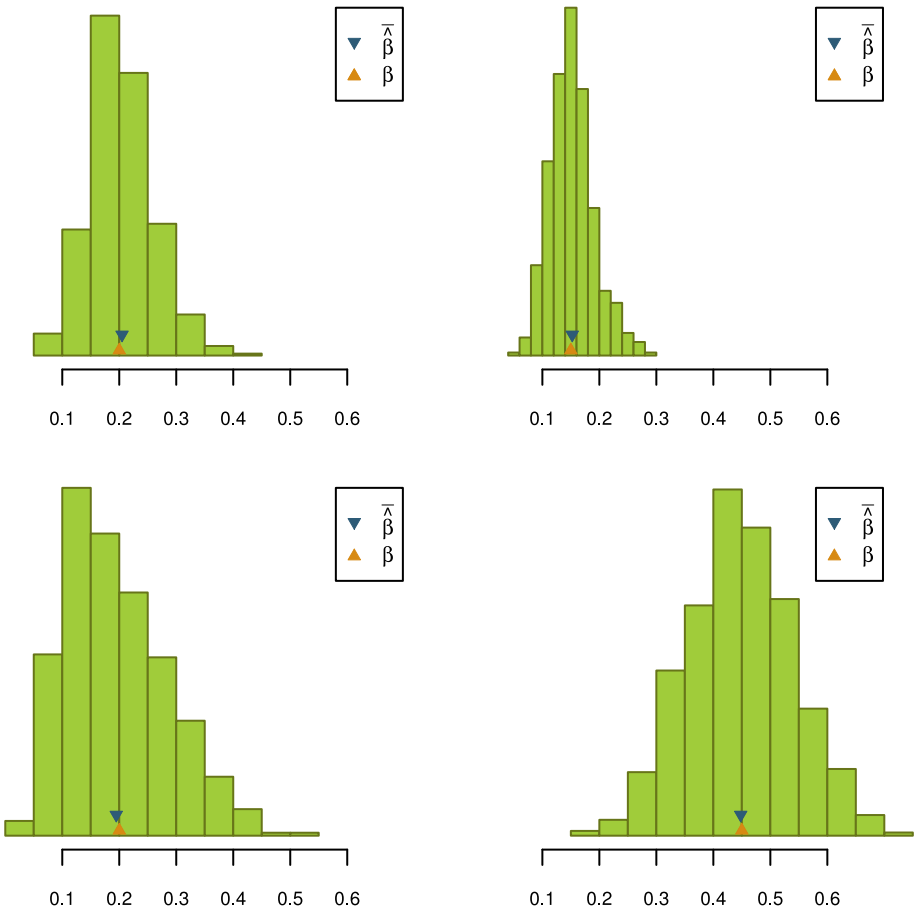

Figure 3.6: Histogram of the estimated parameters in case of an almost extreme urn composition with $\boldsymbol{m} = (50, 80, 10, 20)$ *and* $n = 50$ *where the third value (lower left graph) is highly critical and the fourth (lower right graph) slightly. It demonstrates that on average the estimation is correct. However, the individual cases can result in very different parameter estimators. The estimators for the non-critical ball colours in the urn are basically unaffected.*

- Using a small fractional pseudo count p in all cases. Here, all m_i are increased to $m_i + p$. In contrast to the previous two methods, the effect on non-critical values is insubstantial.

- Neutralising the critical cases. First, the non-critical estimators are calculated as if the critical ones were not there. The critical estimators are then assigned the same weight as the sum of the weights of the non-critical ones. Finally, all estimators are renormalised. This method provides the least bias and a tolerable variance. Therefore, it will be used in the following.

	bias		variance	
method	critical	non-critical	critical	non-critical
always integer pseudo	0.0721	0.0209	0.04289	0.00587
critical integer pseudo	0.2367	0.0583	0.00235	0.00115
fractional pseudo	0.0660	0.0191	0.04688	0.00610
neutralising	0.0653	0.0189	0.04736	0.00612

Table 3.1: Table of the alternative procedures in dealing with extreme events as described in main text. The biases and variances are shown examplaryly for a critical estimator and a non-critical one.

3.2 Modelling

In this part, we develop our model. Hereby, we design a new view on the dinucleotides and combine it with the noncentral multivariate hypergeometric distribution. Firstly, we define a probability space for the dinucleotides.

3.2.1 The Probability Space and Random Variables

Given $n \in \mathbb{N}$, let

$$\begin{aligned}\Omega &= \{\omega = (\omega_1, \omega_2, \ldots, \omega_n) |\, \omega_i \in \{A, C, G, T\}^2 \text{ for } i = 1, \ldots, n\} \\ &= (\{A, C, G, T\}^2)^n\end{aligned}$$

denote the set of all possible sequences of dinucleotides of length n. Furthermore, let $Y = (Y_1, Y_2, \ldots, Y_n)$ be a random variable on Ω with the mapping

$$\begin{aligned}Y_i : \Omega &\longrightarrow \{A, C, G, T\}^2 \\ \omega &\longmapsto \omega_i\end{aligned}$$

for all $i \leq n$. One should note that ω_i itself is twodimensional. Strongly speaking, we have $\omega_i = ((\omega_i)_1, (\omega_i)_2)$ and $Y_i = ((Y_i)_1, (Y_i)_2)$, analogously.

In addition, let $S^0, S^1, \ldots, S^n$ be the sequence of matrices

$$S_g^m := (S_{i,j;g}^m)_{(i,j)} := \left(\sum_{k=1}^{m} \mathbb{1}_{\{Y_k=(g(i),g(j))\}} \right)_{(i,j)} \tag{3.5}$$

for $m = 0, \ldots, n$ with $i, j \in \{1, 2, 3, 4\}$ and g being any bijection between $\{1, 2, 3, 4\}$ and $\{A, C, G, T\}$. Next, we define the ith rows and jth columns of S_g^m as

$$R_{i;g}^m = \left(\sum_{k=1}^{m} \mathbb{1}_{\{Y_k=(g(i),g(j))\}} \right)_{(j=1,\ldots,4)}$$
$$C_{j;g}^m = \left(\sum_{k=1}^{m} \mathbb{1}_{\{Y_k=(g(i),g(j))\}} \right)_{(i=1,\ldots,4)}$$

for $i, j = 1, 2, 3, 4$.

The definition of the summation 3.5 is similar to the sums introduced in 2.1 on page 12. However, the difference is the twodimensional structure.

Looking at the rows or columns of the matrices allows fixing a dimension of ω_i and with it a nucleotide on the first or the second position of the dinucleotides, respectively, and concentrate on its partner nucleotide. The effect is the basis of the definition of the rows and columns and is visible therein. Furthermore, the rows and columns will be used to exploit this trait.

In the following, we rarely need all indices and reduce the notation to S_g where appropriate. Similarly, we use $R_{i;g}$ and $C_{j,g}$.

Probability

We want to have a measure for the pair building behaviour and not the single counts. Therefore, we define the probability of S_g conditional on row and column sums.

To this means, we note that

$$\Omega = \Omega' \cup Q,$$

where for given row and column sums $\boldsymbol{r} = (r_1, r_2, r_3, r_4)$ and $\boldsymbol{c} = (c_1, c_2, c_3, c_4)$ the path $\omega \in \Omega'$, if its respective sums match $\boldsymbol{r}$ and $\boldsymbol{c}$. Otherwise, we have $\omega \in Q$.

Clearly, we want to set the whole measure on those paths which have the correct row and column sums. We thus define $\mathbb{P}(Q) = 0$.

On Ω' we define the probability of S_g in terms of its four rows $R_{i;g}$ using the conditional probabilities

$$\begin{aligned}\mathbb{P}(S_g) &= \mathbb{P}(R_{1;g}, R_{2;g}, R_{3;g}, R_{4;g}) \\ &= \mathbb{P}(R_{1;g}) \cdot \mathbb{P}(R_{2;g}|R_{1;g}) \cdot \mathbb{P}(R_{3;g}|R_{1;g}, R_{2;g}) \cdot \mathbb{P}(R_{4;g}|R_{1;g}, R_{2;g}, R_{3;g})\end{aligned}$$

and the noncentral hypergeometric distribution for each of its rows which we set to

$$\begin{aligned}P_{R_{1;g}} &\sim \operatorname{hyp}(\boldsymbol{c}, &&\boldsymbol{\beta}^{(1)}, r_1) \\ P_{R_{2;g}|R_{1;g}} &\sim \operatorname{hyp}(\boldsymbol{c} - R_{1;g}, &&\boldsymbol{\beta}^{(2)}, r_2) \\ P_{R_{3;g}|R_{1;g},R_{2;g}} &\sim \operatorname{hyp}(\boldsymbol{c} - R_{1;g} - R_{2;g}, &&\boldsymbol{\beta}^{(3)}, r_3) \\ P_{R_{4;g}|R_{1;g},R_{2;g},R_{3;g}} &\sim \operatorname{hyp}(\boldsymbol{c} - R_{1;g} - R_{2;g} - R_{3;g}, &&\boldsymbol{\beta}^{(4)}, r_4).\end{aligned}$$

This way, we have defined a probability on all of Ω.

Model Parameters

It is important to note, that since $\boldsymbol{c}$ is fixed, the distribution of the fourth row is fully determined by the first three rows and therefore

$$\operatorname{hyp}(\boldsymbol{c} - R_{1;g} - R_{2;g} - R_{3;g}, \boldsymbol{\beta}^{(4)}, r_4) \equiv 1$$

yielding to

$$\mathbb{P}(S_g) = \mathbb{P}(R_{1;g}) \cdot \mathbb{P}(R_{2;g}|R_{1;g}) \cdot \mathbb{P}(R_{3;g}|R_{1;g}, R_{2;g}). \tag{3.6}$$

Consequently, the parameters of the model are given by $\beta = (\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \boldsymbol{\beta}^{(3)})$, each of length 4 as defined in 3.3 on page 27, resulting in 9 parameters.

Well-Definedness

There are two more points that need to be taken into account. Firstly, we could have defined the probability in terms of the columns instead of the rows.

Additionally, we have so far defined S_g with a particular bijection g. Making use of a different bijection would result in an altered matrix.

In both cases, the definition of the probability would yield to different model parameters. We have to ensure that different choices would have negligible effects. First, we will consider the effect of a change of bijection and let the transposed approach follow.

3.2.2 Equivalence of the Different Models

In the following, we want to look at the influence of the choice of the bijection g on the models.

When defining the matrix S in 3.5 on page 38 we used an arbitrary bijection

$$g : \{1, 2, 3, 4\} \longrightarrow \{A, C, G, T\}.$$

We have then set the probability of the matrix S as the product of the probabilities of its rows.

However, the probabilities of the rows depend on their order in S, which is given by g. Therefore, we need to show that for two different bijections g_1 and g_2 the resulting models with the matrices S_{g_1} and S_{g_2} and their probabilities are equivalent and our model is indeed well-defined. The figure 3.8 on page 43 displays this for special cases as discussed below.

There are $4! = 24$ different choices for g which yield to $(24-1)! \approx 2.59 \cdot 10^{22}$ pairwise comparisons of the models, or at least 23 when comparing one model to all the others. We do not have to perform this tedious task. Instead, we first analyse the bijections more closely.

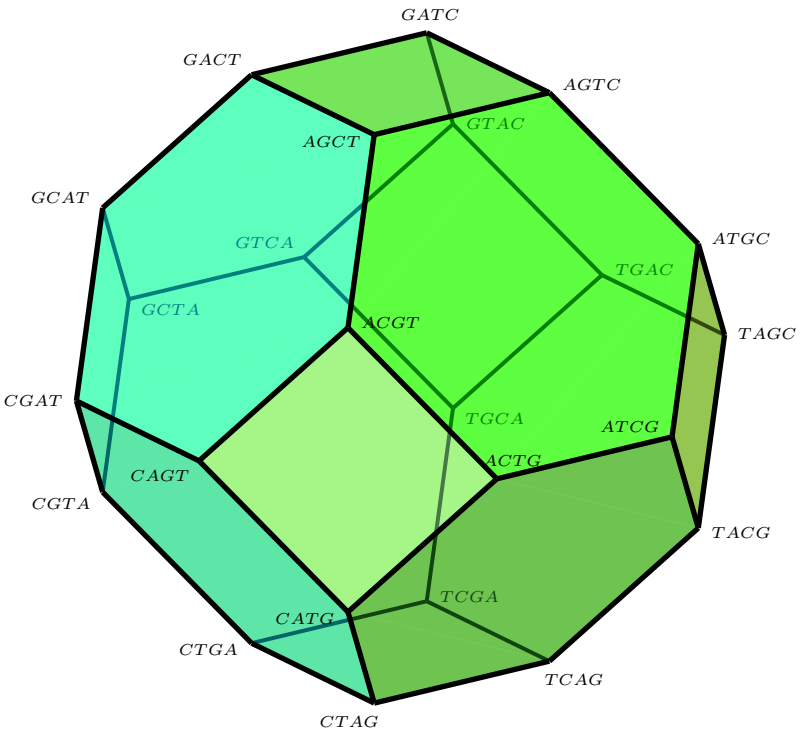

Figure 3.7: A truncated octahedron whose vertices represent the 24 equivalent models. The edge between two vertices stands for a swap of two sequential steps in the process of model building.

Bijections

The differences between two bijections g_1 and g_2 can be described by a permutation on the set $\{1, 2, 3, 4\}$ yielding to

$$g_2 = g_1 \circ \pi$$

We define two bijections g_1 and g_2 as *neighbouring* if there exists a neighbour transposition τ such that

$$\begin{aligned} &\tau(j) = j + 1 \text{ and } \tau(j + 1) = j && \text{for some } j \in \{1, 2, 3, 4\} \\ &\tau(i) = i && \text{for } i \in \{1, 2, 3, 4\} \backslash \{j, j + 1\} \end{aligned}$$

Connecting neighbouring bijections, we can draw a truncated octahedron whose edges are all 24 bijections as in figure 3.7. In fact, we have just mapped

the permutation group $\mathbb{S}_4$ on the 24 different models $\{S_g |\ g : \{1,2,3,4\} \longrightarrow \{A,C,G,T\}$ bijection$\}$.

$$\begin{array}{ccc} \{1,2,3,4\} & \xrightarrow{g_2} & S_{g_2} \\ \Big\downarrow \pi & & \Big\downarrow \text{edge of the octahedron} \\ \{1,2,3,4\} & \xrightarrow{g_1} & S_{g_1} \end{array}$$

Therefore, any two bijections are connected by a path of transpositions. Since every permutation in $\mathbb{S}_4$ is the product of neigbouring transpositions, we only have to scrutinise the equivalence of neigbouring bijections, i. e. adjacent edges on the truncated octadedron.

Hereby, the necessary comparisons of the models are reduced to the following three cases of the three possible transpositions

$$\begin{array}{lll} \tau_1 : & \tau_1(1) = 2 & \\ & \tau_1(2) = 1 & \\ & \tau_1(i) = i \text{ for } i \in \{3,4\} & \end{array} \qquad \begin{array}{lll} \tau_2 : & \tau_2(2) = 3 & \\ & \tau_2(3) = 2 & \\ & \tau_2(i) = i \text{ for } i \in \{1,4\} & \end{array}$$

$$\begin{array}{lll} \tau_3 : & \tau_3(3) = 4 & \\ & \tau_3(4) = 3 & \\ & \tau_3(i) = i \text{ for } i \in \{1,2\} & \end{array}$$

that correspond to the matrix changes as shown in the upper two diagrams in figure 3.8 on the facing page.

