BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

APPLICATIONS OF STATISTICAL PHYSICS AND INFORMATION THEORY TO THE ANALYSIS OF DNA SEQUENCES

by

IVO GROSSE

Diplom, Humboldt University Berlin, 1995

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
2000

Approved by

First Reader

H. Eugene Stanley, Ph.D.
University Professor
Professor of Physics

Second Reader

Sidney Redner, Ph.D.
Professor of Physics

APPLICATIONS OF STATISTICAL PHYSICS AND INFORMATION THEORY TO THE ANALYSIS OF DNA SEQUENCES

(Order No.)

IVO GROSSE

Boston University Graduate School of Arts and Sciences, 2000

Major Professor: H. Eugene Stanley, University Professor, Professor of Physics

ABSTRACT

DNA carries the genetic information of most living organisms, and the goal of genome projects is to uncover that genetic information. One basic task in the analysis of DNA sequences is the recognition of protein coding genes. Powerful computer programs for gene recognition have been developed, but most of them are based on statistical patterns that vary from species to species.

In this thesis I address the question if there exist *universal* statistical patterns that are different in coding and noncoding DNA of all living species, regardless of their phylogenetic origin. In search for such species-independent patterns I study the *mutual information function* of genomic DNA sequences, and find that it shows persistent period-three oscillations. To understand the biological origin of the observed period-three oscillations, I compare the mutual information function of genomic DNA sequences to the mutual information function of stochastic model sequences. I find that the *pseudo-exon model* is able to reproduce the mutual information function of genomic DNA sequences. Moreover, I find that a generalization of the pseudo-exon model can connect the existence and the functional form of long-range correlations to the presence and the length distributions of coding and noncoding regions.

Based on these theoretical studies I am able to find an information-theoretical quantity, the *average mutual information* (AMI), whose probability distributions are significantly different in coding and noncoding DNA, while they are almost identical in all studied species. These findings show that there exist universal statistical patterns that are different in coding and noncoding DNA of all studied species, and they suggest that the AMI may be used to identify genes in different living species, irrespective of their taxonomic origin.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Unraveling the human genome is one of the most supreme challenges in our days. Aside from raising a lot of serious ethical questions, it also confronts us with a series of tremendously hard scientific problems.

These problems mainly arise from the following discrepancy: while sequencing techniques are almost completely automatic controlled, the analysis of the sequenced data is not. Hence, the major scientific goal raised by the Human Genome Project is the extraction of biologically and medically relevant information from almost automatically sequenced DNA and RNA molecules.

In principle, biochemical methods are able to do this job, but since they are extremely expensive and time consuming, the pressing demand for alternative approaches to extract the information hidden in our genome appeared. In this situation, concepts and techniques from statistical physics and information theory turned out to be welcome tools to handle the problem of extracting valuable information from biosequences such as DNA, RNA, or amino acid chains.

The main goal of this work is the presentation of concepts and methods derived from statistical physics that we apply to problems born by a statistical analysis of biosequences.

This thesis is divided into seven parts: a brief presentation of some basic biological background, parts A - E, and an appendix. Part A (chapters 3 - 5) contains an introduction to basic concepts of information theory, a comparison of the mutual information function to correlation functions, and an introduction to Rényi and Tsallis entropies. Part B (chapters 6 - 8) is devoted to the problem of estimating entropies from finite experimental data sets and to the derivation of the Bayes estimator of Rényi, Tsallis, and (as a limiting case)

Shannon entropies. Part C (chapters 9 - 12) consists of a discussion and derivation of finite size effects that occur when estimating (generalized) entropies, the mutual information function, or correlation functions from finite experimental data. In part D (chapters 13 - 15) we study long-range correlations in genomic DNA sequences, and we present a simple statistical model that can reproduce experimentally observed correlations in genomic DNA sequences. In part E (chapters 16 - 18) we study statistical patterns that are universally different in coding and noncoding DNA, and we present four different approaches – all based on statistical physics and information theory – of how these patterns may be used to identify coding regions in un-annotated DNA sequences.

In chapter 2, we provide our reader with a brief molecular biological background, which will suffice to understand the motivation of our statistical analysis of biosequences as well as the application of our results to biological questions.

In chapter 3, we introduce several entropy functions as measures of the information content stored in symbolic sequences such as time series generated by dynamical systems, natural texts, pieces of music, climatic data, or DNA sequences. We show how these information theoretical measures can be beneficially applied to extracting the information content stored in unknown sequences.

Chapter 4 is devoted to a comparison of the mutual information function introduced in chapter 3 to correlation functions. We present two fundamental problems that arise when applying correlation functions to symbolic sequences. First, symbols must be mapped to numbers in order to compute any correlation function, and correlation functions are not invariant under changes of the map. Hence, computing correlation functions from symbolic sequences introduces some arbitrariness of choosing the function by which symbols are mapped to numbers. Second, we show that even the infinite set of all possible autocorrelation functions computable from a given (non-binary) symbolic sequence is unable to detect all statistical dependences in that sequence.

We first analyze how many parameters are necessary and sufficient to store all information about statistical dependences in sequences over an alphabet of λ symbols. Then we show that even the infinite set of all different autocorrelation functions belonging to one symbol sequence cannot determine all of these parameters. This implies that there are functional relations between different autocorrelation functions computed from one symbolic sequence.

The solution of the puzzle of why infinitely many autocorrelation functions (which are

different from each other) are unable to detect the $(\lambda-1)^2$ statistical dependences stored in a λ-ary symbolic sequence is that all autocorrelation functions form a linear space the dimension of which is even smaller than the number of parameters storing the information about the statistical dependence between all λ symbols. We present a method by which a proper basis of autocorrelation functions can be constructed chapter 4. Moreover, we deliver an algorithm, by which the functional dependence of an arbitrarily selected autocorrelation function on the basis functions can be derived.

At the end of chapter 4 we show that the mutual information function does not possess the two shortcomings mentioned above, and hence we choose the mutual information function and related information-theoretic quantities in our study of DNA sequences.

In chapter 5 we introduce two families of generalized entropies, the Rényi entropies and the Tsallis entropies. We illustrate the relation of both the Rényi and the Tsallis entropies to the Shannon entropy as well as some relations between the Rényi and the Tsallis entropies. We discuss some mathematical properties of these generalized entropies, which will become important in our study of DNA sequences.

We learn that serious problems to estimate entropies or other statistical measures arise by the uncircumventable fact that all sequences have a finite length. Hence, chapters 6 - 12 are devoted to the derivation and discussion of statistical properties of (higher order) Shannon, Rényi, and Tsallis entropies, the mutual information function, and correlation functions.

In chapter 6, we review some theorems from sampling theory and realize that estimates of population parameters can be regarded as random variables, the expectation value and variance of which we try to determine in the following chapters. These results allow us to evaluate the accuracy and reliability of those measures introduced in chapters 3 - 5.

A fundamental property of every estimator is its bias, i.e., the systematic error that is induced by estimating population parameters from finite samples. By applying Steiner's theorem about the angular momentum of rigid bodies to the problem of sampling in mathematical statistics, we learn that the maximum likelihood estimator of the variance of a normal population is biased and therefore commonly corrected. Analogously, it is our goal to derive unbiased estimators of higher order Shannon, Rényi, and Tsallis entropies, the mutual information function, and correlation functions, which we display in terms of length correction formulae.

In chapter 6, we introduce two basic statistical concepts: the maximum likelihood

and the minimum variance method. The first method results in the maximum likelihood estimator, whereas the second leads us directly to the Bayes estimator. Regarding the Bayes theorem and applying it to our task to estimate population parameters from finite samples, we come up with a relation between these two independently developed estimators.

Two approaches can be followed to improve the accuracy and reliability of statistical estimators. We can either compute the bias of commonly used estimators and then correct them or try to derive alternative estimators that inherently are more accurate and precise than their classical counterparts. In chapters 7 and 8, we follow the latter approach and derive the Bayes estimator of the Shannon entropy in chapter 7 and the Bayes estimator of the Rényi and Tsallis entropies in chapter 8.

In chapters 9, 10, 11, as well as 12 we compute the means and variances of the frequency estimators of the Shannon entropy, the mutual information function, correlation functions, as well as Rényi and Tsallis entropies, respectively. In all four cases we are successful to present length correction formulae that allow us to estimate the desired quantities without bias in a first order approximation.

Not all of these results are new. The bias of the Shannon entropy estimator, i.e., its mean systematic error, was already derived by [Herzel 1988], [Grassberger 1988], and [Li 1989], as well as [Basharin 1959], [Harris 1975], or [Levitin 1994], who also found an approximation of the variance of this estimator, i.e., its reliability, by expanding the entropy function in a Taylor series. The novel results that we contribute are the proof that we always underestimate the Shannon entropy independently of the underlying probability distribution and a relation between the variance of the entropy estimates and the variance of the corresponding logarithmic probabilities.

By exploiting the same approach as in chapter 9, we derive the bias and the variance of the mutual information estimator in a first order approximation. Although the first approximation of the bias of the mutual information is always negative, we show that the natural estimator does not always overestimate the mutual information. This means that – in contrast to chapter 9 where we could prove that the natural estimator always underestimates the Shannon entropy – such a theorem does not exist for the mutual information, and we present a counterexample in chapter 10.

In order to determine the variance of the mutual information estimates, we have to consider correlations between 1-gram and 2-gram Shannon entropies. Surprisingly, these two entropy estimates are highly correlated in biological sequences, which results in only

tiny fluctuations of the mutual information estimates.

Again, we can relate the sample variance of the mutual information estimates to the population variance of the logarithmic probability ratios. Furthermore, we derive an analogous relation for the correlators between different entropies. The basic approach that we chose to derive all previously mentioned results utilizes the Taylor expansion of the entropy function, which we truncate in order to derive our approximate results. However, since the entropy function is not analytic, the Taylor series diverges in certain parts of our given simplex. Therefore, we present an alternative approximation of the entropy function borrowed from regression theory in appendix G.

There, we derive an approximation that converges in all points of our given simplex extremely quickly. We present a completely analytic solution to this problem, which eventually leads us to the difficult task of inverting the famous Hilbert matrix. Since we can, however, explicitly display the inverse, we obtain the desired expansion coefficients by completely elementary methods.

Chapter 11 is devoted to statistical properties of correlation functions. We show that the natural estimator of correlation functions is always biased by deriving an exact estimation of the corresponding finite sample effect. This analytic result allows us to present an exact length correction formula for the natural correlation function estimator.

In chapter H.5 we derive the means and variances of the natural estimators of Rényi and Tsallis entropies, and we compare these results with the mean and variance of the natural estimator of the Shannon entropy. The significant variance of the natural entropy estimators was the motivation for our search for alternative estimators and the derivation of the Bayes estimators of (higher order) Shannon, Rényi, and Tsallis entropies in chapters 7 and 8.

In chapter 13 we study correlations in DNA sequences. Our results obtained in chapters 10 and 11 allow us to correct for finite size effects in order to detect biologically induced correlations in DNA sequences. We show that periodic oscillations of correlation functions and the mutual information, which dominate all correlations on length scales up to thousands of base pairs, are simply due to a nonuniform codon usage in protein coding DNA. We develop a simple statistical model, the pseudo-exon model, which can qualitatively and quantitatively reproduce the experimentally observed period-3 correlations in genomic DNA of yeast.

In chapter 14 attempt to understand not only the presence of long-range period-3

correlations, but also the decay of the envelop of these correlations. We show that the decay of the envelope may originate from the mosaic structure of genomic DNA, i.e. from the concatenation of alternating coding and noncoding regions in any genomic DNA. Based and the simplifying assumption that coding regions are drawn independently from a given length distribution, we can relate the length distribution of coding regions to the functional form of the decay of correlations. Specifically, we can show that an exponential length distribution will lead to an exponential decay of correlation functions and the mutual information function, while a power-law length distribution of coding regions will lead to a power-law decay of correlations.

We generalize the pseudo-exon model to the pseudo-chromosome model by taking into account the experimentally observed length distribution of coding regions, and we can show that this model is able to reproduce the long-range correlation behavior of genomic DNA in many different organisms.

Chapter 15 is devoted to an more detailed discussion of long-range correlation features in genomic DNA sequences. As we have shown in chapter 4, correlation functions computed from one symbolic sequence by choosing different mappings of symbols to numbers are not necessarily equal to each other. In chapter 15 we study correlation functions for several biologically relevant mappings, and relate the observed patterns to biological features of coding and noncoding DNA. Specifically, we show that the pronounced period-3 in correlation functions for almost any mapping is caused to a significant degree by the nonuniformity of the codon frequency distribution in coding DNA. We also show that the slow decay of the long-range C+G correlation function is related to the existence of dispersed repeats and CpG islands. As a last example, we show that periodicities of 10-11 bp in correlation functions of yeast DNA may originate from an alternation of hydrophobic and hydrophilic amino acids in yeast proteins.

In chapter 18 we present a new approach to the problem of the statistical identification of protein-coding regions in genomic DNA. This approach is based on an entropic segmentation algorithm, which attempts to divide a heterogeneous sequence into homogeneous subsequences based on the Jensen-Shannon divergence as a measure of heterogeneity. We find that this method is highly accurate in finding borders between coding and noncoding regions, and we demonstrate that it is more accurate in identifying the correct location of genes than methods based on sliding windows.

In chapter 16 we investigate if there exist universal statistical patterns that are different

in coding and noncoding DNA of all living species. We find that the probability distribution functions of the average mutual information (AMI) are significantly different in coding and noncoding DNA, while they are almost identical for all living species, ranging from simple bacteria to complex vertebrates. We show that the accuracy by which the AMI can distinguish coding from noncoding DNA is comparable to the accuracy of classical coding measures, which are trained on species-specific data sets.

In an attempt to understand the origin of the observed species-independence of the AMI distributions, we search for statistical models that could reproduce the experimentally obtained AMI distributions. For noncoding DNA we succeed, and we can show that the species-independence of the noncoding AMI distributions reflects the absence of the genetic code in noncoding DNA of any living species. However, for coding DNA we can show that the presence of the genetic code in coding DNA is not sufficient to reproduce the observed AMI distributions. This finding leads us to the conclusion that there exist additional correlations and inhomogeneities in coding DNA of all living species, which are responsible for the observed universality of the AMI distributions.

Let us finally mention that all concepts introduced in this thesis are not restricted to analyses of biological sequences. The derived statistical properties of the Bayes entropy estimators and the natural estimators of correlation functions, the mutual information, or the Shannon, Rényi, and Tsallis entropies apply to all situations in which we are confronted with inferring knowledge from experimental data. Moreover, the techniques introduced to evaluate statistical properties of correlation functions or entropies are also applicable to measures and quantities not discussed in this work.

Chapter 2

Biological Background

In this chapter, we present a brief survey of molecular biology and genetics, as far as it is relevant for motivating why this thesis has been written and for understanding the applications of statistical physics and information theory that will follow. A detailed introduction to molecular biology and genetics can be found in a series of textbooks about molecular biology or genetics such as [Wolkenstein 1983], [Knippers 1990], [Watson 1992], [Berg & Singer 1992], [Lewin 1993], [Griffith 1993], or [Kolchanov & Lim 1994], the study of which we highly recommend to interested readers.

When Robert Hooke analyzed a piece of cork under a light microscope in 1665, he discovered the first organic cells. However, it took another 170 years before Mathias Jacob Schleiden and Theodor Schwann published observations that gave birth to the discipline named cytology. Today we know that all organisms consist of cells, the nuclei of which contain their entire genetic information.

Organisms whose cells possess a nucleus are called eukaryotes, whereas those without a nucleus like, for example, bacteria are called prokaryotes. The absence of a nucleus does not mean that bacteria can live without genetic material. The defining difference between prokaryotes and eukaryotes is just the missing or existing nucleic membrane that protects the genetic material.

If we try to trace back where the genetic material is stored, we will realize a substance called chromatin, which is found in a sometimes compressed and sometimes uncompressed shape depending on the cell's cycle. As we will see in the following, the compressed shape, which we refer to as chromosomes, corresponds to the inactive state, whereas the uncompressed shape corresponds to the active state of the chromatin.

About 80% of the chromatin is built up of proteins, whereas the remaining 20% are contributed by nucleic acids. The proteins, which are mainly basic and called histones, play a fundamental, however, not yet completely discovered role in the complex framework of gene regulation in higher eukaryotes.

Nucleic acids are linear macromolecules synthesized by a polycondensation of nucleotides and thus called polynucleotides. Their building blocks, the nucleotides, are chemical compounds consisting of a base, a sugar molecule, and phosphoric acid. The nucleotides are connected such that we obtain a sugar phosphorus backbone where the bases are attached to the sugar molecules.

The sugar can either be 2-deoxyribose or ribose. In the first case, we obtain deoxyribonucleic acid called DNA, whereas we end up with ribonucleic acid called RNA in the second case. These macromolecules can either be single stranded or complexes of two non covalently bounded polynucleotides. The total length of all DNA molecules stored in one tiny cell of the human organism is about two meters. This unambiguously suggests that there is a highly hierarchical structure of the genome, since otherwise a proper functioning of a randomly compressed thread seems unimaginable.

Let us finally specify the bases that build up the nucleotides of DNA or RNA molecules. The four bases adenine, cytosine, guanine, and thymine occur in DNA nucleotides, whereas thymine is substituted by uracil in RNA molecules. These five bases can be chemically classified to belong to the following two classes: adenine and guanine are purines, whereas cytosine, thymine, and uracil are pyrimidines.

Since Watson and Crick discovered the double helix structure of the DNA [Watson & Crick 1953], we can easily understand why the number of purines equals the number of pyrimidines in all DNA molecules. Because the two complementary strands that DNA consists of are bound by hydrogen bonds between a purine and a pyrimidine, the purine-pyrimidine-ratio has to be constant for all double stranded DNA. Moreover, since adenine (A) always pairs with thymine (T) and cytosine (C) with guanine (G), we further obtain that the number of A equals the number of T and the number of C equals the number of G appearing in all double stranded DNA.

A gene is a region in the DNA that encodes one protein. Therefore, regions between them are commonly called intergenic sequences. Before we go ahead to define what coding and noncoding regions are, let us list three fundamental processes that govern molecular biology and genetics.

The process that produces an identical copy of the DNA is called replication. Today we know that this process is semi-conservative, i.e., each of the two daughter DNA molecules contains one parental and one freshly synthesized strand.

Aside from its multiplication, the production of proteins is essential for the survival of the cell. Two processes are established to perform this task. The first one, which is termed transcription, copies the coding strand of the DNA to a so called messenger RNA. This m-RNA can then leave the nucleus and swim to the ribosomes, where the second process called translation takes place.

Ribosomes can be considered as factories where the demanded polypeptides are synthesized. The genetic message now stored on the m-RNA is here translated into its corresponding amino acid chain. The molecules governing the translation process are again RNAs, however this time so called transport RNAs. They can be regarded as the molecular dictionary of the genetic code, since they guarantee an unambiguous assignment of amino acids to codons, which are defined as triplets of m-RNA nucleotides.

Let us have a closer view on the transcription of genes to m-RNAs. Strictly speaking, we were cheating when we wrote that the coding DNA strand is directly copied to the m-RNA molecule. Let us now consider the transcription process in detail.

In a first step, the DNA polymerase, i.e., the enzyme that copies the coding DNA strand to a so called pre-m-RNA, binds to the promoter. Promoters are functional sites in the genomic DNA that act as transcription start signals. We know that almost all promoters contain a $TATA$ box, which is located about 30 base pairs upstream the transcription start. Furthermore, we know that a $CAAT$ box about 75 base pairs upstream the transcription start point is vital for an efficient gene expression. However, these two features are not at all sufficient to identify promoters in un-annotated DNA sequences. Hence, much effort is paid to discover more features that are typical for real promoters and atypical for all other $TATA$ simulants.

After the polymerase has bound to the promoter, it starts copying the DNA along its coding strand until a transcription stop signal terminates the transcription. The pre-m-RNA molecule is now provided with a cap and a poly-A-tail which both are supposed to impart the m-RNA its stability and resistance against aggressive enzymes. Additionally, a number of pieces called introns are spliced out of the pre-m-RNA.

This observation implies that genes of higher eukaryotes possess a mosaic like structure of alternating exons and introns. The exons are those parts of the genes that carry the

genetic information transferred by m-RNAs to the ribosomes. Here it is then translated into its corresponding amino acid sequence determining the structure and function of the manufactured polypeptide. On the other hand, the purpose of the introns and the intergenic sequences, which make up about 95% of the entire human genome, is much more unclear [Nowak 1994].

This situation caused some people to call the noncoding DNA junk, which we think names just the opposite of what the noncoding pieces really are. Many functional groups such as promoters, enhancers, or poly-A-sites are known to be located in the noncoding DNA. Moreover, the complex network of gene regulation is supposed to be stored in the noncoding DNA, which might be related to recent findings of long range correlations in introns and intergenic sequences [Li 1992],[Peng 1992], [Voss 1992], [Peng 1993], [Peng 1994].

It is still an open question how the cell identifies coding regions in the pre-m-RNA with an accuracy of almost 100%. All artificially created techniques fail to discriminate exons from introns with a probability of at least 5% [Lapedes 1989], [Uberbacher 1991], [Fickett 1992], [Farber 1992], [Mural 1994]. The reliability of their prediction becomes in particular bad if the exons to be identified become smaller than 50 base pairs [Farber 1992].

This fact, which reflects the difficulties that statistical methods have with discriminating between exons and introns or identifying functional sites such as promoters, poly-A-sites, or splice junctions if the available sample size becomes small, perfectly motivates the goal of our following work.

Part A

Chapter 3

Entropies and Information Theory

In the previous chapter, we have seen that the main problem in molecular biology is not the sequencing of DNA, RNA, or protein chains, but the extraction of biologically relevant information from them.

Hence, we devote this chapter to information measures, i.e., quantities that measure the amount of information stored in single symbols (such as nucleotides in DNA or amino acids in protein sequences) or blocks of n of those symbols, which we in general call n-words or n-grams.

We will axiomaticly introduce the Shannon entropy $H_1(\vec{p})$ in section 3.2, which defines the average amount of uncertainty in one symbol of a given sequence. The four axioms, by which the Shannon entropy can be uniquely defined, display some powerful mathematical properties, which we will discuss in the same section.

In section 3.3, we will generalize the Shannon entropy by introducing higher order entropies H_n. As in the previous section, we will devote the second part of section 3.3 to a discussion of some properties of higher order entropies.

Conditional entropies h_n will be defined in section 3.4, in which we will also discuss some mathematical properties of the series h_n computed from stationary and ergodic sources.

The following four sections, 3.5, 3.6, 3.7, and 3.8, are dedicated to the introduction of the mutual information function. In section 3.5, we present a first definition of this function as a measure that quantifies the information gain about a symbol X by obtaining another symbol Y.

Three possible generalizations of this preliminary mutual information make up the contents of sections 3.6, 3.7, and 3.8. In section 3.6, we give up the restriction to single

symbols and thus obtain the higher order mutual information as a measure of correlations between sub-words of arbitrary size.

We will present the mutual information between two cylinders of different sources in section 3.7 and finally introduce the Kullback entropy in section 3.8. We will realize the mutual information as a special case of the Kullback entropy and thus develop a deeper understanding of what the mutual information measures.

In section 3.8, we will also show that the mutual information only vanishes, if really all appearing symbols are statistically independent.

Finally, we will present a short summary of all results collected in this chapter in section 3.9.

Although all results presented in this chapter can be found elsewhere in the literature [Ebeling & Feistel 1982], [Grassberger 1988], [Herzel 1988], [Khinchin 1957a], [Khinchin 1957b], [Kolmogorov 1958], [Kullback 1959], [Leven 1989], [McMillan 1953], [Pompe 1994], [Rényi 1982], [Shannon 1948a], [Shannon 1948b], [Yockey 1992], we consider it beneficial for our reader

- to display in one chapter the definitions of information theoretical measures that we will use in the remainder of this work,
- to present some intuitive explanations of these measures, and
- to exhibit an encyclopedic collection of useful theorems,

since concepts derived from theory have already been proven to be extraordinarily fruitful in various disciplines such as statistical physics, bioinformatics, or linguistics.

Since the realization of the basic concepts behind all definitions presented in this chapter are essential for understanding the methods by which we will analyze biological sequences, we recommend all readers to become familiar with, for example, the concept of an information source, before going over to the following chapters, in which we start our statistical analysis of DNA sequences.

3.1 Introduction to Information Theory

In this section, we want to present some basic ideas behind what is commonly called information theory. We will start with analyzing the communication process and dissect

it into five elementary steps according to [Shannon, 1948b]. We will introduce the concept of an information source and rationalize why we can derive all statistical properties of a stationary and ergodic source from a single, however, in principle infinitely long sample sequence. We will define a probability measure on the set of all cylinders of a given information source and show that, for stationary and ergodic sources, this measure can be interpreted as the probability to find the corresponding substrings in the infinitely long sample sequence generated by the considered source. Finally, we will mention a serious problem arising from the uncircumventable fact that all available sequences always have a finite length and present an example that might first illustrate all definitions and concepts delivered in this section and second give us an insight into techniques designed to distinguish coding from noncoding DNA.

If we want to communicate with others, the following five elementary steps are essential for the transmission of any kind of information:

1. The message that we wish to send has to be produced and emitted.
2. We have to encode the message that we are going to send.
3. We have to transmit this encoded message to our desired receiver through a possibly noisy channel.
4. The receiver has to decode our received message.
5. Our destined subject or object has to receive this decoded message.

Even though this classification, which was introduced by Shannon in 1948, looks slightly artificial, it has been proven to be very valuable for a scientific analysis of the communication process and will be proven to be extremely fruitful for a statistical analysis of biological sequences.

Please note that the *information source* as well as the *destination* do not necessarily have to be humans. The basic concept sketched above also holds for the communication between animals or even between computers, machines, and other technical devices.

The central question in information theory, which we can now understand, is how we can determine the capacity of a channel. Here, the channel capacity is understood as the maximum amount of information that this channel is capable to transmit per time.

Before we however present a proper definition of the *amount of information* stored in a message, we first want to explain what is commonly understood by an *information source*.

A mathematical definition of the information source was presented by McMillan in [1953] and Khinchin in [1957a, 1957b]. Let us at this point just briefly outline their basic ideas.

What an information source basically does is producing a symbolic sequence

$$\{x_t\}_{t\in\mathcal{G}} = \ldots x_{-2}\, x_{-1}\, x_0\, x_1\, x_2 \ldots$$

with an, in principle, infinite length. $\mathcal{G}$ denotes the set of all positive and negative integer numbers and x_t defines the symbol that we find at position t in our given sequence. The finite set of all appearing symbols is called an *alphabet* $A \equiv \{A_1, A_2, ..., A_M\}$ where M is the number of different symbols appearing in our sequence of consideration and therefore called the *alphabet size*.

German and English texts are composed of the 26 letters $\{A, B, C, ..., X, Y, Z\}$, so that, in this particular case, the term we defined as alphabet is identical to the Latin alphabet. In DNA sequences, where the four letters A, C, G, and T denote the four occurring nucleotides *adenine*, *cytosine*, *guanine*, and *thymine*, the alphabet of size 4 would simply be the set $\{A, C, G, T\}$.

There is, of course, an infinite number of sequences belonging to one information source, everyone of which can be understood as one possible realization of this source.

If we now restrict, at some arbitrarily chosen positions t_i, our symbols x_{t_i} to be equal to given symbols $y_i \in A$, then we call the set containing all possible realizations of the source that are compatible with these constraints a *cylinder*. At this point, we can assign probabilities to any given cylinder of a source and thus end up with a probability measure μ reflecting the set of all assignments of probabilities to all cylinders of a source. This assignment, i.e., the probability measure μ, unambiguously characterizes the information source.

Two special cylinders will become the focus of our attention in the remainder of this work. One is the cylinder that is given by n consecutive indices, i.e., by $t_1 = \tau, t_2 = \tau + 1, ..., t_n = \tau + n - 1$. If we choose the symbols $y_1, y_2, ..., y_n$ as those defining our particular cylinder, then its measure μ is interpreted as the probability to find the *n-word* or *n-gram* $(y_1, y_2, ..., y_n)$ at position τ. This probability is, in the following, simply denoted by $p_{(y_1,y_2,...,y_n)}(\tau)$.

The second class of cylinders we will deal with in the following are those defined by two symbols y_1 and y_2 at the two positions t_1 and t_2 that are k positions apart from each

other, i.e., $t_2 - t_1 = k$. By setting $\tau \equiv t_1$, we analogously obtain the probability to find the symbol y_1 at position τ and the symbol y_2 k positions later with its common denotation $p_{(y_1,y_2)}(k,\tau)$.

In case of stationary sources, the probabilities $p_{(y_1,y_2,...,y_n)}(\tau)$ and $p_{(y_1,y_2)}(k,\tau)$ do not depend on τ, i.e., the probability to find a certain string does not depend on the position within the sequence.

If, moreover, the source is also ergodic, then we can consider the relative frequencies of substrings that we derive from one infinitely long realization (which we commonly call a sample sequence) as their probabilities, i.e., then the *time averages* converge (in probability) to the corresponding ensemble averages.

Although it is, in principal, trivial to derive the probabilities $p_{(y_1,y_2,...,y_n)}$ or $p_{(y_1,y_2)}(k)$ for any $y_1, y_2, ..., y_n$ from infinitely long sequences produced by a stationary and ergodic source, the estimation of these probabilities can become extremely hard or even impossible if only finite sequences are available.

Imagine, for example, we want to estimate the probability, by which the 6-mer of nucleotides $ACTTGT$ appears in a given DNA sequence with a length of 10,000 base pairs. Since there are 4096 different 6-mers possible, we cannot expect to find the substring $ACTTGT$ much more frequently than twice in our entire sequence (provided that the probability p_{ACTTGT} does not drastically deviate from 1/4096).

This situation, however, corresponds to the task to estimate the probability of a coin by flipping it twice.

The situation in biology is even worse. Although there are sequences as long as 10,000 base pairs and longer, one of the main goals of computational molecular biology is to develop algorithms that can reliably distinguish between coding and noncoding DNA. As we, however, have learned in chapter 2, exons and introns are often shorter than 100 base pairs.

Since, on the other hand, significant differences between the 6-mer frequencies of eukaryotic exons and introns exist, these differences are desired to be exploited by algorithms that are designed to discriminate coding from noncoding DNA.

Surprisingly, these approaches, which make up the central part of the Gene Recognition Module GRAIL installed at the Oak Ridge National Laboratory [Uberbacher 1991] and [Mural 1994], are really successful. Three of the twelve algorithms that base on statistical differences between coding and noncoding DNA in humans exploit differences in the 6-mer

composition between exons and introns. This means that the 4096 dimensional vector containing the absolute frequencies of the observed 6-mers, which often contains more than 4000 zeros, carries some valuable information about the decision whether or not the underlying DNA strand is protein coding.

3.2 Shannon Entropy

In this section, we will introduce the Shannon entropy as a measure of uncertainty. Strictly speaking, the Shannon entropy characterizes a source and thus a set of produced sequences, but not a particular sequence. This means that the Shannon entropy is a functional of the probability distribution over all possible sequences that our considered source can produce.

Theoreticians like Shannon [1948b], Khinchin [1957b], Rényi [1957], and Kolmogorov [1958] have introduced some requirements that functions have to fulfill in order to be considered a measure of uncertainty. These requirements, which eventually build up a complete system of axioms, are presented below.

Axiom 3.1 *Let $p_1, p_2, ..., p_M$ be the components of the M-dimensional probability vector $\vec{p}$ and $H(\vec{p})$ a scalar function of $\vec{p}$. Then the first axiom requires the function $H(\vec{p})$ be smooth in $\vec{p}$, which means that tiny variations of the probabilities p_i cause only tiny variations of the values of H.*

Axiom 3.2 *Let M be constant and the positive probabilities p_i fulfill the normalization constraint $\sum_{i=1}^{M} p_i = 1$. Then $H(\vec{p})$ becomes a maximum if $p_i = 1/M$ for all $i = 1, 2, ..., M$; i.e.,*

$$H\left(\frac{1}{M}, \frac{1}{M}, ..., \frac{1}{M}\right) \geq H(p_1, p_2, ..., p_M) \tag{3.1}$$

for all $p_1, p_2, ..., p_M$. Hence, the second axiom requires that the source is transmitting symbols with the maximum amount of uncertainty if all symbols are sent with the same probability $1/M$.

Axiom 3.3 *If we formally add an impossible event, i.e., $p_{M+1} = 0$, then the third axiom requires*

$$H(p_1, p_2, ..., p_M, 0) = H(p_1, p_2, ..., p_M). \tag{3.2}$$

This means that a desirable measure should not be affected by the formal addition of impossible events.

Axiom 3.4 *Let p_i be the probabilities for the events x_i, q_j be the probabilities for the events y_j, and P_{ij} be the joint probabilities to observe the events x_i and y_j simultaneously. Let us further denote the vector $(p_1, p_2, ..., p_M)$ by $\vec{p}$, the vector $(q_1, q_2, ..., q_M)$ by $\vec{q}$, and the $M \times N$ matrix containing the elements P_{ij} by $\hat{P}$. If we now define the conditional probabilities $p(x_i|y_j) \equiv P_{ij}/q_j$ and $p(y_j|x_i) \equiv P_{ij}/p_i$ for all non-vanishing p_i and q_j, the fourth axiom requires the following equality:*

$$H\left(\hat{P}\right) = H\left(\vec{p}\right) + \sum_{i=1}^{M} p_i \cdot H\left(p(y_1|x_i), p(y_2|x_i), ..., p(y_N|x_i)\right) \tag{3.3}$$

$$= H\left(\vec{q}\right) + \sum_{j=1}^{N} q_j \cdot H\left(p(x_1|y_j), p(x_2|y_j), ..., p(x_M|y_j)\right) \tag{3.4}$$

for all multivariate distributions $\hat{P}$.

By defining the **conditional entropies**

$$H\left(\vec{p}|\vec{q}\right) \equiv \sum_{j=1}^{N} q_j \cdot H\left(p(x_1|y_j), p(x_2|y_j), ..., p(x_M|y_j)\right) \tag{3.5}$$

and

$$H\left(\vec{q}|\vec{p}\right) \equiv \sum_{i=1}^{M} p_i \cdot H\left(p(y_1|x_i), p(y_2|x_i), ..., p(y_N|x_i)\right), \tag{3.6}$$

we obtain the fourth axiom in the memorizable form

$$H\left(\hat{P}\right) = H\left(\vec{p}\right) + H\left(\vec{q}|\vec{p}\right) = H\left(\vec{q}\right) + H\left(\vec{p}|\vec{q}\right). \tag{3.7}$$

A conclusion of the fourth axiom is that, for independent events X and Y, i.e., if $P_{ij} = p_i \cdot q_j$ for all $i = 1, 2, ..., M$ and $j = 1, 2, ..., N$, the measure of uncertainty is simply additive, i.e.,

$$H\left(\hat{P}\right) = H\left(\vec{p}\right) + H\left(\vec{q}\right). \tag{3.8}$$

In section 3.8, we will show that this equality only holds if the two events X and Y are statistically independent. This means that the single equation $H\left(\hat{P}\right) = H\left(\vec{p}\right) + H\left(\vec{q}\right)$ implies the $M \cdot N$ equalities

$$P_{ij} = p_i \cdot q_j \tag{3.9}$$

for $i = 1, 2, ..., M$ and $j = 1, 2, ..., N$.

On the other hand, we realize $H\left(\vec{q}|\vec{p}\right) = 0$ if X and Y are functionally dependent. In this case, we obtain the equality $H\left(\hat{P}\right) = H\left(\vec{p}\right) = H\left(\vec{q}\right)$.

In general, we can state the following inequalities for all $\hat{p}$:

$$0 \leq H\left(\vec{p}|\vec{q}\right) \leq H\left(\vec{p}\right) \tag{3.10}$$

as well as

$$0 \leq H\left(\vec{q}|\vec{p}\right) \leq H\left(\vec{q}\right) \tag{3.11}$$

and thus

$$max\left(H\left(\vec{p}\right), H\left(\vec{q}\right)\right) \leq H\left(\hat{P}\right) \leq H\left(\vec{p}\right) + H\left(\vec{q}\right). \tag{3.12}$$

As Khinchin shows in [1957b], the four axioms displayed above determine the following functional form of the desired measure of uncertainty $H\left(\vec{p}\right)$:

$$H\left(p_1, p_2, ..., p_M\right) = -C \sum_{i=1}^{M} p_i \cdot \ln(p_i) \tag{3.13}$$

where C is a constant that defines the units in which we desire to measure the amount of uncertainty. This means that the measure of uncertainty compatible to all four axioms displayed above has to be proportional to the average logarithm of the probabilities p_i.

Therefore, we define the *Shannon entropy*

$$H\left(p_1, p_2, ..., p_M\right) \equiv -\sum_{i=1}^{M} p_i \cdot \log(p_i) \tag{3.14}$$

where the base to which we take the logarithm corresponds to the units in which we wish to measure the average amount of uncertainty stored in one symbol of our stationary and ergodic sequence.

If we identify this constant C with the Boltzmann constant k_B, we receive the microscopic definition of the thermodynamic entropy.

Setting $C = 1/\ln(2)$ yields

$$H\left(p_1, p_2, ..., p_M\right) = -\sum_{i=1}^{M} p_i \cdot \log_2(p_i), \tag{3.15}$$

which identifies the elementary units of uncertainty with binary digits. In this case, where we measure the amount of uncertainty in bits, we end up with a very instructive interpretation of the Shannon entropy.

Then the Shannon entropy $H\left(p_1, p_2, ..., p_M\right)$, which is a measure of the average amount of uncertainty in one symbol of our given sequence, is the (average) number of binary

questions that we must ask in order to identify a particular symbol provided we use an optimal strategy of asking.

Let us, at this point, present an example that will explain this interpretation.

Imagine we are sitting in front of a deck of 32 hidden cards, which all are different, and want to guess which card is on top of the pile. The deeper question is how much we gather if somebody tells us the answer. Applying our informal definition of the Shannon entropy, we start asking binary questions about the uppermost card. Let us, for the sake of simplicity, assume that all cards are numbered from 1 through 32.

Since each card exists exactly once, an optimal asking strategy would be the following:

1. Does the card on top of the pile belong to the first 16, i.e., is the number on this card ≤ 16?

2. Depending on the answer on question 1, yes or no, we would continue asking: Does the card belong to the first ... or third ... group of 8?

3. And so on.

After having asked the fifth question, we will know for sure which card is on top of the pile, i.e., the amount of uncertainty is equal to 5 bit in this case.

Calculating

$$H\left(p_1, p_2, ..., p_M\right) = -\sum_{i=1}^{M} \frac{1}{M} \cdot \log_2\left(\frac{1}{M}\right) = \log_2(M) \tag{3.16}$$

yields the same result for $M = 32$ since $2^5 = 32$.

Imagine we are now interested in the amount of uncertainty in a single letter in an English text. A foreign friend[1] whose only knowledge about English be that there are 32 letters would be confronted with the same task of guessing letters as we were with the task of guessing cards.

For him, the amount of uncertainty per letter would be 5 bit as well.

However, if we ask an Englishmen to guess a letter that we randomly picked from an English text, he would definitely need less than 5 questions on average to determine the hidden letter. Two reasons contribute to this decrease of the Shannon entropy per letter.

[1]Please note that this is a hypothetic example since we believe that foreign people might, on average, understand ten-thousand times more about English than we about their foreign language.

First, there are only 26 letters in the Latin alphabet and not 32. And second, these 26 letters are by far not equidistributed. An E appears much more frequently than, for example, an X. Hence, an optimal strategy would probably start with the first question: Is the hidden letter a vowel?

Shannon was the first who analyzed the entropy of English texts [Shannon 1951]. His results, which have not been changed over the last five decades, state that the amount of uncertainty per letter in English texts is about 4.03 bit. This means that we should be able to recover a lost letter by asking somebody who found it about 4 binary questions on average.

However, what happens if we want to guess letters from a context? Imagine for example we are reading a book and suddenly are coming across a spot where black ink hides exactly one letter. The sentence reads as follows: "The sky is blu%." The % marks the letter that we cannot identify and thus have to guess. But do we really have to ask more than four times to guess the letter completely?

Of course, not! The answer is that the amount of uncertainty in this letter (e) is indeed very close to zero. The question that we pose with this statement is where this entropy decrease originates from?

In the following section, we will introduce higher order entropies and conditional entropies, which will eventually allow us to quantify the amount of uncertainty in a symbol that is embedded in a context.

We will learn that correlations between a symbol that we want to guess and its environment always decrease the Shannon entropy of this symbol and hence increase its redundancy.

3.3 Higher Order Entropies

In this section, we will introduce Shannon's n-gram entropies H_n, which define the amount of uncertainty in a sub-word of length n.

As we have already mentioned in section 3.1, a source is completely and unambiguously determined by its probability measure μ. Vice versa, we are automatically given all n-gram probabilities once our source is specified. Please let us neglect the problems that arise from estimating these n-gram probabilities from finite sequences for the next moment.

In section 3.2, we have learned that the function

$$H\left(p_1, p_2, ..., p_M\right) = -\sum_{i=1}^{M} p_i \cdot \log_2(p_i), \tag{3.17}$$

measures the average amount of uncertainty in one single letter.

Let us now formally identify all possible sub-words of length n that can be composed by letters y_i derived from an alphabet of size λ with new symbols A_j. Note that $i = 1, 2, ..., \lambda$, but $j = 1, 2, ..., M$ with $M = \lambda^n$. Of course, there are $M = \lambda^n$ different sub-words of length n composable from letters belonging to a λ-ary alphabet.

Denoting the n-gram probabilities by $p_1^{(n)}, p_2^{(n)}, ..., p_M^{(n)}$, we obtain the *higher order entropies*

$$H_n\left(p_1^{(n)}, p_2^{(n)}, ..., p_M^{(n)}\right) = -\sum_{i=1}^{M} p_i^{(n)} \cdot \log_2\left(p_i^{(n)}\right) \tag{3.18}$$

as a measure of uncertainty in sub-words of length n.

Again, we choose the dual logarithm in order to measure the amount of uncertainty in bit. Then, the a higher order entropy of 12 bit means we have to ask 12 binary questions to get to know the corresponding sub-word.

As Khinchin shows in [1957a], the series $\{H_n\}$ is monotonically increasing for stationary and ergodic sources, i.e.,

$$H_m \leq H_n \tag{3.19}$$

for all $m < n$.

This statement is intuitively clear if we apply our informal definition of the Shannon entropy. If we have to ask H_n times to get to know a sub-word of length n on average, then we need at least the same number of questions to ask for a sub-word of length $n + 1$. The amount of uncertainty in the last letter cannot be negative, since entropies are always nonnegative.

Let us consider a purely random and a periodic sequence to exemplify the definition of higher order entropies.

In the following, we will call a sequence Bernoulli-like or, in short, Bernoulli sequence, if all λ symbols appear statistically independent of each other. This means that the probability to find a sub-word of length n is identical to the product of the probabilities of its components, i.e.,

$$p_{(y_1, y_2, ..., y_n)} = p_{y_1} \cdot p_{y_2} \cdots p_{y_n}. \tag{3.20}$$

Exploiting this equality for Bernoulli sequences immediately yields

$$H_n = n \cdot H_1, \tag{3.21}$$

which is intuitively clear: H_1 is the average amount of uncertainty per letter. This means we have to ask H_1 times to get to know this letter. How often do we then have to ask a binary question in order to guess the second letter of our n-gram? Since all letters are statistically independent in our sequence, we cannot infer anything about the letter at position 2 by getting to know the first letter and thus have to ask again H_1 binary questions. Hence, we have to ask $n \cdot H_1$ questions to obtain the identity of a sub-word of length n, which is in accord to the fourth axiom stating the additivity of the Shannon entropy for independent variables.

The same approach allows us to derive an analytic relation among higher order entropies of periodic sequences. If the sub-words are longer than the period, which we denote by l, i.e., $n > l$, then we do not have to guess the last $n-l$ symbols, because they are completely determined by the first l symbols. Hence, we obtain

$$H_n = H_l \tag{3.22}$$

for all $n \geq l$.

Let us now come back to the question how we can exploit the properties of the series $\{H_n\}$ to detect correlations in our underlying sequence.

Two examples might serve as a motivation why many scientists are interested in correlations that appear in symbolic sequences such as natural texts, pieces of music, computer programs, or DNA.

On the one hand, there is the increasing demand to compress data. However, data can only be compressed if they contain some redundancy. But redundancies are intrinsicly related to correlations within the analyzed sequence.

If there is a set of rules that have to be obeyed such as spelling rules and the grammar of a certain language, then we have less freedom to construct permitted sequences. This, however, decreases the number of questions we have to ask in order to obtain a hidden symbol or sub-word.

Imagine, for example, we are to guess an English word and have already found out one particular letter be a q. Then we do not have to ask at all for the next letter since we know it must be an u. This example clearly demonstrates how correlations (which are weak forms of restrictions) decrease higher order entropies and thus increase redundancies.

Aside from demanding better data compression algorithms, many biologists and physicians are interested in understanding gene regulation in higher eukaryotes. The discovery of the existence of long-range correlations in noncoding DNA sequences [Li 1992], [Peng 1992], [Voss 1992], which however seem to be not present in protein coding pieces, raised a loud controversy and hence the demand for a careful analysis of all measures and techniques used to detect correlations in symbol sequences.

In chapter 11, we will introduce some classical techniques like the analysis of random walk fluctuations, power spectra, or autocorrelation functions, by which correlations are commonly measured. We will show how all of these techniques are related to each other and thus concentrate on the question whether or not correlation functions are able to detect all kinds of correlations hidden in a symbol sequence. We will present some novel results regarding this question and outline a method, by which all statistical dependences between symbols or sub-words within a symbolic sequence can be detected.

We will see that the mutual information, an entropy-related measure derived from information theory, is a welcome tool to detect all kinds of correlations in a symbol sequence. Therefore, we will now come back to our question for a quantity that defines the amount of uncertainty that is remaining in a symbol, if the preceding n symbols are known. If we then apply this measure, which we term *conditional entropy*, to our sentence "The sky is blu%," we will realize that this conditional entropy is indeed much smaller than four bit. This means we do not have to ask about four times for an illiterate letter in English texts, if we are given some preceding letters and know some basic orthographic and grammatical rules.

The effect that we have to ask the less, the more we understand about English, i.e., the more regulations we know, again illustrates the crucial relation between the strength of correlations and redundancy.

Let us in the following section define the conditional entropies h_n, which define the amount of uncertainty in a symbol y_{n+1} if the preceding n symbols $y_1, y_2, ..., y_n$ are known.

3.4 Conditional Entropies

Please recall that we denote Shannon's n-gram entropies, which define the amount of uncertainty in a string of length n, by H_n.

Let us in this chapter define the *conditional entropy* h_n as the difference between the

two higher order entropies H_{n+1} and H_n, i.e.,

$$h_n \equiv H_{n+1} - H_n. \tag{3.23}$$

The quantity h_n measures the amount of uncertainty in one symbol provided we know the preceding n-gram, i.e., the conditional entropy h_n quantifies the number of questions we must ask in order to identify the last letter of our $(n+1)$-gram provided we already know the first n letters.

In the following paragraph, we will motivate the term conditional entropy, i.e., we will show that h_n can be written as the average logarithm of the conditional probabilities $p(y_{n+1}|y_1, y_2, ..., y_n)$, whereas the Shannon entropy H_n is given by the average logarithm of the n-gram probabilities $p(y_1, y_2, ..., y_n)$, as we have seen in section 3.3.

Let $p(S_n) \equiv p(y_1, y_2, ..., y_n)$ be the probability to find the string $S_n \equiv y_1 y_2 ... y_n$ of length n at any arbitrary position τ in our stationary and ergodic sequence. Let further

$$p(y_{n+1}|S_n) \equiv \frac{p(S_n, y_{n+1})}{p(S_n)} \tag{3.24}$$

be the conditional probability to find the symbol y_{n+1} following the string S_n, where $p(S_n, y_{n+1})$ denotes the probability to find the $(n+1)$-gram $y_1 y_2 ... y_{n+1}$.

Then the conditional entropy h_n can be rewritten in the following manner:

$$\begin{aligned}
h_n &= H_{n+1} - H_n && (3.25)\\
&= -\sum_{\{S_n y_{n+1}\}} p(S_n, y_{n+1}) \cdot \log_2(p(S_n, y_{n+1})) + \sum_{\{S_n\}} p(S_n) \cdot \log_2(p(S_n)) && (3.26)\\
&= -\sum_{\{S_n y_{n+1}\}} p(S_n, y_{n+1}) \cdot \log_2(p(S_n, y_{n+1}))\\
&\quad + \sum_{\{S_n y_{n+1}\}} p(S_n, y_{n+1}) \cdot \log_2(p(S_n)) && (3.27)\\
&= -\sum_{\{S_n y_{n+1}\}} p(S_n, y_{n+1}) \cdot \log_2\left(\frac{p(S_n, y_{n+1})}{p(S_n)}\right) && (3.28)\\
&= -\sum_{\{S_n y_{n+1}\}} p(S_n, y_{n+1}) \cdot \log_2\left(p(y_{n+1}|S_n)\right), && (3.29)
\end{aligned}$$

which states that the conditional entropy is given by the average logarithm of the conditional probabilities $p(y_{n+1}|S_n)$ and hence the average uncertainty about the symbol y_{n+1} if the preceding n symbols $y_1, y_2, ..., y_n$ are known.

Since the amount of uncertainty about a symbol y_{n+1} can never be increased by providing the observer with some additional information about the symbol y_0, we can state that

$$h_n \geq h_{n+1} \tag{3.30}$$

for all $n \in \mathcal{N}$, i.e., the series of the conditional entropies h_n is monotonicly decreasing.

This relation between the conditional entropies h_n can be back-translated into a nice property of the higher order entropies H_n. Since the conditional entropy h_n as a function of n can be understood as the first derivative of the higher order entropy H_n as a function of n, the monotonic decrease of the function h_n implies the convexity of H_n as a function of n.

This means that although the higher order entropies are always increasing with rising n, the increase becomes smaller and smaller.

Since, in addition to the monotonicity of h_n, the conditional entropies are always nonnegative, i.e., the set of all h_n is bounded from below, the series $\{h_n\}_{n\in\mathcal{N}}$ converges to a nonnegative number

$$h \equiv \lim_{n\to\infty} h_n. \tag{3.31}$$

This quantity h, which is called the *entropy of the source*, quantifies the amount of uncertainty remaining in a symbol, if all preceding symbols of our infinitely long, stationary, and ergodic sequence are known.

Please note that the quantities H_n/n, which can be defined as the average amount of uncertainty of an n-gram per letter, also converge to the entropy of the source, i.e.,

$$\lim_{n\to\infty} \frac{H_n}{n} = \lim_{n\to\infty} h_n = h. \tag{3.32}$$

Hence, the quantity H_n/n is a good measure of correlations as well. In fact, the number $1 - \frac{H_n}{n\cdot\log(\lambda)}$ is chosen as a measure of redundancy by Gatlin [1972], Ebeling [1982], and Mantegna [1994, 1995], who analyzes linguistic differences between coding and noncoding DNA in these s.

We, however, stick to the conditional entropies h_n for three reasons:

- The numbers h_n converge quicker to the entropy of the source than the entropies per letter H_n/n.

- The conditional entropies h_n with their interpretation as the average uncertainty about a symbol given the preceding n-gram are closer to the correlation measure we are looking for than the quantities H_n/n.

- The mutual information function, which we will introduce in the next section, can easily be defined (and understood) in terms of the conditional entropies h_n.

Before we, however, go over to our next section, where we introduce the mutual information function, let us present two examples in the remainder of this section that shall illustrate the meaning of conditional entropies.

Imagine a Bernoulli sequence. Then all conditional entropies h_n are equal to H_1, i.e.,

$$h_n = H_1, \tag{3.33}$$

since $H_n = n \cdot H_1$ for all $n \in \mathcal{N}$. This result is not at all surprising, because all symbols are statistically independent and hence uncorrelated so that we do not learn anything about the following symbol in our sequence independently on how many preceding symbols we have gotten to know. Therefore, we have to ask H_1 times for a single symbol independently on how many preceding symbols we are given, i.e., the conditional entropies h_n are always identical to H_1.

If we consider a periodic sequence, then we obtain

$$h_n = 0 \tag{3.34}$$

by applying eq. (3.22) for all $n \geq l$, where l denotes the length of the period. This is intuitively clear, since we can predict (without any uncertainty) the symbol y_{n+1} following the n-gram S_n if $n \geq l$.

The series $\{h_n\}$ of all other stationary and ergodic sequences range within these two limiting cases and tell us, in principle, everything about statistical dependences between symbols or substrings within our sequence. The practical problem that we are, however, confronted with while analyzing English texts, pieces of music, or biological sequences is the finite size of all these samples. In order to get a reliable estimate about the conditional probability $p(y_{n+1}|S_n)$, we have to make sure that the corresponding $(n+1)$-gram $S_n y_{n+1}$ occurs at least once in our sample.

This restriction would, however, never allow us to analyze correlations on length scales of hundreds or thousands of symbols, since even in the best case that we are dealing with

binary sequences, the required minimum sequence length would be 2^{1000}, which is a number that exceeds by far the number of atoms in our entire universe.

In the following section, we will introduce the mutual information function, which is not affected by those serious combinatorial problems mentioned above. Since the mutual information will turn out to measure exactly what we understand by redundancy, we will use the mutual information function to measure correlations in symbolic texts.

3.5 Mutual Information

In this section, we will present the mutual information as a measure of correlations between symbols within one sequence. Since we analyze these correlations depending on the distance between the considered symbols, we obtain the mutual information as a function of the inter-symbol-distance, which we will briefly call the mutual information function.

The philosophy behind the definition of this measure is extremely simple. Instead of considering substrings of length n, we will now deal with cylinders that are given by a pair of two symbols in a distance k, as outlined in section 3.1.

Then we can define a conditional entropy $h_1(k)$ by the following equation:

$$h_1(k) \equiv H_2(k) - H_1, \tag{3.35}$$

in which H_1 and $H_2(k)$ are defined by

$$H_1 \equiv -\sum_{i=1}^{\lambda} p_i \cdot \log(p_i) \tag{3.36}$$

and

$$H_2(k) \equiv -\sum_{i,j=1}^{\lambda} P_{ij}(k) \cdot \log\left(P_{ij}(k)\right) \tag{3.37}$$

where p_i denotes the probability to find the symbol A_i at any arbitrary position in our stationary sequence and $P_{ij}(k)$ names the joint probability to find the two symbols A_i and A_j (in this order) k positions apart from each other. Please recall that $k = 1$ corresponds to adjacent positions and note that $p_i = p_{A_i}$ and $P_{ij}(k) = p_{(A_i,A_j)}(k)$ according to our denotation introduced in section 3.1.

In analogy to section 3.4, we obtain that $h_1(k)$ quantifies the average amount of information hidden in a symbol of our stationary and ergodic sequence provided we know the symbol that appears k positions upstream. If this conditional information is measured

in bit, then $h_1(k)$ is equal to the average number of questions we have to ask in order to obtain the information about the second symbol provided we know the first symbol.

Let us now compare this information $h_1(k)$ with the Shannon entropy H_1, which is equal to the average number of binary questions we have to ask about the second symbol provided we do not know the symbol k positions upstream. The difference between these two entropies gives us the information that we gain about the second symbol by getting to know the first one.

This difference is exactly what we are looking for as a correlation measure between two symbols that appear k positions apart from each other in a given stationary and ergodic sequence.

Hence, we define

$$I(k) \equiv H_1 - h_1(k) \tag{3.38}$$

as the *mutual information function* of our given information source.

Applying eqs. (3.35), (3.36), and (3.37) to this definition yields

$$I(k) = 2 \cdot H_1 - H_2(k) = \sum_{i,j=1}^{\lambda} P_{ij}(k) \cdot \log\left(\frac{P_{ij}(k)}{p_i \cdot p_j}\right), \tag{3.39}$$

which is the definition usually found in the literature [Kullback 1951b], [McEliece 1977], [Pompe 1986], [Ebeling 1992], [Yockey 1992], [Herzel 1994a].

Let us at this point again consider a Bernoulli sequence to illustrate the definition given above.

Since, in the case of Bernoulli sequences, $h_n = H_1$ for all n, we obtain

$$I(k) = 0 \tag{3.40}$$

for all $k \in \mathcal{N}\backslash\{0\}$.

This result exactly corresponds to our intuitive understanding of what the mutual information defines. In Bernoulli sequences, where all symbols are statistically independent of each other, we do not gain any information about the current symbol by getting to know any previous one.

In deterministic sequences, were all h_n are zero, i.e., there is no uncertainty about the next symbol provided we know at least one symbol that appeared in the past, the mutual information function is constant at its highest possible value H_1. In mathematical terms,

$$I(k) = H_1 \tag{3.41}$$

for all $k \in \mathcal{N}\backslash\{0\}$.

This means not only that the entire information available about the symbol to be guessed is provided by getting to know any previous symbol of our deterministic sequence, but also that the information H_1 that any symbol of this sequence can provide us with is entirely exploited to predict the symbol to be guessed.

In chapter 13 we will deal with some delicate problems that arise by analyzing correlations in periodic sequences. We will see that even slight periodicities give rise to periodic mutual information functions as well as to periodic correlation functions. Since periodicities in coding DNA sequences are trivially generated by a nonuniform codon usage, which is typical for coding but atypical for noncoding sequences, the periodicity of the mutual information function can be exploited to identify coding DNA pieces.

But before we start discussing some important mathematical properties of the mutual information function, let us outline three possible generalizations of this 2-point correlation measure.

In our first step, we will not only consider symbol-symbol-correlations, but also correlations between a set of n symbols at arbitrary positions and another m symbols at arbitrary positions. A special case of this generalization are correlations between n-grams and m-grams separated by k symbols and, for $n = m = 1$, our mutual information prototype just defined above.

In our second step, we will give up our restriction to one sequence. We will also consider correlations between symbols or sub-words appearing in different sequences. The combination of both step one and step two will eventually provide us with a measure of all *many-point crosscorrelations*.

In step three, we will introduce the *Kullback entropy* [Kullback, 1951b] and show that the mutual information is a special case of this measure.

3.6 Higher Order Mutual Information

The mutual information $I(k)$ quantifies what we will (on average) learn about a hidden symbol if somebody tells us which symbol is located k positions upstream in the sequence. Let us now generalize this mutual information by not only considering pairs of symbols k positions apart from each other, but, for example, a symbol pair at position τ and a single symbol at position $\tau + k$ in our sequence of investigation.

The answer on the question "Which amount of information about the symbol at position $\tau + k$ do we obtain by getting to know the symbol pair at position τ?" is then exactly defined as our generalized mutual information. It can again be written as the average logarithm of the terms $\frac{P_{ij}(k)}{p_i \cdot q_j}$ where p_i is now the probability to find the i-th 2-gram, q_j is the probability to find the j-th symbol, and $P_{ij}(k)$ is the joint probability to find the i-th 2-gram and the j-th symbol k positions downstream.

In mathematical terms,

$$I(k) = \sum_{i=1}^{\lambda^2} \sum_{j=1}^{\lambda} P_{ij}(k) \cdot \log \left(\frac{P_{ij}(k)}{p_i \cdot q_j} \right) . \tag{3.42}$$

Please note that the index i is now counting all 2-grams and thus running from 1 to λ^2.

All following generalizations are straightforward and thus only briefly displayed.

Instead of considering correlations between a 2-gram and a remote symbol, we can as well analyze the mutual information between an n-gram and a symbol k positions downstream in our stationary and ergodic sequence. Then i runs from 1 to λ^n and p_i denotes the probability of the cylinder given by the i-th n-gram. The mathematical expression is analogous to eqs. (3.39) and (3.42) and reveals the following interpretation: if the logarithm is taken to base 2, the mutual information quantifies the information gain about a randomly chosen symbol by getting to know the n-gram k positions upstream.

We will, in the following, also stop to restrict the target of our prediction be a single symbol. Instead we will present a quantity that measures correlations between an n-gram and a 2-gram separated by k symbols. In this case, we denote the 2-gram probabilities by q_j where $j = 1, 2, ..., \lambda^2$ and again apply the same definition. Then the mutual information measures the average amount of information we gain about a 2-gram by getting to know the n-gram k positions upstream.

Correlations between n-grams and m-grams that are separated by k symbols can be analyzed by an analogously defined mutual information that quantifies the information content we obtain about an m-gram if somebody tells us which n-gram is located k positions upstream. Please note that n and m can be any positive integer number, i.e., we do not require $n \geq m$.

The last step will be that we also allow our n-tuple and m-tuple to contain disconnected symbols. The ultimate question that we are now asking is which information do we obtain about m symbols at m arbitrary positions by getting to know another n symbols at another

n arbitrary positions? The answer is given by the mutual information, which is defined as

$$I(\vec{k}) \equiv \sum_{i=1}^{\lambda^m} \sum_{j=1}^{\lambda^n} P_{ij}(\vec{k}) \cdot \log \left(\frac{P_{ij}(\vec{k})}{p_i \cdot q_j} \right) . \tag{3.43}$$

Please note that this mutual information function now depends an a vector $\vec{k}$ that contains the distances between all of the considered $m + n$ symbols.

This generalization will turn out to be extremely fruitful for the analysis of DNA sequences, since there are a lot of correlations of this type known to molecular biologists, which are, however, not yet systematically exploited. As an illustrating example, we might choose the problem to identify promoters in eukaryotic genomes. We already know from molecular biology that a $TATA$ box at position $k_1 = -30$ and a $CAAT$ box at position $k_2 = -75$ are necessary for building up a functioning promoter, if position $k_0 = 1$ marks the transcription start.

The problem by which computational molecular biologists are now confronted with is that these features are by far not sufficient to identify real promoters in eukaryotic DNA. If we assume for the sake of simplicity that all nucleotides appear with the same probability, we come up with the bothering result that the mean distance between two $TATA$ boxes is 256 base pairs on average.

On the other hand, we know the typical lengths of genes and their density in the genome, i.e., we know about the typical lengths of intergenic sequences. The human genome, which consists of about three billion base pairs, is expected to contain about a hundred thousand genes. This means that the mean distance between two promoters is in the order of thirty thousand base pairs.

Hence, a search for all $TATA$ elements will certainly provide us with all promoters, but 99% of them will be $TATA$-simulants and not real promoters. Therefore, many biologists are searching for other features or patterns that distinguish real promoters from $TATA$-simulants [Mural 1994]. A welcome tool for these kinds of studies could be the generalized mutual information. This measure might reveal whether there are additional positions that are highly correlated with appearing $CAAT$ and $TATA$ boxes. These correlations could then be exploited for identifying real promoters with a higher reliability.

3.7 Mutual Information Crosscorrelations

In this section, we will introduce a second generalization of the mutual information defined in section 3.5. We will give up our restriction to one sequence in which we look for correlations. The consideration of two sequences and correlations between them will eventually bring us back to the original questions risen in information theory.

Let us remember the five elementary steps by which a message is transmitted. We are now neglecting the coding and decoding process and only concentrate on the question what happens with a given symbol sequence that is sent through a noisy channel.

Let the original sequence, which we assume to be stationary and ergodic for the sake of simplicity, be denoted by $X := \ldots x_{-2}\, x_{-1}\, x_0\, x_1\, x_2 \ldots$, where the letters x_i are chosen from the alphabet A of size λ. Let then the received sequence be $Y := \ldots y_{-2}\, y_{-1}\, y_0\, y_1\, y_2 \ldots$, where we assume the letters y_i be chosen from the same alphabet A.

If the channel is noiseless, we obtain $X \equiv Y$, i.e., $x_i = y_i$ for all integer i.

However, if the channel is noisy, mutations may happen that randomly change some of the transmitted symbols.

The question that we are now asking is: "How much information do we gain about the sent symbol x_n (that we, as the receiver, do not know) if we receive the symbol y_n?"

The answer is given by the mutual information, which is defined as

$$I(k) = \sum_{i=1}^{\lambda} \sum_{j=1}^{\lambda} P_{ij} \cdot \log \left(\frac{P_{ij}}{p_i \cdot q_j} \right), \tag{3.44}$$

where p_i is the probability that the symbol $A_i \in A$ is sent, q_j is the probability that the symbol $A_j \in A$ is received, and P_{ij} is the joint probability that the symbol $A_i \in A$ is sent and the symbol $A_j \in A$ is received. If the logarithm is taken to base 2, we obtain the mutual information measured in bit.

The motivation and explanation of this definition is analogous to section 3.5 and thus only sketched here.

Let H_X be the Shannon entropy of the emitted sequence[2] X, H_Y be the Shannon entropy of the received sequence Y, and H_2 the Shannon entropy reflecting the average

[2]Please remember that this definition is only correct for stationary and ergodic sources, where we can identify the statistical properties of an infinitely long sample sequence with the statistical properties of the source. In all other cases, the definition of H_X, H_Y, and H_2 would also be possible, but then we would have to spend more words to present a proper definition of the Shannon entropies H_X, H_Y, and H_2 of the information sources X, Y, and (X,Y).

information stored in the symbol pairs (x_i, y_i).

The difference between the mean uncertainty about the symbol pair (x_i, y_i) and the mean uncertainty about the symbol y_i is again identical to the mean uncertainty about the (sent) symbol x_i provided we know the (received) symbol y_i. Therefore, let us again define the conditional entropy

$$h_{X|Y} \equiv H_2 - H_Y \tag{3.45}$$

in analogy to the definition of h_2 in section 3.4.

Of course, we can also define a conditional entropy $h_{Y|X} \equiv H_2 - H_X$, which quantifies the mean uncertainty about y_i provided we know x_i. However, please note that $h_{X|Y}$ is not necessarily equal to $h_{Y|X}$, i.e., $h_{X|Y}$ is not necessarily symmetric in X and Y.

In order to answer the question what amount of information we gather about the sent symbol x_i by receiving the symbol y_i, we just subtract $h_{X|Y}$ from H_X and thus obtain

$$I[X,Y] = H_X - h_{X|Y} = H_X + H_Y - H_2 = \sum_{i,j=1}^{\lambda} P_{ij} \cdot \log\left(\frac{P_{ij}}{p_i \cdot q_j}\right). \tag{3.46}$$

Even though we will, in the remainder of this work, only deal with analyzing correlations between symbols or sets of symbols within one sequence, this two-sequence-generalization will turn out to be essential for correctly calculating the mutual information function of finite sequences.

If we are given an only finite sequence, we cannot determine the corresponding mutual information with absolute accuracy, but only estimate the mutual information values with a certain reliability. Due to finite size effects, the estimates of p_i and q_i (these are the probabilities to find the symbol A_i at the left hand side and at the right hand side of our symbol pair, respectively) do not have to be identical. Therefore, a careful distinction between the two probabilities p_i and q_i is required even in the case of dealing with only one sequence.

Moreover, the two-sequence-generalization will be crucial for understanding statistical and systematic errors by estimating correlation functions from finite samples. We will show in chapter 11 that a careful analysis of the relation between the probabilities p_i and q_i will eventually result in an exact expression of the systematic errors induced by the non-vanishing bias of the natural correlation function estimator.

At the end of this section, we will briefly focus on some possible (and recommendable) generalizations of the mutual information between two sequences.

In the first instance, we give up our restriction that the letters x_i and y_i have to be chosen from the same alphabet. This means that i now runs from 1 to λ_X and j from 1 to λ_Y, where λ_X denotes the size of the alphabet A, from which the symbols x_τ are chosen, and λ_Y denotes the size of the alphabet B, from which the symbols y_τ are chosen.

In the second instance, we apply all generalizations discussed in our previous section. This means that we are now asking how much information we gain about n symbols picked at arbitrary positions in the sent sequence X by getting to see m symbols of the received sequence Y. Let us eventually present our final definition of the mutual information.

Definition 3.1 *Let p_i denote the probability of the cylinder A_i of a given information source X, where $i = 1, 2, ..., M_X$ and $M_X = \lambda_X^n$ in case of our example presented above. Let analogously q_j denote the probability of the cylinder B_j of a source Y, where $j = 1, 2, ..., M_Y$ and $M_Y = \lambda_Y^m$ for example. Let further P_{ij} be the joint probability for the simultaneous occurrence of the two cylinders A_i and B_j.*

Then we define the **mutual information** *as*

$$I[X,Y] = \sum_{i=1}^{M_X} \sum_{j=1}^{M_Y} P_{ij} \cdot \log \left(\frac{P_{ij}}{p_i \cdot q_j} \right). \tag{3.47}$$

Before we will turn to our third generalization of the mutual information by defining the Kullback entropy, let us state an interesting theorem about the mutual information $I[X,Y]$ as a function of the two information sources X and Y, which deals with the relation between $I[X,Y]$ and $I[Y,X]$.

Imagine we substitute the role between the sender and the receiver, i.e., we do not ask what the receiver learns about the sent sequence X by receiving Y, but vice versa what the sender can infer about the received sequence Y by analyzing the channel and the sent sequence X.

It is formally trivial to show that

$$I[X,Y] = I[Y,X] \tag{3.48}$$

independently on the chosen alphabets A and B.

This, however, means that, for example, the amount of information we gain about a received symbol from a binary alphabet B by getting to know a sent 5-gram composed from a ternary alphabet A is equal to the amount of information we obtain about this 5-gram from X by getting to know the binary symbol from Y. This general statement of

the symmetry of the mutual information as a function of the two sources X and Y just reflects the *mutuality* of this information measure.

3.8 Kullback Entropy and Mutual Information

In this section, we will define the Kullback entropy $K\left(\vec{p}, \vec{p^0}\right)$ and show that the mutual information $I(\vec{p})$ is equal to the Kullback entropy in the special case in which we set $\vec{p^0}$ equal to the product of the marginal distributions of $\vec{p}$.

We will present some properties of the Kullback entropy that will turn out to be extremely valuable for the application of the mutual information function as a correlation measure to symbol sequences such as texts, pieces of music, time series, DNA, RNA, or amino acid sequences.

We will, in particular, show in theorem 3.1 that the Kullback entropy is always positive and only vanishes if $\vec{p} \equiv \vec{p^0}$, which implies that the mutual information vanishes if, and only if, all appearing symbols in our sequence are statistically independent.

An extended introduction of Kullback's entropy and detailed discussions of the theorems presented in this section can be found in [Jaglom 1965], [Khinchin 1957a], [Kullback 1951b], [McEliece 1977], [Shannon 1948a], [Völz 1982 & 1983], and [Yockey 1992].

We introduced the mutual information as that amount of information which we gain about a symbol or tuple x_i by receiving a message y_i. Let us now consider the general case that we have some prior assumption about a system X, which we represent by a vector $\vec{p^0} \equiv (p_1^0, p_2^0, ..., p_M^0)$ containing the prior probabilities p_i^0 of the states of our considered system.

Let us then carry out some experiments that provide us with new information about this system. Let our posterior knowledge about the states of the system be reflected by $\vec{p} \equiv (p_1, p_2, ..., p_M)$, which denotes the vector containing the posterior probabilities p_i.

Then the information gain $K\left(\vec{p}, \vec{p^0}\right)$ is given by

$$K\left(\vec{p}, \vec{p^0}\right) \equiv C \sum_{i=1}^{M} p_i \cdot \ln\left(\frac{p_i}{p_i^0}\right) \tag{3.49}$$

and termed the *Kullback entropy*.

Kullback introduced this quantity as a *divergence measure*, since it quantifies the compatibility of the outcomes of a sampling experiment with the prior hypothesis about the considered system.

Let us now illustrate the relationship between the Kullback entropy and the mutual information as defined in the previous chapter. Imagine we are setting up some sampling experiments by which we want to decide whether or not two random variables are statistically independent. Let the two random variables be X and Y with their possible realizations $x_1, x_2, ..., x_{M_X}$ and $y_1, y_2, ..., y_{M_Y}$, respectively.

Let now our prior assumption be the statistical independence of X and Y, i.e., $P_{ij}^0 = p_i \cdot q_j$, where p_i denotes the probability to observe x_i and q_j denotes the probability of the experimental outcome y_j.

If we denote the experimental joint probabilities by P_{ij}, we come up with the special Kullback entropy

$$K\left(\hat{P}, \hat{P}^0 = \vec{p} \cdot \vec{q}\right) = \sum_{i=1}^{M_X} \sum_{j=1}^{M_Y} P_{ij} \cdot \log\left(\frac{P_{ij}}{p_i \cdot q_j}\right), \tag{3.50}$$

where the base of the logarithm is purposely left unspecified, since it is related to the units in which we wish to measure the amount of information.

As we easily recognize, this Kullback entropy is identical to our definition 3.1 of the mutual information.

By definition, this quantity $K\left(\hat{P}, \hat{P}^0 = \vec{p} \cdot \vec{q}\right)$ measures the information we gain about the P_{ij} under the assumption of independent X and Y. If this information is zero or close to zero, this means that our assumption $\hat{P}_{ij}^0$, i.e., the assumption of the statistical independence between X and Y, is perfect or very good, respectively. If, on the other hand, the mutual information is high, this means that our independence assumption is very bad, i.e., X and Y are statistically dependent.

Let us now present a lemma and a theorem that underscore the value of the mutual information $I[X, Y]$ as a measure of correlations between X and Y.

Lemma 3.1 *The Kullback entropy $K\left(\vec{p}, \vec{p^0}\right)$ is a convex function of the probabilities p_i, i.e.,*

$$\frac{\partial^2 K}{\partial p_i \partial p_j} = \frac{\delta_{ij}}{p_i} \geq 0 \tag{3.51}$$

for all $i, j = 1, 2, ..., M$, all $\vec{p}$, and all $\vec{p^0}$.

Since also $f(x, x_0) = -\ln(x/x_0)$ is a convex function in x, i.e., $\frac{\partial^2 f}{\partial x^2} = \frac{1}{x^2} \geq 0$, we can apply Jensen's inequality (see Appendix B) to the definition of the Kullback entropy and

obtain

$$-K\left(\vec{p},\vec{p^0}\right) = \sum_{i=1}^{M} p_i \cdot \log\left(\frac{p_i^0}{p_i}\right) \tag{3.52}$$

$$\leq \log\left(\sum_{i=1}^{M} p_i^0\right) \tag{3.53}$$

$$= \log(1) = 0, \tag{3.54}$$

which yields our announced theorem.

Theorem 3.1 *The Kullback entropy is always positive, i.e.,*

$$K\left(\vec{p},\vec{p^0}\right) \geq 0 \tag{3.55}$$

for all $\vec{p}$ and $\vec{p^0}$, and only vanishes if equality holds in Jensen's inequality, i.e., the Kullback entropy vanishes if, and only if, $\vec{p} \equiv \vec{p^0}$.

This theorem implies the important property of the mutual information $I[X,Y]$ to be always positive and only equal to zero, if X and Y are statistically independent. For exactly this reason, we consider the mutual information a more powerful measure of correlations than, for example, the correlation function $C[X,Y]$ defined in chapter 11.

At this point we turn back to analyzing symbol sequences and present a theorem about the mutual information function $I(k)$ of Markov chains.

Theorem 3.2 *Let the random variables X, Y, and Z form a (first order) Markov chain, i.e., Z depends on X only through Y, i.e., $p(z|y) = p(z|x,y)$. Then we obtain that*

$$I[X,Z] \leq I[X,Y] \tag{3.56}$$

as well as

$$I[X,Z] \leq I[Y,Z] \tag{3.57}$$

for all X, Y, and Z.

This means that the mutual information function $I(k)$ of a first order Markov chain is always monotonicly decreasing. Vice versa, we can state that any deviation from the monotonic decay of the mutual information function indicates the non Markovian character of the underlying sequence.

In information theory, the inequalities (3.56) and (3.57) carry a very important meaning: *discrete memoryless channels* always tend to leak information [McEliece 1977]. If Y is a definite function of X and Z is a definite function of Y, then we can think of the Markov chain (X, Y, Z) as a data processing configuration. Then theorem 3.2 says that data processing can only destroy information.

3.9 Summary

In section 3.1, we analyzed the communication process and introduced the concept of an information source. We showed that the probability measure of a given cylinder can be understood as the probability to find the corresponding substring in stationary and ergodic sequences. We want to emphasize again that we are only in this case allowed to study infinitely long sample sequences instead of sequence ensembles.

In section 3.2, we presented an axiomatic approach to the definition of the Shannon entropy as a measure of uncertainty in single symbols.

Section 3.3 was then devoted to the introduction of higher order entropies, which measure the amount of uncertainty in substrings of, in principle, arbitrary length. We realized that the series of higher order entropies is always monotonicly increasing and convex. Whereas Bernoulli sequences exhibit a linear growth of their higher order entropies, periodic sequences reach a plateau at the length of their period.

In section 3.4, we introduced conditional entropies, which measure the amount of uncertainty in a symbol provided we know some previous ones. We explained why the series of conditional entropies of stationary and ergodic sources always monotonicly decreases and considered a Bernoulli sequence as well as a periodic string as limiting cases of all possible h_n-series of stationary and ergodic sources.

In section 3.5, we motivated the definition of the mutual information as a measure of correlations between two symbols. Since these two symbols are not restricted to neighbor each other, but can instead be separated by a gap of arbitrary length, the mutual information function turned out to be a very powerful tool to detect long-range correlations in symbolic sequences.

Sections 3.6, 3.7, and 3.8 were eventually dedicated to presenting three kinds of generalizations of the mutual information function presented in section 3.5.