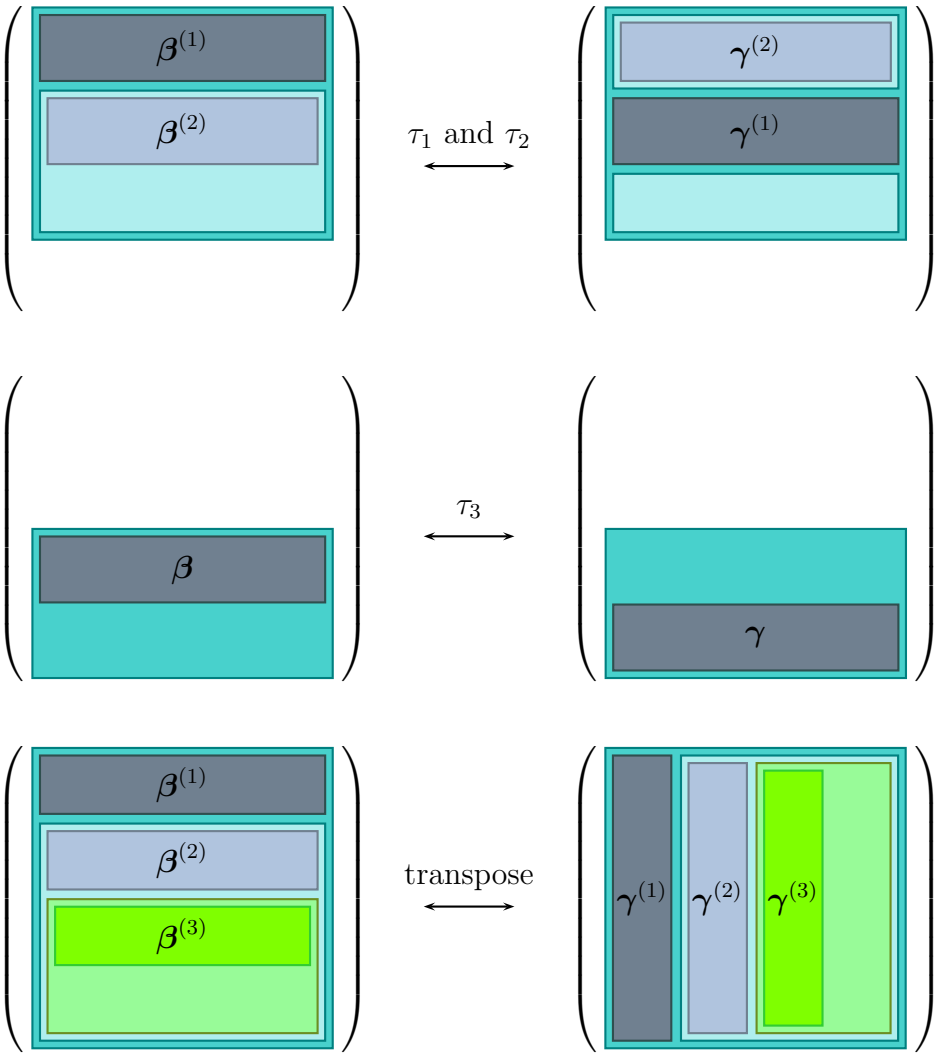

Figure 3.8: The first diagram shows the effect of the transition τ_1 on the models. As the last row of the matrix does not play any role, it could be removed. Then, the boxes can be placed over the 2nd through 4th row, and we get the description of the transition τ_2. The diagrams in the middle display the effect of transition τ_3, where only one set of parameters is involved. The third diagram represents the model change if the matrix is transposed. Hereby, the probabilities are set to the columns instead of the rows of the matrix and we have vertical sets of parameters.

Equivalence

In the following, we want to scrutinise these transpositions and their effect on the probability of the matrix as defined in formula 3.6. Here, the first three rows provide a set of parameters for the noncentral hypergeometric distribution each, whereas the last row does not.

Therefore, the first observation is that τ_1 and τ_2 yield to the same case. They both affect two sets of parameters. Indeed, any previous or following rows and their parameters stay untouched.

Let $g_2 = g_1 \circ \tau_1$. Then

$$\begin{aligned} R_{1;g_1} &= R_{2;g_2} \\ R_{2;g_1} &= R_{1;g_2} \\ R_{3;g_1} &= R_{3;g_2} \\ R_{4;g_1} &= R_{4;g_2} \end{aligned}$$

and with equation 3.6 yielding to

$$\begin{aligned} \mathbb{P}(S_{g_1}) &= \mathbb{P}(R_{1;g_1}) \cdot \mathbb{P}(R_{2;g_1}|R_{1;g_1}) \cdot \mathbb{P}(R_{3;g_1}|R_{1;g_1}, R_{2;g_1}) \\ &= \mathbb{P}(R_{2;g_2}) \cdot \mathbb{P}(R_{1;g_2}|R_{2;g_2}) \cdot \mathbb{P}(R_{3;g_2}|R_{1;g_2}, R_{2;g_2}) \\ &\simeq \mathbb{P}(R_{1;g_2}) \cdot \mathbb{P}(R_{2;g_2}|R_{1;g_2}) \cdot \mathbb{P}(R_{3;g_2}|R_{1;g_2}, R_{2;g_2}) \\ &= \mathbb{P}(S_{g_2}) \end{aligned}$$

we have

$$\begin{aligned} &\mathbb{P}(S_{g_1}) \simeq \mathbb{P}(S_{g_3}) \\ \Leftrightarrow\ &\mathbb{P}(R_{1;g_1}) \cdot \mathbb{P}(R_{2;g_1}|R_{1;g_1}) \simeq \mathbb{P}(R_{1;g_2}) \cdot \mathbb{P}(R_{2;g_2}|R_{1;g_2}). \end{aligned}$$

Analogously, we get for $g_3 = g_1 \circ \tau_2$ that $\mathbb{P}(S_{g_1}) \simeq \mathbb{P}(S_{g_2})$ is equivalent to

$$\mathbb{P}(R_{2;g_1}|R_{1;g_1}) \cdot \mathbb{P}(R_{3;g_1}|R_{1;g_1}, R_{2;g_1}) \simeq \mathbb{P}(R_{2;g_3}|R_{1;g_3}) \cdot \mathbb{P}(R_{3;g_3}|R_{1;g_3}, R_{2;g_3}).$$

Therefore, we observe that τ_1 and τ_2 indeed represent the same case.

The transposition τ_3 however is different as there is only one set of parameters in one of the exchanged rows involved. Following the same arguments with

$g_4 = g_1 \circ \tau_3$ as above, equation 3.6 results in

$$\begin{aligned} &\mathbb{P}\left(S_{g_1}\right) \simeq \mathbb{P}\left(S_{g_4}\right) \\ \Leftrightarrow\ &\mathbb{P}\left(R_{3;g_1} | R_{1;g_1}, R_{2;g_1}\right) \simeq \mathbb{P}\left(R_{3;g_4} | R_{1;g_4}, R_{2;g_4}\right). \end{aligned}$$

Thus, these two cases will be analysed separately, yet the setting is similar. We apply an equivalence test as introduced below.

In the following, we distinguish the models by the names of their respective parameters. We present the first model as β-model with parameters β, while the notation is γ and γ-model for the second set of parameters and the model, respectively.

It should be noted that the model parameters are indeed different, not simply interchanged, as could be thought at a first glance.

Procedure for the Equivalence Test

Before we can start testing for the equivalence, we need two actually corresponding models that are going to be compared, i. e. we need corresponding sets of β and γ.

Let β be any arbitrary set of possible parameters and m the size of the urn. First, we sample $n = 1000$ times a submatrix of S out of an urn of size $1000 \cdot m$. In each of the n repetitions we apply the transposition τ on the matrix and estimate parameters $\tilde{\gamma}_i$ for $i = 1, \ldots, n$ in the corresponding model. Finally, we define $\gamma := \frac{1}{n} \sum_{i=1}^{n} \tilde{\gamma}_i$.

We can now proceed with the equivalence test.

Let m be the size of the urn which we sample independently $n = 100$ times from a submatrix s_i of S according to the β-model. The parameters $\hat{\beta}_i$ and $\hat{\gamma}_i$ are estimated from s_i and the log-likelihood ratio

$$\lambda_i := \log \frac{\mathcal{L}_{\hat{\beta}_i}(s_i)}{\mathcal{L}_{\hat{\gamma}_i}(s_i)}$$

for $i = 1, \ldots, n$, is computed, where $\mathcal{L}_{\hat{\beta}_i}(s_i) = \mathbb{P}\left(S = s_i | \hat{\beta}_i\right) = \mathbb{P}_{\hat{\beta}_i}(S = s_i)$ is

the likelihood function in the submatrix s_i under the model given by $\hat{\beta}_i$, and $\mathcal{L}_{\hat{\gamma}_i}(s_i)$, analogously.

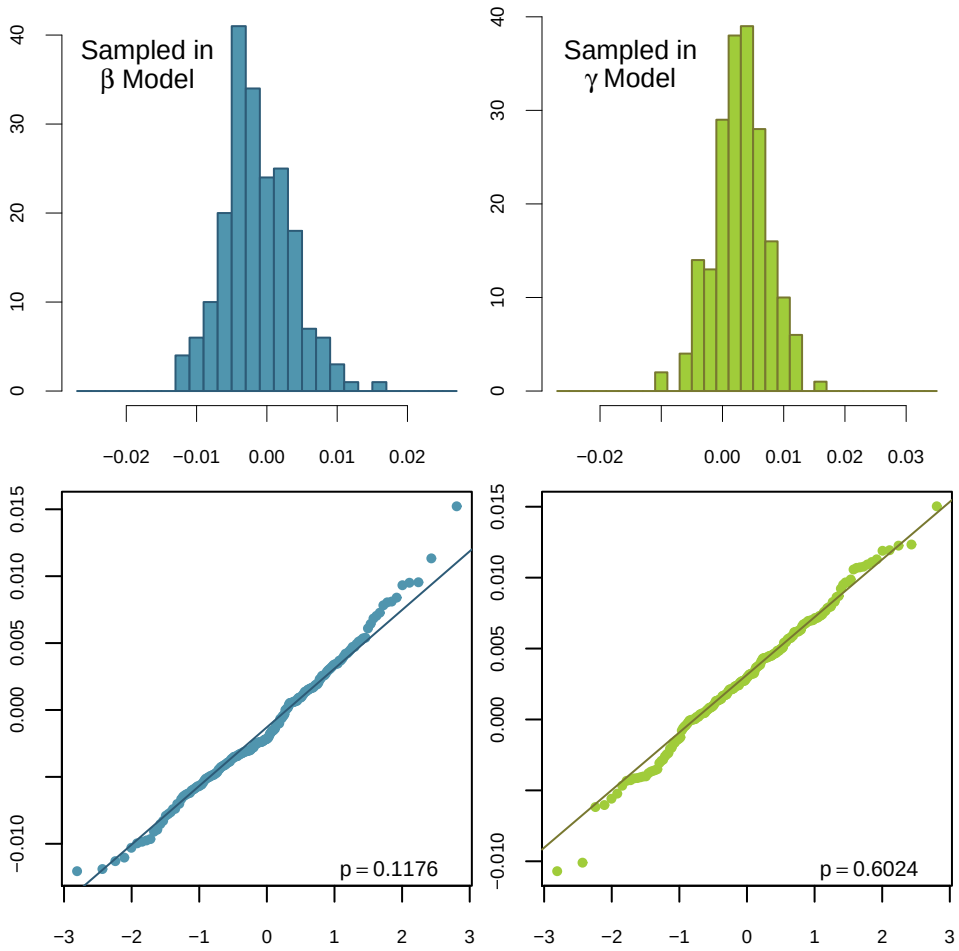

Figure 3.9: The graphs on the left hand side show the histogram of λ_i of the equivalence test in case 1 as described in the main text and their quantile-quantile plot against the normal distribution. The p-value of the Shapiro-Wilk normality test is also given. On the right hand side the roles of β and γ were exchanged.

In the following, λ_i are assumed to be normally distributed. This is backed up by a Quantile-Quantile plot and a Shapiro-Wilk test of normality, see lower left graph in figure 3.9. Furthermore, their variance σ^2 is unknown.

Having the basics settled, we can proceed by using the equivalence test with the confidence interval procedure as defined in (8). Our test statistic on $\lambda = \{\lambda_1, \ldots, \lambda_n\}$ is

$$T(\lambda) = \frac{\sqrt{n}\,\bar{\lambda}}{S_n(\lambda)}$$

and the hypotheses are

$$\begin{aligned} H_0 &: T(\lambda) \notin A \\ H_1 &: T(\lambda) \in A \end{aligned}$$

where $A := [-\varepsilon, \varepsilon]$ is the equivalence region for a given ε.

Furthermore, let K_α be the $(1-\alpha)$ confidence interval of $T(\lambda)$. The hypothesis H_0 will be rejected if $K_\alpha \subset A$.

Case 1: τ_1 and τ_2

In these two settings both models, which are compared, consist of two sets of parameters, e. g. $(\boldsymbol{\beta^{(1)}}, \boldsymbol{\beta^{(2)}})$ or $(\boldsymbol{\beta^{(2)}}, \boldsymbol{\beta^{(3)}})$. This is shown in the upper diagram of figure 3.8. For the equivalence region A, we choose $\varepsilon = \log(1+0.01) \approx 0.00995$ and $\alpha = 0.0001$. Using these values, we perform the equivalence test and receive $K_\alpha^\beta = [-0.00261, 0.00383]$. The obtained values λ_i are shown in a histogram in the upper left graph of figure 3.9. With this result, we can reject the hypothesis H_0 of non-equivalence.

Thereafter, we exchanged the roles of β and γ for a consistency check and performed the test again. The results were indeed similar with the confidence interval $K_\alpha^\gamma = [0.00178, 0.00423]$. The histogram of the values λ_i as well as a quantile-quantile plot against the normal distribution can be admired in the right half of figure 3.9. It can be noted that while the λ_i in the β-model are biased towards negative values, in the γ-model the opposite is the case. This can be explained by the slight bias of the estimators which had to be used to compute the γ-model.

Case 2: τ_3

The results were very similar and comparable to those in case 1. Using the same $\varepsilon = \log(1 + 0.01) \approx 0.00995$ and $\alpha = 0.0001$, we obtain the confidence intervals $K_\alpha^\beta = [-0.00223, -0.0008824]$ and $K_\alpha^\gamma = [0.00625, 0.00910]$ for sampling in the β and γ-model, respectively. In both cases the hypothesis H_0 are rejected.

Case 3: Transposed model

The effect of the columnwise model definition still needs to be considered. A diagram is shown in figure 3.8. Here, the situation is slighly more complicated due to the bias in the estimators. With the same values for ε and α as used above, the test with sampling in the β-model does not reject, while with sampling in the γ-model it does. With a less conservative choise for $\varepsilon = \log(1 + 0.05) \approx 0.0488$ the hypotheses H_0 are rejected in both directions. The confidence intervals are $K_\alpha^\beta = [0.0392, 0.0422]$ and $K_\alpha^\gamma = [0.00644, 0.00788]$, respectively. As the confidence interval of the sampling β-model test indicates, the values of λ_i are indeed biased towards positive values.

It should be pointed out, that in the cases 1 and 2 we looked at the same urn, the difference in the models was only in the order of sampling. In case 3 however, the γ-model is the dual urn to the one in the β-model. Therefore, the bias in the estimators becomes more relevant. Emphatically, this demonstrates the limitations set by the estimators rather than the non-equivalence of the transposed model.