In section 3.6, we gave up the restriction of measuring correlations between single

symbols and defined the mutual information between a set of m and a set of n symbols at arbitrary positions. We demonstrated that this generalization yields a powerful tool for linguistic analyses of DNA sequences.

In section 3.7, we presented a generalization of the mutual information as a measure of the shared between two symbols or sets of symbols stored in different sequences. We underscored the importance of this generalization for properly understanding and correctly calculating the systematic estimation errors induced by the always finite length of real sequences.

Section 3.8 was reserved for introducing the Kullback entropy. We showed that the mutual information is a special Kullback entropy and thus gained a new interpretation of the mutual information between two random variables. By relating this interpretation with our previous understanding, we developed a new insight into what the mutual information measures.

In section 3.8, we derived the theorem that the mutual information is always positive and vanishes if, and only if, the considered random variables are statistically independent. This property is the main reason why we prefer the mutual information function over autocorrelation functions for analyzing symbolic texts such as DNA, RNA, or amino acid sequences.

Chapter 4

Measuring Correlations in Symbol Sequences

This chapter is devoted to relations between correlation functions and the mutual information function. We show that in sequences over an alphabet of λ symbols statistical dependences are measured by $(\lambda-1)^2$ independent parameters. However, not all of them can be determined by autocorrelation functions. Appropriate sets of correlation functions (including crosscorrelations) are introduced, which allow the detection of all dependences. The results are exemplified for binary, ternary, and quaternary symbol sequences. As an application, we discuss in section 13 that a nonuniform codon usage in protein-coding DNA sequences introduces periodic correlations even at distances in the order of 1000 base pairs.

4.1 Introduction

The statistical analysis of symbol sequences is of growing interest in various fields, for example linguistics [Shannon 1951, Jaglom & Jaglom 1965, Schenkel 1993, Ebeling & Pöschel 1994, Levitin & Feingold 1994], analysis of biosequences [Gatlin 1972, Ebeling & Feistel 1982, Ebeling et al. 1987, Herzel 1988, Karlin & Brendel 1992, Peng et al. 1992, Borštnik et al. 1993, Herzel et al. 1994a, Herzel et al. 1994, Herzel et al. 1995], cellular automata [Grassberger 1986], or dynamical systems [Farmer 1982, Eckmann & Ruelle 1985, Ebeling & Nicolis 1991]. Under the assumption of stationarity, probabilities can be assigned to the occurrence of symbols in given sequences. These symbols are, for example,

the four nucleotides A, C, G, and T occurring in DNA sequences, the 26 letters of the Latin alphabet in English texts, or the 20 amino acids in protein sequences.

The aim of this chapter is to study certain characteristics of some widely used measures such as correlation functions and mutual information [Herzel & Ebeling 1985, Pompe et al. 1986, Fraser & Swinney 1986, Li 1990]. In particular, we focus on the interrelation between correlation functions and the mutual information function. Our study is motivated by the continuing analysis of long correlations in biosequences [Ebeling et al. 1987, Herzel 1988, Karlin & Brendel 1992, Peng et al. 1992, Borštnik et al. 1993, Herzel et al. 1994a, Herzel et al. 1994, Herzel et al. 1995], but most of our results are of general importance.

Both, the correlation function and the mutual information measure correlations within one sequence (autocorrelations) or between two sequences (crosscorrelations). However, both measures impose their own meaning of what they define as correlation. There is a clear mathematical expression for statistical independence between random variables — the factorization of the corresponding probabilities. The mutual information and correlation coefficients vanish in case of statistical independence, but they differ in quantifying deviations from statistical independence, i.e., in quantifying statistical dependences.

Correlation coefficients measure only linear dependences, whereas the mutual information is more general in the sense that it detects all kinds of statistical dependences. However, we will argue that vanishing correlation functions can ensure statistical independence under certain circumstances as well.

Both, correlation functions and mutual information have their advantages and disadvantages. The following list might help to compare and contrast some features that are relevant in this context:

Advantages of correlation functions:

- A lot of experience how correlation functions behave has been gathered in statistical physics. For example, there are well known relations between the autocorrelation function and the variance growth of a corresponding random walk [Peng et al. 1992, Stanley et al. 1994].

- Fourier transforms of autocorrelation functions give power spectra, which allow an easy detection of periodicities.

- Correlation coefficients can be either positive or negative. Hence, they can distinguish between correlations and anticorrelations.
- Correlation coefficients are very specific. They can, for example, measure correlations between the nucleotides A and C in DNA sequences or between hydrophilicity and charge in amino acid sequences.

On the other hand,

- correlation functions measure only linear correlations. This can be illustrated by chaotic sequences for which correlation coefficients may vanish, although the iterates are even functionally dependent. For example, the autocorrelation coefficients vanish for the fully developed logistic map $x_{n+1} = 4\,x_n(1 - x_n)$ [Herzel & Ebeling 1985, Grossmann & Thomae 1977].
- Correlation coefficients are not invariant with respect to coordinate transformations. Hence, the assignment of numbers to symbols is somewhat arbitrary and the output of the autocorrelation function depends on the chosen projection.

Advantages of the mutual information:

- The mutual information detects any kind of dependence (not only linear correlations), which is stated by the following theorem: The mutual information vanishes if, and only if, all occurring symbols are statistically independent. (See, e.g., [Mackey 1989] for a proof.)
- The mutual information is invariant under coordinate transformations.
- No assignment of numbers to symbols is required.

Disadvantages:

- The mutual information is less specific than autocorrelation functions since any deviation from statistical independence is mapped onto a single number.
- For all finite sequences, there is a systematic overestimation of the mutual information, which, however, can be corrected. (See [Herzel et al. 1994a] and Appendix II for reviews of finite sample corrections.)

We have already mentioned that a vanishing mutual information is necessary and sufficient for the statistical independence of all symbols. The arising question is now whether there are sets of correlation functions the vanishing of which guarantees statistical independence as well. Moreover, we derive some functional relations between different correlation functions, on the one hand, and correlation functions and the mutual information function, on the other hand. With this respect, this chapter can be regarded as a generalization of previous results by Wentian Li [Li 1990].

In particular, we concentrate on quaternary sequences, which were not studied in [Li 1990], and apply our results to DNA sequences since they are of immanent interest in biology.

Our main results may be summarized as follows: In order to quantify statistical dependences in sequences with λ symbols, $(\lambda - 1)^2$ independent parameters have to be studied for a given distance between symbols.

However, only $\frac{\lambda \cdot (\lambda - 1)}{2}$ of these parameters can be measured by autocorrelation coefficients. We suggest to use crosscorrelations in order to detect all statistical dependences.

4.2 Definitions

Symbol sequences are composed of letters from an alphabet of λ letters $\{A_1, A_2, ..., A_\lambda\}$ (e.g. $\{A, C, G, T\}$ for DNA sequences or $\{A, B, C, ..., X, Y, Z\}$ for English texts). Since we assume stationarity, any symbol A_i $(i = 1...\lambda)$ occurs with a well defined probability p_i at any arbitrary site, i.e., this probability p_i does not depend on the position at which the symbol appears in the sequence. At this point, we remark that the assumption of stationarity is a delicate question for finite sequences. Natural sequences such as languages, DNA, or proteins are often not completely stationary. So we see, for example, significant fluctuations of the G+C content in DNA sequences even on length scales of several million base pairs. For instance, chromosome bands indicate regions of different G+C content [Herzel et al. 1995, Lewin 1997]. However, these problems of instationarity can be reduced by trend elimination techniques [Stanley et al. 1994, Li et al. 1994]. The effects of a nonuniform codon usage in the reading frame on instationarity is discussed in section VII.

Now we define joint probabilities $p_{ij}(k)$ to find the symbol A_i and the symbol A_j (in this order) in a distance k. So $p_{ij}(1)$ refers to adjacent symbols A_i and A_j. Please note

that, due to the assumed stationarity, the joint probabilities $p_{ij}(k)$ do not depend on the positions of A_i and A_j, but only on their distance within the sequence.

Two symbols in a distance k are defined to be statistically independent if, and only if, $p_{ij}(k) = p_i \cdot q_j$ where p_i denotes the probability to find the symbol A_i at an arbitrary position n, and q_j denotes the probability to find the symbol A_j at position $n+k$. Due to stationarity, all p_j are equal to q_j, so that we can call two symbols statistically independent if (and only if) $p_{ij}(k) = p_i \cdot p_j$.

At this point we define a quantity that measures the statistical dependence (and thus correlations) between symbols in a distance k:

$$I(k) \equiv \sum_{i,j=1}^{\lambda} p_{ij}(k) \cdot \log \frac{p_{ij}(k)}{p_i \cdot p_j}. \tag{4.1}$$

This quantity is termed *mutual information* (or transinformation [Herzel & Ebeling 1985]) and is related to the Kullback information or Boltzmann's H-functional [Mackey 1989]. As base of the logarithm we choose 2 and obtain the following interpretation for $I(k)$. The mutual information gives us the information (measured in bit) that we receive about the second symbol by getting to know the first one.

Now we turn to the definition of autocorrelation functions of symbol sequences. Here, we have to assign real numbers $(a_1, a_2, ..., a_\lambda)$ to symbols $A_1, A_2, ..., A_\lambda$. For further convenience, we define the vector $\vec{a}$ as the λ-tuple $\vec{a} \equiv (a_1, a_2, ..., a_\lambda)$. Then we define

$$C_{\vec{a}}(k) \equiv \langle a(n) \cdot a(n+k) \rangle - \langle a(n) \rangle \cdot \langle a(n+k) \rangle \tag{4.2}$$

as the *autocorrelation function* with the chosen projection $\vec{a}$ where $a(n)$ denotes the number assigned to the symbol at position n in the sequence and $\langle ... \rangle$ refers to the average over the entire sequence.

Assuming ergodicity, the above average can be replaced by the average over probabilities $p_{ij}(k)$ to define the autocorrelation function:

$$\begin{aligned} C_{\vec{a}}(k) &= (\sum_{i,j=1}^{\lambda} p_{ij}(k) \cdot a_i \cdot a_j) - (\sum_{i=1}^{\lambda} p_i \cdot a_i)^2 \\ &= \sum_{i,j=1}^{\lambda} (p_{ij}(k) - p_i \cdot p_j) \cdot a_i \cdot a_j. \end{aligned} \tag{4.3}$$

If we now define a quadratic $\lambda \times \lambda$ matrix $\hat{D}(k)$ by setting the entries $D_{ij}(k) \equiv p_{ij}(k) - p_i \cdot p_j$, we can write any autocorrelation function as a bilinear form of this matrix $\hat{D}(k)$ in the following way:

$$C_{\vec{a}}(k) = \vec{a} \cdot \hat{D}(k) \cdot \vec{a}^T. \tag{4.4}$$

If, and only if, all symbols in our sequence are statistically independent, this matrix $\hat{D}(k)$ is identical to $\hat{0}$.

Various relations between different autocorrelation functions are derived in the next sections. For simplicity, the dependence of the joint probabilities $p_{ij}(k)$, the matrix $\hat{D}(k)$, the mutual information $I(k)$, and the correlation functions $C_{\vec{a}}(k)$ on the distance k is dropped in the following. However, relations derived for any of those quantities hold for all $k \in \mathcal{N} \setminus \{0\}$.

4.3 How to guarantee statistical independence

This question was also addressed by Wentian Li [Li 1990] for binary and ternary sequences. However, we do not require the matrix $\hat{P}$ (which is defined by the entries p_{ij}) to be symmetric. Moreover, we discuss statistical properties of sequences that are composed of more than three different letters. Some relations valid for sequences with any alphabet size are derived in Appendix I. Novel results regarding the dependence between autocorrelation functions of ternary and quaternary sequences are derived as special cases in sections IV and V, respectively.

As shown above, any correlation function is a bilinear form of the matrix $\hat{D}$, and $\hat{D} \equiv \hat{0}$ is necessary and sufficient for the statistical independence of all symbols in the sequence. Hence, this matrix $\hat{D}$ contains all information we need to evaluate any statistical dependence within our symbol sequence of interest.

We first raise the following question: How many of the λ^2 entries of the matrix $\hat{D}$ are independent?

For this reason, let us first collect all constraints that might decrease the number of independent entries of $\hat{D}$.

The normalization of the joint probabilities p_{ij}

$$\sum_{i,j=1}^{\lambda} p_{ij} = 1 \tag{4.5}$$

and of the probabilities p_i

$$\sum_{i=1}^{\lambda} p_i = 1 \tag{4.6}$$

yields

$$\sum_{i,j=1}^{\lambda} D_{ij} = \sum_{i,j=1}^{\lambda} P_{ij} - (\sum_{i=1}^{\lambda} p_i) \cdot (\sum_{j=1}^{\lambda} p_j) = 0. \tag{4.7}$$

So the normalization constraint for our *dependence matrix* $\hat{D}$ reads as follows: The sum over all entries of $\hat{D}$ has to vanish.

Furthermore, the equations

$$\sum_{i=1}^{\lambda} p_{ij} = p_j \tag{4.8}$$

and

$$\sum_{j=1}^{\lambda} p_{ij} = p_i \tag{4.9}$$

hold due to stationarity for all $j = 1...\lambda$ and for all $i = 1...\lambda$, respectively. Hence, we obtain for all $j = 1...\lambda$

$$\sum_{i=1}^{\lambda} D_{ij} = \sum_{i=1}^{\lambda} (p_{ij} - p_i \cdot p_j) = 0 \tag{4.10}$$

and for all $i = 1...\lambda$

$$\sum_{j=1}^{\lambda} D_{ij} = \sum_{j=1}^{\lambda} (p_{ij} - p_i \cdot p_j) = 0. \tag{4.11}$$

In other words, this means that the sum of the matrix entries D_{ij} in each column and in each row has to vanish. This gives us another $2 \cdot \lambda$ constraints besides the normalization.

At this point, it is important to realize that these
$2 \cdot \lambda + 1$ constraints are not independent of each other. In the following paragraph, we are showing that two of the constraints mentioned above depend on a set of only $2 \cdot \lambda - 1$ equations.

In a first step, we show that the equations

$$\sum_{i=1}^{\lambda} D_{i\lambda} = 0 \tag{4.12}$$

and

$$\sum_{j=1}^{\lambda} D_{\lambda j} = 0 \tag{4.13}$$

can be derived from the remaining $2 \cdot \lambda - 1$ constraints. This means we have to prove that both, the sum in the last column and the sum in the last row vanish, provided normalization

holds and all sums in the first $\lambda-1$ columns and in the first $\lambda-1$ rows vanish. By using the $\lambda-1$ row-constraints and realizing the normalization condition, we obtain:

$$\sum_{i=1}^{\lambda} D_{i\lambda} = \sum_{i,j=1}^{\lambda} D_{ij} - \sum_{j=1}^{\lambda-1} \underbrace{(\sum_{i=1}^{\lambda} D_{ij})}_{0} = 0. \tag{4.14}$$

Analogously, we proceed and yield:

$$\sum_{j=1}^{\lambda} D_{\lambda j} = \sum_{i,j=1}^{\lambda} D_{ij} - \sum_{i=1}^{\lambda-1} \underbrace{(\sum_{j=1}^{\lambda} D_{ij})}_{0} = 0. \tag{4.15}$$

Now we have to show that the $2\cdot\lambda-1$ constraints, which have just been found to be sufficient to guarantee stationarity and normalization, are indeed independent. By now we know that there are not more than $2\cdot\lambda-1$ independent equations constraining the λ^2 dimensional space of all parameters D_{ij}. Thus, there are $\lambda^2-2\cdot\lambda+1=(\lambda-1)^2$ parameters sufficient for determining the matrix $\hat{D}$.

In order to prove that these parameters are also necessary, let us fill the matrix $\hat{D}$ starting with the first line. We define the entries $D_{11}, D_{12}, ..., D_{1\lambda-1}$. At this point, $D_{1\lambda}$ is also defined by the first row-sum-constraint. We proceed with the second row and define $D_{21}, D_{22}, ..., D_{2\lambda-1}$. Again, $D_{2\lambda}$ is then determined by the second row-sum-constraint. We repeat this procedure until we arrive at the element $D_{\lambda-1\lambda-1}$. By now, we have filled the first $\lambda-1$ rows completely by having exploited all of the $\lambda-1$ row-sum-constraints.

We can determine the matrix element $D_{\lambda 1}$ by applying the first column-sum-constraint, $D_{\lambda 2}$ by applying the second one, ..., and $D_{\lambda\lambda-1}$ by using the $(\lambda-1)$th column-sum-constraint. We have just determined all entries D_{ij} except $D_{\lambda\lambda}$ by having defined $(\lambda-1)^2$ parameters and having used $2\cdot(\lambda-1)$ constraints. The last parameter $D_{\lambda\lambda}$ can now be determined by using the remaining normalization constraint. The whole procedure, which is exemplified for ternary sequences in section V, is illustrated by the following sketch:

$$\hat{D} = \begin{pmatrix} D_{11} & \cdots & D_{1\lambda-1} & * \\ \vdots & \ddots & \vdots & \vdots \\ D_{\lambda-11} & \cdots & D_{\lambda-1\lambda-1} & * \\ * & \cdots & * & * \end{pmatrix} \tag{4.16}$$

The matrix elements marked by stars can be determined by the upper left quadratic matrix of type $(\lambda-1)\times(\lambda-1)$ together with the constraints. At this point, we can answer

the question how many parameters are required to determine the matrix $\hat{D}$ unambiguously. It is obvious that exactly $(\lambda-1)^2$ entries (all elements of the upper left quadratic matrix) can be chosen independently. So we have learned that $(\lambda-1)^2$ parameters are necessary and sufficient to define the matrix $\hat{D}$.

Consequently, we concentrate on the $(\lambda-1)\times(\lambda-1)$ matrix composed of those $(\lambda-1)^2$ independent parameters in the upper left corner, which we denote by $\tilde{D}$. The remaining $2\cdot\lambda-1$ entries of the matrix $\hat{D}$ can then be calculated by all of the $2\cdot\lambda-1$ constraints discussed above.

Now we turn to another relevant question. How many of those $(\lambda-1)^2$ parameters (which are required to identify all correlations in a sequence built up of an alphabet containing λ letters) can be determined by autocorrelation functions?

To evoke the reader's interest in this question, let us call that there is an infinite number of autocorrelation functions since we can assign real numbers to all λ symbols, but there are only $(\lambda-1)^2$ parameters to be determined.

We remember that every matrix can unambiguously be decomposed into a sum of a symmetric and an antisymmetric matrix in the following way. If $\hat{D}$ is the matrix to be decomposed, and $\hat{S}$ and $\hat{A}$ denote its symmetric and antisymmetric parts, then

$$S_{ij}=\frac{D_{ij}+D_{ji}}{2} \tag{4.17}$$

and

$$A_{ij}=\frac{D_{ij}-D_{ji}}{2} \tag{4.18}$$

for all $i,j=1...\lambda$.

It follows immediately from eqs. (10), (11), and (18) that the sum over all entries, row-sums, and column-sums of the antisymmetric matrix $\hat{A}$ vanish. Consequently, our constraints for $\hat{D}$ apply to $\hat{S}$ as well.

It can easily be seen that autocorrelation functions are determined solely by the symmetric part:

$$\begin{aligned} C_{\vec{a}} &= \sum_{i=1}^{\lambda} D_{ii}\cdot a_i^2+\sum_{\substack{i,j=1\\ i<j}}^{\lambda}(D_{ij}+D_{ji})\cdot a_i\cdot a_j \\ &= \vec{a}\cdot\hat{S}\cdot\vec{a}^{\,T}. \end{aligned} \tag{4.19}$$

Hence, autocorrelation functions cannot reveal any information about antisymmetric components of $\hat{D}$. We stress that also autocorrelations of moments of the numbers a_i depend only on the symmetric ingredients of the matrix $\hat{D}$.

The symmetric matrix corresponding to the matrix $\tilde{D}$ (remember that only the upper left matrix has to be considered) contains $\frac{\lambda\cdot(\lambda-1)}{2}$ independent parameters, S_{ij} $(i,j = 1...\lambda-1; i \leq j)$, which is the number of elements in the upper triangle matrix including the diagonal elements. On the other hand, the corresponding antisymmetric matrix contains $\frac{(\lambda-1)\cdot(\lambda-2)}{2}$ independent parameters, A_{ij} $(i,j = 1...\lambda-1; i < j)$ (upper triangle matrix excluding the diagonal elements).

Summarizing, we state that $(\lambda-1)^2$ independent parameters are necessary and sufficient for the estimation of all correlations in a given sequence, $\frac{(\lambda-1)\cdot(\lambda-2)}{2}$ of which cannot be determined by any autocorrelation function. The question that we are going to answer in the following sections is how all of the remaining $\frac{\lambda\cdot(\lambda-1)}{2}$ parameters can be calculated by autocorrelation functions.

4.4 Binary Sequences

In this section, we want to apply our results to binary sequences and compare them with studies by Wentian Li [Li 1990]. In binary stationary sequences, the dependence matrix $\hat{D}$ has the form:

$$\hat{D} = \begin{pmatrix} p_{11} - p_1^2 & p_{12} - p_1 \cdot p_2 \\ p_{21} - p_1 \cdot p_2 & p_{22} - p_2^2 \end{pmatrix} \tag{4.20}$$

Due to the constraints of vanishing row-sums and column-sums, we get

$$\hat{D} = \begin{pmatrix} D_{11} & -D_{11} \\ -D_{11} & D_{11} \end{pmatrix} \tag{4.21}$$

where we have introduced the notation $D_{11} = p_{11} - p_1^2$. Since $p_{12} = p_1 \cdot p_2 - D_{11}$ and $p_{21} = p_1 \cdot p_2 - D_{11}$, we realize that Li's requirement, $p_{12} = p_{21}$, is a simple consequence of the constraints resulting from stationarity.

As shown in section II, any autocorrelation function can be expressed by the following bilinear form:

$$\begin{aligned} C_{(a_1,a_2)} &= (a_1, a_2) \cdot \begin{pmatrix} D_{11} & -D_{11} \\ -D_{11} & D_{11} \end{pmatrix} \cdot \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \\ &= (a_1 - a_2)^2 \cdot D_{11} \end{aligned} \tag{4.22}$$

Now we consider the autocorrelation function where we have assigned the numbers 1 and 0 to the symbols A_1 and A_2, respectively, i.e, $(a_1, a_2) = (1, 0)$. Then, the corresponding autocorrelation function is identical to D_{11}:

$$C_{(1,0)} = D_{11}. \tag{4.23}$$

Thus, we can rewrite eq. (22) as follows:

$$C_{(a_1,a_2)} = (a_1 - a_2)^2 \cdot C_{(1,0)}. \tag{4.24}$$

This equation reveals that all autocorrelation functions are dependent on only one (e.g., on $C_{(1,0)}$) which can be considered a basis of the 1-dimensional space spanned by all autocorrelation functions of binary sequences.

We see that $C_{(1,0)} = 0$ implies $\hat{D} \equiv \hat{0}$ and thus statistical independence. This means that $C_{(1,0)}$ (or any other autocorrelation function of binary sequences) vanishes if, and only if, all symbols are statistically independent.

Thus, we realize that autocorrelation coefficients are as good as the mutual information with respect to detecting correlations in binary sequences.

These results also apply to a problem arisen in communication theory [Bernasconi 1987]: Extensive optimization studies have been performed [Krauth & Mezard 1995] to find *low autocorrelation binary sequences* by minimizing $C_{(1,-1)}(k)$. It turns out that a vanishing autocorrelation function

$$C_{(1,-1)} = 4 \cdot C_{(1,0)} = 4 \cdot D_{11} \tag{4.25}$$

implies also statistical independence of letters in the corresponding binary strings.

4.5 Ternary Sequences

Sequences based on an alphabet of three letters are widely used to encode natural languages [Ebeling & Pöschel 1994] or music [Ebeling & Nicolis 1992, Ebeling et al. 1995]. The well known Morse Code can also be regarded as being composed of three symbols: $\{short, long, pause\}$. In this section, we show that

- not all correlations between symbols can be detected by autocorrelation functions

and that

- the three autocorrelation functions $\{C_{(1,0,0)}, C_{(0,1,0)}, C_{(0,0,1)}\}$ are a basis in the sense that all possible autocorrelation functions of ternary sequences are a linear combination of those "basic" ones.

From section III we know that there are $(\lambda-1)^2 = 4$ independent entries of the matrix $\hat{D}$, $\frac{\lambda\cdot(\lambda-1)}{2} = 3$ of which belong to the symmetric matrix $\hat{S}$. We show in this section that these three parameters can be determined by three autocorrelation functions, for example, by $\{C_{(1,0,0)}, C_{(0,1,0)}, C_{(0,0,1)}\}$. However, the antisymmetric parameter remains hidden and cannot be determined by any autocorrelation function.

For illustration, let us consider the following first order Markov process: If the current symbol is A_1, the following one is A_1 with probability 1/3 and A_2 with probability 2/3; A_3 cannot occur right after A_1. Analogously, the transition probabilities from A_2 to A_2 and A_3 to A_3 are 1/3 and those from A_2 to A_3 and A_3 to A_1 are 2/3. Due to symmetry, we obtain $p_1 = p_2 = p_3 = 1/3$. This Markov process can be represented by the following matrix:

$$\hat{P}(k=1) = \begin{pmatrix} 1/9 & 2/9 & 0 \\ 0 & 1/9 & 2/9 \\ 2/9 & 0 & 1/9 \end{pmatrix} \tag{4.26}$$

containing the joint probabilities defined in section II. This process introduces some periodicity $A_1 \to A_2 \to A_3 \to A_1$ leading to non-vanishing antisymmetric components, which is not atypical for DNA sequences. (Cf. the nonuniform codon usage discussed in [Ebeling et al. 1987, Fickett 1982, Staden 1984] and in section VII.) All information regarding correlations in this Markov chain is stored in the matrix

$$\hat{D}(k=1) = \begin{pmatrix} 0 & 1/9 & -1/9 \\ -1/9 & 0 & 1/9 \\ 1/9 & -1/9 & 0 \end{pmatrix} \tag{4.27}$$

which obviously contains a non-vanishing antisymmetric part.

Since

$$C_{(1,0,0)} = D_{11}, \tag{4.28}$$

$$C_{(0,1,0)} = D_{22}, \tag{4.29}$$

and

$$C_{(0,0,1)} = D_{33}, \tag{4.30}$$

we realize that these three autocorrelation functions, which were introduced by Voss, are all equal to zero in this example and thus fail to measure any correlation.

In the following, we are going to show that the three autocorrelation functions $\{C_{(1,0,0)}, C_{(0,1,0)}, C_{(1,1,0)}\}$ are sufficient to determine the three independent parameters S_{11}, S_{12}, and S_{22} required to identify the symmetric matrix $\hat{S}$. By definition, we have

$$C_{(1,0,0)} = S_{11}, \tag{4.31}$$

$$C_{(0,1,0)} = S_{22}, \tag{4.32}$$

and

$$C_{(1,1,0)} = S_{11} + S_{12} + S_{21} + S_{22}. \tag{4.33}$$

Eqs.(31) and (32) together with

$$S_{12} = S_{21} = \frac{C_{(1,1,0)} + C_{(1,0,0)} + C_{(0,1,0)}}{2} \tag{4.34}$$

give the entries of $\tilde{S}$ (the upper left 2×2 matrix of $\hat{S}$) determined by $\{C_{(1,0,0)}, C_{(0,1,0)}, C_{(0,0,1)}\}$: Due to the constraints, $\hat{S}$ is completely determined by $\{S_{11}, S_{12}, S_{22}\}$:

$$S_{13} = S_{31} = -S_{11} - S_{12} \tag{4.35}$$

$$S_{23} = S_{32} = -S_{21} - S_{22} \tag{4.36}$$

$$S_{33} = S_{11} + S_{12} + S_{21} + S_{22}. \tag{4.37}$$

We note in passing that eqs. (33) and (37) reveal the identity

$$C_{(0,0,1)} = C_{(1,1,0)}. \tag{4.38}$$

So we can summarize that the three autocorrelation functions $\{C_{(1,0,0)}, C_{(0,1,0)}, C_{(0,0,1)}\}$ are sufficient to identify the matrix $\hat{S}$ and constitute a basis for all autocorrelation functions of ternary sequences, since

$$C_{\vec{a}} = \vec{a} \cdot \hat{S} \cdot \vec{a}^T \tag{4.39}$$

as pointed out in section III.

However, we have to remember that autocorrelation functions cannot determine all $(\lambda - 1)^2 = 4$ parameters of the matrix $\hat{D}$. This means that the statement "there are no correlations between symbols if all possible autocorrelation functions vanish" is not valid for

ternary sequences. Hence, there are sequences for which all autocorrelation functions are identical to zero even though correlations do exist (as shown in our introductory example). In these situations, the mutual information function could reveal those correlations not detectable by autocorrelation functions.

Another concept to display the antisymmetric correlations is to determine the $\frac{(\lambda-1)\cdot(\lambda-2)}{2}$ independent entries of the antisymmetric matrix $\hat{A}$ by calculating crosscorrelation functions in the following manner. We introduce the *crosscorrelation function*

$$\begin{aligned} C_{\vec{a},\vec{b}}(k) &= \langle a(n)\cdot b(n+k)\rangle - \langle a(n)\rangle\cdot\langle b(n+k)\rangle \\ &= \vec{a}\cdot\hat{D}\cdot\vec{b}^{T} \end{aligned} \tag{4.40}$$

by using two different assignments $\vec{a}$ and $\vec{b}$ to the symbols of a given sequence. If we choose the basis $\vec{a}=(1,0,0)$ and $\vec{b}=(0,1,0)$, we obtain

$$C_{\vec{a},\vec{b}} = D_{12}, \tag{4.41}$$

which is (in addition to S_{11}, S_{12}, and S_{22}) sufficient to determine the four independent parameters of the matrix $\hat{D}$:

$$D_{11} = S_{11} \tag{4.42}$$

$$D_{12} = C_{(1,0,0),(0,1,0)} \tag{4.43}$$

$$D_{21} = 2\cdot S_{12} - D_{12} \tag{4.44}$$

$$D_{22} = S_{22}. \tag{4.45}$$

The application of all $2\cdot\lambda-1=5$ constraints gives the remaining entries of $\hat{D}$.

4.6 Quaternary Sequences

A statistical analysis of quaternary sequences is of fundamental importance in biology, since all nucleic acid sequences are composed of four nucleotides with the four bases: Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). All kinds of RNA (m-RNA, t-RNA, r-RNA) are built up of A, C, G, U (Uracil). Measuring and understanding correlations in DNA sequences or sub-sequences like exons, introns, or intergenic regions is one of

the current goals. So it is the particular purpose of this chapter to apply our results to quaternary sequences in order to foster further analyses of DNA sequences.

There are $(\lambda - 1)^2 = 9$ independent parameters in the matrix $\hat{D}$, only $\frac{\lambda \cdot (\lambda - 1)}{2} = 6$ of which can be determined by autocorrelation functions. The arising question is now to find an appropriate basis of six independent autocorrelation functions, the linear combination of which yields any imaginable autocorrelation function of quaternary sequences.

There are seven possible projections of quaternary sequences onto binary ones, which are indeed studied in [Stanley et al. 1994]. For convenience, we choose the DNA alphabet $\{A, C, G, T\}$ instead of $\{A_1, A_2, A_3, A_4\}$ for the quaternary sequence and the alphabet $\{X, Y\}$ for the binary sequence in this section. Then we can visualize the seven projections in the following list:

- $\{A, C\} \Longrightarrow X$ and $\{G, T\} \Longrightarrow Y$:
  "alphabetical AC-GT rule", no biological interpretation known; corresponding to $C_{(1,1,0,0)}$
- $\{A, G\} \Longrightarrow X$ and $\{C, T\} \Longrightarrow Y$:
  "purine-pyrimidine rule"; corresponding to $C_{(1,0,1,0)}$
- $\{A, T\} \Longrightarrow X$ and $\{C, G\} \Longrightarrow Y$:
  "hydrogen bond rule"; corresponding to $C_{(1,0,0,1)}$
- $A \Longrightarrow X$ and $\{C, G, T\} \Longrightarrow Y$:
  "Adenine rule"; corresponding to $C_{(1,0,0,0)}$
- $C \Longrightarrow X$ and $\{A, G, T\} \Longrightarrow Y$:
  "Cytosine rule"; corresponding to $C_{(0,1,0,0)}$
- $G \Longrightarrow X$ and $\{A, C, T\} \Longrightarrow Y$:
  "Guanine rule"; corresponding to $C_{(0,0,1,0)}$
- $T \Longrightarrow X$ and $\{A, C, G\} \Longrightarrow Y$:
  "Thymine rule"; corresponding to $C_{(0,0,0,1)}$

In this way we have obtained seven different binary sequences, the correlation functions of which we want to calculate. Remembering that there is but one autocorrelation function to be determined in binary sequences and choosing $(x, y) = (1, 0)$ for the sake of simplicity,

we receive seven different autocorrelation functions. Since we know that there are only six independent parameters in the symmetric matrix $\hat{S}$, we can already conclude that these *binary autocorrelation functions* cannot be independent.

In the following, we are checking whether six of these seven binary autocorrelation functions are sufficient to determine the matrix $\hat{S}$ entirely. In this case, they would form a basis in the (usual) sense that all quaternary autocorrelation functions could be linearly combined by them.

Now we are showing that the upper six autocorrelation functions are indeed sufficient to identify all six independent parameters of $\hat{S}$. Applying eq. (24) to $C_{(1,0,0,1)}$ yields:

$$C_{(1,0,0,1)} = C_{(0,1,1,0)} \tag{4.46}$$

We can use the same strategy as in section V and write down all six autocorrelation functions in terms of the symmetric elements of $\hat{D}$. In the next step, we obtain the $\frac{\lambda \cdot (\lambda - 1)}{2} = 6$ entries of the symmetric matrix $\tilde{S}$ in terms of the autocorrelation functions $C_{(1,0,0,0)}$, $C_{(0,1,0,0)}$, $C_{(0,0,1,0)}$, $C_{(1,1,0,0)}$, $C_{(1,0,1,0)}$, and $C_{(0,1,1,0)}$:

$$S_{11} = C_{(1,0,0,0)} \tag{4.47}$$

$$S_{12} = \frac{C_{(1,1,0,0)} - C_{(1,0,0,0)} - C_{(0,1,0,0)}}{2} \tag{4.48}$$

$$S_{13} = \frac{C_{(1,0,1,0)} - C_{(1,0,0,0)} - C_{(0,0,1,0)}}{2} \tag{4.49}$$

$$S_{22} = C_{(0,1,0,0)} \tag{4.50}$$

$$S_{23} = \frac{C_{(0,1,1,0)} - C_{(0,1,0,0)} - C_{(0,0,1,0)}}{2} \tag{4.51}$$

$$S_{33} = C_{(0,0,1,0)}. \tag{4.52}$$

The remaining entries of $\hat{S}$ can be calculated by applying the constraints

$$S_{14} = -S_{11} - S_{12} - S_{13} \tag{4.53}$$

$$S_{24} = -S_{21} - S_{22} - S_{23} \tag{4.54}$$

$$S_{34} = -S_{31} - S_{32} - S_{33} \tag{4.55}$$

$$S_{44} = \sum_{i,j=1}^{3} S_{ij} \tag{4.56}$$

and using the symmetry

$$S_{ij} = S_{ji}. \tag{4.57}$$

Obviously, the chosen autocorrelation functions form a basis since they contain the entire information about all correlations that autocorrelation functions could detect in principle. For example, $C_{(0,0,0,1)}$ is the following linear combination of the basis functions used above:

$$\begin{aligned} C_{(0,0,0,1)} &= -C_{(1,0,0,0)} - C_{(0,1,0,0)} - C_{(0,0,1,0)} \\ &\quad +C_{(1,1,0,0)} + C_{(1,0,1,0)} + C_{(0,1,1,0)}. \end{aligned} \tag{4.58}$$

After replacing $C_{(0,1,1,0)}$ by $C_{(1,0,0,1)}$, we obtain

$$\begin{aligned} C_{(1,0,0,0)} &+ C_{(0,1,0,0)} + C_{(0,0,1,0)} + C_{(0,0,0,1)} = \\ C_{(1,1,0,0)} &+ C_{(1,0,1,0)} + C_{(1,0,0,1)}. \end{aligned} \tag{4.59}$$

This relation allows to select an appropriate basis of six autocorrelation functions out of those seven. We suggest to choose $\{C_{(1,0,0,0)}, C_{(0,1,0,0)}, C_{(0,0,1,0)}, C_{(0,0,0,1)}, C_{(1,0,1,0)}, C_{(1,0,0,1)}\}$ since all of the autocorrelation functions in this set are biologically interpretable. The "AC-GT rule" as well as others such as the molecular mass rule [Stanley et al. 1994] are just a linear combination of those autocorrelation functions mentioned above.

A general way to find a simple basis for any alphabet size λ is presented in Appendix I.

After we have extensively discussed the dependence between quaternary autocorrelation functions, we are now coming back to our original task to study how all of the $(\lambda - 1)^2 = 9$ independent parameters building up the matrix $\hat{D}$ can be determined.

As already pointed out in section V, one possibility is the calculation of crosscorrelation functions. In quaternary sequences, where $\frac{\lambda \cdot (\lambda - 1)}{2} = 6$ of 9 parameters have been determined by autocorrelation functions, the remaining $\frac{(\lambda - 1) \cdot (\lambda - 2)}{2} = 3$ antisymmetric

entries can be estimated by three independent crosscorrelation functions, for example: $C_{(1,0,0,0),(0,1,0,0)}$, $C_{(1,0,0,0),(0,0,1,0)}$, and $C_{(0,1,0,0),(0,0,1,0)}$.

In order to measure all existing correlations by a single function, one can calculate the mutual information function as discussed in sections I and II. We study the mutual information function of genomic DNA sequences in chapter 13, and we dedicate chapters 14, 15, 18, and 16 to further applications of the mutual information to the analysis of DNA sequences.

4.7 Summary

Several statistical measures can detect long-range correlations in symbolic strings such as English texts or DNA sequences. We understand all of them as functions that quantify the degree of statistical dependence between symbols in a distance k. Hence, we raised the question how many parameters have to be estimated in order to determine all correlations in a given distance. We could show that, for any distance k, $(\lambda - 1)^2$ parameters are required to identify all correlations in a sequence composed by an alphabet of λ symbols.

We turned to the question whether autocorrelation functions could determine these $(\lambda - 1)^2$ parameters and realized that they can detect only $\frac{\lambda \cdot (\lambda - 1)}{2}$ parameters. Moreover, we learned that all autocorrelation functions are linearly dependent on a set of $\frac{\lambda \cdot (\lambda - 1)}{2}$ properly chosen autocorrelation functions that we termed a basis.

Sections IV, V, and VI were devoted to selecting an appropriate basis for binary, ternary, and quaternary sequences, respectively.

We could show that one autocorrelation function is sufficient to measure all correlations in binary sequences. Thus we were able to state that all symbols (in a given distance) are statistically independent if, and only if, the chosen autocorrelation function vanishes.

In section V, we considered ternary sequences to illustrate our general results summarized in Appendix I. We realized that only three of four parameters required to identify all correlations can be determined by autocorrelation functions. Hence, statistical independence cannot be guaranteed by the disappearance of all autocorrelation functions as illustrated by our example in that section. However, crosscorrelation functions and the mutual information function were found to be a welcome tool to detect all correlations.

Section VI was dedicated to quaternary sequences, which are of particular interest in biology. We showed that all autocorrelation functions are linearly dependent on a basis

of $\frac{\lambda\cdot(\lambda-1)}{2} = 6$ functions, all of which carry a biological interpretation. In particular, we derived eq. (59) connecting seven autocorrelation functions that are widely used to analyze DNA sequences. At the end of that section, we emphasized again that crosscorrelations or mutual information are required to determine the remaining $\frac{(\lambda-1)\cdot(\lambda-2)}{2} = 3$ parameters of the dependence matrix $\hat{D}$.

In Appendix A, we derived that a properly chosen set of $\frac{\lambda\cdot(\lambda-1)}{2}$ autocorrelation functions is sufficient to identify all independent parameters of the symmetric matrix $\hat{S}$. Hence, we proved that these $\frac{\lambda\cdot(\lambda-1)}{2}$ autocorrelation functions form a basis. We delivered an algorithm that allows to construct the symmetric matrix $\hat{S}$ in terms of a properly chosen basis.

Chapter 5

Generalized Entropies

The order-q Tsallis (H_q) and Rényi entropies (K_q) receive broad applications in the statistical analysis of complex phenomena. Here we provide a brief introduction to both H_q and K_q, and we present some connections between these two families of generalized entropies.

5.1 Introduction

Building on the works of Shannon [Shannon 1948] and Khinchin [Khinchin 1957a], generalized entropies have witnessed an increasing interest in their application to characterize complex behavior in models and real systems. As the Shannon entropy is formally defined as an average value, the idea underlying a generalization is to replace the average of logarithms by an average of powers. Then this gives rise to the order-q Tsallis entropy H_q [Tsallis 1988, Curado & Tsallis 1991] or, similarly, the Rényi entropy K_q [Rényi 1970]. The external parameter q applies to describe inhomogeneous structures of the probability distribution and whence the associated process under consideration. From both order-q entropies, H_q and K_q, the Shannon entropy is obtained in the limit $q \to 1$. Applications of order-q entropies occur in a variety of fields of sciences like, e.g., non-linear dynamical systems [6-10], statistical thermodynamics [11-16], classical mechanics [Plastino 1994], or evolutionary programming [Stariolo & Tsallis 1995, Penna 1995].

5.2 Definitions and Properties

This section is aimed at introducing the notation used throughout this work as well as giving the definitions of the order-q Tsallis entropy, H_q, and the Rényi entropy, K_q. We then review some basic properties of these entropies, which will finally allow us the derivation of an indirect Bayes estimator of the Rényi entropy in section 5.

Consider a random variable A that can take on M different discrete values a_i, $i = 1, \ldots, M$, with an associated probability-vector $\vec{p} \equiv \{p_1, \ldots, p_M\}$ with components $p_i \equiv p(a_i)$. The probabilities satisfy the two constraints $0 \leq p_i \leq 1$ and $\sum_{i=1}^{M} p_i = 1$. It is customary to refer to the set of all possible outcomes as the alphabet $\mathcal{A}$ with cardinality M. Then the Shannon entropy of A is defined as

$$H(A) = -\sum_{i=1}^{M} p_i \log_2 p_i \equiv -\langle \log_2 p_i \rangle. \tag{5.1}$$

Since the base of the logarithm is chosen to be 2, the Shannon entropy is measured in units of bits. One distinctive property of H, which is not shared by the generalized entropies, is worth mentioning: the entropy of a composite event can be given as the sum of the marginal and the conditional entropy.

By equation (5.1), events having either a particularly high or low occurrence do not contribute much to the Shannon entropy. In order to weight particular regions of the probability-vector $\vec{p}$, one can consider the following partition function:

$$Z_q(A) = \sum_{i=1}^{M} p_i^q \equiv \langle p_i^{q-1} \rangle. \tag{5.2}$$

In contrast to equation (5.1), the average of logarithms is now replaced by an average of powers of q. Clearly, a change of the order q will change the relative weights of how the event i contributes to the sum. Therefore, varying the parameter q allows to monitor the inhomogeneous structure of the distribution $\vec{p}$: the larger q, the more heavily the larger probabilities enter into Z_q, and vice versa. Obviously, Z_0 equals the number of events i with non-vanishing probability, and Z_1 introduces normalization. Then the order-q Tsallis entropy is defined as

$$H_q(A) = \frac{1}{\ln 2} \frac{Z_q(A) - 1}{1 - q} \equiv \frac{1}{\ln 2} \frac{\langle p_i^{q-1} - 1 \rangle}{1 - q}. \tag{5.3}$$

Since the prefactor is chosen to be $1/\ln 2$, the Tsallis entropy is measured in units of bits. This can be seen by considering the limit $q \to 1$: we easily verify that $\lim_{q \to 1} H_q = H$

holds.

The order-q entropy due to Rényi is given by

$$K_q(A) = \frac{1}{1-q}\log_2 Z_q(A) \equiv -\log_2 \left\langle p_i^{q-1} \right\rangle^{1/(q-1)}. \tag{5.4}$$

Here the argument of the logarithm is the generalized average of the numbers p_i. By equation (5.3), the relationship connecting both order-q entropies reads as

$$K_q(A) \;=\; \frac{1}{1-q}\log_2 \Big[1 + (1-q)\ln 2\; H_q(A)\Big]. \tag{5.5}$$

From equation (5.5) we see that for fixed q, K_q and H_q are monotonic functions of one another and that $\lim_{q\to 1} K_q = H$ holds.

Let us summarize the following features of order-q entropies:

1. $H_q \geq 0$ and $K_q \geq 0$. For given M, the global maxima (minima) are attained at $p_i = 1/M \;\forall i$ for $q > 0$ ($q < 0$). In particular, we have that $K_q^{max} = H^{max}$.

2. H_q and K_q are monotonically decreasing functions of q for arbitrary probability-vectors $\vec{p}$: $H_q \geq H_{q'}$ and $K_q \geq K_{q'}$ for $q < q'$.

3. $H_q(A)$ is a concave (convex) function of the probabilities given $q > 0$ ($q < 0$). The curvature dependence of K_q upon q and $\vec{p}$ is non-trivial [Curado & Tsallis 1991]. Yet the following two inequalities hold: K_q is a convex (concave) function of p_i for $q < 0$ ($0 < q \leq 1$).

4. Considering two subsets, $\mathcal{A}$ and $\mathcal{B}$, then $K_q(A,B)$ obeys additivity for independent random variables, whereas $H_q(A,B)$ is pseudo-additive. That is, we find $H_q(A,B) = H_q(A) + H_q(B) + (1-q)H_q(A)H_q(B)$. Furthermore, $H_q(A,B)$ generalizes the Shannon-additivity to the order q (see, e.g., [Shannon 1948] or [Curado & Tsallis 1991] for a definition and discussion).

By the above properties, the whole set of order-q entropies (which generalize the Shannon entropy) provides us with a whole spectrum of entropies, in which $q = 1$ is singled-out by the property of composite events. In the light of the fact that K_q is indeed additive but, in general, not a concave (convex) function of the probabilities p_i on the entire simplex, it is remarkable that via the non-linear transformation (5.5) we are able to switch between two types of entropies of order q, either having the property of additivity or of well-defined concavity (convexity).

5.3 Summary

In this chapter we presented a brief introduction to the definition and statistical properties of generalized entropies. These generalizations of the Shannon entropy will turn out to be of practical importance in later chapters. Before we apply these generalized entropies to DNA or amino acid sequences, we will first study how to estimate generalized entropies from finite data sets. We will present an introduction to the theory of estimating population parameters from finite data sets in the following chapters.

Part B

Chapter 6

Estimating Population Parameters from Finite Samples

In our previous chapter we have introduced some measures designed to determine the information content stored in symbolic sequences such as time series, natural texts, pieces of music, econometric data, or DNA strings.

In chapter 2, we have learned that the extraction of biologically relevant information from DNA, RNA, or amino acid sequences is the main goal in the stormingly evolving discipline called computational molecular biology.

The problems that we are always confronted with arise from the uncircumventable fact that all existing sequences have a finite length and thus do often not allow us to estimate quantities that we desire to determine with sufficient reliability. However, this dilemma is identical to the basic task dealt with in mathematical statistics.

The general problem in mathematical statistics is the estimation of certain population parameters θ or of functions $f(\theta)$ of those parameters from samples, which always have a finite size. These parameters θ can, for example, be the probability p of a coin, the mean μ, or the variance σ^2 of a normal population.

For the purpose of illustration, imagine we want to estimate the probability of a coin. Strictly speaking, we want to estimate the probability for the event that the coin shows its head after having been tossed. Certainly, we would flip the coin N times, count how often we have observed a head, and calculate the relative frequency as k/N if the absolute frequency (the number of heads) is denoted by k.

Then we would pretend that this relative frequency k/N is a very good estimator for

the probability p of this coin at least for sufficiently large N. The questions that arise at this point are, however, whether there is a motivation behind this recipe to accept the relative frequency k/N as an estimator for the probability p, and whether this recipe is indeed valuable in cases when the number of sample points N is very small, which often occurs when analyzing biosequences.

In section 6.2, we will present the maximum likelihood method, which will eventually allow us to answer the question arisen above. We will discuss some general statistical properties shared by maximum likelihood estimators and apply them to our primary problem to estimate entropies from finite samples.

In section 6.5, we will introduce the Bayes estimator, which — instead of maximizing the likelihood — minimizes the variance of the estimate. Our final goal is the derivation of the Bayes estimator for the Shannon entropy in chapter 7 and the discussion of its statistical properties.

Section 6.6 is devoted to comparing the maximum likelihood estimator with the Bayes estimator. In this section, we will show that the maximum likelihood estimator chooses, under the Bayes hypothesis, the maximum of the posterior density function of the quantity to be estimated, whereas the Bayes estimator chooses the expectation value of the posterior distribution as its estimate.

Recommendable books on this field are, for example, [Martin 1971], [McEliece 1977], [Gnedenko 1981], [Rényi 1982], [Fisz 1989], or [Wickmann 1990].

6.1 Sampling Theory and the Analysis of Time Series

In this section, we will introduce some helpful definitions and notations, which allow us to formulate our task to estimate population parameters or functions of them from finite samples in a mathematical language.

A necessary requirement for any statistical analysis is a set of initial conditions that are reproducible. These conditions define an *experiment*, and by making an observation (or a set of observations) we produce an *outcome* of this experiment. Let's denote the outcomes, which are either single numbers or possibly sets of numbers, by x_i.

Definition 6.1 *The set of all possible outcomes* x_i $(i = 1, 2, ..., n)$ *of an experiment is called the* **sample space** *or* **population**.

The random numbers x_i *are called* **sample points** *in the sample space.*

Definition 6.2 *A subset of the sample points, e.g.* x_i $(i = 1, 2, ..., N)$ *is called an* **event** *and denoted by*

$$E \equiv \{x_i : i = 1, 2, ..., N\}. \tag{6.1}$$

Definition 6.3 *If* $x_1, x_2, ..., x_N$ *denotes a set of numerical values of* N *observations selected from a larger set, then the set of values is called a* **sample of size N**.

Definition 6.4 *A numerical value determined from some, or all, of the values of a sample is called a* **statistic**.

The basic task of mathematical statistics is now to provide us with deliberate tools of estimating the values of the parameters of a population by calculating certain statistics from a sample.

As an example, we consider the estimation of the probability of a given coin. Our experiment can be defined as "tossing a given coin" resulting in outcomes head or tail up. Thus, our sample space is discrete and contains but two elements, which we like to denote by 1 and 0, respectively. A sample of size N is then a list of N values $x_i \in \{0, 1\}$ corresponding to the observed events tail or head up.

The *natural estimator* of the Shannon entropy, which is defined as

$$\hat{H}(x_1, x_2, ..., x_N) = -\frac{k}{N} \cdot \log \frac{k}{N} - (1 - \frac{k}{N}) \cdot \log(1 - \frac{k}{N}) \tag{6.2}$$

with

$$k = k(x_1, x_2, ..., x_N) = \sum_{i=1}^{N} x_i \tag{6.3}$$

is a function of the sample points $x_1, x_2, ..., x_N$ and thus commonly termed a statistic. This statistic can formally be regarded as a random variable and then be analyzed with respect to its statistical properties, which are referred to as systematic and statistical errors.

The statistical properties of this statistic $\hat{H}(x_1, x_2, ..., x_N)$ are the focus of our investigations in chapter 9.

The probability p of the coin to show its head is the population parameter, which we generally denote by θ. Of course, nobody knows the value of θ. It is just the value of θ that we want to estimate by sampling.

6.2 Maximum Likelihood Estimator

To prepare the introduction of the maximum likelihood method, let us define the *likelihood function.*

Definition 6.5 *Let $L(x;\theta)$ denote the density function of a random variable x, where the form of L is known, but not the value of θ, which is to be estimated. Let $x_1, x_2, ..., x_N$ be a random sample of size N. The joint density function $L(x_1, x_2, ..., x_N;\theta)$ of the independent random variables $x_1, x_2, ..., x_N$ is then given by*

$$L(x_1, x_2, ..., x_N;\theta) = \prod_{i=1}^{N} L(x_i;\theta) \tag{6.4}$$

and called the **likelihood function of** θ.

In the following, we suppress the dependence of L on $x_1, x_2, ..., x_N$ and come to the definition of the maximum likelihood estimator of the population parameter θ from a sample of size N.

Definition 6.6 *The* **maximum likelihood estimator** *of the population parameter θ is that statistic $\hat{\theta}$ which maximizes $L(\theta)$ for variations of θ.*

Here, we stick to the common notation that $\hat{\theta}$ denotes the estimator of the the population parameter θ.

If L is sufficiently smooth, i.e., if the second derivative exists, then the maximum likelihood estimator can be defined to be that $\hat{\theta}$ for which the following necessary equations hold:

$$\frac{\partial L(\theta)}{\partial \theta} = 0 \tag{6.5}$$

and

$$\frac{\partial^2 L(\theta)}{\partial \theta^2} < 0. \tag{6.6}$$

Obviously, the maximum likelihood method leads us directly to an optimization problem in the parameter space of θ, which can be perfectly dealt with by physicists, since variational principles make up the fundamental laws of physics.

Since $L(\theta) > 0$ and $\ln(\theta)$ is a monotonic function of θ, the first equation is equivalent to

$$\frac{\partial \ln L(\theta)}{\partial \theta} = 0, \tag{6.7}$$

where the function $\ln L(\theta)$ is often referred to as the *log-likelihood function.*

Before we start to derive some general theorems about maximum likelihood estimators, we want to calculate the maximum likelihood estimator for the probability of a coin.

Let p be the probability for the observation of a coin's head after a toss, and q the probability of observing the coin's tail. Since $p+q=1$, the only population parameter is p.

The likelihood for tossing the sequence $x_1, x_2, ..., x_N$ is $p^{\left(\sum_{i=1}^{N} x_i\right)} \cdot (1-p)^{\left(\sum_{i=1}^{N} 1-x_i\right)}$ and thus the likelihood function for this Bernoulli process of N tosses is

$$L(p) = p^k \cdot (1-p)^{N-k} \tag{6.8}$$

if k is the number of heads we observed and $N-k$ is the number of tails, i.e., $k = \sum_{i=1}^{N} x_i$.

For $k=0$ or $k=N$, we immediately see that

$$\hat{p} = 0 \tag{6.9}$$

and

$$\hat{p} = 1 \tag{6.10}$$

maximizes the likelihood function $L(p)$ for $k=0$ and $k=N$, respectively.

For all other k, $(0<k<N)$, setting the partial derivative of the log-likelihood function

$$\frac{\partial \ln L(p)}{\partial p} = \frac{\partial (k \cdot \ln p + (N-k) \cdot \ln(1-p))}{\partial p} = \frac{k}{p} - \frac{N-k}{1-p} \tag{6.11}$$

identical to zero, yields the solution

$$\hat{p} = \frac{k}{N}, \tag{6.12}$$

which is neither 0 nor 1 for $0<k<N$ so that all operations performed above are possible, i.e., no division by zero occurs.

Summarizing eqs. (6.9), (6.10), and (6.12), we can state that the maximum likelihood estimator $\hat{p}$ for the probability p of a coin[1] is given by the relative frequency k/N of the outcomes associated with p.

[1] In German texts, we find the term *0 - 1 - population* for the population corresponding to the distribution function

$$F(x) = \begin{cases} 0 & for \quad -\infty < x < 0 \\ p & for \quad 0 \leq x < 1 \\ 1 & for \quad 1 \leq x < \infty \end{cases} \tag{6.13}$$

Since we do not like to term the M-dimensional generalization a *0 - 1 - 2 - ... - (M − 1) - population*, we decided to name the population corresponding to the first distribution a *coin-population* and the sample space of an M-sided die an *M-sided-die-population.*

In other words, the assumption that the (hidden) probability p is indeed equal to k/N has the highest likelihood compared to all other estimates.

It is shown in appendix D.1 that the maximum likelihood estimator of the probability vector $\vec{p}$ of an M-sided-die yields exactly the same recipe, namely, the identification of the components p_i of the vector $\vec{p}$ with the relative frequencies of the observations, k_i/N. Hence, if we are to estimate the six probabilities of a die, we start rolling the die N times, count the absolute frequencies k_i $(i = 1, 2, ..., 6)$ of the six sides of the die, and then estimate the probabilities $\hat{p}_i = k_i/N$, since this is the estimate with the highest likelihood.

Let's now generalize the maximum likelihood concept on cases where we want to estimate several parameters $\theta_1, \theta_2, ..., \theta_M$ of a population from a sample of size N.

We define the likelihood function of the parameters $\theta_1, \theta_2, ..., \theta_M$ in a straightforward way as $L(x_1, x_2, ..., x_N; \theta_1, \theta_2, ..., \theta_M)$, suppress the dependence on the sample points $x_1, x_2, ..., x_N$, and define the maximum likelihood estimator of the parameters $\theta_1, \theta_2, ..., \theta_M$ to be the solution of the following system of M equations:

$$\frac{\partial \ln L(\theta_1, \theta_2, ..., \theta_i, ..., \theta_M)}{\partial \theta_i} = 0 \tag{6.14}$$

for all $i = 1, 2, ..., M$.

To illustrate how the maximum likelihood method works for a set of parameters $\theta_1, \theta_2, ..., \theta_M$ that we want to estimate simultaneously from a sample of size N, let us consider the example of a normal population where both the mean μ and the variance σ are to be estimated.

The likelihood function reads as

$$L(\mu, \sigma) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp(-\frac{(x_i - \mu)^2}{2\sigma^2}) \tag{6.15}$$

and thus the two equations

$$\frac{\partial L(\mu, \sigma)}{\partial \mu} = 0 \tag{6.16}$$

and

$$\frac{\partial L(\mu, \sigma)}{\partial \sigma} = 0 \tag{6.17}$$

can be solved analytically:

$$\frac{\partial \ln L(\mu, \sigma)}{\partial \mu} = -\sum_{i=1}^{N} \frac{2 \cdot (x_i - \mu)}{2 \cdot \sigma^2} = 0 \tag{6.18}$$

gives

$$\hat{\mu} = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{6.19}$$

and

$$\frac{\partial \ln L(\mu,\sigma)}{\partial \sigma} = -\frac{N}{\sigma} + \sum_{i=1}^{N} \frac{(x-\mu)^2}{\sigma^3} = 0 \tag{6.20}$$

gives

$$\hat{\sigma^2} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2. \tag{6.21}$$

Please realize that we have derived the maximum likelihood estimator of σ and not of σ^2. Theorem 6.1 will however prove the identity of the maximum likelihood estimator of σ^2 and the squared maximum likelihood estimator of σ.

We note that the maximum likelihood estimator for the variance of a normal population is equal to the sum of the quadratic deviations from the sample mean divided by N and not by $(N-1)$. Before we however can give a reason why the estimator

$$\hat{\sigma^2} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 \tag{6.22}$$

should better be used to estimate the variance of normally distributed populations, we have to define two properties of point estimators.

6.3 Consistency and Bias

It is intuitively obvious that a desirable property of any estimator is that its estimates tend to the value of its population parameter as the sample size N increases. Any other behavior would clearly be misleading. This property is commonly called *consistency* and defined as follows:

Definition 6.7 *An estimator $\hat{\theta}_N$ (note that N is the index indicating the sample size) is said to be a* **consistent estimator** *of the population parameter θ if, for any positive (arbitrarily small) ϵ and η, there exist some N such that*

$$P\left(\left|\hat{\theta}_N - \theta\right| < \epsilon\right) > 1 - \eta. \tag{6.23}$$

This means that an estimator $\hat{\theta}$ is consistent if $\{\hat{\theta}_N\}_{N\in\mathcal{N}}$ converges in probability to θ for $N \to \infty$.

We may further restrict the possible estimators by requiring that for all N the expectation value of $\hat{\theta}_N$ is equal to θ. Such an estimator is termed *unbiased.*

Definition 6.8 *An estimator $\hat{\theta}_N$, computed from a sample of size N, is called an* **unbiased estimator** *if*

$$E(\hat{\theta}_N) = \theta \tag{6.24}$$

for all $N \in \mathcal{N}$.

At this point we can judge that the maximum likelihood estimator of the probability vector $\vec{p}$ of a multinomial population is unbiased since

$$E(k_i/N) = p_i \tag{6.25}$$

for all $i = 1, 2, ..., M$ as shown in appendix I.

The maximum likelihood estimator for the mean μ of a normal population is also unbiased, whereas the maximum likelihood estimator for the variance σ^2 of a normal population is not as we will show below.

In appendix C, we will calculate the expectation value of the maximum likelihood estimator for the variance of a normal population and show that it is by a factor of $\frac{N-1}{N}$ smaller than the population variance; in mathematical terms:

$$E(\hat{\sigma^2}_N) = \frac{N-1}{N} \cdot \sigma^2. \tag{6.26}$$

To illustrate the difference between $\hat{\sigma^2}$ and σ^2, imagine we want to estimate the variance of a normal population. We consider a sample of size N and determine σ^2 by calculating the maximum likelihood estimator given in eq. (6.19). We chose this estimator, because it maximizes the likelihood for our estimate, i.e., it maximizes the probability that our estimate is indeed identical with the 'true' sample variance.

Of course, the maximum likelihood estimator does not guarantee that we are always correct with our estimate, i.e., we can understand this estimate as a random variable and then ask for its expectation value. What we will show in appendix C is that, by applying the maximum likelihood estimator to the variance of a normal population, we systematically underestimate the population variance by a factor of $\frac{N-1}{N}$ on average.

These introductory examples were presented to illustrate the basic ideas of our approach to evaluate measures and algorithms currently used to analyze DNA, RNA, or protein sequences by estimating (or approximating) their statistical properties like systematic and statistical errors induced by an always finite size of available samples.

6.4 Properties of Maximum Likelihood Estimators

Let us eventually present a theorem about the maximum likelihood estimator of a function $f(\theta)$ of a population parameter θ that is of crucial interest for all practical applications dealing with estimating functions of probabilities from finite samples.

Theorem 6.1 *If $f(\theta)$ is a monotonic function of θ, then the maximum likelihood estimator of f is given by*

$$\hat{f}(\theta) = f(\hat{\theta}) \tag{6.27}$$

if $\hat{\theta}$ is the maximum likelihood estimator of θ.

This means that the estimator $\hat{f}(\theta)$ does not necessarily maximize the likelihood of f if f is not a monotonic function.

Let us illustrate this message by the following example:

The Shannon entropy of a coin is given by

$$H(p) = -p \cdot \log(p) - (1-p) \cdot \log(1-p), \tag{6.28}$$

which is a non-monotonic function of p.

It is shown in appendix D.2 that the *natural estimator* of the Shannon entropy,

$$\hat{H}(k) = -\frac{k}{N} \cdot \log(\frac{k}{N}) - (1 - \frac{k}{N}) \cdot \log(1 - \frac{k}{N}), \tag{6.29}$$

is not the maximum likelihood estimator of H although $\hat{p} = \frac{k}{N}$ is the maximum likelihood estimator of p.

Before we go over to introduce the Bayes estimator in the following section, let us mention three theorems that might emphasize the importance of maximum likelihood estimators in general.

Theorem 6.2 *Maximum likelihood estimators are consistent.*

Theorem 6.3 *Maximum likelihood estimators have a distribution that tends to normality for large samples.*

Theorem 6.4 *Maximum likelihood estimators have a minimum variance in the limit of large samples.*

The value of theorem 6.4 is very questionable for analyzing biosequences as well as time series in meteorology, economics, or theoretical physics since the studied samples are often so small that we have to estimate population parameters or functions of them in the range of extremely poor statistics.

At this point we are ready to rise the question whether or not there is an alternative to the maximum likelihood estimator that has the property of minimizing the variance not only for infinitely large N but for all $N \in \mathcal{N}$. The next chapter will be devoted to answering this question and deriving a minimum variance estimator.

6.5 Bayes Estimator

In this section, we will derive an estimator that minimizes the variance of its estimates for all sample sizes $N \in \mathcal{N}$. We will introduce and solve a variation principle in subsection 6.5.1. In subsection 6.5.2, we will then calculate the Bayes estimator for the probability of a coin and compare the results with the maximum likelihood estimator of the same population. Finally, subsection 6.6 is devoted to stressing some general relations between maximum likelihood and Bayes estimators.

6.5.1 Minimum Variance Principle

In this subsection, we will introduce the Bayes estimator by considering the following example:

Imagine we have two estimators $\hat{p}_1(k)$ and $\hat{p}_2(k)$ for the probability p of a coin. The question "Which of these estimators is the better one?" cannot be answered in general, because we have not yet specified what a good estimator should look like.

Certainly, it should predict the 'true' sample probability as precise as possible, which means $(\hat{p}(k) - p)^2$ be minimal.

Since we, however, do not know the value of p — this is just the parameter that we are going to estimate — we define the functional

$$F[\hat{p}(k)] \equiv \int_0^1 (\hat{p}(k) - p)^2 \cdot P(k|p) \cdot P(p)\, dp, \tag{6.30}$$

which we are going to minimize in order to obtain our desired estimator $\hat{p}(k)$. Let us, in the following, motivate this step.

The term $(\hat{p}(k) - p)^2$ quantifies the quadratic deviation of the estimate $\hat{p}(k)$ from the true population parameter p, which we choose as our penalty. Now realize that it is not only important by which magnitude our estimates deviate from the theoretical population parameter, but also how often a certain quadratic deviation appears.

The value of an estimator that is always a bit wrong might be comparable to the value of another one that is almost always right, but sometimes extremely far from the true population parameter. Hence, we multiply the penalty $(\hat{p}(k) - p)^2$ by the probability $P(k|p)$ by which this deviation appears.

$P(k|p)$ is the conditional probability of obtaining k heads from a sample of N tosses under the assumption that the probability of the coin be p.

The product $(\hat{p}(k) - p)^2 \cdot P(k|p)$ does still depend on the population parameter p, i.e., before we integrate this term over the entire parameter space, which is the interval $[0, 1]$ in our case, we might assume that some p appear more frequently than others. This prior assumption is expressed by the *prior probability density* $P(p)$ of the population parameter p.

Consequently, we also multiply our integrand by $P(p)$ and thus end up with eq. (6.30) for the functional that we desired to minimize.

The minimization of the functional $F[\hat{p}(k)]$ is trivial and we immediately obtain

$$\hat{p}(k) = \frac{\int_0^1 p \cdot P(k|p) \cdot P(p)\, dp}{\int_0^1 P(k|p) \cdot P(p)\, dp} \tag{6.31}$$

as that estimator which minimizes $F[\hat{p}(k)] \equiv \int_0^1 (\hat{p}(k) - p)^2 \cdot P(k|p) \cdot P(p)\, dp$ for all $N \in \mathcal{N}$.

At this point, we are ready to state the minimum variance principle and to define the resulting Bayes estimator of the population parameter θ.

Definition 6.9 *Let θ be the only parameter of our given population, $P(\theta)$ its prior probability density, and $P(x_1, x_2, ..., x_N|\theta)$ the conditional probability to realize the sample $x_1, x_2, ..., x_N$.*

Then we define the **Bayes estimator** *of the population parameter θ to be that statistic $\hat{\theta}(x_1, x_2, ..., x_N)$ which minimizes*

$$F[\hat{\theta}(x_1, x_2, ..., x_N)] \equiv \int (\hat{\theta}(x_1, x_2, ..., x_N) - \theta)^2 \cdot P(x_1, x_2, ..., x_N|\theta) \cdot P(\theta)\, d\theta \tag{6.32}$$

where the integral stretches over the whole parameters space of θ.

Minimizing the functional $F[\hat{\theta}(x_1, x_2, ..., x_N)]$ leads us again to a simple quadratic equation the solution of which is

$$\hat{\theta}(x_1, x_2, ..., x_N) = \frac{\int \theta \cdot P(x_1, x_2, ..., x_N|\theta) \cdot P(\theta)\, d\theta}{\int P(x_1, x_2, ..., x_N|\theta) \cdot P(\theta)\, d\theta}. \tag{6.33}$$

In section 6.6, we will show that this estimator $\hat{\theta}(x_1, x_2, ..., x_N)$ is identical to the expectation value of the *posterior probability distribution* of θ.

At the end of this section, we present the definition of the Bayes estimator for populations described by more than one parameter.

Let $\vec{\theta} = (\theta_1, \theta_2, ..., \theta_M)$ be the M-dimensional vector containing the population parameters $\theta_1, \theta_2, ..., \theta_M$, $P(\vec{\theta})$ its prior probability density, and $P(x_1, x_2, ..., x_N|\vec{\theta})$ the conditional probability to realize the sample $x_1, x_2, ..., x_N$.

Then the Bayes estimator of the population parameter vector $\vec{\theta} = (\theta_1, \theta_2, ..., \theta_M)$ is that vector function $\hat{\vec{\theta}}(x_1, x_2, ..., x_N)$ which minimizes the functional

$$F[\hat{\vec{\theta}}(x_1, x_2, ..., x_N)] \equiv \int\int \cdots \int \left(\hat{\vec{\theta}}(x_1, x_2, ..., x_N) - \vec{\theta}\right)^2 \tag{6.34}$$

$$\cdot \; P(x_1, x_2, ..., x_N|\vec{\theta}) \cdot P(\vec{\theta})\, d\vec{\theta} \tag{6.35}$$

where $\left(\hat{\vec{\theta}}(x_1, x_2, ..., x_N) - \vec{\theta}\right)^2$ denotes the scalar product, i.e., the squared norm of the vector $\hat{\vec{\theta}}(x_1, x_2, ..., x_N) - \vec{\theta}$ and the integral stretches over the entire parameters space of $\vec{\theta}$.

Rewriting the scalar product

$$\left(\hat{\vec{\theta}}(x_1, x_2, ..., x_N) - \vec{\theta}\right)^2 = \sum_{i=1}^{M} \left(\hat{\theta}_i(x_1, x_2, ..., x_N) - \theta_i\right)^2 \tag{6.36}$$

leads to

$$F[\hat{\vec{\theta}}] = \sum_{i=1}^{M} \int\int \cdots \int \left(\hat{\theta}_i(x_1, x_2, ..., x_N) - \theta_i\right)^2 \tag{6.37}$$

$$\cdot \; P(x_1, x_2, ..., x_N|\vec{\theta}) \cdot P(\vec{\theta})\, d\vec{\theta} \;\Rightarrow min. \tag{6.38}$$

If we now minimize each of the M summands individually, i.e., independently on the other $M-1$ terms, we obtain a lower bound for the minimum of the sum. If we can then show that the individual estimates of the parameters θ_i are compatible to the estimate

of $\vec{\theta}$, we have found the Bayes estimator for $\vec{\theta}$ and can display it in the following explicit expression:

$$\hat{\vec{\theta}}(x_1, x_2, ..., x_N) = (\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_M)(x_1, x_2, ..., x_N) \tag{6.39}$$

with

$$\hat{\theta}_i(x_1, x_2, ..., x_N) = \frac{\int\int\cdots\int \theta_i \cdot P(x_1, x_2, ..., x_N|\vec{\theta}) \cdot P(\vec{\theta})\, d\vec{\theta}}{\int\int\cdots\int P(x_1, x_2, ..., x_N|\vec{\theta}) \cdot P(\vec{\theta})\, d\vec{\theta}} \tag{6.40}$$

for all $i = 1, 2, ..., M$.

The whole procedure is exemplified in appendix E.1, where we derive the Bayes estimator for the probabilities of an M-sided die that has been rolled N times under the prior assumption of a uniform distribution of the probabilities p_i $(i = 1, 2, ..., M)$.

However, we recommend to read the following section 6.5.2, in which we derive the Bayes estimator for the probability of a coin under the same assumptions, before going to appendix E.1 and studying the M-dimensional generalization.

6.5.2 Laplace Estimator

In this subsection, we will derive the Bayes estimator for the probability of coin and compare it to the corresponding maximum likelihood estimator.

Our task is to calculate

$$\hat{p}(k) = \frac{\int_0^1 p \cdot P(k|p) \cdot P(p)\, dp}{\int_0^1 P(k|p) \cdot P(p)\, dp} \tag{6.41}$$

with

$$P(k|p) = \binom{N}{k} \cdot p^k \cdot (1-p)^{N-k} \tag{6.42}$$

and

$$P(p) = 1. \tag{6.43}$$

The choice of a uniform prior distribution for p reflects the *maximum entropy principle* applied to the weakest possible assumption that we can make, namely that we do not know anything about the probability p except that it ranges in the interval $[0, 1]$.

The assumption of a uniform prior distribution of p is also called the *Bayes hypothesis*, which leads to

$$\hat{p} = \frac{\int_0^1 p \cdot \binom{N}{k} \cdot p^k \cdot (1-p)^{N-k}\, dp}{\int_0^1 \binom{N}{k} \cdot p^k \cdot (1-p)^{N-k}\, dp} \tag{6.44}$$

for the Bayes estimator of the probability p.

In appendix H.1, we will derive that

$$\int_0^1 p^s \cdot (1-p)^t\, dp = \frac{s! \cdot t!}{(s+t+1)!} \tag{6.45}$$

for any $s \in \mathcal{N}$ and $t \in \mathcal{N}$, which yields

$$\hat{p} = \frac{k+1}{N+2}. \tag{6.46}$$

This is the Bayes estimator for the probability of a coin under the Bayes hypothesis, which is also called the *Laplace estimator*.

We see that, for large N, the difference between the maximum likelihood estimator and the Laplace estimator vanishes. However, for small sample sizes N, the difference becomes significant.

Imagine, for example, we want to estimate the probability of a coin, but we must not flip the coin more than once. Then there are only two outcomes of this experiment possible. Either we obtain a head or a tail. The maximum likelihood estimator would, in this situation, suggest to assume $p = 1$ or $p = 0$, respectively, which is extremely unrealistic.

The Bayes estimator, on the other hand, would recommend you to assume the probabilities $p = 2/3$ or $p = 1/3$, respectively. Note, that it can even deliver a meaningful prediction for the case that we did not flip the coin at all. Then, the estimated probability would be $p = 1/2$, which corresponds exactly to our prior assumption.

In appendix E.1, we will calculate the Laplace estimator for the probability vector $\vec{p} = (p_1, p_2, ..., p_M)$ of an M-sided die. In that case, the vector $\vec{k} = (k_1, k_2, ..., k_M)$ of the absolute frequencies is multinomially distributed, which leads to the Laplace estimator

$$\hat{p}_i = \frac{k_i + 1}{N + M} \tag{6.47}$$

for all $(i = 1, 2, ..., M)$ under the assumption of a uniform prior distribution on the $M-1$ dimensional simplex spanned by all $\vec{p}$.

It is easy to show that the Laplace estimator is consistent, since $\hat{p}_i = \dfrac{k_i + 1}{N + M}$ converges in probability to p.

However, the disadvantage of the Laplace estimator is that it is not unbiased for $p \neq 1/2$, since then

$$E(\hat{p}) - p = \frac{1 - 2 \cdot p}{N + 2} = \frac{1 - 2p}{N + 2} \neq 0. \tag{6.48}$$

6.5.3 Bayes Estimator of Functions of Population Parameters

Remembering our original task to estimate higher order entropies from finite samples, we will apply the minimum variance principle to estimators of functions $f(\vec{\theta})$ of population parameters $\theta_1, \theta_2, ..., \theta_M$ in this section and thus introduce the Bayes estimator $\hat{f}(\vec{k})$.

Let $P(\vec{\theta})$ be the posterior probability density of the vector $\vec{\theta} = (\theta_1, \theta_2, ..., \theta_M)$ containing the M population parameters θ_i, $P(\vec{k}|\vec{\theta})$ be the likelihood of this hypothesis $\vec{\theta}$ to generate the sample of size N represented by the vector $\vec{k}$, and $\hat{f}(\vec{k})$ be the estimator of the function $f(\vec{\theta})$.

Then the minimum variance principle, which recommends us to choose that estimator $\hat{f}(\vec{k})$ which minimizes the mean quadratic deviation from the theoretical function $f(\vec{\theta})$, i.e., which minimizes

$$F[\hat{f}(\vec{k})] \equiv \int \left(\hat{f}(\vec{k}) - f(\vec{\theta})\right)^2 \cdot P(\vec{k}|\vec{\theta}) \cdot P(\vec{\theta})\, d\vec{\theta}, \tag{6.49}$$

defines the Bayes estimator $\hat{f}(\vec{k})$ of the function $f(\vec{\theta})$.

In equivalence to section 6.5.1, we finally obtain

$$\hat{f}(\vec{k}) = \frac{\int\int\cdots\int f(\vec{\theta}) \cdot P(\vec{k}|\vec{\theta}) \cdot P(\vec{\theta})\, d\vec{\theta}}{\int\int\cdots\int P(\vec{k}|\vec{\theta}) \cdot P(\vec{\theta})\, d\vec{\theta}} \tag{6.50}$$

as the Bayes estimator of $f(\vec{\theta})$.

In chapter 7, we will derive the Bayes estimator of the Shannon entropy $H(\vec{p})$, which will turn out to be extremely powerful, if higher order entropies have to be estimated from finite sequences.

Please note in passing that the Bayes estimator of the function $f(\vec{\theta})$ is not necessarily equal to the function f of the Bayes estimator of $\vec{\theta}$.

6.6 Sampling and the Bayes Theorem

In subsection 6.6.1, we want to introduce the Bayes formula, which we then apply to our general task to make a decision about possible values of a function f of population parameters that we are to estimate from a sample of size N.