Conclusion

We have presented the noncentral hypergeometric distribution and its application to dinucleotide sequences. Moreover, we have demonstrated that the developed model is well-defined. In the next chapter, we evaluate the performance of the model applied to the real world data presented in the previous chapter.

4 Bridging Model and Data

In this chapter we combine the real world data and our model from the previous chapter. First, we approach only a few distantly related organisms but do this in detail. Thereafter, organisms are investigated that have previously been studied in literature. The classification capability of the model is evaluated. As a highlight, an overall pairwise distinction rate is computed before a scenario based on confidence regions providing an insight into genetic neigbourhoods is introduced.

4.1 Principal Component Analysis on Selected Organisms

In the following, we want to investigate the parameters of our model on real world data. Hereby, we want to keep two points in mind. Firstly, we need to demonstrate that the full number of parameters is indeed indispensible and cannot be reduced to a subset that contains the full information.

Besides that, we would like to take a first glance on the parameters for different organisms. The arising questions are if these really vary across the species and, in case they do, if this variation does manifest itself in the individual genes strongly enough to allow for classification.

To this end, we use a principal component analysis (71; 29) and concentrate on the 3-1 dinucleotide.

org	A-A	A-C	A-G	C-A	C-C	C-G	G-A	G-C	G-G
b. s.	0.319	0.172	0.251	0.312	0.202	0.252	0.312	0.317	0.195
m. l.	0.245	0.232	0.177	0.296	0.145	0.288	0.377	0.268	0.208
t. t.	0.573	0.099	0.234	0.225	0.303	0.111	0.304	0.230	0.356
p. t.	0.364	0.176	0.125	0.318	0.229	0.169	0.264	0.355	0.215

Table 4.1: The nine genome wide parameters of the 3-1 dinucleotides of the organisms bacillus subtilis *(as b. s.),* mesorhizibium loti *(as m. l.),* thermus thermophilus *(as t. t.) and* picrophilus torridus *(as p. t.).*

Selected Organisms

For this analysis, we have chosen three different bacteria from very non-similar environments that belong to three different phyla, namely *Bacillus subtilis (strain 168)* (63), *Mesorhizibium Loti (strain MAFF303099)* (57) and *Thermus Thermophilus (strain HB27)* (48), as well as the archaeon *Picrophilus Torridus (strain DSM 9790)* (37) for comparison.

We first have a look at the nine genome wide parameters of these organisms in table 4.1. It can clearly be seen, that the organisms differ in all nine parameters. Nonetheless, it is not easy to picture the situation in more detail due to the high dimensionality.

Number of Parameters and its Reduction

Our parameters are defined to model the pair building behaviour of dinucleotides. As there are $|\{A, C, G, T\}|^2 = 16$ dinucleotides one could argue that the same amount of parameters is necessary. However, in chapter 3.2.1 we have assigned our model conditional on the single nucleotide counts. On both positions of the dinucleotide the single nucleotides can be counted yielding to four numbers on the first and four on the second position. Yet, their sums are equal and therefore, we condition our model on seven quantities with nine free parameters remaining.

At this point, it is not a priori clear that all of these are in fact necessary. Therefore, we analyse the eigenvalues in the principal component analysis. These are shown in figure 4.1 on the next page for all four organisms. The decreasing eigenvalues do not offer space for a canonical cut-off. Furthermore,

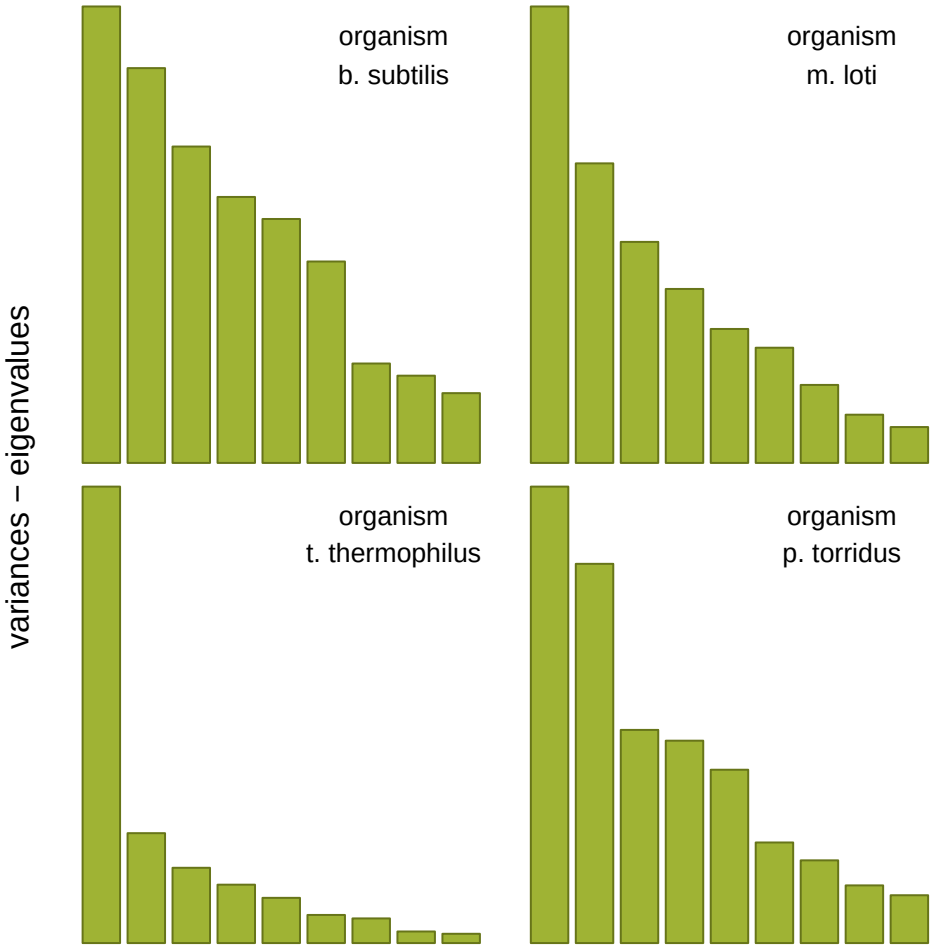

Figure 4.1: Eigenvalues obtained in the principal component analysis of the 3-1 dinucleotide for the organisms b. subtilis, m. loti, t. thermophilus and p. torridus. The progression in the eigenvalues does not yield to a canonical cut-off. The analysis of the corresponding eigenvectors will reveal a dissimilar underlying structure for the different organisms (see main text).

the principal components consist of different loadings of the parameters, i. e. correspond to different directions in the 9-dimensional space. This can be seen exemplarily for the case of the largest component in table 4.2. In fact, these directions may even be orthogonal. In figure 4.2 on the following page, we have a closer look at the relative directions of all eigenvectors of two pairs of organisms. It can be seen, that no lower dimensional subspace occurs. Even

	b. s.	m. l.	t. t.	p. t.
b. s.		0.96	0.33	0.74
m. l.	16		0.54	0.69
t. t.	71	57		0.21
p. t.	42	48	78	

Table 4.2: Relative directions of largest principal components for the different organisms. The upper diagonal part presents the scalar product of the eigenvectors, while the lower diagonal part gives the corresponding angles. (The abbreviations of the organisms are the same as in table 4.1.)

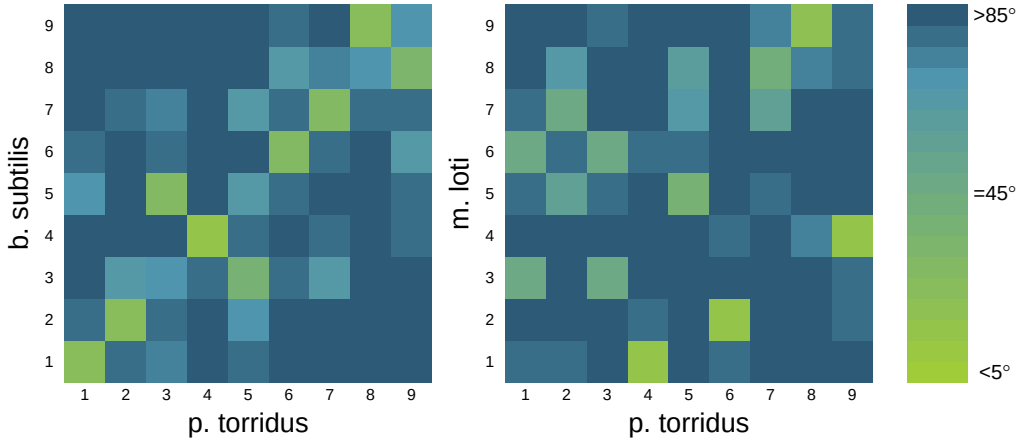

Figure 4.2: Relative directions of eigenvectors for two pairs of organisms. The angles between two vectors are shown in the colour coding given on the right hand side. While p. torridus provides the columns in both cases, it is paired with different organisms. The pairing with b. subtilis shows similarity of the directions in the components, although with different ordering. However, the pairing with m. loti demonstrates that this finding cannot be generalised.

more clearly, no such structure persists equivalently in both cases. Together, these results reveal that the nine parameters of our model all contribute significantly to the distribution.

We have so far focussed on the 3-1 dinucleotide. In figure 4.3 the eigenvalues obtained in the principal component analyses for all three dinucleotides as well as their combination are displayed exemplarily for t. thermophilus. The lower

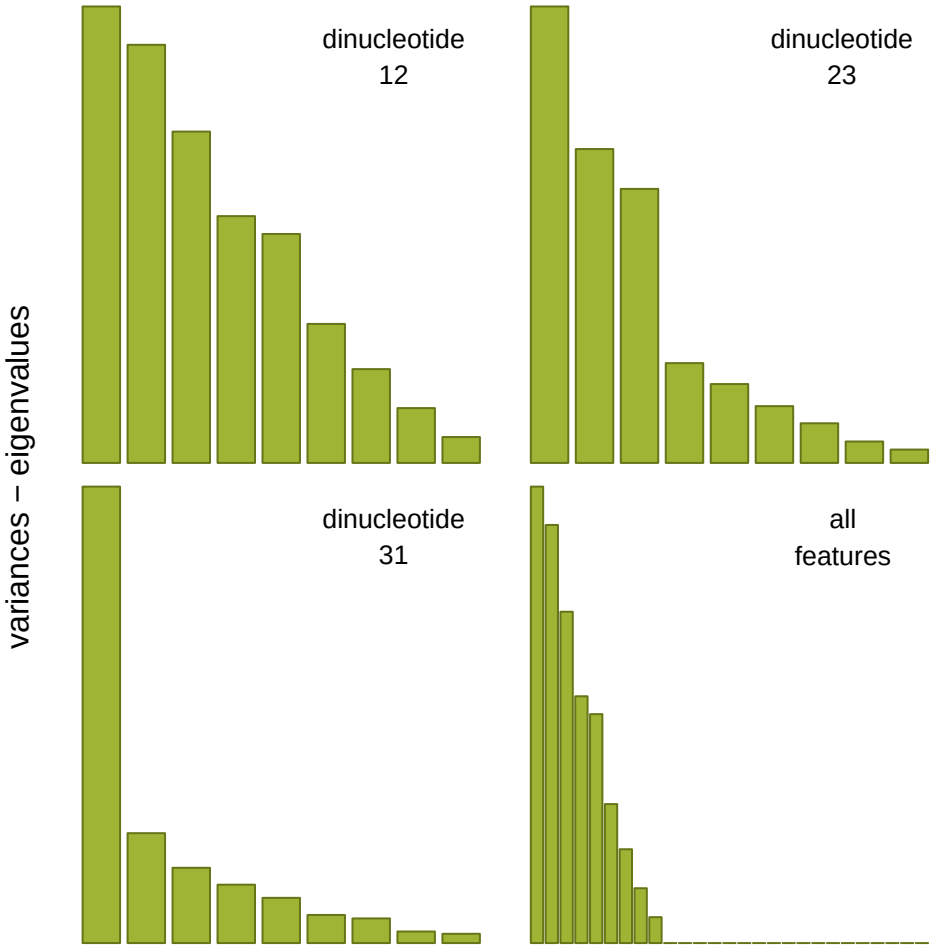

Figure 4.3: Eigenvalues obtained for the different dinucleotides as well as their combination when concentrating on the organism t. thermophilus. Their decrease does not show any characteristic pattern. The most remarkable observation is the vanishing of the eigenvalues beyond the first 9 *in the combined dinucleotides analysis. This is due to the circular structure of the dinucleotides in the genomic sequence (see main text).*

graph on the right hand side shows that all but 9 eigenvalues vanish. This can be explained by the circular structure of the 1-2, 2-3 and 3-1 dinucleotides, e. g. a certain given 1-2 dinucleotide already contains information about the following 2-3 dinucleotide. The high degree of dependency in our model makes it hard to reveal the effect in detail. While each of the dinucleotides is found to

contribute equally to the resulting eigenvectors, it is still desirable to consider them separately because of the different roles they play in biology.[1]

Comparison of Organisms in the Components

The results of the principal component analysis are also used for a first comparison of the genes of the organisms. The full 9-dimensional space can never be visualised. Therefore, we have to resort to some kind of projection. The principal component analysis provides a hierarchy of relevance of the different components which can serve as a guide in this process.

First, we choose an organism arbitrarily, whose components will provide a basis for our visualisation. Subsequently, we perform a transformation of bases of the genes of all organisms under consideration. We now project the 9-dimensional data onto the plane spanned by the first two components and obtain figure 4.4 on the facing page. Obviously, the reference organism shows the highest variance as its genes are projected on the plane defined by the two vectors of highest variance. However, this may or may not coincide with components of highest variance of the other organisms. In fact, table 4.1 on page 50 has shown, that this is not the case for the organisms in our example.

Clearly, the regions occupied by the different organisms overlap, at least in the projected view. The overlap is enhanced by the projection itself, since we have neglected a possibly large part of the information available. However, it is apparent from figure 4.4, that the notion of each organism occupying a certain region does make sense. Indeed, the centers of those regions are clearly apart from each other.

Conclusion

Using the principal component analysis, we have demonstrated that the number of parameters in our model cannot be reduced any further. The computed eigenvalues did not provide a justification for a cut-off. Additionally, the components were dependent on the organism and no common subspace could be identified. The above holds for all three dinucleotides.

[1] See also figure 4.6 on page 65.