In subsection 6.6.2, we will then show that the Bayes estimator of $f(\vec{\theta})$ (with $\vec{\theta} = (\theta_1, \theta_2, ..., \theta_M)$ and $\theta_1, \theta_2, ..., \theta_M$ being the population parameters of our considered sample space) is identical to the expectation value of f over its posterior distribution. This, however, means that the expectation value of $f(\vec{\theta})$ over the posterior probability density $P(\vec{\theta}|x_1, x_2, ..., x_N)$ is that estimator $\hat{f}(\vec{\theta})$ which minimizes the mean quadratic deviation from the real population parameter function $f(\vec{\theta})$.

In subsection 6.6.3, we will derive that, under the Bayes hypothesis, i.e., under the assumption of a uniform prior density $P(\vec{\theta})$, the maximum likelihood estimator chooses that value $f(\vec{\theta})$ which maximizes the posterior probability density $P(f|x_1, x_2, ..., x_N)$ if $x_1, x_2, ..., x_N$ are the outcomes of our sampling experiments.

The interesting point is that, under the Bayes assumption, the maximum likelihood estimator does not only maximize the conditional probability $P(x_1, x_2, ..., x_N|f)$, which this estimator is supposed to do by definition, but also the posterior probability density $P(f|x_1, x_2, ..., x_N)$.

Finally, we will compare and contrast the Bayes estimator and the maximum likelihood estimator in subsection 6.6.4

6.6.1 Bayes Formula

After we will have introduced the Bayes formula for discrete random variables, we will apply it to the question what we can learn about a population parameter by sampling. In this context, we will introduce the terms prior and posterior probability and relate them to the Bayes Formula for continuous variables at the end of this subsection.

Let $A, B_1, B_2, ..., B_{M-1}$, and B_M be $M + 1$ events, $P(A), P(B_1), P(B_2), ..., P(B_{M-1})$ and $P(B_M)$ their probabilities, and $P(A|B_i)$ the conditional probability of the event A given that the event B_i has occurred for $i = 1, 2, ..., M$.

The question that Thomas Bayes, an English clergyman, raised and answered in 1763, was whether we can now calculate the conditional probability $P(B_i|A)$ for the event B_i given that the event A has occurred.

Theorem 6.5 (Bayes Theorem) *If the events B_i $(i = 1, 2, ..., L)$ are mutually exclusive and exhaustive (i.e. all possible events are included in the B_i) events, and if A can occur only in combination with one of the L events B_i, then the conditional probabilities $P(B_i|A)$ can be calculated as follows:*

$$P(B_i|A) = \frac{P(A|B_i) \cdot P(B_i)}{\sum_{j=1}^{L} P(A|B_j) \cdot P(B_j)} \tag{6.51}$$

for all $i = 1, 2, ..., L$.

The proof is extremely simple.

$$P(A \cap B_i) = P(A|B_i) \cdot P(B_i) \tag{6.52}$$

and

$$P(B_i \cap A) = P(B_i|A) \cdot P(A) \tag{6.53}$$

by definition for all $i = 1, 2, ..., L$.

However, since

$$P(A \cap B_i) = P(B_i \cap A), \tag{6.54}$$

we obtain

$$P(B_i|A) = \frac{P(A|B_i) \cdot P(B_i)}{P(A)}. \tag{6.55}$$

Now,

$$P(A) = \sum_{j=1}^{L} P(A|B_j) \cdot P(B_j), \tag{6.56}$$

and hence,

$$P(B_i|A) = \frac{P(A|B_i) \cdot P(B_i)}{\sum_{j=1}^{L} P(A|B_j) \cdot P(B_j)} \tag{6.57}$$

for all $i = 1, 2, ..., L$.

In the following part of this section, we will introduce the terms *prior* and *posterior probability* and relate them to the Bayes Formula for discrete variables. Let us, for the sake of simplicity, consider the following example.

We are given two coins about which we only know that one coin is a fair coin, i.e., the probability to show its head is $p_1 = 1/2$, whereas the other coin has a probability of $p_2 = 4/5$ to show its head. Now it is our task to decide which coin the fair one is.

Before we start tossing one of the coins, we do not know anything about them. We cannot at all distinguish between them without starting to toss one of the coins. Hence, our intuitive assumption about the *prior probability* that we have chosen the fair coin is 1/2, which we also obtain by exploiting the maximum entropy principle.

In order to increase our knowledge about the question which coin we have chosen, we start tossing the coin N times. Let k be the number of heads we have observed. Then

$$P(k|p_1) = \binom{N}{k} \cdot p_1^k \cdot (1-p_1)^{N-k} \tag{6.58}$$

is the probability that the first coin has produced our observed sample whereas

$$P(k|p_2) = \binom{N}{k} \cdot p_2^k \cdot (1-p_2)^{N-k} \tag{6.59}$$

is the probability that the unfair coin has generated it.

The important question that we, however, ask is what is the probability that our sample was generated by the fair coin? In other words, what is the probability that we have chosen the fair coin?

This question can easily answered by applying the Bayes formula and calculating the *posterior probability* $P(p_1|k)$:

$$P(p_1|k) = \frac{P(k|p_1) \cdot P(p_1)}{P(k|p_1) \cdot P(p_1) + P(k|p_2) \cdot P(p_2)}. \tag{6.60}$$

Analogously, we obtain

$$P(p_2|k) = \frac{P(k|p_2) \cdot P(p_2)}{P(k|p_1) \cdot P(p_1) + P(k|p_2) \cdot P(p_2)}, \tag{6.61}$$

which immediately reveals that

$$P(p_1|k) + P(p_2|k) = 1. \tag{6.62}$$

Therefore, the posterior probability $P(p_1|k)$ reflects the knowledge about the chosen coin that we have gained by our sampling experiment of tossing the coin N times.

Finally we can state that the Bayes formula relates the prior probabilities $P(B_i)$ of possible hypotheses B_i with their posterior probabilities $P(B_i|A)$, if we understand A as the outcome of a given sampling experiment.

If the set of all possible hypotheses we are taking into account before setting up an experiment is not discrete but continuous, we have to modify the Bayes formula for continuous variables. Then, the prior probabilities become probability densities as well as the

posterior probabilities become posterior probability densities and the Bayes formula reads as

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{\int P(A|B) \cdot P(B)\, dB}. \tag{6.63}$$

6.6.2 Bayes Estimator - Revisited

The Bayes formula allows us to derive the posterior probability distribution for any set of hypotheses and any possible experimental outcomes provided we can express our prior assumption as well as the conditional probabilities $P(A|B)$ in mathematical terms and can then calculate the occurring sum or integral. Often, we will not be able to derive a closed form expression of the posterior density, but let us in the following analyze what we can state about its maximum and expectation value.

The expectation value of $\vec{\theta}$ over the posterior probability distribution $P(\vec{\theta}|\vec{x})$ reads as

$$E(\vec{\theta}) = \int \vec{\theta} \cdot P(\vec{\theta}|\vec{x})\, d\vec{\theta} \tag{6.64}$$

$$= \frac{\int \vec{\theta} \cdot P(\vec{x}|\vec{\theta}) \cdot P(\vec{\theta})\, d\vec{\theta}}{\int P(\vec{x}|\vec{\theta}) \cdot P(\vec{\theta})\, d\vec{\theta}}, \tag{6.65}$$

which is identical to the Bayes estimator of $\vec{\theta}$.

In general, we can state that the Bayes estimator of a population parameter vector $\vec{\theta}$ or a function $f(\vec{\theta})$ of them can be understood as the expectation value of $\vec{\theta}$ or $f(\vec{\theta})$ over the posterior distribution of $\vec{\theta}$. Vice versa, we have shown in section 6.5 that the expectation value of $\vec{\theta}$ or $f(\vec{\theta})$ over the posterior distribution of $\vec{\theta}$ automatically minimizes the sample deviation from the theoretical value of $\vec{\theta}$ or $f(\vec{\theta})$.

6.6.3 Maximum Likelihood Estimator - Revisited

For the sake of simplicity, let p be the only population parameter in the following section. The maximum of the posterior probability density $P(p|k)$ can then be obtained by solving the equation

$$\frac{\partial P(p|k)}{\partial p} = 0. \tag{6.66}$$

Applying the Bayes formula and realizing that the denominator does not depend on p yields

$$\frac{\partial P(k|p)}{\partial p} \cdot P(p) + P(k|p) \cdot \frac{\partial P(p)}{\partial p} = 0 \tag{6.67}$$

if the denominator

$$\int_0^1 P(k|p) \cdot P(p)\, dp \neq 0. \tag{6.68}$$

If, as under the Bayes hypothesis, the prior probability density $P(p)$ does not depend on p, i.e., the partial derivative vanishes, then we immediately come up with

$$\frac{\partial P(k|p)}{\partial p} = 0. \tag{6.69}$$

This, however, is exactly the Maximum-Likelihood-Condition. Consequently, the maximum likelihood estimator favors that guess about the hypothetic population parameter p which maximizes its posterior probability density $P(p|k)$.

6.6.4 Maximum Likelihood versus Minimum Variance

Whereas the Bayes estimator of a population parameter θ chooses the expectation value of all hypothetic θ values over the posterior probability density $P(\theta|k)$ and thus minimizes the quadratic deviation of the estimates from the true parameter θ, the maximum likelihood estimator searches for the maximum of the posterior probability density $P(\theta|k)$ under the Bayes hypothesis of a uniform prior $P(\theta)$.

The conclusion that we are enforced to draw here is the following. If we are to estimate a certain population parameter from a finite sample, we first have to inquire about whether it is important to estimate this parameter precisely or whether our goal should be to come as close as possible to the true parameter value.

The first case means that we get penalized (with the same penalty) if we do not estimate the true parameter value correctly; i.e., it only counts whether we estimate correctly or not. In this case, we should preferably use the maximum likelihood estimator, which searches for the parameter value with the highest posterior probability density.

If we are, however, penalized proportionally to the squared distance that our estimate deviates from the true parameter value, then we should undoubtedly prefer the Bayes estimator. Since we, in almost all practical situation, are not confronted with the demand to determine a certain parameter precisely, but normally are required to find an estimate that is as close as possible to the true parameter value, we should prefer the Bayes estimator over the maximum likelihood estimator in those situations.

6.7 Summary

This chapter was devoted to present some statistical definitions, theorems, and techniques by which information about theoretical population parameters or functions of them can be inferred from finite samples.

After we had presented some basic definitions in section 6.1, we introduced the maximum likelihood method in section 6.2. In section 6.3, we defined another two estimator properties, while section 6.4 was dedicated to presenting some valuable theorems about statistical properties of maximum likelihood estimators in general.

We raised the question whether we cannot develop an estimator that has the desirable property to predict a population parameter or a set of them from a finite sample as precisely as possible. This question lead us to the formulation of the minimum variance principle and the resulting Bayes estimator in section 6.5. In that section, we also calculated the Bayes estimator of the probability of a coin in order to exemplify the minimum variance method. Finally, we presented the Bayes estimator of functions of population parameters, which will allow us to derive this desirable estimator of Shannon entropies in the following chapter.

In section 6.6, we investigated relations between the maximum likelihood and minimum variance estimators, and realized that the Bayes estimator of a population parameter θ always chooses the expectation value of the posterior distribution of θ, whereas the maximum likelihood estimator chooses its maximum under the Bayes hypothesis. In subsection 6.6.4, we stated our conclusion that Bayes estimators should be preferred over maximum likelihood estimators in situations where we want to approximate the theoretical population parameter as closely as possible by our estimates from finite samples.

Chapter 7

Bayes Estimator of the Shannon Entropy

7.1 Motivation

In the previous chapter, we have learned that Bayes estimators have to be preferred over maximum likelihood estimators in all cases where the goal of the considered estimation is to approximate the true but hidden parameter as close as possible. Since, in general, the difference between the maximum likelihood and the Bayes estimator vanishes for large sample sizes N, maximum likelihood estimators are often preferred due to their simple expressibility. In small samples, however, the application of maximum likelihood estimators becomes questionable, and should better be replaced by their Bayes counterparts.

The estimation of higher entropies from finite samples, such as DNA sequences, English texts, pieces of music, or time series generated by dynamical or stochastic processes, is a wide-spread method to analyze linguistic structures hidden in those sequences mentioned above. However, entropy estimates are often required in situations where sufficiently large samples are not available. Imagine, for example, we have realized significant differences of the 6-mer Shannon entropy between coding and noncoding pieces in eukaryotic DNA sequences. Our ultimate goal would then be to exploit this different statistical behavior to distinguish between coding and noncoding DNA.

The problem that we are confronted with in this example is to estimate the 6-mer Shannon entropy H_6 from sequences not much longer than 100 base pairs. In our previously introduced notation this means $M = 4^6 = 4096$ and $N \approx 100$. By imagining the size of

our 4096 dimensional state space, we immediately realize that the 100 sample points that we are given cannot at all guarantee a reliable estimate of any of the 4096 probabilities p_i. Moreover, we know that approximately 4000 dicodons do not at all appear in our DNA sequence.

Since the maximum likelihood estimator of the probability assigns those non-observed dicodons the probability 0 and the natural entropy estimator is defined as the entropy of those maximum likelihood estimates, we clearly realize why the natural entropy estimator systematically fails to deliver any reasonable estimate of the Shannon entropy if the sample size is small.

Aside from linguistic analyses, there is a wide spectrum of problems in physics where entropies and related statistics have to be estimated from finite samples the size of which is often extremely small [Wolpert & Wolf 1993].

The following section will be dedicated to deriving an exact expression for the Bayes estimator of the Shannon entropy.

7.2 Derivation

In this section, we will derive the estimator for the Shannon entropy

$$H(p_1, p_2, ..., p_M) = -\sum_{i=1}^{M} p_i \cdot \ln(p_i) \tag{7.1}$$

that minimizes the quadratic deviation from the true value of $H(p_1, p_2, ..., p_M)$ under the assumption of a uniform prior probability density.

According to eq. (6.50), the Bayes estimator of the Shannon entropy, which is a function of the probabilities $p_1, p_2, ..., p_M$, is given by

$$\hat{H}(\vec{k}) = \frac{\int H(\vec{p}) \cdot P(\vec{k}|\vec{p}) \cdot P(\vec{p})\, d\vec{p}}{\int P(\vec{k}|\vec{p}) \cdot P(\vec{p})\, d\vec{p}}, \tag{7.2}$$

where $\vec{p} \equiv (p_1, p_2, ..., p_M)$, $\vec{k} \equiv (k_1, k_2, ..., k_M)$, and the integrals are taken over the whole simplex given by $\sum_{i=1}^{M} p_i = 1$ and $p_i \geq 0$ for all $i = 1, 2, ..., M$.

By interchanging the integrals with the finite sum appearing in the numerator, we obtain

$$\hat{H}(\vec{k}) = \sum_{i=1}^{N} \frac{\int -p_i \cdot \ln(p_i) \cdot P(\vec{k}|\vec{p}) \cdot P(\vec{p})\, d\vec{p}}{\int P(\vec{k}|\vec{p}) \cdot P(\vec{p})\, d\vec{p}}. \tag{7.3}$$

Exploiting our results obtained in appendices H.1, H.2, and H.3, we yield

$$\int\limits_{simplex} -p_i \cdot \ln(p_i) \cdot P(\vec{k}|\vec{p}) \cdot P(\vec{p})\, d\vec{p} \tag{7.4}$$

$$= \int\limits_0^1 -p_i \cdot \ln(p_i) \cdot K(p_i; k_i, M, N)\, dp_i \tag{7.5}$$

$$= -\frac{N!}{k_i! \cdot (N - k_i + M - 2)!} \cdot \int\limits_0^1 \ln(p_i) \cdot p_i^{k_i+1} \cdot (1 - p_i)^{N-k_i}\, dp_i \tag{7.6}$$

$$= -\frac{N!}{k_i! \cdot (N - k_i + M - 2)!} \cdot J(k_i + 1, N - k_i + M - 2) \tag{7.7}$$

$$= \frac{N!}{k_i! \cdot (N - k_i + M - 2)!} \cdot \frac{(k_i + 1)! \cdot (N - k_i + M - 2)!}{(N + M)!} \cdot \sum_{j=k_i+2}^{N+M} \frac{1}{j}. \tag{7.8}$$

Analogously we proceed and obtain

$$\int\limits_{simplex} P(\vec{k}|\vec{p}) \cdot P(\vec{p})\, d\vec{p} = \int\limits_0^1 K(p_i; k_i, M, N)\, dp_i \tag{7.9}$$

$$= \frac{N!}{k_i! \cdot (N - k_i + M - 2)!} \cdot \int\limits_0^1 p_i^{k_i} \cdot (1 - p_i)^{N-k_i+M-2}\, dp_i \tag{7.10}$$

$$= \frac{N!}{k_i! \cdot (N - k_i + M - 2)!} \cdot \frac{k_i! \cdot (N - k_i + M - 2)!}{(N + M - 1)!}. \tag{7.11}$$

Eventually, dividing both terms leads to

$$\frac{\int -p_i \cdot \ln(p_i) \cdot P(\vec{k}|\vec{p}) \cdot P(\vec{p})\, d\vec{p}}{\int P(\vec{k}|\vec{p}) \cdot P(\vec{p})\, d\vec{p}} = \frac{k_i + 1}{N + M} \sum_{j=k_i+2}^{N+M} \frac{1}{j} \tag{7.12}$$

for all $i = 1, 2, ..., M$.

If we define

$$\hat{H}_i(k_i) \equiv \frac{k_i + 1}{N + M} \sum_{j=k_i+2}^{N+M} \frac{1}{j}, \tag{7.13}$$

we obtain the Bayes estimator of the Shannon entropy as

$$\hat{H}(\vec{k}) = \sum_{i=1}^{M} \hat{H}_i(k_i). \tag{7.14}$$

7.3 Properties

This section, which is dedicated to the discussion of and the comparison between some statistical properties of the Bayes entropy estimator and the so called natural entropy estimator, will reveal striking advantages of using the Bayes estimator instead of the natural one if the available sample size N is small compared to the number M of possible states in our system of consideration.

These statistical properties of the Bayes entropy estimator, which are perfectly exemplified in [Wolpert & Wolf 1993] for the binomial case, unambiguously suggest to prefer the Bayes estimator over the natural one in cases where N is small compared to M.

Since the algorithm for calculating the Bayes entropy estimator only consists of calculating finite harmonic sums, it might be even quicker than the natural estimator algorithm, where logarithms of relative frequencies have to be taken. Hence, we highly recommend to substitute all subroutines calculating the natural entropy estimator by subroutines that calculate the Bayes entropy estimator from a given finite sample.

Let us finally outline a relation between the

- Bayes estimator of the Shannon entropy $H(p_1, p_2, ..., p_M)$

and the

- Shannon entropy H of the Bayes estimator of the probabilities $p_1, p_2, ..., p_M$.

In section D.2, we learned that, due to the non-monotonicity of the function $H(p) = -p \cdot \ln(p) - (1-p) \cdot \ln(1-p)$, the maximum likelihood estimator of the Shannon entropy is not equal to the Shannon entropy H of the maximum likelihood estimator of the probability p. The same statement holds for the Bayes estimator of the Shannon entropy; i.e., the Shannon entropy of the Bayes estimators of the probabilities $p_1, p_2, ..., p_M$,

$$\hat{H}_G \equiv -\sum_{i=1}^{M} \frac{k_i + 1}{N + M} \cdot \ln\left(\frac{k_i + 1}{N + M}\right), \tag{7.15}$$

is not the Bayes estimator of the Shannon entropy [Wolpert & Wolf 1993].

In the following, we will, however, derive a relation between these two entropy estimators.

Let us, for this reason, approximate the finite harmonic sum $\sum_{j=k_i+2}^{N+M} \frac{1}{j}$ by the corre-

sponding definite integrals:

$$\int_{k_i+1}^{N+M} \frac{1}{x}\,dx > \sum_{j=k_i+2}^{N+M} \frac{1}{j} > \int_{k_i+2}^{N+M+1} \frac{1}{x}\,dx. \tag{7.16}$$

Multiplying the upper bound approximation

$$\int_{k_i+1}^{N+M} \frac{1}{x}\,dx = -\ln\left(\frac{k_i+1}{N+M}\right) \tag{7.17}$$

by $\dfrac{k_i+1}{N+M}$ and summing over all terms yields

$$\hat{H} < -\sum_{i=1}^{M} \frac{k_i+1}{N+M} \cdot \ln\left(\frac{k_i+1}{N+M}\right). \tag{7.18}$$

Now we see that the right hand side of this inequality is the Shannon entropy of the Bayes probability estimator, which we briefly call Grassberger's entropy estimator according to [Grassberger 1994].

The difference between the the Bayes entropy estimator and Grassberger's entropy estimator is equal to the error of replacing the finite harmonic sums by the corresponding integrals.

Considering the asymptotic behavior of the Bayes entropy estimator, we can state that

$$\hat{H}(\vec{k}) = \sum_{i=1}^{M} \frac{k_i+1}{N+M} \sum_{j=k_i+2}^{N+M} \frac{1}{j} \to -\sum_{i=1}^{M} \frac{k_i}{N} \cdot \ln\left(\frac{k_i}{N}\right) \tag{7.19}$$

for $N \to \infty$. This means that not only the natural entropy estimator, but also the Bayes entropy estimator is consistent [Wolpert & Wolf 1993].

Let us in the remainder of this section study some simulations that are to reveal statistical properties of the three entropy estimators introduced above. In order to choose adequate parameters of our simulations, let us first study the following biological situation, which is typical for raising the problem of estimating entropies from finite samples.

While analyzing biological symbol sequences, we are almost always confronted with situations in which the sample size N can not be guaranteed to be significantly greater than the number M of possibly occurring words, since M is blowing up exponentially with the word-length n, namely $M = \lambda^n$.

The following simulations are motivated by the task to estimate 5-mer entropies of several DNA sequences. Since $M = \lambda^n = 1024$, we need sequences of at least 10,000 base

pairs to obtain reliable entropy estimates by the natural entropy estimator. However, since biological tasks often demand to analyze shorter sequences as well, we are confronted with estimating entropies from samples in which $N \approx M$. Hence, we chose $N = 5000$, $N = 2000$, and $N = 1000$ for our simulations.

Another parameter that we must specify is the probability vector $\vec{p} = (p_1, p_2, ..., p_{1024})$. We know that, if we compare two estimators A and B, A can be on average much closer to the true population parameter than B for some particular $\vec{p}$, but then B can be more accurate than A for some other $\vec{p}$. Since each estimator has its favorite $\vec{p}$, we have to choose this probability vector very carefully in order to make sure that our following analysis does not become useless just because we simulate in a point on the $\vec{p}$-simplex that is biologically irrelevant.

As representation of a typical probability vector $\vec{p}$, we choose the the 5-mer distribution derived from the 2,181,032 base pairs long sequence of the *Caenorhabditis elegans* chromosome III. The rank ordered distribution of these 1024 probabilities is displayed in Figure 7.1 as a double-logarithmic plot.

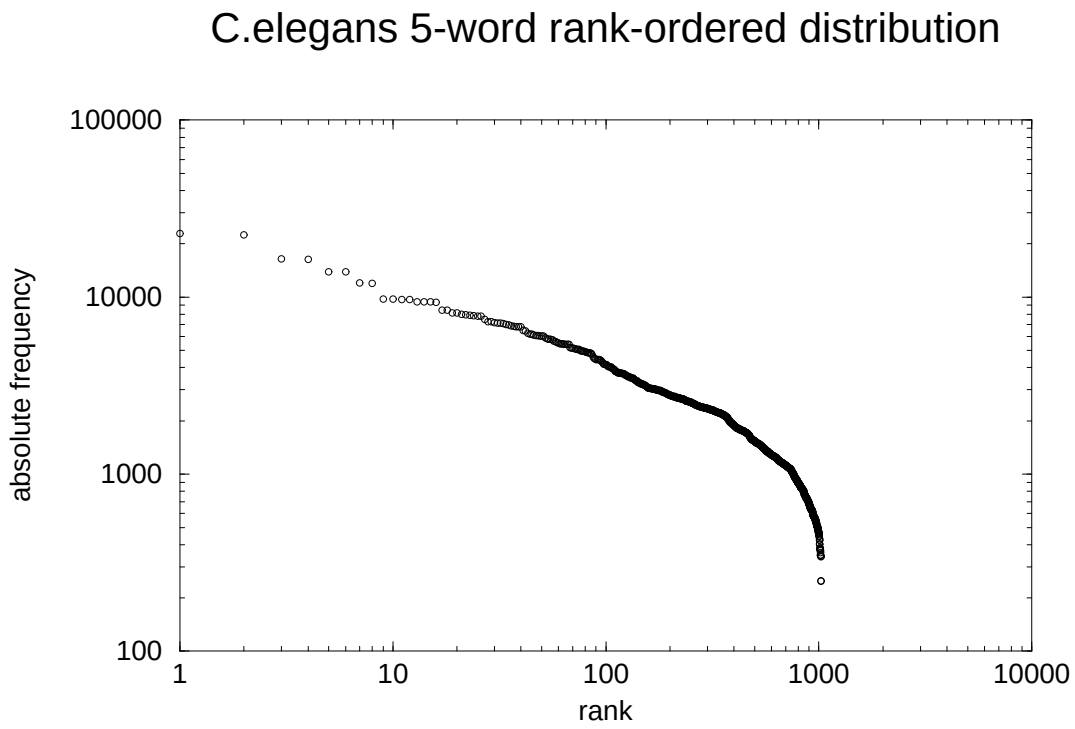


Figure 7.1: Rank-ordered 5-mer distribution of the Caenorhabditis elegans chromosome III. We estimate the 5-mer probabilities as the corresponding relative frequencies of all overlapping 5-mers in the 2,181,032 base pairs long sequence of the Caenorhabditis elegans chromosome III.

This probability vector is now chosen for the following three simulations corresponding to $N = 5000$, $N = 2000$, and $N = 1000$. According to these 1024 given probabilities, a sequence of N 5-mers is randomly composed. Then we apply three different entropy estimators, namely the natural entropy estimator, the Bayes entropy estimator, and Grassberger's entropy estimator [Grassberger 1994] given by eq. (7.15), to this sequence with the finite length N.

Since we know the true population parameters $p_1, p_2, ..., p_{1024}$, we can as well calculate the theoretical 5-mer entropy. The differences between the estimated entropy values and the theoretical entropy value define our three random variables, which we call *entropy estimate deviation from true*. Now we repeat the experiment of generating a sequence 10,000 times and thus obtain a time series where we denote the time step by *number of simulation*.

Figures 7.2 – 7.7 demonstrate the accuracy and reliability of the Bayes estimator over the natural estimator for the *C.elegans* probability vector $\vec{p}$.

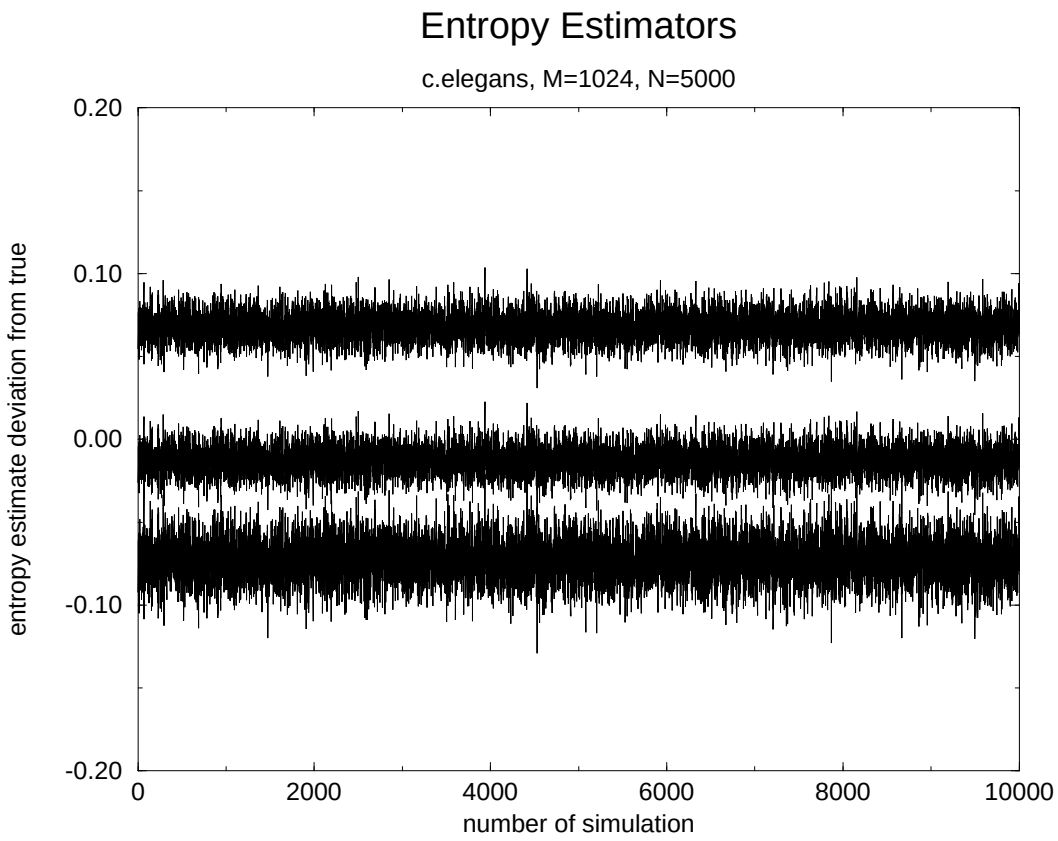


Figure 7.2: Comparison of three entropy estimators for $M = 1024$, $N = 5000$, and $\vec{p}$ derived from C.elegans. The upper curve corresponds to Grassberger's estimator, the curve in the middle to the Bayes entropy estimator, and the lower curve to the natural estimator of the Shannon entropy H_5. We see that the natural estimator and the Bayes estimator systematically underestimate the theoretical Shannon entropy, whereas Grassberger's estimator systematically overestimates the true entropy value for our chosen $\vec{p}$. However, the biases of the natural estimator as well as Grassberger's estimator are so strong that there is even not a single event among our 10,000 trials where their estimates come close to the theoretical value, i.e., their biases are larger than their standard deviations. The Bayes estimator, on the other hand, is almost unbiased compared to its variance. Please note that our length correction formula presented in chapter 6 can almost perfectly correct the bias of the natural estimator. Hence, the significant difference between the quality of the natural and the Bayes estimator is not given by their biases, but by their variances, which we will study in the next figure.

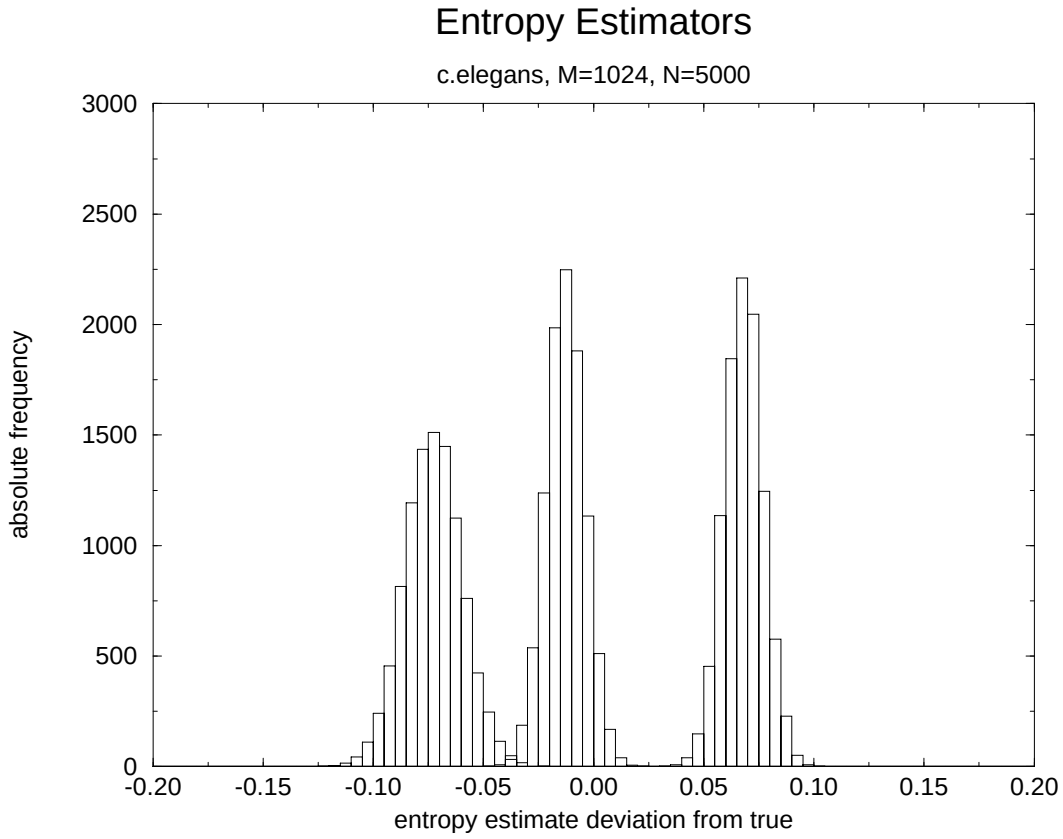


Figure 7.3: Comparison of three entropy estimators for $M = 1024$, $N = 5000$, and $\vec{p}$ derived from C.elegans. This figure displays the three histograms corresponding to Figure 5.2. We see that the variances of Grassberger's estimator and the Bayes estimator are of comparable size, whereas the fluctuations of the natural entropy estimator are slightly larger. This implies to prefer the Bayes estimator over the natural one even though we can almost perfectly correct the significant bias of the natural entropy estimator by our length correction formula derived in chapter 6.

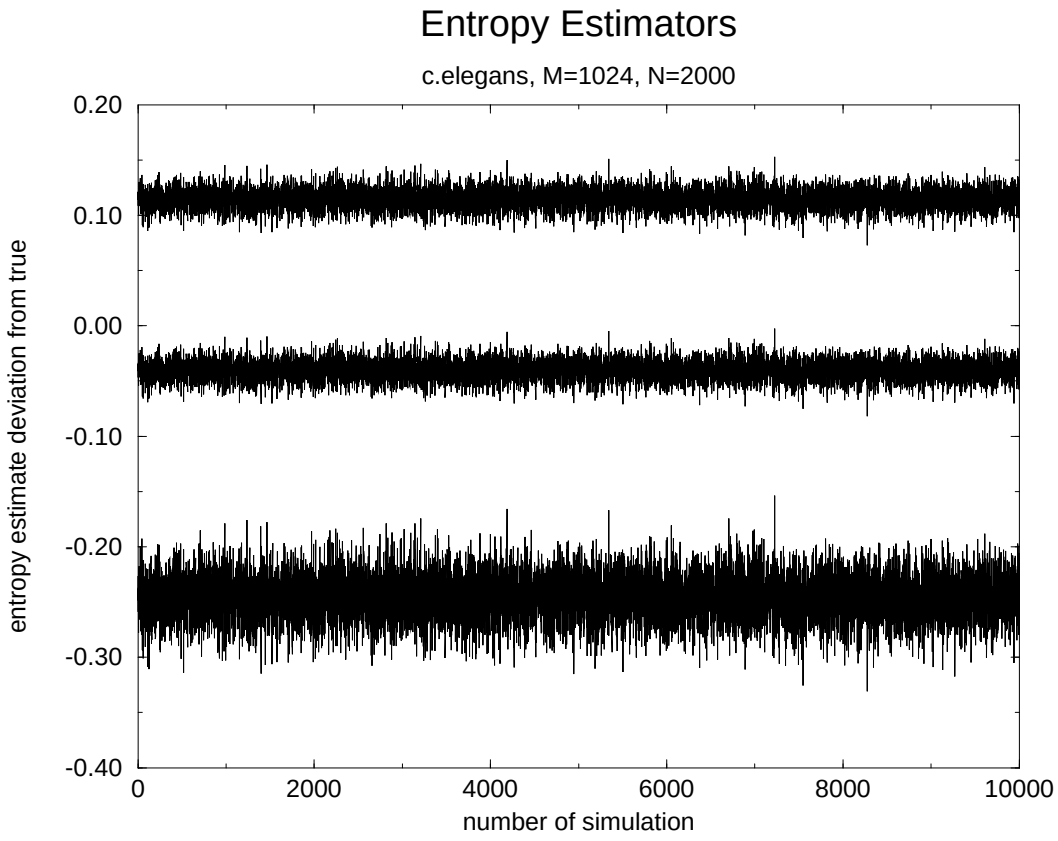


Figure 7.4: Comparison of three entropy estimators for $M = 1024$, $N = 2000$, and $\vec{p}$ derived from C.elegans. Again, the upper curve reflects Grassberger's estimates, the curve in the middle the Bayes ones, and the lower curve the natural estimates of the Shannon entropy. In comparison with our previous two figures, we see that the biases of all three estimators become larger as the sample size N, i.e., the sequence length, decreases. Again, the natural and the Bayes estimator systematically underestimate the theoretical Shannon entropy, whereas Grassberger's entropy estimator systematically overestimates the theoretical value. We note again that the bias of the Bayes estimator is significantly smaller than the biases of its competitors. However, please keep in mind that we can correct the strong bias of the natural estimator of the Shannon entropy.

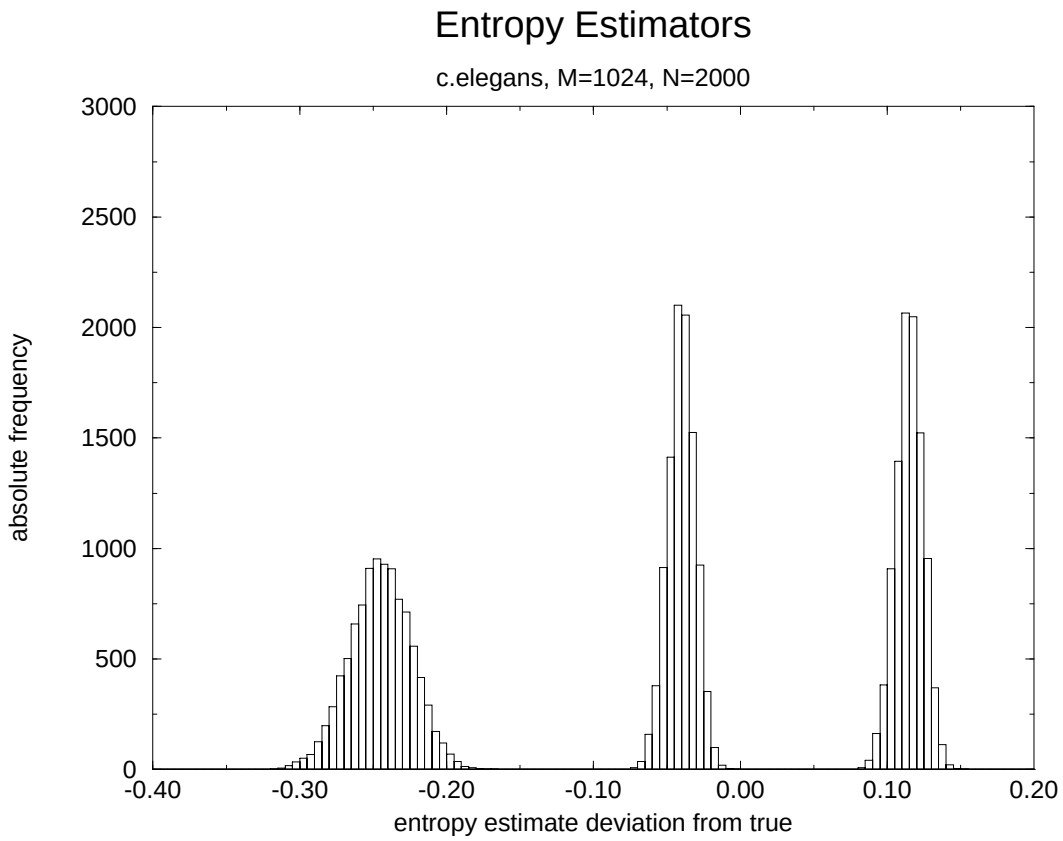


Figure 7.5: Comparison of three entropy estimators for $M = 1024$, $N = 2000$, and $\vec{p}$ derived from C.elegans. These histograms exhibit that not only the biases, but also the variances grow as N decreases. Since biases of estimators can often be corrected, the comparatively small variance of the Bayes estimator is its main advantage over the natural estimator of the Shannon entropy.

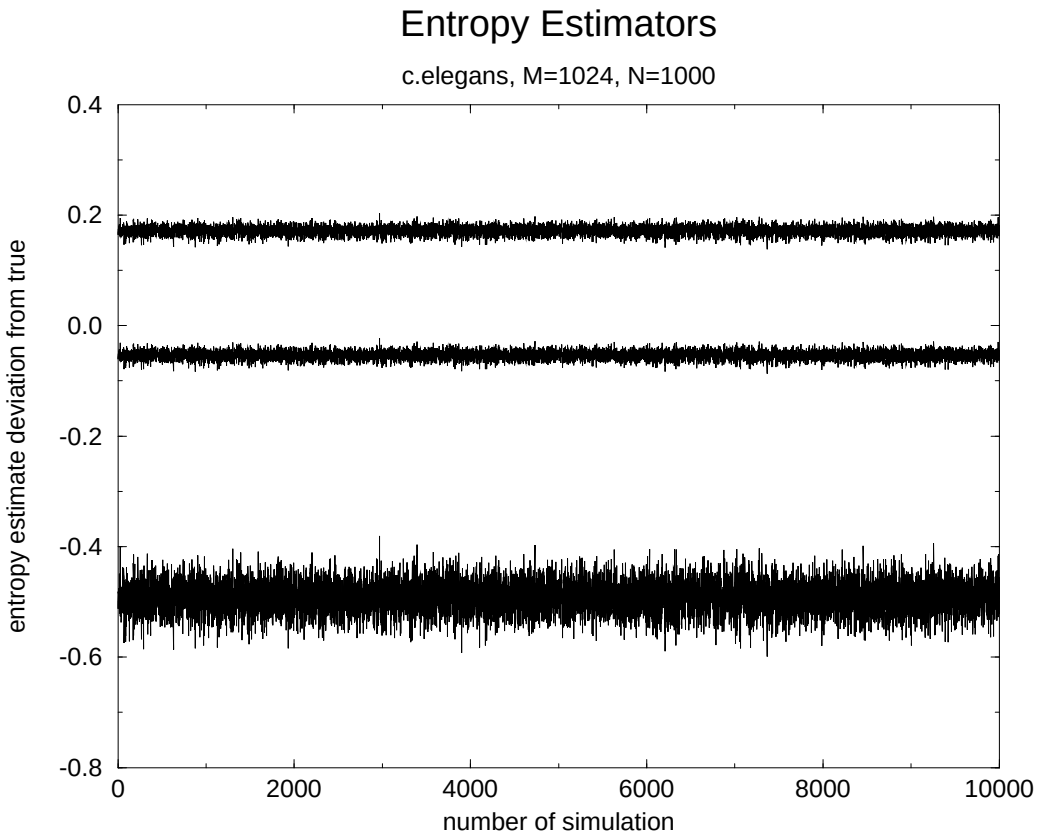


Figure 7.6: Comparison of three entropy estimators for $M = 1024$, $N = 1000$, and $\vec{p}$ derived from C.elegans. Grassberger's estimator again yields the highest estimates followed by the Bayes and then by the natural entropy estimator. All biases become larger with a decreasing sequence length N, as a comparison with Figures 5.4 and 5.5 reveals. In particular, the bias of the natural estimator becomes outstandingly large. However, recall that our length correction formula can correct it even in this situation where $N < M$.

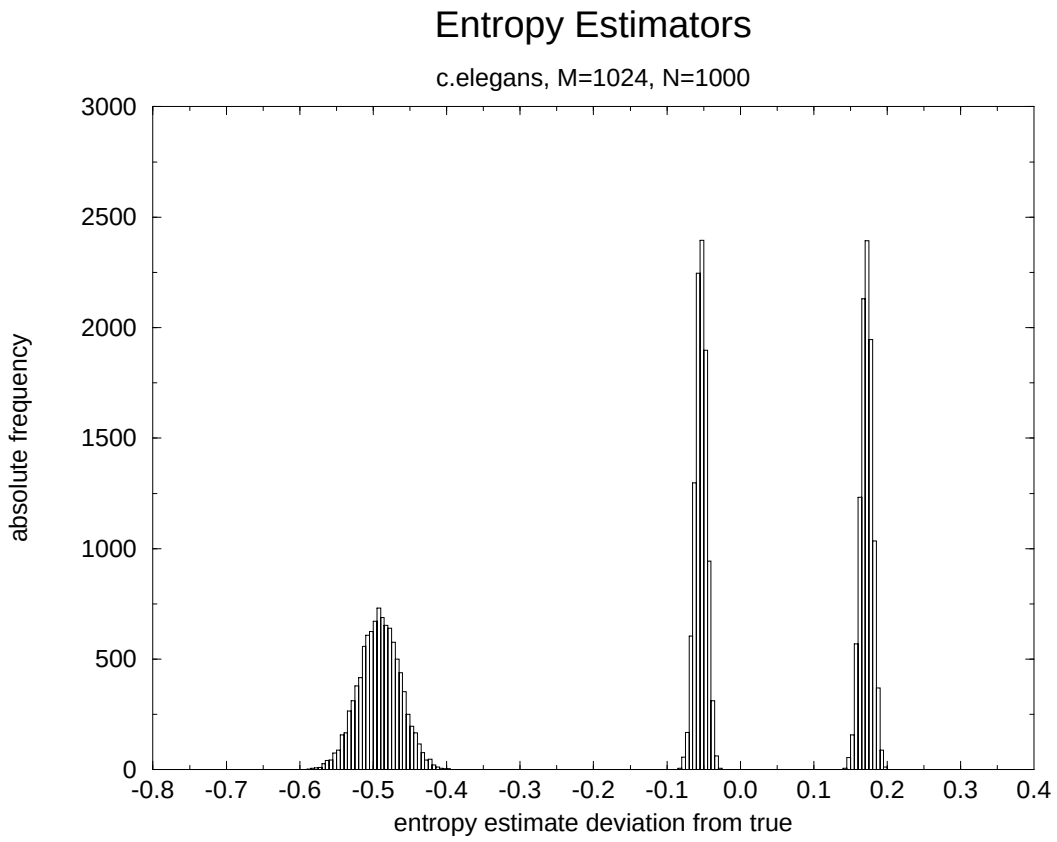


Figure 7.7: Comparison of three entropy estimators for $M = 1024$, $N = 1000$, and $\vec{p}$ derived from C.elegans. These histograms show that the variances of the Bayes estimator is comparable to the fluctuations of Grassberger's entropy estimator, which both are significantly smaller than the variance of the natural entropy estimates. Since we realize that the ratio of the variance of the natural entropy estimates and the variance of the Bayes estimates becomes larger for decreasing N, we highly recommend to substitute the natural estimator by the Bayes estimator of the Shannon entropy in order to get more reliable estimates of the theoretical Shannon entropy in samples where N cannot be guaranteed to be much greater than M.

Even though we could reasonably motivate the selection of the probability vector $\vec{p}$ in our previous simulations, this does not free us from presenting more simulations with different $\vec{p}$. However, instead of going this way and filling the next pages with pictures similar to the previous ones, we will design the following experiment.

Instead of fixing the probability vector arbitrarily to any $\vec{p}$, we also randomly pick this vector from any point of the simplex given by the normalization constraint of the p_i. In order not to overreach any probability vector, we assume an uniform probability density of the vectors $\vec{p}$. Please note that this assumption does not mean that we simulate with equidistributed p_i. An uniform density on the $\vec{p}$-simplex rather means that the equidistributed vector $\vec{p} = (\frac{1}{M}, \frac{1}{M}, ..., \frac{1}{M})$ has the same right to get selected as the c.elegans probability vector displayed in Figure 7.1 or any other imaginable probability vector $\vec{p}$. In mathematical terms, all probability vectors have the same probability $P(\vec{p})$ to get selected for our simulation experiment.

The outcome of these simulations will then be an average distribution of the three estimates introduced above. Remember that our original task was to evaluate our three entropy estimators in terms of their statistical properties. What we have shown with our previous simulations was that the Bayes estimator should be preferred over the natural estimator of the Shannon entropy in the particular case of rank-ordered distributions similar to Figure 7.1.

Although this rank ordered distribution might count as a typical example representing the 5-mer probabilities in eukaryotic DNA, we now ask the question what the distributions of the entropy estimates look like in general, i.e., for randomly chosen $\vec{p}$-vectors uniformly spread upon the simplex. In our following simulation, we choose $N = 1000$, since, due to their consistency, all estimators can be shown to converge to each other for increasing N.

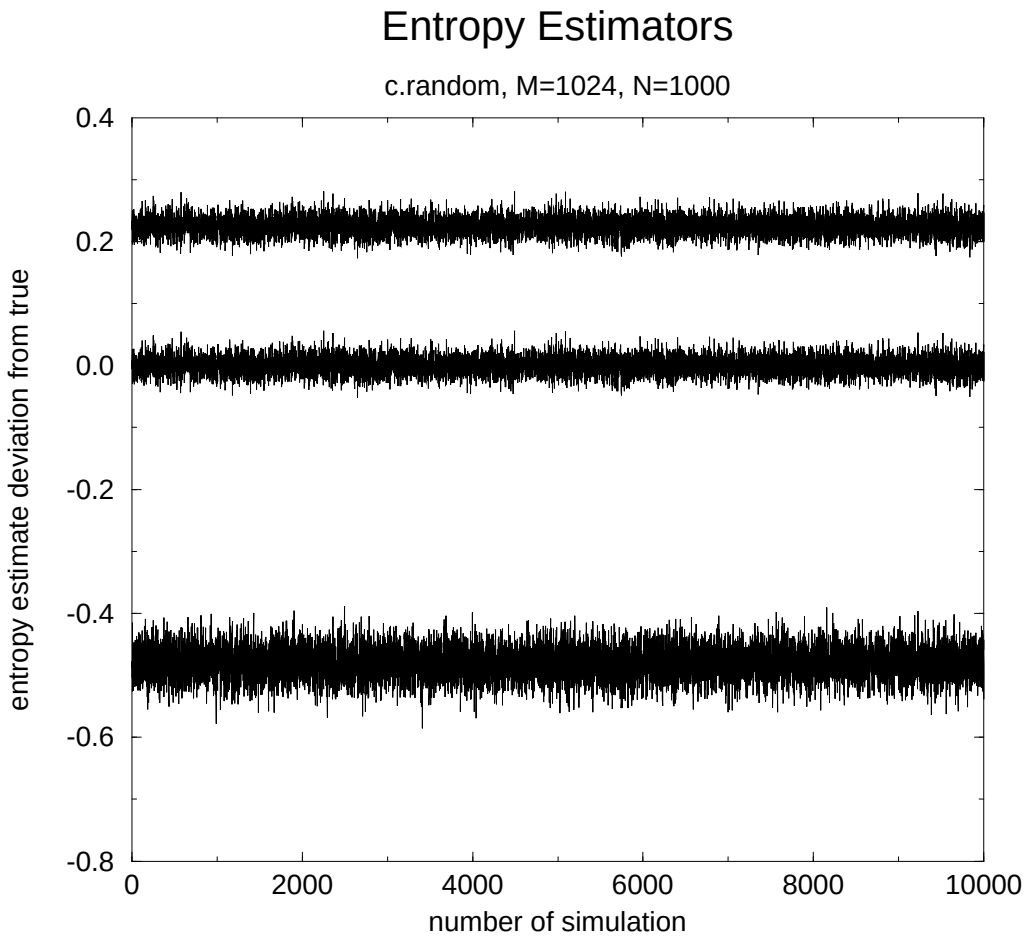


Figure 7.8: Comparison of three entropy estimators for $M = 1024$, $N = 1000$, and $\vec{p}$ randomly selected. From top to bottom, we display the estimates of Grassberger's estimator, the Bayes estimator, and the natural estimator of the 5-mer Shannon entropy, respectively. Please note that a symmetric distribution of the estimated entropies around the theoretical values does not imply that the corresponding estimator is unbiased. It rather means that the estimator is sometimes positively biased and sometimes negatively such that the average bias just vanishes. Applied to the Bayes estimator, this means that the cases where we systematically overestimate the theoretical Shannon entropy yield the same absolute deviation from true as those cases in which we systematically underestimate the true value. Hence, two effects simultaneously contribute to the variance of the entropy estimates: the average variance over all fixed vectors $\vec{p}$ and the fluctuations of the biases for different $\vec{p}$.

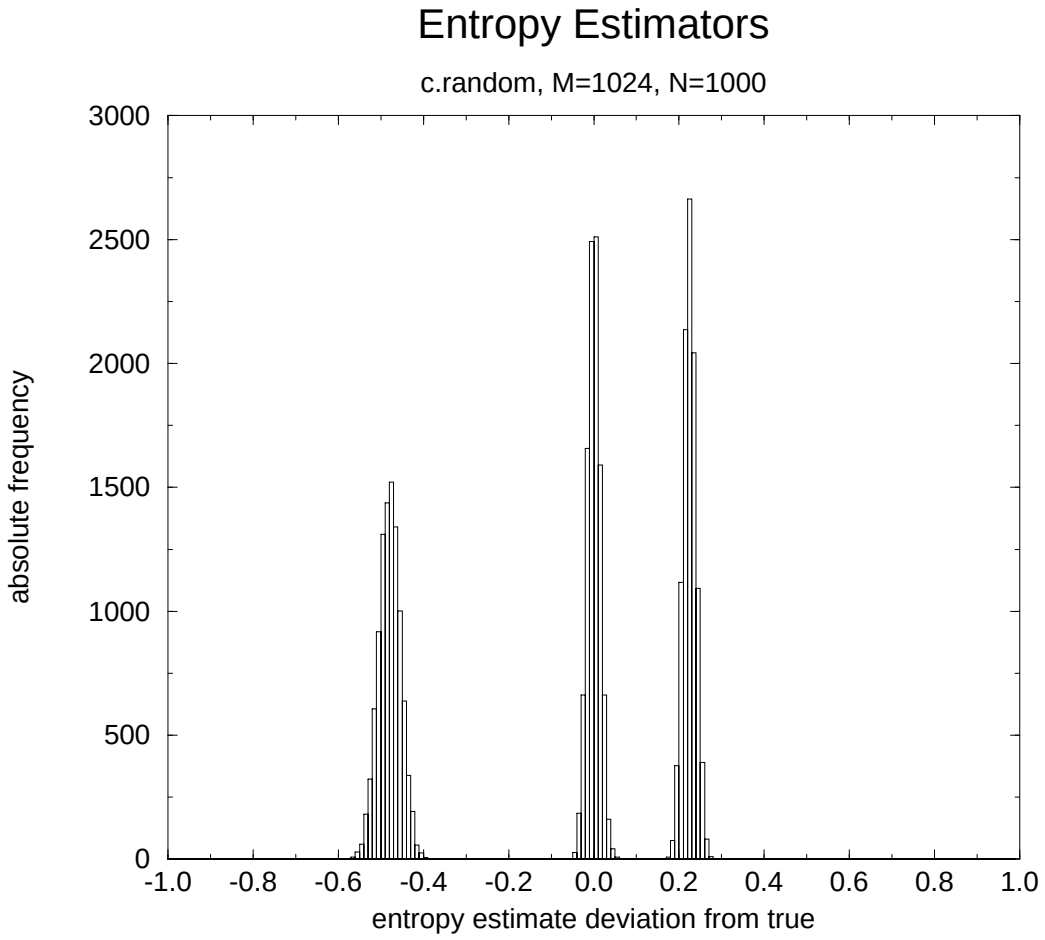


Figure 7.9: Comparison of three entropy estimators for $M = 1024$, $N = 1000$, and $\vec{p}$ randomly selected. Studying these histograms belonging to Figure 5.8, we realize that the natural estimator is systematically underestimated on average. This result is not at all surprising, since we can show in chapter 6 that the natural entropy estimator is always biased independently on the underlying vector $\vec{p}$. In chapter 6, we also derive a length correction formula, which perfectly corresponds to this simulation. Grassberger's estimator, on the other hand, systematically overestimates the theoretical entropy for all 10,000 vectors $\vec{p}$ generated in this simulation. Finally, the Bayes estimator does not show a systematic deviation from the theoretical values on average, which does however not mean that this estimator is unbiased (c.f. Figure 5.8). Regarding the variance of the estimates, the Bayes estimator has obviously to be preferred over the natural one independently on the availability of powerful bias correction formulae.

7.4 Summary

In this chapter we derived the Bayes estimator of the Shannon entropy,

$$\hat{H}(\vec{k}) = \sum_{i=1}^{M} \frac{k_i + 1}{N + M} \sum_{j=k_i+2}^{N+M} \frac{1}{j}. \tag{7.20}$$

In section 7.1, we motivated the demand for an entropy estimator that has the desirable property to be very close to the theoretical Shannon entropy that we intend to determine. Since Bayes estimates have, in general, a smaller mean quadratic deviation from the theoretical population parameter than the estimates of any other estimator, we tried to derive the Bayes estimator of the Shannon entropy.

Section 7.2 was entirely devoted to deriving the Bayes estimator of the Shannon entropy under the prior assumption of a uniform distribution on the simplex of the underlying probability vectors. Please remember that this uniform prior does not mean that we prefer equidistributed probability vectors. Conversely, the assumption of a uniform prior is identical to the demand of equal rights for all possible probability vectors.

In section 7.3, we compared the statistical properties of the natural entropy estimator, Grassberger's entropy estimator, which is defined as the entropy function of the Laplace probability estimator, and our in section 7.2 derived Bayes entropy estimator.

These studies unambiguously revealed the strength of the Bayes entropy estimator, which were its small bias and, more importantly, the small fluctuations of its estimates. Therefore, we recommend everybody who is confronted with the problem of estimating entropies from finite samples and has no idea about the underlying probability distribution to substitute all procedures that calculate the natural entropy estimates by subroutines computing their Bayes counterparts.

Additionally, we could relate Grassberger's estimator to the Bayes estimator of the Shannon entropy by showing that we obtain Grassberger's estimator if we substitute the finite sum in eq. (7.20) by the corresponding definite integral. This finding finally allowed us to present the inequality that Grassberger's estimator is always greater than the Bayes estimator of the Shannon entropy.

Chapter 8

Bayes Estimators of Generalized Entropies

One practical and common problem in the analysis of experimental data is the estimation of probabilities or functions of probabilities from finite sets of observed data. The finite size of data sets can lead to serious systematic and statistical errors in numerical estimates. In this chapter we address the problem of estimating generalized entropies from finite samples, and we derive the Bayes estimator of the order-q Tsallis entropy, including the order-1 (i.e., the Shannon) entropy, under the assumption of a uniform prior probability density. The Bayes estimator yields, in general, the smallest mean-quadratic deviation from the true parameter as compared to any other estimator. Exploiting the functional relationship between the Tsallis entropy H_q and the Rényi entropy K_q, we use the Bayes estimator of H_q to estimate K_q. We compare the Bayes estimators with the frequency-count estimators for H_q and K_q. We find by numerical simulations that the Bayes estimator reduces statistical errors of order-q entropy estimates for Bernoulli as well as for higher-order Markov processes derived from the complete genome of the prokaryote *Haemophilus influenzae*.

8.1 Introduction

Here we address the estimation of these entropies from a finite set of experimental data. Under the assumption of a stationary process generating the data, the data set is composed out of N data points chosen from M possible different outcomes. The problem that

arises when entropies are to be estimated from these finite data sets is that the probabilities are a priori unknown. Naively replacing these probabilities by the sampled relative frequencies produces large statistical and systematic deviations of estimates from the true value [Basharin 1959, Harris 1975]. This problem becomes severe when the number of data points N is in the order of magnitude of the number of different states M, which occurs in many practical applications, e.g. in the estimations of correlations and dimensions. In such cases, the choice of an estimator with small deviations from the true value becomes important. Several different estimators have thus been developed, mainly devoted to the estimation of the Shannon entropy [22-27]. Specific estimators for the Rényi entropy and for the dimensions associated to them have also been derived, as well as for upper bounds on entropy estimates [Grassberger 1988, Schürmann & Grassberger 1996].

While one can, in principle, calculate the systematic errors arising from frequency-counts, less fluctuating entropy estimates can only be obtained by employing a different entropy estimator. The estimator which possesses the optimal property to minimize the mean-quadratic deviation of the estimate from the true value, subject to a certain prior assumption, is customarily referred to as the Bayes estimator. In this work, we derive the Bayes estimator of the Tsallis entropy and discuss its statistical properties. We then exploit this Bayes estimator to measure the Rényi entropy.

8.2 Tsallis Entropy Estimator

In this section we focus upon the first task stated in the preceding section, by deriving the Bayes estimator of the generalized Tsallis entropy H_q. The total number of symbols available in a sample for the estimation is given by $N = \sum_{i=1}^{M} N_i$. Let further $P(\vec{N}|\vec{p}) = N![\prod_{i=1}^{M} p_i^{N_i}/N_i!]$ be the underlying conditional probability distribution to obtain the (multinomially distributed) observable-vector $\vec{N}$ with components N_i. Finally, $Q(\vec{p})$ denotes the prior probability density of the probability-vector $\vec{p}$. It satisfies the constraint $\int_{\mathcal{S}} d\vec{p}\, Q(\vec{p}) = 1$ where the integration extends over the whole simplex $\mathcal{S} \equiv \{\vec{p} \mid \forall i\ p_i \geq 0, \sum_{i=1}^{M} p_i = 1\}$. Then the Bayes estimator of H_q reads as

$$\widehat{H_q}(\vec{N}) = \frac{1}{W(\vec{N})} \int_{\mathcal{S}} d\vec{p}\, H_q(\vec{p})\ P(\vec{N}|\vec{p})\ Q(\vec{p}) \tag{8.1}$$

where the normalization constant is given by

$$W(\vec{N}) = \int_{\mathcal{S}} d\vec{p}\, P(\vec{N}|\vec{p})\ Q(\vec{p}). \tag{8.2}$$

According to the Bayes theorem, $P(\vec{p}|\vec{N}) = P(\vec{N}|\vec{p})Q(\vec{p})/Q(\vec{N})$. Thus equation (8.1) is equivalent to $\widehat{H_q}(\vec{N}) = \int_{\mathcal{S}} d\vec{p}\, H_q(\vec{p}) P(\vec{p}|\vec{N})$, which is the average of $H_q(\vec{p})$ over the posterior distribution $P(\vec{p}|\vec{N})$.

In what follows, we will derive the Bayes estimator of H_q under the assumption of a uniform prior probability density $Q(\vec{p}) = \text{const}$. That is to say, we regard all possible probability-vectors $\vec{p} \in \mathcal{S}$ to be relevant.

If we write down the Bayes estimator of H_q as

$$\widehat{H_q}(\vec{N}) = \frac{1}{\ln 2}\frac{1}{1-q}\Big[\widehat{Z_q}(\vec{N}) - 1\Big] \quad \text{with} \quad \widehat{Z_q}(\vec{N}) = \frac{1}{W(\vec{N})}\int_{\mathcal{S}} d\vec{p}\, \sum_{i=1}^{M} p_i^q\; P(\vec{N}|\vec{p}) \tag{8.3}$$

then it can be seen that the derivation of $\widehat{H_q}$ reduces to the derivation of the Bayes estimator of the partition function Z_q. The normalization constant W and the quantity W' (see later) will be evaluated in appendix B. Interchanging the integral with the finite sum, $\widehat{Z_q}$ may be cast into the form

$$\widehat{Z_q}(\vec{N}) = \frac{\Gamma(N+M)}{\prod_{j=1}^{M}\Gamma(N_j+1)} \times \left\{ \sum_{i=1}^{M} \int_{\mathcal{S}} \prod_{j=1}^{M} dp_j\; p_j^{(N_j+\delta_{ij}q)} \right\}. \tag{8.4}$$

Integrating over $M-1$ of the M components, we obtain

$$\widehat{Z_q}(\vec{N}) = \frac{\Gamma(N+M)}{\prod_{j=1}^{M}\Gamma(N_j+1)} \times \left\{ \int_{p_i=0}^{1} dp_i\; W'(p_i;\vec{N})\; p_i^q \right\}. \tag{8.5}$$

Evaluating the remaining integral, we arrive at

$$\widehat{Z_q}(\vec{N}) = \frac{\Gamma(N+M)}{\Gamma(N+M+q)} \times \left[\sum_{i=1}^{M} \frac{\Gamma(N_i+1+q)}{\Gamma(N_i+1)} \right] \tag{8.6}$$

and thus, composing all above expressions, we eventually obtain

$$\widehat{H_q}(\vec{N}) = \frac{1}{\ln 2}\frac{1}{1-q}\left\{ \frac{\Gamma(N+M)}{\Gamma(N+M+q)} \times \left[\sum_{i=1}^{M} \frac{\Gamma(N_i+1+q)}{\Gamma(N_i+1)} \right] - 1 \right\}. \tag{8.7}$$

Expression (8.7) constitutes a central result of this work: the Bayes estimator of the Tsallis entropy of order q. To illustrate the differences between the fluctuations of the Bayes estimator and the frequency-count estimator of H_q, in the following we will simplify expression (8.7) for the special cases $q = 1$ and $q = 2$. The motivation for this parameter choice stems from the following. We recall that H_q is indeed a generalization of H, providing upper and lower bounds for the Shannon entropy. As such, we wish to make contact with the Bayes estimator $\widehat{H}$ for the Shannon entropy. This is realized in the limit $q \to 1$. The second

example, $q = 2$, plays an important role in the statistical analysis of non-linear dynamical systems. Here $q = 2$ gives rise to quantities like the correlation dimension and the second-order Kolmogorov entropy (see, e.g., [Ruelle 1989] and references therein) as well as a generalization of the mutual information which preserves positivity [Pompe 1993].

To obtain $\widehat{H_1}$, we introduce the auxiliary function

$$F(q) = \left[\sum_{i=1}^{M} \frac{\Gamma\left(N_i + 1 + q\right]}{\Gamma(N_i + 1)}\right] \Big/ \Gamma\left(N + M + q\right). \tag{8.8}$$

This will become useful due to the necessary consideration of the limit $q \to 1$, since expression (8.7) is not defined otherwise. Introducing $\alpha_q = N_i + 1 + q$ and $\beta_q = N + M + q$, we may write at the limit point

$$\widehat{H_1}(\vec{N}) = \lim_{q \to 1} \hat{H}_q(\vec{N}) = -\frac{\Gamma(\beta_0)}{\ln 2} \left.\frac{\partial F\left(q\right)}{\partial q}\right|_{q=1} \tag{8.9}$$

where

$$\frac{\partial F\left(q\right)}{\partial q} = \sum_{i=1}^{M} \left\{ \frac{\Gamma\left(\alpha_q\right)}{\Gamma(\alpha_0)\Gamma\left(\beta_q\right)} \Big(\psi^{(1)}(\alpha_q) - \psi^{(1)}(\beta_q) \Big) \right\}. \tag{8.10}$$

Here $\psi^{(1)}(z) = d \ln \Gamma(z)/dz$ is the Digamma-function. Since α_1 and β_1 are integers, we may express $\psi^{(1)}(z)$ in terms of the finite harmonic sum $\psi^{(1)}(z) = \sum_{l=1}^{z-1} 1/l - E_c$, with $E_c = \lim_{R \to \infty} \left(\sum_{r=1}^{R} 1/r - \ln R \right)$ being Euler's constant. Inserting this expression into equation (8.10), we get

$$\frac{\partial F\left(1\right)}{\partial q} = -\sum_{i=1}^{M} \left\{ \frac{\Gamma(\alpha_1)}{\Gamma(\alpha_0)\Gamma(\beta_1)} \left(\sum_{l=\alpha_1}^{\beta_0} \frac{1}{l} \right) \right\} \tag{8.11}$$

and hence we obtain

$$\widehat{H_1}(\vec{N}) = \frac{1}{\ln 2} \left[\sum_{i=1}^{M} \frac{N_i + 1}{N + M} \left(\sum_{l=N_i+2}^{N+M} \frac{1}{l} \right) \right]. \tag{8.12}$$

Equation (8.12) defines the Bayes estimator of the order-1 Tsallis entropy under a uniform prior probability density. Comparing the above expression with results derived in [Wolpert & Wolf 1995] and [Grosse 1996], we verify the consistency of expression (8.7) in the limit $q \to 1$. That is, the Bayes estimator of the order-1 Tsallis entropy is identical to the Bayes estimator of the Shannon entropy: $\widehat{H_1} \equiv \hat{H}$.

We now turn to the case $q = 2$. From equation (8.6) we can read off the Bayes estimator of p_i^q to be

$$\widehat{p_i^q} = \frac{\Gamma(N + M)}{\Gamma(N + M + q)} \times \frac{\Gamma(N_i + 1 + q)}{\Gamma(N_i + 1)}. \tag{8.13}$$

Thus we write down $\widehat{H_2}$ in the form

$$\widehat{H_2}(\vec{N}) = \frac{1}{\ln 2}\left(1 - \sum_{i=1}^{M} \widehat{p_i^2}\right) \quad \text{with} \quad \widehat{p_i^2} = \left(\frac{N_i+1}{N+M}\right)\left(\frac{N_i+2}{N+M+1}\right). \tag{8.14}$$

In general, we find the following characteristics of the Bayes estimator to be noteworthy:

1. $\widehat{H_q}$ is defined in the parameter range $q \in (-1, \infty)$. Apparently, cases of particular interest (and simplicity) are given when q takes on integer values $n \in \mathcal{N}$ (set of non-negative integer numbers) which allow one to replace Gamma-functions by factorials. Similar simple expressions can also be obtained when $q = (n+1)/2$.

2. Given $q = n$, then equation (8.13) factorizes into a product of n terms, which takes on the following singled-out form:

$$\widehat{p_i^n} = \left(\frac{N_i+1}{N+M}\right)\left(\frac{N_i+2}{N+M+1}\right)\left(\frac{N_i+3}{N+M+2}\right)\cdots\left(\frac{N_i+n}{N+M-1+n}\right).$$

   As we have shown above, $\widehat{H_{q}}_{|q=1}$ includes the Bayes estimator of the Shannon entropy. Setting now $n = 1$, we furthermore re-obtain Laplace's (successor rule) estimator (see, e.g., [Schürmann & Grassberger 1996]): $\widehat{p_i^n}_{|n=1} = (N_i+1)/(N+M)$. Moreover, for $q = n$ the asymptotic approach $\widehat{p_i^n} \to {f_i}^n$ is realized by allowing $N \to \infty$, i.e., $\widehat{H_n}$ converges towards the frequency-count estimator of H_n. Thus the Bayes estimator is consistent.

3. We note that the Bayes estimator of H_q is not equal to the estimator obtained by inserting the Bayes estimator of the probability-vector $\vec{p}$, i.e., $\widehat{H_q}(\vec{N}) \neq H_q(\widehat{\vec{p}})$.

8.3 Rényi Entropy Estimator

In this section, we consider the Bayes estimator of the Rényi entropy K_q. Substituting K_q for H_q in equation (8.1), the problem of deriving the estimator is the calculation of the integral

$$\widehat{K_q}(\vec{N}) = \frac{1}{1-q}\frac{1}{W(\vec{N})}\int_{\mathcal{S}} d\vec{p}\ \log_2 Z_q(\vec{p})\ P(\vec{N}|\vec{p})\ Q(\vec{p}). \tag{8.15}$$

Even in the simple case for $M = 2$, finding the explicit analytical solution of the above integral turns out to be very complicated. In appendix C we will show that the Bayes estimator of the binary Rényi entropy (under the assumption of a uniform prior probability

density) can be written as

$$\widehat{K_q}(N_1, N_2) = \frac{1}{\ln 2}\frac{1}{1-q}\left(I_q\left(N_1, N_2\right) - q\sum_{l=N_1}^{N}\frac{1}{l+1}\right) \tag{8.16}$$

for all $N_1 + N_2 = N$. In the above expression we have introduced the following notation:

$$I_q\left(N_1, N_2\right) = \frac{\Gamma(N+2)}{\Gamma(N_1+1)\Gamma(N_2+1)}\int_0^\infty dx\ \frac{x^{N_2}}{(1+x)^{N+2}}\ \ln\left(1+x^q\right). \tag{8.17}$$

Though the integrand in the above integral is well-defined and thus this integral exists for all q, we could not obtain a closed analytical expression for arbitrary given N_1, N_2 and q. This does also hold for the case $M > 2$. So the explicit evaluation of equation (8.15) remains a challenge.

Although equation the binary case, c.f. equation (8.17), could be solved numerically to give $\widehat{K_q}$, we may seek another strategy which is of practical use also in the multi-variate case $M > 2$. We recall that H_q and K_q are intimately related to each other via equation (5.5). Therefore, a natural way to estimate K_q would be to estimate H_q and then use relation (5.5) to compute K_q of corresponding order. Hence we may write down the (indirect) Bayes estimator $\widetilde{K_q}$, c.f. equation (8.6), in the form[1]

$$\widetilde{K_q}(\vec{N}) = \frac{1}{1-q}\log_2\left\{\frac{\Gamma(N+M)}{\Gamma(N+M+q)}\times\left[\sum_{i=1}^{M}\frac{\Gamma(N_i+1+q)}{\Gamma(N_i+1)}\right]\right\}. \tag{8.18}$$

Since $\lim_{q\to 1}\widetilde{K_q} = \lim_{q\to 1}\widehat{H_q}$, the limit $\widetilde{K_1} = \widehat{H}$ holds and we again re-obtain the Bayes estimator of the Shannon entropy. The motivation to proceed in this way is lead by the fact that both $\widehat{H_q}$ and $\widetilde{K_q}$ can be understood as entropies computed from the Bayes estimator of the partition function Z_q. As such, we gain a significant reduction of the entropy variance due to $\widehat{Z_q}$.

8.4 Numerical Tests

In this section we compare the variances of the direct and indirect Bayes estimators, $\widehat{H_q}$ and $\widetilde{K_q}$, with the variances of the frequency-count estimators, $\bar{H}_q$ and $\bar{K}_q$. To investigate and contrast the performance of the two different estimators we choose m-step memory Markov processes belonging to the following cases: (a) generated by a process with

[1] Please note that we distinguish the indirect from the direct Bayes estimator by a tilde.

no memory, i.e., $m = 0$, and (b) generated by a process with memory $m = 5$. In (a) we choose a process with equidistributed probabilities (henceforth denoted as Bernoulli process), whereas in the latter case we use the fifth-order transition probabilities taken from the complete $1,830,240$ nucleotides long *Haemophilus influenzae* DNA sequence [Fleischman et al. 1995] to generate a Markov chain with fifth-order memory. Figure 2 shows the rank ordered statistics obtained from the above DNA sequence and from a sequence of same length derived from a Bernoulli process. It can be seen that the DNA sequence is far more inhomogeneous than the realization of the Bernoulli process. The derived rank-order frequencies might count as a typical example representing hexamer distributions in (prokaryotic) DNA. The entropy analysis of biosequences has received applications in order to distinguish between coding and non-coding DNA [Fickett & Tung 1992], to detect repeated nucleotide sequences [Herzel et al. 1994b], and to characterize protein sequences [Herzel 1988, Strait & Dewey 1996]. A prerequisite to the application of generalized entropies in biosequence analysis are reliable estimators. Therefore we consider a probability-vector derived from a DNA sequence to test the performance of the Bayes estimators, given by the expressions (8.7) and (8.18), versus the frequency-counts estimators, which are obtained by defining $\bar{Z}_q \to \sum_{i=1}^{M} f_i^q$ with $f_i = N_i/N$.

Since we are particularly interested in the case where the size of the sequence length is in the order of magnitude of the cardinality of the alphabet, $M = 4^6$, we perform our numerical simulations with $N_{(\mathrm{a})} = 4 \cdot 10^3$ and $N_{(\mathrm{b})} = 8 \cdot 10^3$. Then, according to the probability-vector $\vec{p} \equiv (p_1, \ldots, p_{4096})$, a sequence S is randomly generated from which we estimate the entropy values. In both cases we can also compute the theoretical hexamer entropies (since we take the relative frequencies obtained from the DNA sequence as probabilities by definition). Hence, the difference between the estimated and the theoretical values defines a random variable, which we define as "entropy estimate deviation from true". Generating an ensemble of $10,000$ sequences and estimating the entropies from each sequence, we obtain the histograms displayed in figures 3, 4 and 5. These studies demonstrate the merit of the Bayes order-2 entropy estimators as compared to the frequency-count estimators. Indeed, the variances of the Bayes estimates are significantly smaller than the variances of the frequency-count estimates for both Markov processes with memory $m = 0$ and $m = 5$. In repeated simulations with different sequence lengths and different values of q ranging from -1 to 50 we could observe similar results: the Bayes estimator of H_q and K_q produces significantly smaller variances than the frequency-count estimator.

As analytical calculations and numerical simulations reveal, the Bayes estimator of Z_q (and hence for H_q and K_q) is biased. As we will show in appendix A, this bias can be approximated within $\mathcal{O}(1/N)$, by using a straightforward analytical approach.

8.5 Summary

In this chapter we derived the direct Bayes estimator $\widehat{H_q}$ of the order-q Tsallis entropy and the indirect Bayes estimators $\widetilde{K_q}$ of order-q Rényi entropy of a finite, discrete data set.