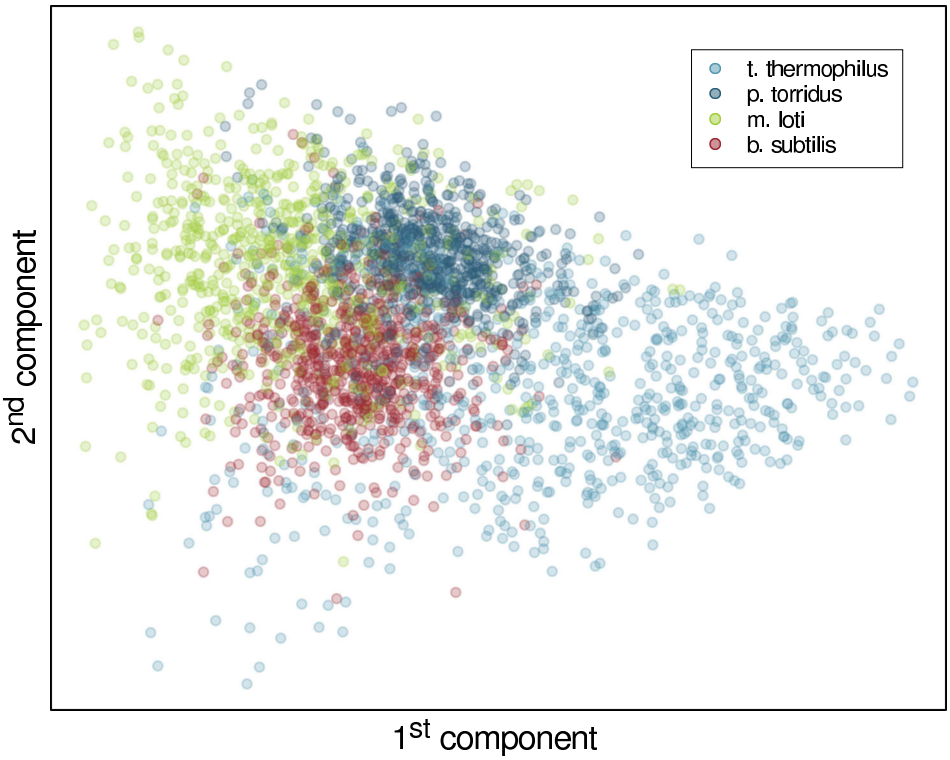

Figure 4.4: Genes from several organisms, shown in the components of the organism t. thermophilus and projected on the first two. The considered feature was the 3-1 dinucleotide. In each of the four organisms described in the legend, 750 genes were randomly selected for the figure to maintain clarity. Although they are projected on a 2-dimensional space, a separation is already visible.

Furthermore, we have considered a projection onto the leading two components, whereby we have shown that regions manifest themselves for different organisms. Although these regions show overlap and will therefore not yet allow for a single gene to be uniquely identified, they provide a characterisation of the organism. For example, it seems conceivable to identify a group of genes based on these regions. We expect the separation between organisms to be more pronounced in higher dimensions, which will be investigated in section 4.3.

Most importantly, the PCA has established that our model outlined in chapter 3.2 has proven its ability to distinguish different organisms.

4.2 Benchmarking against the Literature

Dinucleotide bias has been investigated previously using various signatures (88; 11). However, the signatures were computed on the whole sequences and no distinction of the positions of the dinucleotides was made. In coding regions, requirements caused by the coded amino acids conflict with this mix-up procedure. Nevertheless, dinucleotides on the different positions were compared against each other (45), and first interorganism comparisons based on this distinction were attempted (78; 50). The latter paper focusses on prokaryotic organisms and therefore lends itself for benchmarking of our model.

We shall first give a short outline of the procedure used in the work of Hooper and Berg (50). Based on the observation that in the absence of dinucleotide bias a hypergeometric distribution arises, they calculate the mean and variance within this model. Thereafter, they measure the discrepancy between the observed and expected dinucleotide counts obtaining a 16-dimensional distance vector. This vector can be computed for individual genes and compared to genome wide mean for different organisms using Hotelling's T^2 test (51). This way, the dinucleotide bias is exploited to identify the source of the genes, assumed to be the organism to which the T^2 distance is minimal.

Before actually comparing the results, we would like to express concern about some steps and arguments used in the procedure outlined above.

- The reduction of the 16-dimensional parameter space to 15 dimensions is not fully justified. Particularly, the arbitrary reduction by omitting a parameter without consideration of its variance or the information within bears the risk of producing uneven distance bias.
- Hotelling's T^2 test requires independence which clearly is not available in this case. The authors are aware of this problem. Yet, their way of dealing with it is purely empirical and might thus not work with different data.
- Moreover, there is some confusion about the meaning of confidence. The confidence level of a test and the proportion of genes for which the test rejects are used quasi-synonymously. This has no consequences for the pairwise comparison.

- Finally, the model consists of 135 parameters, given in the mean vector of length 15 as well as the corresponding covariance matrix. Although, the number of independent parameters remains unclear, the application to genes whose length is not considerably longer than that value leads to overfitting.

For a performance comparison we choose the same organisms as the authors. These include besides the prokaryotes *Escherichia coli (strain K-12)* (7), *Bacillus subtilis (strain 168)* (63), *Chlamydia trachomatis (strain D/UW-3/CX)* (103), *Chlamydophila pneumoniae (strain CWL 029)* (56) and *Rickettsia prowazekii (strain Madrid E)* (2) also the archaea *Aeropyrum pernix (strain K1)* (62) and *Methanobacterium thermoautotrophicum (strain Delta H)* (101), as well as the chromosomes IV and VI from the yeast genome *Saccharomyces cerevisiae (strain S288C)* (52; 77).

As the authors reduced the considered genes to those of a minimal length larger than 400 nucleotides, i. e. 133 dinucleotides, we proceed analogously. However, it can not be guaranteed that the very same set of genes was used for the comparison, as the state of the published genome might have changed considerably since the original work by Hooper and Berg. Moreover, as the authors did not cite the full source of the genomes, we could only guess the exact organisms on the sub-species level according to the date of publication.

Starting with one genome, we compute the likelihood of all n genes under consideration in the parameter model $\boldsymbol{\beta}$ of this source genome as well as the models $\boldsymbol{\gamma}_k$ of the other genomes for $k = 1, \ldots, 7$. Then, a likelihood ratio for each of the 7 pairs of organisms is calculated as

$$L_k(g_i) := \frac{\mathcal{L}_{\boldsymbol{\beta}}(g_i)}{\mathcal{L}_{\boldsymbol{\gamma}_k}(g_i)}$$

for the genes g_i for $i = 1, \ldots, n$. Finally, the normalised count

$$\Lambda_k := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{L_k(g_i) \geq 1\}}(g_i)$$

is considered. This proportion is computed for each organism serving as source organism against the models of the others. The results are presented in table 4.3 on the following page.

source organism against	e. coli	b. subtilis	c. trachomatis	c. pneumoniae	r. prowazekii	a. pernix	m. thermoauto.	s. cerevisiae
e. coli		86.2	96.5	94.9	83.5	95.7	98.6	85.0
b. subtilis	87.1		95.3	92.8	92.0	97.4	98.1	89.5
c. trachomatis	91.6	88.5		65.4	91.5	91.9	97.8	91.8
c. pneumoniae	94.8	93.5	56.4		93.7	90.3	98.0	94.4
r. prowazekii	91.9	94.5	96.2	95.8		96.0	99.3	95.6
a. pernix	93.5	94.0	86.9	89.4	86.5		97.8	94.3
m. thermoauto.	98.4	96.8	99.5	99.0	97.0	98.8		96.6
s. cerevisiae	82.4	82.6	96.1	93.0	89.1	93.7	94.9	
e. coli		88.2	96.5	94.9	83.5	95.8	98.8	85.1
b. subtilis	88.9		95.3	92.8	92.0	97.3	98.3	89.6
c. trachomatis	92.5	90.6		65.4	91.5	91.8	98.0	91.9
c. pneumoniae	95.7	95.2	56.4		93.8	90.3	98.2	94.4
r. prowazekii	92.5	95.8	96.2	95.8		96.1	99.4	95.7
a. pernix	94.7	95.9	86.9	89.4	86.5		98.0	94.3
m. thermoauto.	99.4	98.1	99.5	99.0	97.0	98.8		96.7
s. cerevisiae	86.9	85.6	96.1	93.0	89.1	93.6	95.2	

Table 4.3: Pairwise separation of genes, the structure of the table is analogous to Table 1 in Hooper and Berg (50) to allow direct comparison. $(\Lambda_1, \ldots, \Lambda_7)$ *is shown as a column vector for each source organism. All values are percentages. In the upper half all genes were taken into account while in the lower half the genes were restricted to those not classified as alien by Colombo. The name of the organism* methanobacterium thermoautotrophicum *was abbreviated for clarity. Particularly,* e. coli *and* b. subtilis *show systematically larger values.*

As can be seen, our model is able to reproduce approximately the same identification rates for most pairs of organisms while variations to either side occur. It is worth mentioning, that in the present case the parameters for our model were always obtained from all genes of the genome, while Hooper and Berg calculate the mean bias vectors as well as the covariance matrix, which together constitute their model, only from the gene-set of restricted length. Following this approach, we obtain an increase in the identification rate percentages of 0.5 to 5 for some pairs while others remain unaffected.

It should be pointed out, that in the presence of horizontal gene transfer, a level of 100% is not necessarily a desirable result as it can only be achieved if alien genes are not detected. Clearly, in the present case, the optimum value to be reached is unknown.

However, codon usage is independent of the 3-1 dinucleotide as it models the complementary transitions only. Therefore, we can reduce the set of considered genes by those, which have been found to be alien by a codon usage model. To this means, we use the tool Colombo (111).

Alien genes found by Colombo which do not have their source in the comparison organism affect the number of considered genes. This yields to a slight decrease in the values of Λ_k of order 10^{-3}. Genes identified by Colombo as alien which can be regarded as a HGT event between the two genomes under consideration result in removal of the gene from the numerator in the calculation of $1 - \Lambda_k$. Therefore, the fraction given in the table 4.3 increases, possibly by a large amount, as

$$\sum_{i=1}^{n} \mathbb{1}_{\{L_k(g_i)<1\}}(g_i)$$

is small. If the ratio on the reduced gene set still does not reach 100%, this can be explained as an additional detection event beyond the results given by Colombo, which of course can be either a false positive or a successfully identified HGT event.

It can be seen in table 4.3 that in *e. coli* and *b. subtilis* the values of Λ_k systematically increase after the Colombo preprocessing.

A general characteristic of statistical identification methods is their weekness in distinguishing very closely related species. This effect is visible in the organisms *c. trachomatis* and *c. pneumoniae* as these both belong to the taxonomic family *Chlamydiaceae*. In this case, the increase in the identification rate was phenomenal when computing the model parameters only from the longer genes of minimal length 134. The value obtained for comparison of the genes of *c. trachomatis* against the *c. pneumoniae* model changes from 56.4% to 73.0%.

Hooper and Berg set 134 dinucleotides as the minimal gene length in their work. We have so far adapted this number. Still, for application purposes it is very interesting to investigate the sensitivity of the method to this gene length cut-off. We therefore consider the generic case of minimum gene lengths between 0 and 500 dinucleotides. Examplarily, we present the results for *e. coli* in figure 4.5. The identification rate by our model typically displays an increase between approximately 100 and 200. Above this length, most curves show saturation. Four out of seven curves exceed the values given by Hooper and Berg for cut-off lengths below or in the vicinity of their magic cut-off of 134. However, our model gives a lower identification rate in the case of *s. cerevisiae* independent of the cut-off length. As we have observed a large effect of alien genes for this pair, when comparing to the Colombo-filtered genomes, and we expect that our method identifies further alien genes beyond Colombo, it seems plausible that the low identification rates computed with our model reflect the biological situation.

Conclusion

Our benchmark has shown that the performance of our model is comparable to the results of Hooper and Berg, when applied to their test organisms. Furthermore, by applying it to Colombo preselected genes, we have identified additional suspects for HGT events. A minor increase in the gene length cut-off to approximately 200 could result in identification rates which are typically superior to those given by Hooper and Berg. Finally, it should be kept in mind that all these comparisons were made directly, without penalising the 135 parameters of the Hooper and Berg approach against our 9 parameter model.

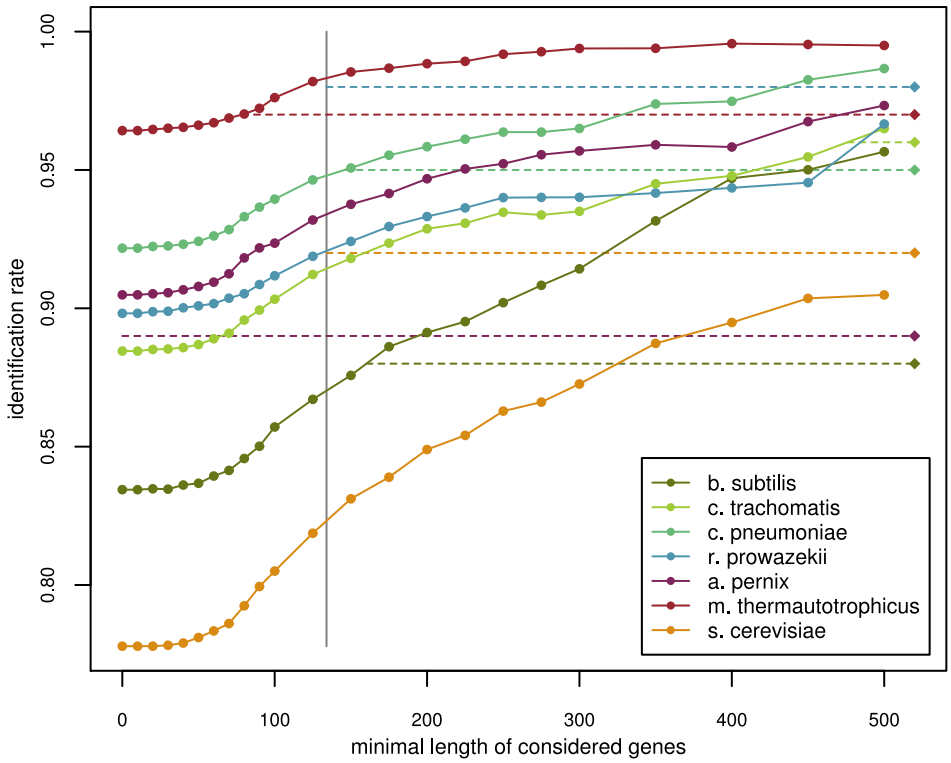

Figure 4.5: Identification rate of the genes of e. coli. *The scenario is similar to the one in table 4.3, except for the minimal length of the genes taken into consideration. Each curve corresponds to the pairwise comparison with a specific organism as given in the legend. On the right hand side, we show the values obtained by Hooper and Berg in (50), which were calculated for the minimal gene length* 134 *marked by the vertical grey line. Additionally, we indicate the minimum gene length at which our model exceeds the Hooper and Berg values by the dashed lines, where appropriate.*

4.3 Global Pairwise Distinction

In the following, we want to scrutinise the separation qualities of our model between organisms of different taxonomic branches. Hereby, we concentrate our analysis on the phylum level. First, we evaluate the capability to identify the origin of a gene when two phyla are compared. Next, we use a related scenario, in which we determine a gene's origin between all available phyla.

Implications of Hoeffding's Inequality

Before we proceed with the actual test, we need to assess the classification risk R, which is the expectation value of the comparison outcome function X, a random variable to be defined later. Upon frequent independent repetition of the test, we obtain a random process $\{X_i : i \in \mathbb{N}\}$ of the individual outcomes.

For any $n \in \mathbb{N}$, the risk R can be bounded in probability using the estimated mean of X via

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - R > \varepsilon\right) < \delta$$

for the error ε with uncertainty δ.

A useful functional form of δ can be obtained using Hoeffding's Inequality (49, Theorem 1), yielding

$$\delta = e^{-2\,n\,\varepsilon^2}$$

given that $X \in [0, 1]$.

Assume that we have set ε. In this case, Hoeffding's inequality relates the number of repetitions of the test and the uncertainty. Due to the monotonicity of the logarithm, we can derive the necessary number of repetitions, if a given uncertainty level should be guaranteed, namely

$$n = \left\lceil -\frac{\ln \delta}{2\,\varepsilon^2} \right\rceil .$$

For the purpose of our tests, we have set $\varepsilon = 0.0025$ and $\delta = 5 \cdot 10^{-6}$ and obtain $n = 976486$. Therefore, we work with $n = 10^6$.

Binary Decision Scenario

We shall first focus on a scenario where we are given a gene and two organisms from different phyla and have to decide to which of the two organisms the gene belongs.

More specifically, a test phylum and a comparison phylum are randomly chosen from the set of available phyla. Subsequently, a representing organism is

picked for each of the two with uniform probability. From the two organisms, we obtain the models parametrised by β and γ for the test and comparison organism, respectively. The test gene g is sampled from the set of the allowed genes of the test organism, again uniformly. In order to assign the gene to one of the two organisms, its likelihood ratio

$$L(\beta, \gamma; g) := \frac{\mathcal{L}_\gamma(g)}{\mathcal{L}_\beta(g)} \tag{4.1}$$

is computed. It should be pointed out that this definition differs from the one used in part 4.2. Clearly, an assignment of the gene to its own organism is equivalent to a low value of $L(\beta, \gamma; g)$. The outcome of the test is then given by the outcome function

$$\lambda(L(\beta, \gamma; g)) := \mathbb{1}_{\{L(\beta,\gamma;g)<1\}},$$

taking values 0, for successful identification, and 1, in case of failure.

This gives rise to a random variable

$$X := \lambda \circ L$$

with values in $[0, 1]$ on the probability space

$$\Omega := \mathcal{T}^2 \otimes \left(\bigcup_{i=1}^{\#\mathcal{T}} \{\mathcal{G} \,|\, \mathcal{G} \in t_i\} \right)^2 \otimes \left(\bigcup_{i=1}^{\#\tilde{\mathcal{G}}} \{g \,|\, g \in \mathcal{G}_i\} \right),$$

where $\mathcal{T} = \{t_1, \ldots, t_k\}$ is the set of the taxonomic units on a given level, e. g. phyla, and the set of all genomes in the different taxonomic units is $\tilde{\mathcal{G}} = \{\mathcal{G} \,|\, \mathcal{G} \in t_i \text{ for } t_i \in \mathcal{T}\}$.

For an element $\omega = (t', t'', \mathcal{G}', \mathcal{G}'', g) \in \Omega$ we define the probability

$$\mathbb{P}(\omega) := u_{\mathcal{T}}(t') \cdot u_{\{\mathcal{T}\setminus t'\}}(t'') \cdot u_{t'}(\mathcal{G}') \cdot u_{t''}(\mathcal{G}'') \cdot u_{\mathcal{G}'}(g)$$

for $t' \neq t'', \mathcal{G}' \in t', \mathcal{G}'' \in t'', g \in \mathcal{G}'$, and

$$\mathbb{P}(\omega) := 0$$

otherwise, where u_A denotes the uniform distribution on the finite set A.

feature	$\overline{X}$
1-2 dinucleotide	11.5 %
2-3 dinucleotide	5.6 %
3-1 dinucleotide	7.1 %
1-2 and 2-3	3.5 %
1-2 and 3-1	4.0 %
2-3 and 3-1	3.0 %
1-2, 2-3 and 3-1	2.2 %
codon usage	1.5 %
codon usage and 3-1	1.2 %

Table 4.4: Classification risk of the binary decision for different likelihood functions. In the case of two-fold dinucleotide combinations, the likelihood is given by the product of the individual likelihoods, examplarily $\mathcal{L}_{\beta(1-2,2-3)} = \mathcal{L}_{\beta(1-2)} \cdot \mathcal{L}_{\beta(2-3)}$. The three-fold likelihood is defined analogously. The codon usage is computed as in (111, equation 3).

In the present case, $\mathcal{T}$ consists of 11 phyla containing different numbers of organisms. We only consider genes from the reduced gene set restricted to those not classified as alien by Colombo. Due to this restriction the influence of known alien genes on the classification risk is avoided.

Led by the considerations at the beginning of this section, we sample $n = 10^6$ elements $\omega_i \in \Omega$. From these, we obtain realisations of X after fixing a likelihood function. In addition to the different dinucleotide features, we can also use combinations of those as well as take codon usage (111, equation 3) into account. The resulting $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ are compared in table 4.4. As can be seen, the classification risk can be quite different for the different dinucleotide features. This can be understood, if the relation of the dinucleotide features to codon usage is taken into account, see also figure 4.6 on the facing page. Specifically, the 2-3 dinucleotide partially profitates from the codon usage bias associated with the use of the Colombo preselection of genes. The combination of two dinucleotides yields to an improvement. Making use of all three dinucleotides, the classification is comparable to the one based on codon usage. However, the residual difference is influenced by the starting conditions which are strongly favourable for codon usage. Therefore, a decisive statement cannot be made regarding the superiority in the classification risk

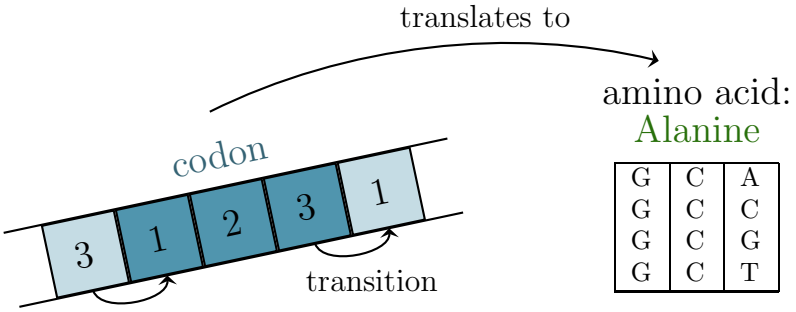

Figure 4.6: Relation between codon and dinucleotides on the different positions. While the transient 3-1 dinucleotide complements the codon, the 1-2 and 2-3 dinucleotides lie within. As can examplarily be seen in the table for Alanine*, the 1-2 dinucleotide remains mostly constant for the several codons translating into the same amino acid. The 2-3 dinucleotide on the other hand varies with the occurances depending on the codon usage. Therefore, they can be considered as* amino acid usage light *and* codon usage light*, respectively.*

of the dinucleotide versus the codon usage approach. In any case, a strong point of dinucleotide bias is the ability to trade between accuracy and the number of parameters in a controllable way.

Full Decision Scenario

Encouraged by the results obtained in the previous part, we will now consider a slightly more general case. In most conceivable applications, a test will not involve a single binary comparison but rather multiple comparisons with organisms from different phyla. In order to quantify the classification risk in such a situation, we shall make use of a modified version of the test scenario outlined earlier.

The adapted procedure, which we apply in this case, consists of the following steps:

- From each phylum an organism is sampled according to the uniform distribution.

- One of these organisms is picked as the test organism, again uniformly. The others will serve as comparison organisms.
- A test gene is chosen in the genome of the test organism.
- Pairwise binary comparison is carried out for the test gene between the models of the test organism and each of the comparison organisms.
- The test is successful if each comparison is in favour of the test organism.

The test result is then again expressed by an outcome function

$$\lambda(L(\beta,\gamma_1;g),\ldots,L(\beta,\gamma_m;g)) := \mathbb{1}_{\{L(\beta,\gamma_1;g)<1,\ldots,L(\beta,\gamma_m;g)<1\}}$$

where L is defined as in equation 4.1 and m denotes the number of available phyla for comparison, i. e. the total number of phyla under consideration is $k = m + 1$. Analogously to the binary decision scenario and making use of $\boldsymbol{L} = (L_1,\ldots,L_m)$, we can now define the random variable

$$X := \lambda \circ \boldsymbol{L}$$

on the probability space

$$\Omega := \mathcal{T} \otimes \bigotimes_{i=1}^{\#\mathcal{T}} \{\mathcal{G} \,|\, \mathcal{G} \in t_i\} \otimes \left(\bigcup_{i=1}^{\#\tilde{\mathcal{G}}} \{g \,|\, g \in \mathcal{G}_i\} \right),$$

where $\mathcal{T} = \{t_1,\ldots,t_k\}$ is the set of the taxonomic units on a given level, here again chosen to be the phylum level, and the set of all genomes in the different taxonomic units is $\tilde{\mathcal{G}} = \{\mathcal{G} \,|\, \mathcal{G} \in t_i \text{ for } t_i \in \mathcal{T}\}$.

The probability for an element $\omega = (t,\mathcal{G}_1,\ldots,\mathcal{G}_k,g) \in \Omega$ is defined as

$$\mathbb{P}(\omega) := u_{\mathcal{T}}(t) \cdot \left(\prod_{i=1}^{k} u_{t_i}(\mathcal{G}_i) \right) \cdot u_{\mathcal{G}'}(g)$$

for $g \in \mathcal{G}'$ and $\mathcal{G}' \in t$, and

$$\mathbb{P}(\omega) := 0$$

otherwise, where u_A denotes again the uniform distribution on A.

feature	$\overline{X}$	upper bound
1-2 dinucleotide	54 %	71 %
2-3 dinucleotide	31 %	44 %
3-1 dinucleotide	36 %	52 %
1-2 and 2-3	22 %	30 %
1-2 and 3-1	25 %	34 %
2-3 and 3-1	19 %	26 %
1-2, 2-3 and 3-1	15 %	20 %
codon usage	11 %	14 %
codon usage and 3-1	9 %	11 %

Table 4.5: Classification risk for the full decision scenario. For comparison, the bound given by the assumption of independent and identically distributed individual binary comparisons based on formula 4.2 and the values from table 4.4 on page 64 are shown in the right most column. The definition of the likelihood functions for combination of features is analogous to the binary decision scenario.

Hoeffding's inequality can be used to determine the necessary number of repetitions. However, it should be noted that in the present scenario the computational effort for a given number of repetitions is higher by roughly a factor of 5 as the most costly operation is the calculation of the likelihood function. Therefore, it would be desirable to reduce the number of repetitions. On the other hand, the classification risk is supposed to be larger. For instance, if we assume the individual likelihood ratios to be independent and identically distributed, the classification risk would be given by

$$R = 1 - \prod_{i=1}^{m} (1 - R') \tag{4.2}$$

where R' is the classification risk in the binary decision case obtained before. Exemplarily, for $R' = 2.2\%$ we find $R = 20\%$, while $R' = 7.1\%$ leads to $R = 52\%$. This suggests that an accuracy on the sub-percent level will not be needed in this scenario allowing us to lower the number of repetitions to $n = 250000$ for $\varepsilon = 0.005$ and $\delta = 5 \cdot 10^{-6}$. The results for this case are given in table 4.5.

As a central result, we conclude that our model is capable of distinguishing the genetic signature of organisms from different phyla. In this context, one

should first note that if a feature was incapable of such a distinction,

$$1 - \prod_{i=1}^{m}(1 - 0.5) = 1 - 0.5^{10} = 0.999\,.$$

On the other hand, assume that a single pair of phyla was indistinguishable. In that case, at least $0.5 \cdot \frac{2}{11} = 0.091$ of all tests would fail, even if all other phyla would be perfectly distinguished. Although our results exceed this value, the assumption of perfect distinction is violated for any pair of phyla. We therefore conclude that, using the combination of the different dinucleotide features, we can tell apart all phyla.

Furthermore, it should be noted that in both scenarios the combination of codon usage and the 3-1 dinucleotide displays an improvement over codon usage alone. This shows that our dinucleotide model is capable of additional detection of genes and works in a complementary way, as anticipated. Therefore, we are indeed able to use the dinucleotide bias for detection of alien genes.

Conclusion

The results obtained in this section indicate the applicability of our model in alien gene detection. Several methodological variants are conceivable and each have their own advantages. The decision for a specific variant will therefore have to be based on the biological question to be investigated. The most promising options are the combination of all three dinucleotides or the combination of codon usage together with the 3-1 dinucleotide as its complement. Furthermore, it might be interesting to work with a single dinucleotide for drastic reduction in parameter number. At this point, the 3-1 or 2-3 dinucleotides recommend themselves due to their biological role.

4.4 Genetic Neighbourhoods

If we concentrate on a particular dinucleotide, the 9 parameters of the model for a specific organism can be thought of as a point in the hypercube $[0, 1]^9$ assigned to that organism. As we have a large number of genes for each

organism, we can also think of all the points corresponding to their individual models. Based on the properties of the genetic data investigated in chapter 2, we expect those points to lie within a region around the organism center. We are now going to give a possible definition of such regions and study the implications.

To this end, consider a genome $\mathcal{G}$ and its set of genes[2]. We then make use of the following bootstrap procedure.

From this gene set we draw with replacement a collection $(g_i)_{i=1,\dots,N}$. To each g_i, we apply the β-estimator and obtain a collection $(\beta^i)_{i=1,\dots,N}$. It should be noted, that the β-estimator as defined in 3.4 and used according to 3.6 yields to a 9-tupel.

We estimate the genome mean by the sample mean $\hat{\beta}$ of $(\beta^i)_{i=1,\dots,N}$ and the covariance matrix $\hat{\Sigma}$ analogously.

Confidence Ellipsoids

As the matrix $\hat{\Sigma}$ is symmetric, it can be used to define an ellipsoid E centered around $\hat{\beta}$, given by

$$E := \{p \,|\, (p-\hat{\beta})^T \,\hat{\Sigma}^{-1}\,(p-\hat{\beta}) \leq c^2\} \tag{4.3}$$

with $c^2 = (\chi^2)^{-1}(1-\alpha)$, where α is the confidence level. Note that these ellipsoids are not necessarily contained in $[0,1]^9$. We therefore consider the intersection region, where appropriate.

The ellipsoid E defined above has a volume given by

$$\mathcal{V}(E) = (c^2)^{9/2} \cdot D_9 \cdot \prod_{i=1}^{9} \frac{1}{\sqrt{\lambda_i}} = c^9 \cdot \frac{D_9}{\sqrt{\det \hat{\Sigma}^{-1}}} = c^9 \cdot D_9 \cdot \sqrt{\det \hat{\Sigma}} \tag{4.4}$$

where $\lambda_1, \dots, \lambda_9$ are the eigenvalues of $\hat{\Sigma}^{-1}$, related to the radii r_i via $\lambda_i = 1/r_i^2$, and the dimensional constant

$$D_9 = \frac{\sqrt{\pi^{\,9}}}{\Gamma(\frac{9}{2}+1)} = \frac{\pi^4 \cdot 2^{\,9} \cdot 4!}{9!}\,.$$

[2] In the following, we restrict ourselves to those genes not classified as alien by Colombo.

Furthermore, the ratio

$$\frac{\min_i r_i}{\max_i r_i} = \sqrt{\frac{\min_i \lambda_i}{\max_i \lambda_i}}$$

defines the ellipticity.

Using these definitions, we can now start to explore the geometric configuration of the ellipsoids for all organisms.

Heuristic Determination of Ellipsoids

As a starting point, we need to fix the bootstrap procedure. To this end, we repeat the bootstrapping k times at fixed parameters and compare the resulting confidence ellipsoids. We take into account both the center points and the volumes at a given confidence level. We then look for a size of the bootstrap sample for which the run-to-run variation becomes acceptable. This is illustrated for 5 different organisms in figure 4.7.

We can proceed with the evaluation of the confidence level α to set the sizes of the ellipsoids. Figure 4.8 displays the parameters obtained for all genes of the organism *Escherichia Coli (strain K12)*. As can be seen, those correspond nicely to the genome-wide parameters. Although the center is clearly defined, considerable variation over the genes is visible. In figure 4.9, we show the fraction of genes of an organism which are actually contained within the corresponding ellipsoid. In order to avoid large overlap between the ellipsoids, we choose a rather large $\alpha = 0.67$ and obtain 21% to 25% of genes within the ellipsoid. With these parameters, we can now compute the volumes according to equation 4.4 for all $\nu = 284$ organisms under consideration[3]. We obtain the vast majority of volumes bounded by $1.0 \cdot 10^{-7}$ and set a cut-off at $2.0 \cdot 10^{-7}$. The histogram of the volumes is shown in figure 4.10. We observe that the total volume occupied by the ellipsoids of all organisms can be bounded according to

$$\mathcal{V}\left(\bigcup_{i=1}^{\nu} E_i\right) \leq \sum_{i=1}^{\nu} \mathcal{V}(E_i),$$

yielding a total volume below $1.3 \cdot 10^{-5}$. Therefore, essentially all of the hypercube can be considered empty. Using this knowledge, we are now ready

[3]We have excluded multiple strains of the same organism from the dataset to prevent misleading collisions of their ellipsoids.