Our approach of deriving the Bayes estimators of H_q and K_q has been been motivated by the requirement to estimate generalized entropies from realizations where the total sample size N available may only be in the order of magnitude of the cardinality M. The central result of this work, namely the Bayes estimator of the Tsallis entropy H_q, is stated in expression (8.7). As we could not arrive at a closed form expression of the direct Bayes estimator of the Rényi entropy, we proposed an indirect Bayes estimator by the transformation-formula which connects the Tsallis with the Rényi entropy. In fact, both estimators, $\widehat{H_q}$ and $\widetilde{K_q}$, are based on the Bayes estimator of the partition function Z_q, which may be exploited to estimate related quantities like generalized dimensions or order-q Kolmogorov entropies. In the case of $q = (n+1)/2$, $n \in \mathcal{N}$, these estimators are easy to implement for numerical purposes.

A comparative study of the accuracy by which both the Bayes and the frequency-count estimators extract the order-2 entropies of m-step memory Markov-chains demonstrated the strength of the Bayes estimator. Over the whole parameter range $q \in (-1, \infty)$ the Bayes estimator outperforms the frequency-count estimator by a significantly smaller variance of its estimates. This makes the Bayes estimator appropriate to measure generalized entropies in a sample, whose size N may be as small as the cardinality M of the alphabet.

The Bayes estimators $\widehat{H_q}$ and $\widetilde{K_q}$ have been derived under the assumption of a uniform prior probability density. Clearly, the specific choice of an assumption for the prior probability density is application-dependent. Given no other constraint except $\vec{p} \in \mathcal{S}$, we assumed a constant prior probability density over the simplex. Note that this does not mean that the probabilities p_i are equidistributed, but rather that all probability-vectors $\vec{p}$ on the simplex $\mathcal{S}$ are equiprobable. Nevertheless, the numerical simulations demonstrate that for the probability-vectors considered is this work, which are by no means equidistributed on the simplex, the Bayes estimator with $Q(\vec{p}) = \text{const}$ leads to variances which

are significantly smaller as compared to the variances of the frequency-counts estimators of generalized entropies of order q.

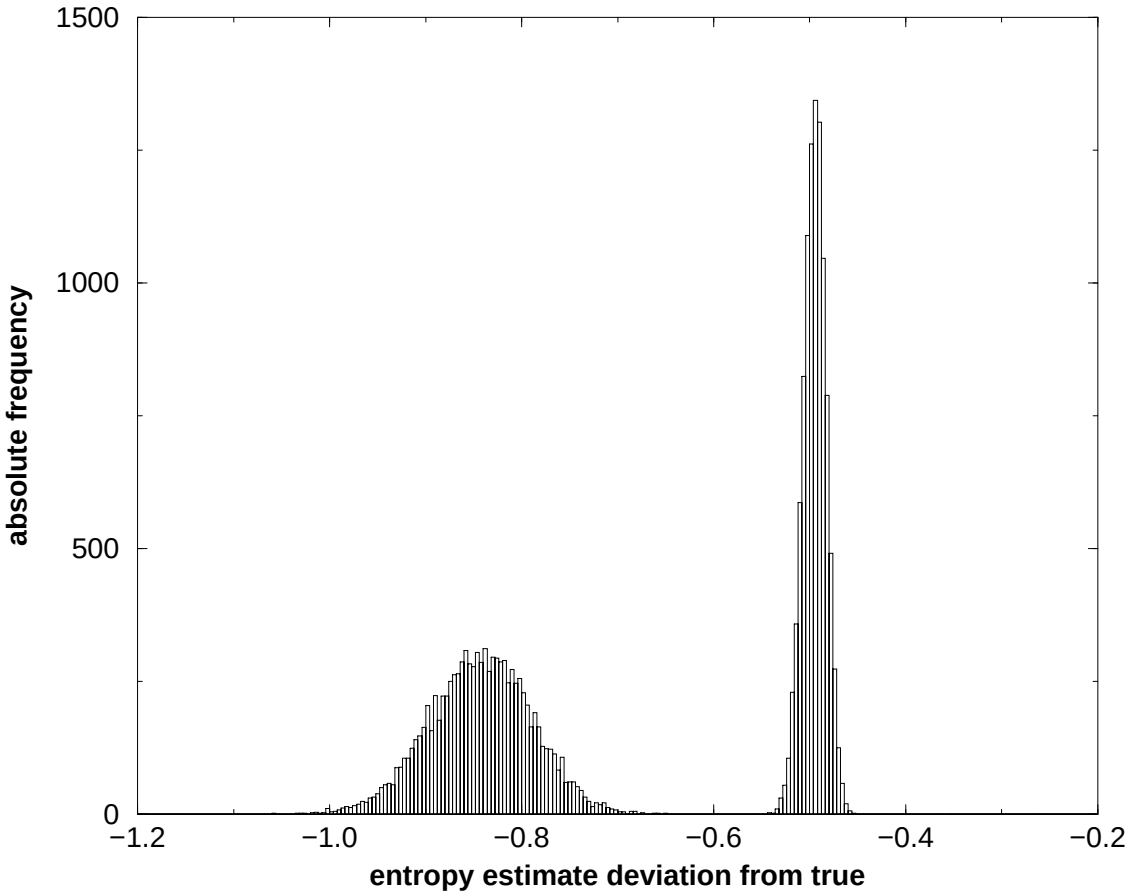


Figure 8.1: Comparison of the Bayes and the frequency-count estimator of the Shannon entropy, $\widehat{H_1}$ and $\bar{H}_1$ respectively. We generate an ensemble of $10,000$ sequences, each of which composed of $N = 250$ data points chosen from an alphabet with cardinality $M = 256$. The 256 possible outcomes were samples from a uniform distribution, $p_i = 1/M$. From each such sequence, the entropy is estimated by the Bayes estimator and the frequency-count estimator. Figure 1 displays the corresponding histograms of $\widehat{H_1}$ (right) and $\bar{H}_1$ (left). It can be seen that the variance of $\widehat{H_1}$ is about one order of magnitude smaller as compared to the variance of $\bar{H}_1$. Note that the significant negative bias can, in principle, be approximated by length correction formulae. Therefore, it is the smaller variance of $\widehat{H_1}$ that makes this estimator superior.

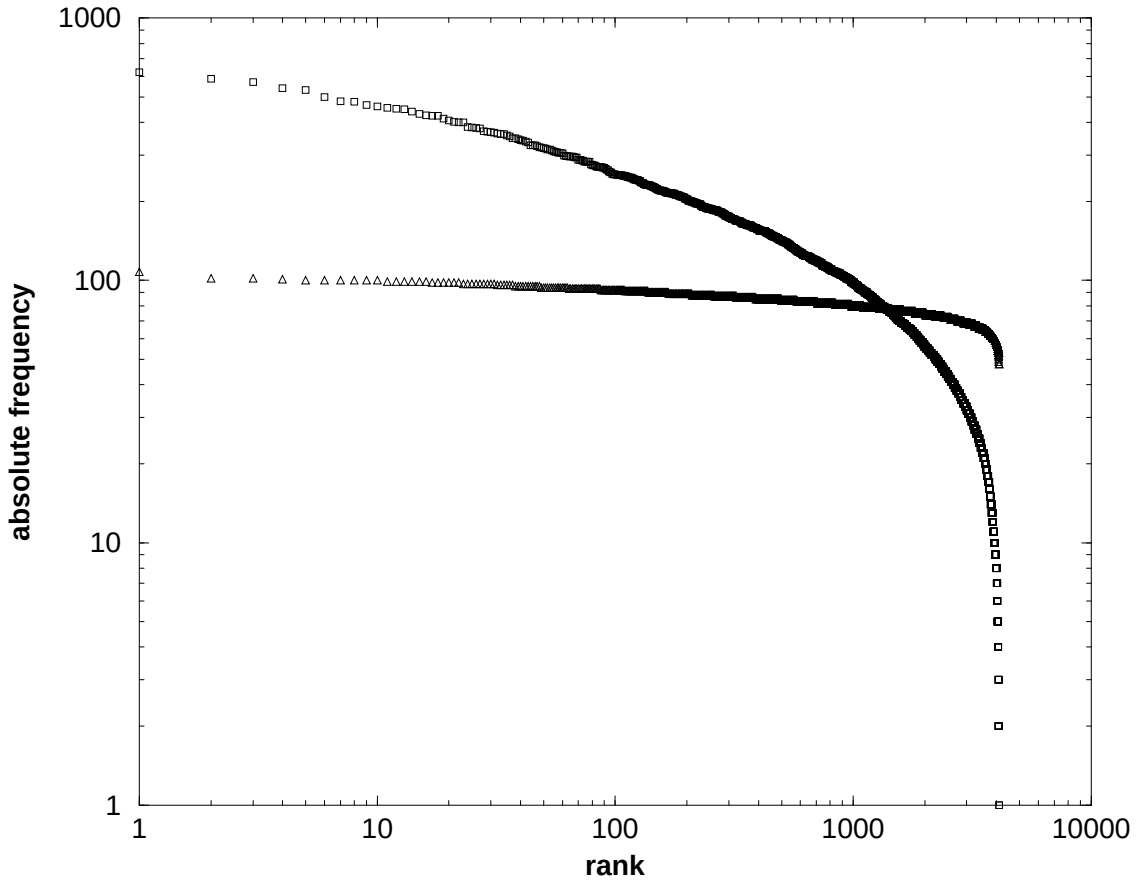


Figure 8.2: The rank ordered hexamer distribution of the complete *Haemophilus influenzae* DNA sequence displayed as a double-logarithmic plot (□). For a comparison, the rank ordered hexamer distribution of a Bernoulli-sequence of same length has been included in the figure (△).

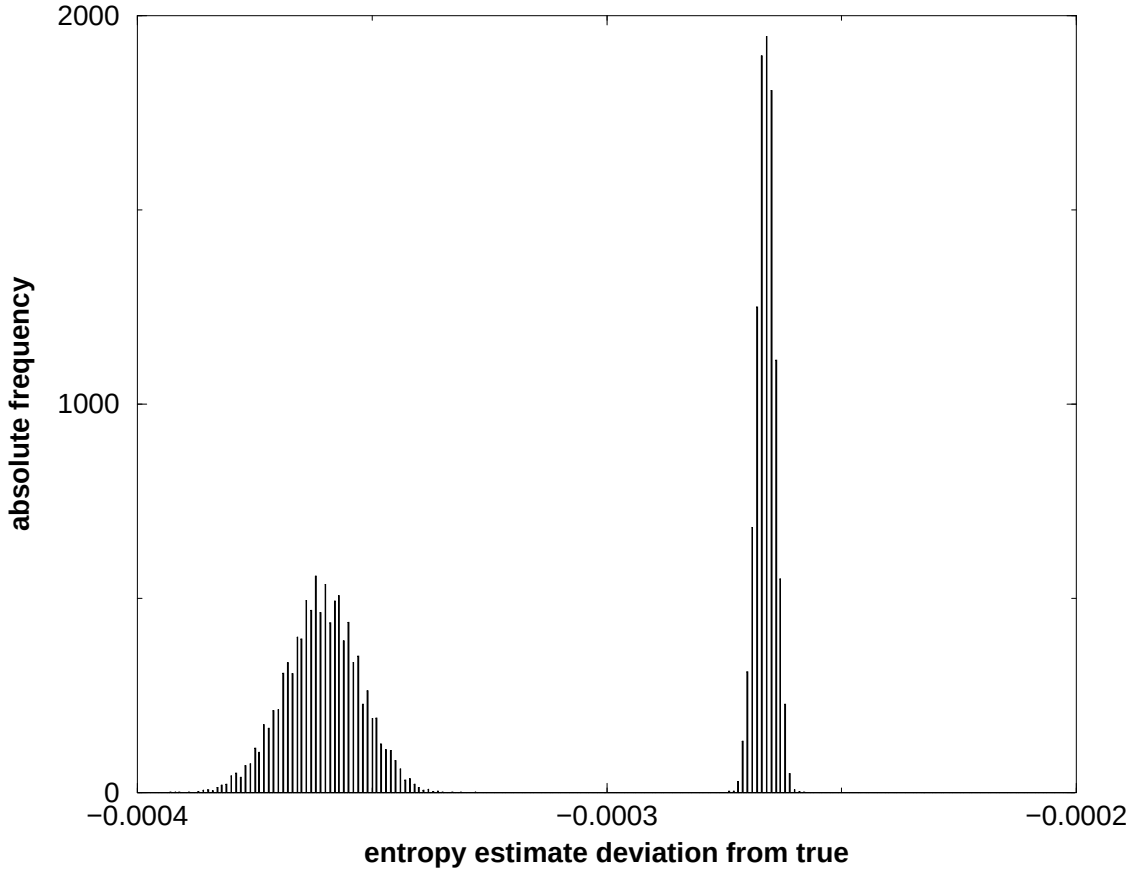


Figure 8.3: Comparison of the entropy estimators $\widehat{H_2}$ (right) and $\bar{H}_2$ (left) with $M = 4096$, $N = 4000$ and equidistributed $p_i = 1/M$. We observe the small width of the variance of the Bayes estimator $\widehat{H_2}$ as compared to the frequency-count estimator $\bar{H}_2$. Equation (12.3) predicts the entropy bias with $\Delta\widehat{H_2} = -2.66 \cdot 10^{-4}$ (bits/symbol), in good agreement to the observed value. According to [Holste 1997], the bias of $\bar{H}_2$ can be approximated to be $\Delta\bar{H}_1 = -0.36 \cdot 10^{-3}$ (bits/symbol), which is in good agreement to the observed value, too. In samples where N is in the order of magnitude of M, the reliability of the Bayes estimator is significantly higher than the reliability of the frequency-count estimator.

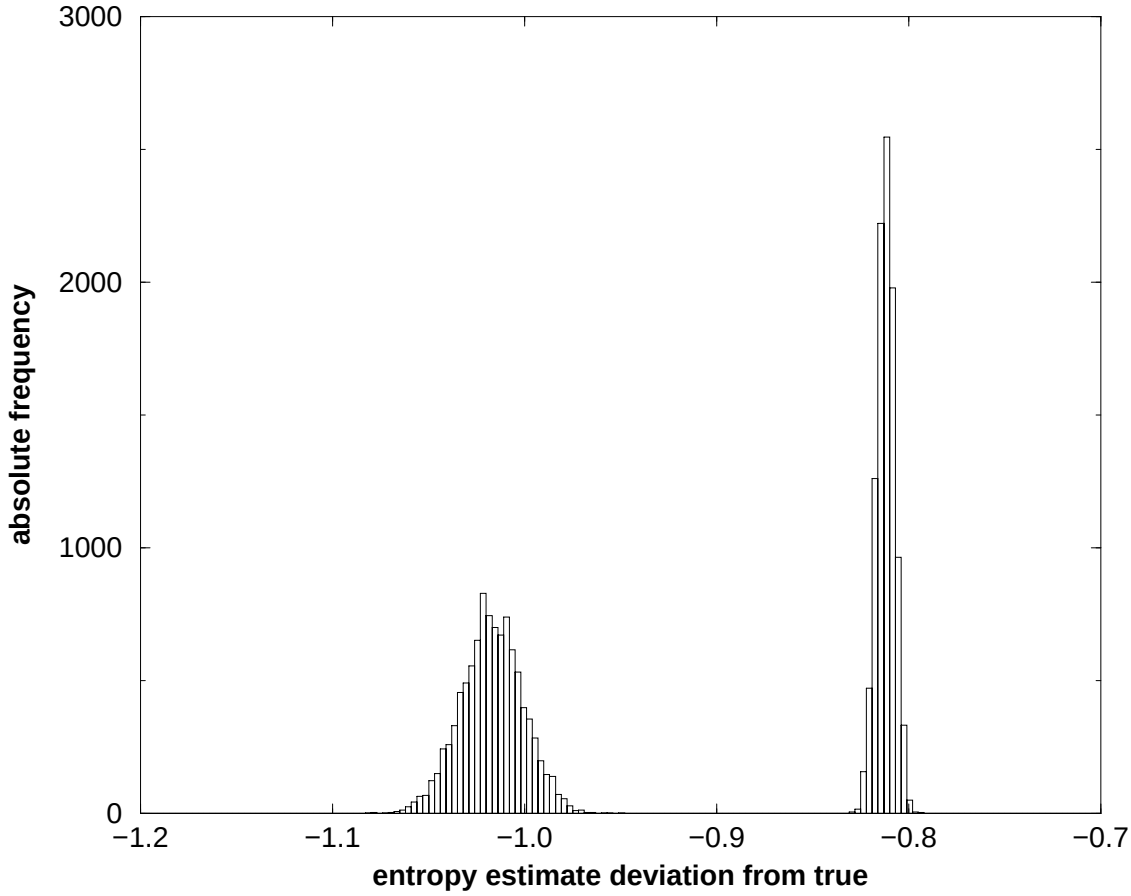


Figure 8.4: Comparison of the entropy estimators $\widetilde{K_2}$ (right) and $\bar{K}_2$ (left) with $M = 4096$, $N = 4000$ and equidistributed $p_i = 1/M$. We observe that fluctuations of the Bayes estimator $\widetilde{K_2}$ are strongly suppressed as compared to the frequency-count estimator $\bar{K}_2$. Equation (12.5) predicts the entropy bias with $\Delta\widetilde{K_2} = -0.81$ (bits/symbol), in good agreement to the observed value. According to [Holste 1997], the bias of $\bar{K}_2$ can be approximated to be $\Delta\bar{K}_2 = -1.02$ (bits/symbol), which is in good agreement to the observed value as well

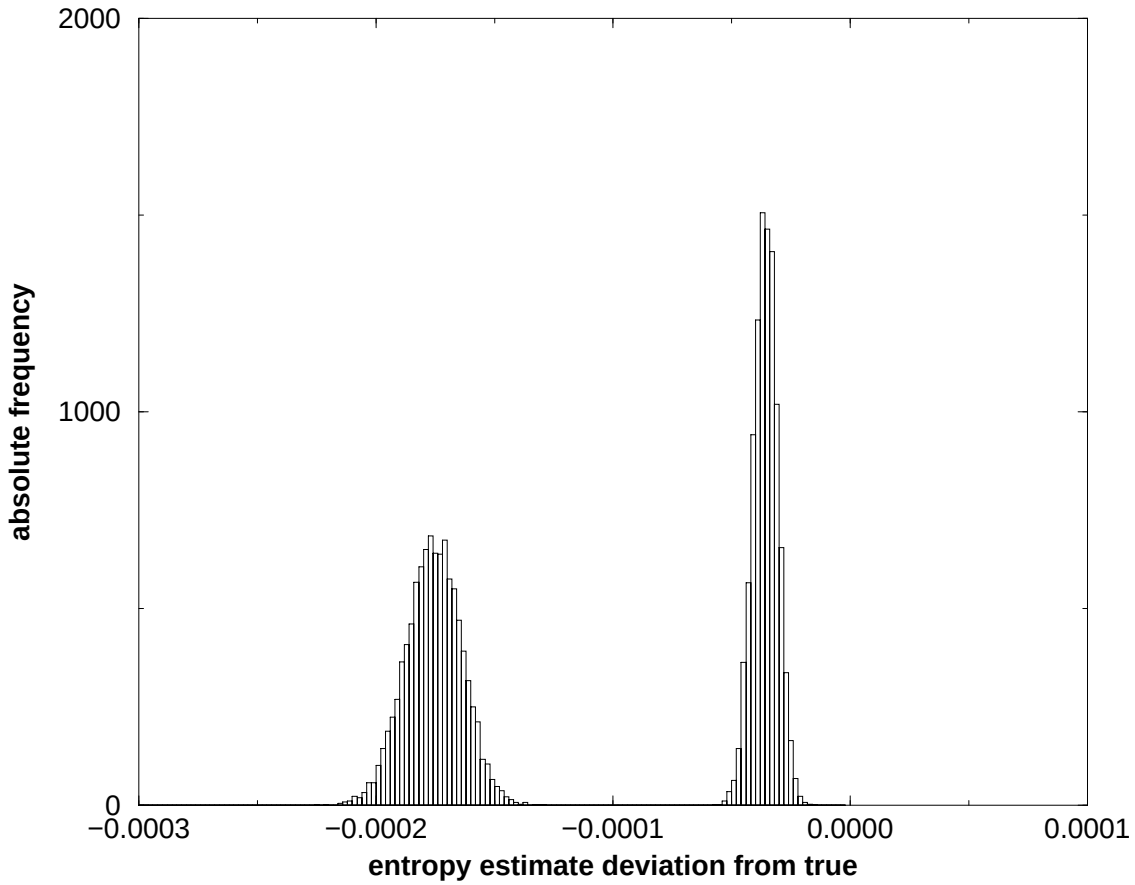


Figure 8.5: Comparison of the entropy estimators $\widehat{H_2}$ (right) and $\bar{H}_2$ (left) with $M = 4096$, $N = 8000$ and p_i derived from the *H. influenzae* DNA sequence. We observe the smaller variance of the Bayes estimator $\widehat{H_2}$ as compared to the frequency-count estimator $\bar{H}_2$. Equation (12.3) predicts the entropy bias with $\Delta\widehat{H_2} = -0.38 \cdot 10^{-4}$ (bits/symbol) and, according to [Holste 1997], the bias of $\bar{H}_2$ can be approximated to be $\Delta\bar{H}_1 = -0.18 \cdot 10^{-3}$ (bits/symbol).

Part C

Chapter 9

Statistical Properties of the Natural Estimator of the Shannon Entropy

The following section is devoted to the derivation and discussion of certain statistical properties of the natural entropy estimator,

$$\hat{H}(\vec{k}) = \sum_{i=1}^{M} -\frac{k_i}{N} \cdot \ln\left(\frac{k_i}{N}\right), \tag{9.1}$$

such as its expectation value and variance.

Remember that the estimate of a population parameter can be viewed as a random variable with a well defined probability distribution. The problem that occurs in practical applications is, however, that this probability distribution is often unknown and only hard to derive. Nevertheless, it can, in principle, be sufficiently described by all of its moments. Often its first two moments, namely its expectation value and its variance, suffice to give us a basic idea of what the distribution looks like.

The bias of an estimator is, as we have learned in section 6.3, defined as the deviation of the expected value of its estimates from the real population parameter value. The variance of the estimates, on the other hand, is defined as the mean quadratic distance from the expectation value, and thus reflects the magnitude of fluctuations of this estimator. We commonly relate the term reliability with the second moment of the estimate's probability distribution.

Section 9.1 is devoted to showing that the natural entropy estimator is biased in all cases where our underlying probability distribution is not pathologically concentrated in a single point. In section 9.2, we then derive an approximation for this bias, which has turned out to be fundamental for all statistical analyses of biological symbol sequences. Eventually, section 9.3 is dedicated to presenting an approximation for the variance of the natural entropy estimates.

Please note that neither the approximation of the expectation value nor the approximation of the variance of the Shannon entropy are new. These and similar approximations can be found in [Basharin 1957], [Harris 1975], [Pompe 1986], [Herzel 1988], [Li 1989], [Herzel 1994a], or [Levitin 1994].

9.1 The Natural Entropy Estimator is Biased

In this section, we will show that we systematically underestimate the Shannon entropy $H(\vec{p})$ by using the natural entropy estimator $\hat{H}(\vec{k})$ independently on the underlying probability distribution $p_1, p_2, ..., p_M$.

This bias of the natural entropy estimator is due to the nonlinearity, i.e., the convexity, of the entropy function and can easily be understood in the case of equidistributed p_i.

Let us imagine we are to estimate the Shannon entropy

$$H(p) = -p \cdot \ln(p) - (1-p) \cdot \ln(1-p) \tag{9.2}$$

of a coin and assume $p = 1/2$. Let further be the sample size $N = 2$, i.e., we are tossing the coin only twice.

Then the possible outcomes of our tossing experiment are:

1. We toss two heads, i.e., $k = 2$.
2. We toss one head and one tail (or one tail and one head), i.e., $k = 1$.
3. We obtain two tails, i.e., $k = 0$.

In the first and the third case, we estimate the Shannon entropy of the coin be zero, which reflects our estimates of the probability $\hat{p} = 1$ or $\hat{p} = 0$, respectively. In the second case, we estimate the probability $\hat{p} = 1/2$ and thus the entropy $\hat{H} = \ln(2)$.

Comparing all of our entropy estimates with the theoretical Shannon entropy $H(p) = \ln(2)$, we see that all estimates, which we also call the observed entropies, are smaller than

or equal to the theoretical entropy. In our case, this only reflects the well known statement that a uniform distribution yields the highest possible entropy. Our realizations, however, cannot be more uniform than the underlying uniform probability distribution is. Moreover, they often deviate from uniformity, which happens in two of three cases in our example.

Thus, it is not surprising that the expected value of the observed entropies, i.e., the ensemble mean of all observed entropies, is smaller than the theoretical entropy value in our example:

$$\begin{aligned} E(\hat{H}) &= P(k=0|p=1/2)\cdot\hat{H}(k=0) + P(k=1|p=1/2)\cdot\hat{H}(k=1) \\ &+ P(k=2|p=1/2)\cdot\hat{H}(k=2) = 1/2\cdot\ln(2) < \ln(2) = H. \end{aligned} \tag{9.3}$$

The question we are raising at this point is whether this inequality always holds. Let us, for example, consider the same experiment, but now under the assumption of $p = 0.7$. Again, we obtain the three possible estimates for $\hat{H}$, namely 0, 1, and 0, for the three possible outcomes $k = 2$, $k = 1$, and $k = 0$, respectively, of our tossing experiment.

However, now we realize that the entropy observed in the second case is larger than the theoretical value. Nevertheless, we can prove that the expected entropy (which we define as the expectation value of all observed entropies) is smaller than the theoretical entropy:

$$\begin{aligned} E(\hat{H}) &= P(k=0|p=0.7)\cdot\hat{H}(k=0) + P(k=1|p=0.7)\cdot\hat{H}(k=1) \\ &+ P(k=2|p=0.7)\cdot\hat{H}(k=2) = 2\cdot 0.7\cdot 0.3\cdot\ln(2) \\ &< -0.3\cdot\ln(0.3) - 0.7\cdot\ln(0.7) = H. \end{aligned} \tag{9.4}$$

The interesting question that appears at this moment is whether the expected entropy is always smaller than the theoretical entropy independently on our assumption about the underlying distribution and its dimensionality.

The following section is devoted to deriving the theorem that the expected value of the observed entropies is always smaller than the theoretical entropy.

Let $\vec{p} = (p_1, p_2, ..., p_M)$ be the vector containing the probabilities of an M-sided die and N be the given sample size. Let further $\binom{N}{\vec{k}}$ denote the multinomial coefficient given by

$$\binom{N}{\vec{k}} \equiv \frac{N!}{k_1!\cdot k_2!\cdots k_M!}. \tag{9.5}$$

Then the expected entropy reads as

$$E\left(\hat{H}(\vec{k})\right) = \sum_{\{\vec{k}\}} P(\vec{k}|\vec{p}) \cdot \hat{H}(\vec{k}) \tag{9.6}$$

with

$$P(\vec{k}|\vec{p}) = \binom{N}{\vec{k}} \prod_{i=1}^{M} p_i^{k_i} \tag{9.7}$$

and

$$\hat{H}(\vec{k}) = \sum_{i=1}^{M} -\frac{k_i}{N} \cdot \ln\left(\frac{k_i}{N}\right). \tag{9.8}$$

Interchanging the finite sum over all possible states $\vec{k}$ with the sum over all state indices i yields

$$E\left(\hat{H}(\vec{k})\right) = -\sum_{i=1}^{M} \sum_{\{\vec{k}\}} P(\vec{k}|\vec{p}) \cdot \frac{k_i}{N} \cdot \ln\left(\frac{k_i}{N}\right). \tag{9.9}$$

Now we exploit eq. (2.12), which we derive in appendix B by applying Jensen's inequality to the function $f(p) = -p \cdot \ln(p)$, and obtain

$$-\sum_{\{\vec{k}\}} P(\vec{k}|\vec{p}) \cdot \frac{k_i}{N} \cdot \ln\left(\frac{k_i}{N}\right) \leq -p_i \cdot \ln(p_i) \tag{9.10}$$

for all $i = 1, 2, ..., M$, since

$$p_i = \sum_{\{\vec{k}\}} P(\vec{k}|\vec{p}) \cdot \frac{k_i}{N}. \tag{9.11}$$

By applying this result to eq. (9.9), we end up with the inequality

$$\begin{aligned} E\left(\hat{H}(\vec{k})\right) &= -\sum_{i=1}^{M} \sum_{\{\vec{k}\}} P(\vec{k}|\vec{p}) \cdot \frac{k_i}{N} \cdot \ln\left(\frac{k_i}{N}\right) & (9.12) \\ &\leq -\sum_{i=1}^{M} p_i \cdot \ln(p_i) = H(\vec{p}) & (9.13) \end{aligned}$$

where equality holds only in the pathological case that P is concentrated in one point, i.e., if one $p_i = 1$ and all others be 0.

In other words, the natural entropy estimator is always biased and on average underestimates the theoretical Shannon entropy.

9.2 The Entropy Bias

In the last section, we have shown that the natural entropy estimator is biased. Furthermore, we could show that this bias is always negative, i.e., we systematically underestimate the theoretical entropy independently on the underlying probability distribution $p_1, p_2, ..., p_M$. In this section, we will quantify this bias and derive a length correction formula that approximates the bias in the order of $1/N$.

The expected value of the natural entropy estimates is given by

$$E(\hat{H}(\vec{k})) = \sum_{\{\vec{k}\}} \hat{H}(\vec{k}) \cdot P(\vec{k}|\vec{p}). \tag{9.14}$$

We know that the vector $\vec{k}$ containing the absolute frequencies $k_1, k_2, ..., k_M$ is multinomially distributed according to

$$P(\vec{k}; \vec{p}, N) = \binom{N}{\vec{k}} \cdot \prod_{i=1}^{M} p_i^{k_i}. \tag{9.15}$$

However, we do neither know the distribution of $\hat{H}(\vec{k})$ nor the expectation value of this statistic.

But following [Harris 1975] and [Herzel, 1988], we can expand the function $\hat{H}(\vec{k})$ in a Taylor series and thus derive first and second order approximations of the expectation value of $\hat{H}(\vec{k})$.

We define the relative frequencies as the positive real variables $x_1, x_2, ..., x_M$ by

$$x_i \equiv k_i/N \tag{9.16}$$

for all $i = 1, 2, ..., M$ and then compute the power series expansion of $H(x_1, x_2, ..., x_M)$ about the point $(x_1, x_2, ..., x_M) = (p_1, p_2, ..., p_M)$.

Since the partial derivatives read as

$$\frac{\partial H}{\partial x_i} = -\ln(x_i) - 1 \tag{9.17}$$

$$\frac{\partial^2 H}{\partial x_i \partial x_j} = 0 \tag{9.18}$$

$$\frac{\partial^k H}{\partial {x_i}^k} = (-1)^{k-1} \cdot \frac{(k-2)!}{x_i^{k-1}} \tag{9.19}$$

for all $i,j = 1,2,...,M$, $j \neq i$, and all $k = 2,3,...,\infty$, we obtain the following Taylor series for the observed Shannon entropy $H(x)$:

$$\begin{aligned}\tilde{H}(\vec{x}) &= \sum_{i=1}^{M} -p_i \cdot \ln(p_i) - (\ln(p_i)+1)\cdot(x_i - p_i) \\ &+ \sum_{k=2}^{\infty} \frac{(-1)^{k-1}}{(k-1)\cdot k} \cdot \frac{1}{p_i^{k-1}} \cdot (x_i - p_i)^k \qquad (9.20) \\ &= \sum_{i=1}^{M} -p_i \cdot \ln(p_i) - \ln(p_i)\cdot(x_i - p_i) \\ &+ \sum_{k=2}^{\infty} \frac{(-1)^{k-1}}{(k-1)\cdot k} \cdot \frac{1}{p_i^{k-1}} \cdot (x_i - p_i)^k. \qquad (9.21)\end{aligned}$$

Please realize that we denoted the limes of this power series by $\tilde{H}$ and not by H because we have not yet analyzed whether the function $H(x)$ is indeed analytic, which means, whether it is expressible by its Taylor series.

We know that all power series in x about a point x_0 converge in an open interval $(x_0 - r, x_0 + r)$, where r is called the radius of convergence. Our next task will be to determine the radius of convergence r of the series given above.

The ratio test reveals that the radius of convergence r of a power series $\sum_{k=0}^{\infty} a_k \cdot x^k$ can be determined by

$$r = \lim_{k\to\infty} \left| \frac{a_k}{a_{k+1}} \right|. \qquad (9.22)$$

Hence, we obtain, for the above series and $k \geq 2$,

$$\left| \frac{a_k}{a_{k+1}} \right| = \frac{k+1}{k-1} \cdot p_i \to p_i \qquad (9.23)$$

for $k \to \infty$ and all $i = 1,2,...,M$.

Since power series converge inside their convergence intervals, but diverge outside, the Taylor expansion derived above is only convergent on the whole interval $(0,1)$ in the case that we are dealing with binary sequences in which the two symbols occur with the same probability $1/2$. Let us incidentally remark that in this particular case, the series also converges at the points $x = 0$ and $x = 1$ and that its limits are even identical to the values of the function H at this point. This means we may identify H with $\tilde{H}$ in this case.

In all other cases, the Taylor series becomes divergent in certain intervals, and the only statement we are allowed to derive is that the power series given above converges in the open parallel epiped given by the two points $(0,0,...,0)$ and $2\cdot(p_1,p_2,...,p_M)$. Here we can then prove $H(x) = \tilde{H}(x)$.

For all $\vec{x}$ that fall into this parallel epiped, the Taylor approximation can be improved with any higher order term whereas the opposite is the case for all other $\vec{x}$. If we, however, keep in mind the probabilities by which vectors $\vec{x}$ appear in our sample of size N, then we will realize that this Taylor approximation is a really good and valuable approach in many practical situations. It only fails if the underlying statistics becomes extremely poor.

For those situations, we present a different approach in appendix G. There, we exploit a technique derived from regression theory, which finally provides us with a power series for the Shannon entropy that converges in the entire interval $[0, 1]$.

Let us in this section continue with the Taylor approach and write down the expectation value of the natural entropy estimator $H(x)$.

$$\begin{aligned} E(\tilde{H}(\vec{x})) &= \sum_{i=1}^{M} -p_i \cdot \ln(p_i) - \ln(p_i) \cdot E(x_i - p_i) \\ &+ \sum_{k=2}^{\infty} \frac{(-1)^{k-1}}{(k-1) \cdot k} \cdot \frac{1}{p_i^{k-1}} \cdot E\left((x_i - p_i)^k\right). \end{aligned} \tag{9.24}$$

In Appendix I, we introduce generating functions and present a method by which we can, in principle, derive all moments of a multinomial distribution. For the sake of simplicity, we truncate the Taylor expansion after the quadratic term, i.e., we neglect all moments higher than the second, and thus obtain

$$E(\tilde{H}(\vec{x})) = \sum_{i=1}^{M} -p_i \cdot \ln(p_i) - \sum_{i=1}^{M} \frac{1}{2 \cdot p_i} \cdot \frac{p_i \cdot (1 - p_i)}{N} + \mathcal{O}\left(\frac{1}{N^2}\right) \tag{9.25}$$

$$= H(\vec{p}) - \frac{M-1}{2 \cdot N} + \mathcal{O}\left(\frac{1}{N^2}\right) \tag{9.26}$$

since

$$E(x_i - p_i) = 0 \tag{9.27}$$

and

$$E\left((x_i - p_i)^2\right) = \frac{p_i \cdot (1 - p_i)}{N} \tag{9.28}$$

as derived in appendix I.

In our previous section, we have learned that we always systematically underestimate the Shannon entropy independently on the underlying probability distribution $p_1, p_2, ..., p_M$. In this section, we could now derive the surprising result that the bias of the natural entropy estimator

$$E(\hat{H} - H) = -\frac{M-1}{2 \cdot N} + \mathcal{O}\left(\frac{1}{N^2}\right) \tag{9.29}$$

does also not depend on the underlying probability distribution in a first order approximation.

However, this interesting finding is not new and has only been presented here for exemplifying our basic approach. We highly recommend our readers to review the original papers mentioned above, where these so called finite sample effects are studied and the value of the length correction formula corresponding to eq. (9.29) are perfectly illustrated.

9.3 The Entropy Variance

In this section, we will derive an approximation of the variance of the natural entropy estimates $H(\vec{k})$ where the random vector $\vec{k}$ is multinomially distributed according to

$$P(\vec{k}|\vec{p}) = \binom{N}{\vec{k}} \prod_{i=1}^{M} p_i^{k_i}. \tag{9.30}$$

In section 6.3, we have discussed some desirable properties of point estimators; namely, to be consistent and unbiased. Another, perhaps even more relevant question is whether our estimator is also reliable. Whereas the question for the bias of a given estimator confronted us with calculating the expectation value of its estimates, we are now confronted with computing the mean sample variance of its estimated values.

Fluctuations of their corresponding estimates are immanent for all estimators due to always finite sample sizes. The magnitude of these fluctuations, however, is an important criterion to evaluate statistical measures and to determine their reliability.

Since we can approximate the systematic sampling errors of all estimators considered in this work and thus correct their bias, the remaining statistical errors decide about the power of the considered estimators.

In chapter 7, we have already discussed the variance of the natural entropy estimates. Let us here have a closer view on the size of the variance and compare it with analytical results derived below.

The simulations performed here exactly correspond to those experiments carried out in section 7.3. However, instead of fixing the 1024-dimensional probability vector $\vec{p}$ to the 5-mer probabilities derived from the C.elegans chromosome III, we simply choose $\vec{p} = (\frac{1}{M}, \frac{1}{M}, ..., \frac{1}{M})$ for our following simulations. We generate sequences of $N = 1000$ 5-mers and repeat this experiment 10,000 times, which yields the following histogram of the observed

Shannon entropy estimates

$$\hat{H}(\vec{k}) = -\sum_{i=1}^{M} \frac{k_i}{N} \cdot \ln\left(\frac{k_i}{N}\right). \tag{9.31}$$

Although we revealed an approximate value of the sample variance of the natural Shannon entropy estimates, we have to admit that simulations of those statistics are tedious and lethal for any scientific progress in a long run. Thus our goal will be to derive an analytic expression that approximates the variance of the natural entropy estimates as a function of $p_1, p_2, ..., p_M$, and N.

A naive approach yields

$$\sigma^2\left(\hat{H}(\vec{n})\right) = \sigma^2\left(\sum_{i=1}^{M} -n_i \cdot \ln(n_i)\right) \approx \sum_{i=1}^{M} \sigma^2\left(-n_i \cdot \ln(n_i)\right), \tag{9.32}$$

if we assume the relative frequencies $n_i = k_i/N$ be statistically independent.

In a first order approximation, we obtain

$$-n_i \cdot \ln(n_i) = -p_i \cdot \ln(p_i) - \ln(p_i) \cdot (n_i - p_i) + \mathcal{O}\left((n_i - p_i)^2\right) \tag{9.33}$$

for all $i = 1, 2, ..., M$ and thus

$$\sigma^2\left(\hat{H(\vec{n})}\right) \propto \sum_{i=1}^{M} -\ln^2(p_i) \cdot \sigma^2\left((n_i - p_i)\right). \tag{9.34}$$

In this linear approximation, we could obviously relate the fluctuations of the Shannon entropy to the fluctuations of the relative frequencies n_i.

Assuming equidistributed p_i, we obtain

$$\sigma^2\left(\hat{H}(\vec{n})\right) \propto \frac{M \cdot \ln^2(M) \cdot (M-1)}{N \cdot M^2} \tag{9.35}$$

since

$$\sigma^2\left((n_i - p_i)\right) = \frac{p_i \cdot (1 - p_i)}{N}. \tag{9.36}$$

Comparing this first order approximation, which yields

$$\sigma^2\left(\hat{H}(\vec{n})\right) \propto \frac{\ln^2(M)}{N} \approx 0.05, \tag{9.37}$$

with Figure 9.1 reveals the shocking fact that our variance calculation is by a factor of 20 too big.

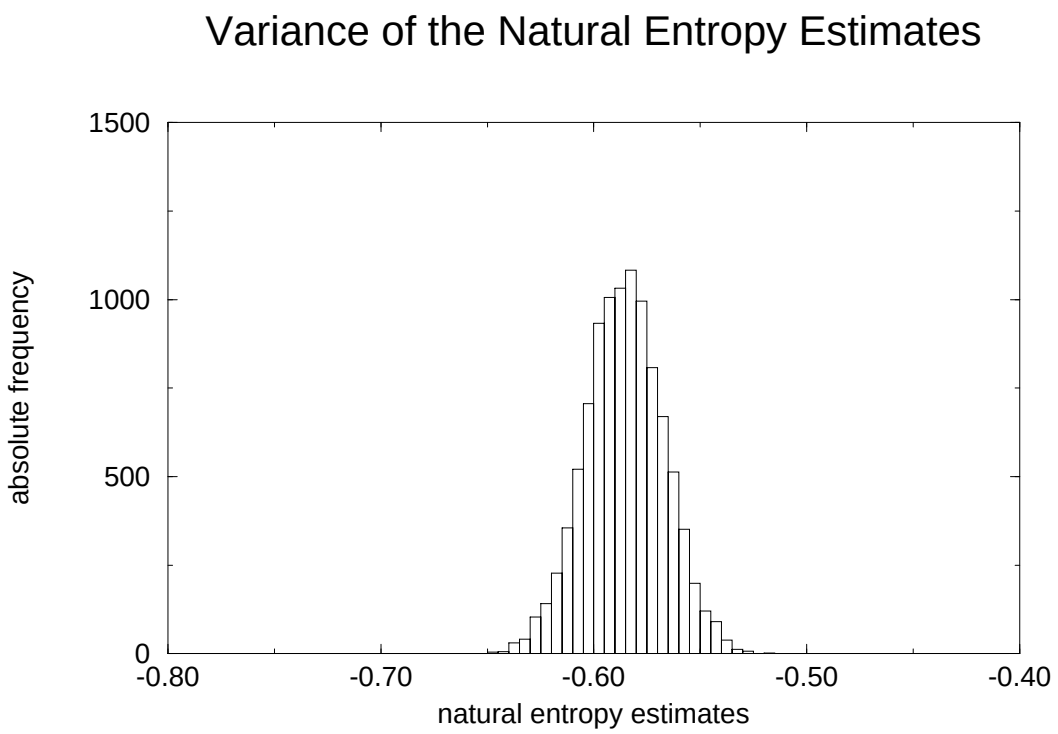


Figure 9.1: Variance of the observed 5-gram entropies H_5. We generated $10,000$ sequences each of which contained $N = 1000$ out of 1024 equidistributed 5-mers. The bias of the natural Shannon entropy estimator of about -0.58 is almost perfectly predicted by our linear approximation $\frac{M}{2 \cdot N}$. Note that the standard deviation of the natural entropy estimates, which we approximate by 0.05, is significantly smaller than the their systematic error. The variance, i.e., the squared standard deviation, can be roughly estimated as $3 \cdot 10^{-3}$.

The serious question is whether this error was caused by the naive assumption all n_i be independent or by our first order approximation.

In the following, we will show which dramatic effect the constraint $\sum_{i=1}^{M} n_i = 1$ imposes on the sum of the fluctuations of the otherwise independent n_i.

To present this paradoxon as clearly as possible, let us consider the following example. We are rolling a normal six-sided die the faces of which show the six letters $A, B, ..., F$. Let our sample size N be 100, i.e., we are rolling the die 100 times and then count how often we obtained an A, a B, a C, and so on. Eventually we are wondering whether the number of As we rolled has anything to do with the number of Bs.

The naive answer would be, that there is no statistical dependence between the frequency of rolling an A and the frequency of rolling a B. By definition, there is no correlation between the outcomes of our experiment!

Nevertheless, we know about the normalization constraint, which requires the sum over all absolute frequencies be equal to $N = 100$, or the sum over the relative frequencies be equal to 1.

In mathematical terms, the constraint

$$\sum_{i=1}^{M} n_i = 1 \tag{9.38}$$

induces slight correlations between the random variables n_i and thus our assumption of statistical independence between them is questionable.

Let us in the following prove that these weak normalization correlations can indeed dramatically change our result obtained above.

We still stick to the first order approximation given by eq. (9.33) and hence obtain

$$\sigma^2\left(\hat{H}(\vec{n})\right) = \sigma^2\left(\sum_{i=1}^{M} -n_i \cdot \ln\left(n_i\right)\right) \propto \sigma^2\left(\sum_{i=1}^{M} -\ln(p_i) \cdot (n_i - p_i)\right). \tag{9.39}$$

Assuming uniformity among the p_i, i.e., $p_i = 1/M$ for all $i = 1, 2, ..., M$, yields

$$\sigma^2\left(\hat{H}(\vec{n})\right) = \ln^2(M) \cdot 0 \tag{9.40}$$

since

$$\sum_{i=1}^{M} n_i = \sum_{i=1}^{M} p_i = 1. \tag{9.41}$$

What we have just learned is that the weak correlations we induced by regarding the normalization constraint for the observed frequencies destroy all fluctuations of the observed Shannon entropies in a linear approximation.

This example was to serve as an illustration that the proper summation on the simplex given by eq. (9.38) cannot be substituted by a summation over all possible states $\vec{k}$ disregarding their normalization constraint.

In the remainder of this section, we will derive an analytic expression for the variance of the observed entropies derived from a sample of size N for arbitrary $p_1, p_2, ..., p_M$.

Let us split this task according to

$$\sigma^2\left(\hat{H}\right) = E\left(\left(\hat{H} - H\right)^2\right) - \left(E\left(\hat{H}\right) - H\right)^2 \tag{9.42}$$

and start approximating the first term in the order of $1/N$.

Since

$$\hat{H}(\vec{n}) - H(\vec{p}) = \sum_{i=1}^{M} -\ln(p_i) \cdot (x_i - p_i) + \sum_{k=2}^{\infty} \frac{(-1)^{k-1}}{(k-1)\cdot k} \cdot \frac{1}{p_i^{k-1}} \cdot (x_i - p_i)^k, \tag{9.43}$$

we obtain

$$\left(\hat{H}(\vec{n}) - H(\vec{p})\right)^2 = \sum_{i,j=1}^{M} \ln(p_i) \cdot \ln(p_j) \cdot (x_i - p_i) \cdot (x_j - p_j) \tag{9.44}$$

$$+ \quad h.\ o.\ t. \tag{9.45}$$

Calculating the expectation values of these terms yields

$$E\left(\left(\hat{H}(\vec{n}) - H(\vec{p})\right)^2\right)$$

$$= \sum_{i,j=1}^{M} \ln(p_i) \cdot \ln(p_j) \cdot E\left((n_i - p_i) \cdot (n_j - p_j)\right) + E\left(h.\ o.\ t.\right) \tag{9.46}$$

$$= \sum_{i=1}^{M} \ln^2(p_i) \cdot E\left((n_i - p_i)^2\right)$$

$$+ \sum_{i,j=1}^{M} (1 - \delta_{ij}) \cdot \ln(p_i) \cdot \ln(p_j) \cdot E\left((n_i - p_i) \cdot (n_j - p_j)\right) + E\left(h.\ o.\ t.\right) \tag{9.47}$$

$$= \sum_{i=1}^{M} \ln^2(p_i) \cdot \frac{p_i \cdot (1 - p_i)}{N}$$

$$- \sum_{i,j=1}^{M} (1 - \delta_{ij}) \cdot \ln(p_i) \cdot \ln(p_j) \cdot \frac{p_i \cdot p_j}{N} + E\left(h.\ o.\ t.\right) \tag{9.48}$$

$$= \frac{1}{N} \sum_{i=1}^{M} \ln^2(p_i) \cdot p_i$$

$$- \frac{1}{N} \sum_{i,j=1}^{M} \ln(p_i) \cdot \ln(p_j) \cdot p_i \cdot p_j + E\left(h.\ o.\ t.\right) \tag{9.49}$$

$$= \frac{1}{N}\sum_{i=1}^{M} p_i \cdot \ln^2(p_i) - \frac{1}{N}\left(\sum_{i=1}^{M} \ln(p_i)\cdot p_i\right)^2 + E\left(h.\ o.\ t.\right) \tag{9.50}$$

$$= \frac{1}{N}\left(\sum_{i=1}^{M} p_i \cdot \ln^2(p_i) - H^2\right) + E\left(h.\ o.\ t.\right). \tag{9.51}$$

Considering $E\left(h.\ o.\ t.\right)$ reveals that all higher order terms only contribute fluctuations in the order of $\mathcal{O}\left(1/N^2\right)$.

This is not at all trivial, which all readers might experience who derive an approximation for the variance of the natural entropy estimates in the order of $\mathcal{O}\left(1/N^2\right)$. In that case, a Taylor expansion of $H(\vec{n})$ in the order of $\mathcal{O}\left((\vec{n}-\vec{p})^2\right)$ does not suffice to collect all terms contributing to an $1/N^2$ approximation of $\sigma^2(\hat{H})$.

At this point, we renounce to display the second order approximation $\sigma^2(\hat{H})$, but rather recommend the interested reader to study the wonderful review article by Harris [1975].

Realizing that the square of the natural entropy estimator bias goes with $\frac{(M-1)^2}{4\cdot N^2} + \mathcal{O}\left(1/N^3\right)$ and thus neglecting the term $\left(E(\hat{H})-H\right)^2$ in our $1/N$ approximation of its variance leads to our final result:

$$\sigma^2\left(\hat{H}\right) = \frac{1}{N}\left(\sum_{i=1}^{M} p_i \cdot \ln^2(p_i) - H^2\right) + \mathcal{O}\left(\frac{1}{N^2}\right). \tag{9.52}$$

At this point, we eventually understand why the fluctuations of the observed entropies are of negligible magnitude compared to the bias of the natural entropy estimator in our introductory example. If the probabilities p_i approach an uniform distribution, the term $\sum_{i=1}^{M} p_i \cdot \ln^2(p_i) - H^2$ vanishes.

Finally, we will display this well known result in a new fashion, which might delight all readers who can sometimes hear the harmonic cords of the great music that nature plays.

Rewriting eq. (9.52) yields

$$\sigma^2\left(\hat{H}\right) = \frac{1}{N}\left(\sum_{i=1}^{M} p_i \cdot \ln^2(p_i) - H^2\right) + \mathcal{O}\left(\frac{1}{N^2}\right) \tag{9.53}$$

$$= \frac{1}{N}\cdot\left(E\left(\ln^2(p_i)\right) - E^2\left(\ln(p_i)\right)\right) + \mathcal{O}\left(\frac{1}{N^2}\right), \tag{9.54}$$

and hence,

$$\sigma^2\left(\hat{H}\right) = \frac{1}{N}\cdot\sigma^2\left(\ln(p_i)\right) + \mathcal{O}\left(\frac{1}{N^2}\right). \tag{9.55}$$

Theorem 9.1 *The sample variance of the observed entropies is given by the population variance of the M numbers $\ln(p_i)$ divided by the sample size N.*

Although this is a clear mathematical statement, we are far from intuitively understanding the relation between the fluctuations of the Shannon entropy estimates on the one hand and the variance of the logarithms of the theoretical probabilities on the other hand.

For equidistributed p_i, we immediately realize

$$\sigma^2\left(\ln(p_i)\right) = \left(E\left(\ln^2(p_i)\right) - E^2\left(\ln(p_i)\right)\right) = 0, \tag{9.56}$$

which means that the sample variance of the Shannon entropy decays with $1/N^2$ in this case.

Chapter 10

Statistical Properties of the Natural Estimator of the Mutual Information

In chapter 3, we introduced the mutual information as a correlation measure which possesses the desirable property that it vanishes if, and only if, the considered random variables are statistically independent.

The estimation of this function from finite samples will, however, again affect the precise determination of the hidden correlations. Hence, we are again confronted with the task to derive some basic statistical properties of the mutual information estimator such as its expectation value and variance.

In the following section we will show that the natural estimator of the mutual information is usually biased. However, unlike in the case of the natural entropy estimator, which always underestimates its theoretical Shannon entropy, the direction of the mutual information bias depends on the underlying probability distribution p_{ij} and the sample size N.

In section 10.2, we will derive a first order approximation of this bias and discuss first applications of the corresponding correction formula. Finally, we will approximate the population variance of the natural mutual information estimates in section 10.6.

10.1 The Mutual Information Estimator is Biased

Unlike in section 9.1 where we could prove that the natural entropy estimator $\hat{H}$ always underestimates the theoretical Shannon entropy H, there is no such theorem for the mutual information. However, this does not imply that the natural estimator of the mutual information is unbiased. It rather means that the mutual information is overestimated by its natural estimator in some cases (i.e. for some probability distributions p_{ij} and some sample sizes N) and underestimated in others. The aim of this section is to develop an understanding of why there is no strict inequality about the direction of the bias (as in the case of the natural entropy estimator) and to exemplify the two cases in which the natural estimator typically overestimates and underestimates the mutual information.

The reason why the natural entropy estimator always underestimates its theoretical Shannon entropy is the concavity of the entropy function $H(\vec{p}) = -\sum_{i=1}^{M} p_i \cdot \ln(p_i)$. In contrast, the mutual information

$$I(\hat{p}) = \sum_{i,j=1}^{M} p_{ij} \cdot \ln\left(\frac{p_{ij}}{p_i \cdot q_j}\right) \tag{10.1}$$

is neither convex nor concave in its arguments p_{ij}. Here and in the following, $\hat{p}$ denotes the $M \times M$ matrix containing the elements p_{ij}. Before we start to illustrate the *local* convexity and concavity of the mutual information I by two examples, let us see what we can learn from the relation

$$I = H_X + H_Y - H_2 \tag{10.2}$$

about the convexity of I as a function of $\hat{p}$.

As we learned in section 9.1, the concavity of the Shannon entropy implies that the entropy of the arithmetic mean of two arbitrary probability vectors $\vec{p}_1$ and $\vec{p}_2$ is greater than the arithmetic mean of the corresponding Shannon entropies, i.e.

$$H(\bar{p}) \geq \frac{H(\vec{p}_1) + H(\vec{p}_2)}{2} \tag{10.3}$$

where the arithmetic mean of the probability vectors $\vec{p}_1$ and $\vec{p}_2$ is denoted by

$$\bar{p} \equiv \frac{\vec{p}_1 + \vec{p}_2}{2}. \tag{10.4}$$

Of course, the same statement holds for higher order entropies: let $\hat{p}_1$ and $\hat{p}_2$ be two arbitrary probability matrices with M rows and M columns, and let us denote the arithmetic

mean of these probability matrices by

$$\bar{p} \equiv \frac{\hat{p}_1 + \hat{p}_2}{2}. \tag{10.5}$$

Then the entropy of the average distribution, $H(\bar{p})$, is greater than the average over the entropies $H(\hat{p}_1)$ and $H(\hat{p}_2)$, i.e.

$$H(\bar{p}) \geq \frac{H(\hat{p}_1) + H(\hat{p}_2)}{2} \tag{10.6}$$

for *arbitrary* $\hat{p}_1$ and $\hat{p}_2$.

From eq. (10.4) we see that the mutual information of the average distribution, $I(\bar{p})$, is smaller than the average over the mutual information values $I(\hat{p}_1)$ and $I(\hat{p}_2)$, i.e.

$$I(\bar{p}) \leq \frac{I(\hat{p}_1) + I(\hat{p}_2)}{2} \tag{10.7}$$

if the probability matrices $\hat{p}_1$ and $\vec{p}_2$ have the same marginal distributions.

If, however, the marginal distributions of $\hat{p}_1$ and $\hat{p}_2$ are different, then the marginal entropies H_X and H_Y will increase by averaging, which diminishes (and possibly overcompensates) the increase of H_2. As we will see in the following two examples, it depends on the underlying probability distributions $\hat{p}_1$ and $\hat{p}_2$ which of the two forces ($H_X + H_Y$ versus H_2) wins over the other.

Let us start with an example that illustrates the case in which we lose mutual information in X about Y by averaging over two probability distributions. For the sake of simplicity, let X and Y be *binary* random variables, i.e., let $\hat{p}_1$ and $\hat{p}_2$ be two 2×2 matrices.

Assume the matrices $\hat{p}_1$ and $\hat{p}_2$ be those for which the mutual information is exactly 1 bit, i.e.

$$\hat{p}_1 \equiv \frac{1}{2} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{10.8}$$

and

$$\hat{p}_2 \equiv \frac{1}{2} \cdot \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{10.9}$$

In this case, the average distribution becomes

$$\bar{p} \equiv \frac{1}{2} \cdot \left(\frac{1}{2} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \frac{1}{2} \cdot \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \right) = \frac{1}{4} \cdot \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}. \tag{10.10}$$

The mutual information of this probability distribution is clearly zero, which we can intuitively understand by the following information-theoretical considerations.

Assume we have two binary sources and two perfect channels. The binary sources emit the two input symbols x_1 and x_2 with the same probability $p_1 = p_2 = 1/2$. In channel 1, the input symbol x_1 is mapped to the output symbol y_1, and the input symbol x_2 is mapped to the output symbol y_2. This yields the joint probability matrix $\hat{p}_1$. In channel 2, x_1 is mapped to y_2, and x_2 is mapped to y_1, which yields the joint probability matrix $\hat{p}_2$. Since both channels are perfect, the mutual information between the input signal X and the output signal Y is 1 bit in both cases.

Now assume that we are to receive messages from both channels. If we can distinguish both channels, i.e., if we know the channel through which the message was sent, then we can uniquely reconstruct the input signal from the received signal. In this situation, the mutual information per transmitted signal is exactly 1 bit, i.e.

$$\frac{I(\hat{p}_1) + I(\hat{p}_2)}{2} = \ln 2 = 1bit. \tag{10.11}$$

In a second experiment, we do not have knowledge about the channel through which the incoming message was sent. Somebody else is receiving the message from either of the two channels, and we are only told the received bit (y_1 or y_2), but we are not told the channel from which the signal was received. In this situation, we do not have any clue about the sent signal. If the received symbol is y_1, the input signal was x_1 with probability 1/2 (transfer through channel 1) and x_2 with probability 1/2 (transfer through channel 2). If the received symbol is y_2, the story is the same. In total, we do not obtain any information about the sent symbol x_i by obtaining y_j. Hence, the mutual information between X and Y is zero, i.e.

$$I(\bar{p}_1) = 0. \tag{10.12}$$

As we will see in later stages of this work, we *usually* lose mutual information by averaging over probability matrices. This implies that we *usually* overestimate the mutual information from finite samples. However, these statements are not strict, which we will illustrate by the following example.

Let us now consider the matrices

$$\hat{p}_1 \equiv \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \tag{10.13}$$

and

$$\hat{p}_2 \equiv \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}. \tag{10.14}$$

The mutual information between X and Y is exactly zero in both cases. However, by averaging we obtain

$$\bar{p} \equiv \frac{1}{2} \cdot \left(\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \right) = \frac{1}{2} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{10.15}$$

which yields a mutual information of 1 bit.

We can again develop an intuitive understanding of this result by imagining two senders of binary signals and two perfect channels. This time, sender 1 can only emit the input signal x_1, while sender 2 can only emit the input signal x_2. Both channels are perfect and map x_1 to y_1 and x_2 to y_2. Each channel alone (coupled to its sender) cannot transmit any information. Therefore

$$\frac{I(\hat{p}_1) + I(\hat{p}_2)}{2} = 0. \tag{10.16}$$

In contrast, the mutual information of the average probability distribution $\bar{p}$ is greater than zero. Assume again that we (as the receiver) are not told the channel through which the message was sent. The possible output symbols are y_1 and y_2, both appearing with probability $q_1 = q_2 = 1/2$. The input symbols are x_1 and x_2, and in our experiment they are sent with equal probability $p_1 = p_2 = 1/2$.

If we obtain y_1, we know that this signal traveled through channel 1. Therefore we know the input symbol must have been x_1. If the output symbol is y_2, we know the signal traveled through channel 1. Then we know the input signal must have been x_2. In both cases we can uniquely reconstruct the input symbol; hence, the mutual information between X and Y is 1 bit, i.e.

$$I(\bar{p}) = \ln 2 = 1bit. \tag{10.17}$$

What we have illustrated so far is that the mutual information is neither convex nor concave. There are regions on the simplex of the joint probability distributions $\hat{p}$ where the the mutual information $I(\hat{p})$ is *locally* convex and there are other regions for which the mutual information is *locally* concave. In the following we will study how this complicated convexity pattern influences the direction of the bias of the natural information estimator.

The natural mutual information estimator is defined by

$$\hat{I}(\hat{p}) \equiv \sum_{i,j=1}^{M} \hat{p}_{ij} \cdot \ln \left(\frac{\hat{p}_{ij}}{\hat{p}_i \cdot \hat{q}_j} \right) \tag{10.18}$$

where $\hat{p}$ be the $M \times M$ matrix containing the elements $\hat{p}_{ij}$,

$$\hat{p}_{ij} = \frac{k_{ij}}{N} \tag{10.19}$$

be the relative frequency to observe the symbol pair (x_i, y_j),

$$\hat{p}_i = \frac{k_i}{N} = \frac{1}{N}\sum_{j=1}^{M} k_{ij} = \sum_{j=1}^{M} \hat{p}_{ij} \tag{10.20}$$

be the relative frequency to observe the symbol x_i in the input sequence, and

$$\hat{q}_j = \frac{l_j}{N} = \frac{1}{N}\sum_{i=1}^{M} k_{ij} = \sum_{i=1}^{M} \hat{p}_{ij} \tag{10.21}$$

be the relative frequency to observe the symbol y_j in the output sequence.

Now consider an ensemble of observers who estimate the mutual information from a sample of size N. Let us in our first case assume the theoretical mutual information be zero, i.e., all joint probabilities p_{ij} factorize according to $p_{ij} = p_i \cdot q_j$ for all $i, j = 1, 2, ...M$. However, the estimated values of the mutual information are always greater than (or equal to) zero. Therefore, the mutual information of a Bernoulli-sequence is always overestimated by its natural estimator.

This theorem is easy to understand. Due to the finite size N of the sample, most realizations mimic statistical dependences between the symbols x_i and y_j. Only those realization for which (by chance) the observed joint frequencies $\hat{p}_{ij}$ factorize according to $\hat{p}_{ij} = \hat{p}_i \cdot \hat{q}_j$ for all $i, j = 1, 2, ...M$ yield a mutual information equal to zero, while all other realizations give a mutual information greater than zero. Therefore, the expected mutual information $I^{exp} \equiv E(\hat{I})$ is always greater than zero for Bernoulli-sequences.

In the case of weakly correlated sequences, this inequality remains the same. On average, we systematically overestimate the mutual information in finite sequences. The intuitive reason for this behavior is that the contribution to the finite size effect of the joint entropy H_2 still dominates over the contribution of the two marginal entropies H_X and H_Y. However, let us in the following paragraphs analyze an extreme case in which the natural mutual information estimates are always smaller than the true mutual information value regardless of the sample size N.

Assume the binary random process given by the joint probability matrix

$$\hat{p} \equiv \frac{1}{2} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{10.22}$$

The theoretical mutual information in X about Y is 1 bit. Let us now study the ensemble of possible observations if we restrict the sample size to $N = 1$. The only two outcomes of this sampling experiment are: observing the pair (x_1, y_1) with probability 1/2 and: observing

the pair (x_2, y_2) with probability 1/2. In case 1, the joint frequency matrix $\hat{p}_1$ estimated from a sample of size $N = 1$ is

$$\hat{p}_1 \equiv \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \tag{10.23}$$

while in case 2, the joint frequency matrix $\hat{p}_2$ is

$$\hat{p}_2 \equiv \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}. \tag{10.24}$$

In both cases the estimated mutual information is zero. Therefore, the expected mutual information I^{exp} is equal to zero. This result can be generalized to an alphabet of size M and to arbitrary joint-probability distributions $\hat{p}$: the expected mutual information drawn from a sample of size $N = 1$ is zero.

Unfortunately, this result is only of pathological interest. Let us therefore study the change of the expected mutual information with increasing sample size N. In our extreme example, where the theoretical mutual information assumes its maximum of 1 bit, the expected mutual information (i.e. the average over the observed mutual information values) is monotonically increasing with N and asymptotically approaching the theoretical mutual information from below.

In all other cases (in which the theoretical mutual information does not assume its maximum), we will observe an interesting crossover effect. For small sample sizes N, the natural estimator underestimates the mutual information, while for large sample sizes, it overestimates the theoretical mutual information value. This crossover point lies at large N when the theoretical mutual information I is large and at small N when I is small. Since statistical dependences between nucleotides in DNA sequences or between amino acids in protein sequences are usually weak, we systematically overestimate the mutual information of most DNA and protein sequences, even if they are as short as, say, 50 nucleotides or amino acids.

In the next section, we will turn to the magnitude of this systematic error and try to approximate the bias of the mutual information estimator.

10.2 The Mutual Information Bias

In this section, we will derive a first order approximation of the expected mutual information, i.e., we will approximate the expectation value of the mutual information observed in a sample of size N.

Since

$$I(\hat{p}) = -H_2(\hat{p}) + H_x(\vec{p}) + H_y(\vec{q}) \tag{10.25}$$

as shown in chapter 3, we obtain

$$\hat{I}(\hat{x}) = -\hat{H}_2(\hat{x}) + \hat{H}_x(\vec{x}) + \hat{H}_y(\vec{y}) \tag{10.26}$$

with I being the mutual information, H_2 being the 2-gram Shannon entropy, H_x being the 1-symbol Shannon entropy of the transmitter, H_y being the 1-symbol Shannon entropy of the receiver, $\hat{I}$, $\hat{H}_2$, $\hat{H}_x$, $\hat{H}_y$ being their natural estimators based on a sample of size N, $\hat{p}$ being the joint probability matrix, $\vec{p}$ and $\vec{q}$ being its corresponding marginal probability vectors, and $\hat{x}$, $\vec{x}$, and $\vec{y}$ being their corresponding relative frequencies.

Note that we are dealing with two sequences here, between which we want to estimate the mutual information as a measure of how closely these two sequences are correlated. Therefore, we distinguish between the two marginal distributions $\vec{p}$ and $\vec{q}$ as well as between $\vec{x}$ and $\vec{y}$. In chapter 11, where we introduce correlation functions, we will specify all of our results for the case of dealing with correlations between different sites within one sequence, i.e., for analyzing autocorrelations.

Exploiting eq. (10.26) and our results from chapter 9, we obtain

$$\begin{aligned} E\left(\hat{I}(\hat{x})\right) &= -E\left(\hat{H}_2(\hat{x})\right) + E\left(\hat{H}_x(\vec{x})\right) + E\left(\hat{H}_y(\vec{y})\right) && (10.27) \\ &= -H_2(\hat{p}) + \frac{M^2-1}{2\cdot N} + \mathcal{O}\left(1/N^2\right) \\ &+ H_x(\vec{p}) - \frac{M-1}{2\cdot N} + \mathcal{O}\left(1/N^2\right) \\ &+ H_y(\vec{q}) - \frac{M-1}{2\cdot N} + \mathcal{O}\left(1/N^2\right), && (10.28) \end{aligned}$$

and hence,

$$E\left(\hat{I}(\hat{x})\right) = I(\hat{p}) + \frac{(M-1)^2}{2\cdot N} + \mathcal{O}\left(1/N^2\right). \tag{10.29}$$

In the following section, we want to analyze the fluctuations of the observed mutual information values around their expectation value $E\left(\hat{I}(\hat{x})\right)$.

10.3 The Mutual Information Variance - Part I

In this section, we will deliver a first order approximation of the mean quadratic deviation of the observed mutual information values $\hat{I}(\hat{x})$ from their expectation value.

According to eq. (10.26) and denoting the covariance between two random variables a and b by $C(a,b)$, we obtain

$$\begin{aligned}\sigma^2\left(\hat{I}(\hat{x})\right) &= \sigma^2\left(\hat{H}_2(\hat{x})\right) + \sigma^2\left(\hat{H}_x(\hat{x})\right) + \sigma^2\left(\hat{H}_y(\hat{x})\right) \\ &- 2\cdot C(\hat{H}_2,\hat{H}_x) - 2\cdot C(\hat{H}_2,\hat{H}_y) + 2\cdot C(\hat{H}_x,\hat{H}_y) \qquad (10.30)\end{aligned}$$

for the mean sample variance of the mutual information between two sequences.

Naively assuming $\hat{H}_2$, $\hat{H}_x$, and $\hat{H}_y$ be mutually independent yields

$$\begin{aligned}\sigma^2\left(\hat{I}(\hat{x})\right) &= \sigma^2\left(\hat{H}_2(\hat{x})\right) + \sigma^2\left(\hat{H}_x(\hat{x})\right) + \sigma^2\left(\hat{H}_y(\hat{x})\right) \qquad (10.31)\\ &= \frac{1}{N}\cdot\left(\sum_{i,j=1}^{M} p_{ij}\cdot\ln^2(p_{ij}) - H_2^2(\hat{p})\right) + \mathcal{O}\left(1/N^2\right) \\ &+ \frac{1}{N}\cdot\left(\sum_{i=1}^{M} p_i\cdot\ln^2(p_i) - H_x^2(\vec{p})\right) + \mathcal{O}\left(1/N^2\right) \\ &+ \frac{1}{N}\cdot\left(\sum_{j=1}^{M} q_j\cdot\ln^2(q_j) - H_y^2(\vec{q})\right) + \mathcal{O}\left(1/N^2\right), \qquad (10.32)\end{aligned}$$

which, however, is much larger than the mutual information fluctuations we really observe, e.g., in all three figures displayed in chapter 13.

Remember that we understood in section 9.3 why the fluctuations of the observed Shannon entropies were about twenty times smaller than naively expected. Now we are confronted with the strange phenomenon that the real mutual information fluctuations are again smaller than the already tiny fluctuations of $\hat{H}_2(\hat{x})$.

The arising question is whether the three occurring Shannon entropies are indeed mutually uncorrelated, i.e., whether the assumption all three covariances vanish is reasonable or not.

In the following two sections, we will calculate $C(\hat{H}_2,\hat{H}_x)$, $C(\hat{H}_2,\hat{H}_y)$, as well as $C(\hat{H}_x,\hat{H}_y)$ and show

- that $\hat{H}_2$ and $\hat{H}_x$ as well as $\hat{H}_2$ and $\hat{H}_y$ are highly correlated,

which finally leads to the effect

- that the variance of the mutual information estimates almost vanishes.

10.4 Correlations between the 2-gram Shannon entropy $\hat{H}_2$ and its marginal 1-symbol Shannon entropy $\hat{H}_1$

Let us start in this section with deriving a first order approximation of the covariance $C(H_2, H_x)$ between the 2-gram Shannon entropy H_2 and the 1-symbol Shannon entropy H_x.

$$\begin{aligned} C(\hat{H}_2, \hat{H}_x) &= E\left(\hat{H}_2 \cdot \hat{H}_x\right) - E(\hat{H}_2) \cdot E(\hat{H}_x) = \\ & E\left((\hat{H}_2 - H_2) \cdot (\hat{H}_x - H_x)\right) - E(\hat{H}_2 - H_2) \cdot E(\hat{H}_x - H_x) \end{aligned} \tag{10.33}$$

Since the second term

$$E(\hat{H}_2 - H_2) \cdot E(\hat{H}_x - H_x) = \frac{(M^2-1)\cdot(M-1)}{4 \cdot N^2} + \mathcal{O}\left(1/N^3\right) \tag{10.34}$$

is negligible in the order of $1/N$, we concentrate on the first term and obtain:

$$C(\hat{H}_2, \hat{H}_x) \propto E\left(\sum_{i,j=1}^{M} \ln(p_{ij}) \cdot (x_{ij} - p_{ij}) \cdot \sum_{k=1}^{M} \ln(p_k) \cdot (x_k - p_k)\right) \tag{10.35}$$

$$= \sum_{i,j,k,l=1}^{M} \ln(p_{ij}) \cdot \ln(p_k) \cdot E\left((x_{ij} - p_{ij}) \cdot (x_{kl} - p_{kl})\right) \tag{10.36}$$

$$= \sum_{i,j=1}^{M} \ln(p_{ij}) \cdot \ln(p_i) \cdot \frac{p_{ij} \cdot (1 - p_{ij})}{N}$$

$$- \sum_{i,j,k,l=1}^{M} (1 - \delta_{ik}) \cdot (1 - \delta_{jl}) \cdot \ln(p_{ij}) \cdot \ln(p_k) \cdot \frac{p_{ij} \cdot p_{kl}}{N} \tag{10.37}$$

$$= \frac{1}{N} \sum_{i,j=1}^{M} \ln(p_{ij}) \cdot \ln(p_i) \cdot p_{ij}$$

$$- \frac{1}{N} \sum_{i,j,k,l=1}^{M} \ln(p_{ij}) \cdot \ln(p_k) \cdot p_{ij} \cdot p_{kl} \tag{10.38}$$

$$= \frac{1}{N} \sum_{i,j=1}^{M} p_{ij} \cdot \ln(p_{ij}) \cdot \ln(p_i) - \frac{1}{N} H_2 \cdot H_x \tag{10.39}$$

$$= \frac{1}{N} \cdot C\left(\ln(\hat{p}) \cdot \ln(\vec{p})\right) \tag{10.40}$$

in surprising analogy to eq. (9.55), which relates the variance of the natural Shannon entropy estimates to the logarithmic variance of their underlying probabilities.

Analogously, we obtain

$$C(\hat{H}_2, \hat{H}_y) = \frac{1}{N} \cdot C\left(\ln(\hat{p}) \cdot \ln(\vec{q})\right) + \mathcal{O}\left(1/N^2\right). \tag{10.41}$$

In other words, the covariance between the observed 2-gram Shannon entropy and one of its marginal 1-symbol Shannon entropy observed from the same sample of size N is, in a first order approximation, equal to the covariance between the logarithms of the joint probabilities p_{ij} and the logarithms of their marginal symbol probabilities p_i divided by N.

Let us eventually derive an approximation for the correlation coefficient r between $\hat{H}_1$ and $\hat{H}_2$, which is defined as the normalized covariance:

$$\begin{aligned} r(\hat{H}_1, \hat{H}_2) &\equiv \frac{C(\hat{H}_1, \hat{H}_2)}{\sqrt{\sigma^2(\hat{H}_1) \cdot \sigma^2(\hat{H}_2)}} && (10.42) \\ &\propto \frac{C(\ln(\vec{p}, \hat{p}))}{\sqrt{\sigma^2(\ln(\vec{p})) \cdot \sigma^2(\ln(\hat{p}))}} && (10.43) \\ &= r(\ln(\vec{p}), \ln(\hat{p})) && (10.44) \end{aligned}$$

This is a really noticeable result, since the right hand side of this equality does not depend on the sample size N. It states that the correlation coefficient between the natural estimates of the statistics H_1 and H_2 is independent of the number of sample points and given by the correlation coefficient between the logarithm of the joint probabilities p_{ij} and the logarithm of their marginal symbol probabilities p_i.

Although this is a clear mathematical result, we are again very far from intuitively understanding the relation between the sample correlator on the one side and the population correlator on the other side of the equality.

10.5 Correlations between the two marginal 1-symbol Shannon entropy estimates $\hat{H}_x$ and $\hat{H}_y$

In this section, we will derive a first order approximation for the covariance $C(\hat{H}_x, \hat{H}_y)$ between the Shannon entropy estimates $\hat{H}_x$ and $\hat{H}_y$.

Since we again realize that the leading term of the bias is proportional to $1/N$ for both estimators, $\hat{H}_x$ and $\hat{H}_y$, we obtain:

$$C(\hat{H}_x, \hat{H}_y) = E\left((\hat{H}_x - H_x) \cdot (\hat{H}_y - H_y)\right)$$

$$
\begin{aligned}
- \quad & E(\hat{H}_x - H_x) \cdot E(\hat{H}_y - H_y) && (10.45)\\
\propto \quad & E\left(\sum_{i=1}^{M} \ln(p_i) \cdot (x_i - p_i) \cdot \sum_{j=1}^{M} \ln(q_j) \cdot (x_j - q_j)\right) && (10.46)\\
= \quad & \sum_{i,j=1}^{M} \ln(p_i) \cdot \ln(q_j) \cdot E\left((x_i - p_i) \cdot (x_j - q_j)\right) && (10.47)
\end{aligned}
$$

by neglecting all $\mathcal{O}(1/N^2)$ terms.

If we assume the two considered sequences be independent, we end up with

$$C(\hat{H}_x, \hat{H}_y) = 0 + \mathcal{O}\left(1/N^2\right) \tag{10.48}$$

as expected.

If we, however, consider the opposite limiting case, namely the two sequences be identical, which occurs if we are interested in autocorrelations, then we obtain:

$$C(\hat{H}_x, \hat{H}_y) = \sqrt{\sigma^2(\hat{H}_x) \cdot \sigma^2(\hat{H}_y)} + \mathcal{O}\left(1/N^2\right). \tag{10.49}$$

In the remainder of this section, we will derive a first order approximation for the general case where we analyze autocorrelations in a sequence of length N'. Let the correlation length in which we are interested be k. Then, our sample size reduces to $N = N' - k$, since we do not want to introduce artificial correlations by applying cyclic boundary conditions. Hence, we count the first and the last k symbols in our given sequence once and the remaining $N' - 2 \cdot k$ symbols in the middle of the sequence twice, since we count $N' - k$ pairs of symbols in total.

How this partially overlapping counting affects the expectation value of correlation functions is presented in chapter 11.

Here, we exploit the finding

$$E\left((x_i - p_i) \cdot (y_i - p_i)\right) = \frac{N' - 2 \cdot k}{(N' - k)^2} \cdot p_i \cdot (1 - p_i) \tag{10.50}$$

for all $i = 1, 2, ..., M$ and

$$E\left((x_i - p_i) \cdot (y_j - p_j)\right) = -\frac{N' - 2 \cdot k}{(N' - k)^2} \cdot p_i \cdot p_j \tag{10.51}$$

for all $i, j = 1, 2, ..., M$ with $i \neq j$ to obtain

$$C(\hat{H}_x, \hat{H}_y) \quad = \quad \frac{N' - 2 \cdot k}{(N' - k)^2} \sum_{i=1}^{M} \ln^2(p_i) \cdot p_i \cdot (1 - p_i) \tag{10.52}$$

$$
\begin{aligned}
- & \frac{N' - 2 \cdot k}{(N' - k)^2} \sum_{i,j=1}^{M} (1 - \delta_{ij}) \cdot \ln(p_i) \cdot \ln(p_j) \cdot p_i \cdot p_j && (10.53) \\
= & \frac{N' - 2 \cdot k}{(N' - k)^2} \sum_{i=1}^{M} \ln^2(p_i) \cdot p_i && (10.54) \\
- & \frac{N' - 2 \cdot k}{(N' - k)^2} \sum_{i,j=1}^{M} \ln(p_i) \cdot \ln(p_i) \cdot p_i \cdot p_j && (10.55) \\
= & \frac{N' - 2 \cdot k}{(N' - k)^2} \cdot \left(\sum_{i=1}^{M} p_i \cdot \ln^2(p_i) - H_1^2 \right) && (10.56)
\end{aligned}
$$

and thus

$$
C(\hat{H}_x, \hat{H}_y) = \frac{N' - 2 \cdot k}{(N' - k)^2} \cdot \sigma^2 \left(\ln(\vec{p}) \right) + \mathcal{O}\left(1/N^2\right). \tag{10.57}
$$

This result perfectly produces both of our limiting cases, namely

$$
C(\hat{H}_x, \hat{H}_y) = \sigma^2 \left(\ln(\vec{p}) \right) + \mathcal{O}\left(1/N^2\right) \tag{10.58}
$$

for $k/N' \to 0$ and

$$
C(\hat{H}_x, \hat{H}_y) = 0 + \mathcal{O}\left(1/N^2\right) \tag{10.59}
$$

for $k \geq N'/2$.

10.6 The Mutual Information Variance - Part II

In section 10.4, we have seen that the observed 2-gram Shannon entropies are highly correlated with the corresponding 1-symbol Shannon entropy observations. This is the reason why the fluctuations of the observed mutual information are by far smaller than naively expected.

In this section, we will finish the derivation of the mutual information variance observed from a sample of size N, i.e., from a sequence of N units where the units might, for example, be amino acids in proteins or dicodons in the case of DNA sequences.

According to eq. (10.30) and exploiting our results from sections 10.3, 10.4, and 10.5, we obtain:

$$
\begin{aligned}
& \sigma^2 \left(\hat{I}(\hat{x}) \right) \\
= & \sigma^2 \left(\hat{H}_2(\hat{x}) \right) + \sigma^2 \left(\hat{H}_x(\hat{x}) \right) + \sigma^2 \left(\hat{H}_y(\hat{x}) \right) \\
- & 2 \cdot C(\hat{H}_2, \hat{H}_x) - 2 \cdot C(\hat{H}_2, \hat{H}_y) + 2 \cdot C(\hat{H}_x, \hat{H}_y) && (10.60)
\end{aligned}
$$

$$\propto \frac{1}{N}\cdot\sigma^2\left(\ln(\hat{p})\right)+\frac{1}{N}\cdot\sigma^2\left(\ln(\vec{p})\right)+\frac{1}{N}\cdot\sigma^2\left(\ln(\vec{q})\right) \tag{10.61}$$

$$-\frac{2}{N}\cdot C\left(\ln(\hat{p}),\ln(\vec{p})\right)-\frac{2}{N}\cdot C\left(\ln(\hat{p}),\ln(\vec{q})\right)+\frac{2}{N}\cdot C\left(\ln(\vec{p}),\ln(\vec{q})\right) \tag{10.62}$$

$$=\frac{1}{N}\cdot\sigma^2\left(\ln(\hat{p})-\ln(\vec{p})-\ln(\vec{q})\right) \tag{10.63}$$

and thus

$$\sigma^2\left(\hat{I}(\hat{x})\right)=\frac{1}{N}\cdot\sigma^2\left(\ln\left(\frac{\hat{p}}{\vec{p}\cdot\vec{q}}\right)\right)+\mathcal{O}\left(1/N^2\right) \tag{10.64}$$

for our limiting case $k << N$.

10.7 Summary of Chapters 9 and 10

Chapters 9 and 10 were entirely devoted to presenting approximations of statistical and systematic errors that uncircumventably occur by estimating entropies or the mutual information function from finite samples.

In both sections, we payed much effort to outline the conceptional techniques by which we analyzed these errors, since these concepts are, in principle, applicable to approximate statistical and systematic errors of any measure of interest.

The purpose of this summary is now to display the main results, in particular those that we found new, on two pages.

Let $\vec{p}=(p_1,p_2,...,p_{M_X})$ and $\vec{q}=(q_1,q_2,...,q_{M_Y})$ be two probability vectors and $\hat{P}$ the $M_X\times M_Y$ matrix containing the elements P_{ij} where $i=1,2,...M_X$ and $j=1,2,...M_Y$.

Then the negative average logarithms of the probabilities p_i, q_j, and P_{ij} are defined as the Shannon entropies H_X, H_Y, and H_2, respectively.

$$H_X = -E\left(\ln\left(\vec{p}\right)\right) \tag{10.65}$$

$$H_Y = -E\left(\ln\left(\vec{q}\right)\right) \tag{10.66}$$

$$H_2 = -E\left(\ln\left(\hat{P}\right)\right) \tag{10.67}$$

The natural estimators, $\hat{H}_X$, $\hat{H}_Y$, and $\hat{H}_2$, of these Shannon entropies are defined as the entropy function applied to the maximum likelihood estimators $\hat{p}_i$, $\hat{q}_j$, and $\hat{P}_{ij}$ of the corresponding probabilities p_i, q_j, and P_{ij}.

$$\hat{H}_X = -E\left(\ln\left(\hat{p}_i\right)\right) \tag{10.68}$$

$$\hat{H}_Y = -E\left(\ln\left(\hat{q}_j\right)\right) \tag{10.69}$$

$$\hat{H}_2 = -E\left(\ln\left(\hat{P}_{ij}\right)\right) \tag{10.70}$$

The biases, $\Delta\hat{H}_X$, $\Delta\hat{H}_Y$, and $\Delta\hat{H}_2$, of these entropy estimators are given by

$$\Delta\hat{H}_X = -\frac{M_X - 1}{2 \cdot N} \tag{10.71}$$

$$\Delta\hat{H}_Y = -\frac{M_Y - 1}{2 \cdot N} \tag{10.72}$$

$$\Delta\hat{H}_2 = -\frac{M_X \cdot M_Y - 1}{2 \cdot N} \tag{10.73}$$

by neglecting $\mathcal{O}(1/N^2)$ terms if N denotes the sample size.

In the following, we will present all results in this first order approximation and thus drop mentioning the higher order terms $\mathcal{O}(1/N^2)$.

The variances, $\sigma^2\left(\hat{H}_X\right)$, $\sigma^2\left(\hat{H}_Y\right)$, and $\sigma^2\left(\hat{H}_2\right)$, of the corresponding natural entropy estimators are given by the variances of the logarithms of the corresponding probabilities[1] divided by N.

$$\sigma^2\left(\hat{H}_X\right) = \frac{1}{N} \cdot \sigma^2\left(\ln\left(\vec{p}\right)\right) \tag{10.74}$$

$$\sigma^2\left(\hat{H}_Y\right) = \frac{1}{N} \cdot \sigma^2\left(\ln\left(\vec{q}\right)\right) \tag{10.75}$$

$$\sigma^2\left(\hat{H}_2\right) = \frac{1}{N} \cdot \sigma^2\left(\ln\left(\hat{P}\right)\right) \tag{10.76}$$

Please note that the left hand side of the above equalities are functions of the sample, whereas the variances of the logarithms of the probabilities on the right hand side are functions of the population parameters p_i, q_j, or P_{ij}, respectively, and thus sampling independent.

The only dependence of the right hand sides on the sample is introduced by the factor $1/N$, i.e., by dividing by the sample size.

The covariances $C\left(\hat{H}_X, \hat{H}_Y\right)$, $C\left(\hat{H}_X, \hat{H}_2\right)$, and $C\left(\hat{H}_Y, \hat{H}_2\right)$ between the three entropy estimates $\hat{H}_X$, $\hat{H}_Y$, and $\hat{H}_2$ are given by the covariances between the logarithms of the

[1] The variance of the entropy estimates is really proportional to the variance of the logarithms of the corresponding probabilities and not to the variance of the logarithms of the probability estimates. The interesting story about these equalities is that they relate the variance of estimates with the variance of population parameters.

corresponding probabilities[2] divided by N.

$$C\left(\hat{H}_X, \hat{H}_Y\right) = \frac{1}{N} \cdot C\left(\ln\left(\vec{p}\right), \ln\left(\vec{q}\right)\right) \tag{10.77}$$

$$C\left(\hat{H}_X, \hat{H}_2\right) = \frac{1}{N} \cdot C\left(\ln\left(\vec{p}\right), \ln\left(\hat{P}\right)\right) \tag{10.78}$$

$$C\left(\hat{H}_Y, \hat{H}_2\right) = \frac{1}{N} \cdot C\left(\ln\left(\vec{q}\right), \ln\left(\hat{P}\right)\right) \tag{10.79}$$

The corresponding correlation coefficients $r\left(\hat{H}_X, \hat{H}_Y\right)$, $r\left(\hat{H}_X, \hat{H}_2\right)$, and $r\left(\hat{H}_Y, \hat{H}_2\right)$ between the three entropy estimates $\hat{H}_X$, $\hat{H}_Y$, and $\hat{H}_2$, which all are functions of the given sample, do not depend on the sample size N and are given by the correlation coefficients between the logarithms of the corresponding probabilities in a first order approximation.

$$r\left(\hat{H}_X, \hat{H}_Y\right) = r\left(\ln\left(\vec{p}\right), \ln\left(\vec{q}\right)\right) \tag{10.80}$$

$$r\left(\hat{H}_X, \hat{H}_2\right) = r\left(\ln\left(\vec{p}\right), \ln\left(\hat{P}\right)\right) \tag{10.81}$$

$$r\left(\hat{H}_Y, \hat{H}_2\right) = r\left(\ln\left(\vec{q}\right), \ln\left(\hat{P}\right)\right) \tag{10.82}$$

The mutual information I between two random variables X and Y is defined as the expectation value of the numbers

$$\ln\left(\frac{P_{ij}}{p_i \cdot q_j}\right), \tag{10.83}$$

where we copy all denotations from above, i.e.,

$$I = E\left(\ln\left(\frac{P_{ij}}{p_i \cdot q_j}\right)\right). \tag{10.84}$$

The natural estimator $\hat{I}$ of the mutual information is defined as the mutual information function calculated from the maximum likelihood estimators $\hat{p}_i$, $\hat{q}_j$, and $\hat{P}_{ij}$ of the corresponding probabilities p_i, q_j, and P_{ij}.

$$\hat{I} = E\left(\ln\left(\frac{\hat{P}_{ij}}{\hat{p}_i \cdot \hat{q}_j}\right)\right) \tag{10.85}$$

The bias $\Delta\hat{I}$ of this estimator is given by

$$\Delta\hat{I} = \frac{(M_X - 1) \cdot (M_Y - 1)}{2 \cdot N}. \tag{10.86}$$

[2]Please note that we could again relate the covariance of estimates to the covariance of population parameters, i.e., the covariances on the right hand side of the displayed equalities do not mean the covariances of the logarithmic probability estimates.

Please note the missing minus sign, i.e., we systematically overestimate the mutual information whereas we systematically underestimate all Shannon entropies.

The variance of the mutual information estimates as a function of the given sample, i.e., the variance of the numbers $E\left(\ln\left(\frac{\hat{P}_{ij}}{\hat{p}_i \cdot \hat{q}_j}\right)\right)$, is equal to the variance of the numbers $\ln\left(\frac{P_{ij}}{p_i \cdot q_j}\right)$, the expectation value of which is defined as the mutual information, divided by the sample size N.

$$\sigma^2\left(\hat{I}\right) = \frac{1}{N} \cdot \sigma^2\left(\ln\left(\frac{P_{ij}}{p_i \cdot q_j}\right)\right) \tag{10.87}$$

Please remember at this point the valuable property that the right hand side of this equality vanishes for random numbers X and Y that are statistically independent, which implies the disappearance of all mutual information fluctuations in the order of $1/N$.

Chapter 11

Statistical Properties of Natural Estimators of Correlation Functions

Power spectra, random walk studies, or the direct calculation of autocorrelation functions make up most of the tools commonly used to detect long-range correlations in time series, natural texts, pieces of music, econometric data, or DNA sequences. However, since the power spectrum is defined as the Fourier transform of the corresponding correlation function and the growth of random walk fluctuations corresponds to an integrated autocorrelation function [Stanley 1994], we will concentrate on statistical properties of correlation function estimators from finite samples in this chapter.

In analogy to our previous chapters 9 and 10, we will analyze the bias of the natural covariance estimator $\hat{C}_{ij}(k)$ and the variance of its estimates.

The covariance $C(x,y)$ between two numerical random variables x and y is defined by

$$C(x,y) \equiv \sum_{i,j=1}^{M} x_i \cdot y_j \cdot P(x_i,y_j) - \sum_{i=1}^{M} x_i \cdot p(x_i) \cdot \sum_{j=1}^{M} y_j \cdot q(y_j) \tag{11.1}$$

$$= E(x \cdot y) - E(x) \cdot E(y) = E\left((x - E(x)) \cdot (y - E(y))\right), \tag{11.2}$$

if the state space is discrete and M-dimensional for both variables x and y.

Since we are interested in measuring correlations between symbols within one sequence, we will specify this definition in the following and introduce the dependence of the covariance on the integer number k denoting the correlation length.

Let a_i be the real numbers we assigned to the symbols A_i where $i = 1, 2, ..., M$ and M be the alphabet size, e.g., $M = 20$ in amino acid sequences or $M = 4^6 = 4096$ if we consider dicodons in DNA or RNA sequences.

By formally identifying the position n in our sequence with the time t in a time series $a(t)$, we can easily introduce the correlation function

$$C_{\vec{a}}(k) \equiv E(a(n) \cdot a(n+k)) - E(a(n)) \cdot E(a(n+k)) \tag{11.3}$$

where the average is taken over all positions n, and k is the size of the gap between the two multiplied numbers.

The subscript $\vec{a}$ is to remember that we originally want to analyze correlations within a symbolic sequence and just chose the projection vector $\vec{a} = (a_1, a_2, ..., a_M)$ to assign numbers to symbols in order to define a correlation function.

The deeper sense is that correlation functions are not invariant under coordinate transforms, i.e., if we change the assignment of numbers to symbols, we may (and in general do) obtain a different correlation function.

Before we are ready to rewrite eq. (11.3), we have to introduce the probabilities p_i to find the symbol A_i at any position in our given sequence as well as the probabilities $P_{ij}(k)$ to find the symbol pair (A_i, A_j) in a distance k. Note that $P_{ij}(k = 1)$ reflects the probability to find the two adjacent symbols A_i and A_j at any position in our considered sequence.

If we now replace the sum over the entire sequence by the sum over all possible combinations of two symbols multiplied by their corresponding probabilities to occur, we obtain:

$$C_{\vec{a}}(k) = \vec{a} \cdot \hat{D}(k) \cdot \vec{a}^T \tag{11.4}$$

by defining $\hat{D}(k)$ as the $M \times M$ matrix containing the elements

$$D_{ij}(k) \equiv P_{ij}(k) - p_i \cdot p_j. \tag{11.5}$$

The elements $D_{ij}(k)$ are often referred to as correlation functions between the symbols A_i and A_j, but in order not to mismatch $C_{ij}(k)$ and $C_{\vec{a}}(k)$, we denote the $C_{ij}(k)$ correlation functions by $D_{ij}(k)$ and call the matrix $\hat{D}(k)$ *dependence matrix.*

Since all correlation functions $C_{\vec{a}}(k)$ are bilinear forms of the dependence matrix, we can easily derive all statistical properties of the $C_{\vec{a}}(k)$ estimator, if we know the statistical behavior of the $D_{ij}(k)$ estimator.

Let us thus define the natural estimator $\hat{D}_{ij}(k)$ of the matrix element $D_{ij}(k)$ by

$$\hat{D}_{ij}(k) = \hat{P}_{ij}(k) - \hat{p}_i \cdot \hat{p}_j \tag{11.6}$$

where $\hat{P}_{ij}(k)$, $\hat{p}_i$, and $\hat{p}_j$ are the maximum likelihood estimators the probabilities $P_{ij}(k)$, p_i, and p_j, respectively.

Please note that we introduce some finite size effects by identifying the estimates $\hat{p}_i$ and $\hat{p}_j$: since we always deal with finite sequences, the possible difference between the frequency composition of the first and the last k symbols is not negligible.

Hence, we prefer to distinguish between the probability p_i to observe the symbol A_i 'on the left hand side' and the probability q_j to observe the symbol A_j 'on the right hand side' although we know $p_i = q_j$ in all stationary sequences.

We will see that this distinction is essential for understanding that the bias of the natural $\hat{D}_{ij}(k)$ estimator does not only depend on the sample size N, but also on the correlation distance k.

In the following section, we will derive an exact analytic expression for the expectation value of the $\hat{D}_{ij}(k)$ estimates and show that autocorrelations are always underestimated whereas crosscorrelations are always overestimated.

Section 11.2 is then devoted to deriving a first order approximation of the $\hat{D}_{ij}(k)$ variance.

11.1 The Correlation Function Bias

In this section, we will derive an exact analytic expression for the expectation value of the natural correlation function estimates $\hat{D}_{ij}(k)$ and show that the corresponding natural estimators are always biased for $k < N'/2$ where N' denotes the sequence length.

Let $P_{ij}(k)$, p_i, and q_j be the probabilities defined in the previous section. Then the expectation value — strictly speaking, the expectation function — of the natural correlation function estimates is given by

$$E\left(\hat{D}_{ij}(k)\right) = E\left(\hat{P}_{ij}(k)\right) - E\left(\hat{p}_i \cdot \hat{q}_j\right). \tag{11.7}$$

If we denote the number of symbol pairs we are counting by

$$N = N' - k \tag{11.8}$$

for $k < N$, we obtain

$$E\left(\hat{P}_{ij}(k)\right) = P_{ij}(k) \tag{11.9}$$

and

$$E\left(\hat{p}_i \cdot \hat{q}_i\right) - p_i \cdot q_i = \frac{N - 2 \cdot k}{(N-k)^2} \cdot p_i \cdot (1 - q_i) \tag{11.10}$$

for all $i = 1, 2, ..., M$ as well as

$$E\left(\hat{p}_i \cdot \hat{q}_j\right) - p_i \cdot q_j = -\frac{N - 2 \cdot k}{(N-k)^2} \cdot p_i \cdot q_j \tag{11.11}$$

for all $i, j = 1, 2, ..., M$ and $i \neq j$.

Let us start the proof with denoting the absolute frequency by which we find the symbol pair (A_i, A_j) in our sequence of length N' by $N_{ij}(k)$. Then the absolute frequency N_i^l of observing the symbol A_i on the left hand side is given by

$$N_i^l = \sum_{j=1}^{M} N_{ij}(k) \tag{11.12}$$

and the frequency of the symbol A_j that we find on the right hand side is

$$N_j^r = \sum_{i=1}^{M} N_{ij}(k) \tag{11.13}$$

Since we count $N' - 2 \cdot k$ symbols twice, N_i^l and N_j^r are not independent. It is exactly this dependence that introduces the bias of the correlation function.

To get a handle of this bias, let us decompose the frequencies N_i^l and N_j^r into the frequencies N_i^0 and N_j^0 of the symbols that we have counted twice and the frequencies N_i^{l-0} and N_j^{r-0} corresponding to the first and last k symbols in our sequence.

The following sketch might illustrate these definitions.

$$\underbrace{ACCT}_{N^{l-0}}\underbrace{GTATACGGGTTATCCATACTGGT}_{N^0}\underbrace{CTGG}_{N^{r-0}} \tag{11.14}$$

The length of this given DNA sequence is $N' = 30$ bp. Let us assume we are interested in the correlation function between the nucleotides A and G and want to estimate its value for $k = 4$. Then, $N = N' - k = 26$, i.e., we have to count 26 symbol pairs, and therefore, we will count $N' - 2 \cdot k = 22$ symbols twice.

Hence, $N_A^{l-0} = 1$, $N_A^0 = 5$, $N_G^0 = 6$, and $N_G^{r-0} = 2$ in our example.

Now we see:

$$\begin{aligned} E\left(N_i^l \cdot N_j^r\right) &= E\left((N_i^{l-0} + N_i^0) \cdot (N_j^{r-0} + N_j^0)\right) && (11.15) \\ &= E\left(N_i^{l-0} \cdot N_j^{r-0}\right) + E\left(N_i^{l-0} \cdot N_j^0\right) \\ &+ E\left(N_i^0 \cdot N_j^{r-0}\right) + E\left(N_i^0 \cdot N_j^0\right) && (11.16) \\ &= k^2 \cdot p_i \cdot q_j + k \cdot (N' - 2 \cdot k) \cdot p_i \cdot q_j \\ &+ (N' - 2 \cdot k) \cdot k \cdot p_i \cdot q_j + (N' - 2 \cdot k)^2 \cdot p_i \cdot q_j \\ &+ (N' - 2 \cdot k) \cdot Q_{ij} && (11.17) \end{aligned}$$

with

$$Q_{ij} = -p_i \cdot q_j \tag{11.18}$$

for all $i, j = 1, 2, ..., M$ with $i \neq j$ and

$$Q_{ii} = p_i \cdot (1 - q_i) \tag{11.19}$$

for all $i = 1, 2, ..., M$, which proves eqs. (11.10) and (11.11).

Hence, we can state that the natural correlation function estimator is biased for all $P_{ij}(k)$ where its bias is given by

$$E\left(\hat{D}_{ii}(k)\right) - D_{ii}(k) = -\frac{N' - 2 \cdot k}{(N' - k)^2} \cdot p_i \cdot (1 - q_i) \tag{11.20}$$

for all $i = 1, 2, ..., M$ and

$$E\left(\hat{D}_{ij}(k)\right) - D_{ij}(k) = \frac{N' - 2 \cdot k}{(N' - k)^2} \cdot p_i \cdot q_j \tag{11.21}$$

for all $i, j = 1, 2, ..., M$ with $i \neq j$.

In other words, all diagonal terms of any given auto-covariance matrix are systematically underestimated whereas all non-diagonal terms are systematically overestimated.

11.2 The Variance of the Natural Correlation Function Estimates

In this section, we will derive a first order approximation for the variance of the correlation functions $D_{ij}(k)$. We will learn that, in contrast to the mutual information variance, these fluctuations do not vanish as the P_{ij} approach $p_i \cdot q_j$. This qualitative difference will give

rise to some important conclusions about the preferable measure of statistical dependences when DNA sequences (which are known to exhibit only weak 2-point-correlations) are to be analyzed.

In the following, we will restrict ourselves to the limiting case $k << N'$ and drop the k dependence for the sake of simplicity.

Let us now start deriving the sample variance of the natural correlation function estimates by defining

$$\Delta P_{ij} \equiv \hat{P}_{ij} - P_{ij} \tag{11.22}$$

$$\Delta p_i \equiv \hat{p}_i - p_i \tag{11.23}$$

$$\Delta q_j \equiv \hat{q}_j - q_j \tag{11.24}$$

for all $i, j = 1, 2, ..., M$, which denote the deviations between the relative frequencies observed in a sample of size N and their corresponding probabilities.

The squared correlation estimate fluctuations can be rewritten as

$$\sigma^2\left(\hat{D}_{ij}\right) = E\left(\hat{D}_{ij} - D_{ij}\right)^2 - E^2\left(\hat{D}_{ij} - D_{ij}\right) \tag{11.25}$$

$$= E\left(\hat{P}_{ij} - P_{ij} - (\hat{p}_i \cdot \hat{q}_j - p_i \cdot q_j)\right)^2 + \mathcal{O}\left(1/N^2\right), \tag{11.26}$$

since the squared bias

$$E^2\left(\hat{D}_{ij} - D_{ij}\right) \propto \frac{1}{N^2} \cdot Q_{ij}^2 \tag{11.27}$$

is zero in the order of $1/N$.

This yields

$$\begin{aligned} \sigma^2\left(\hat{D}_{ij}\right) &\propto E\left(\Delta P_{ij} - p_i \cdot \Delta q_j - q_j \cdot \Delta p_i - \Delta p_i \cdot \Delta q_j\right)^2 \\ &\propto E\left(\Delta^2 P_{ij} - 2 \cdot p_i \cdot \Delta P_{ij} \cdot \Delta q_j - 2 \cdot q_j \cdot \Delta P_{ij} \cdot \Delta p_i\right) \\ &+ E\left(p_i^2 \cdot \Delta^2 q_j + 2 \cdot p_i \cdot q_j \cdot \Delta p_i \cdot \Delta q_j + q_j^2 \cdot \Delta^2 p_i\right), \end{aligned} \tag{11.28}$$

where the symbol $\propto$ means that we neglect all terms in the order of $1/N^2$.

From appendix I and by some simple arithmetics we obtain:

$$E\left(\Delta^2 P_{ij}\right) = \frac{1}{N} \cdot P_{ij} \cdot (1 - P_{ij}) \tag{11.29}$$

$$E\left(\Delta^2 p_i\right) = \frac{1}{N} \cdot p_i \cdot (1 - p_i) \tag{11.30}$$

$$
\begin{aligned}
E\left(\Delta^2 q_j\right) &= \frac{1}{N}\cdot q_j\cdot(1-q_j) && (11.31)\\
E\left(\Delta P_{ij}\cdot\Delta p_i\right) &= \sum_{k\neq j} E\left(\Delta P_{ij}\cdot\Delta P_{ik}\right)+E\left(\Delta P_{ij}\cdot\Delta P_{ij}\right)\\
&= \sum_{k\neq i} -\frac{1}{N}\cdot P_{ij}\cdot P_{ik}+\frac{1}{N}\cdot P_{ij}\cdot(1-P_{ij})\\
&= \frac{1}{N}\cdot P_{ij}\cdot(1-p_i) && (11.32)\\
E\left(\Delta P_{ij}\cdot\Delta q_j\right) &= \frac{1}{N}\cdot P_{ij}\cdot(1-q_j) && (11.33)\\
E\left(\Delta p_i\cdot\Delta q_j\right) &= \sum_{k\neq j} E\left(\Delta P_{ik}\cdot\Delta q_j\right)+E\left(\Delta P_{ij}\cdot\Delta q_j\right)\\
&= \sum_{k\neq j}\sum_{l} E\left(\Delta P_{ik}\cdot\Delta P_{lj}\right)+E\left(\Delta P_{ij}\cdot\Delta q_j\right)\\
&= \sum_{k\neq j}\sum_{l} -\frac{1}{N}\cdot P_{ik}\cdot P_{lj}+\frac{1}{N}\cdot P_{ij}\cdot(1-q_j)\\
&= \sum_{k\neq j} -\frac{1}{N}\cdot P_{ik}\cdot q_j+\frac{1}{N}\cdot P_{ij}\cdot(1-q_j)\\
&= \frac{1}{N}\cdot(P_{ij}-p_i\cdot q_j)=\frac{1}{N}\cdot D_{ij} && (11.34)
\end{aligned}
$$

We derived eq. (11.33) in complete analogy to eq. (11.32), and we used eq. (11.33) to derive eq. (11.34). Note that eq. (11.33) is similar to the equations we obtained for the variance of the entropy and mutual information estimates in chapters 9 and 10: the covariance between the estimates of the probabilities p_i and q_j is equal to the covariance between the symbols A_i and A_j divided by the sample size N. This implies that the covariance between the estimates of the probabilities p_i and q_j vanishes in the order of $1/N$ if the symbols A_i and A_j are un-correlated, i.e., if $D_{ij}=0$ for all $i,j=1,2,...M$.

By summing up all terms in eq. (11.28) we obtain

$$
\begin{aligned}
\sigma^2\left(\hat{D}_{ij}\right) &\propto \frac{1}{N}\cdot\left(P_{ij}\cdot(1-P_{ij})-2\cdot P_{ij}\cdot p_i\cdot(1-q_j)-2\cdot P_{ij}\cdot q_j\cdot(1-p_i)\right.\\
&+ \left. p_i^2\cdot q_j\cdot(1-q_j)+q_j^2\cdot p_i\cdot(1-p_i)+2\cdot p_i\cdot q_j\cdot(D_{ij})\right). && (11.35)
\end{aligned}
$$

By eliminating P_{ij} in favor of D_{ij} and after some algebra we can express this result for the variance of the correlation function estimates as follows:

$$
\begin{aligned}
\sigma^2\left(\hat{D}_{ij}\right) &\propto \frac{1}{N}\cdot D_{ij}\cdot\left((1-2\cdot p_i)\cdot(1-2\cdot q_j)-D_{ij}\right)\\
&+ \frac{1}{N}\cdot p_i\cdot q_j\cdot(1-p_i)\cdot(1-q_j) && (11.36)
\end{aligned}
$$

for all $i, j = 1, 2, ..., M$.

For Bernoulli sequences, i.e. sequences for which $D_{ij} = 0$ for all $i, j = 1, 2, ..., M$, this result simplifies to

$$\sigma^2\left(\hat{D}_{ij}\right) = \frac{1}{N} \cdot p_i \cdot q_j \cdot (1 - p_i) \cdot (1 - q_j) + \mathcal{O}\left(1/N^2\right) \tag{11.37}$$

for all $i, j = 1, 2, ..., M$.

In contrast to the fluctuations of the mutual information function, which vanish in the order of $1/N$ if the corresponding random variables approach statistical independence, the fluctuations of the correlation functions remain in the order of $1/\sqrt{N}$ even in the case of statistical independence.

11.3 Summary

In this chapter we investigated statistical properties, specifically the bias and the variance, of the natural estimator of correlation functions. We derived an exact analytic expression of the correlation bias, and thus we could show that correlation functions are always biased for $k < N/2$. We derived a first-order approximation of the correlation variance, which will turn out to be of practical importance in later chapters, in which we will estimate correlation functions from experimental data sets, such as DNA and amino acid sequences.

Chapter 12

Statistical Properties of Natural Estimators of Generalized Entropies

12.1 The Generalized Entropy Bias

This chapter is devoted to asymptotic length corrections of the entropy bias of $\widehat{H_q}$ and $\widetilde{K_q}$. As shown by numerical simulations in sections 4 and 5, though the variance is significantly small both Bayes estimators produce biased entropy estimates. It is a general feature that many estimators, in particular those minimizing the variance, share this property of being biased. Consequently, the systematic deviation of the expectation value of the estimated entropies from the true entropy value, namely the bias, has to be calculated and taken into account in order to correct the bias of the observed estimates. Explicitly,

$$\Delta\widehat{H_q} = \mathrm{E}\widehat{H_q}(\vec{N}) - H_q\left(\vec{p}\right) = \frac{1}{\ln 2}\frac{1}{1-q}\left(\sum_{i=1}^{M}\Delta\widehat{p_i^q}\right) \tag{12.1}$$

defines the bias of the estimator $\widehat{H_q}$. Here by E we denote the expectation value with respect to the multinomial distribution: $\mathrm{E}(\cdot) = \sum_{(N_1,\ldots,N_M)} P(\vec{N}|\vec{p})(\cdot)\delta(\sum_{i=1}^{M} N_i - N)$. Clearly, an unbiased statistic satisfies $\Delta(\cdot) = 0$.

The problem encountered in deriving the bias of entropy estimators is that it is difficult to obtain a closed form expression. However, in this case one may still obtain an approximation to the exact bias, for example, by expanding a power-series around the true

values of $\widehat{p_i^q}$ and applying E to each individual term within this series. The underlying idea exploits the fact that any probability distribution can, in principle, be extensively described by all of its moments. For the Bayes estimators derived in this work, this applies to $q = n, n > 1$. Expanding the exact entropy bias as a series in terms of $(1/N)^d$ with $d = 1, 2, \ldots$, we arrive at a $d = 1$ approximation by Taylor-expanding the entropies in powers of $(f_i - p_i)^m$, $m \in \mathcal{N}$, and truncating this series after the quadratic term.

As principles of this technique have been discussed in detail, e.g. in [Harris 1975, Holste 1997], we will not elaborate on this in further detail here, but only present the final results of the $\mathcal{O}(1/N)$ approximation of the entropy bias for the case $q = 2$. Since

$$\Delta\widehat{p_i^q}_{|q=2} = \left[2\left(2Np_i + 1\right) - \left(2N + M\right)Mp_i^2\right] \Big/ \left(N + M\right)^2 \tag{12.2}$$

we obtain the entropy bias of the order-2 Tsallis entropy as

$$\Delta\widehat{H}_2 = -\frac{1}{\ln 2}\frac{\left(2N + M\right)\left(2 - MZ_2\right)}{\left(N + M\right)^2} \tag{12.3}$$

(note: the approximation is exact for the case $q = 2$). In order to obtain order-2 Tsallis entropy estimates that are unbiased in $\mathcal{O}(1/N)$, we define the estimator

$$\widehat{H}_2^{(d=1)} = \widehat{H}_2 + \frac{1}{\ln 2}\frac{\left(2N + M\right)\left(2 - M\widehat{Z}_2\right)}{\left(N + M\right)^2}. \tag{12.4}$$

For the Bayes estimator of the Rényi entropy it is more difficult to calculate the entropy bias. Given equidistributed states, we find that $\mathrm{E}\log(\cdot) \approx \log\mathrm{E}(\cdot)$ holds, and thus we can get an approximation to the bias of $\widetilde{K_2}$, which reads as:

$$\Delta\widetilde{K_2} = -\log_2\left[\frac{N^2 + 2\left(2N + M\right)/Z_2}{\left(N + M\right)^2}\right] \; + \; \mathcal{O}\left(\frac{1}{N^2}\right). \tag{12.5}$$

And hence, in analogy to equation (12.4), we obtain Rényi entropy estimates $\widehat{K_2}^{(1)}$ that are unbiased in $\mathcal{O}(1/N)$. For non-Bernoullian distributions fluctuations increase, which render the above approximation to be, in general, no longer reliable. In this case, unbiased estimates of K_2 (in the order of $\mathcal{O}(1/N)$) may be obtained by a transformation of the unbiased Tsallis entropy $\widehat{H}_2^{(1)}$.

According to the correction terms, c.f. expressions (12.3) and (12.5), the systematic error depends on the individual probability components p_i as well as the cardinality of the alphabet M. Since the simulation performed in this investigation are not aimed at a detailed analysis of finite-size effects but rather a study of the variances of the Bayes

entropy estimator versus the frequency-count estimator, we insert the theoretical values of p_i in the above correction terms, i.e., we set $\hat{p}_i = p_i$. In any attempt to estimate these quantities from a sample of data points, it is crucial to the entropy bias by which method we estimate the unknown variables p_i (see, e.g., [Schmitt et al. 1993]). A study of the quantification of the order-q entropy bias, using asymptotic length corrections, deserves further investigations and will be undertaken in forthcoming work.

Part D

Chapter 13

Long-Range Correlations in DNA Sequences and the Pseudo-Exon Model

In this chapter we study the mutual information function introduced in chapter 3 as well as correlation functions to DNA sequences. We find long-range period-3 oscillations of both the mutual information function and several autocorrelation functions in genomic DNA of yeast. We hypothesize that these triplet periodicities are caused by the nonuniformity of the codon frequency distribution in coding DNA, and we test this hypothesis by a simple stochastic model, which we term *pseudo-exon model*. The pseudo-exon model concatenates codons that are drawn statistically independently from a given (nonuniform) probability distribution. We show that this simple model is sufficient to reproduce (even quantitatively) the period-3 oscillations observed in the mutual information function and autocorrelation functions of genomic DNA from yeast.

13.1 Introduction

As pointed out in chapter 4, six autocorrelation functions and three crosscorrelation functions are required to detect all statistical dependences in quaternary sequences. We emphasize that the mutual information function may serve as a natural and convenient alternative or supplement to correlation functions [Ebeling et al. 1987, Herzel 1988, Herzel et al. 1994a, Herzel et al. 1995]. With the aid of mutual information, a pronounced

period three has been found at distances of more than 1000 base pairs in [Herzel et al. 1995].

Traditionally, long-range correlations are quantified by a power-law decay of the autocorrelation functions

$$C(k) \propto k^{-\gamma}, \tag{13.1}$$

the intimately related $1/f$-spectra, or by anomalous diffusion. (These tools are reviewed in [Stanley et al. 1994].) For DNA sequences, long-range correlations are relatively weak [Herzel et al. 1994a], and therefore, we expand the mutual information in terms of our $D_{ij}(k)$ that measure the deviations from statistical independence:

$$p_{ij}(k) = p_i \cdot p_j + D_{ij}(k). \tag{13.2}$$

With the normalization (7), we obtain from a Taylor expansion

$$I(k) = \frac{1}{2 \cdot \ln 2} \sum_{i,j=1}^{\lambda} \frac{D_{ij}(k)^2}{p_i \cdot p_j} + \mathcal{O}(D_{ij}^3). \tag{13.3}$$

The linear terms vanish, since $I(k)$ exhibits a minimum at $D_{ij}(k) = 0$. Therefore, absolute values of the mutual information are often very small compared to autocorrelation functions [Herzel et al. 1994a]. In order to interpret these small values, a careful analysis of statistical and systematic errors is required [Herzel et al. 1994a].

Autocorrelation functions are just bilinear forms of the matrix $\hat{D}(k)$ according to eq. (4), and consequently, a power-law (60) together with eq. (62) implies

$$I(k) \propto k^{-2\gamma}. \tag{13.4}$$

Hence, long-range correlations can easily be quantified by the mutual information as well. Moreover, mutual information functions have been used to calculate *symbolic spectra*, and an integrated version of $I(k)$ has been studied [Herzel et al. 1995] in analogy to random walk studies, which are directly related to integrated autocorrelation functions [Peng et al. 1992].

Long-range power-law correlations in DNA sequences have been found in *introns* and *intergenic regions*, which do not code for proteins [Peng et al. 1992, Stanley et al. 1994]. Their interpretation in terms of biological structure and function is, however, still debated [Herzel et al. 1995, Stanley et al. 1994, Li et al. 1994, Grosberg 1993].

13.2 Nonuniform Codon Usage

In the remainder of this chapter, we focus on the role of a nonuniform codon usage in exons and demonstrate that, in sequences with long protein-coding segments (e.g., in yeast DNA), the resulting periodicity plays a significant role.

Let us recall some well-known facts about the genetic code [Lewin 1997, Watson et al. 1992, Berg & Singer 1992, Kolchanov & Lim 1994]: 61 codons (3-symbol-words) of the possible 64 encode 20 different amino acids whereas the remaining 3 are used as stop codons. For several reasons, the codon distribution is very nonuniform in exons (e.g. $p(\mathrm{CGA})$=0.4% and $p(\mathrm{CTG})$=3.3% according to [Staden 1984]):

- The number of triplets coding for an amino acid is different. (For instance, Tryptophan is coded only by TGG whereas Leucine, Serine, and Arginine are coded by even six codons.)
- There are specific amino acid compositions for proteins.
- For any amino acid, a preference for certain codons with respect to others exists. (These preferences are assumed to be related to the availability of t-RNAs and correlate with the expression rate of genes.)

The different codon usage in exons and introns is widely exploited to detect protein-coding segments in unknown DNA [Fickett 1982, Staden 1984, Lapedes et al. 1990, Uberbacher & Mural 1991, Fickett & Tung 1992]. In the following we discuss the implications of a specific codon usage on correlation measures.

13.3 Positional Nucleotide Frequencies

A nonuniform codon usage introduces in general peculiarities of the base composition at different positions in the reading frame. For example, the nucleotide G is more frequent at position 1 (referring to the first symbol of codons) than at position 2 according to the tables in [Staden 1984]. In order to demonstrate the effect of the reading frame on autocorrelations and mutual information, we start with a representative table of relative frequencies of A, C, G, and T in the three positions of the frame.

	position 1	position 2	position 3
A	0.326	0.337	0.335
C	0.179	0.217	0.164
G	0.262	0.100	0.171
T	0.233	0.346	0.330

These relative frequencies are obtained from a 6324 base pair long exon of the yeast chromosome III [Oliver et al. 1992]. Obviously, there is only a week dependence of A on the position, but a significant one of G and T. In the following the frequency of the i-th nucleotide at the l-th position is denoted by $p_i^{(l)}$. The overall probability of symbol i follows directly by averaging over the three positions

$$p_i = \frac{p_i^{(1)} + p_i^{(2)} + p_i^{(3)}}{3} \qquad (i = 1...4). \tag{13.5}$$

13.4 Pseudo-exon model

In order to estimate the effect of a nonuniform codon usage on correlation measures, we make the simplifying assumption that subsequent codons are statistically independent. This allows the direct calculation of the joint probabilities $p_{ij}(k)$ from tables as shown above. For $k \geq 3$, the corresponding probabilities factorize due to our assumption of independence. First we consider $k = 3, 6, 9, ...$, i.e., the two symbols are in the same position within the frame:

$$p_{ij}(k) = \frac{p_i^{(1)} p_j^{(1)} + p_i^{(2)} p_j^{(2)} + p_i^{(3)} p_j^{(3)}}{3}. \tag{13.6}$$

For $k = 4, 7, 10, ...$ we obtain

$$p_{ij}(k) = \frac{p_i^{(1)} p_j^{(2)} + p_i^{(2)} p_j^{(3)} + p_i^{(3)} p_j^{(1)}}{3}, \tag{13.7}$$

and distances $k = 5, 8, 11, ...$ lead to

$$p_{ij}(k) = \frac{p_i^{(1)} p_j^{(3)} + p_i^{(2)} p_j^{(1)} + p_i^{(3)} p_j^{(2)}}{3}. \tag{13.8}$$

Inspection of the last two expressions reveals that

$$p_{ij}(k = 4, 7, ...) = p_{ji}(k = 5, 8, ...). \tag{13.9}$$

Consequently, the values of the mutual information at these positions are identical. Hence, the mutual information function exhibits a rather specific feature: It oscillates between the two values $I(3) = I(6) = I(9) = ...$ and $I(4) = I(5) = I(7) = I(8) = I(10) = ...$. Such a period three of probabilities $p_{ij}(k)$ [Fickett 1982] and of the mutual information has indeed been observed in DNA sequences [Ebeling et al. 1987, Herzel et al. 1995]. It follows from eqs. (64)-(68) that

$$D_{ij}(k) + D_{ij}(k + 1) + D_{ij}(k + 2) = 0 \tag{13.10}$$

for all i and j and for all $k \geq 3$. Consequently, the sum over three consecutive values of any correlation function vanishes, if the distance k is longer than or equal to three, i.e., correlation and anticorrelation counterbalance each other within one period of three nucleotides. This is relevant for random walk studies [Peng et al. 1992, Stanley et al. 1994], which are related to integrated correlation functions.

13.5 Analysis of Yeast DNA

From the table given above, we obtain

$$I(k = 3, 6, ...) = 7.9 \cdot 10^{-4} \tag{13.11}$$

and

$$I(k = 4, 5, 7, 8, ...) = 2.9 \cdot 10^{-4}. \tag{13.12}$$

Figure 13.5 shows results of a Monte-Carlo simulation of a *pseudo-exon* generated by concatenating independent codons. This is a Bernoulli-like process with 61 non-vanishing probabilities corresponding to the relative frequencies of codons in the exon chosen to generate the above table.

The graphs clearly demonstrate the expected periodicity of the GG-correlation function, the small amplitude of the "non-biological" AC-GT-function, and the predicted feature of the mutual information (high-low-low). The GG-autocorrelation function demonstrates eq. (69): the sum over three consecutive k vanishes for $k \geq 3$. Due to finite sample effects, there are random fluctuations around the analytical values given in eqs. (70) and (71) (cf. Appendix II).

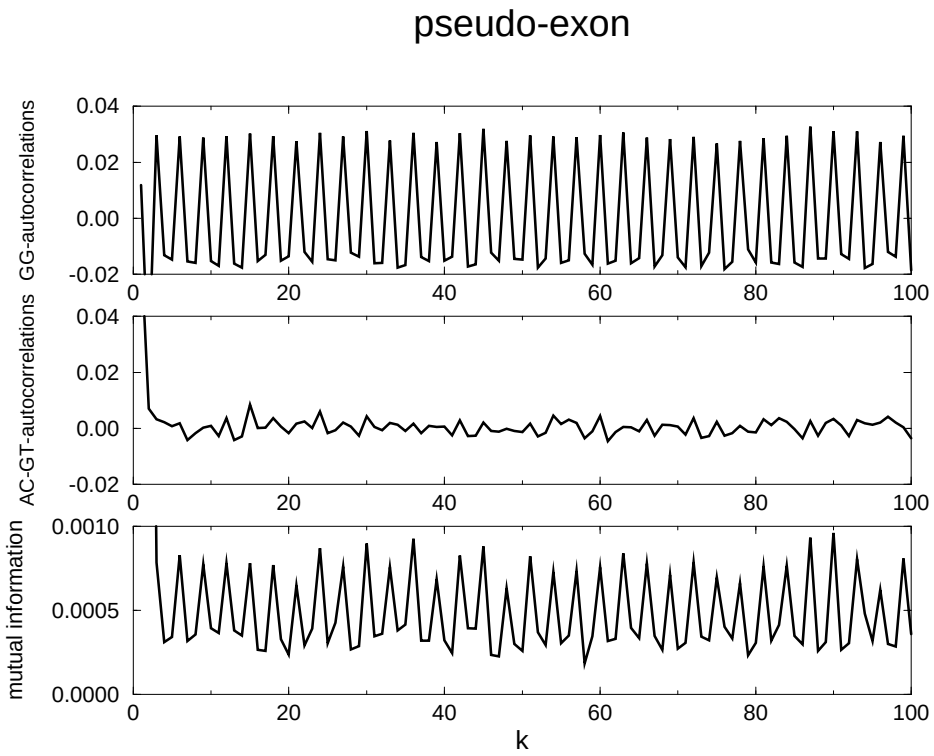


Figure 13.1: Two autocorrelation functions and the mutual information for a pseudo-exon without codon-codon-interactions. We chose the 6324 base pair long exon starting at position 278851 on chromosome III of the yeast *Saccharomyces cerevisiae* to derive a representative codon usage table for yeast DNA. Then, we generated a 300000 base pair long pseudo-exon by a random concatenation of 100000 codons according to this table. The GG-autocorrelation function (upper graph) exposes a very strong periodicity, since the probability to find the nucleotide G varies tremendously with its position in the reading frame. The "non-biological" AC-GT-autocorrelation (middle graph) reveals only a faint periodicity due to the fact that the corresponding probabilities are almost uniformly distributed over the three possible positions in the frame. The pronounced periodicity exhibited by the mutual information corresponds exactly to the predicted behavior. Note that, despite the absolute values of the mutual information are really tiny, the differences between the maxima at $k = 3, 6, ...$ and the minima at $k = 4, 5, 7, 8, ...$ are by far higher than random fluctuations.

13.6 Discussion

A nonuniform codon usage implies (even for independent codons) a persistent oscillation of the joint probabilities $p_{ij}(k)$, and hence, of the correlation functions and the mutual information. These oscillations can be used to design algorithms that discriminate coding from noncoding sequences. A widely used technique considers the maximal and minimal frequencies of A, C, G, and T in all three frames [Fickett 1982]. We emphasize that the mutual information has some striking advantages compared to traditional methods:

- It detects any deviation from statistical independence.
- It takes into account all 16 joint probabilities.
- Due to the above properties of $I(k)$, a single number can be chosen for classification: the difference between $I(k)$ for $k = 3, 6, 9, ...$ and the remaining values.
- The statistical properties (bias, variance) of entropy-estimators have been extensively studied (see Appendix II).

However, as shown below, correlation functions usually oscillate with larger amplitudes than the mutual information function.

The period-three oscillations induced by the reading frame have consequences for the interpretation of long-range correlations. Strictly speaking, DNA sequences are not stationary within exons since the probabilities p_i depend on the position in the reading frame (comparable to seasonal periodicities in climatic data). For distances much longer than exons they should vanish, but protein-coding segments may extend over thousands of base pairs, a distance which is typically analyzed in the context of long-range correlations [Stanley et al. 1994, Li et al. 1994]. However, there are also other sources of long-range correlations since they have been found in introns as well [Peng et al. 1992].

Finally, we exemplify that the theoretical results mentioned above are indeed relevant for DNA sequences. For this purpose we present correlation functions and mutual information derived from the complete DNA sequence of the yeast chromosome III [Oliver et al. 1992].

Figure 13.6 demonstrates that the shape of the mutual information and autocorrelation functions of real DNA sequences is dominated by period-3-oscillations. As pointed out in this section and illustrated in Figure 13.5, these oscillations are due to a nonuniform codon

usage. We realize that the high specificity of autocorrelation functions and their dependence on the chosen projection is not only of theoretical interest but also of practical importance for analyzing biological sequences.

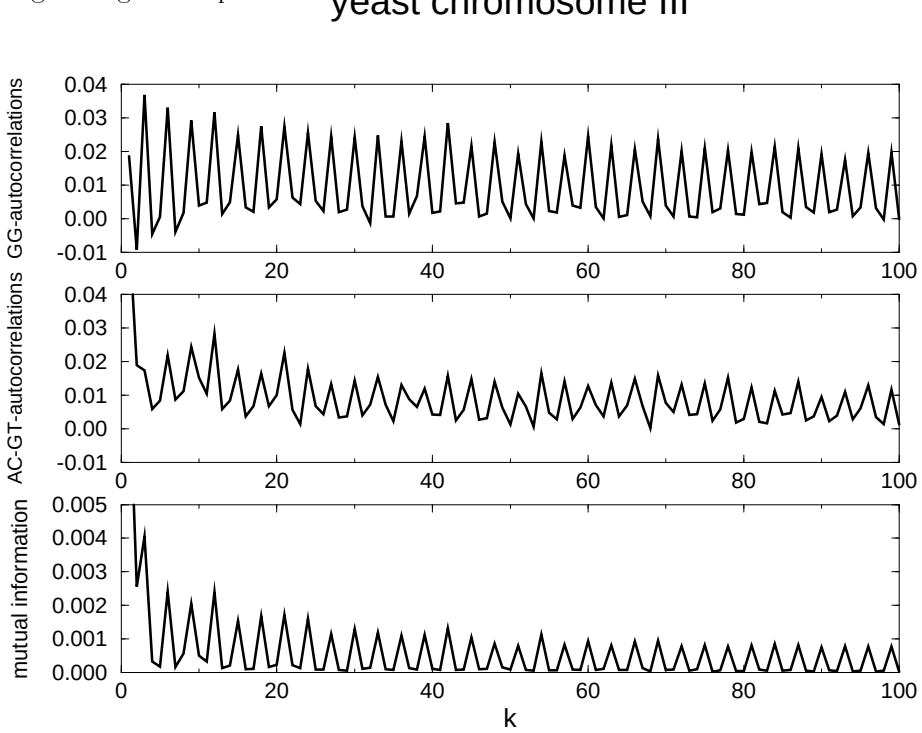


Figure 13.2: Correlations of the 315338 base pair long DNA sequence of the yeast *Saccharomyces cerevisiae* chromosome III for distances k between 1 and 100 base pairs. The dominating triplet-periodicity that is induced by the nonuniform codon usage in yeast can easily be observed. The comparison of the upper two graphs reveals that the A+C content is indeed a poor indicator for a nonuniform codon usage. The mutual information function displays its typical high-low-low-pattern, which can be exploited to discriminate exons from introns. Note that the peak at $k = 3$ reveals correlations between neighboring codons in yeast DNA.

Figure 13.6 also reveals the interesting finding that the AC-GT-autocorrelation function (middle graph) shows a much weaker performance than, for instance, the GG-autocorrelation function (upper graph), which is biologically interpretable. The predicted periodicity of the mutual information function (high-low-low) is strikingly confirmed by the lower graph (cf. Figure 1). In addition to those features of autocorrelations and the mutual information function that are caused by a nonuniform codon usage, we can easily observe further biological correlations. The high $I(3)$-value indicates, for example, existing

correlations between adjacent codons in yeast DNA.

The decay of the envelopes of the mutual information and the autocorrelation functions indicate biologically relevant correlations that, however, diminish with an increasing distance k.

By increasing the correlation length k, we realize that all of the features discussed above stay persistent up to distances of about 1000 base pairs. Figure 3 illustrates that the oscillation behavior typical for exonic sequences is still observable at these length scales.

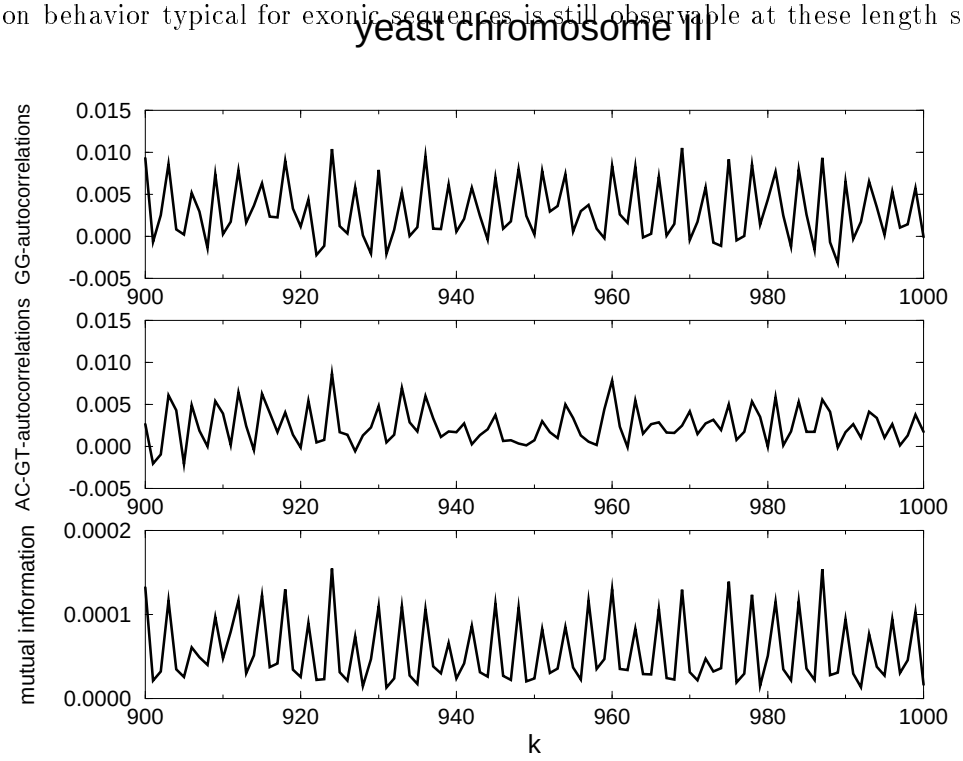


Figure 13.3: Correlations of the complete DNA sequence of yeast chromosome III for distances k between 900 and 1000 base pairs. The GG-autocorrelation function as well as the mutual information maintain their dominating period-3-oscillations up to 1000 base pairs. Remember that the reduced amplitudes are due to the small number of exons longer than 1000 base pairs.

However, the absolute values of the mutual information derived from the DNA of yeast chromosome III are by far lower than the theoretical values calculated in eqs. (70) and (71). This can easily be understood if one takes into account that there are only a few exons on the considered chromosome which are longer than 1000 base pairs. Hence, only a small percentage of all coding regions contribute to the explained periodicity of the mutual information function for these distances.

Both, Figure 2 and Figure 3 clearly demonstrate that the nonuniform codon usage in coding regions yields the main contribution to correlations in yeast DNA, since its induced period-3-oscillations are the dominating pattern on almost all length scales.

13.7 Summary

We derived that a power-law decay of autocorrelation functions corresponds to a power-law decay of the mutual information, where the exponent δ in the mutual information function is twice the exponent γ of the autocorrelation functions.

We studied the mutual information function of protein-coding sequences, which are characterized by specific codon usage tables. We showed that these tables allow the analytical calculation of correlation measures under the assumption of independent codons. We confirmed the predicted triplet-peak structure by direct analyses of yeast DNA. We found that the nonuniform codon usage induces persistent oscillations up to scales of the maximum exon length, which extends over thousands of base pairs.

Chapter 14

Correlations in DNA Sequences – the Role of Protein Coding Segments

In this chapter we propose one possible explanation for the recent observation of long-range correlations in genomic DNA sequences. Protein coding segments (exons) exhibit persistent correlations between their nucleotides with a pronounced period three. We show that this periodicity, which is induced by the nonuniform codon usage, implies long-range correlation over hundreds of base pairs if the length distribution of exons is taken into account. We derive analytic expressions which relate the length distribution of exons to the correlation decay, and we find agreement with numerical simulations. Finally, we analyze the decay of the mutual information function in yeast chromosomes, in an E. coli chromosome region, and in myosin heavy chain genes as representative examples. It turns out that in these cases we can explain most of the long-range statistical dependences even quantitatively.

14.1 Introduction

The statistical analysis of DNA sequences is of importance for understanding the structure and function of genomes [Gatlin 1972, Trifonov & Brendel 1986, von Heijne 1987, Bell & Marr 1990, Watson et al. 1992, Yockey 1992, Kolchanov & Lim 1994, Lewin 1997]. Statistical dependences between nucleotides have been analyzed for decades in various contexts [Shepherd 1981, Fickett 1982, Ebeling et al. 1987, Herzel 1988, Trifonov 1989,

Fickett et al. 1992, Grosse 1999]. Among physicists the detection of long-range correlations has attracted much attention during the past years [Li & Kaneko 1992, Peng et al. 1992, Voss 1992, Borštnik et al. 1993, Herzel et al. 1994a, Arneodo 1995, Herzel & Grosse 1995, Herzel et al. 1995]. Using mutual information functions [Li & Kaneko 1992, Herzel et al. 1994a, Herzel et al. 1995], autocorrelation functions [Borštnik et al. 1993, Herzel & Grosse 1995], spectra [Voss 1992, Chechetkin & Turygin 1994], and random walk analyses [Peng et al. 1992, Dreismann & Larhammer 1993, Stanley et al. 1994], correlations ranging from a few base pairs (bp) up to 10^4 bp have been analyzed. However, the biological interpretation of most of these findings remains still speculative.

From a molecular biological point of view, long-range correlations are not surprising since the complex organization of genomes involves many different scales. In fact, large variations in base composition on scales of thousands of base pairs have been discussed extensively in the literature (see, e.g., [Elton 1974, Bernardi et al. 1985, Korenberg & Rykowski 1988, Bernardi et al. 1985, Ikemura et al. 1990, Fickett et al. 1992, Karlin & Brendel 1993]). For example, Elton [Elton 1974] reviews experimental data showing that DNA fragments up to 10^4 bp have rather large variances of the G+C content. He points out that these variations cannot be explained by short-correlated fluctuations. In this way, long-range correlations have been indicated already decades ago. Explicit examples of pronounced fluctuations of the G+C content together with the gene distribution with an approximate period of 10^5 bp were provided by the recent sequencing of yeast chromosomes [Feldmann et al. 1994, Dujon et al. 1994].

It has been pointed out by several authors that the mosaic structure of genomes is presumably responsible for long-range correlations [Herzel et al. 1994a, Bernardi et al. 1985, Karlin & Brendel 1993]. Indeed, the organization of the genome is very complex: eukaryotic genes usually consist of several protein coding segments (*exons*) interrupted by intervening sequences (*introns*). Moreover, there are regulatory elements such as promoters, splice sites, enhancers, and silencers, which are sometimes up to thousands of base pairs away from exons. Genomes of higher eukaryotes also comprise long stretches of DNA without any obvious biological function containing, e.g., *pseudo genes* and various types of *repeats* [Lewin 1997, Herzel et al. 1995, Herzel et al. 1994b].

There are several models of DNA where a segmentational structure is postulated [Elton 1974, Fickett et al. 1992, Churchill 1989, Buldyrev et al. 1993, Li et al. 1994]. Elton discusses, for example, the variance of the G+C content for a model with constant and

exponentially distributed fragments [Elton 1974], and Buldyrev et al. study a Lévy-walk model [Buldyrev et al. 1993]. However, hypothetical length distributions of fragments have to be postulated in these papers.

Contrarily, we will show in this chapter that already the well-known length distribution of exons generates long-ranging correlations. As a first step we demonstrate in section III that a nonuniform *codon usage* in protein coding segments induces persistent period-three oscillations. In that section we introduce a model by which we generate artificial DNA sequences called *pseudo-exons*—a concatenation of statistically independent codons chosen randomly from a given codon usage probability table. In sections IV and V, we emphasize the central role of the length distribution of exons. We derive analytic expressions which relate the exon length distribution to the correlation decay and show that these analytic results are in perfect agreement with numerical simulations. In section VI, we apply these theoretical considerations to several DNA sequences (yeast chromosomes, E. coli DNA, and a myosin heavy chain gene).

We show that correlations on scales of hundreds of base pairs can be simulated even quantitatively by taking into account solely the non-uniformity of the codon usage and the length distribution of exons. In this way we relate well known biological facts to observed long-range correlations between nucleotides.

14.2 Correlation measures

DNA sequences can be viewed as symbolic strings composed of the four "letters" $(A_1, A_2, A_3, A_4) \equiv (A, C, G, T)$ referring to the nucleotides adenine, cytosine, guanine, and thymine. The probability to find the nucleotide A_i is denoted by p_i $(i = 1, 2, 3, 4)$. Pair correlations within sequences can be measured by the joint probabilities $p_{ij}(k)$ to find the symbol A_i and k letters downstream the symbol A_j. Then, statistical independence of symbols in a distance k is defined by $p_{ij}(k) = p_i \cdot p_j$, which leads to the mutual information function $I(k)$ [Yockey 1992, Herzel 1988, Herzel & Grosse 1995, Kullback 1959, Herzel & Ebeling 1985, Li 1990] as a measure of statistical dependence:

$$I(k) = \sum_{i,j=1}^{4} p_{ij}(k) \log_2 \frac{p_{ij}(k)}{p_i \cdot p_j}. \tag{14.1}$$

By choosing the logarithm to base 2, $I(k)$ is measured in bit and gives the information on the letter A_j knowing the letter A_i. The mutual information $I(k)$ vanishes if, and

only if, statistical independence holds, i.e., if all 16 joint probabilities $p_{ij}(k)$ factorize. Consequently, the mutual information allows to detect any pair correlation.

More specific indicators of dependences are correlation functions. Their definition requires an assignment of numbers a_i to the corresponding symbols A_i. Assuming ergodicity and stationarity, the usual estimation of autocorrelation functions via averages over the sequence

$$C(k) = \langle a(n)\, a(n+k) \rangle - \langle a(n) \rangle \, \langle a(n+k) \rangle \tag{14.2}$$

can be written in terms of the probabilities defined above:

$$\begin{aligned} C(k) &= (\sum_{i,j=1}^{4} p_{ij}(k) \cdot a_i \cdot a_j) - (\sum_{i=1}^{4} p_i \, a_i)(\sum_{j=1}^{4} p_j \, a_j) \\ &= \sum_{i,j=1}^{4} (p_{ij}(k) - p_i \cdot p_j) \cdot a_i \cdot a_j . \end{aligned} \tag{14.3}$$

By definition, correlation functions measure only linear dependences. However, for quaternary sequences such as DNA, six properly chosen autocorrelation functions and three cross-correlation functions can guarantee the statistical independence between all nucleotide-pairs [Herzel & Grosse 1995].

Long-range correlations are often characterized by power-laws

$$C(k) \propto k^{-\gamma}. \tag{14.4}$$

Such a scaling behavior can also be analyzed by using power-spectra or the random walk approach with related scaling exponents [Stanley et al. 1994]. A power-law (4) implies also a power-law decay of the mutual information function:

$$I(k) \propto k^{-2\gamma}. \tag{14.5}$$

This can be easily derived using a Taylor expansion in terms of

$$D_{ij}(k) = p_{ij}(k) - p_i \cdot p_j, \tag{14.6}$$

which measure deviations from statistical independence. Since $I(k)$ has a minimum at $D_{ij} \equiv 0$, the sum over all linear terms vanishes, and we obtain

$$I(k) = \frac{1}{2 \cdot \ln 2} \sum_{i,j=1}^{4} \frac{D_{ij}^2(k)}{p_i \cdot p_j} + \mathcal{O}(D_{ij}^3). \tag{14.7}$$

In the Appendix we use this relation to discuss finite sample effects. Equation 7 illustrates that the mutual information $I(k)$ accumulates all pair correlations in a distance k. For DNA sequences, the above 2nd order approximation is extremely close to the actual mutual information because of the weakness of correlations (see, e.g., Figure 1). Since correlation functions can be written as quadratic forms of the *dependence matrix* D_{ij} (cf. Eq. (3) and (6)), a scaling exponent γ of correlation functions leads to an exponent 2γ for the mutual information.

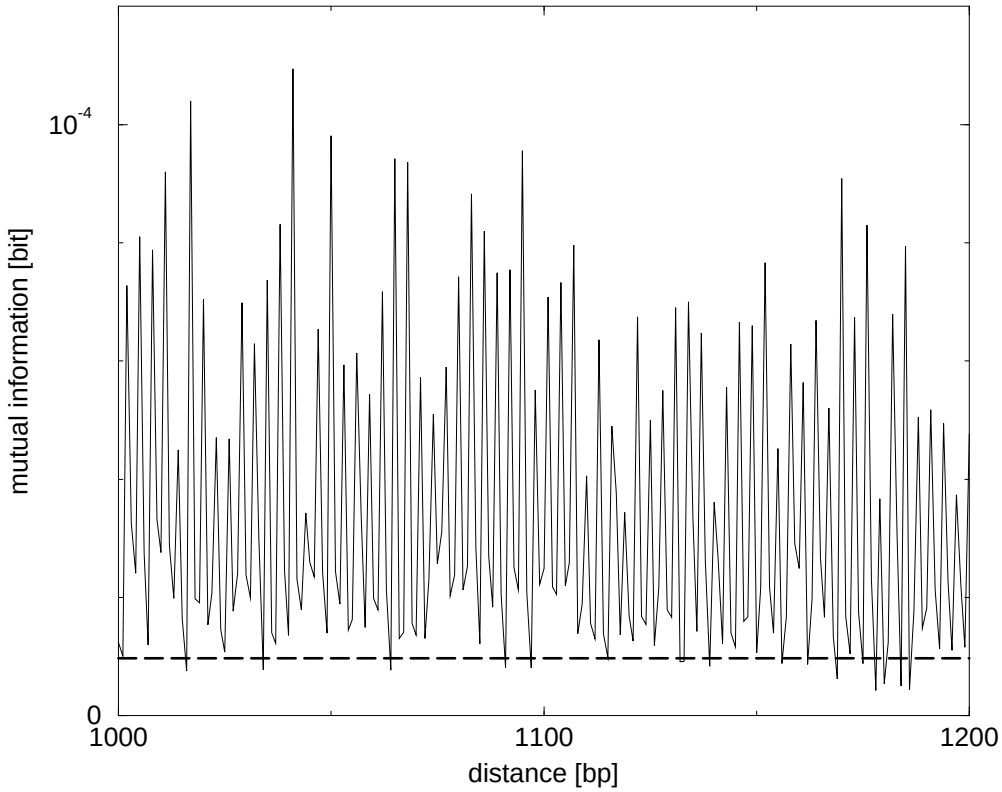


Figure 14.1: Mutual information function of the yeast chromosome XI (666,448 bp). The periodicity due to the triplet code is visible even for distances above 1000 bp. The dashed line marks the bias according to Eq.(8).

In this chapter we study mainly the decay of the mutual information function as an overall measure of statistical dependences. In contrast to entropies of long "words"

[Herzel et al. 1994a, Herzel et al. 1994b] the statistical and systematic errors of the mutual information are relatively small since only 16 probabilities have to be estimated from samples of thousands of nucleotides. For example, the bias of the mutual information for a sample of size N has been calculated [Herzel 1988, Herzel & Grosse 1995] to be

$$\Delta I = \frac{9}{2 \ \ln 2 \ N}, \tag{14.8}$$

which is marked in some figures by a dashed line. Though this bias is small, it becomes relevant for very weak correlations. Therefore, we discuss finite sample effects in some detail in an Appendix.

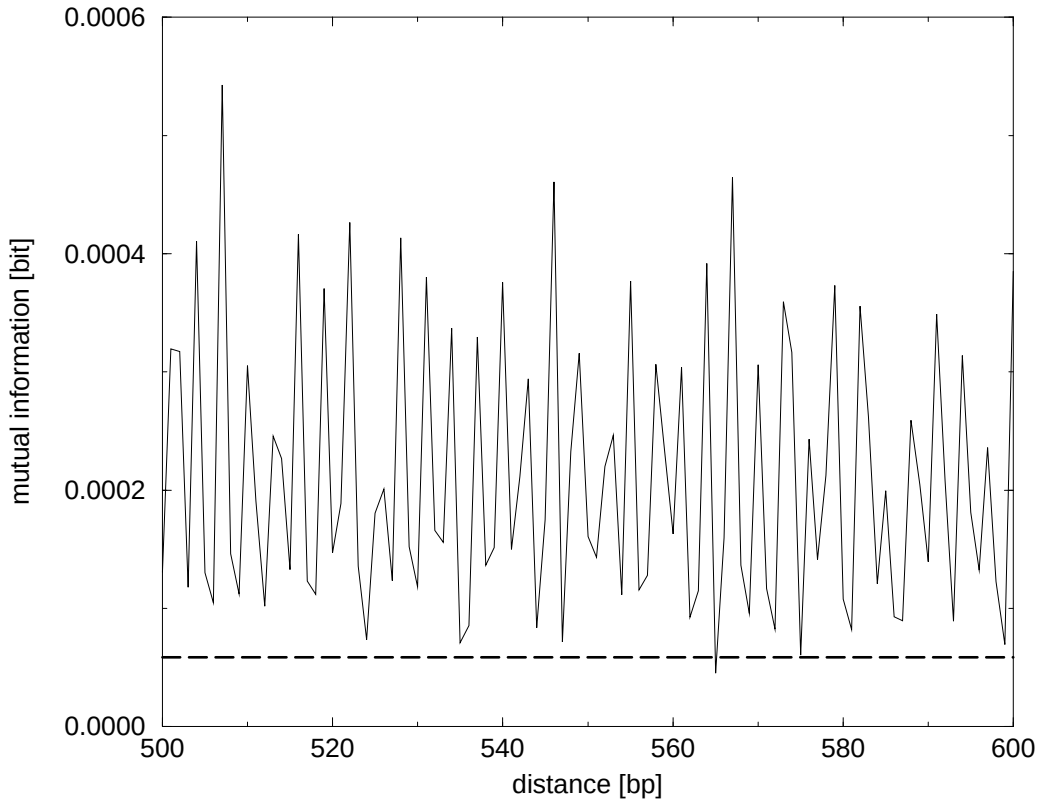


Figure 14.2: Period-three oscillations of the mutual information for a chromosome region of Escherichia coli (strain K-12, 111,401 bp).

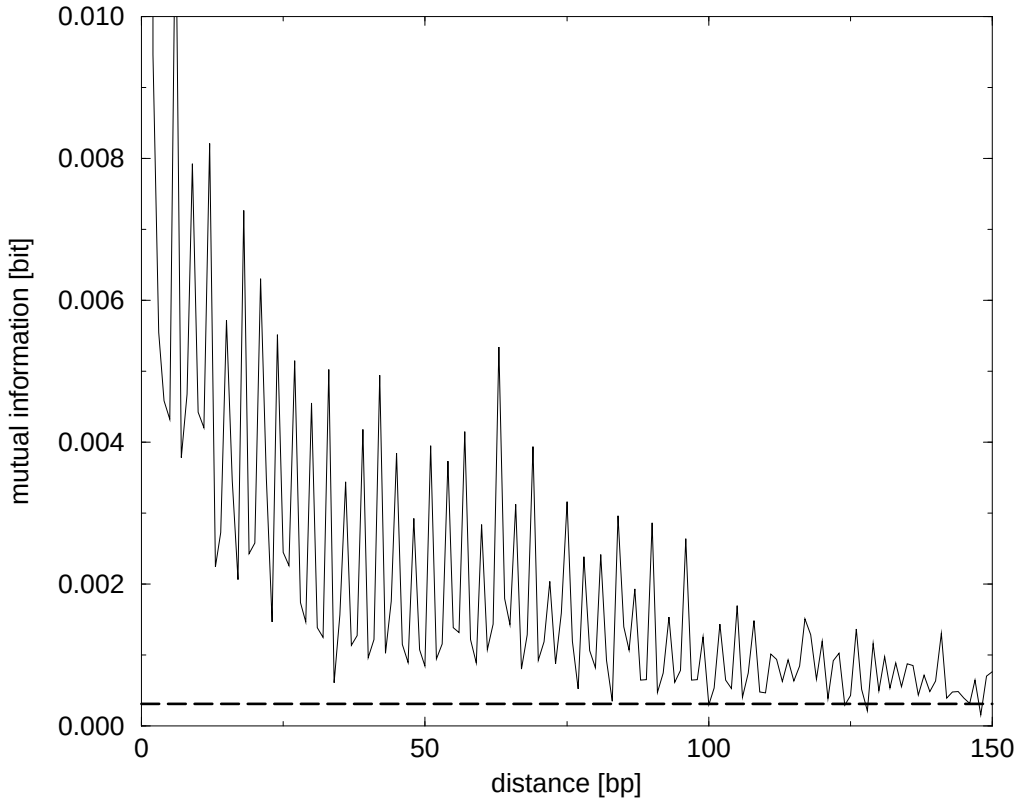


Figure 14.3: Mutual information function of the HUMBMYH7 gene (20,855 bp from the first to the last exon). The mean exon length is about 150 bp which is the characteristic length of the decay of the pronounced period-three oscillations.

14.3 Effects of a Nonuniform Codon Usage

Analyses of DNA sequences revealed that their correlation functions often exhibit strong period-three components, which are induced by the genetic code [Shepherd 1981, Fickett 1982, Ebeling et al. 1987, Chechetkin & Turygin 1994, Luo & Li 1991]. Figures 1-3 exemplify these periodicities for yeast chromosome XI, an E. coli chromosome region, and a myosin heavy chain gene.

In protein coding segments, 61 of the possible 64 codons (3-symbol-words) encode 20 different amino acids whereas the remaining 3 are used as stop codons. It has been discussed [Bernardi et al. 1985, Ikemura 1981, Staden 1984, Sharp & Li 1987] that these codons are used with quite different frequencies for several reasons:

- There are specific amino acid compositions for proteins.
- The number of triplets encoding an amino acid is different.
- For any amino acid, a preference of certain codons over others exists.
- The G+C content of the third codon position is correlated to the G+C content of the surrounding DNA region [Dujon et al. 1994].

In general, a nonuniform codon usage causes the concentration of each nucleotide to be different in all three positions of the reading frame. As we will show in the following, it is exactly this *position asymmetry* of all 4 nucleotides that introduces the pronounced period-three pattern of correlation functions as well as the mutual information function.

In order to quantify the effect of a nonuniform codon usage on correlation measures, we introduce a stochastic model that randomly concatenates subsequent codons. In the following, we term the model sequences of independent codons chosen from a given codon usage table *pseudo-exon*. As we will see, these pseudo-exons, which consist of statistically independent codons, display periodic long-range correlations between their nucleotides. Our next task is to analytically calculate the strength of these correlations, which was shown to be a prominent long-range correlation pattern of real DNA (cf. Figures 1-3).

Let us start with the calculation of the mutual information of an infinitely long pseudo-exon generated by such a Bernoulli-like process on the level of codons.

We denote the frequency of the i-th nucleotide at the m-th position by $p_i^{(m)} (m = 1, 2, 3)$.