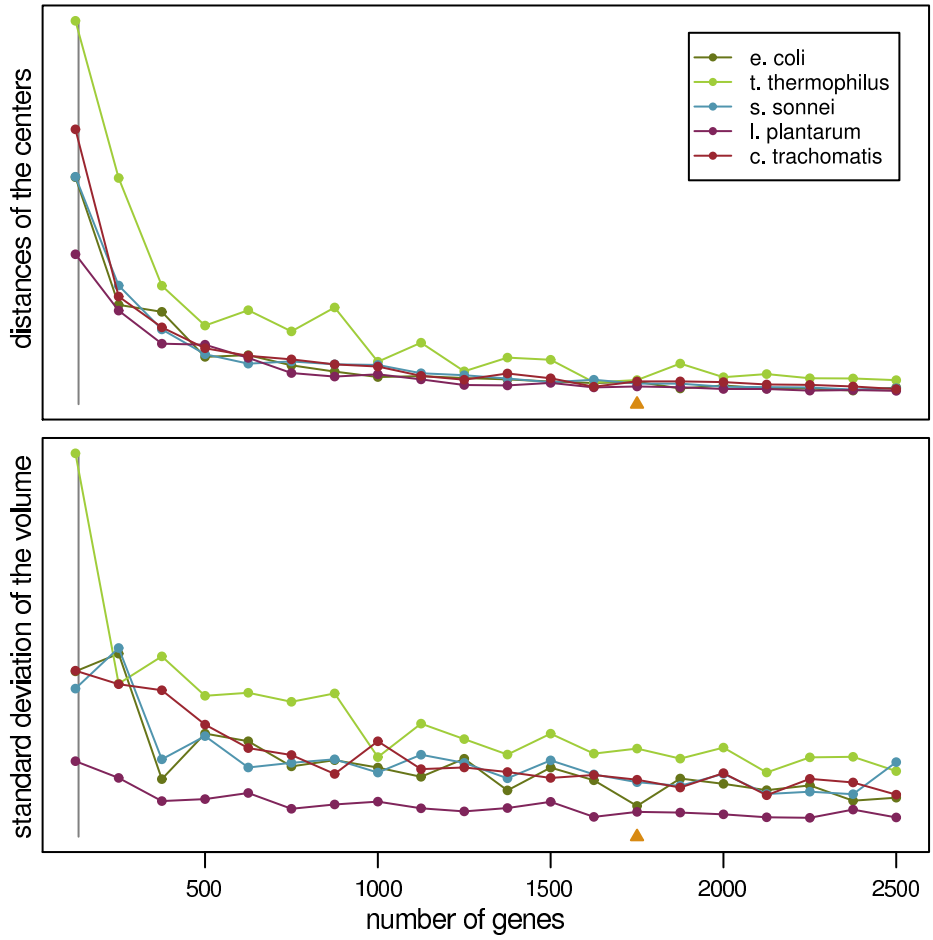

Figure 4.7: Determination of bootstrap sample size. For each sample size, $k = 20$ repetitions of the draw procedure were evaluated. The upper graph shows the average distance of the centers of two such instances. The lower graph presents the standard deviations of their volumes. In both graphs, the chosen sample size of 1750 *genes is highlighted. The organisms used are* Escherichia coli (strain K12)*,* Thermus thermophilus (strain HB27)*,* Shigella sonnei (strain Ss046)*,* Lactobacillus plantarum (strain WCFS1) *and* Chlamydia trachomatis (strain D/UW-3/CX)*.*

to interpret the results shown on the right hand side of figure 4.9. Clearly, the fraction of genes within the ellipsoid of their own organism suggests that most genes from a genome are localised within a region. Examplarily, this claim

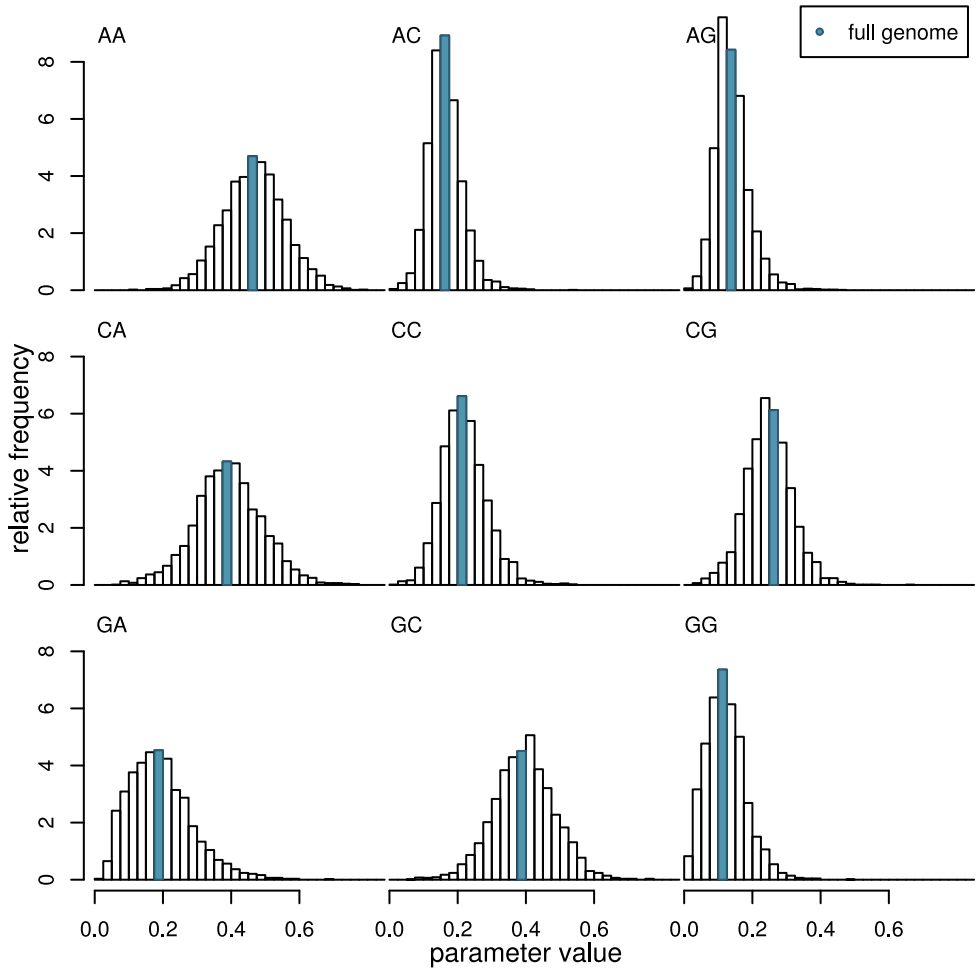

Figure 4.8: The genewise parameter estimates of Escherichia Coli (strain K12) *together with the estimates from the genome. The centers are clearly visible and correspond nicely to the genomewide parameters. The centers extend from* 0.1 *to* 0.4.

is supported by figure 4.4 on page 55. Imagine that we add a few uniformly distributed ellipsoids to the scenario. In that case, we do not expect to increase the total number of genes contained within all of them considerably. On the other hand we have observed that this number increases by more than a factor of 2 for approximately 50 ellipsoids. This can only be understood if many of those are located close to one another.

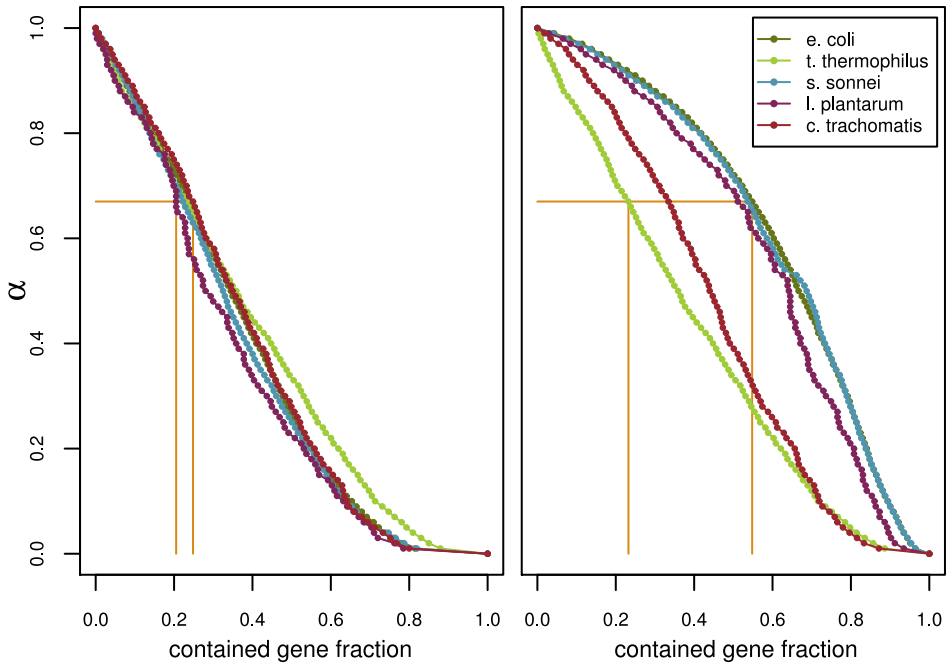

Figure 4.9: Fraction of genes within the ellipsoid for different confidence levels α. On the left hand side, only the organisms own ellipsoid is taken into account, while on the right hand side, all ellipsoids corresponding to organisms of the same phylum are considered. The organisms are the same as in figure 4.7. The orange lines at $\alpha = 0.67$ indicate our choice of confidence level, with the vertical lines showing the minimal and maximal fraction of genes within this confidence level, respectively. While the fraction of genes contained in their organism's own ellipsoid lie between 21% and 25%, the fraction of genes within any ellipsoid of their own phylum is highly dependent on the size of the respective phylum and can reach up to 55%.

This consideration motivates a change of focus from an organism's own ellipsoid to the union of ellipsoids of its own phylum. This will be of particular importance in the next section 4.5. Moreover, it would be interesting to check to which degree the proximities in the β-hypercube can be traced back to biology. Or, putting it the other way around, if exploration of the β-hypercube can help to answer biological questions.

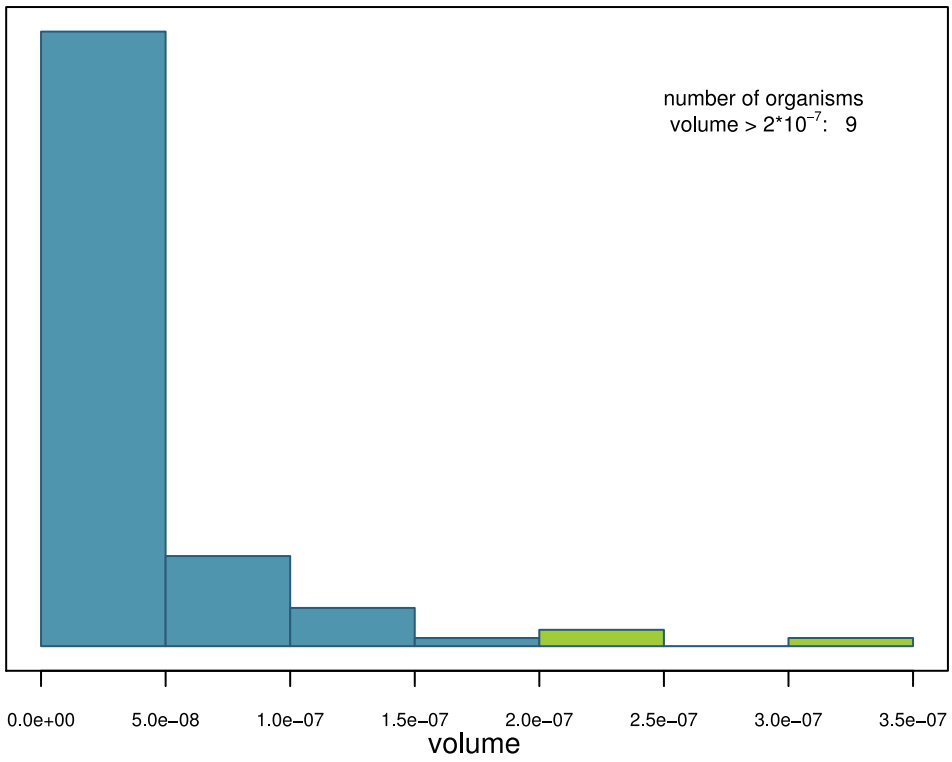

Figure 4.10: Histogram of the volumes of the confidence ellipsoids for $\alpha = 0.67$*. The green area contains the organisms with large volumes above* $2.0 \cdot 10^{-7}$*.*

In addition to the volumes, we also compute the ellipticities of the ellipsoids. These can be seen in figure 4.11, again as a histogram. Most ellipsoids are only moderately elliptic, with an ellipticity ratio around 0.25. We consider an ellipsoid as extremely elliptic if its ellipticity is below 0.15, and set a cut-off at this value. It turns out that the organisms which are beyond the cut-off for either the volume or the ellipticity mostly coincide.