The overall probability of symbol i follows directly by averaging over the three positions

$$p_i = \frac{p_i^{(1)} + p_i^{(2)} + p_i^{(3)}}{3} \qquad (i = 1...4). \tag{14.9}$$

The following table displays the 12 frequencies $p_i^{(m)}$, which are obtained from the 5805 bp of the protein coding segments from the intensively studied [Grosse 1999, Buldyrev et al. 1993] human beta-myosin heavy chain (HUMBMYH7) gene.

	position 1	position 2	position 3
A	0.296	0.437	0.079
C	0.248	0.184	0.343
G	0.351	0.123	0.471
T	0.105	0.256	0.107

The joint probabilities $p_{ij}(k)$ can be obtained directly from tables as shown above. For $k \geq 3$, the corresponding probabilities factorize due to our assumption of independence. First we consider $k = 3, 6, 9, ...$, i.e., the two symbols of the pair are in the same position within the frame:

$$p_{ij}(k) = \frac{p_i^{(1)} p_j^{(1)} + p_i^{(2)} p_j^{(2)} + p_i^{(3)} p_j^{(3)}}{3}. \tag{14.10}$$

For $k = 4, 7, 10, ...$ we obtain

$$p_{ij}(k) = \frac{p_i^{(1)} p_j^{(2)} + p_i^{(2)} p_j^{(3)} + p_i^{(3)} p_j^{(1)}}{3}, \tag{14.11}$$

and distances $k = 5, 8, 11, ...$ lead to

$$p_{ij}(k) = \frac{p_i^{(1)} p_j^{(3)} + p_i^{(2)} p_j^{(1)} + p_i^{(3)} p_j^{(2)}}{3}. \tag{14.12}$$

Inspection of the last two expressions reveals that $p_{ij}(k = 4, 7, ...) = p_{ji}(k = 5, 8, ...)$. Consequently, the values of the mutual information at these positions are identical. The above expressions allow us to calculate the *in-frame mutual information*

$$I_{in} \equiv I(k = 3, 6, 9, ...) \tag{14.13}$$

and the *out-of-frame mutual information*

$$I_{out} \equiv I(k = 4, 5, 7, 8, 10, 11, ...). \tag{14.14}$$

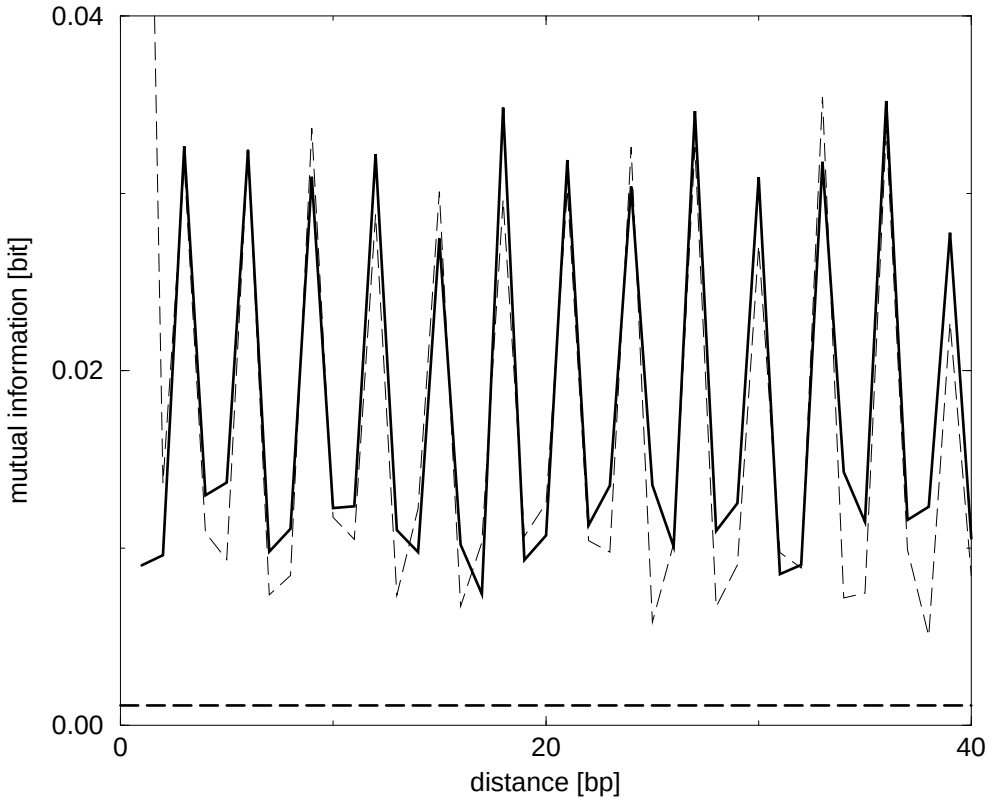


Figure 14.4: Dashed line: Mutual information of a concatenation of all 40 exons (5,805 bp) of the HUMBMYH7 gene (compare Figure 3). Full line: Corresponding pseudo-exon (5,805 bp) generated from the codon usage table of the HUMBMYH7 gene.

For example, the table given above yields

$$\begin{aligned} I_{in} &= 0.0247, \\ I_{out} &= 0.0083. \end{aligned}$$

The corresponding high-low-low pattern is indeed obvious in the examples graphed in Figures 1-4. Figure 4 displays the period-three oscillations of a pseudo-exon that is indeed quite similar to the mutual information of the corresponding exons.

In summary, for a single protein coding segment, a given codon usage table allows us to analytically calculate the resulting period-three oscillations. However, genomes contain many exons, introns, and intergenic sequences. Moreover, protein coding segments are found in all three reading frames and on both DNA strands. Therefore, we are not surprised by the fact that the mutual information function plotted in Figures 1-3 are decaying and thus deviate from a purely repeated high-low-low pattern.

The next section is devoted to the role of the length distribution of exons, which indeed strongly affects the decay properties of correlation measures. Taking into account these length distributions, we can generalize the pseudo-exon model to stochastic models of genes and even of whole chromosomes termed *pseudo-chromosomes*.

14.4 Length Distribution of Exons

We have discussed in the preceding section that the joint probabilities $p_{ij}^{(k)}$ calculated within an exon reflect the nonuniform codon usage. Since long stretches of DNA include many different exons, only a fraction of pairs A_i and A_j are located on the same exon. More precisely, an exon of length l contains $l-k$ pairs contributing to the codon usage induced periodicity. Consequently, the length distribution $\rho(l)$ of exons in a given DNA will be considered in this section.

We define $\rho(l)$ as the probability distribution that an exon has a length l. In the next section we discuss, for instance, a fixed length $l = L$, exponential, and power-law distributions $\rho(l)$. Figure 5 shows a histogram of the lengths of exons for yeast chromosomes. It can be seen that there are rather long protein coding segments. Regression reveals that the empirical distribution can be approximated by an exponential decay (full line) and by a power-law (dashed line) as well. Hence we discuss both cases in some detail.

For the sake of simplicity, we assume below that all exons are characterized by a single

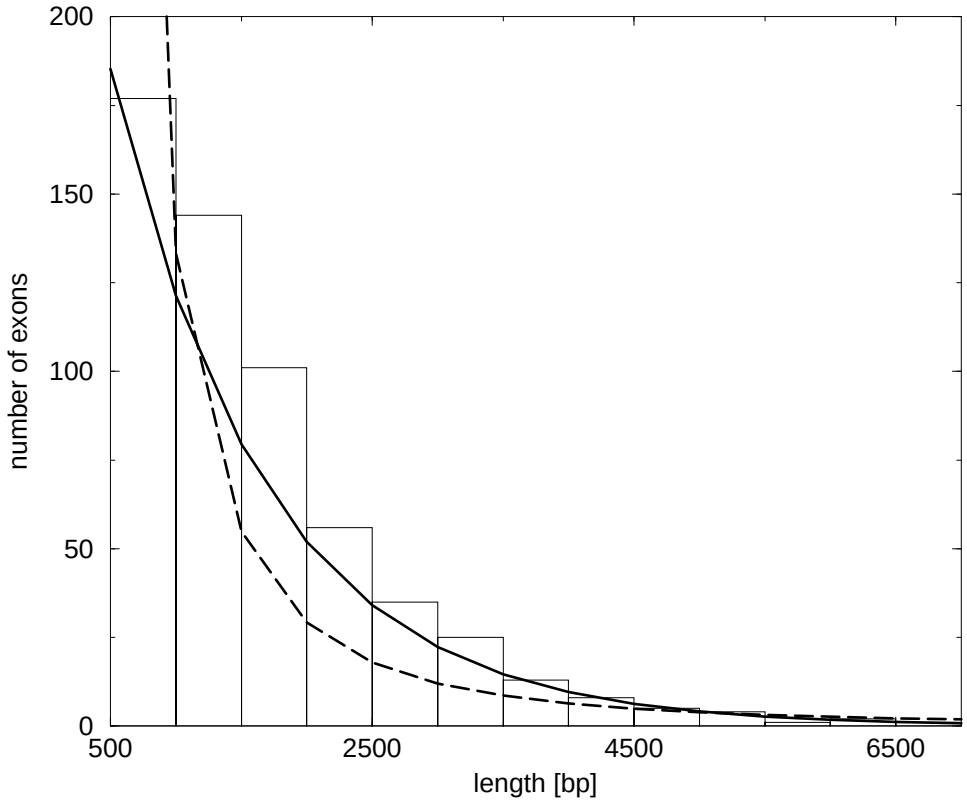


Figure 14.5: Histogram of open reading frames (ORF's) longer than 500 bp from the yeast chromosomes III, IX, and XI. Regression by an exponential function and a power-law decay are indicated by full and dashed lines, respectively.

codon usage table. This is, of course, a strong assumption since it is known that the codon usage depends, e.g., on the degree of gene expression [Ikemura 1981, Sharp & Li 1987]. However, Sharp and Li claim that "within species the differences are largely in the degree rather than the direction of codon usage bias" [Sharp & Li 1987]. If whole chromosomes are analyzed, one has to take into account that genes are located on both strands. Therefore we use in our simulations of *pseudo-chromosomes* (see section VI) also complementary codon usage tables.

As discussed in the preceding section the nonuniform codon usage leads to specific statistical dependences within exons. These are quantified below by the dependence matrix

$$D_{ij}^{exon}(k) = p_{ij}^{exon}(k) - p_i^{exon} \cdot p_j^{exon} \tag{14.15}$$

In the following we denote the total fraction of protein coding sequences in a given DNA sequence by F. For the yeast chromosomes we have, for example, $F \approx 0.7$ [Feldmann et al. 1994]. The task is now to estimate the correlation decay for a given sequence length N, fraction of coding segments F, and probability distribution $\rho(l)$.

The mean exon length is given by

$$\bar{l} = \sum_l l \cdot \rho(l). \tag{14.16}$$

For yeast DNA, where genes exhibit only a few introns, the mean exon length is about 1400 bp. The typical length scale of human exons is a few hundred base pairs. However, there are also exons with a length of several thousand base pairs (e.g. exon 11 of the BCRA1 gene comprises 3426 bp).

The expectation value $\bar{n}$ of the number of exons in a sequence of length N and an exon fraction F is

$$\bar{n} = \frac{F \cdot N}{\bar{l}}. \tag{14.17}$$

Consequently, the average number of exons with a length l is given by

$$n(l) = \rho(l) \cdot \bar{n} = \frac{F \cdot N \cdot \rho(l)}{\sum_l l \cdot \rho(l)}. \tag{14.18}$$

Since we focus in this chapter on correlations due to the nonuniform codon usage, we neglect statistical dependences of pairs A_i and A_j which are not within the same exon. This implies, for example, that the base composition in exons and introns is considered to be the same. Generalizations of this simplified approach are discussed in the final section.

Since every exon contributes $l-k$ pairs, we obtain the number $Z(k)$ of pairs which are located in the same exon:

$$Z(k) = \sum_{l=k+1}^{l_{max}} (l-k)\, n(l). \tag{14.19}$$

Here, overlaps of protein coding segments have been neglected. The total number of pairs in a distance k is $N-k$, and hence, the overall deviations $D_{ij}(k)$ from statistical independence are given by

$$D_{ij}(k) = \frac{Z(k)}{N-k} \cdot D_{ij}^{exon}(k). \tag{14.20}$$

This result can now explain the decay of correlation functions and the mutual information function, since both measures can be obtained from the decay of the $D_{ij}(k)$ (cf. Eqs.(3)-(7)). It can be seen that beside the internal period-3 oscillations described by $D_{ij}^{exon}(k)$ we obtain a k-dependent pre-factor related via $Z(k)$ to the exon length distribution. Since we are primarily interested in the long-ranging correlations, we focus in the following on the envelope

$$E(k) \propto \sum_{l} \frac{l-k}{N-k} \cdot n(l). \tag{14.21}$$

This formula is a central result of this chapter. It elucidates the immediate effect of the length distribution of exons on the decay properties of correlation measures, which we will exemplify in the following section.

14.5 Models of Length Distributions

Now we illustrate the considerations of the preceding section for 3 representative probability distributions $\rho(l)$, namely a uniform, an exponential, and a power-law distribution.

We analytically derive the corresponding decay laws and test the predictions using *pseudo-genes*[1]. These consist of interspersed pseudo-exons within a *random sea*, i.e., statistically independent letters with the same base composition as the pseudo-exons. The length of each exon is chosen randomly from the distribution $\rho(l)$ under consideration. In all simulations in this section we have chosen the codon usage table of the HUMBMYH7 gene studied in section III. Details of the simulations are described in the figure captions.

[1] We use this terminus in analogy to *pseudo-exons* and *pseudo-chromosomes* for corresponding stochastic model sequences. It should not be confused with knocked out genes which are termed pseudo genes as well.

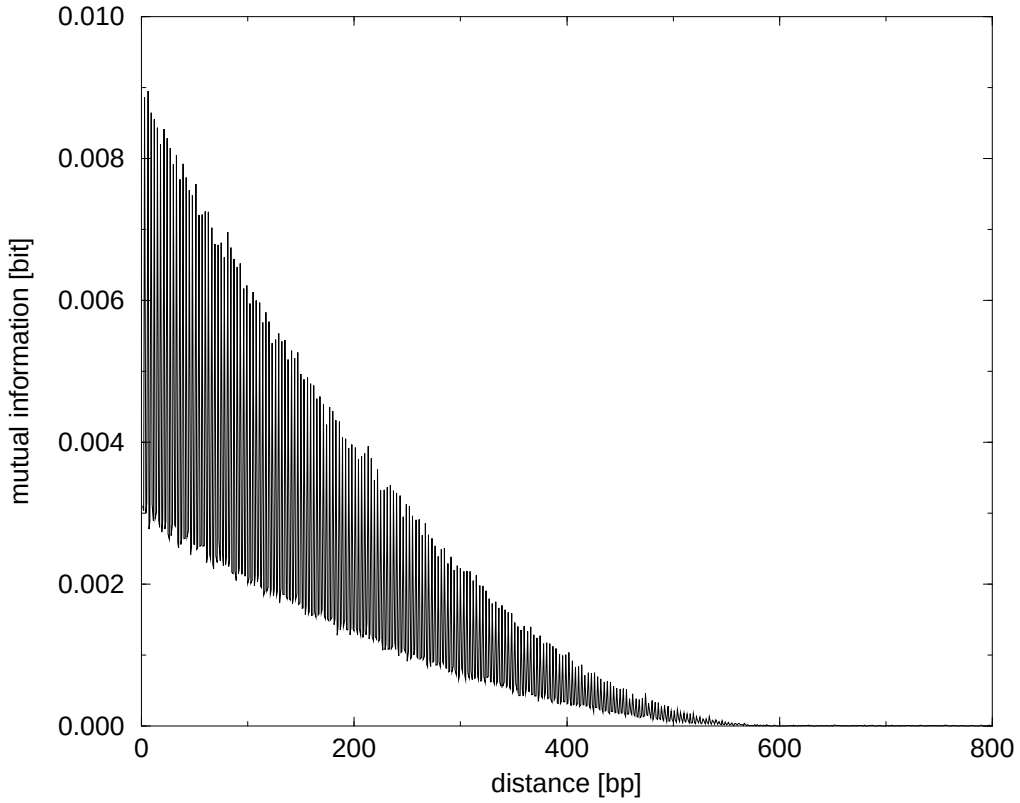


Figure 14.6: Mutual information of a 10^6 bp long random sequence. Within a "random sea" of independent letters A, C, G, and T, 1000 pseudo-exons of a length 600 bp have been interspersed. For small k, we observe the expected period-three oscillations between $F^2 I_{in}$ and $F^2 I_{out}$ (see Eqs. (7) and (20)). Please note that Eq. (24) predicts exactly the parabolic decay between $k = 0$ and $k = 600$.

As a first model we discuss a fixed length of all protein coding segments $l = L$,

$$\rho(l) = \delta_{lL}. \tag{14.22}$$

This yields

$$\bar{n} = n(L) = \frac{F \cdot N}{L}. \tag{14.23}$$

For $k < L$ we obtain essentially a linear decay of the envelope $E(k)$:

$$E(k) \propto \frac{F \cdot N}{N-k} \frac{L-k}{L} \approx F \cdot \frac{L-k}{L}. \tag{14.24}$$

The k-dependence of the denominator can be neglected for $N \gg L$. The resulting linear decay of $D_{ij}(k)$ implies a quadratic decay of the mutual information function (cf. Eq. (7)). Such a parabola is seen in Figure 6 for a pseudo-gene with constant exon length.

Of course, it is more realistic to assume an exponentially decaying length distribution (compare Figure 5). As above in Eq. (24), we neglect the k-dependence of the denominator. For the sake of simplicity, we further replace the summation in Eq. (21) by an integration from k to infinity. Then an exponential length distribution

$$\rho(l) = \lambda \ \exp(-\lambda\, l) \tag{14.25}$$

gives an exponential decay of the envelope:

$$E(k) \propto F \cdot \lambda \int_k^\infty (l-k) \cdot \rho(l)\, dl = F \ \exp(-\lambda\, k). \tag{14.26}$$

Figure 7 displays the results for a corresponding simulation of a pseudo-gene.

As a last example we consider a power-law decay from a lower cut-off length L_{min} with an exponent $\beta > 2$:

$$\rho(l) = (\beta - 1)\, L_{min}^{\beta-1}\, l^{-\beta} \qquad \text{for} \qquad l \geq L_{min} \tag{14.27}$$

and zero otherwise. The mean value of the length is then given by

$$\bar{l} = \frac{\beta - 1}{\beta - 2}\, L_{min}. \tag{14.28}$$

After integration we obtain a power-law decay of the envelope for $k > L_{min}$:

$$E(k) \propto \frac{F \cdot L^{\beta-2}}{\beta - 1}\, k^{2-\beta}. \tag{14.29}$$

The log-log presentation of the mutual information in Figure 8 indicates indeed a power-law for a simulation of a corresponding pseudo-gene.

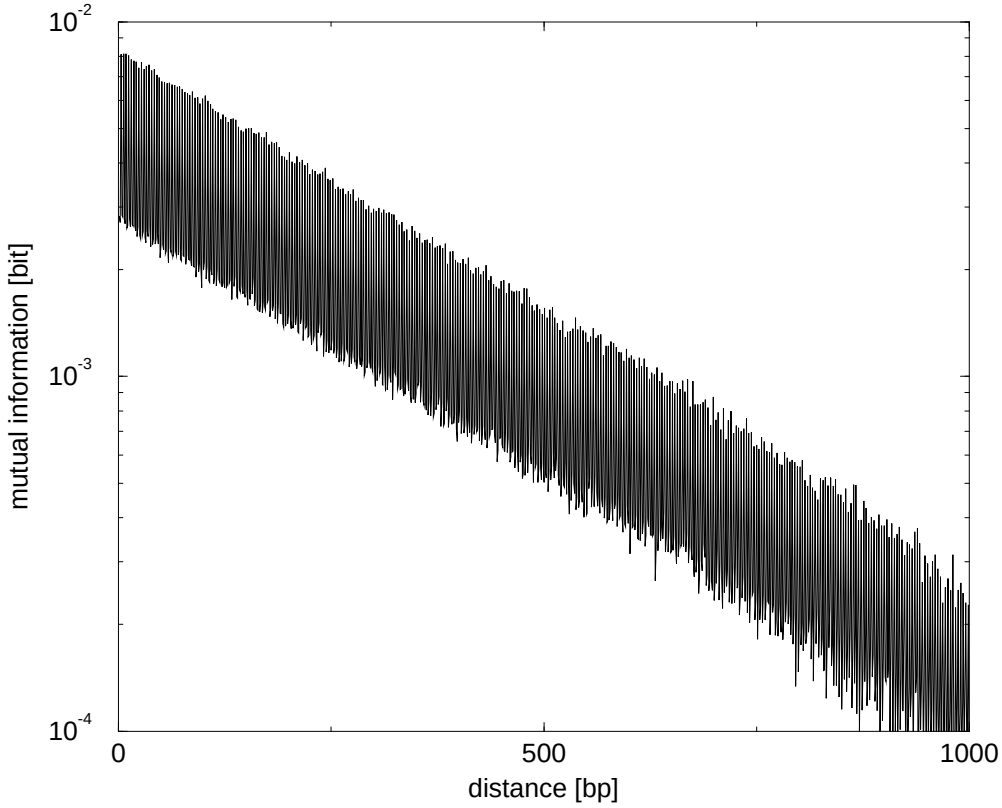


Figure 14.7: Mutual information of a 10^6 bp long sequence containing 1000 pseudo-exons with exponentially distributed lengths (mean value 600 bp). The logarithmic vertical scale reveals the predicted exponential decay.

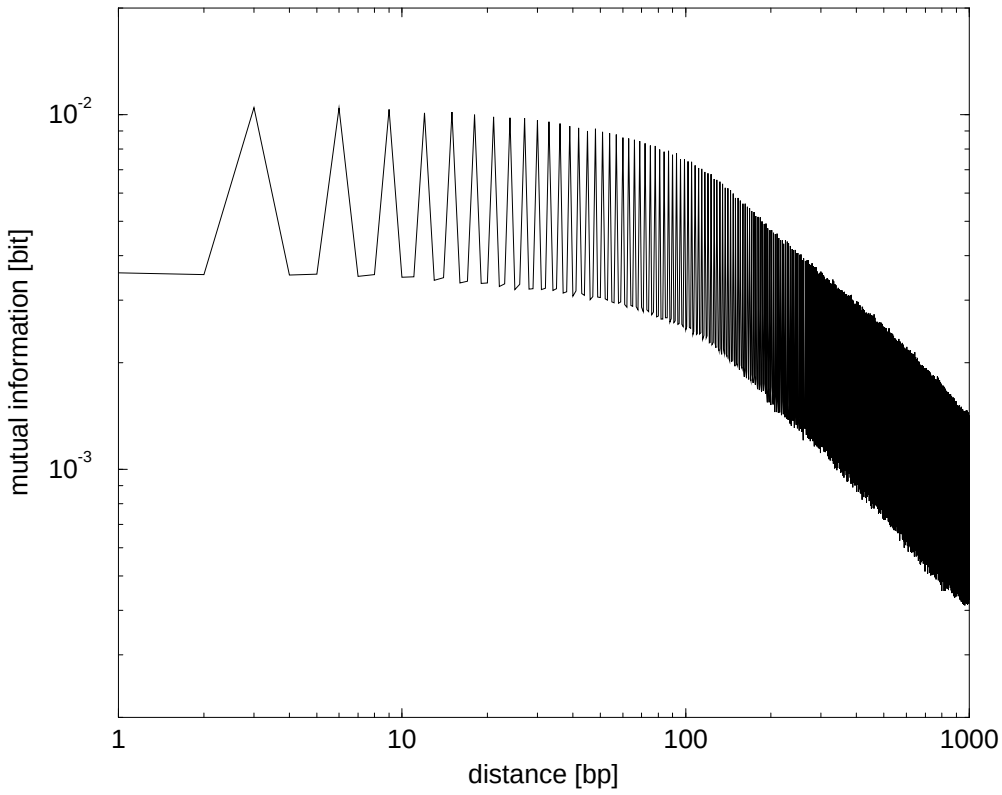


Figure 14.8: Mutual information of a $7 \cdot 10^6$ bp long sequence with 7000 pseudo-exons. The parameters of the exon length distribution are $L_{min} = 150$ and $\beta = \frac{9}{4}$.

These examples show how the length distribution of exons affects the decay of correlations, which are due the nonuniform codon usage. In the next section we show that our considerations apply to DNA sequences and that a considerable amount of observed correlations can be predicted just by knowing codon usage tables and the length distribution of exons.

14.6 Applications to DNA Sequences

In this section we apply our concept to representative DNA sequences. It was already demonstrated in Figure 1 that the periodicity due to the nonuniform codon usage plays a significant role in yeast DNA. This is due to large fraction of coding sequences ($F \approx 0.7$) and rather long exons (compare Figure 5). In order to quantify the effect of exons on correlations we generate *pseudo-chromosomes* as follows: codon usage tables are taken from long yeast genes as a basis for the simulation of pseudo-exons (see section III). In order to simulate strand symmetry, 50 % of the pseudo-exons are generated with the complementary codon usage table. The empirical histogram from the corresponding chromosome is taken as length distribution for the interspersed pseudo-exons. In between the pseudo-exons Bernoulli sequences with the same base composition are inserted. In this way a stochastic model of a chromosome is defined which incorporates only well-known features – the nonuniform codon usage and the alternation of coding segments and intervening sequences.

Figure 9 reveals that the decay for the pseudo-chromosomes is quite similar to the actual decay for the yeast chromosomes. Only for small k additional correlations can be seen which are discussed in the final section.

Similar agreement was also found for codon usage tables from other protein coding segments and for some strand asymmetry.

Significant long-range correlations in the yeast chromosome III up to several kilo base pairs have been reported by Munson et al. [Munson 1992]. The existence of such correlations is indeed corroborated by our mutual information analysis. However, they exist also in a pseudo-chromosome (see Figure 9), and hence, the length distribution of exons is sufficient to explain these correlations.

In the same way as for the yeast chromosome, we generated a stochastic model of a DNA region of E. coli (see Figure 2). Figure 10 shows a comparison of the mutual information

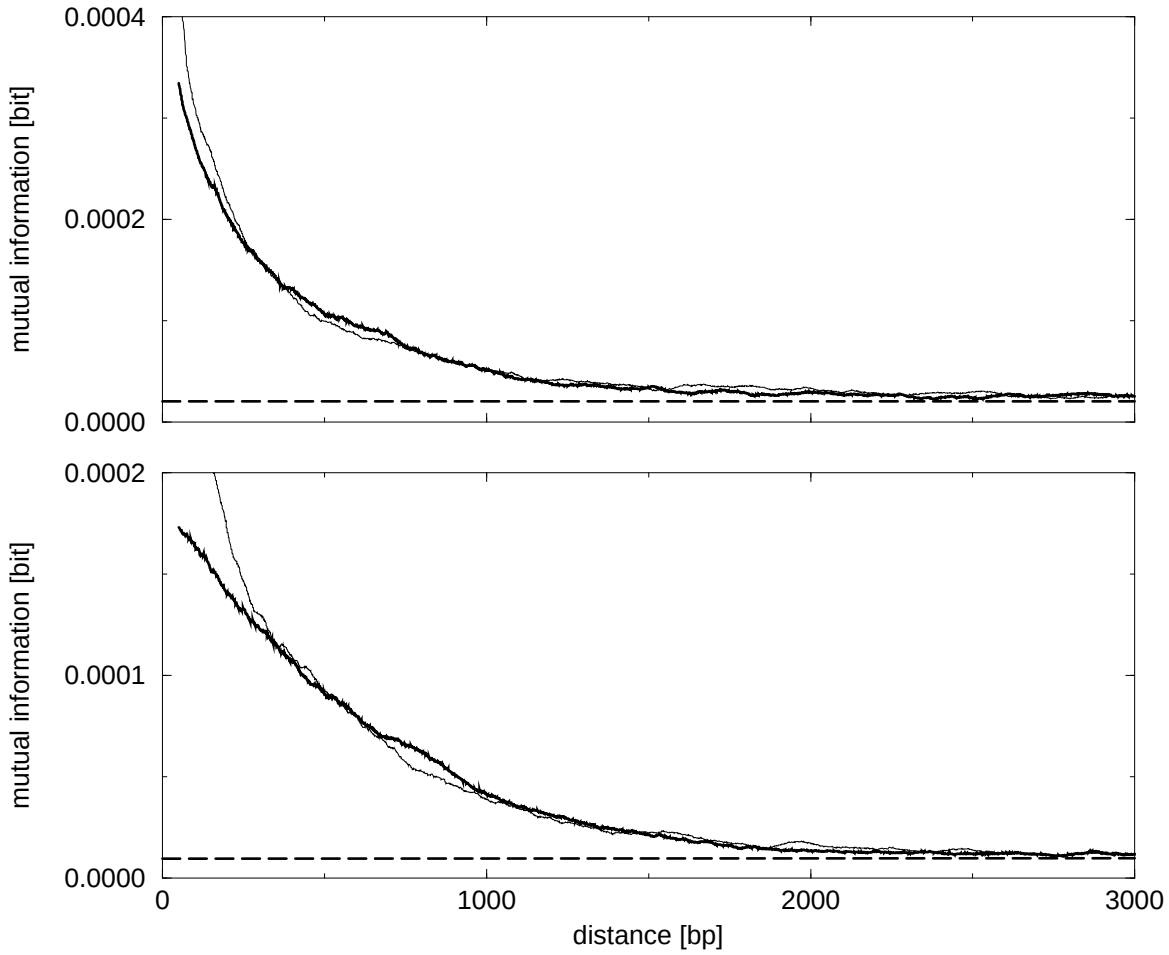


Figure 14.9: Decay of the mutual information function for yeast chromosomes (thin lines) and the corresponding pseudo-chromosomes (thick lines). In order to reduce the strong fluctuations (compare Figure 1) and to focus on the decay we have applied a 99 bp running average. Upper graph: Chromosome III. The codon usage table was taken from the temperature-sensitive lethal TSM1 protein (4,221 bp). Lower graph: Chromosome XI, table from the ORF which encodes dynein (12,276 bp).

functions.

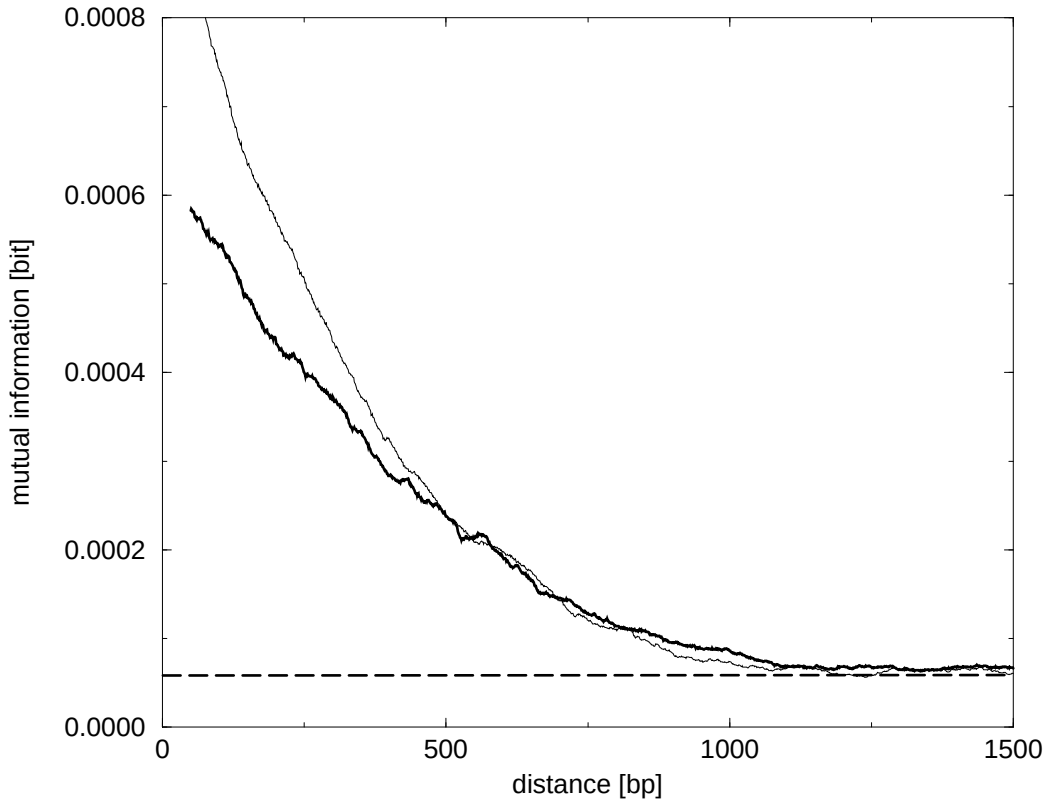


Figure 14.10: Mutual information decay for the E. coli chromosome region (thin line) and a corresponding pseudo-region with the same length distribution of pseudo-exons. The codon usage table was taken from the isoleucil-tRNA ligase (2,811 bp). As in Figure 9 a 99 bp running average was applied.

Finally, we discuss the correlation decay in the myosin heavy chain gene M74000 of *Brugia malayi*. We have chosen this gene since the 15 exons constitute about 68 % of the total gene. Consequently, the correlations due to the exons and their length distribution are more pronounced then in genes with only a few percent of exons.[2] The codon usage-table and empirical length distribution of the analyzed gene are taken to generate a pseudo-

[2]In fact, the decay for the human myosin heavy chain depicted in Figure 3 is also strongly influenced by correlations within its introns.

gene as described in section V. Since there are fairly long exons in this gene, Figure 11 displays the expected long tail of the envelope. Quite similar correlations are found in the corresponding pseudo-gene (thick line) pointing to the fact that most correlations are solely due to the length distribution of exons.

It turns out that for such relatively short DNA sequences a careful calculation of the bias (dashed line in Figure 11) is necessary for a correct interpretation of the decay.

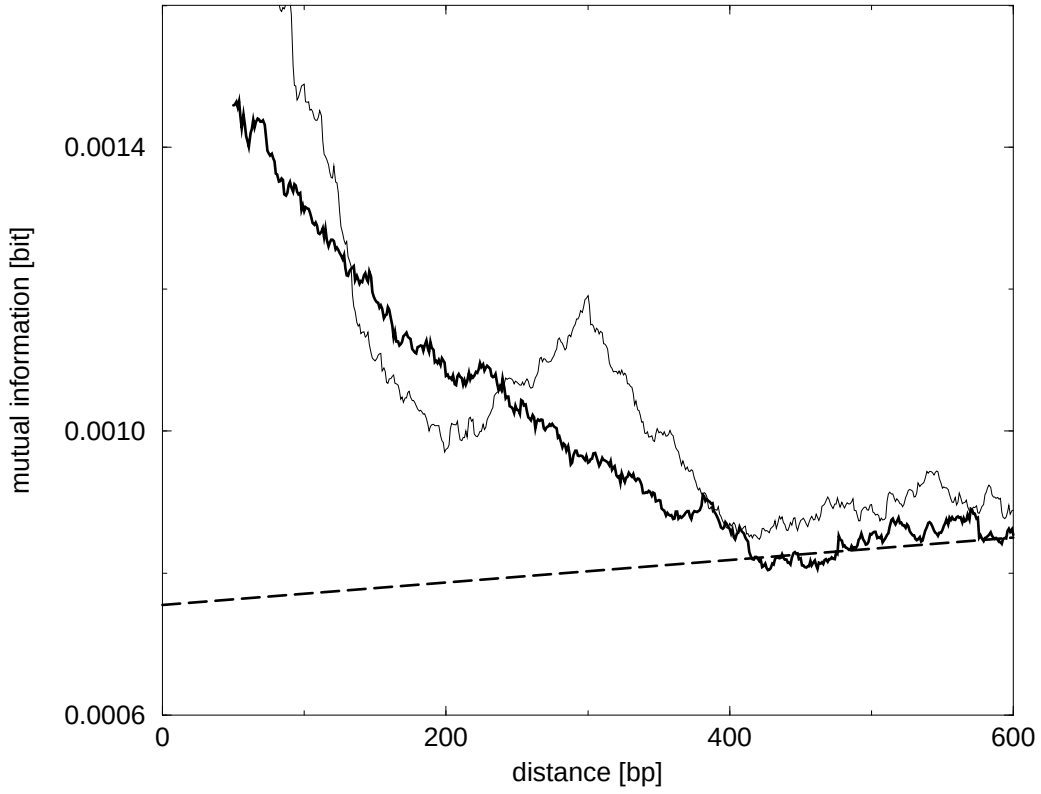


Figure 14.11: Comparison of the smoothed mutual information (99 bp running average) of Brugia malayi myosin heavy chain gene (8,600 bp from the first to the last exon) and a corresponding random sequence with the same exon length distribution and codon usage. Since the sample size decreases with the distance there is a clear increase of the bias (see also the Appendix).

14.7 Summary and Discussion

This chapter was devoted to relate a significant part of observed long-range correlations to the pattern of protein coding segments. We have shown that the triplet code induces via a nonuniform codon usage persistent oscillations of correlation measures. By taking into account the length distributions of exons, a long-ranging decay of the mutual information function and correlation functions could be predicted. For example, a power-law distribution of the exon length implies a power-law decay of correlation measures.

Pseudo-chromosomes based on the empirical length distribution in yeast chromosomes exhibit a quite similar decay of correlations and, therefore, most of the correlations in yeast DNA could be traced back to a simple origin. Our considerations apply to all parts of genomes where coding segments constitute a significant portion of the DNA such as bacteria or retroviruses. This was exemplified for a DNA region of E. coli and for a myosin heavy chain gene with a large fraction of exons.

Typically, in higher eukaryotes only a few percent of the DNA are protein coding regions. Consequently, observed long-range correlation in DNA as the human β-globin region [Peng et al. 1992] or in genes with very long introns [Li & Kaneko 1992] cannot be explained simply by the nonuniform codon usage within exons. Moreover, the well-known compositional variations along chromosomes on scales above 10^5 bp [Bernardi et al. 1985, Feldmann et al. 1994, Dujon et al. 1994] are beyond the scope of our analysis.

Our concept is however more generally applicable. It can be formulated as follows:

- look for fragments of differing statistical properties,
- analyze its length distribution,
- define appropriate (stochastic) pseudo-sequences,
- analyze their correlation decay,
- compare it with the empirical mutual information.

Related stochastic models of the DNA heterogeneity have a long tradition [Elton 1974, Fickett et al. 1992, Herzel et al. 1994b, Buldyrev et al. 1993], but these models are based on hypothetical length distributions of fragments. Contrarily, our approach simply exploits the well-known length distribution of exons.

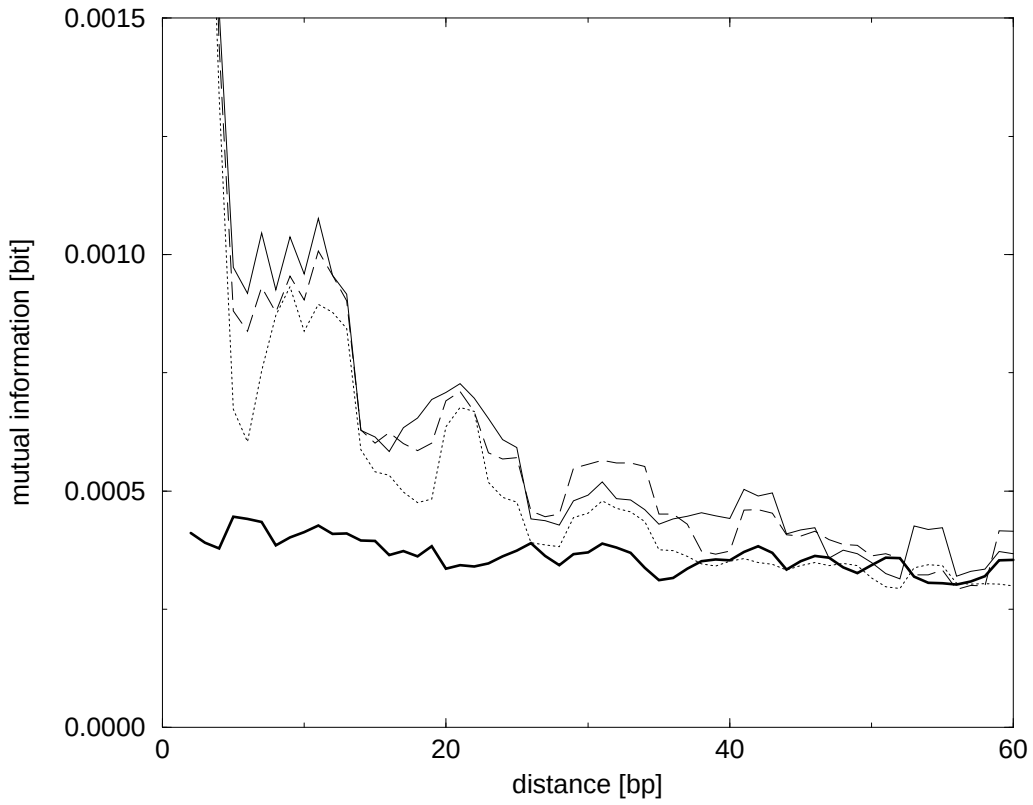


Figure 14.12: Mutual information of yeast chromosomes III (full line), IX (dashed line), XI (dotted line) for short distances. In order to eliminate the dominating period-three oscillations, we apply a running average over 3 bp. The comparison with a pseudo-chromosome (thick line) reveals additional correlations (in particular a 10-11 bp period).

As a first step of a more general approach, Schmitt et al. [Schmitt et al. 1996] recently studied length distributions of over-represented "words" termed *modules*. We suggest to analyze also length distributions of–for example–isochores [Bernardi 1989], gene clusters, dispersed repeats, simple-sequence DNA, or CpG islands. If one takes into account different compositions of exons and introns, the length distribution of introns comes into play as well. We expect that stochastic models which include the actual length distributions of all these segments can relate most observed long-range correlations to known biological structures.

Though we have quantitatively explained the origin of long-range correlations in mostly protein coding sequences, many questions remain open. For example, correlations within introns and intergenic sequences were not the subject of this chapter. Moreover, we have seen in Figure 9 additional correlations in yeast DNA for small distances, which cannot be explained by our pseudo-exon concept. Figure 12 displays an example of such a peak structure with a periodicity of about 10 bp. These peaks may reflect the *pitch* of DNA, i.e., a 10.5 bp periodicity that has been found in curved DNA [Trifonov & Sussman 1980, Konopka & Smythers 1987] and DNA folded into nucleosomes [Ioshikhes et al. 1992]. Additionally, the well-known 3-4 amino acid periodicities in α-helical proteins [Herzel 1988, Kanehisa & Tsong 1980, White 1994, Schmitt et al. 1997] are a possible source of the observed peak structure.

In summary, we have shown in this chapter that the length distribution of exons in real DNA induces long-range correlations which can be described by appropriate stochastic models. We stress, finally, that beyond these correlations other DNA base-pair fluctuations exist on various scales [Trifonov 1989, Li & Kaneko 1992, Peng et al. 1992, Voss 1992, Bernardi 1989, Dujon et al. 1994]. Their role for the chromosome organization and gene expression has still to be explored.

Chapter 15

Interpreting Correlations in DNA and Protein Sequences

Understanding the complex organization of genomes as well as predicting the location of genes and the possible structure of the gene products are some of the most important problems in current molecular biology. Many statistical techniques are used to address these issues. A central role among them play correlation functions. In this chapter we study the decay of the entire 4×4 dimensional covariance matrix of DNA sequences. We apply this covariance analysis to human chromosomal regions, yeast DNA, and bacterial genomes, and we interpret the three most pronounced statistical features – long-range correlations, a period 3, and a period 10-11 – using known biological facts about the structure of genomes. For example, we relate the slowly decaying long-range G+C correlations to dispersed repeats and CpG islands. We show quantitatively that the 3-base-pair-periodicity is caused to a significant degree by the nonuniformity of the codon usage in protein coding segments. We also show that periodicities of 10-11 base-pairs in yeast DNA may possibly originate from an alternation of hydrophobic and hydrophilic amino acids in their corresponding protein sequences.

15.1 Introduction

Correlation functions of DNA and protein sequences have been widely studied in order to understand the complex organization of genomes. Many correlation measures, such as binary correlation functions [Trifonov & Sussman 1980, Shepherd et al. 1981], power spectra

[Voss 1992], random walk variances [Peng et al. 1992], or the mutual information function [Ebeling et al. 1987, Li & Kaneko 1992] have been employed to address these questions.

Correlation functions of DNA sequences (as well as power spectra or the intimately related random walk approaches [Stanley et al. 1994]) are based on the assignment of numbers to the nucleotides A, C, G, and T. Analyzing only single mappings of numbers to symbols limits the utility of those approaches. In particular, it can be shown that even the infinite set of all possible autocorrelation functions of a given quaternary sequence cannot measure all statistical dependences between the four symbols A, C, G, and T. As pointed out in [Herzel & Grosse 1995], it is necessary to analyze at least three crosscorrelation functions in addition to at least six autocorrelation functions, in order to gather all information about statistical dependences in quaternary sequences. Therefore, we study the full spectrum of 16 elementary correlation functions and emphasize in this chapter that the comparison of different correlation functions can give insight into the biological meaning of observed correlations.

First, we discuss long-range correlations in long human DNA sequences. It turns out that the correlation decay depends strongly on the parameters of the chosen correlation function.

For the interpretation of observed periodicities of 10-11 base-pairs (bp), a comparison of different correlation functions points to their biological origin: an alternation of hydrophobic and hydrophilic amino acids, which induces specific periodicities in correlation functions.

15.2 Symbols and Definitions

Correlation functions of numerical time series $\{x_\ell\}$ are defined by

$$C_{xx}(k) = \langle x_\ell x_{\ell+k} \rangle - \langle x_\ell \rangle \langle x_{\ell+k} \rangle, \tag{15.1}$$

where $\langle ... \rangle$ denotes the time average over the entire sequence $\{x_\ell\}$. In case of analyzing symbolic sequences, e.g. if $x_\ell \in \{\mathrm{A, C, G, T}\}$, it is necessary to map those symbols $\{x_\ell\}$ to numbers $\{y_\ell\}$, before these numbers $\{y_\ell\}$ can then be multiplied. In the following, we will denote these symbol-to-number-assignments by the vectors $\vec{a} \equiv (a, c, g, t)$ or $\vec{b}$, where a stands for the real number mapped to nucleotide A, etc.

In our figures we present, for example, adenine (A) autocorrelation functions $C_{AA}(k)$ with $\vec{a} = \vec{b} = (1, 0, 0, 0)$, crosscorrelation functions $C_{AT}(k)$ with $\vec{a} = (1, 0, 0, 0)$ and $\vec{b} =$

$(0,0,0,1)$, autocorrelation functions of the weakly binding nucleotides (A and T) C_{WW} with $\vec{a} = \vec{b} = (1,0,0,1)$, or purine autocorrelation functions C_{RR} with $\vec{a} = \vec{b} = (1,0,1,0)$.

In order to express the autocorrelation function of a given sequence $\{x_\ell\}$ and a given mapping $\vec{a} = \vec{b}$ in terms of the statistical dependences between all 4×4 pairs of nucleotides, let us denote the probability of a nucleotide to occur in the sequence $\{x_\ell\}$ by p_i $(i = 1,2,3,4)$ and the pair probability to find nucleotide i at a certain position and the nucleotide j exactly k nucleotides downstream by $p_{ij}(k)$ $(i,j = 1,2,3,4)$.

Then, the statistical dependences between all nucleotides can be quantified by the 4×4 numbers $D_{ij}(k) \equiv p_{ij}(k) - p_i p_j$, which form the so called "dependence matrix" [Herzel & Grosse 1995].

Using these notations, any correlation function $C_{\vec{a}\vec{b}}(k)$ can be expressed as a bilinear form of the dependence matrix, i.e. [Weiss & Herzel 1997]

$$C_{\vec{a}\vec{b}}(k) = \sum_{i,j=1}^{4} a_i b_j p_{ij}(k) - (\sum_{i=1}^{4} a_i p_i)(\sum_{j=1}^{4} b_j p_j) = \sum_{i,j=1}^{4} a_i b_j D_{ij}(k). \tag{15.2}$$

Another measure of pair correlations, which has the mathematical property to map *all* statistical dependences onto a single number, is the mutual information [Kullback 1959]. Here we define the mutual information function as the amount of information (in bit) that one obtains about a hidden nucleotide by getting to know the nucleotide k positions upstream. In mathematical terms, the mutual information function $I(k)$ is defined by

$$I(k) = \sum_{i,j=1}^{4} p_{ij}(k) \log_2 \frac{p_{ij}(k)}{p_i p_j} \tag{15.3}$$

This function vanishes if and only if all $D_{ij}(k)$ are equal to zero. Therefore, $I(k) \equiv 0$ implies the statistical independence of all nucleotides in all distances k.

By Taylor expanding the mutual information function $I(k)$ as a function of the joint probabilities $p_{ij}(k)$ about the points of statistical independence $p_i \cdot p_j$, we can establish an analytic relation between $I(k)$ and the covariance functions $D_{ij}(k)$:

$$I(k) = \frac{1}{2 \cdot \ln 2} \sum_{i,j=1}^{4} \frac{D_{ij}(k)^2}{p_i p_j} + \mathcal{O}(D_{ij}(k)^3) \tag{15.4}$$

Since statistical dependences are usually extremely weak in biosequences, the $\mathcal{O}(D_{ij}(k)^3)$ terms can be neglected. This means that the mutual information function is approximately equal to the weighed sum over the squares of all 16 covariances.

15.3 Long-range correlations in human DNA

This section is devoted to the analysis of correlations in two long chromosome regions: HSFLNG6PD (219,447 bp) – a chromosome X region – and HUMTCRB (684,973 bp) – the human T-cell receptor beta locus. We exemplify, that different "alphabets", i.e. different assignments of numbers to nucleotides, lead to quite different correlation functions in single sequences. A particular feature, the slow decay of the G+C correlations is related to well-known biological structures.

Earlier studies of chromosome regions (e. g. [Peng et al. 1992, Peng et al. 1994]) using the random walk approach provided indications of a power law decay:

$$C_{RR}(k) \sim k^{-\gamma} \tag{15.5}$$

In [Peng et al. 1992], autocorrelations corresponding to our notation $\vec{a} = \vec{b} = (1,0,1,0)$ have been analyzed. Since the random walk exponent α [Peng et al. 1992] is intimately related to γ via

$$\gamma = 2 - 2\alpha, \tag{15.6}$$

we can directly compare their observations with correlation patterns of $C_{RR}(k)$.

Consequently, values of $\alpha = 0.61$ (human T-cell receptor alpha/delta locus [Peng et al. 1994]) or $\alpha = 0.71$ (human beta-globin region [Peng et al. 1992]) correspond to $\gamma = 0.78$ or $\gamma = 0.58$ respectively. According to Eq. (15.4) one expects the mutual information to have a power-law exponent of 2γ, since $I(k)$ is in a good approximation the sum of squared correlation functions. The upper curve in Figure 1, however, reveals that the mutual information function decays relatively slowly (corresponding to $\gamma \approx 0.35$). In order to understand which correlations govern the decay of the mutual information, we analyze all 16 auto- and crosscorrelations. It turns out that their decay differs drastically. In Figure 1 two representative examples are shown: In accordance with earlier studies [Peng et al. 1992, Peng et al. 1994] the correlation function C_{RR} decays relatively fast ($\gamma \approx 0.66$), whereas C_{WW} decays slowly ($\gamma \approx 0.29$). Similarly, the T-cell receptor beta locus exhibits a wide range of exponents for different entries of the covariance matrix, such as $\gamma \approx 0.49$ for C_{WW}, $\gamma \approx 0.70$ for C_{RR}, and $\gamma \approx 0.48$ for $I(k)$. Obviously, the decay of the mutual information is dominated by the slow decay of C_{WW}, i. e. autocorrelations of the A+T or – equivalently – the G+C content.

The observed slow decay of the G+C correlations quantifies various earlier indications of long-ranging variations of the G+C content in DNA sequences [Elton 1974,

Bernardi et al. 1985, Fickett et al. 1992, Bettecken et al. 1992]. There are many well-known sources of G+C variations in human DNA:

- difference in G+C content in coding and noncoding regions [Fickett 1982]
- Alu repeats (G+C rich, up to 300 bp long [Korenberg & Rykowski 1988])
- L1 repeats (G+C poor, up to 6400 bp long [Korenberg & Rykowski 1988])
- CpG islands (typically 500-2000 bp long [Clay et al. 1995])

The length distribution of these structures and their clustering in isochores [Bernardi et al. 1985, Korenberg & Rykowski 1988, Ikemura et al. 1990] induce long-ranging fluctuations of the G+C content. Using the documentation of the chromosome regions studied, we can discuss these features more specifically: The G+C rich chromosome X region from which Figure 1 is derived contains more than 300 Alu repeats and 17 CpG islands with a mean length of about 1000 bp. These regions can easily explain the slow decay of the C_{WW} correlation function.

15.4 Periodicities in yeast and bacterial DNA

In the following we will show that the comparison of different correlation functions is also helpful to understand short-range periodicities of DNA sequences. It is well known that correlation functions of coding segments exhibit pronounced period-three oscillations [Trifonov & Sussman 1980, Shepherd et al. 1981]. This is a simple consequence of the nonuniform frame dependent nucleotide probabilities, which in turn is caused by the nonuniform codon usage. In particular, there is typically an excess of guanine (G) in the position 1 of the reading frame [Staden 1984] which leads to strong period-3-oscillations of the GG-autocorrelation (compare Figure 2). If the open reading frame is not interrupted by introns the period-three oscillations may extend over thousands of base-pairs [Herzel & Grosse 1995]. The decay of the envelope can be calculated from the length distribution of coding segments. It was shown in [Herzel & Grosse 1997] that long-range correlations in yeast DNA can be explained by the long tail in the exon length distribution.

In addition to the period 3, an oscillation of the mutual information with a period of 10-11 base pairs has recently been detected in several yeast chromosomes [Herzel & Grosse 1997]. These periodicities become visible in Figure 2 after applying 3

bp running averages leading to the thick lines. It turns out that the 10-11 bp periodicity is strong in C_{AA} (or C_{TT}) autocorrelation functions but weak in the C_{GG} (or C_{CC}) functions. Moreover, the crosscorrelation function C_{AT} exhibits a phase shift of half a period. We will discuss these features in the next section in connection with protein sequences.

Since the autocorrelation function of weak the nucleotides A and T can be written as

$$C_{WW} = C_{AA} + C_{TT} + C_{AT} + C_{TA} \tag{15.7}$$

one expects particularly strong oscillations for C_{WW}. Indeed, we find a rather strong periodicity in bacterial sequences, which is illustrated by Fig 3.

In [Herzel & Grosse 1997] possible explanations for the observed 10-11 bp periodicities have been proposed: DNA bending [Trifonov & Sussman 1980] or nucleosomal signals (see [Marini et al. 1982, Ioshikhes et al. 1996] for details) or periodicities of the corresponding protein sequences [Weiss & Herzel 1997, Kanehisa & Tsong 1980, Schmitt et al. 1997, Zhurkin 1981]. If the nucleosomal pattern would be the major source of the 10-11 base-pair oscillation, they should be detectable in noncoding DNA as well. Since we find no pronounced peaks 10,11 or 21 bp in introns or human chromosome regions (compare Figure 1), we now test the hypothesis [Zhurkin 1981] that the observed periodicities are induced by correlations in protein sequences.

15.5 Correlations in protein sequences

Proteins are composed of 20 amino acids. Therefore, correlations can be characterized by a dependence matrix $D_{ij}(k)$ with 20×20 entries. After taking into account normalization constraints for the elements $D_{ij}(k)$, still 19×19 of those D_{ij} can be independent [Herzel & Grosse 1995].

Consequently, there is a lot of freedom in the choice of the assignment vectors $\vec{a}$ and $\vec{b}$, and correlations depend strongly on the chosen vectors $\vec{a}$ and $\vec{b}$. If $\vec{a}$ and $\vec{b}$ contain only zeros and ones, any vector can be considered as a classification of the 20 amino acids into two groups. In [Weiss & Herzel 1997] correlation functions of 10^7 classifications have been studied. It turned out that the strongest correlation signals are associated with two groups of amino acids: L, I, V, F, M (mostly hydrophobic and non-polar) and E, K, D, R, Q (hydrophilic and polar). In Figure 4 the corresponding auto- and crosscorrelations are shown for a global set of 2912 proteins [White 1994].

Averaging over many protein sequences reduces statistical fluctuations and reveals representative correlation patterns. It can be seen in Figure 4 that there are significant autocorrelations at 3 and 4 residues. This means that hydrophobic (and –likewise– hydrophilic) amino acids have a preferred distance of 3-4 amino acids. The peak of the crosscorrelation at $k = 2$ indicates that there is a tendency of L, I, V, F, M to alternate with E, K, D, R, Q. These patterns are related to α-helices - a frequent secondary structure element of folded proteins [Schmitt et al. 1997]. A typical α-helix has a helical repeat of about 3.6 residues [Creighton 1993]. If the helix would have, say, hydrophobic residues preferentially on one side, the corresponding protein sequence would have a 3.6 residue periodicity corresponding to a 10.8 nucleotides in the associated DNA sequence.

The genetic code is strongly degenerated, since 64 triplets of nucleotides are mapped onto stop signals plus 20 amino acids. For example, leucine is encoded by 6 different codons. It is therefore not evident that the relatively weak correlations of protein sequences can induce significant periodicities in the corresponding DNA sequences. It turns out that the middle letter of the codons is closely related to the biochemical properties of the amino acid. More specifically, all five amino acids L, I, V, F, and M have at that position a T and E, K, D, and Q posses an A in the middle of all their codons. Consequently, a distance of 3 hydrophobic (or hydrophilic) amino acids implies an occurrence of an TT (or AA) pair in a distance of $k = 9$ nucleotides. Note that there is indeed a peak at $k = 9$ in C_{AA} in Figure 1. Moreover, an alternation of hydrophobic and hydrophilic amino acids implies a phase shifted oscillation of C_{AT} as in Figure 1.

In order to test the hypothesis that the observed periodicities in DNA sequences (see Figures 2, 3) are caused by the correlations in the amino acid sequences of the expressed proteins, we design the following experiment: We start with the global set of 2912 protein sequences, and translate the amino acid sequences back to pseudo-DNA sequences by using a uniform codon usage, i. e. codons are selected randomly with equal probability for each codon. In this way we guarantee that no additional correlation (e. g. another overlapping code in the third position) is introduced. Therefore, the resulting periodicities will be a consequence of the correlations in the protein sequences only.

The resulting Figure 5 resembles clearly the periodicities shown in Figure 2. The strong period-three oscillations of C_{GG}, the 10-11 bp periodicities of C_{AA} and C_{AT} , and even the phase shift of C_{AA} and C_{AT} (that we observed in *real* DNA) are reproduced. This provides a strong indication that the observed oscillations of 10-11 bp are mainly due to correlations

in protein sequences.

15.6 Discussion

We demonstrated that an analysis of specific correlation functions of biosequences allows the interpretation of several observed correlations in terms of well-known biological structures. A careful detection of specific correlation patterns provides also a basis for the detection of structural elements. For example, a quantification of the period three by the mutual information function allows to distinguish coding from noncoding regions [Grosse et al. 1999]. Correlations in protein sequences can be exploited to predict the structural class of protein [Weiss & Herzel 1997].

A central result of this chapter is the explanation of the 10-11 base-pair periodicities in yeast and bacterial DNA. The close relation of the middle nucleotide of the codon to the biochemical properties hydrophobicity and polarity induces a direct correspondence between hydrophobicity oscillations of protein sequences and correlations of protein coding DNA sequences.

Interestingly, it has been found that nucleosomal patterns are quite similar to the protein-induced correlation patterns [Ioshikhes et al. 1996, Denisov et al. 1997]: there is a preference of AA and TT pairs in a distance of about 10.3 bp and a phase shift of 5-6 nucleotides between A and T to assist bending around nucleosomes. This implies that nucleosomes can use nucleotide patterns induced by the encoded proteins for their positioning. This coincidence might reflect a coevolution of packaging DNA into chromatin and the amino acid and codon usage.

So far we have not found significant periodicities of 10-11 bases in introns. This would be an indication against the nucleosome formation in the intron regions of genomic DNA. However, since introns constitute a substantial proportion of the genomic DNA, and most of the genomic DNA is packed in chromatin [Holde 1988], the nucleosomes must also form in the intron regions. Experimental work on the reconstruction of chromatin on cDNA has demonstrated that the cDNA fails to fold into a unique organized chromatin structure [Liu et al. 1995]. There is the following explanation of this apparent contradiction between the presence of the nucleosomes on the intronic sequences and the absence of a corresponding sequence pattern: the nucleosome sequence pattern is very weak. Only 3 - 5 dinucleotides AA and TT, properly distributed along the sequence, are sufficient

to direct the nucleosome to its position [Bolshoy et al. 1996]. These signal dinucleotides may occupy any 3 - 5 positions of 12 available preferred positions for AA and 12 for TT [Ioshikhes et al. 1996]. Actually, such a weak and degenerate pattern is not infrequent even in random sequences. This means that even without specific sequence biases intronic regions of DNA may (and do) contain many sites with sufficient nucleosome specificities. An analysis of the sequence patterns in a large database of nucleosome sequences experimentally mapped in the introns may eventually clarify this point.

The similarity between Figure 2 from yeast DNA and Figure 5 from protein sequences is a strong indication that the amino acid correlations are the dominating origin of the 10-11 bp periodicities in yeast. The oscillations in Figure 4 from prokaryotes exhibit somewhat different features: They are stronger, persist over more than 100 bp, and have a slightly larger period of about 11 bp. Consequently, additional sources of these periodicities have to be taken into account.

One attractive explanation could be that the period 11 reflects folding of the naturally supercoiled prokaryotic DNA. The inter-wound right-handed super-helices in bacteria have a normal super-helical density of 0.04 - 0.05 which corresponds to 200 - 250 base pairs per one super-helical turn [Vologodskii 1992]. In order to stabilize and/or synchronize this writhe, the prokaryotic DNA molecules may contain regions of intrinsically curved and twisted DNA with corresponding periodic biases. The right-handed super-helix would require a sequence period higher than the helical repeat of free DNA, very much like the left-handed super-helix of the nucleosomal DNA requires a period lower than the free DNA helical repeat [Ulanovsky & Trifonov 1983]. The actual excess would depend on the particular geometry of the super-helix, which, however, is largely unknown.

While we restricted ourselves in this chapter to the analysis and explanation of known correlation features, it is our hope that this conceptual approach might help to find yet unknown patterns in DNA and protein sequences.

Part E

Chapter 16

Species Independence of Mutual Information in Coding and Noncoding DNA

We explore if there exist universal statistical patterns that are different in coding and noncoding DNA and can be found in all living organisms, regardless of their phylogenetic origin. We find that (i) the *mutual information function* $\mathcal{I}$ has a significantly different functional form in coding and noncoding DNA. We further find that (ii) the probability distributions of the *average mutual information* $\overline{\mathcal{I}}$ are significantly different in coding and noncoding DNA, while (iii) they are almost the same for organisms of all taxonomic classes. Surprisingly, we find that $\overline{\mathcal{I}}$ is capable of predicting coding regions as accurately as organism-specific coding measures.

16.1 Introduction

DNA carries the genetic information of most living organisms, and the goal of genome projects is to uncover that genetic information. Hence, genomes of many different species, ranging from simple bacteria to complex vertebrates, are currently being sequenced. As automated sequencing techniques have started to produce a rapidly growing amount of raw DNA sequences, the extraction of information from these sequences becomes a scientific challenge. A large fraction of an organism's DNA is not used for encoding proteins [Lewin 1997, Lodish et al. 1995, Alberts et al. 1994]. Hence, one basic task in the analysis

of DNA sequences is the identification of coding regions. Since biochemical techniques alone are not sufficient for identifying all coding regions in every genome, researchers from many fields have been attempting to find statistical patterns that are different in coding and noncoding DNA [Fickett 1982, Staden 1982, Guigó et al. 1992, Fickett & Tung 1992, Fickett 1996, Burset & Guigó 1996, Claverie 1997]. Such patterns have been found, but none seems to be species independent. Hence, traditional coding measures [Comment 1] based on these patterns need to be trained on organism-specific data sets before they can be applied to identify coding DNA. This training-set dependence limits the applicability of traditional coding measures, as many new genomes are currently being sequenced for which training sets do not exist.

16.2 Mutual Information Function

In search for *species-independent* statistical patterns that are different in coding and noncoding DNA, we study the *mutual information function* $\mathcal{I}(k)$, which quantifies the amount of information (in units of bits) that can be obtained from one nucleotide X about another nucleotide Y that is located k nucleotides downstream from X [Comment 2]. Within the framework of statistical mechanics $\mathcal{I}$ can be interpreted as follows. Consider a compound system (X, Y) consisting of the two subsystems X and Y. Let p_i denote the probability of finding system X in state i, let q_j denote the probability of finding system Y in state j, and let P_{ij} denote the joint probability of finding the compound system (X, Y) in state (i, j). Then the entropies of the systems X, Y, and (X, Y) are defined by

$$\begin{aligned} \mathcal{H}[X] &\equiv -k_B \sum_i p_i \ln p_i, \\ \mathcal{H}[Y] &\equiv -k_B \sum_j q_j \ln q_j, \text{ and} \\ \mathcal{H}[X,Y] &\equiv -k_B \sum_{i,j} P_{ij} \ln P_{ij}, \end{aligned}$$

where k_B denotes the Boltzmann constant. If X and Y are *statistically independent*, then $\mathcal{H}[X] + \mathcal{H}[Y] = \mathcal{H}[X, Y]$, since the Boltzmann entropy is extensive. If X and Y are *statistically dependent*, then the sum of the entropies of the subsystems X and Y is strictly greater than the entropy of the compound system (X, Y), i. e. $\mathcal{H}[X] + \mathcal{H}[Y] > \mathcal{H}[X, Y]$. The *mutual information* $\mathcal{I}[X, Y]$ is defined as the difference of the sum of the entropies of

the subsystems and the entropy of the compound system,

$$\mathcal{I}[X,Y] \equiv \mathcal{H}[X]+\mathcal{H}[Y]-\mathcal{H}[X,Y].$$

If k_B is replaced by $1/\ln 2$, then $\mathcal{I}[X,Y]$ quantifies the amount of information in X about Y in units of bits [Shannon 1948].

The following two examples may serve to illustrate the intuitive (and information theoretic) meaning of $I(k)$. Consider a random, uncorrelated sequence, in which each nucleotide occurs independently of any other nucleotide in the sequence. Intuitively it is clear that we cannot obtain any information from any nucleotide X about any nucleotide Y, so $I(k)$ should be zero for all distances k. Indeed $I(k)=0$ for all k according to Eq. (16.1), since the statement that all nucleotides are statistically independent can be mathematically formulated by the set of equalities: $P_{ij}(k)=p_j\cdot p_j$ for all i, j, and k. From these equalities it follows that all the logarithms appearing in Eq. (16.1) are zero, and hence the sum in Eq. (16.1) is equal to zero.

As a second example consider a sequence in which each nucleotide occurring with equal probability $p_i=1/4$ is determined by the previous nucleotide. In this case we will be able to *determine* the identity of nucleotide Y by learning the identity of X. Intuitively we say we obtain an information of 2 bits about Y by learning the identity of X. Indeed $I(k)=2$ by Eq. (16.1), so again Eq. (16.1) agrees with our intuition. For quaternary sequences $I(k)$ always ranges from 0 to 2, and for most DNA sequences $I(k)$ is close to 0, which states that in a typical DNA sequence the information in nucleotide X about nucleotide Y is small. If $I(k)$ is monotonically decreasing with k, it means that the information in nucleotide X about nucleotide Y gets smaller as the distance k between X and Y increases.

Two obvious but noteworthy properties of $\mathcal{I}[X,Y]$ are (i) $\mathcal{I}[X,Y]=\mathcal{I}[Y,X]$, so the amount of information in X about Y is equal to the amount of information in Y about X, and (ii) $\mathcal{I}[X,Y]\geq 0$, so the amount of information is always non-negative, and it is equal to zero if and only if X and Y are statistically independent. We choose $P_{ij}(k)$ to denote the joint probability of finding the pair of nucleotides n_i and n_j $(n_i,n_j\in\{A,C,G,T\})$ spaced by a gap of $k-1$ nucleotides, and we define $p_i\equiv\sum_j P_{ij}(k)$ and $q_j\equiv\sum_i P_{ij}(k)$. Then

$$\mathcal{I}(k)\equiv\sum_{i,j=1}^{4} P_{ij}(k)\log_2\frac{P_{ij}(k)}{p_i q_j} \tag{16.1}$$

quantifies the degree of statistical dependence between the nucleotides X and Y spaced by a gap of $k-1$ nucleotides, and we study $\mathcal{I}$ as a function of k for coding and noncoding

DNA of all eukaryotic organisms available in GenBank release 111 [Comment 3].

Figure 1 shows $\mathcal{I}(k)$ for coding and noncoding human DNA. We find that for noncoding DNA $\mathcal{I}(k)$ decays to zero, whereas for coding DNA $\mathcal{I}(k)$ oscillates between two values, the *in-frame* mutual information $\mathcal{I}_{\rm in}$ at distances k that are multiples of 3 and the *out-of-frame* mutual information $\mathcal{I}_{\rm out}$ at all other values of k.

16.3 Average Mutual Information

The oscillatory behavior of $\mathcal{I}(k)$ in coding DNA is a consequence of the presence of the genetic code (which maps non-overlapping nucleotide triplets (codons) to amino acids) and the non-uniformity of the codon frequency distribution. The fact that the codon frequencies are non-uniformly distributed in almost all genes of all organisms is well known to biologists, and arises because (i) the frequency distribution of amino acids is non-uniform, (ii) the number of synonymous codons [Comment 4] that encode one amino acid varies from 1 to 6, and (iii) the frequency distribution of synonymous codons is non-uniform [Ikemura 1981]. A simple model that reflects the non-uniformity of the codon frequency distribution, but neglects any other correlation, is the *pseudo-exon model* [Herzel et al. 1995], which concatenates codons randomly chosen from a given probability distribution $(Q_{AAA}, ..., O_{TTT})$, where Q_{XYZ} denotes the probability of codon XYZ $(X, Y, Z \in \{A, C, G, T\})$. As the pseudo-exon model has been shown to reproduce the period-3 oscillations in genomic DNA [Herzel et al. 1995], we use the model assumption of neglecting weak correlations between codons in order to express the joint probabilities $P_{ij}(k)$ in terms of the 12 *positional nucleotide probabilities* $p_i^{(m)}$ [Comment 5] of finding nucleotide n_i at position $m \in \{1, 2, 3\}$ in an arbitrarily chosen reading frame [Comment 6] as follows [Staden 1982, Herzel et al. 1995]:

$$P_{ij}(k) = \frac{1}{3} \cdot \begin{cases} p_i^{(1)} p_j^{(1)} + p_i^{(2)} p_j^{(2)} + p_i^{(3)} p_j^{(3)} & \text{for} \quad k = 3, 6, 9, ... \\ p_i^{(1)} p_j^{(2)} + p_i^{(2)} p_j^{(3)} + p_i^{(3)} p_j^{(1)} & \text{for} \quad k = 4, 7, 10, ... \\ p_i^{(1)} p_j^{(3)} + p_i^{(2)} p_j^{(1)} + p_i^{(3)} p_j^{(2)} & \text{for} \quad k = 5, 8, 11, ... \end{cases} . \qquad (16.2)$$

It is clear that $P_{ij}(k)$ is invariant under shifts of the reading frame, because the expressions on the r. h. s. of Eq. (16.2) are invariant under cyclic permutations of the upper indices $(1, 2, 3)$. Since the second and third line on the r. h. s. of Eq. (16.2) are identical after transposition of the lower indices (i, j), we obtain $P_{ij}(k = 4, 7, 10, ...) = P_{ji}(k = 5, 8, 11, ...)$, which implies that $\mathcal{I}(k)$ computed from $P_{ij}(k)$ of Eq. (16.2) will assume only two different

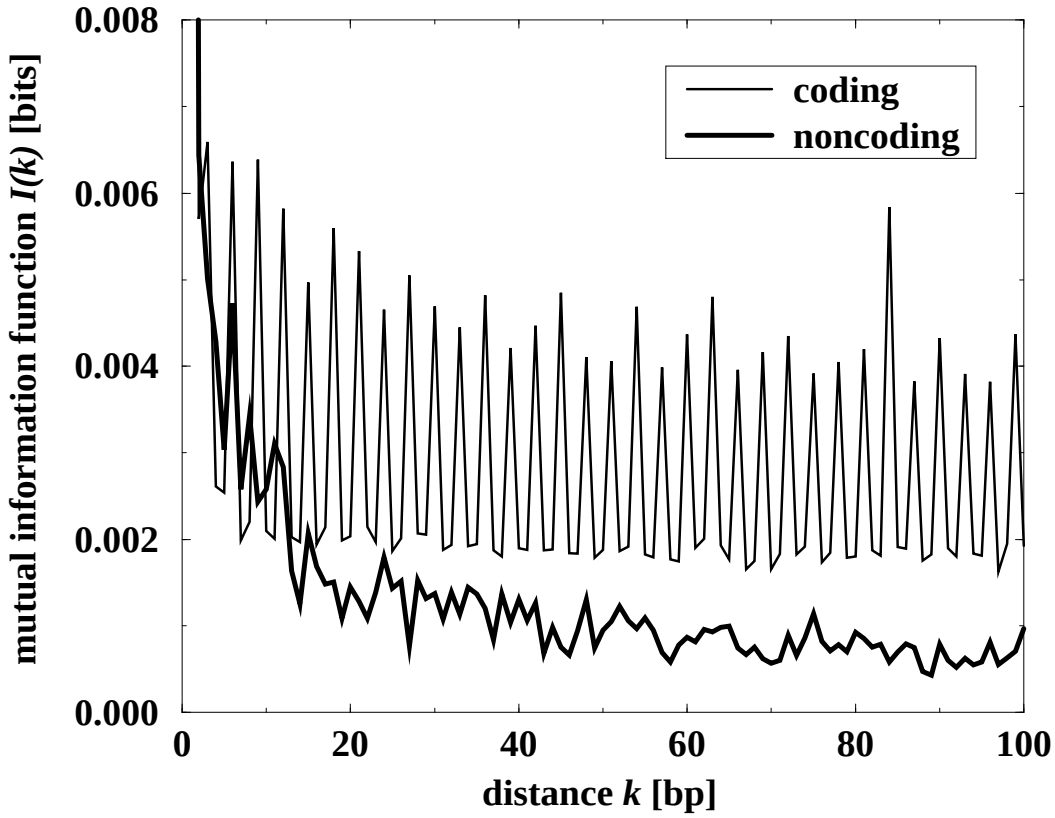


Figure 16.1: Mutual information function, $\mathcal{I}(k)$, of coding (thin line) and noncoding (thick line) human DNA, from GenBank release 111 [Comment 3]. We cut all human, non-mitochondrial DNA sequences into non-overlapping fragments of length 500 bp, starting at the 5'-end. We compute the mutual information function of each fragment, correct for the finite length effect [Herzel et al. 1995], and display the average over all mutual information functions (of coding and noncoding DNA separately). While $\mathcal{I}(k)$ for noncoding DNA monotonically decays to zero as k increases, $\mathcal{I}(k)$ of coding DNA shows persistent period-3 oscillations.

values, $\mathcal{I}_{\rm in} = \mathcal{I}(3,6,9,...)$ and $\mathcal{I}_{\rm in} = \mathcal{I}(4,5,7,8,10,11,...)$.

In order to construct a coding measure that can predict whether a single sequence is coding or noncoding, we sample from each sequence the 12 frequencies $p_i^{(m)}$, compute $P_{ij}(k)$ from $p_i^{(m)}$ by using Eq. (16.2), and then compute

$$\mathcal{I}_{\rm in} \equiv \mathcal{I}(3) \quad \text{and} \quad \mathcal{I}_{\rm out} \equiv \mathcal{I}(4) = \mathcal{I}(5) \tag{16.3}$$

by plugging $P_{ij}(k)$ and $p_i = (p_i^{(1)} + p_i^{(2)} + p_i^{(3)})/3$ into Eq. (16.1).

We find that $\ln(I_{\rm in})$ and $\ln(I_{\rm out})$ are almost linearly dependent, and are thus highly correlated (correlation coefficient $C = 0.96$ for both coding and noncoding DNA). This simplifies the question of how to combine $I_{\rm in}$ and $I_{\rm out}$ into a single quantity, as almost any combination will yield approximately the same accuracy. For the sake of obtaining a simple coding measure with a natural and intuitive interpretation, we compute from $\mathcal{I}_{\rm in}$ and $\mathcal{I}_{\rm out}$ the *average mutual information*

$$\overline{\mathcal{I}} \equiv \mathcal{P}_{\rm in} \cdot \mathcal{I}_{\rm in} + \mathcal{P}_{\rm out} \cdot \mathcal{I}_{\rm out}, \tag{16.4}$$

where $\mathcal{P}_{\rm in} = 1/3$ and $\mathcal{P}_{\rm out} = 2/3$ denote the occurrence probabilities of $\mathcal{I}_{\rm in}$ and $\mathcal{I}_{\rm out}$. $\overline{\mathcal{I}}$ quantifies the *average* amount of information one obtains about a nucleotide X by learning both the identity of any other nucleotide Y in the same DNA sequence and whether the distance k between X and Y is a multiple of 3. We expect that due to the presence of the genetic code $\overline{\mathcal{I}}$ will be typically greater in coding than in noncoding DNA.

The practical implementation of the algorithm looks as follows:

1. Count the number of occurrences of nucleotide $n_i \in \{A, C, G, T\}$ in position $m \in \{1, 2, 3\}$ of an arbitrarily chosen reading frame in a given DNA sequence of length[1] N. Denote that number by $N_i^{(m)}$.
2. Divide $N_i^{(m)}$ by $N/3$, the total number of nucleotides occurring in position m, and define the *positional nucleotide frequency* $p_i^{(m)} \equiv 3 \cdot N_i^{(m)}/N$. Note that the positional nucleotide frequencies are normalized to 1, that is $\sum_{i=1}^{4} p_i^{(m)} = 1$ for all m.
3. Compute $P_{ij}(3)$ and $P_{ij}(4)$ from $p_i^{(m)}$ by using Eq. (16.2).
4. Define $p_i \equiv \sum_{m=1}^{3} p_i^{(m)}/3$, which is the overall, normalized frequency of nucleotide n_i.

[1]For the sake of simplicity, assume N be a multiple of 3.

5. Compute $I(3)$ and $I(4)$ from $P_{ij}(3)$, $P_{ij}(4)$, and p_i by using Eq. (16.1). Define $I_{\rm in} \equiv I(3)$ and $I_{\rm out} \equiv I(4)$ as well as $P_{\rm in} \equiv 1/3$ and $P_{\rm out} \equiv 2/3$.
6. Compute the average mutual information (AMI) by using Eq. (16.4).

The source code is available upon request from `ivo@bu.edu`.

16.4 Accuracy of the Average Mutual Information

First, we investigate how accurately $\overline{\mathcal{I}}$ can distinguish coding from noncoding DNA. In order to compare the accuracy by which $\overline{\mathcal{I}}$ can distinguish coding from noncoding DNA with the accuracy of traditional coding measures, we use the standard benchmark test and data sets of Fickett and Tung [Fickett & Tung 1992]. Figure 2 shows the $\overline{\mathcal{I}}$-histograms for coding and noncoding human DNA sequences of length 108 bp from the data sets of Fickett and Tung [Fickett & Tung 1992]. Since $\overline{\mathcal{I}}$ does not require prior training, we show the $\overline{\mathcal{I}}$-histograms for both the training and the test set. We find that for both data sets the AMI distributions are significantly different for coding and noncoding DNA.

The accuracy $\mathcal{A}$ is defined as follows: Denote by $\rho_c(\overline{\mathcal{I}})$ and $\rho_n(\overline{\mathcal{I}})$ the probability density functions of $\overline{\mathcal{I}}$ for coding and noncoding DNA (see Figure 16.4). Define the overlap integral $\mathcal{O}(\overline{\mathcal{I}}) \equiv \int \mathcal{M}(\overline{\mathcal{I}}) d\overline{\mathcal{I}}$, where $\mathcal{M}(\overline{\mathcal{I}})$ denotes the maximum of the two values $\rho_c(\overline{\mathcal{I}})$ and $\rho_n(\overline{\mathcal{I}})$ at position $\overline{\mathcal{I}}$. In statistical terms, $\mathcal{O}(\overline{\mathcal{I}})$ can be expressed as the sum of T_p and T_n, $\mathcal{O}(\overline{\mathcal{I}}) = T_p + T_n$, where T_p (T_n) denotes the fraction of true positives (true negatives) over all positives (all negatives) [Comment 8]. Hence, the accuracy, defined by $\mathcal{A}(\overline{\mathcal{I}}) \equiv \mathcal{O}(\overline{\mathcal{I}})/2$, ranges from from 1/2 (no discrimination) to 1 (perfect discrimination) [Comment 9].

Table 1 shows the accuracy of the top 8 phase-independent coding measures as ranked in Fickett and Tung [Fickett & Tung 1992] and the accuracy of $\overline{\mathcal{I}}$ computed on exactly the same data sets.

We find that the AMI is as accurate as many of the traditional coding measures, which are trained on organism-specific data sets [Fickett & Tung 1992], in contrast to the AMI, which does not require prior training.

We find that the accuracy of $\overline{\mathcal{I}}$ ($\mathcal{A}(\overline{\mathcal{I}}) = 0.69, 0.76, 0.81$ for human DNA sequences of lengths $N = 54, 108, 162$ bp) is higher than the accuracy of many of the 21 traditional coding measures from Ref. [Fickett & Tung 1992]. In particular, $\mathcal{A}(\overline{\mathcal{I}})$ is comparable to the accuracy of the hexamer measure H, ($\mathcal{A}(H) = 0.70, 0.73, 0.74$), which is the most accurate of the 21 frame-independent [Comment 6] coding measures from Ref. [Fickett & Tung 1992].

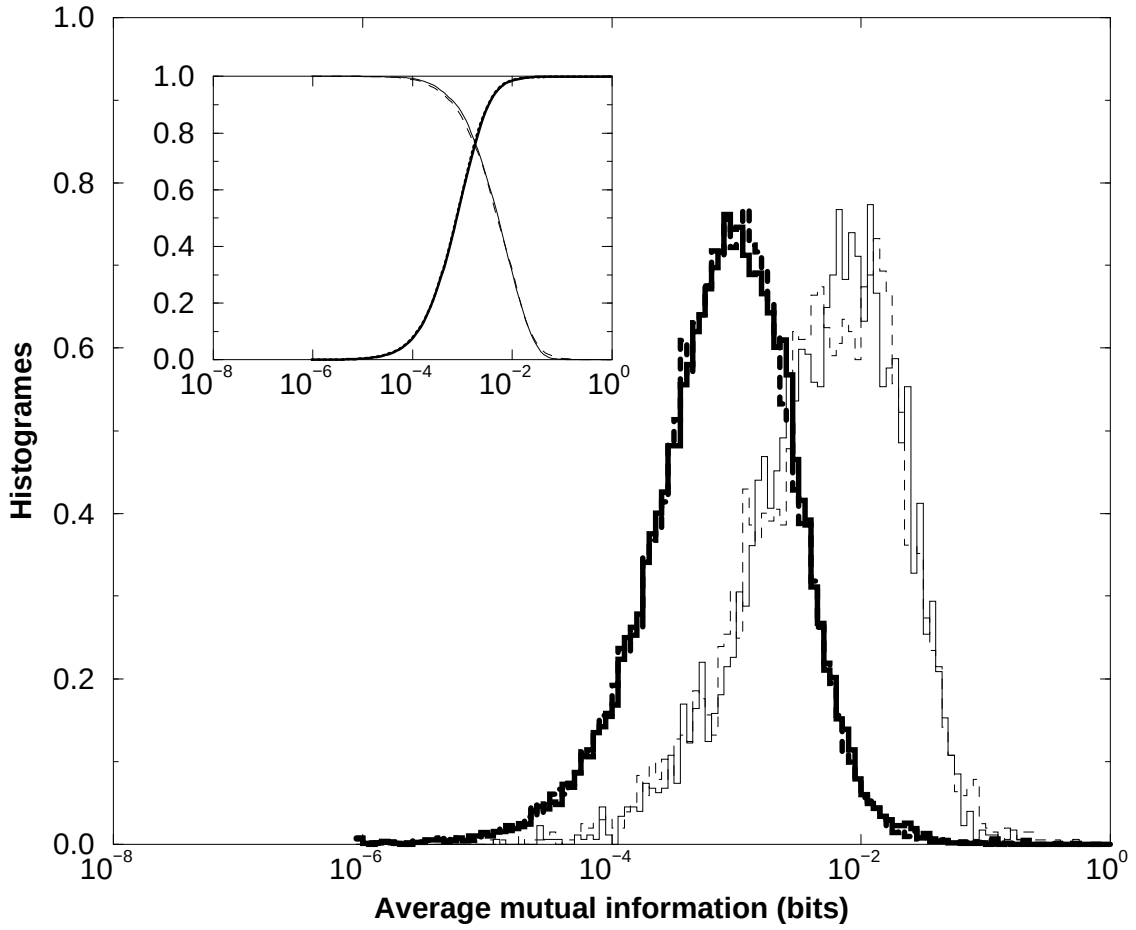


Figure 16.2: $\overline{\mathcal{I}}$-distributions of data sets `humg108a` (solid lines) and `humg108b` (dashed lines) of Fickett and Tung [Fickett & Tung 1992] for coding DNA (thin lines) and noncoding DNA (thick lines). In both data sets the $\overline{\mathcal{I}}$-distribution of noncoding DNA is centered at significantly smaller values than the $\overline{\mathcal{I}}$-distribution of coding DNA. The cumulative distribution functions of $\overline{\mathcal{I}}$ presented in the inset show that $\overline{\mathcal{I}}$ allows a discrimination of coding and noncoding DNA with an accuracy of approximately 76%.

	Coding Measure	54 bp	108 bp	162 bp
1.	Hexamer	70.5%	73.1%	74.2%
2.	Position Asymmetry	70.2%	76.6%	80.6%
3.	Dicodon Usage	70.2%	72.9%	73.9%
4.	Fourier	69.9%	76.5%	80.8%
5.	Hexamer-1	69.9%	72.6%	73.8%
6.	Hexamer-2	69.9%	72.6%	73.8%
7.	Run	66.6%	70.3%	71.3%
8.	Codon Usage	65.2%	68.0%	69.5%
9.	AMI	69.2%	76.1%	80.7%

Table 16.1: Accuracy of 8 coding measures and the AMI. We compare the accuracies of the best 8 phase-independent coding measures as evaluated by Fickett and Tung [Fickett & Tung 1992] to the accuracy of the AMI for three sets of coding and noncoding human DNA sequences of lengths 54 bp, 108 bp, and 162 bp. We find that, on all three length scales, the accuracy of the AMI (without prior training) is comparable to the accuracy of traditional coding measures (after prior training).

This finding is interesting, because H (like all other 20 traditional coding measures) is trained on species-specific data sets, and $\overline{\mathcal{I}}$ is not. If the $\overline{\mathcal{I}}$-distributions turn out to be species independent, then $\overline{\mathcal{I}}$ could be used without prior training to distinguish coding from noncoding DNA in all species, regardless of their taxonomic origin.

16.5 Species Independence of the Average Mutual Information

After having found that—without prior training—the AMI can distinguish coding from noncoding DNA as accurately as traditional coding measures, the question arises if the probability distribution functions of the AMI are species-independent. Figure 2 shows the $\overline{\mathcal{I}}$-distributions for coding and noncoding DNA sequences from species of different taxonomic orders, phyla, and kingdoms. We find that the $\overline{\mathcal{I}}$-distributions are significantly different for coding and noncoding DNA, while they are almost identical for all taxonomic sets. In order to supplement this qualitative finding by a quantitative analysis, we present in Table 1 the means and variances of $\log_{10}\overline{\mathcal{I}}$. Table 1 shows that the means are significantly

Table 16.2: Means (variances) of $\log_{10}\overline{\mathcal{I}}$ for coding and noncoding DNA of 6 taxonomic sets. While the means of $\log_{10}\overline{\mathcal{I}}$ are significantly different in coding and noncoding DNA, they are almost the same for all taxonomic sets. Also the variances of $\log_{10}\overline{\mathcal{I}}$ are almost the same for all taxonomic sets, supplementing the visual finding from Figure 2 that the $\overline{\mathcal{I}}$-distributions are nearly species-independent.

	noncoding	coding
primates	-2.52 (0.31)	-2.04 (0.30)
non-primate vertebrates	-2.54 (0.39)	-2.06 (0.30)
vertebrates	-2.53 (0.34)	-2.05 (0.30)
invertebrates	-2.50 (0.33)	-2.04 (0.32)
animals	-2.52 (0.34)	-2.05 (0.31)
plants	-2.48 (0.31)	-2.09 (0.31)

different for coding and noncoding DNA, and that the means and variances are almost the same for all species. This finding is in agreement with the visual finding based on Figure 2 that the $\mathcal{I}$-distributions are species independent and significantly different in coding and noncoding DNA.

16.6 Quantification of the Species Independence

In order to quantitatively compare the "species independence" of the AMI to the "species independence" of the codon usage, we introduce a quantity that we call the *degree of species dependence* (DSD). Define x_i and y_i $(i = 1, \ldots, M)$ to be the usage frequencies of the $M = 64$ codons for two non-overlapping sets of 1024 DNA sequences of length 108 bp. Denote by

$$\chi^2(X, Y) \equiv \sum_{i=1}^{M} \frac{(x_i - y_i)^2}{x_i + y_i} - (M + 1) \tag{16.5}$$

the normalized "distance" between two histograms $X \equiv (x_1, \ldots, x_M)$ and $Y \equiv (y_1, \ldots, y_M)$. Let A_c, A_n, B_c, and B_n denote the four possible histograms for coding and noncoding DNA from the taxonomic groups A and B. We define the DSD to be the ratio of the average distance between species and the average distance between coding and noncoding DNA,

$$\mathrm{DSD} \equiv \frac{\chi^2(A_c, B_c) + \chi^2(A_n, B_n)}{\chi^2(A_c, A_n) + \chi^2(B_c, B_n)} \ . \tag{16.6}$$

We analyze the degree of species dependence of the codon usage on four taxonomic

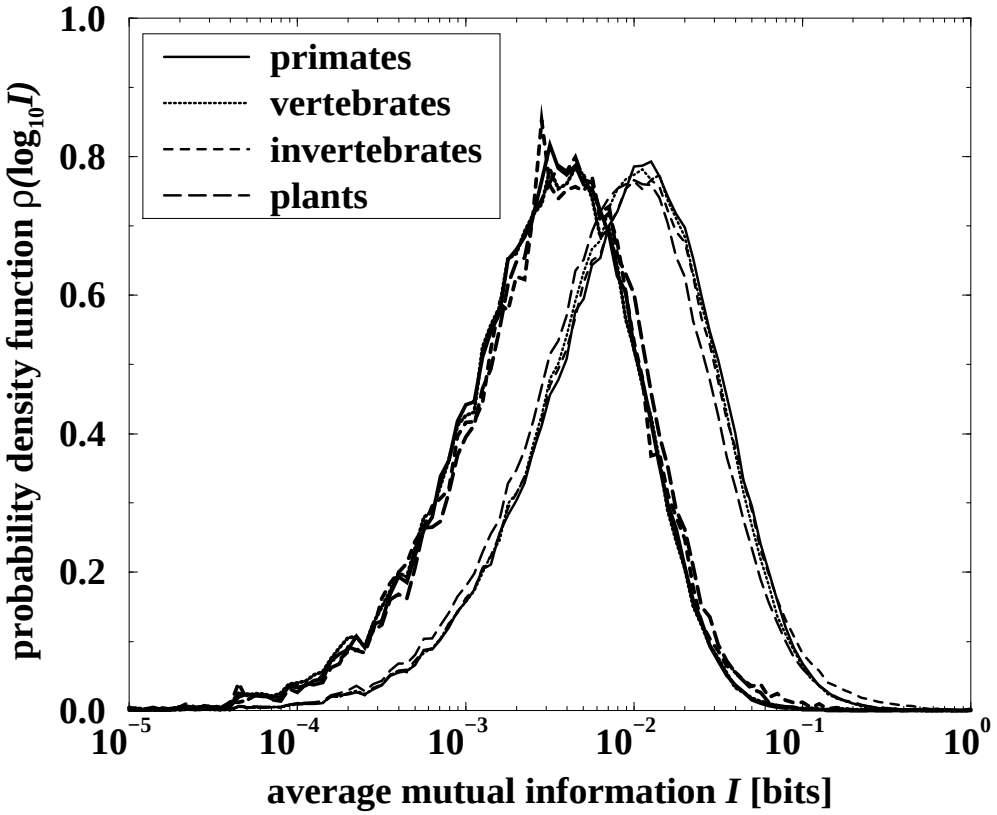


Figure 16.3: $\overline{\mathcal{I}}$-distributions of coding DNA (thin lines) and noncoding DNA (thick lines) from all eukaryotic DNA sequences in GenBank release 111 [Comment 3]. We cut all sequences into non-overlapping fragments of length 54 bp [Comment 7], starting at the 5'-end. We compute $\overline{\mathcal{I}}$ of each DNA fragment and show the $\overline{\mathcal{I}}$-histograms for coding and noncoding DNA, for each of the 4 disjoint taxonomic sets (primates, non-primate vertebrates, invertebrates, plants) separately. We find that (i) for all taxonomic sets $\rho_n(\overline{\mathcal{I}})$ is centered at significantly smaller values than $\rho_c(\overline{\mathcal{I}})$, while (ii) $\rho_c(\overline{\mathcal{I}})$ and $\rho_n(\overline{\mathcal{I}})$ of different taxonomic sets are almost identical. The close similarity of the $\overline{\mathcal{I}}$-distributions for different taxonomic classes, phyla, and kingdoms illustrates the species independence of $\rho_c(\overline{\mathcal{I}})$ and $\rho_n(\overline{\mathcal{I}})$.

levels by comparing primates with non-primate mammals, mammals with non-mammalian vertebrates, vertebrates with invertebrates, and animals with plants. We randomly partition the set of all GenBank-111 sequences into non-overlapping blocks of 1024 sequences, and compare all possible combinations of these blocks. Table 2 shows the average DSD over these combinations.

Table 16.3: The degree of species dependence of the codon usage and the AMI. Column 1 displays the DSD of the codon usage; the value of 0.01 in row 1 states that the codon usage differences between primates and non-primate mammals are only 1% of the differences between coding and noncoding DNA. When DNA is analyzed from species belonging to different taxonomic classes, phyla, or kingdoms (rows 2, 3, and 4), the DSD becomes larger, which quantifies the well-known fact that the codon usage is strongly species dependent. Columns 2 displays the degree of species dependence of the AMI, which we compute in the same way (and for the same sets of sequences) as for the codon usage. The degree of species dependence of the AMI never exceeds 0.02, quantifying the finding from Figure 3 that the AMI distributions are species independent.

class	of	organism	codon usage	AMI
primates	–	non-primate mammals	0.01	0.01
mammals	–	non-mammalian vertebrates	0.10	0.01
vertebrates	–	invertebrates	0.69	0.01
animals	–	plants	0.58	0.02

Column 1 of Table 2 shows that the degree of species dependence of the codon usage is quite small (0.01) when primates are compared to non-primate mammals. This states that the codon usage is not identical in primates and non-primate mammals, but it is so similar that the codon usage differences between primates and non-primate mammals is about 100 times smaller than the differences between exons and introns. When we compare vertebrates to invertebrates, the degree of species dependence increases to about 0.69, which states that the differences between species are approximately 2/3 as large as the differences between exons and introns. The data from column 1 are consistent with the well-known fact that the codon usage is species dependent [Fiers & Grosjean 1979, Ikemura 1981, Sharp & Li 1987, Bulmer 1987, Bernardi 1989, Nakamura et al. 1996, Karlin & Mrazek 1997].

Next, we analyze the degree of species dependence of the AMI by discretizing the continuous AMI distributions as follows: when comparing two AMI distributions X and Y (see Figure 3), we map the AMI values into $M = 64$ bins in such a way that each bin $i \in \{1, \ldots, M\}$ contains the same number of data points $x_i + y_i$. We then compute the DSD of these discretized AMI distributions X and Y for the same blocks of 1024 sequences of length 108 bp as we used to calculate the DSD of the codon usage distributions.

We find (column 2 of Table 2) that the AMI differences between primates and non-primate mammals are about 100 times smaller than the AMI differences between exons and introns. It is surprising that the degree of species dependence remains of the order of 0.01 when mammals are compared to non-mammalian vertebrates, or when vertebrates are compared to invertebrates. Even when DNA from animals is compared to DNA from plants, the AMI yields a degree of species dependence of only 0.02. The data from column 2 are in agreement with the observation, based on Figure 2 and Figure 3, that the AMI distributions are species independent. This species independence, in connection with the finding that the accuracy of the AMI is comparable to the accuracy of traditional coding measures, suggests that the AMI might possibly be useful for the recognition of protein-coding regions in genomes for which training sets do not exist.

In search for a possible origin of the observed species independence, we attempt to develop simple models that are able to reproduce the $\overline{\mathcal{I}}$-distributions for coding and non-coding DNA.

16.7 Understanding the Species Dependence for Noncoding DNA

We first present a model that reproduces the $\overline{\mathcal{I}}$-distributions for noncoding DNA. For a random, uncorrelated sequence of arbitrary composition $(p_1, p_2, ..., p_4)$, we can derive the asymptotic form of the probability density function $\rho(\overline{\mathcal{I}})$ as follows: expand $\mathcal{I}(k)$ about $P_{ij}(k) - p_i p_j$, and truncate the Taylor series after the quadratic term. The constant term vanishes because $\mathcal{I}(k) = 0$ at $P_{ij}(k) = p_i p_j$, and the linear terms vanish because $\mathcal{I}(k)$ achieves its minimum at $P_{ij}(k) = p_i p_j$. Hence, we obtain

$$\mathcal{I}(k) \propto \frac{1}{\ln 2} \sum_{i,j} \frac{(P_{ij}(k) - p_i p_j)^2}{2 p_i p_j}, \tag{16.7}$$

where the symbol $\propto$ indicates that we neglect terms of $O\left((P_{ij}-p_ip_j)^3\right)$. Substituting $P_{ij}(k)$ (for $k=3,4,5$) by the expressions on the r. h. s. of Eq. (16.2) and expressing $\overline{\mathcal{I}} \equiv (\mathcal{I}(3)+\mathcal{I}(4)+\mathcal{I}(5))/3$ in terms of $p_i^{(m)}$ yields

$$\overline{\mathcal{I}} \propto \frac{1}{\ln 2}\left(\sum_{i,m}\frac{(p_i^{(m)}-p_i)^2}{2p_i}\right)^2. \tag{16.8}$$

For a random, uncorrelated sequence the probability density function of $N\sum_{i,m}(p_i^{(m)}-p_i)^2/p_i$ converges, for asymptotically large sequence length N, to a χ^2-distribution with 6 degrees of freedom [Cramer 1946]. Hence, we obtain that $\rho(\overline{\mathcal{I}})$ converges, for asymptotically large N, to

$$\rho(\overline{\mathcal{I}}) = \frac{(N\sqrt{\ln 2})^3}{4}\cdot\sqrt{\overline{\mathcal{I}}}\cdot e^{-N\sqrt{\ln 2}\sqrt{\overline{\mathcal{I}}}}. \tag{16.9}$$

Figure 3(a) shows $\rho(\overline{\mathcal{I}})$ from Eq. (16.9) and the $\overline{\mathcal{I}}$-histograms for noncoding human DNA for $N=54$ bp, 108 bp, and 162 bp. We find that (i) the $\overline{\mathcal{I}}$-distributions for noncoding DNA collapse after rescaling with a factor of N^2, and that (ii) the $\overline{\mathcal{I}}$-distributions can be approximated by Eq. (16.9). The agreement of the theoretical with the experimental $\overline{\mathcal{I}}$-distributions states that the species independence of the $\overline{\mathcal{I}}$-distributions for noncoding DNA may be attributed to the absence of the genetic code in noncoding DNA of all living species.

16.8 Understanding the Species Dependence for Coding DNA

We now test if the species-independence of the $\overline{\mathcal{I}}$-distributions for coding DNA may be reproduced by a simple model that incorporates the presence of a reading frame. We generate a random, uncorrelated sequence where the probability of obtaining nucleotide n_i at position m is given by $p_i^{(m)}$ [Comment 10]. Figure 3(b) shows the $\overline{\mathcal{I}}$-histograms for the model sequences and for human coding DNA sequences of length $N=54$ bp. We find that the $\overline{\mathcal{I}}$-distribution of the model sequences is significantly different from the $\overline{\mathcal{I}}$-distribution of human coding DNA sequences. We perform the same analyses for different organisms, ranging from simple bacteria to complex vertebrates, as well as for different N, and we find that in all cases the modeled $\overline{\mathcal{I}}$-distributions cannot reproduce the $\overline{\mathcal{I}}$-distributions of experimental, coding DNA. This result shows that the presence of a reading frame in

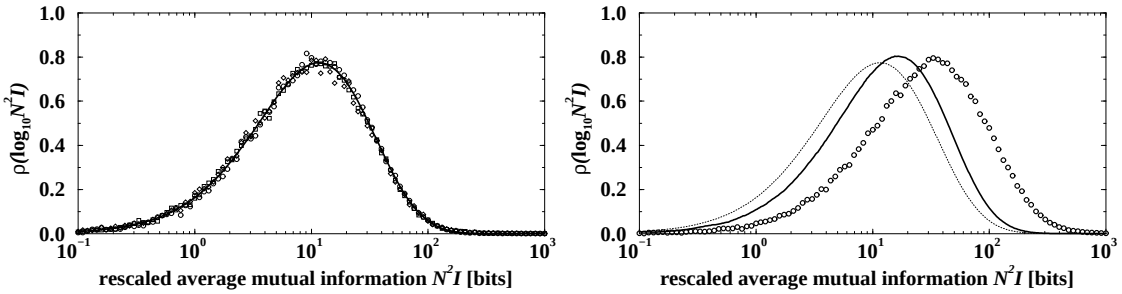


Figure 16.4: Rescaled $\overline{\mathcal{I}}$-distributions of model and experimental, coding and noncoding DNA [Comment 3]. Figure 3(a) shows the histograms of $\log_{10} N^2\overline{\mathcal{I}}$ for noncoding human DNA for $N = 54$ bp ($\circ$), 108 bp ($\Box$), and 162 bp ($\diamond$), and the corresponding χ^2 probability density function with 6 degrees of freedom (thick line). In addition to the observation (Figure 2) that the $\overline{\mathcal{I}}$-distributions are almost identical for different species, we find that (i) the rescaled $\overline{\mathcal{I}}$-distributions collapse for all taxonomic sets and for all N, and that (ii) they agree with the χ^2 probability density function. Hence, the species independence of the $\overline{\mathcal{I}}$-distributions for noncoding DNA may be explained by the absence of a reading frame in noncoding DNA of all species. Figure 3(b) shows the histograms of $\log_{10} N^2\overline{\mathcal{I}}$ for coding human DNA sequences of length $N = 54$ bp ($\circ$), the corresponding non-central χ^2 probability density function (thick line), and the central χ^2 probability density function (thin dotted line). We find that (i) the modeled $\overline{\mathcal{I}}$-distribution (thick line) is indeed shifted to higher $\overline{\mathcal{I}}$-values than the $\overline{\mathcal{I}}$-distribution of noncoding DNA (thin dotted line), but that (ii) the $\overline{\mathcal{I}}$-distribution of the model sequences ($\circ$) is significantly different from the $\overline{\mathcal{I}}$-distribution of coding human DNA. The significant difference between the modeled and the experimental $\overline{\mathcal{I}}$-distribution states that the presence of a reading frame is not sufficient to reproduce the species-independent $\overline{\mathcal{I}}$ distributions for coding DNA (Figure 2).

coding DNA is not sufficient to reproduce the $\overline{\mathcal{I}}$-distributions of experimental, coding DNA, and thus cannot explain the observed species-independence for coding DNA. This finding leads us to the conclusion that there must exist additional correlations or inhomogeneities [Comment 11] in coding DNA, which are responsible for the observed species-independence of the $\overline{\mathcal{I}}$-distributions.

16.9 Conclusions

We reported the finding of a species-independent statistical quantity, the average mutual information $\overline{\mathcal{I}}$, whose probability distribution function is significantly different in coding and noncoding DNA. We showed that $\overline{\mathcal{I}}$ can distinguish coding from noncoding DNA as accurately as traditional coding measures, which all require prior training on species-specific DNA data sets. The capability of $\overline{\mathcal{I}}$ to distinguish coding from noncoding DNA without prior training and irrespective of its phylogenetic origin suggests that $\overline{\mathcal{I}}$ might be useful to identify coding regions in genomes for which training sets do not exist. In an attempt to understand the origin of the observed species-independence of $\overline{\mathcal{I}}$, we found that the species-independence of $\rho_n(\overline{\mathcal{I}})$ may result from the absence of a reading frame in noncoding DNA. We derived analytically the $\overline{\mathcal{I}}$-distribution for an ensemble of random, uncorrelated sequences of arbitrary composition, and we showed that this distribution is consistent with the observed $\overline{\mathcal{I}}$-distribution of noncoding DNA for all species and all sequence lengths N. For coding DNA, we could show that the presence of a reading frame in coding DNA sequences is not sufficient to reproduce the observed $\overline{\mathcal{I}}$-distributions of coding DNA. This finding makes it tempting to conjecture that additional correlations or inhomogeneities are a vital and species-independent ingredient of coding DNA sequences of any living organism.

Chapter 17

Optimization of Coding Measures Using Positional Dependence of Nucleotide Frequencies

In this chapter we study the discrimination accuracy of coding measures based on the positional dependence of nucleotide frequencies, and we analyze the statistical dependences between coding measures and different A+T content. We introduce two generalized coding measures, the position asymmetry D_p and the position information function I_q, and we study how accurately D_p and I_q can distinguish coding from non-coding DNA as a function of the parameters p and q. We determine the parameter values p^* and q^* for which D_p and I_q distinguish coding from non-coding DNA most accurately. We find that p^* and q^* vary only little with the length of the studied DNA sequence. Moreover, we find that D_{p^*} and I_{q^*} yield accuracies comparable to the accuracy of customarily employed coding measures.

17.1 Introduction

Recognition of protein-coding regions in novel sequenced DNA by statistical and information-theoretic means constitutes a challenging problem in computational molecular biology (Fickett 1996; Searls, 1998). This task has received considerable attention in the last decade, as large-scale sequencing projects generate primary un-annotated DNA sequences in an exponentially increasing amount. Genes of higher eukaryotes consist of expressed regions (exons) which are interrupted by intragenic regions (introns). Exons and

introns, respectively, posses distinctive statistical features which can be exploited to discriminate protein-coding versus non-coding DNA. Conventional algorithms for gene-finding integrate heterogeneous types of information, namely the search by content and the search by signal. A third type of information is derived from database similarity searches. Gene search by content signifies for a given sequence the potential to which it is coding. Gene search by signal involves the detection of binding sites and other signals in the surroundings of a gene, such as promoters, splice sites, translation initiation and termination sites, and poly(A)-sites. To predict the most likely gene structure from the primary transcript gene search by content is typically combined with the search by signal using probabilistic (hidden Markov) models of DNA, discriminant analysis, or neural networks. Several pertinent models have been proposed and applied to the gene-identification problem in a computer program. Some examples of such programs include GeneID (Guigó *et al.* 1992), GeneParser (Snyder & Stormo 1993), GENMARK (Borodovsky & McIninch 1993), GenLang (Dong & Searls 1994), FGENEH (Solovyev *et al.* 1994), GRAIL II (Xu *et al.* 1994), MZEF (Zhang 1997); GENSCAN (Burge & Karlin 1997), GeneGenerator (Kleffe *et al.* 1998), GLIMMER (Salzberg *et al.* 1998). The advantages and disadvantages of several gene-identification algorithms have been evaluated recently (Fickett 1996; Burset & Guigó 1996; Claverie 1997) as well as the application of algorithms in conjunction (Murakami & Takagi 1998). Computer-based prediction is nowadays customarily used to provide the initial annotation of genes (Fleischman *et al.* 1995; Nelson *et al.* 1999).

The most prominent statistical feature in exons is the existence of a reading frame with an associated codon usage. The reading frame induces a triplet periodicity in coding sequences, while it is absent in non-coding sequences. In general, a nonuniform codon usage gives rise to a different concentration of each nucleotide in the positions of the reading frame. Possible reasons for the nonuniformity of the codon usage are (i) the nonuniform amino acid composition of proteins, (ii) the unequal number of codons encoding different amino acids, and (iii) the nonuniform distribution of synonymous codons.

We study coding measures based on the positional dependence of nucleotide frequencies. Coding measures correlate with the likelihood that a certain region in DNA is protein-coding and, apart from gene identification, they are applicable to shotgun sequencing experiment (whether non-assembled DNA fragments occur in coding regions) and mRNA/cDNA matching analysis (whether pieces of cDNA from expressed mRNAs occur in the translated segment of mRNA). Evaluated coding measures can be applied to DNA sequences

without prior training. We show that coding measures can be directly derived from the corresponding DNA sequence. The the number of variables used to discriminate coding versus non-coding DNA is 4×3 (nucleotide $\times$ triplet position). On the one hand, this has the advantage to overcome statistical noise, but on the other hand the disadvantage to neglect information not encoded in the codon composition, such as codon-codon interactions. Various coding measures have been developed to quantify the positional dependence of nucleotide frequencies and applied in computer programs. Some of them include the prevalence for the use of codons of the form RNY (here R = purine, Y = pyrimidine, and N = any nucleotide) (Shepherd 1981), the nonuniform positional nucleotide usage (Fickett 1982; Staden 1984; Fickett & Tung 1992), the different G+C content (Bibb *et al.* 1984), the detection of characteristic periodicities (Silvermann & Linsker 1986; Tiwari *et al.* 1997), the higher concentration of G in the first codon position (Trifonov 1987), the positional dependence of entropy (Amalgor 1985; Grosse *et al.*, 1999a) and of nucleotide pair correlations (Grosse *et al.* 1999b).

We introduce generalized coding measures and study how accurately the position asymmetry function D_p and the position information function I_q distinguish coding from non-coding DNA in dependence on the parameter values p and q. We search for parameter values which maximize the discrimination accuracy, and we contrast and compare these optimum parameters with conventionally used parameters. Our study shows that exon recognition with optimum parameters yields comparable results in accuracy as currently employed coding measures, some of which require a much higher number of parameters. We make use of two different sequence data sets to evaluate the accuracy: (i) to connect our studies to previous work we use data from standardized benchmarks of Fickett & Tung (1992), and in addition (ii) we use recent data from GenBank (release 111 1999). Since the discrimination accuracy of many coding measures has been shown to depend strongly on the A+T content (Guigó & Fickett 1995), we examine also the statistical dependences of coding measure functions on different A+T content of DNA sequences. We conduct this analysis using graphical methods and tests for non-linear and linear statistical dependences.

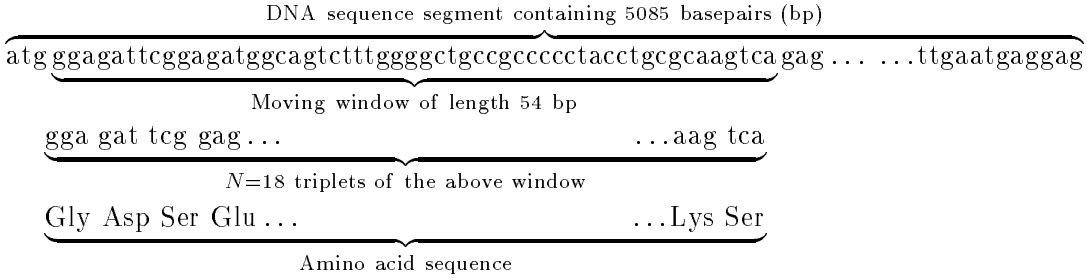
The content of this study is as follows. In section 2 we introduce the frame dependence matrix from which we derive current coding measures. We discuss how a certain measure captures the positional dependence of nucleotide frequencies and supplement the discussion by an example. In this context, we examine the statistical dependence of coding measures on the A+T content of the query sequence. In section 3 we introduce the data sets prepared.

We turn to the performance of coding measures and evaluate explicitly the discrimination accuracy of the positional information measure I_1. In section 4 we introduce generalized coding measures as function of parameters. We discuss properties of these functions and evaluate the discrimination accuracy as a parameter-dependent function. We search for those parameters which extract optimally the discriminating features to distinguish coding versus non-coding DNA. We contrast and compare our results with conventionally used parameter values and with the performance of established coding measures.

17.2 Currently Applied Coding Measures

We consider a moving window of length $3 \cdot N$ along a DNA sequence and decompose the window into N successive non-overlapping triplets. Let $F(b|l)$ denote the number of occurrences of base b $(= 1,2,3,4 \equiv \mathrm{A,C,G,T})$ at the triplet position $l \in (1,2,3)$. The corresponding relative frequencies are defined by $f(b|l) = F(b|l)/N$. Let the *frame dependence matrix* $\mathbf{F}$ be the matrix in which each element contains $f(b|l)$. Then, $\mathbf{F}$ has 4×3 elements where due to normalization 9 suffice to fully determine it. Furthermore, we separate the mean $f(b)$ of base b according to $f(b|l) = f(b) + d(b|l)$ as the elements of the vector $\mathbf{f}$. The mean of base b is calculated from $f(b) = \sum_{l=1}^{3} f(b|l)/3$. The residuals $d(b|l)$ enter as elements of the matrix $\mathbf{D}$. The entries of $\mathbf{D}$ represent the positional deviations from occurrences expected by chance. Obviously, $\mathbf{F}$ is complementary to $\mathbf{D}$ and $\mathbf{f}$.

The notations are visualized in the following sketch, in which the data are obtained from the 5085 bp long protein-coding sequence from the intensively studied (Buldyrev *et al.* 1993) human beta-myosin heavy chain (HUMBMYH7) gene:



From the decomposition we calculate the elements $f(b|l)$, $f(b)$, and $d(b|l)$. For the above

window they read (for illustrative purposes the elements have round-off errors of 0.01):

$$\mathbf{F} = \begin{pmatrix} 0.11 & 0.22 & 0.17 \\ 0.17 & 0.39 & 0.39 \\ 0.50 & 0.17 & 0.33 \\ 0.22 & 0.22 & 0.11 \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} 0.17 \\ 0.32 \\ 0.33 \\ 0.18 \end{pmatrix}, \text{ and } \quad \mathbf{D} = \begin{pmatrix} -0.06 & 0.05 & 0.00 \\ -0.15 & 0.07 & 0.07 \\ 0.17 & -0.16 & 0.00 \\ 0.04 & 0.04 & -0.07 \end{pmatrix}.$$

Note that in this way, *e.g.*, the excess (lack) of G in the first (second) codon position as well as the high G+C content in the third codon position become apparent.

At this stage, we briefly reconsider currently applied coding measures proposed to capture the discriminating features in $\mathbf{F}$, or $\mathbf{D}$ and $\mathbf{f}$, in a single scalar coding potential. We outline the mechanism concerning how a nonuniform codon usage affects a certain measure. Later on this will allow us to introduce a framework of generalized coding measures. For illustration, we include calculations for the above window.

- *TestCode* (Fickett 1982). The base compositional asymmetry is quantified in a linear weighted sum containing eight parameters, four of them are given by $f(b)$. The asymmetry $A(b)$ of base b is defined by the ratios of the maximal and minimal values of $F(b|l)$. The resulting weighted overall coding measure is calculated to be

$$A = \sum_{b=1}^{4} \left\{ W(b) \cdot \frac{\max_{l\in(1,2,3)}\{F(b|l)\}}{\min_{l\in(1,2,3)}\{F(b|l)\}+1} + w(b)\cdot f(b) \right\} \tag{17.1}$$

  where the weights $W(b)$ and $w(b)$ are determined by training sets of exons and introns, and the additional 1 in the denominator has been introduced to avoid the divergence in the case $F(b|l)=0$. In order to construct a parameter-independent measure, we set $W(b)\equiv 1$ and $w(b)\equiv 0$. Then,

$$A = \sum_{b=1}^{4} \left\{ \frac{f(b)+\max_{l\in(1,2,3)}\{d(b|l)\}}{f(b)+\min_{l\in(1,2,3)}\{d(b|l)\}+1/N} \right\} \tag{17.2}$$

  measures the largest positional deviation from the mean. If we substitute the values from $\mathbf{f}$ and $\mathbf{D}$ according to the window given above, we obtain $A=9.23$ as a measure for its coding potential.

- *Uneven positional base frequencies* (Staden 1984). The deviation of the positional base concentration from its mean are used to define the distances $|d(b|l)|$. This coding measure reads

$$D_1 = \sum_{b=1}^{4}\sum_{l=1}^{3} |d(b|l)|. \tag{17.3}$$

The introduction of the index "1" will become clear in the context later on. Substituting the values from the sample window, we obtain $D_1 = 0.88$. To contrast this outcome with non-coding DNA, we make the simplifying assumption that each base b shows no dependence on the triplet position, $f(b|1) = f(b|2) = f(b|3)$. Hence $d(b|l) = 0$ due to the absence of any frame dependence, and the coding potential according to this model vanishes.

- *Position asymmetry* (Fickett & Tung 1992). The sample variance is used to quantify the positional spread of the mean frequencies $f(b)$, using

$$D_2 \quad = \quad \sum_{b=1}^{4} \sum_{l=1}^{3} d^2(b|l). \tag{17.4}$$

Obviously, this coding measure is closely related to the one proposed by Staden (1984). Applying eqn (17.4) to the sample window yields $D_2 = 0.10$, and for the model of non-coding sequences $D_2 = 0$.

- *Fourier measure* (Silverman & Linsker 1986). The square of the Fourier transform (the power spectrum) is derived from 4 (each for one base) binary translated DNA sequences. Assigning some base b at the nth window position $U_n(b) = \delta_{a,b}$ (where $\delta_{a\,b} = 1$ if $a = b$, and it is 0 otherwise), the power spectrum can be calculated from

$$P(f) \quad = \quad \sum_{b=1}^{4} \left| \frac{\sum_{n=1}^{3 \cdot N} U_n(b) \cdot \mathrm{e}^{-i \cdot 2\pi f \cdot n}}{3 \cdot N} \right|^2 \tag{17.5}$$

where $f = \frac{m}{3 \cdot N}$ with $m = 1, \ldots, \frac{3 \cdot N}{2}$. Commonly $m = N$ is used to extract one scalar coding measure $P(\frac{1}{3})$ from the spectrum signifying the period-3 amplitude. Interestingly, it has been observed (Guigó 1999) that $P(\frac{1}{3})$ and D_2 measure the same coding potential. Since $\frac{1}{3 \cdot N} \cdot \sum_{n=1}^{3 \cdot N} U_n(b) = f(b)$, it can be analytically shown that by using $\mathrm{e}^{-i\frac{2\pi}{3}}$ as weights, $P(\frac{1}{3})$ is up to a constant equal to D_2 (Grosse, unpublished). Consequently, $P(\frac{1}{3})$ can be derived from the frame dependence matrix $\mathbf{F}$. The full spectrum can in fact be used to overcome statistical noise (Tiwari *et al.* 1997) using the ratio $\frac{P(\frac{1}{3})}{\langle P(f) \rangle}$, where the average $\langle P(f) \rangle$ can computed from $f^2(b)$.

- *Positional information.* For each position l, the position entropy $H_1(l)$ can be calculated from $f(b|l)$ (Shannon 1948; Schneider 1997). Likewise, the entropy corresponding to the mean H_1 can be calculated from $f(b)$. Considering each $H_1(l)$ in isolation (Amalgor 1985) results to poor discrimination accuracy (Fickett&Tung 1992). To

obtain an accurate coding measure it turns out that difference between H_1 and the average of $H_1(l)$ is appropriate, defining the positional information as (Grosse *et al.* 1999a)

$$\begin{aligned} I_1 &= H_1 - \frac{1}{3}\cdot\sum_{l=1}^{3} H_1(l) \\ &= -\sum_{b=1}^{4} f(b)\cdot\log_2 f(b) + \frac{1}{3}\cdot\sum_{l=1}^{3}\sum_{b=1}^{4} f(b|l)\cdot\log_2 f(b|l). \end{aligned} \tag{17.6}$$

If we use the relation $f(b,l) = f(b|l)\cdot f(l)$ and introduce $f(l) = \sum_{b=1}^{4} f(b,l)$ as the relative frequency to find b at the lth triplet position, the positional information can be written as a mutual information between positional nucleotides. This gives rise to

$$I_1 = \sum_{b=1}^{4}\sum_{l=1}^{3} f(b,l)\cdot\log_2\left(\frac{f(b,l)}{f(b)\cdot f(l)}\right). \tag{17.7}$$

I_1 is the average information in base b about the position l (and vice versa) measured in units of bits. In case of the window above, we obtain $I_1 = 0.09$ (bits/bp). For non-coding DNA, $f(b,l)$ factorizes to $f(b)\cdot f(l)$, and $I_1 = 0$ vanishes.

- *Average mutual information* (Grosse *et al.* 1999b). The mutual information $I(k)$ as a function of the base pair (a,b) separated by a distance k is used. Under the simplifying assumption that the DNA sequence consists of statistically independent codons (Herzel & Grosse 1995), the corresponding frequencies $f_k(a,b)$ factorize and can be obtained from $\mathbf{F}$ according to

$$f_k(a,b) = \frac{1}{3}\cdot\begin{cases} f(a|1)\cdot f(b|1) + f(a|2)\cdot f(b|2) + f(a|3)\cdot f(b|3) & k = 3,6,9,\ldots \\ f(a|1)\cdot f(b|2) + f(a|2)\cdot f(b|3) + f(a|3)\cdot f(b|1) & k = 4,7,10,\ldots \\ f(a|1)\cdot f(b|3) + f(a|2)\cdot f(b|1) + f(a|3)\cdot f(b|2) & k = 5,8,11,\ldots \end{cases}$$

Transposition of the subscripts a and b shows that $f_k(a,b) = f_{k+1}(b,a)$ for $k = 4,7,10,\ldots$ and, consequently, $I(k)$ assumes only two values: the in-frame and the out-of-frame mutual information, I_{in} for $k = 3,6,9\ldots$ and I_{out} for $k = 4,5,7,8,\ldots$. The average is used to define

$$\bar{I} = \frac{I_{\text{in}} + 2\cdot I_{\text{out}}}{3}. \tag{17.8}$$

Eqn (17.8) quantifies the average mutual information shared by a and b given the distance k between a and b is a multiple of 3. For the window above, the overall

result is calculated to be $\bar{I} = 0.005$ (bits/bp), whereas we find $\bar{I} = 0$ (bits/bp) for non-coding DNA.

In the following section we will evaluate how accurate the coding measure I_1 distinguishes coding versus non-coding DNA. We introduce the data on which this study will be conducted, sketch the evaluation technique, and study the statistical dependence of I_1 on different A+T content.

17.3 Discrimination Accuracy of Positional Information

If we apply a coding measure to sets of exons and introns, we can derive histograms as estimates for the associated probability distributions. A perfect coding measure would generate two non-overlapping histograms. In practical applications, however, histograms overlap due to the finite length of windows. We evaluate the performance of coding measures as follows:

1. Using the abbreviations T (true), F (false), P (positives), and N (negatives), we denote the relative number of coding sequences correctly predicted as coding by the sensitivity $\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$. In the same manner, we denote for non-coding sequences the specificity $\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN+FP}}$.
2. We determine the threshold above which a sample window is assigned to the exon class or the intron class by imposing equal relative errors on exons and on introns, writing $\mathrm{Sn} = \mathrm{Sp}$.
3. Eventually, we quantify the accuracy of a coding measure to yield correct predictions as $\frac{\mathrm{Sn+Sp}}{2}$, ranging from $\frac{1}{2}$ (no discrimination) to 1 (perfect discrimination).

To compare the accuracy by which the positional information I_1 can distinguish coding versus non-coding DNA with the accuracy of other coding measures, we use standard benchmarks and recent data sets of human DNA. By set A, we refer to sequences from the Fickett & Tung (1992) benchmark data set (extracted from GenBank databases in 1992). Since I_1 does not require prior training, we evaluate I_1 for both the training ($\mathrm{A_{training}}$) and the test set ($\mathrm{A_{test}}$). Recently, this data set has been used by Yan *et al.* (1998) to re-evaluate the accuracy of the Fourier measure using a length-shuffled version. Instead of the complete benchmark data, there a subset of 1000 (500 coding, 500 non-coding) sequences for training and (500, 500) for testing was used. In order to avoid the possibility that

one data set might not reflect real performance, we make use of an additional sequence set B, by which we refer all human sequence files extracted from GenBank release 111. For this application, non-overlapping fully protein-coding and fully non-coding sequence windows of length L bp were obtained by partitioning the above sequences longer than L into windows[1] of length L. When Fickett & Tung (1992) evaluated the accuracy of the entropy measure $H_1(l)$ (Amalgor 1985), it performed only moderately in distinguishing coding versus non-coding DNA. We already mentioned in the previous section that this is mainly due to the use of $H_1(l)$ in isolation, rather than to consider I_1, namely the difference $H_1 - \langle H_1(l) \rangle_l$. In this section, we will show that the positional information I_1 is superior to the entropy $H_1(l)$ by evaluating the accuracy of I_1 on datasets $\mathrm{A_{training}}$, $\mathrm{A_{test}}$, and B.

I_1 varies from one window to another. We estimate the probability density function $\mathcal{P}(I_1)$ ($\mathcal{Q}(I_1)$) by applying I_1 to coding (non-coding) DNA. Figure 1 shows the resulting I_1-histograms for sets $\mathrm{A_{training}}$, $\mathrm{A_{test}}$, and B. We find that both coding and non-coding DNA have unimodal I_1-histograms with distinct maxima. We detect that for either data set the I_1-histograms are significantly different for coding and non-coding DNA, although the histograms overlap owing to the finite window length. In each data set I_1-histograms for non-coding DNA are centered at significantly smaller values than the histograms of coding DNA. Figure 1 also shows that the I_1 histograms for sets $\mathrm{A_{traning}}$, $\mathrm{A_{test}}$, and B are similar. Hence, the accuracy by which I_1 distinguishes between coding and non-coding DNA does not sensitively vary when evaluated on human DNA sequences from GenBank releases in 1992 and 1999. Fickett & Tung (1992) state the accuracy of the entropy measure $H_1(l)$ for windows of length 108 bp with 63.1%. The inset shows that I_1 allows a discrimination of coding versus non-coding DNA with an accuracy of approximately 76% evidencing that a novel application of entropy constitutes indeed a powerful coding measure (Grosse *et al.* 1999a). We also evaluate the accuracy of the coding measures $A, \bar{I}, D_1$, and D_2 discussed in the previous section. Table 1 shows in part (a) the accuracy as obtained by application to $\mathrm{A_{traning}}$, $\mathrm{A_{test}}$, and B.

In Figure 1, we observe for coding sequences a small but detectable shift causing the overlap between $\mathcal{P}(I_1)$ and $\mathcal{Q}(I_1)$ to become larger when using set B. We examine whether

[1]Number of coding (non-coding) windows: $\mathrm{A_{training}}$ 54 bp: 20,456 (125,132), 108 bp: 7,086 (58,118), and 162 bp: 3,512 (36,502). $\mathrm{A_{test}}$ 54 bp: 22,902 (122,138), 108 bp: 8,192 (57,032), and 162 bp: 4,266 (35,602). Set B 54 bp: 595,194 (171,677), 108 bp: 282,876 (81,110), 162 bp: 178,520 (51,078), and 1080 bp: 4482 (15,984).

this slight decrease in accuracy is possibly related to a biased A+T content toward low regions. This would signify the sensibility of the performance of I_1 with respect to A+T content variations in databases. Figure 2 shows the A+T content of the data sets and displays mean and variance of I_1 binned to 20 intervals. From the figure we can conclude that one possible explanation for the worsening in accuracy is indeed the difference in the individual A+T content of the data sets. A comparison of the top graphs shows that set B contains more sequences with high A+T content as compared to sets $\mathrm{A}_{\mathrm{training}}$ or $\mathrm{A}_{\mathrm{test}}$. The bottom graphs show that the discrimination is more accurate for sequences with low A+T content. It is a general feature that many coding measures share this property of being dependent on the A+T content (see, *e.g.*, Guigó & Fickett 1995). While I_1 is practically independent on the A+T content of non-coding sequences, it shows a dependence on the A+T content of coding sequences and decays with increasing content.

The linear and non-linear statistical dependences captured by Figure 2, which is in essence a scatter plot binned to 20 ranges, can be quantified. We therefore calculate the *correlation coefficient* $C(X,Y)$ between I_1 $(=X)$ and the window A+T content $(=Y)$ to measure linear correlations. Nonlinear correlations can be quantified by calculating the *uncertainty coefficient* $U(X,Y)$. Details for the calculation of correlation and uncertainty coefficients are consigned to the Appendix. Table 2 shows in part (a) results of $C(X,Y)$ and in part (b) results of $U(X,Y)$ applied to data samples derived from $X = A, \bar{I}, D_1$, D_2, and I_1. We observe for all coding measures weak linear and non-linear dependences for introns, whereas we observe clear linear anti-correlations and higher non-linear correlation for exons. Because $U(X,Y)$ accumulates all pair correlations (Herzel & Grosse 1997), the weakness of correlations between introns and A+T content implies that $U(X,Y) \propto C^2(X,Y)$. We also note that values of $C(X,Y)$ and $U(X,Y)$ are in general higher for sets $\mathrm{A}_{\mathrm{training}}$ or $\mathrm{A}_{\mathrm{test}}$ than for set B, a feature related to the lower overall A+T content of set B.

17.4 Optimization of Coding Measures

In this section, we introduce generalized coding measures. Inspecting the measures D_1 and D_2, as defined in eqn (17.3) and (17.4), we realize that these are actually just two special values of the coding measure function

$$D_p \;=\; \sum_{b=1}^{4}\sum_{l=1}^{3} |d(b|l)|^p \,, \tag{17.9}$$

where p can take on any real number. D_p recovers the measure of Staden (1984) for $p = 1$ and the measure of Fickett&Tung (1992) for $p = 2$. We refer to eqn (17.9) as *position asymmetry function*.

Another coding measure function is obtained in the realm of entropy. Recall that I_1 can be expressed as the difference $H_1 - \langle H_1(l) \rangle_l$. According to Rényi (1970), there exists a natural extension of the ordinary entropy to entropies H_q characterized by a parameter q which can take on any real number. Introducing $Z_q = \sum_{b=1}^{4} f^q(b)$, the Rényi entropies are defined as $H_q = \frac{\log_2 Z_q}{1-q}$. If we now substitute H_q for H_1 in eqn (17.6), we define the *position information function* as

$$\begin{aligned} I_q &= H_q - \frac{1}{3} \cdot \sum_{l=1}^{3} H_q(l) \\ &= \frac{\log_2 \left(\sum_{b=1}^{4} f^q(b) \right)}{1-q} - \frac{1}{3} \cdot \sum_{l=1}^{3} \frac{\log_2 \left(\sum_{b=1}^{4} f^q(b|l) \right)}{1-q}. \end{aligned} \tag{17.10}$$

Equivalently, we can write the above expression using $f(l)$ as

$$I_q = \frac{1}{1-q} \cdot \sum_{l=1}^{3} f(l) \cdot \log_2 \left(\frac{\sum_{b=1}^{4} f^q(b) \cdot f^q(l)}{\sum_{b}^{4} f^q(b,l)} \right). \tag{17.11}$$

Taking the limit $q \to 1$, I_q recovers the positional information I_1 as defined in eqn (17.7).

The generalizations D_p and I_q comprise the focus of the remaining studies. The motivation to introduce coding measure functions stems from the insight that up until now the coding potential of most DNA sequences has been studied by computing, *e.g.*, D_1, D_2, and I_1. That is, just one or two points in an infinite spectrum of discriminating functions D_p and I_q. Hence, one is effectively neglecting discriminating features which may be present in the whole spectrum of D_p and I_q. Here we generalize the method of coding measures in such a way that all coding measures D_p and I_q can be computed from a sequence window. We study the discrimination accuracy of D_p (I_q) as a function of p (q) and evaluate optimal parameters which distinguish coding versus non-coding DNA most accurately.

Note the mechanism of how a parameter change affects a coding measure. Consider, for instance, I_q. The parameter q plays a role similar to a weight. A change of q will change the relative weight that $f(b)$ and $f(b|l)$ contribute to the coding potential. The higher q, the more heavily the larger frequencies contribute, and vice versa. In the limit $q \to \infty$, only the highest frequencies contribute. On the other hand, $q < 0$ weights the contribution of lower frequencies. Hence, varying q allows for all discriminant features captured by

the compositional bias of codon positions in I_q. In a similar manner, the same reasoning applies to D_p by considering moments of order p.

We study the discrimination accuracy of coding measure functions by applying D_p and I_q to $\mathrm{A}_{\mathrm{training}}$, $\mathrm{A}_{\mathrm{test}}$, and B. We compute the accuracy as a function of p and q, respectively, and compare the results with the accuracy of current coding measures. In order to compare the accuracy with previously obtained results on benchmarks, in this evaluation windows have lengths 54, 108, and 162 bp. Figure shows the accuracy for D_p s a function of $p \in [-10, 10]$ and Figure 4 for I_q a and $q \in [-10, 10]$ for windows of length 108 bp. The accuracy exhibits in either case a strong dependence on the parameters values, Both D_p and I_q have a significant higher accuracy for $p, q > 0$ as compared to $p, q < 0$. After passing through zero, the accuracy of D_p attains its global maximum and then varies only little for $p \in [0, 10]$. The accuracy of I_q shows a pronounced peak at about $q = 2$ and decays steeply when q increases. If we consider D_p and I_q as coding measure functions of integer parameters, we find that D_2 and I_2 yield the maximum accuracy with I_2 being the most accurate measure. While $p = 2$ regains the position asymmetry (Fickett & Tung 1992) as the most effective D_p coding measure, for $q = 2$, we obtain that I_2 constitutes a novel coding measures.

In numerical experiments with windows of length 54 and 162 bp we could observe qualitatively similar results. Table 1 summarizes our findings. Part (a) shows the accuracy of current and generalized coding measures using positional dependence of nucleotide frequencies for distinguishing coding versus non-coding DNA for the three different window sizes. Part (b) lists the most effective (phase-independent) coding measures after prior training on set $\mathrm{A}_{\mathrm{training}}$ for set $\mathrm{A}_{\mathrm{test}}$ according to Fickett & Tung (1992). Part (b) also shows the number of parameters required for discrimination (for the Fourier it depends on the number of spectral lines used from the total spectrum, and for the Run measure it depends on the number of runs of "non-trivial" subsets). The results point out that both D_p and I_q yield a comparable accuracy as customarily used coding measures after prior training, some of which require a much higher number of parameters. Since D_p and I_q can be computed efficiently (we extract 12 positional nucleotide frequencies from a window) and need not be trained on prior data sets, generalizations of coding measures could easily be incorporated into existing algorithms.

Having determined the optimal parameters p_{opt} and q_{opt}, we examine the statistical dependence of D_p and I_q at p_{opt} and q_{opt} on the A+T content. Table 2 shows for several

parameter values adjacent to $p_{\rm opt}$ and $q_{\rm opt}$ in part (a) linear and in part (b) non-linear statistical dependences of D_p and I_q on the A+T content. D_p shows overall weak correlations for non-coding DNA, but persistent anti-correlations for coding DNA. On the other hand, I_q correlates in the linear realm lesser with coding DNA when q is increased, albeit correlations decrease in the non-linear realm for non-coding DNA.

We further investigate the dependence of the maximum accuracy at $p_{\rm opt}(N)$ and $q_{\rm opt}(N)$ on the number of triplets N in a window. To conduct this study we evaluate the accuracy of D_p (I_q) as a function of p (q) in consideration of the length N. We partition all human sequences from GenBank into non-overlapping fully protein-coding and fully non-coding sequences longer than 1080 bp into windows of 1080 bp. We then study the accuracy for windows of length which differ two orders of magnitude, ranging from 1080 to 27 bp, while keeping the overall number of bases constant. Figures 5 and 6 represent our findings. Figure 5 shows that $p_{\rm opt}$ is approximately independent on N, while Figure 6 shows this approximate independence for $q_{\rm opt}$. We note, however, that the sharp accuracy profile derived from I_q becomes less pronounced with increasing length N, reflecting that I_q becomes practically parameter-independent in the limit of large window length. As such, usage of $q_{\rm opt}$ plays an important role when using a window with only moderate length, *e.g.* in the order of $N \sim 100$ bp. Indeed, this situation is quite relevant, for human DNA sequences have an average exon length of $\sim$150 bp and have even been reported to range down to the order of 10 bp (Deutsch & Long,1999).

17.5 Conclusions

We generalized coding measures as functions of parameters. We study the position asymmetry function, D_p, and the position information function, I_q. Within this framework, for special values of p and q the functions D_p and I_q include several current coding measures: the uneven positional frequency ($p = 1$), the position asymmetry ($p = 2$), respectively the Fourier measure, and the position entropy ($q = 1$).

We study the discrimination accuracy of D_p and I_q as a function of the parameters p and q. We find that the accuracy of how well D_p and I_q distinguish protein-coding versus non-coding DNA shows a strong dependence on p and q. Our findings reveal that the accuracy of D_p (I_q) as a function of p (q) can be used to search for optimal parameters that extract most discriminating features to distinguish coding versus non-coding DNA.

The accuracy of coding measures presented here can thus be maximized using optimal parameters $p_{\rm opt}$ or $q_{\rm opt}$. In integer parameter space we find D_2 and I_2 to be the most accurate coding measures, where $I_2 \equiv \sum_{l=1}^{3} f(l) \cdot \log_2 \left[\sum_{b=1}^{4} f^2(b,l) / \sum_{b=1}^{4} \left(f(b) \cdot f(l) \right)^2 \right]$. Whereas D_2 establishes the position asymmetry as the most accurate coding measure, I_2 constitutes a novel coding measure.

It has been noted that DNA sequences available in current databases are biased, *e.g.* towards atypical codon usage or low A+T compositional samples as shown in Figure 1, and this in turn could affect the performance of gene identification algorithms. In this regard, coding measures using positional dependent nucleotide frequencies are advantageous in that they can be applied without prior training. To access the A+T content dependence, we examine linear and non-linear statistical relationships between coding measures and the A+T content of windows extracted from recent data. We find both current and generalized coding measures D_q and I_q (at $p_{\rm opt}$ and $q_{\rm opt}$) relatively insensitive to the A+T content of non-coding sequences and clearly anti-correlated to the A+T content of coding sequences. While D_p exhibits persistent correlations between the A+T content and parameter values adjacent to $p_{\rm opt}$, linear (non-linear) correlations between the A+T content and I_q decrease (increase) for parameter values $q > q_{\rm opt}$. Further research directions will include a comprehensive analysis of the dependence of coding measures on the A+T content of model and database sequence sets. The applied combination of linear and non-linear analysis can easily be used to detect statistical dependences in more sophisticated coding measures, *e.g.* based on dicodon interrelations.

We study the accuracy profile of D_p (I_q) at their optimal parameters $p_{\rm opt}$ ($q_{\rm opt}$) as a function of the number of triplets N used to evaluate the coding potential. Our findings show $p_{\rm opt}(N) \approx p_{\rm opt}$ and $q_{\rm opt}(N) \approx q_{\rm opt}$ approximately independent on N, thus avoiding the expendable fine-tuning of parameters. The computation of D_p and I_q is efficient and can be supplemented by existing algorithms that search for gene signals, *e.g.*, start and stop codons, splice junctions, and promoter sequences. The coding measures presented here are universal in that they are not species-specific in detecting positional dependences of nucleotide frequencies from $\mathbf{F}$. The optimization of D_p and I_q does not sensitively depend on any predefined window length and perform reasonably well for window sizes as moderate as 54 bp.

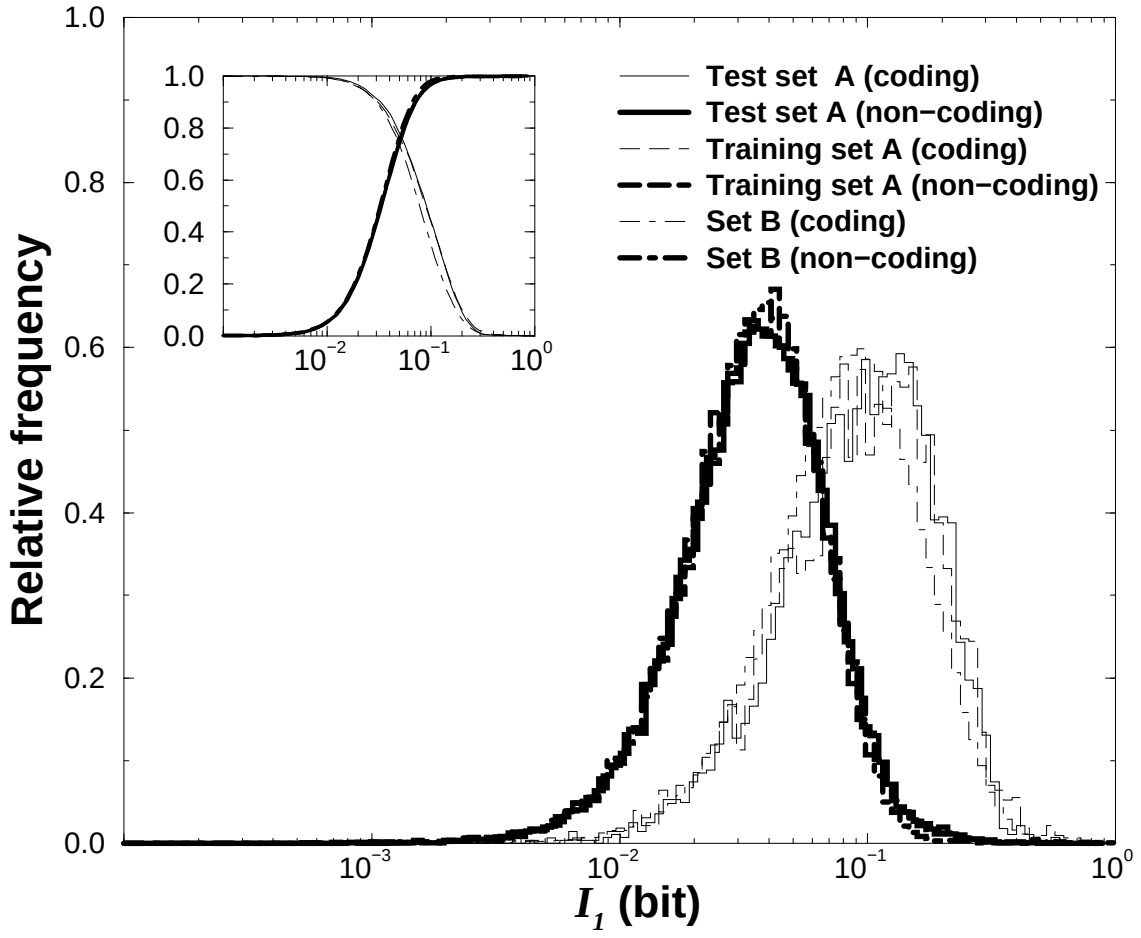
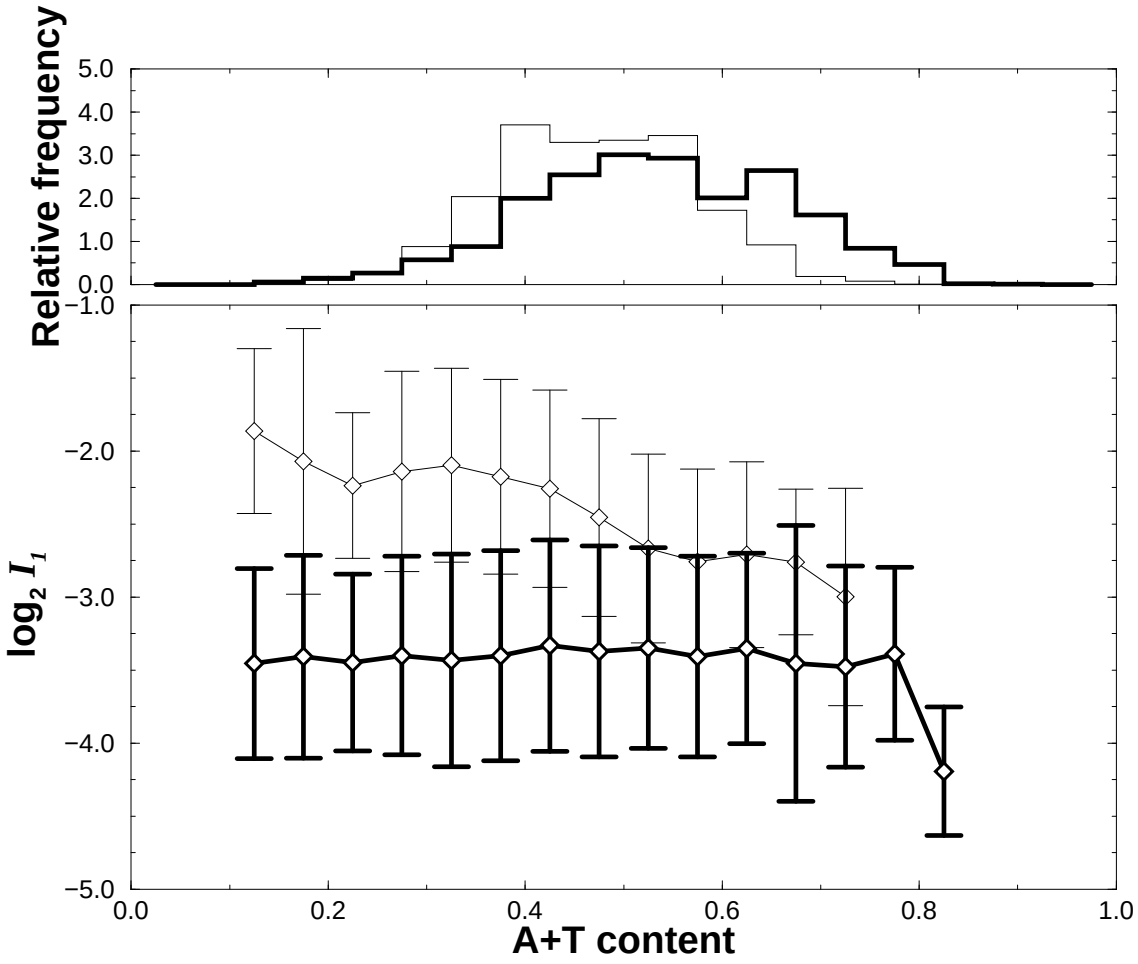
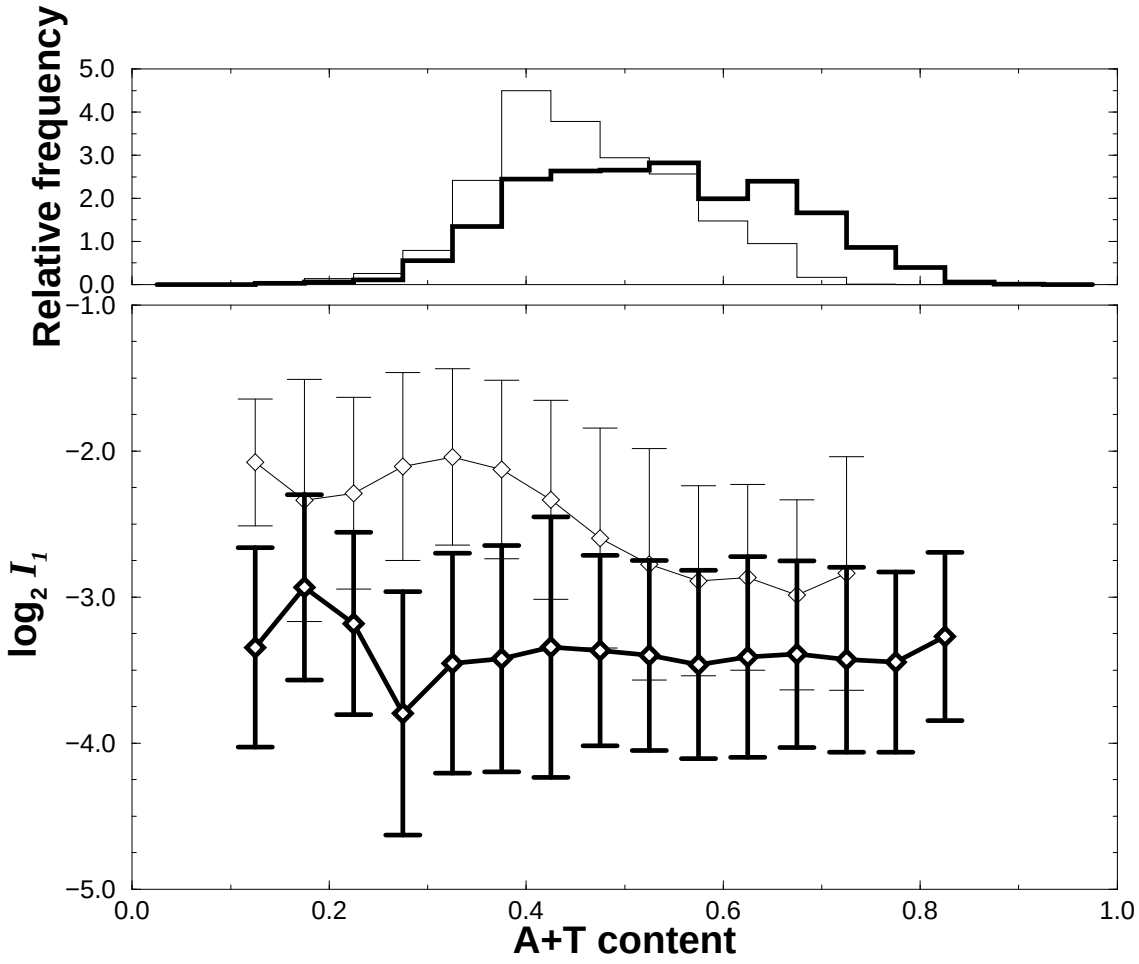


Figure 17.1: Histograms of I_1 for human exons (right) and introns (left) of length 108 bp. The corresponding cumulative distributions are shown in the inset. While the values of I_1 fluctuate from window to window, their distribution is almost the same irrespective of $\mathrm{A}_{\mathrm{training}}$, $\mathrm{A}_{\mathrm{test}}$, or B. For all three data sets, the I_1-histograms of non-coding DNA are centered at significantly smaller values as compared to coding DNA. For $\mathrm{A}_{\mathrm{training}}$ the mean μ and standard deviation σ for exons (introns) are $\mu(\log I_1) = -2.39$ (-3.36) and $\sigma(\log I_1) = 0.70$ (0.68), for $\mathrm{A}_{\mathrm{test}}$ $\mu(\log I_1) = -2.41$ (-3.38) and $\sigma(\log I_1) = 0.74$ (0.69), and for B $\mu(\log I_1) = -2.52$ (-3.41) and $\sigma(\log I_1) = 0.69$ (0.66). For coding DNA, we observe a small but significant shift of set B with respect to both $\mathrm{A}_{\mathrm{training}}$ and $\mathrm{A}_{\mathrm{test}}$. Both distributions show an overlap, the enclosed area of which specifies how accurately we can distinguish coding versus non-coding DNA. The inset shows that I_1 performs on $\mathrm{A}_{\mathrm{training}}$ and $\mathrm{A}_{\mathrm{test}}$ with approximately 76%, and on set B with approximately 75% accuracy.


Relative frequency
5.0
4.0
3.0
2.0
1.0
0.0
−1.0
−2.0
−3.0
−4.0
−5.0
log2 I1
0.0
0.2
0.4
0.6
0.8
1.0
A+T content


Relative frequency
5.0
4.0
3.0
2.0
1.0
0.0
−1.0
−2.0
−3.0
−4.0
−5.0
log2 I1
0.0
0.2
0.4
0.6
0.8
1.0
A+T content

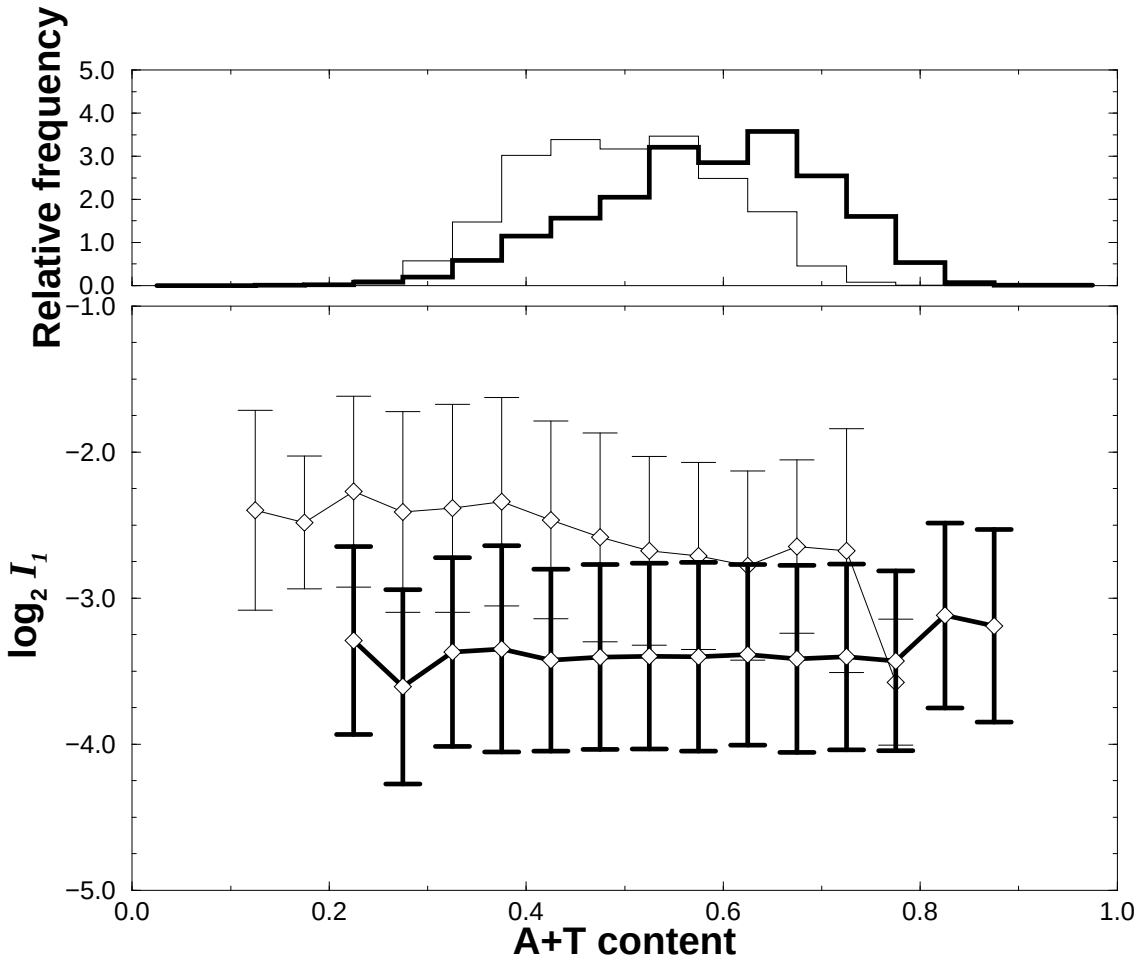


Figure 17.2: We study the dependence of I_1 on the A+T content for exons (thick graph) and introns (thin) in $A_{training}$ (a), A_{test} (b), and B (c). To calibrate error bars, we equate the number of coding sequences with the number of non-coding sequences in all three figures (each class comprising 7000 sequences). In the top, we display the histograms of the overall A+T content as derived from the data sets. The overall A+T content of exons is approximately 45% in $A_{training}$, 43% in A_{test}, and 47% in B. For introns, it is 52%, 51%, and 54%. In the bottom, we show the dependence of $\log I_1$ on the A+T content, by binning the A+T values to 20 bins and computing the mean and standard deviation of I_1 per bin. The overlap of error bars indicates that the discrimination of exons from introns is less accurate for high A+T content.