Computing the Overlap

Finally, it is interesting to ask to which extent the ellipsoids corresponding to the different organisms overlap. While it might be easy to exclude overlap in some situations, the actual computation of non-vanishing overlap volumes is demanding. We will therefore first reduce the number of suspects before we

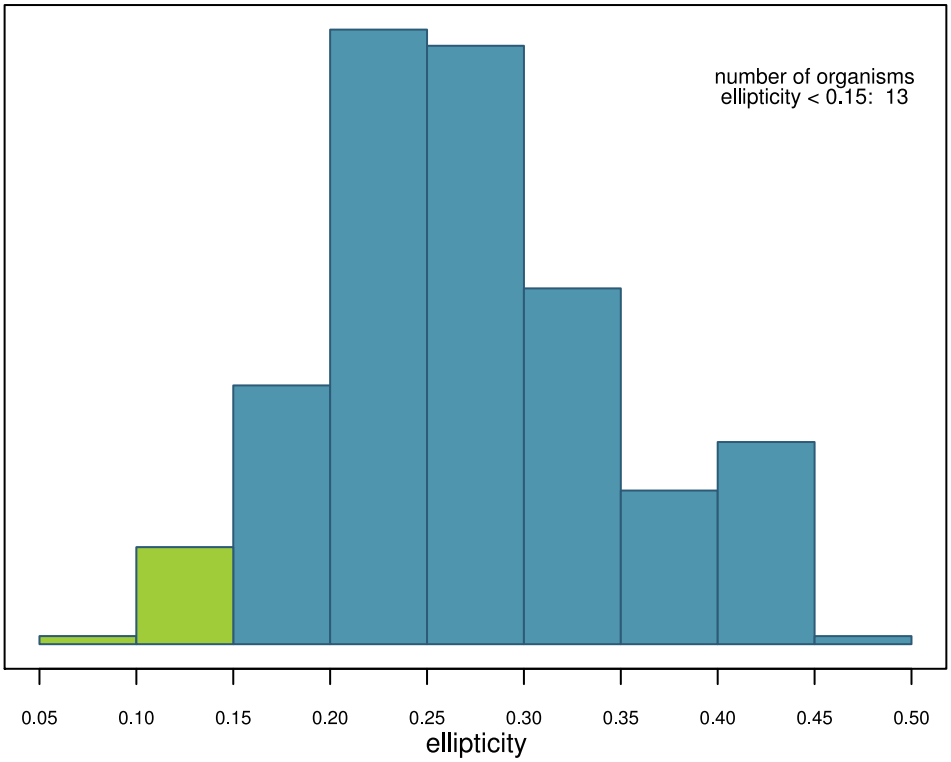

Figure 4.11: Histogram of the ellipticities of the organisms under consideration. The green area contains the organisms with extreme ellipticity below 0.15.

actually perform this computation. Moreover, we leave out those organisms that lie beyond one of the cut-offs defined before. Given two ellipsoids E_1 and E_2, we consider spheres S_1 and S_2 in which they are fully contained, given by

$$S_{1,2} = \{\, p \,|\, (p - \hat{\beta}_{1,2})^2 \leq c^2 \cdot R_{1,2}^2 \,\},$$

where $R = 1/\sqrt{\min_i \lambda_i}$ is the largest radius of the respective ellipsoid and c^2 is defined as before. Obviously, from

$$c^2 \cdot \left(R_1^2 + R_2^2\right) \leq \left(\hat{\beta}_1 - \hat{\beta}_2\right)^2$$

we can conclude

$$S_1 \cap S_2 = \emptyset \qquad \Longrightarrow \qquad E_1 \cap E_2 = \emptyset.$$

The above condition, which is computationally easy to check, allows us to exclude pairwise overlap in all but 80 out of $\nu{\cdot}(\nu{-}1)\,/\,2 \approx 4{\cdot}10^4$ possible cases. For the remaining cases, we really have to compute the overlap. This task can be performed using Monte Carlo integration. To this end, we compute bounding boxes containing $B_1 \supset E_1$ and $B_2 \supset E_2$ of the form

$$\bigotimes_{i=1}^{9} [x_{i,\min}, x_{i,\max}]$$

with $0 \leq x_{i,\min} \leq x_{i,\max} \leq 1$ for $i = 1 \ldots 9$. Clearly, we then have

$$E_1 \cap E_2 \subset B_1 \cap B_2,$$

where the right hand side is straightforward to calculate. Subsequently, we uniformly sample $N = 5 \cdot 10^5$ points p_i inside this box intersection, which typically has a volume of $\mathcal{V}(B_1 \cap B_2) \approx 1 \cdot 10^{-5}$. For each of those points, we verify whether or not it is contained within the two ellipsoids $E_{1,2}$ using equation 4.3. It should be noted that the application of the equation 4.3 involves matrix multiplication and is therefore computationally expensive, thus limiting the number of points we were able to calculate. The volume of the intersection is then approximated by

$$\mathcal{V}(E_1 \cap E_2) \approx \frac{\sum_{i=1}^{N} \mathbb{1}_{E_1}(p_i) \cdot \mathbb{1}_{E_2}(p_i)}{N} \cdot \mathcal{V}(B_1 \cap B_2)$$

We did not find any common points for either pair of organism in the integration procedure. However, the box intersection volume is fairly large with respect to the expected intersection volumes, so that typically only around ten points come to lie in either ellipsoid alone. We can therefore bound the intersection volumes to a few percent of the ellipsoid volumes. We have examplarily carried out the integration procedure with $N = 2.5 \cdot 10^5$ points for two pairs of organisms, namely *Synechocystis (strain PCC 6803)* versus *Pseudoalteromonas haloplanktis (strain TAC125)* and *Mycoplasma pneumoniae (strain M129)* versus *Leptospira borgpetersenii serovar Hardjo-bovis (strain L550)*. Still, we have not found any common points, yet, we can now give an upper bound for the intersection volume in those particular cases of approximately 1% of the individual volumes.

Conclusion

Our investigation of the geometry arising from the definition of confidence ellipsoids has shown that organisms can be seen as separate regions with a

typical range of volumes and shapes. Our choice of parameters has enabled us to keep their overlap small. From these properties, one can expect that gene identification based on position in the β-hypercube should be feasible, although with a limited identification rate for each individual case due to the chosen confidence level. Moreover, our findings suggest geometrical neighbourhoods of organisms from the same phylum. This property can then be exploited to map biological questions to geometrical problems in the in the hypercube.

4.5 Metagenomics

Metagenomics is a biological setup that is used if the organisms in question cannot be sequenced and if some information can be given up. It is not uncommon for microbes that they cannot be multiplied under lab conditions or that they interfere with the sequencing techniques (41; 104). Instead of sequencing the organisms individually, everything from a certain environment is taken and short sequences are obtained (109). The aim is to give a profile of these sequences based on phyla in percentage (16; 21). It is thus not necessary to identify each sequence individually. This compensates for the fact that little information is given, in particular, no predefined codon structure is available.

These specifications resemble the properties which we have found for the identification based on the confidence ellipsoids in the previous section. Therefore, it is peculiarly interesting to explore the potential of this geometrical procedure for the application of metagenomics.

Sampling Procedure

Our starting point is a set $\mathcal{S}$ of genomic sequences; in the first step, they will be genes. These are drawn from the reduced set of genomes used for the confidence ellipsoids according to a prescribed distribution. For simplicity, we first consider the case where the distribution corresponds to the numbers of organisms contributing the phylum.

For each sequence $s \in \mathcal{S}$, we try to identify the source phylum, using the following prescription:

- If $s \in E$ for a single confidence ellipsoid E, it is counted for the phylum to which the ellipsoid belongs.
- If $s \notin \bigcup_i E_i$, i. e. s does not lie in any of the given ellipsoids, then s is considered as *unknown*.
- If $s \in E_i \cap E_j$ for E_i, E_j in the same phylum, we can again count it for this phylum. The same applies for more than two ellipsoids, as long as all of them belong to the same phylum.
- Otherwise, s is disputed between several phyla. This situation can be dealt with in several ways, here we consider the possibilities of either classifying s as unidentifiable and not counting it for either phylum (exclusive approach), or to count it for all matching phyla, with a weight equal to the inverse of their number (sharing approach).

Finally, we obtain a profile consisting of the relative frequencies of counts for all phyla, as well as the fraction of unknown and unidentifiable sequences. The results of a test run using a confidence level $\alpha = 0.67$ with 20 genes sampled from each organism can be seen in table 4.6.

Slightly more than half of the genes are unknown, if we decide to share sequences falling within more than one phylum. Those are found to constitute about ohne quarter of the total sample. Both values are consistent with our findings from the previous section, although the overlap volume of the ellipsoids suggested slightly less disputed sequences. It is interesting to note that some phyla, like *Actinobacteria* or *Firmicutes* give good results if the profiles are renormalised according to the number of identified sequences, while others, like *Chlorobi* give satisfying results without normalisation. A possible explanation for the latter observation could be that this phylum lives in an otherwise sparse region of the parameter space. Exclusive profiling seems biased towards large phyla, which can be understood geometrically, as few double-classification events strongly reduce the counts of a small phylum, while having negligible influence on the large phyla.

We have also performed first tests with initial profiles strongly deviating from the relative phylum sizes, and have found evidence for a volume bias effect: Genes that are not classified in their own organism's ellipsoid will have an

phylum	input profile	shared profiling		exclusive profiling	
		raw	norm.	raw	norm.
Acidobacteria	0.7 %	0.7 %	1.4 %	0.0 %	0.0
Actinobacteria	10.0 %	5.5 %	11.7%	2.0 %	10.1
Aquificae	0.4 %	0.1 %	0.2 %	0.1 %	0.5
Bacteroidetes	2.2 %	1.9 %	4.0 %	0.2 %	1.0
Chlamydiae	2.6 %	1.1 %	2.3 %	0.3 %	1.6
Chlorobi	1.5 %	1.4 %	3.0 %	0.1 %	0.6
Chloroflexi	0.7 %	0.2 %	0.5 %	0.0 %	0.3
Cyanobacteria	3.3 %	2.3 %	4.9 %	0.4 %	1.8
Deinococcus-Th	0.7 %	1.3 %	2.7 %	0.1 %	0.4
Firmicutes	18.2 %	9.6 %	20.3%	3.3 %	17.3
Proteobacteria	56.5 %	21.2 %	44.8%	12.1 %	62.5
Spirochaetes	2.6 %	1.9 %	4.0 %	0.7 %	3.4
Thermotogae	3.7 %	0.1 %	0.2 %	0.1 %	0.4
unknown	–	52.7 %	–	80.7 %	–

Table 4.6: Results of profiling runs on sample data using either a sharing or an exclusive approach to the final profiling step. The raw columns show the result counting the unknown *sequences as their own class, while for the normalised columns, the phyla add up to* 100%.

increased probability to fall into the larger volume spanned by the ellipsoids of large phyla. The output profiles are therefore driven in the direction of the relative phylum sizes. This effect needs to be dealt with when formulating the biological question one wants to answer using this model. Finally, we composed sequences dropping the codon structure. Of course, the fraction of unknown sequences rises strongly in this situation. However, the resulting profiles for the rest of the sample still give reasonable results. The significance of these profiles remains unclear due to the smallness of the sample.

Conclusion

The results obtained on the sample data provide a framework for the analysis of real-world data. As the *true* profile for these is not known, and the sequence composition is supposedly biased against those organisms which can be sequenced and whose data and information is therefore available, metagenomic analysis has so far only been feasible on reduced datasets for narrow questions. This kind of narrowing-down appears also in our results, therefore, we see a potential for application of our model in this context. However, it is very important to get a deeper understanding of systematic uncertainties, which is beyond the scope of this work.

5 Conclusion

In this work, we have constructed a model for dinucleotide bias in genomic sequences, based on a biased urn described by the noncentral hypergeometric distribution by Wallenius. We have analysed the prerequisites in the data, and examined the applicability of the model in a broad range of scenarios.

Our model improves on existing work in terms of its well-defined stochastical basis, which also provides a lot of flexibility for application even in yet not foreseen scenarios. Especially, it can be applied to each dinucleotide feature and also allows for combinations of these, which enhance its predictive power. These combinations may be motivated by and adapted to biological questions.

There is a freedom of choice in the construction of the model. We have shown the mathematical equivalence of the resulting models. However, they need not be equivalent on the biological side. In particular, it is conceivable that the four-step models starting with the nucleotides G and C can be more accurate. This is due to the fact, that the estimators perform best for large urns, i. e. in the first two steps. Thus, they favour a precise value for the GC bias which is known to be biologically important.

The model is optimal in the number of parameters as demonstrated by a principal component analysis. We have evaluated the performance of our model in the identification of the origin of individual genes and have obtained results that are comparable to existing methods using literature data as a benchmark. We have also verified that the model predicts alien genes beyond those identified by a codon usage approach. Using a binary decision scenario, involving a large number of organisms and codon usage based preselection, we have verified that the different dinucleotides indeed correlate differently with the codon usage. Moreover, this scenario demonstrates that the model is sensitive also with respect to the phylum.

We have constructed confidence ellipsoids for all organisms using a bootstrap technique. These give rise to a geometrical view and a notion of neighbourhood and proximity. The concepts so obtained seem to hold also on the phylum level. Therefore, an application to metagenomics comes into reach. We have thus performed first tests using verifiable sample data finding reasonable agreement with the known input profile. Still, further investigations will be necessary to get an understanding of the potentials of the method.

Bibliography

[1] Anderson, T. W. & Goodman, L. A. Statistical inference about markov chains. *Ann. Math. Statist.* **28**, 89–110 (1957).

[2] Andersson, S. G. *et al.* The genome sequence of rickettsia prowazekii and the origin of mitochondria. *Nature* **396**, 133–140 (1998).

[3] Aravind, L., Tatusov, R. L., Wolf, Y. I., Walker, D. R. & Koonin, E. V. Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. *Trends Genet.* **14**, 442–444 (1998).

[4] Bapteste, E. *et al.* Do orthologous gene phylogenies really support tree-thinking? *BMC Evol. Biol.* **5**, 33–33 (2005).

[5] Bapteste, E. & Walsh, D. A. Does the 'ring of life' ring true? *Trends Microbiol.* **13**, 256–261 (2005).

[6] Bates, S., Cashmore, A. M. & Wilkins, B. M. IncP plasmids are unusually effective in mediating conjugation of escherichia coli and saccharomyces cerevisiae: Involvement of the tra2 mating system. *J. Bacteriol.* **180**, 6538–6543 (1998).

[7] Blattner, F. R. *et al.* The complete genome sequence of escherichia coli K-12. *Science* **277**, 1453–1474 (1997).

[8] Brunner, E. Angewandte Statistik. Teil 1. (Lecture Notes). (1998).

[9] Brüssow, H., Canchaya, C. & Hardt, W. D. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol. Mol. Biol. Rev.* **68**, 560–602 (2004).

[10] Burge, C., Campbell, A. M. & Karlin, S. Over- and under-representation of short oligonucleotides in DNA sequences. *Proc. Natl. Acad. Sci. U.S.A.* **89**, 1358–1362 (1992).

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

[12] Canchaya, C., Fournous, G., Chibani-Chennoufi, S., Dillmann, M. L. & Brüssow, H. Phage as agents of lateral gene transfer. *Curr. Opin. Microbiol.* **6**, 417–424 (2003).

[13] Chesson, J. A non-central multivariate hypergeometric distribution arising from biased sampling with application to selective predation. *J. Appl. Probab.* **13**, 795–797 (1976).

[14] Chiter, A., Forbes, J. M. & Blair, G. E. DNA stability in plant tissues: Implications for the possible transfer of genes from genetically modified food. *FEBS Lett.* **481**, 164–168 (2000).

[15] Connolly, J. H., Herriott, R. M. & Gupta, S. Deoxyribonuclease in human blood and platelets. *Br. J. Exp. Pathol.* **43**, 392–401 (1962).

[16] Daniel, R. The metagenomics of soil. *Nat. Rev. Microbiol.* **3**, 470–478 (2005).

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

[18] Daubin, V. & Perrière, G. G+C3 structuring along the genome: A common feature in prokaryotes. *Mol. Biol. Evol.* **20**, 471–483 (2003).

[19] DeFlaun, M. F. & Paul, J. H. Detection of exogenous gene sequences in dissolved DNA from aquatic environments. *J. Microb. Ecol.* **18**, 21–28 (1989).

[20] Deuflhard, P. & Hohmann, A. *Numerische Mathematik I. Eine algorithmisch orientierte Einfhrung.* (de Gruyter, Berlin, 1991).

[21] Dinsdale, E. A. *et al.* Functional metagenomic profiling of nine biomes. *Nature* **452**, 629–632 (2008).

[22] Doolittle, R. F. Searching for the common ancestor. *Res. Microbiol.* **151**, 85–89 (2000).

[23] Doolittle, R. F., Feng, D. F., Anderson, K. L. & Alberro, M. R. A naturally occurring horizontal gene transfer from a eukaryote to a prokaryote. *J. Mol. Evol.* **31**, 383–388 (1990).

[24] Doolittle, W. F. Phylogenetic classification and the universal tree. *Science* **284**, 2124–2129 (1999).

[25] Dufraigne, C., Fertil, B., Lespinats, S., Giron, A. & Deschavanne, P. Detection and characterization of horizontal transfers in prokaryotes using genomic signature. *Nucleic Acids Res.* **33** (2005).

[26] Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* (Chapman and Hall, New York, 1993), 3rd edn.

[27] Einspanier, R. *et al.* The fate of forage plant DNA in farm animals: A collaborative case-study investigating cattle and chicken fed recombinant plant material. *European Food Research and Technology* **212**, 129 – 134 (2001).

[28] EMBL. Nucleotide Sequence Database. URL `www.ebi.ac.uk/embl`.

[29] Falk, M., Becker, R. & Marohn, F. *Angewandte Statistik mit SAS* (Springer, Heidelberg, 1995).

[30] Feller. *An Introduction to Probability Theory and Its Applications*, vol. 1 (John Wiley & Sons, New York, 1968), 3rd edition edn.

[31] Flores, E. & Wolk, C. P. Identification of facultatively heterotrophic, N2-fixing cyanobacteria able to receive plasmid vectors from escherichia coli by conjugation. *J. Bacteriol.* **162**, 1339–1341 (1985).

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

[33] Fog, A. `BiasedUrn` – R package (2006).

[34] Fog, A. Calculation methods for Wallenius' noncentral hypergeometric distribution. *Comm. Statist. Simulation Comput.* **27**, 258:273 (2008).

[35] Franco, A. A. *et al.* Molecular evolution of the pathogenicity island of enterotoxigenic bacteroides fragilis strains. *J. Bacteriol.* **181**, 6623–6633 (1999).

[36] Friedlander, A. M. DNA release as a direct measure of microbial killing. I. Serum bactericidal activity. *J. Immunol.* **115**, 1404–1408 (1975).

[37] Fütterer, O. *et al.* Genome sequence of picrophilus torridus and its implications for life around pH 0. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 9091–9096 (2004).

[38] Gevers, D. *et al.* Opinion: Re-evaluating prokaryotic species. *Nat. Rev. Microbiol.* **3**, 733–739 (2005).

[39] Grantham, R., Gautier, C. & Gouy, M. Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. *Nucleic Acids Res.* **8**, 1893–1912 (1980).

[40] Grantham, R., Gautier, C., Gouy, M., Mercier, R. & Pavé, A. Codon catalog usage and the genome hypothesis. *Nucleic Acids Res.* **8**, 49–49 (1980).

[41] Green, B. D. & Keller, M. Capturing the uncultivated majority. *Curr. Opin. Biotechnol.* **17**, 236–240 (2006).

[42] Grimmett, G. & Stirzaker, D. *Probability and Random Processes* (Oxford University Press, Oxford, 2001), 3rd edn.

[43] Hacker, J., Blum-Oehler, G., Mühldorfer, I. & Tschäpe, H. Pathogenicity islands of virulent bacteria: Structure, function and impact on microbial evolution. *Mol. Microbiol.* **23**, 1089–1097 (1997).

[44] Hacker, J. & Kaper, J. B. Pathogenicity islands and the evolution of microbes. *Annu Rev Microbiol* **54**, 641–679 (2000).

[45] Hanai, R. & Wada, A. Doublet preference and gene evolution. *J. Mol. Evol.* **30**, 109–115 (1990).

[46] Hao, W. & Golding, G. B. The fate of laterally transferred genes: life in the fast lane to adaptation or death. *Genome Res* **16**, 636–643 (2006).

[47] Heinemann, J. A. & Sprague, G. F. Bacterial conjugative plasmids mobilize DNA transfer between bacteria and yeast. *Nature* **340**, 205–209 (1989).

[48] Henne, A. *et al.* The genome sequence of the extreme thermophile thermus thermophilus. *Nat. Biotechnol.* **22**, 547–553 (2004).

[49] Hoeffding, W. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58**, 13–30 (1963).

[50] Hooper, S. D. & Berg, O. G. Detection of genes with atypical nucleotide sequence in microbial genomes. *J. Mol. Evol.* **54**, 365–375 (2002).

[51] Hotelling, H. The generalization of student's ratio. *Ann. Math. Statist.* **2**, 360–378 (1931).

[52] Jacq, C. *et al.* The nucleotide sequence of saccharomyces cerevisiae chromosome IV. *Nature* **387**, 75–78 (1997).

[53] Jain, R., Rivera, M. C. & Lake, J. A. Horizontal gene transfer among genomes: The complexity hypothesis. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 3801–3806 (1999).

[54] Jain, R., Rivera, M. C., Moore, J. E. & Lake, J. A. Horizontal gene transfer in microbial genome evolution. *Theor. Popul. Biol.* **61**, 489–495 (2002).

[55] Jonas, D. A. *et al.* Safety considerations of DNA in food. *Ann. Nutr. Metab.* **45**, 235–254 (2001).

[56] Kalman, S. *et al.* Comparative genomes of chlamydia pneumoniae and c. trachomatis. *Nat. Genet.* **21**, 385–389 (1999).

[57] Kaneko, T. *et al.* Complete genome structure of the nitrogen-fixing symbiotic bacterium mesorhizobium loti. *DNA Res.* **7**, 331–338 (2000).

[58] Karaolis, D. K., Somara, S., Maneval, D. R., Johnson, J. A. & Kaper, J. B. A bacteriophage encoding a pathogenicity island, a type-IV pilus and a phage receptor in cholera bacteria. *Nature* **399**, 375–379 (1999).

[59] Karl, D. M. & Bailiff, M. D. The measurement and distribution of dissolved nucleic acids in aquatic environments. *Limnol. Oceanogr.* **34**, 543–558 (1989).

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

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

[62] Kawarabayasi, Y. *et al.* Complete genome sequence of an aerobic hyperthermophilic crenarchaeon, aeropyrum pernix K1. *DNA Res.* **6**, 83–101 (1999).

[63] Kunst, F. *et al.* The complete genome sequence of the gram-positive bacterium bacillus subtilis. *Nature* **390**, 249–256 (1997).

[64] Lake, J. A., Jain, R. & Rivera, M. C. Mix and match in the tree of life. *Science* **283**, 2027–2028 (1999).

[65] Lawrence, J. G. & Ochman, H. Molecular archaeology of the escherichia coli genome. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 9413–9417 (1998).

[66] Lewin, B. *Genes IX* (Jones & Bartlett Publishers, London, 2007), 9 edn.

[67] Lorenz, M. G., Gerjets, D. & Wackernagel, W. Release of transforming plasmid and chromosomal DNA from two cultured soil bacteria. *Arch. Microbiol.* **156**, 319–326 (1991).

[68] Lorenz, M. G. & Wackernagel, W. Bacterial gene transfer by natural genetic transformation in the environment. *Microbiol. Rev.* **58**, 563–602 (1994).

[69] Lyons, N. I. Closed expressions for noncentral hypergeometric probabilities. *Comm. Statist. Simulation Comput.* **9**, 313–314 (1980).

[70] Manly, B. F. J. A model for certain types of selection experiments. *Biometrics* **30**, 281–294 (1974).

[71] Mardia, K. V., Kent, J. T. & Bibby, J. M. *Multivariate Analysis* (Academic Press, London, 1979).

[72] Marri, P. R., Hao, W. & Golding, G. B. Gene gain and gene loss in streptococcus: Is it driven by habitat? *Mol. Biol. Evol.* **23**, 2379–2391 (2006).

[73] Martin, W. Mosaic bacterial chromosomes: A challenge en route to a tree of genomes. *Bioessays* **21**, 99–104 (1999).

[74] Martin, W. & Embley, T. M. Early evolution comes full circle. *Nature* **431**, 134–137 (2004).

[75] Merkl, R. A survey of codon and amino acid frequency bias in microbial genomes focusing on translational efficiency. *J. Mol. Evol.* **57**, 453–466 (2003).

[76] Moscoso, M. & Claverys, J. P. Release of DNA into the medium by competent streptococcus pneumoniae: Kinetics, mechanism and stability of the liberated DNA. *Mol. Microbiol.* **54**, 783–794 (2004).

[77] Murakami, Y. *et al.* Analysis of the nucleotide sequence of chromosome VI from saccharomyces cerevisiae. *Nat. Genet.* **10**, 261–268 (1995).

[78] Nakashima, H., Nishikawa, K. & Ooi, T. Differences in dinucleotide frequencies of human, yeast, and escherichia coli genes. *DNA Res.* **4**, 185–192 (1997).

[79] Nelson, K. E. *et al.* Evidence for lateral gene transfer between archaea and bacteria from genome sequence of thermotoga maritima. *Nature* **399**, 323–329 (1999).

[80] Nölling, J. *et al.* Genome sequence and comparative analysis of the solvent-producing bacterium clostridium acetobutylicum. *J. Bacteriol.* **183**, 4823–4838 (2001).

[81] Norris, J. R. *Markov Chains* (Cambridge University Press, Cambridge, 1997).

[82] Ochman, H. & Jones, I. B. Evolutionary dynamics of full genome content in escherichia coli. *EMBO J.* **19**, 6637–6643 (2000).

[83] Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299–304 (2000).

[84] Ogram, A., Sayler, G. S. & Barkay, T. The extraction and purification of microbial DNA from sediments. *J. Microbiol. Methods* **7**, 57–66 (1987).

[85] Paget, E. & Simonet, P. On the track of natural transformation in soil. *FEMS Microbiol. Ecol.* **15**, 109–117 (1994).

[86] Palmen, R. & Hellingwerf, K. J. Acinetobacter calcoaceticus liberates chromosomal DNA during induction of competence by cell lysis. *Curr. Microbiol.* **30**, 7–10 (1995).

[87] Pearson, K. On the criterion that a given system of deviations from the probable in the case of correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Phil. Mag.* **50**, 157–175 (1900).

[88] Pietrokovski, S. & Trifonov, E. N. Imported sequences in the mitochondrial yeast genome identified by nucleotide linguistics. *Gene* **122**, 129–137 (1992).

[89] Power, P. M., Jones, R. A., Beacham, I. R., Bucholtz, C. & Jennings, M. P. Whole genome analysis reveals a high incidence of non-optimal codons in secretory signal sequences of escherichia coli. *Biochem. Biophys. Res. Commun.* **322**, 1038–1044 (2004).

[90] Redfield, R. J. Genes for breakfast: The have-your-cake-and-eat-it-too of bacterial transformation. *J. Hered.* **84**, 400–404 (1993).

[91] Redfield, R. J. Do bacteria have sex? *Nat. Rev. Genet.* **2**, 634–639 (2001).

[92] Reeves, P. Evolution of salmonella O antigen variation by interspecific gene transfer on a large scale. *Trends Genet.* **9**, 17–22 (1993).

[93] Revuz, D. *Markov Chains* (North-Holland Publishing, Amsterdam, 1975).

[94] Rivera, M. C. & Lake, J. A. The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature* **431**, 152–155 (2004).

[95] Rocha, E. P. & Danchin, A. Base composition bias might result from competition for metabolic resources. *Trends Genet.* **18**, 291–294 (2002).

[96] Rozenberg-Arska, M., Salters, E. C., van Strijp, J. A., Hoekstra, W. P. & Verhoef, J. Degradation of escherichia coli chromosomal and plasmid DNA in serum. *J. Gen. Microbiol.* **130**, 217–222 (1984).

[97] Sandberg, R., Bränden, C. I., Ernberg, I. & Cöster, J. Quantifying the species-specificity in genomic signatures, synonymous codon choice, amino acid usage and G+C content. *Gene* **311**, 35–42 (2003).

[98] Schubbert, R., Renz, D., Schmitz, B. & Doerfler, W. Foreign (M13) DNA ingested by mice reaches peripheral leukocytes, spleen, and liver

via the intestinal wall mucosa and can be covalently linked to mouse DNA. *Proc. Natl. Acad. Sci. U.S.A.* **94**, 961–966 (1997).

[99] Shapiro, S. S. & Wilk, M. B. An analysis of variance test for normality. *Biometrika* **52**, 591–611 (1965).

[100] Sharp, P. M. & Li, W. H. The codon adaptation index – a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**, 1281–1295 (1987).

[101] Smith, D. R. *et al.* Complete genome sequence of methanobacterium thermoautotrophicum deltaH: Functional analysis and comparative genomics. *J. Bacteriol.* **179**, 7135–7155 (1997).

[102] Smith, J. M. The evolution of prokaryotes: Does sex matter? *Ann. Rev. Ecol. Syst.* **21**, 1–12 (1990).

[103] Stephens, R. S. *et al.* Genome sequence of an obligate intracellular pathogen of humans: Chlamydia trachomatis. *Science* **282**, 754–759 (1998).

[104] Streit, W. R. & Schmitz, R. A. Metagenomics–the key to the uncultured microbes. *Curr. Opin. Microbiol.* **7**, 492–498 (2004).

[105] Sueoka, N. Directional mutation pressure and neutral molecular evolution. *Proc. Natl. Acad. Sci. U.S.A.* **85**, 2653–2657 (1988).

[106] Surovcik, K. `MarkovTest` – R package (to be published).

[107] Takami, H. *et al.* Complete genome sequence of the alkaliphilic bacterium bacillus halodurans and genomic sequence comparison with bacillus subtilis. *Nucleic Acids Res.* **28**, 4317–4331 (2000).

[108] Thomas, C. M. & Nielsen, K. M. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol.* **3**, 711–721 (2005).

[109] Tringe, S. G. *et al.* Comparative metagenomics of microbial communities. *Science* **308**, 554–557 (2005).

[110] von Heijne, G. Signal sequences. The limits of variation. *J. Mol. Biol.* **184**, 99–105 (1985).

[111] Waack, S. *et al.* Score-based prediction of genomic islands in prokaryotic genomes using hidden markov models. *BMC Bioinformatics* **7**, 142–142 (2006).

[112] Wallenius, K. T. *Biased Sampling: The Noncentral Hypergeometric Distribution.* Ph.D. thesis, Stanford University, Stanford, CA (1963).

[113] Whitchurch, C. B., Tolker-Nielsen, T., Ragas, P. C. & Mattick, J. S. Extracellular DNA required for bacterial biofilm formation. *Science* **295**, 1487–1487 (2002).

[114] Woese, C. The universal ancestor. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 6854–6859 (1998).

[115] Woese, C. R. & Fox, G. E. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5088–5090 (1977).

[116] Woese, C. R., Kandler, O. & Wheelis, M. L. Towards a natural system of organisms: Proposal for the domains archaea, bacteria, and eucarya. *Proc. Natl. Acad. Sci. U.S.A.* **87**, 4576–4579 (1990).

[117] Wolk, C. P., Vonshak, A., Kehoe, P. & Elhai, J. Construction of shuttle vectors capable of conjugative transfer from escherichia coli to nitrogen-fixing filamentous cyanobacteria. *Proc. Natl. Acad. Sci. U.S.A.* **81**, 1561–1565 (1984).

[118] Zuckerkandl, E. & Pauling, L. Molecules as documents of evolutionary history. *J. Theor. Biol.* **8**, 357–366 (1965).

Lebenslauf

Name	Katharina Surovcik
Geburtsdatum	7. Oktober 1977
Geburtsort	Bratislava
Staatsangehörigkeit	deutsch

Schule

1984 bis 1988	Grundschule in Bratislava
1988 bis 1989	Grundschule in Wald
1989 bis 1998	Peter-Dörfler-Gymnasium Marktoberdorf Abschluss: Allgemeine Hochschulreife

Studium

1998 bis 2004	Diplomstudium der Mathematik mit Nebenfach Philosophie an der Katholischen Universität Eichstätt-Ingolstadt Abschluss: Diplom
2002 bis 2003	Auslandsstudium am The Queen's College an der University of Oxford
seit 2004	Promotionsstudium im Graduiertenkolleg 'Identification in Mathematical Models: Synergy of Stochastic and Numerical Methods' an der Georg-August-Universität Göttingen