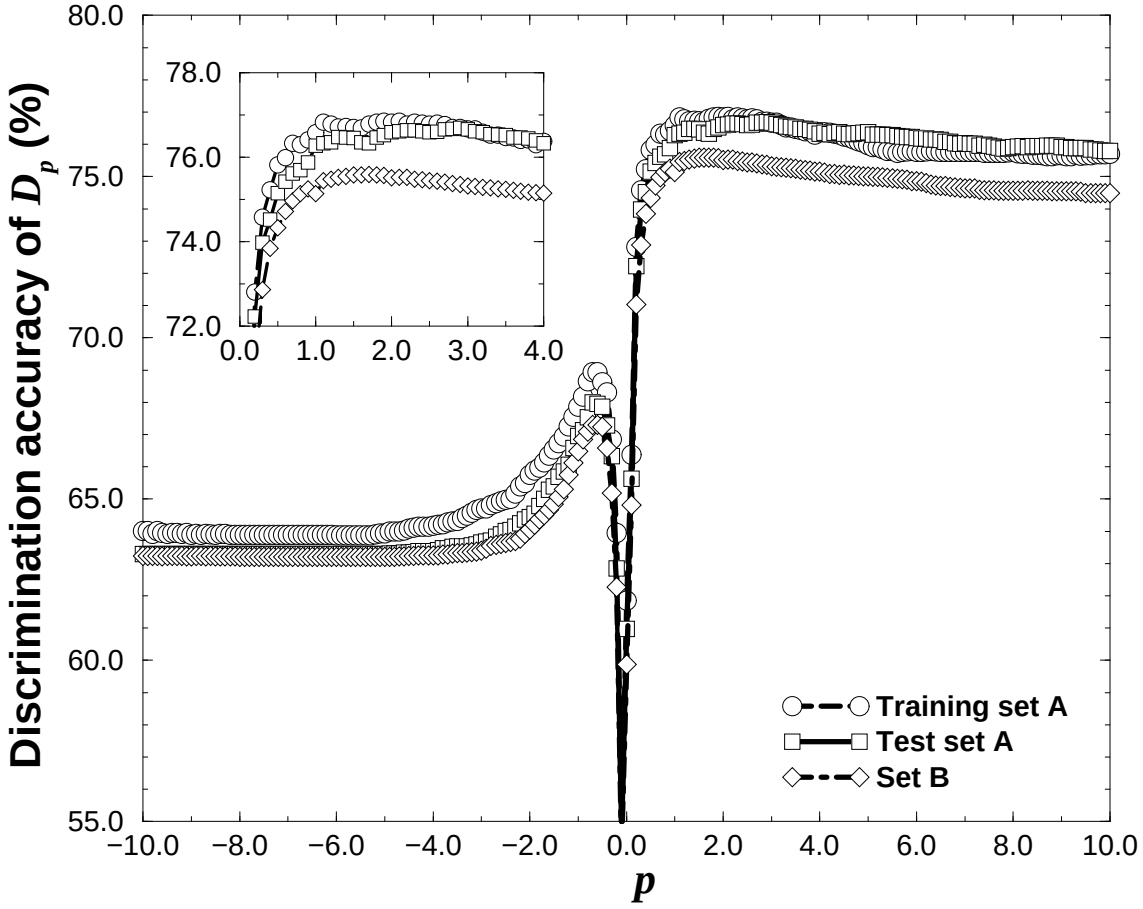


Figure 17.3: The accuracy of D_p as a function of the parameter p evaluated on $\mathrm{A}_{\mathrm{training}}$, $\mathrm{A}_{\mathrm{test}}$, and B. The region around p_{opt} is shown in the inset. The accuracy shows a strong dependence on the parameter p, dropping to nearly 50% (no discrimination) while zero-crossing. The accuracy exhibits two distinct maxima, one local for $p < 0$. For $p > 0$, the accuracy of D_p reaches its global maximum, and it shows a plateau-like behavior for $p > p_{\mathrm{opt}}$ while decaying flatly.

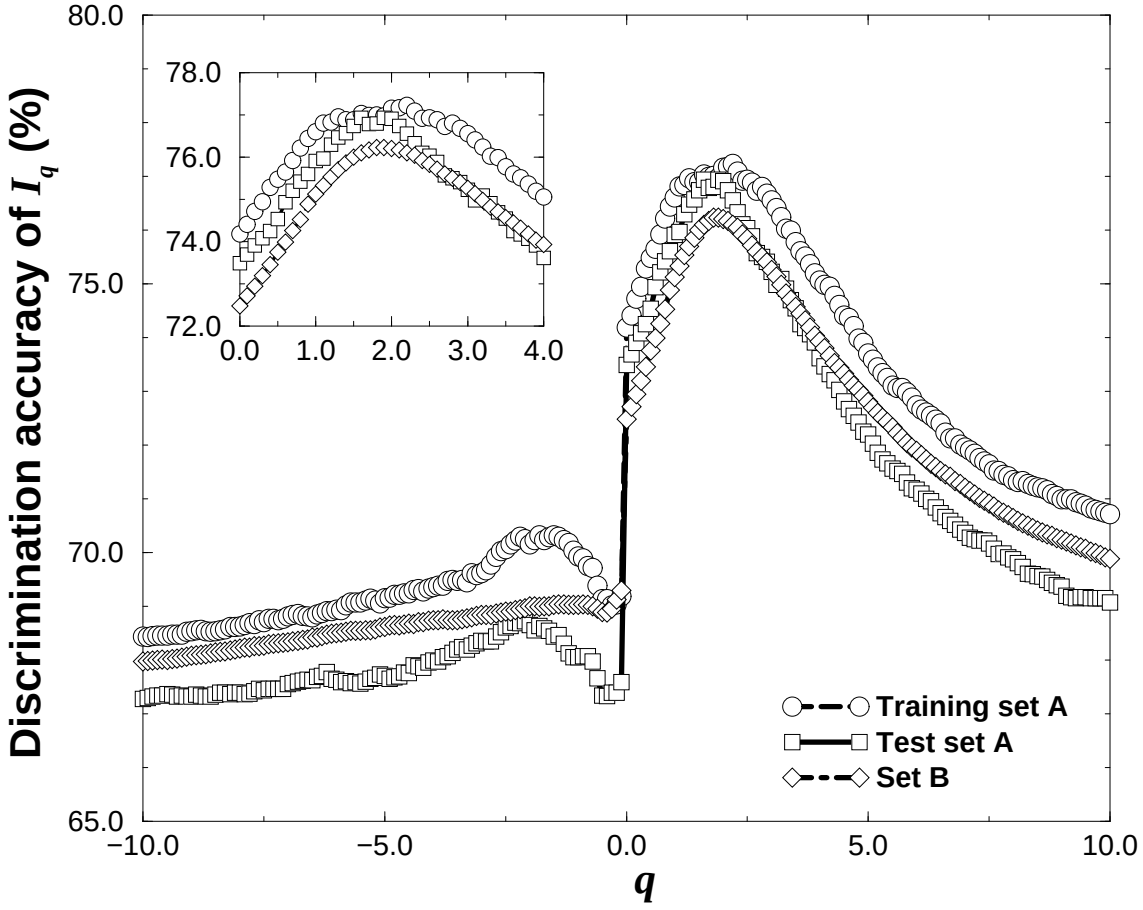


Figure 17.4: The accuracy of I_q as a function of the parameter q evaluated on $\mathrm{A}_{\mathrm{training}}$, $\mathrm{A}_{\mathrm{test}}$, and B. The region around q_{opt} is shown in the inset. The accuracy depends on the parameter q, and exhibits a clear maximum which decays steeply for $q > q_{\mathrm{opt}}$.

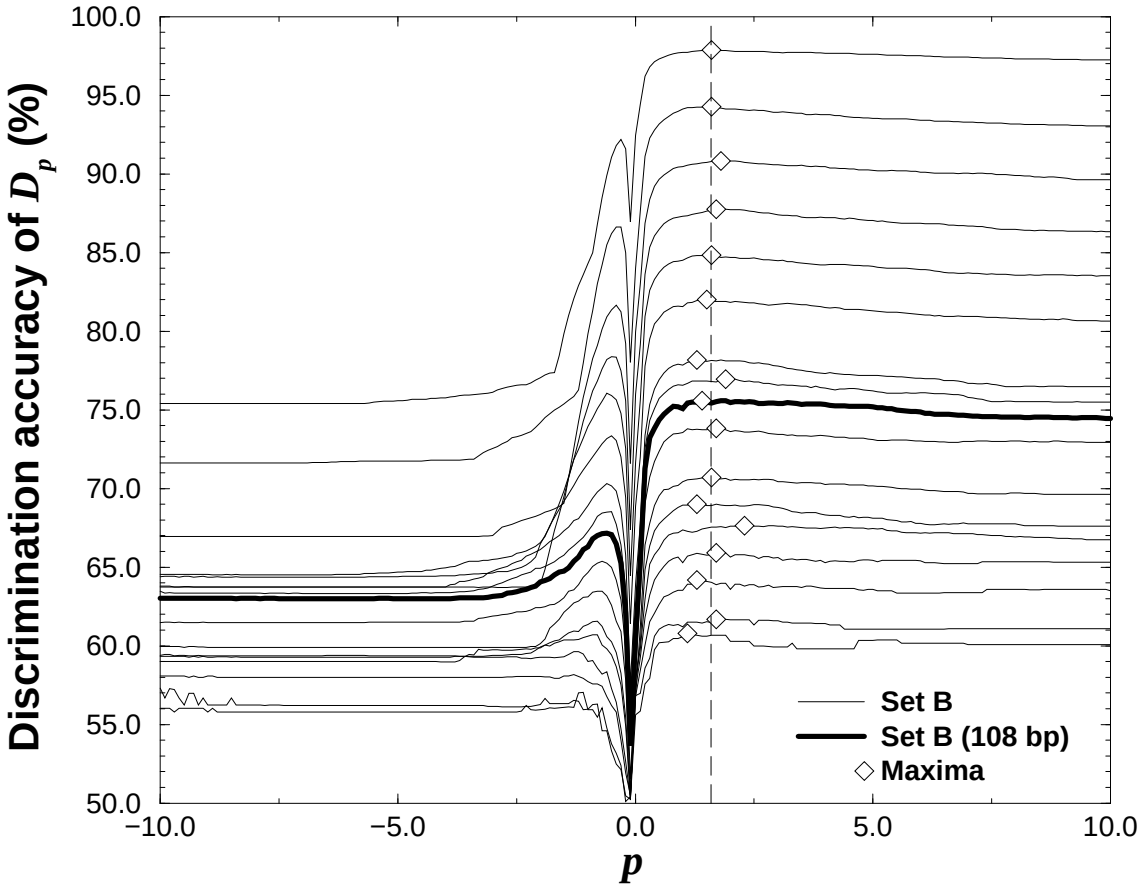


Figure 17.5: Accuracy of D_p as a function of the parameter p for different window lengths (from bottom to top the lengths read: 27, 30, 36, 45, 54, 60, 72, 90, 108, 120, 135,180, 216, 270, 360, 540, and 1080 bp). To guide the eye, the 108 bp length is graphed as a thick line. The value $p_{\rm opt}$ ($\diamond$) at which D_p distinguishes most accurately coding from versus non-coding DNA is relatively insensitive to the window length. The broken line indicates the mean value of all optimal p with $\langle p_{\rm opt}\rangle = 1.6$ and standard deviation $\Delta(p_{\rm opt}) = 0.3$ as estimated on all lengths shown. The fluctuations increase for small (< 60) windows. We note the shape of the accuracy spectrum (two maxima, distinct better performance for $p > 0$) remains overall unchanged when varying the window size.

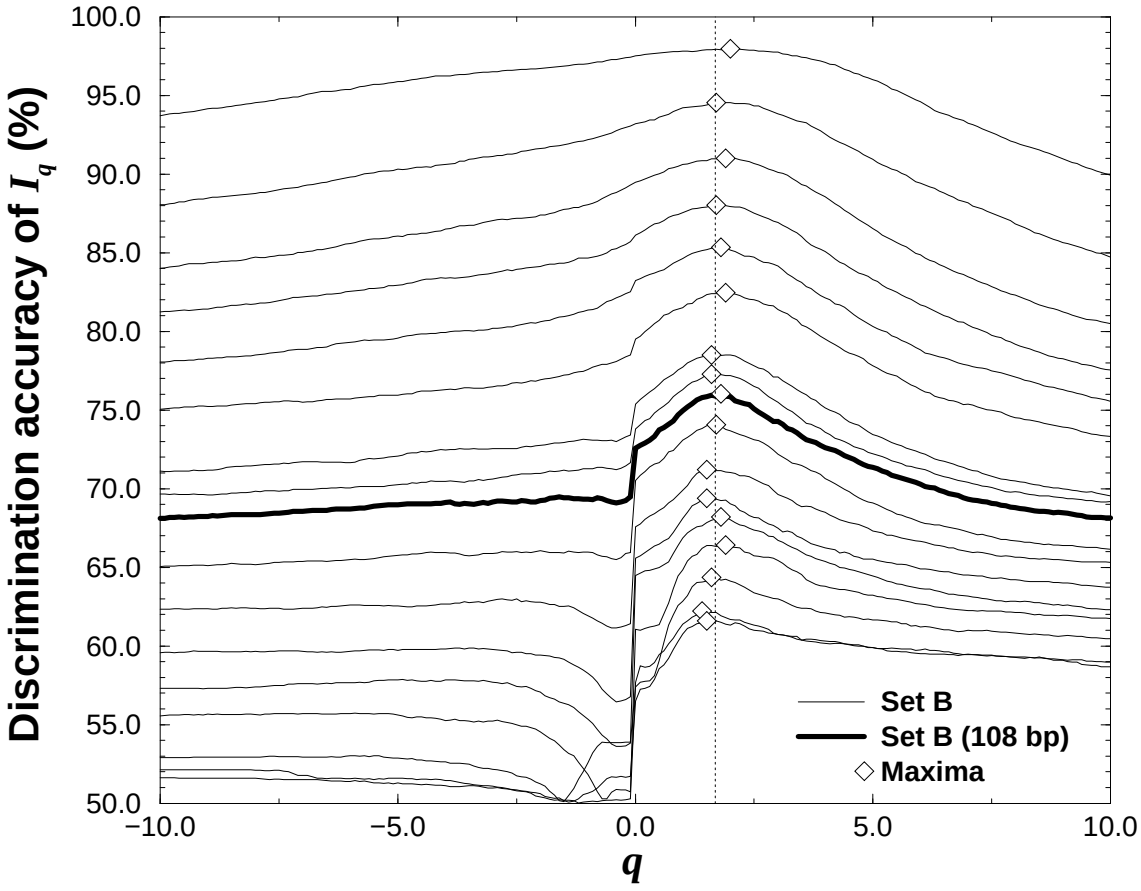


Figure 17.6: Accuracy of I_q as a function of the parameter q for different window lengths for set B (as described in Figure 5). The value $q_{\rm opt}$ ($\diamond$) at which I_q distinguishes most accurately coding versus non-coding DNA is approximately length-independent. The mean value of all optimal q results to $\langle q_{\rm opt}\rangle = 1.7$ and standard deviation to $\Delta(q_{\rm opt}) = (0.2)$. We observe that for increasing length the sharp profile flattens out and becomes a unimodal function of q.

Tab. 1. Performance of coding measures - human DNA

a. Coding measures based on positional dependence of nucleotide frequencies

	54 bp	108 bp	162 bp	54 bp	108 bp	162 bp	54 bp	108 bp	162 bp
A	68.9	75.9	79.6	68.3	75.1	79.3	66.9	74.2	79.3
D_1	69.9	76.6	80.9	69.4	76.3	79.8	68.0	75.1	79.4
D_2	70.2	76.8	80.8	70.0	76.6	80.1	68.0	75.5	80.5
D_3	69.8	76.7	80.4	69.9	76.7	80.3	68.1	75.3	80.3
D_4	69.5	76.4	80.1	69.5	76.3	80.0	67.9	75.1	80.0
$\bar{I}$	69.7	76.4	80.6	69.6	76.1	80.1	67.6	75.2	80.3
I_1	69.2	76.6	80.7	69.0	75.9	80.0	67.1	75.1	80.2
I_2	70.6	77.2	81.1	70.2	76.9	80.6	69.1	76.2	80.9
I_3	69.6	76.6	80.4	68.9	75.2	79.2	68.4	75.3	79.9
I_4	68.7	75.1	78.8	67.9	73.6	77.2	67.6	73.9	78.2

b. Most accurate coding measures

Coding measure	Number of input parameters	Set A_{test}		
		54 bp	108 bp	162 bp
Hexamer	4096	70.5	73.1	74.2
Position asymmetry	12	70.2	76.6	80.6
Dicodon usage	4096	70.2	72.9	73.9
Fourier	8	69.9	76.5	80.8
Hexamer-1	4096	69.9	72.6	73.8
Hexamer-2	4096	69.9	72.6	73.8
Run	6	66.6	70.3	71.3
Codon usage	64	65.2	68.0	69.5

Comparison of the generalized coding measures, D_p and I_q with currently used algorithms. Part (a) shows the accuracy as obtained by applying current (A, $\bar{I}$) and generalized (D_p, I_q) coding measures evaluated on the benchmark test of Fickett & Tung (1992) for three different window sizes. Since these measures require no prior training, we also apply them to the training set. Part (b) shows the percentage average of correctly predicted coding and non-coding DNA regions for 8 coding measures evaluated in Fickett & Tung (1992) as the most accurate discriminator of coding versus non-coding windows of lengths equal

to 54, 108, and 162 bp. The corresponding number of parameters is given. Note that for optimum parameters, p_{opt} and q_{opt}, D_p and I_q are as accurate as conventional measures (after prior training).

Tab. 2. Statistical dependences on the A+T content

a. Correlation coefficient

Data set	A	D_1	D_2	D_3	D_4	$\bar{I}$	I_1	I_2	I_3	I_4
$A_{training}$	-0.39	-0.31	-0.31	-0.31	-0.31	-0.35	-0.36	-0.20	-0.08	0.00
	0.01	-0.02	-0.03	-0.03	-0.03	0.00	0.00	-0.04	-0.05	-0.06
A_{test}	-0.42	-0.33	-0.35	-0.36	-0.35	-0.37	-0.38	-0.22	-0.07	-0.02
	-0.03	-0.05	-0.05	-0.05	-0.05	-0.04	-0.04	-0.05	-0.03	-0.02
B	-0.23	-0.20	-0.20	-0.20	-0.20	-0.21	-0.22	-0.14	-0.07	-0.03
	0.03	-0.04	-0.04	-0.04	-0.04	-0.01	0.00	-0.13	-0.20	-0.23

b. Uncertainty coefficient

Data set	A	D_1	D_2	D_3	D_4	$\bar{I}$	I_1	I_2	I_3	I_4
$A_{training}$	0.06	0.04	0.04	0.04	0.04	0.05	0.05	0.03	0.03	0.03
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.04
A_{test}	0.06	0.05	0.05	0.05	0.05	0.05	0.05	0.03	0.02	0.02
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.04
B	0.02	0.02	0.02	0.02	0.02	0.02	0.00	0.01	0.02	0.02
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.04

Linear (a) and non-linear (b) statistical dependences of current and generalized coding measures on the A+T content evaluated for exons and introns of length 108 bp from $A_{training}$, A_{test}, and B. For each set we choose for each class a subset of 7000 sequences. We state results within 0.01 precision ($|C(X,Y)|$ or $|U(X,Y)| < 0.01$ are stated as zero), and the calculation of the uncertainty coefficient was performed by distributing the data on an array covered by $M = 4 \times 4$ bins. We note that all currently applied coding measures correlate only weakly to the A+T content of non-coding sequences, while they show clear anti-correlations to the A+T content of coding sequences. The statistical dependence of D_p remains approximately persistent for parameter values adjacent to $p_{\rm opt}$. Linear correlations between I_q and the A+T content of coding sequences decay, whereas non-linear correlations between I_q and the A+T content of non-coding sequences rise for $q > q_{\rm opt}$.

Chapter 18

Identification of Protein-Coding Regions in DNA Sequences using an Entropic Segmentation Method

In this chapter we present a new approach to the problem of the statistical identification of DNA coding regions. This approach has two features: (i) DNA sequences are described by using a 12-letter alphabet instead of the usual 4-letter alphabet, and (ii) compositional domains are used instead of moving windows. We find that this method is highly accurate in finding borders between coding and noncoding regions and requires no "prior training" on known data sets. We also obtain more accurate results than those obtained with moving windows in the discrimination of coding from noncoding DNA.

18.1 Introduction

With the availability of the complete genomes of great number of prokaryotic organisms and several eukaryotic ones (yeast, worm), the second phase of genome projects is starting: the functional genomics, i.e. the search for the different functional elements which make up the DNA sequences [McKusick 1998]. The computational identification of functional regions (genes, exons, promoters, enhancers, etc.) is an arduous task and the predictive accuracy of current methods, although adequate in genomes for which a significant fraction of genes are previously known [Burset & Guigó 1996], remains rather low when faced with the large anonymous genomic sequences now being generated by genome projects. So, new

strategies are needed.

A variety of information is used by current gene identification methods, as potential sequence signals involved in gene specification or sequence similarity database searches. In addition, and more importantly, at the core of all gene identification programs, always exist one or more coding measures (see [Fickett 1998, Guigó 1999] for reviews). These measures are typically computed on a moving window, which slides along the sequence. The window length and the step size become thus critical parameters for gene identification. Unfortunately, there do not exist criteria to choose appropriate values for these parameters. Whether a larger window is more or less statistically significant than a shorter one cannot be decided, and the same is true for the step. Consequently, fixing both parameters introduces an unavoidable subjectivity in the analysis.

18.2 Entropic Segmentation Algorithm

We propose here a new method for the computational recognition of protein coding regions in genomic sequences, based on the compositional segmentation of the sequence [Bernaola-Galván et al. 1996, Román-Roldán et al. 1998, Oliver et al. 1999]. The compositional domains obtained are homogeneous and are defined statistically, so they lack the arbitrariness of moving windows. Once domains are obtained, any of the existing standard methods of discriminating coding from noncoding DNA can be applied.

One of the most relevant and well-known statistical features of coding regions is the nonuniform codon usage [Grantham et al. 1981]. This means that inside coding regions not all triplets of nucleotides (called codons) occur with the same probability. In particular the probability of appearance of a nucleotide is different in each of the three positions of the triplets [Shepherd et al. 1981, Staden 1982, Fickett 1982, Herzel & Grosse 1995]. This may be due to the restrictions imposed by the genetic code and also probably to some kind of preferences in the synonymous codon usage, but no matter what its origin is, this feature is not present in noncoding DNA, so this property can be used to distinguish between coding and noncoding DNA. In fact, based on these differences, the first generation of gene prediction programs, designed to identify approximate locations of coding regions in genomic DNA were developed [Staden 1982, Fickett 1982].

Here we develop a segmentation algorithm based on a 12-symbol alphabet. We define the phase of position, i, of a nucleotide to the number $j = i \bmod 3$, where $j \in \{0, 1, 2\}$. So,

each of the nucleotides of the DNA sequences can be substituted by one of the following symbols: $\mathcal{A}_{12} = \{A_0, A_1, A_2,\ T_0, T_1, T_2,\ C_0, C_1, C_2,\ G_0, G_1, G_2\}$, where, for example, T_2 means that we have found a nucleotide T with phase $= 2$.

Our aim is to divide a DNA sequence into segments in such a way as to maximize the difference in composition between them, and where the composition is measured by the frequency vector related to this 12-symbol alphabet. We hope these segments will correspond to alternating coding and noncoding regions. The method we describe here is the same introduced in [Bernaola-Galván et al. 1996] with several improvements. This method has been used to define a measure of DNA sequence compositional complexity [Román-Roldán et al. 1998], to describe the complexity of several DNA sequence models [Bernaola-Galván et al. 1999], to study the compositional structure of the sixteen chromosomes of Yeast [Li et al. 1998] and, recently, to determine statistically the mobility edge of one-dimensional disordered materials [Carpena & Bernaola-Galván 1999].

To compute the difference in composition between two regions of DNA, in order to decide whether they are different domains or not we use the Jensen-Shannon measure (JS).

Consider a DNA sequence composed of symbols belonging to $\mathcal{A}_{12}$, and define the 12-symbol frequency vector $\mathcal{F} \equiv \{f_{\ell,j}\}$, where $\ell \in \{A, T, C, G\}$ and $j \in \{0, 1, 2\}$, where $f_{\ell,j}$ is the relative number of nucleotides of type ℓ with phase j. Given two sequences of lengths n_1 and n_2 with frequency vectors $\mathcal{F}_1$ and $\mathcal{F}_2$, the JS is defined as:

$$\mathrm{JS}\left(\mathcal{F}_1, \mathcal{F}_2\right) = \mathrm{H}\left(\mathcal{F}\right) - \left\{ \frac{n_1}{N}\mathrm{H}\left(\mathcal{F}_1\right) + \frac{n_2}{N}\mathrm{H}\left(\mathcal{F}_2\right) \right\}, \tag{18.1}$$

where $N = n_1 + n_2$, $\mathcal{F} = \frac{n_1}{N}\mathcal{F}_1 + \frac{n_2}{N}\mathcal{F}_2$ is the frequency vector of the entire sequence obtained concatenating both subsequences, and $\mathrm{H}\left(\mathcal{F}\right)$ is the Shannon entropy, defined by:

$$\mathrm{H}\left(\mathcal{F}\right) = -\sum_{\ell,i} f_{\ell,j} \log_2 f_{\ell,j}. \tag{18.2}$$

Among other interesting properties [Bernaola-Galván et al. 1996], JS is almost not affected by the different size of the sequences being compared.

18.3 Applications to DNA Sequences

To test the ability of JS to separate coding from noncoding DNA, we do the following control experiments. First we take a known coding DNA sequence and a known noncoding one,

concatenate them, and go along the resulting sequence with a moving pointer, computing JS for the subsequence at the left and the subsequence at the right of the pointer. The results are shown in Figure 18.1(a) (solid line). Note that the maximum is clearly obtained in the boundary between both regions (vertical dashed line). We also test the effect of inserting one and two nucleotides between the original sequences. This does not affect the global composition but changes the phase of the nucleotides of the second sequence and hence, the resulting frequency vector $\mathcal{F}$ of the right hand side sequence can be considerably different from the original. As can be seen, the maximum value of JS is again obtained in the boundary between the two subsequences and the values are very close to the ones obtained without shift. This is due to the fact that in noncoding DNA all three phases are almost indistinguishable.

Sometimes, especially in prokaryotic genomes, the coding regions are separated by a very small noncoding region, too small to be separately identified on an statistical basis. JS is able to distinguish such coding regions, provided they are in different phases, even if they are in consecutive subregions. The only drawback is that if the regions are in the same phase they would be identified as only one coding region, but they could be easily separated by other methods. To show this, we analyze in Figure 18.1(b) two coding DNA regions, following the same as in Figure 18.1(a). The solid line (the two regions are in phase) reach very low values and does not seem to present a maximum in the boundary of both regions (vertical dashed line). On the other hand, when we introduce a phase shift (dashed and dotted line) we obtain very high values of JS and the maximum is clearly reached in the vicinity of the boundary.

To partition a natural DNA sequence which, in general, will be composed of several coding and noncoding regions, we search for the partition that maximizes the compositional difference between segments, as measured by JS. If the number of such regions is large, the problem presents high complexity, so we can use the heuristic algorithm [Román-Roldán et al. 1998]. In brief the procedure works as follows: we move a sliding pointer along the sequence which divides at each position the sequence into two subsequences and we compute JS. We select the point at which JS reaches its maximum value ($\mathrm{JS}_{\max}$) and compute its statistical significance (see below). If this significance exceeds a given threshold s, then the sequence is cut at this point. Otherwise the sequence remains undivided. The procedure continues recursively for each of the two resulting subsequences created by each cut. Before a new cut is accepted, we check that the subsequences formed

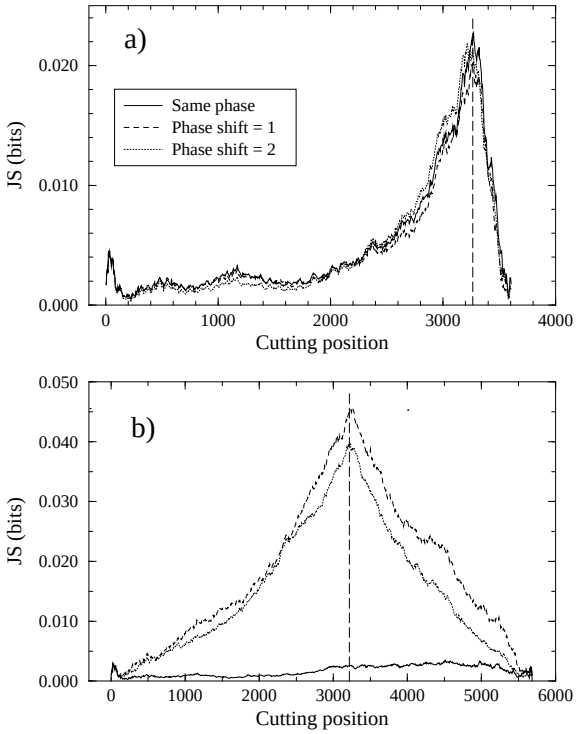


Figure 18.1: (a) JS vs. cutting position for a sequence obtained by joining a coding region (gene *carB* of bacteria *E. coli*, 3,222 bp long) and a noncoding region (intergenic region between genes *leuO* and *ilvI* of bacterium *E. coli*, 389 bp long); the dashed vertical line is the border between both regions. (b) JS vs. cutting position for a sequence obtained by joining two coding regions: genes *carB* (3,222 bp) and *polB* (2,463 bp) of the bacterium *E. coli*. The dashed vertical line is the boundary between the two regions.

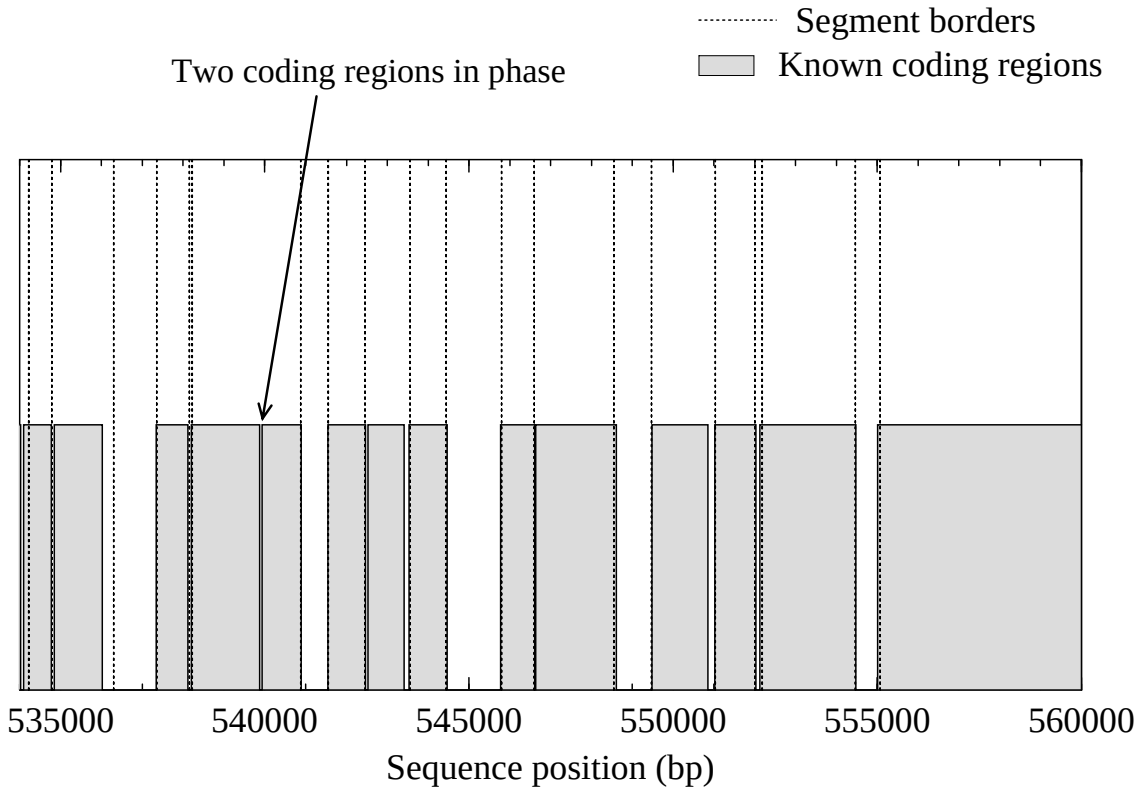


Figure 18.2: Comparison between the known coding regions of *Rickettsia* (shaded areas) and the cuts obtained at significance level $s = 99\%$ (dotted lines).

by the cut remain significantly different from their neighbors. The process stops when none of the possible cutting points has a significance level exceeding s. We say that such a sequence is segmented at the "s significance level."

The significance level $s_{\max}(x)$ of a possible cutting point with $\mathrm{JS}_{\max} = x$ is defined as the probability of obtaining this value or lower within a random sequence:

$$s_{\max}(x) = \mathrm{Prob}\left\{\mathrm{JS}_{\max} \leq x\right\}. \tag{18.3}$$

As $s_{\max}(x)$ does not seem to admit an easy analytical expression, we have obtained an approximation [Comment 12].

In Figure 18.2 we show the results of the segmentation of a region of the genome of the bacterium *Rickettsia prowazekii*. The shaded areas correspond to the coding regions obtained from annotations (GenBank acc. AJ235269 [Anderson et al. 1998]), and the vertical dotted lines are the positions of the cuts produced by the segmentation algorithm.

Note the good agreement between cuts and known coding region borders. Note also that, as can be inferred from Figure 18.1(b), the algorithm does not detect the border between two very close coding regions in the same phase (marked with an arrow in Figure 18.2).

In order to quantify the coincidence between cuts obtained using the segmentation algorithm and known borders between coding and noncoding regions, we introduce the following quantity:

$$D \equiv \frac{1}{2}\left[\sum_i \frac{\min_j |b_i - c_j|}{N_T} + \sum_j \frac{\min_i |b_i - c_j|}{N_T}\right], \tag{18.4}$$

where $\{b_i\}$ is the set of all borders between coding and noncoding regions, and $\{c_j\}$ is the set of all cuts produced by the segmentation, and N_T the total length of the sequence. The first summation measures the discrepancy between cuts and borders by adding for each real border the distance to the closest cut. The second summation performs the same operation but now including for each cut the distance to the closest real border. Both summations are needed to take into account not only the correctness in the position of the cuts (D would be zero just when cuts and borders coincide), but also the different number of borders and cuts. For instance, if the number of cuts is large, the first summation would be very small (because it would be easy to find a cut near any border) but the second summation would be very big. On the contrary, if the number of cuts is very small, the second summation would be also small (one has to sum just a few terms) but the first one would reach a very big value. D can be viewed as an average of the error in the determination of the correct boundaries between coding and noncoding regions, so $(1 - D)$ is a reasonable measure of the accuracy of the method.

Figure 18.3 plots $100(1 - D)$ for the segmentations of three bacterial complete genomes at several significance levels. The accuracy of the method is reasonably good (between 70–80%), especially since the method cannot separate adjacent phase-coding regions (see Figure 18.2).

18.4 Discussion

For the sake of comparison with other methods, we also include results obtained for the same bacterium with a sliding window, which moves along the sequence and, at each position, some discriminant function is evaluated [Comment 13]. The central nucleotide of the window is considered to be coding when the value of the discriminant function is above

a certain threshold, and noncoding when it is below. The positions where the discriminant function equals the threshold are proposed to be borders between coding and noncoding regions. The main problem with this method is the determination of the threshold: the only way to obtain it is to do a *training*, i.e. to analyze a sequence for which coding and noncoding regions are known and to choose the value which maximizes the number of matches for each window size.

In Figure 18.3(a) we also include the values of $100(1-D)$ obtained with the sliding window approach. These values are always clearly below those obtained using the segmentation algorithm. One advantage of our method is that the segmentation algorithm is not very sensitive to a change of significance level. In fact, any segmentation with a significance level within the range $90-95\%$ (the usual range) gives similar results. On the other hand, the choice of the window size seems to be critical, and the optimal values are different for each bacteria. At this point, it is important to recall that the segmentation method does not use any *a priori* biological information, i.e. it does not require previous training.

Up to now we have used the segmentation algorithm only to detect borders between coding and noncoding DNA, but there is no criterion to decide whether a domain is a coding or a noncoding region. In order to do this we can evaluate a discriminant function [Comment 13] and decide whether a segment is composed by coding or noncoding DNA. In Figure 18.3(b) we compare the accuracy of segmentation and moving windows approaches. Here the accuracy is defined [Burset & Guigó 1996] as:

$$1 - D^* \equiv \frac{1}{2}\left[\frac{tp}{p} + \frac{tn}{n}\right], \tag{18.5}$$

where tp (tn) is the number of nucleotides correctly identified as coding (noncoding) and p (n) is the total number of nucleotides identified as coding (noncoding). Again, the results for segments are better than those obtained with moving windows and the choice of significance level is much less critical than the choice of window length.

In conclusion, we have introduced a new method capable of locating borders between coding and noncoding regions without using any *a priori* biological information. In addition the domains obtained by means of the segmentation procedure improve the accuracy of the known discriminant functions.

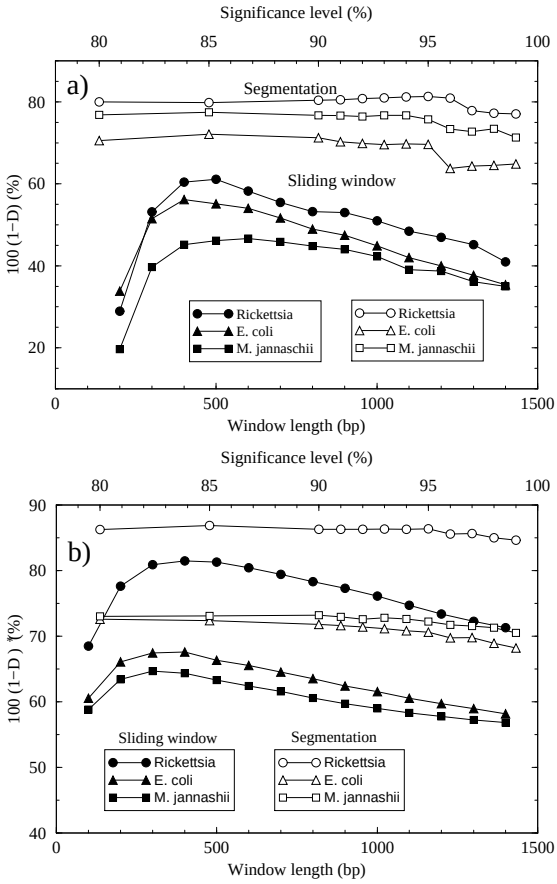


Figure 18.3: (a) Comparison of the accuracy of segmentation (open symbols) and sliding window (closed symbols) approaches in finding borders between coding and noncoding regions for three complete bacterial genomes: *Rickettsia prowazekii* (◯), *Escherichia coli* (△), and *Methanococcus jannaschii* (□); we find the best results when the training of the windows is carried out using the same sequence as the one analyzed. (b) Comparison of the accuracy of segmentation and sliding window approach in identifying coding DNA. The discriminant function and the training is the same as used in (a), the threshold value used with the segments is obtained by interpolating the values obtained for the moving windows.

Appendix A

Constructing a Basis of Linearly Independent Correlation Functions

In this Appendix, we ask whether the $\frac{\lambda\cdot(\lambda-1)}{2}$ free symmetric parameters S_{ij} $(i,j=1...\lambda-1, i\leq j)$ can be estimated from a certain set of $\frac{\lambda\cdot(\lambda-1)}{2}$ autocorrelation functions of sequences composed by λ symbols.

Let us consider autocorrelation functions where we assign the numbers 1 and 0 to all symbols $A_1...A_\lambda$. In our notation, $a_i \in \{0,1\}$ for all $i=1...\lambda$. In order to simplify further considerations for λ-ary sequences, we introduce the following denotations:

$$C^{(i)} \equiv C_{(0,...,0,a_i=1,0,...,0)} \tag{1.1}$$

$$C^{(i,j)} \equiv C_{(0,...,0,a_i=1,0,...,0,a_j=1,0,...,0)}. \tag{1.2}$$

In words, we assign the number 1 to the symbol A_i and the number 0 to all other symbols to define $C^{(i)}$. Analogously, we assign the number 1 to the symbols A_i and A_j and the number 0 to all other symbols to define $C^{(i,j)}$.

Now we use relation (19) to express the $\lambda-1$ autocorrelation functions $C^{(i)}$ for $i=1...\lambda-1$ and the $\binom{\lambda-1}{2}=\frac{(\lambda-1)\cdot(\lambda-2)}{2}$ autocorrelation functions $C^{(i,j)}$ for $i,j=1...\lambda-1, i<j$ in terms of S_{ij}.

$$C^{(i)} = S_{ii} \tag{1.3}$$

$$C^{(i,j)} = S_{ii} + S_{ij} + S_{ji} + S_{jj}. \tag{1.4}$$

This leads directly to the following expressions for the independent entries S_{ij}

$$S_{ij} = \frac{C^{(i,j)} - C^{(i)} - C^{(j)}}{2} \tag{1.5}$$

for all $i, j = 1...\lambda - 1$.

This means that all entries of the symmetric matrix $\hat{S}$ are unambiguously determined by the $(\lambda - 1) + \frac{(\lambda-1)\cdot(\lambda-2)}{2} = \frac{\lambda\cdot(\lambda-1)}{2}$ autocorrelation functions $C^{(i)}$ and $C^{(i,j)}$. Hence, this set of $\frac{\lambda\cdot(\lambda-1)}{2}$ autocorrelation functions can be regarded as a basis in the sense that all autocorrelation functions of λ-ary sequences are a linear combination of those $\frac{\lambda\cdot(\lambda-1)}{2}$ basic ones.

Appendix B

Jensen's Inequality

In this appendix, we will present Jensen's inequality, which can also be found in [Jaglom 1965] or [McEliece 1977]. In a second step, we will apply Jensen's inequality to the functions $f(x) = -\ln(x)$ and $g(x) = x \cdot \ln(x)$.

Definition B.1 *The real-valued function $f(x)$ is said to be* **convex** *on the interval I, if*

$$f(t \cdot x_1 + (1-t) \cdot x_2) \leq t \cdot f(x_1) + (1-t) \cdot f(x_2) \tag{2.1}$$

for all $x_1, x_2 \in I$ and all $t \in [0,1]$.

If strict equality holds whenever $x_1 \neq x_2$ and $0 < t < 1$, $f(x)$ is said to be **strictly convex** *on I.*

If f is sufficiently smooth, we can test f for convexity by calculus and obtain that f is convex iff $f''(x) \geq 0$ on I. If, furthermore, $f''(x) > 0$ except for a finite number of points, then f is strictly convex on I.

Theorem B.1 (Jensen's inequality) *Let X be a random variable and $F(X)$ its distribution function being concentrated on I. If the expectation value $E(X)$ exists and if $f(x)$ is a convex function on I, then Jensen's inequality says that*

$$E(f(X)) \geq f(E(X)). \tag{2.2}$$

Furthermore, if $f(x)$ is strictly convex, Jensen's inequality is strict unless X is concentrated at a single point x_0.

Geometrically, Jensen's inequality says that if a mass distribution is placed on the graph of the convex function $f(x)$, its resulting center of mass will lie above or on the graph of $f(x)$.

Let us now consider, for example, the function $f(x) = -\log(x)$, which is strictly convex for all positive x, since its second derivative $1/x^2$ exists and is positive for all $x > 0$.

Let $S = \{x_1, x_2, ..., x_M\}$ be a discrete set of real numbers, and let $p(x_i)$ a nonnegative function that obeys

$$\sum_{i=1}^{M} p(x_i) = 1. \tag{2.3}$$

Then S becomes a discrete sample space in the obvious way. Let $q(x_i)$ be any other nonnegative function defined on S, and define the random variable X by

$$X(x_i) \equiv \frac{q(x_i)}{p(x_i)}. \tag{2.4}$$

Since

$$E(\log(X)) = \sum_{i=1}^{M} p(x_i) \cdot \log\left(\frac{q(x_i)}{p(x_i)}\right) \tag{2.5}$$

and

$$E(X) = \sum_{i=1}^{M} q(x_i) \cdot \delta_{p(x_i),0}, \tag{2.6}$$

we obtain

$$-\sum_{i=1}^{M} p(x_i) \cdot \log(p(x_i)) \leq -\sum_{i=1}^{M} p(x_i) \cdot \log(q(x_i)) + \log(\alpha), \tag{2.7}$$

if α denotes the sum in $E(X)$, and we exploit Jensen's inequality

$$E(\log(X)) \leq \log(E(X)). \tag{2.8}$$

Furthermore, since $f(x) = -\log(x)$ is strictly convex, equality holds if, and only if, $q(x_i) = \alpha \cdot p(x_i)$ for all i such that $p(x_i) \neq 0$.

Hence, we can state the following theorem.

Theorem B.2 *Let I be a discrete set of integers, and let p_i be a set of positive real numbers such that*

$$\sum_{i \in I} p_i = 1. \tag{2.9}$$

If q_i is any other set of positive real numbers with

$$\sum_{i \in I} q_i = \alpha, \tag{2.10}$$

then

$$-\sum_{i\in I} p_i \cdot \log(p_i) \leq -\sum_{i\in I} p_i \cdot \log(q_i) + \log(\alpha), \tag{2.11}$$

with equality if, and only if, $q_i = \alpha \cdot p_i$ *for all* $i \in I$.

Let us finally display Jensen's inequality for $g(x) = x \cdot \log(x)$: since g is strict convex, we obtain

$$\sum_{i\in I} x_i \cdot \log(x_i) \cdot P(x_i) \geq p_i \cdot \log(p_i), \tag{2.12}$$

which will turn out to be very helpful in chapters 3, 9, and 10.

Appendix C

Steiner's Theorem and the Expectation Value of the Maximum Likelihood Variance Estimator

The task that we are going to solve in this section is the calculation of the expectation value of the maximum likelihood estimator of the variance of a normal population.

Following eq. (6.19)

$$\hat{\sigma^2} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \alpha)^2 - (\bar{x} - \alpha)^2 \tag{3.1}$$

for all real constants α.

By setting $\alpha = \mu$, we obtain an expression for the variance that is, in our examples, easier to calculate than the standard form for $\alpha = 0$.

The expectation value of the variance estimator is then

$$E\left(\hat{\sigma^2}\right) = E\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2\right) - E\left((\bar{x} - \mu)^2\right). \tag{3.2}$$

Since $E\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2\right) = \sigma^2$ by definition, we come up with

$$E\left(\hat{\sigma^2}\right) - \sigma^2 = -E\left((\bar{x} - \mu)^2\right). \tag{3.3}$$

In other words, the mean difference between the true sample variance and its maximum likelihood estimate is not zero but equal to the mean quadratic deviation of the estimated mean from the sample mean.

This is exactly Steiner's Theorem, which states that the momentum of inertia of a rigid body with respect to an arbitrary point μ is equal to its momentum of inertia with respect to its center of mass plus its mass multiplied by the squared distance between μ and its center of mass.

Our remaining task is now to calculate the mean quadratic fluctuations of $\bar{x}$ around μ drawn from a sample of size N of a normal population.

Since

$$E(\bar{x}) = E(\mu), \tag{3.4}$$

we can rewrite

$$E((\bar{x} - \mu)^2) = \sigma^2(\bar{x}), \tag{3.5}$$

which can be decomposed as

$$\sigma^2(\bar{x}) = \sigma^2\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) = \frac{1}{N^2}\sum_{i=1}^{N}\sigma^2(x_i) = \frac{1}{N}\cdot\sigma^2, \tag{3.6}$$

since the events x_i $(i = 1, 2, ..., N)$ are statistically independent.

Hence, we obtain from eqs. (6.19) and (3.6):

$$E\left(\hat{\sigma^2}\right) = \frac{N-1}{N}\cdot\sigma^2. \tag{3.7}$$

Appendix D

Maximum Likelihood Estimators

D.1 Maximum Likelihood Estimator of the Probability Vector in the Multinomial Case

In this section, we will derive the maximum likelihood estimator for the probability vector $\vec{p} = (p_1, p_2, ..., p_M)$ containing the probabilities p_i of an M-sided die to show up its face i. Let the sample size, i.e., the number of times we are rolling the die, be N. Then the likelihood to sample the sequence $x_1, x_2, ..., x_N$ is given by

$$L(x_1, x_2, ..., x_N; \vec{p}) = \prod_{j=1}^{M} p_j^{k_j} \tag{4.1}$$

with $\vec{k}$ being the vector containing the absolute frequencies k_j of rolling an j, i.e.,

$$\vec{k} \equiv (k_1, k_2, ..., k_M) \tag{4.2}$$

with

$$k_j \equiv \frac{1}{j} \sum_{i=1}^{N} \delta(x_i - j) \tag{4.3}$$

for all $j = 1, 2, ..., M$.

Recall that $x_i \equiv j$ if the die shows up its face j after getting rolled the i-th time and that $\delta(x - y)$ denotes the Kronecker symbol, which is 1 for $x = y$ and 0 otherwise.

Maximizing $L(x_1, x_2, ..., x_N; \vec{p}) = \prod_{j=1}^{M} p_j^{k_j}$ under the constraint $\sum_{j=1}^{M} p_j = 1$ leads to the following M equations with λ being the Lagrange multiplier belonging to the normalization

constraint:

$$\frac{\partial \left(\prod_{j=1}^{M} p_j^{k_j} - \lambda \cdot \left(\sum_{j=1}^{M} p_j - 1 \right) \right)}{\partial p_i} = 0 \tag{4.4}$$

for all $i = 1, 2, ..., M$.

This yields for $k_i > 0$ and $p_i > 0$

$$\frac{k_i}{p_i} \cdot \prod_{j=1}^{M} p_j^{k_j} - \lambda = 0 \tag{4.5}$$

or

$$\frac{p_i}{k_i} = \frac{\prod_{j=1}^{M} p_j^{k_j}}{\lambda} = C \tag{4.6}$$

where C is a constant for all $i = 1, 2, ..., M$.

Since

$$\sum_{i=1}^{M} p_i = 1 \tag{4.7}$$

and

$$\sum_{i=1}^{M} k_i = N, \tag{4.8}$$

we immediately obtain

$$C = \frac{1}{N} \tag{4.9}$$

and thus

$$\hat{p}_i = \frac{k_i}{N} \tag{4.10}$$

for all $i = 1, 2, ..., M$ and $k_i > 0$.

For $k_i = 0$, we realize that the likelihood $L(\vec{k}; \vec{p}) = L(x_1, x_2, ..., x_N; \vec{p})$ becomes maximal by estimating the corresponding $p_i = 0$.

Hence, the maximum likelihood estimator can be expressed in the following closed form for all k_i and $i = 1, 2, ..., M$:

$$\hat{p}_i = \frac{k_i}{N}, \tag{4.11}$$

which reflects the standard recipe to choose the relative frequency $\frac{k_i}{N}$ as an estimator of the probability p_i.

D.2 Maximum Likelihood Estimator of the Shannon Entropy

In this section, we will derive the maximum likelihood estimator for the Shannon Entropy $H(p) = -p \cdot \ln(p) - (1-p) \cdot \ln(1-p)$ of a coin with the probability p to show up its head and $1-p$ to show up its tail.

Let $x_i \in \{0, 1\}$ be the discrete random variables corresponding to the outcomes tail or head up in the i-th experiment of flipping the coin.

The likelihood to observe the sequence $x_1, x_2, ..., x_N$ is then

$$L(x_1, x_2, ..., x_N; p) = p^k \cdot (1-p)^{N-k} \tag{4.12}$$

with $k = \sum_{i=1}^{N} x_i$ being the absolute frequency of tossing a head.

We are, however, interested in the likelihood $L(x_1, x_2, ..., x_N; H)$ to produce a certain sequence $x_1, x_2, ..., x_N$ under the condition of a given H. Introducing the conditional probability $P(p|H)$ that a coin has the probability p provided its Shannon entropy is H, this likelihood can be rewritten as

$$\begin{aligned} & L(x_1, x_2, ..., x_N; H) \\ = & \int L(x_1, x_2, ..., x_N; p) \cdot P(p|H)\, dp & (4.13) \\ = & \frac{1}{2} \cdot (L(x_1, x_2, ..., x_N; p) + (L(x_1, x_2, ..., x_N; 1-p)) & (4.14) \\ = & \frac{1}{2} \cdot \left(p^k \cdot (1-p)^{N-k} + (1-p)^k \cdot p^{N-k} \right) & (4.15) \end{aligned}$$

since the function $H(p)$ is symmetric with respect to $p = 1/2$ and thus

$$H(p) = H(1-p). \tag{4.16}$$

Maximizing $2 \cdot L(x_1, x_2, ..., x_N; H)$ yields the equation

$$\begin{aligned} & k \cdot p^{k-1} \cdot (1-p)^{N-k} - (N-k) \cdot p^k \cdot (1-p)^{N-k-1} \\ - & k \cdot (1-p)^{k-1} \cdot p^{N-k} + (N-k)(1-p)^k \cdot p^{N-k-1} = 0 \end{aligned} \tag{4.17}$$

for the maximum likelihood estimator of $H = H(p)$.

At this point, we realize that the maximum likelihood estimator of the Shannon entropy, $\hat{H}(x_1, x_2, ..., x_N) = H(p(x_1, x_2, ..., x_N))$ with $p(x_1, x_2, ..., x_N)$ which can be obtained as the solution of eq. (4.17), is not equal to the so called natural estimator given by

$$\hat{H}(x_1, x_2, ..., x_N) = H\left(\frac{k(x_1, x_2, ..., x_N)}{N} \right). \tag{4.18}$$

Epilogue: Please realize that the solution of eq. (4.17) yields the maximum likelihood estimator for all symmetric functions $f(p)$, i.e., all functions for which the equality $f(p) = f(1-p)$ holds, since this is the only assumption we need to come from eq. (4.13) to eq. (4.14).

Hence, the maximum likelihood estimator of the variance σ^2 of the binomially distributed k, which is given by

$$\sigma^2 = \frac{p \cdot (1-p)}{N} \tag{4.19}$$

can be determined by solving eq. (4.17) and then calculating

$$\hat{\sigma^2} = \frac{p(x_1, x_2, ..., x_N) \cdot (1 - p(x_1, x_2, ..., x_N))}{N}. \tag{4.20}$$

Appendix E

Bayes Estimators

E.1 Laplace Estimator

In this section, we want to derive the Bayes estimator for the probability vector $\vec{p}$ of an M-sided die from a sample of size N under the prior assumption of a uniform distribution of $\vec{p}$.

Let us define the random variables x_i to be j if the outcome of the i-th experiment of rolling the die is that the die shows up its face j ($j = 1, 2, ..., M$).

Our task is to minimize the functional

$$\begin{aligned} F[\vec{f}(x_1, x_2, ..., x_N)] &\equiv \int\int \cdots \int \left(\vec{f}(x_1, x_2, ..., x_N) - \vec{p}\right)^2 \\ &\cdot \quad P(x_1, x_2, ..., x_N|\vec{p}) \cdot P(\vec{p})\, d\vec{p} \end{aligned} \tag{5.1}$$

where

$$\vec{p} = (p_1, p_2, ..., p_M) \tag{5.2}$$

and p_i is the probability of the die to show its face i.

The integral is taken over the entire parameter space of $\vec{p}$ which is here the $M-1$ dimensional simplex given by

$$p_i \geq 0 \tag{5.3}$$

for all $i = 1, 2, ..., M$ and

$$\sum_{i=1}^{M} p_i = 1. \tag{5.4}$$

Since $\vec{k} = (k_1, k_2, ..., k_M)$ with

$$k_j = \frac{1}{j} \sum_{i=1}^{N} x_i \cdot \delta(x_i - j), \tag{5.5}$$

and $\delta(x_i - j)$ being the Kronecker symbol defined by $\delta(x - y) = 1$ for $x = y$ and 0 otherwise is multinomially distributed, the conditional probabilities can be explicitly given by

$$P(x_1, x_2, ..., x_N|\vec{p}) = P(\vec{k}|\vec{p}) = \binom{N}{\vec{k}} \cdot \prod_{i=1}^{M} p_i^{k_i} \tag{5.6}$$

with $\binom{N}{\vec{k}}$ being the multinomial coefficient defined as

$$\binom{N}{\vec{k}} = \frac{N!}{k_1! \cdot k_2! \cdots k_M!}. \tag{5.7}$$

The vector $\vec{k}$, which contains the absolute frequencies k_j of the outcomes j of our experiment, turns out to be a better handle of describing our sample than the vector $(x_1, x_2, ..., x_N)$ and is thus preferred in the future.

Rewriting the scalar product in eq. (5.1) leads to

$$F[\vec{f}(\vec{k})] = \sum_{j=1}^{M} \int \int \cdots \int (f_j(\vec{k}) - p_j)^2 \cdot P(\vec{k}|\vec{p}) \cdot P(\vec{p})\, d\vec{p} \Rightarrow min, \tag{5.8}$$

which can be solved by

$$f_j(\vec{k}) = \frac{\int \int \cdots \int p_i \cdot P(\vec{k}|\vec{p}) \cdot P(\vec{p})\, d\vec{p}}{\int \int \cdots \int P(\vec{k}|\vec{p}) \cdot P(\vec{p})\, d\vec{p}} \tag{5.9}$$

$$= \frac{\int \int \cdots \int p_j \cdot \binom{N}{\vec{k}} \cdot \prod_{i=1}^{M} p_i^{k_i} \cdot P(\vec{p})\, d\vec{p}}{\int \int \cdots \int \binom{N}{\vec{k}} \cdot \prod_{i=1}^{M} p_i^{k_i} \cdot P(\vec{p})\, d\vec{p}} \tag{5.10}$$

if an individual minimizing of all summands in eq. (5.8) is compatible with the minimization of the sum.

Before we can try to calculate the integrals above, we have to specify our prior assumption about the probability density of the vector $\vec{p}$. Again, as described in section 6.5.2, we start with the weakest possible assumption about the $\vec{p}$, namely that we do not know anything about this vector except the p_j be normalized. Then, the maximum entropy

principle suggests the prior probability density be uniform on the simplex given in eqs. 5.3 and 5.4.

Under this assumption, which is also called the Bayes hypothesis, the Bayes estimator for the probability p_j of an M-sided die becomes

$$f_j(\vec{k}) = \frac{\int\int\cdots\int \binom{N}{\vec{k}} \cdot p_j \cdot \prod_{i=1}^{M} p_i^{k_i}\, d\vec{p}}{\int\int\cdots\int \binom{N}{\vec{k}} \cdot \prod_{i=1}^{M} p_i^{k_i}\, d\vec{p}}. \tag{5.11}$$

From appendix H.3, we obtain

$$\begin{aligned} &\int\int\cdots\int \binom{N}{\vec{k}} \cdot p_j \cdot \prod_{i=1}^{M} p_i^{k_i}\, d\vec{p} \\ = &\int_0^1 \frac{N!}{k_j! \cdot (N - k_j + M - 2)!} \cdot p_j^{k_j+1} \cdot (1-p_j)^{N-k_j+M-2}\, dp_j \end{aligned} \tag{5.12}$$

and

$$\begin{aligned} &\int\int\cdots\int \binom{N}{\vec{k}} \cdot \prod_{i=1}^{M} p_i^{k_i}\, d\vec{p} \\ = &\int_0^1 \frac{N!}{k_j! \cdot (N - k_j + M - 2)!} \cdot p_j^{k_j} \cdot (1-p_j)^{N-k_j+M-2}\, dp_j \end{aligned} \tag{5.13}$$

since

$$\sum_{i=1}^{M} p_i = 1 \tag{5.14}$$

and

$$\sum_{i=1}^{M} k_i = N. \tag{5.15}$$

As we show in appendix H.1,

$$\int_0^1 p_j^{k_j+1} \cdot (1-p_j)^{N-k_j+M-2}\, dp_j = \frac{(k_j+1)! \cdot (N - k_j + M - 2)!}{(N+M)!} \tag{5.16}$$

and

$$\int_0^1 p_j^{k_j} \cdot (1-p_j)^{N-k_j+M-2}\, dp_j = \frac{k_j! \cdot (N - k_j + M - 2)!}{(N+M-1)!}. \tag{5.17}$$

Hence,

$$f_j(\vec{k}) = \frac{\int\int\cdots\int p_j \cdot \prod_{i=1}^{M} p_i^{k_i} \cdot d\vec{p}}{\int\int\cdots\int \prod_{i=1}^{M} p_i^{k_i} \cdot d\vec{p}} = \frac{k_j + 1}{N + M}. \tag{5.18}$$

At this point, we realize that the individual minimization of the summands in eq. (5.8) reveals estimators for the probabilities p_j, which only depend on k_j but not on the remaining $M - 1$ components of the vector $\vec{k}$. Hence, the estimator

$$\vec{f}(\vec{k}) = (f_1(\vec{k}), f_2(\vec{k}), ..., f_M(\vec{k})) \tag{5.19}$$

is the Bayes estimator of the probability vector $\vec{p}$ of an M-sided die under the assumption of a uniform prior probability density of $\vec{p}$, which is also called the *Laplace estimator*.

E.2 Bayes Estimator of the Shannon Entropy - Part II

In chapter 7, we have derived the Bayes estimator of the Shannon entropy $H(\vec{p}) = -\sum_{i=1}^{M} p_i \cdot \ln(p_i)$ under the assumption of a uniform prior probability density on the simplex $\{\vec{p}\}$.

The same job was done in a brilliant, however, yet unpublished work by David Wolpert [1993], who calculated the appearing integrals by applying Laplace's convolution theorem and ended up with the result that the Bayes estimator of the Shannon entropy is given by the following equation:

$$\hat{H} = -\sum_{i=1}^{M} \frac{k_i + 1}{N + M} \cdot \Delta\Phi^{(1)}(k_i + 2, N + M + 1), \tag{5.20}$$

in which

$$\Delta\Phi^{(1)}(k_i + 2, N + M + 1) \equiv \Phi^{(1)}(k_i + 2) - \Phi^{(1)}(N + M + 1), \tag{5.21}$$

and

$$\Phi^{(1)}(z) \equiv \frac{\partial \ln(\Gamma(z))}{\partial z}, \tag{5.22}$$

where $\Gamma(z)$ is the gamma-function of the real positive numbers z.

In this section we will show that this finding and our result displayed in eq. (7.14) are indeed identical.

The scientific value that we see in this proof of consistency arises from being convinced that only a few people would like to implement a subroutine computing the logarithm of the

gamma-function and its derivative, but all programmers are certainly able to implement a subroutine calculating a finite harmonic sum.

Starting with the first poly-gamma-function, we obtain

$$\Phi^{(1)}(z) = \frac{\partial \ln(\Gamma(z))}{\partial z} = \frac{\Gamma'(z)}{\Gamma(z)} \tag{5.23}$$

$$= \int_0^1 \frac{1-t^{z-1}}{1-t}\,dt - C \tag{5.24}$$

by exploiting the Gaussian formula [Bronstein & Semendjajew 1989] and

$$C = \lim_{n\to\infty}\left(\sum_{k=1}^{n}\frac{1}{k} - \ln(n)\right) \tag{5.25}$$

defining Euler's constant.

For integer z,

$$\frac{1-t^{z+1}}{1-t} = t^z + t^{z-1} + ... + t + 1, \tag{5.26}$$

and hence,

$$\int_0^1 \frac{1-t^{z-1}}{1-t}\,dt = \int_0^1 t^{z-2} + t^{z-3} + ... + t + 1\,dt \tag{5.27}$$

$$= \frac{1}{z-1} + \frac{1}{z-2} + ... + \frac{1}{2} + 1 \tag{5.28}$$

for integer $z \geq 2$.

Thus we obtain

$$\Delta\Phi^{(1)}(k_i+2, N+M+1) = -\sum_{j=k_i+2}^{N+M}\frac{1}{j}, \tag{5.29}$$

which proves the desired identity.

E.3 Binary Rényi Entropy Estimator

Under the assumption of a uniform prior probability density, $Q(\vec{p}) = \text{const}$, the Bayes estimator of the binary Rényi entropy of order q can be written as

$$\widehat{K_q}(N_1, N_2) = \frac{1}{1-q}\frac{1}{W'(N_1, N_2)}\int_0^1 dp\; p^{N_1}(1-p)^{N_2}\log_2\left[p^q + (1-p)^q\right] \tag{5.30}$$

for all $N_1 + N_2 = N$. Using the normalization constant (8.38), we rewrite equation (5.30) in the form

$$\widehat{K_q}(N_1, N_2) = \frac{1}{\ln 2}\frac{1}{1-q}\frac{\Gamma(N+2)}{\Gamma(N_1+1)\Gamma(N_2+1)} \times \left\{ q \int_0^1 dp\; p^{N_1}\,(1-p)^{N_2} \ln p \right. \\ \left. + \int_0^1 dp\; p^N \left(\frac{1-p}{p}\right)^{N_2} \ln\left[1 + \left(\frac{1-p}{p}\right)^q\right] \right\}. \qquad (5.31)$$

The first term on the right hand side of (5.31) can be calculated to become

$$q \int_0^1 dp\; p^{N_1}\,(1-p)^{N_2} \ln p \qquad \equiv q\; \tfrac{\partial}{\partial N_1}\left(\textstyle\int_0^1 dp\; p^{N_1}\,(1-p)^{N_2}\right) \\ = q\; \tfrac{\partial}{\partial N_1} \mathrm{B}\,(N_1+1, N_2+1) \\ = -q\; \mathrm{B}\,(N_1+1, N_2+1)\left(\textstyle\sum_{l=N_1}^{N} \frac{1}{l+1}\right). \qquad (5.32)$$

In the remaining term in equation (5.31), we change the coordinate $x = (1-p)/p$ and thus arrive at

$$\widehat{K_q}(N_1, N_2) = \frac{1}{\ln 2}\frac{1}{1-q}\left(I_q\,(N_1, N_2) - q \sum_{l=N_1}^{N} \frac{1}{l+1} \right) \qquad (5.33)$$

where we define

$$I_q\,(N_1, N_2) = \frac{\Gamma(N+2)}{\Gamma(N_1+1)\Gamma(N_2+1)} \left\{ \int_0^\infty dx\; x^{N_2}\,[1+x]^{-(N+2)}\; \ln\,(1+x^q) \right\}. \qquad (5.34)$$

Appendix F

Bayes Estimators of Generalized Entropies

In this appendix, we present a generalized derivation of the Bayes estimator of generalized entropies. This derivation is more general than the derivation from chapter 8 in the sense that here we give up the restriction to a uniform a-priori probability density function. In the following we compute the Bayes estimator of Rényi and Tsallis entropies under the a-priori assumption of a Dirichlet probability density function.

F.1 Motivation

The demand made upon computational analysis of observed symbolic sequences has been increasing in the last decade. Here, the concept of entropy receives applications, and the generalizations according to Tsallis $H_q^{(\mathrm{T})}$ and Rényi $H_q^{(\mathrm{R})}$ provide whole-spectra of entropies characterized by an order q.

An enduring practical problem lies in the estimation of these entropies from observed data. The finite size of data sets can lead to serious systematic and statistical estimation errors. We focus on the problem of estimating generalized entropies from limited data samples and derive a Bayes estimator of the Tsallis entropy, $\mathcal{H}_q^{(\mathrm{T})}$, including the $(q = 1)$ Shannon entropy.

By extending our previous results on statistical entropy estimation of symbol sequences [Holste et al. 1998], we use a prior distribution over the probabilities which is of Dirichlet-type. Using the relationship between $H_q^{(\mathrm{T})}$ and $H_q^{(\mathrm{R})}$, we utilize the Bayes entropy estimator

$H_q^{(\mathrm{T})}$ to estimate the Rényi entropy $H_q^{(\mathrm{R})}$ from observed data. The Bayes estimator yields the smallest mean-squared deviation from the true parameter as compared with any other estimator. We compare the Bayes entropy estimators with the frequency-count estimators of $H_q^{(\mathrm{T})}$ and $H_q^{(\mathrm{R})}$. Numerical simulations reveal that the Bayes entropy estimator reduces statistical estimation errors of generalized entropies for statistical processes such as generated by higher-order Markov models.

F.2 Introduction

Building on the works of Shannon [Shannon 1948], generalized entropies have been extensively applied to characterize complex behavior in models and real systems. As the Shannon entropy, H, is formally defined as an average value, the idea underlying a generalization is to replace the average of logarithms by an average of powers. This gives rise to generalized Rényi and Tsallis entropies of order q, $H_q^{(\mathrm{R})}$ and $H_q^{(\mathrm{T})}$, respectively [Rényi 1970, Tsallis 1988, Curado & Tsallis 1991].

The order q applies to describe inhomogeneous structures of the probability distribution associated with the stochastic process under consideration. From both order-q entropies, $H_q^{(\mathrm{R})}$ and $H_q^{(\mathrm{R})}$, the Shannon entropy is obtained in the limit $q \to 1$. Applications of order-q entropies occur in a variety of fields of sciences like, *e.g.*, nonlinear dynamical systems [Beck & Schlögl 1993, Grassberger et al. 1991, Pompe 1993], statistical thermodynamics or evolutionary programming [Tsallis Entropies], and computational molecular biology [Amalgor 1985, Herzel et al. 1994b, Li 1997, Strait & Dewey 1996]. Here we address the problem of estimating generalized entropies from samples of observed data.

Entropy is based on probabilities which are, however, *a priori* unknown. Under the assumption of a stationary process, we consider data sets to be composed of N data points, each chosen from M possible different outcomes. In almost any practical situation the observer only obtains a snapshot of the stochastic process, and hence an enduring problem arises when entropies are to be estimated from limited samples. Replacing these probabilities by the sampled relative frequencies produces large statistical and systematic deviations of estimates from the true value [Harris 1975, Herzel 1988].

This problem becomes severe when the number of data points N is of the order of magnitude of the number of different states, $N \sim \mathcal{O}(M)$, which can occur *e.g.*, in practical estimations of correlations [Grassberger et al. 1991, Schürmann & Grassberger 1996], the vari-

ability in neural spike trains [Steveninck et al. 1997, Schneider et al. 1986], or in computational biology [Li 1997]. In those cases the choice of an estimator with small deviations from the true (though unknown) value becomes important. Several different estimators of the Shannon entropy have been developed [Ebeling et al. 1995, Grassberger 1988, Herzel 1988, Schmitt et al. 1993]. Specific estimators for the Rényi entropy and for the dimensions related to them have also been derived [Grassberger 1988, Pawelzik & Schuster 1987].

Using Bayes estimation, we derive an estimator of generalized entropies. The Bayes estimator possesses the optimal property to minimize the mean-squared deviation of the estimate from the true value, subject to a prior distribution. We derive the Bayes estimator of the order-q Tsallis entropy $H_q^{(\mathrm{T})}$ under the prior assumption of a Dirichlet distribution. We discuss properties of the estimator $H_q^{(\mathrm{T})}$ and contrast the result obtained with the usual frequency-count estimator. Using the relationship connecting $H_q^{(\mathrm{T})}$ with $H_q^{(\mathrm{R})}$, we propose a method on how to extract the Rényi entropy from observed data. We test the performance of the Bayes entropy estimator, by using both homogeneous and inhomogeneous Markov processes with zero- and five-step memories derived from biological sequence data of the prokaryote *H. influenzae*.

F.3 Generalized entropy analysis

In this section, we outline the concept of generalized entropies, and we add some remarks about their similarities and differences.

Consider a random variable A that assumes M different discrete values a_i, $i = 1...M$. Associated with A is a probability vector $\mathbf{P} = (p_1, \ldots, p_M)$ with components $p_i \equiv p(A = a_i)$. The p_i's are defined on a simplex comprised by $\{\mathbf{P} \mid \forall i : p_i > 0 \wedge \sum_i^M p_i = 1\}$. The set of all possible outcomes is referred to as the state space of size M. The Shannon entropy of A is defined as

$$H(\mathbf{P}) = -\langle \mathrm{ld} p_i \rangle_{\mathbf{P}} = -\sum_{i=1}^{M} p_i \ \mathrm{ld} p_i \tag{6.1}$$

where $\langle \cdot \rangle_{\mathbf{P}}$ is the average over $\mathbf{P}$. Due to the dual base of the logarithm (ld), the entropy is measured in units of bits per symbol. A distinctive property of Eqn. 6.1, which is not shared by generalized entropies, is that the entropy of a composite event can be given as the sum of the marginal and the conditional entropy.

It follows from Eqn. 6.1 that events having either a particularly high or low frequency of occurrence do not contribute much to the Shannon entropy. In order to weight particular

regions, one can consider the following partition function

$$Z_q(\mathbf{P}) = \langle p_i^{q-1} \rangle_{\mathbf{P}} = \sum_{i=1}^{M} p_i^q \quad \Big(q \in \mathcal{R}\Big). \tag{6.2}$$

In the spirit of statistical mechanics, by introducing the "energy" $E_i = -\mathrm{ld} p_i$ and identifying q as inverse temperature β, Z_q plays the same role as the partition function of a system which is embedded in a heath bath.

Varying q allows for the monitoring of the inhomogeneous structure of $\mathbf{P}$: for larger q, the larger individual probabilities p_i dominate Z_q. On the other hand, for smaller q, the smaller p_i become dominant. Note that $Z_0 = M$ and $Z_1 = 1$. Building upon the above partition function, the order-q Tsallis entropy is defined as [Tsallis 1988, Curado & Tsallis 1991]

$$H_q^{(\mathrm{T})}(\mathbf{P}) = \frac{1}{\ln 2} \frac{Z_q(A) - 1}{1 - q} = \frac{1}{\ln 2} \left\langle \frac{p_i^{q-1} - 1}{1 - q} \right\rangle_{\mathbf{P}}. \tag{6.3}$$

Similarly, the order-q Rényi entropy is given by [Rényi 1970]

$$H_q^{(\mathrm{R})}(\mathbf{P}) = \frac{\mathrm{ld} Z_q(A)}{1 - q} = -\mathrm{ld} \langle p_i^{q-1} \rangle_{\mathbf{P}}^{\frac{1}{q-1}}. \tag{6.4}$$

In the above expression, $\langle p_i^q \rangle_{\mathbf{P}}^{1/q}$ is the generalized q-average of the numbers p_i. By inspection, $H_q^{(\mathrm{T})}$ and $H_q^{(\mathrm{R})}$ are functionally related through

$$H_q^{(\mathrm{R})}(\mathbf{P}) \;=\; \frac{1}{1 - q} \ln \Big[1 + (1 - q) \ln 2 \; H_q^{(\mathrm{T})}(\mathbf{P})\Big] \tag{6.5}$$

and

$$H_q^{(\mathrm{T})}(\mathbf{P}) \;=\; \frac{1}{1 - q} \exp \Big[(1 - q) \ln 2 \; H_q^{(\mathrm{R})}(\mathbf{P})\Big] - 1. \tag{6.6}$$

We see that $H_q^{(\mathrm{T})}$ and $H_q^{(\mathrm{R})}$ are monotonic functions of one another for fixed q, and they posses the following properties:

- $H_q^{(\mathrm{T})}$ and $H_q^{(\mathrm{R})} \geq 0$. Global maxima (minima) are attained at $p_i = 1/M \;\forall i$ for $q > 0$ $(q < 0)$.
- $H_q^{(\mathrm{T})}$ and $H_q^{(\mathrm{R})}$ are monotonically decreasing functions of q for arbitrary $\mathbf{P}$.
- $H_q^{(\mathrm{T})}$ is a concave (convex) function of p_i, for any $q > 0$ $(q < 0)$. The curvature of $H_q^{(\mathrm{R})}$ upon q is nontrivial [Tsallis Entropies]. Yet we find two inequalities hold: $H_q^{(\mathrm{R})}$ is a convex (concave) function of $\mathbf{P}$ for $q < 0$ $(0 < q \leq 1)$.

- For two independent random variables, $H_q^{(R)}(A,B)$ is additive, *i.e.*, $H_q^{(R)}(\mathbf{P},\mathbf{Q}) = H_q^{(R)}(\mathbf{P}) + H_q^{(R)}(\mathbf{Q})$, and $H_q^{(\mathrm{T})}$ is pseudo-additive, *i.e.*, $H_q^{(T)}(\mathbf{P},\mathbf{Q}) = H_q^{(T)}(\mathbf{P}) + H_q^{(T)}(\mathbf{Q}) + (1\text{–q})H_q^{(T)}(\mathbf{P})H_q^{(T)}(\mathbf{Q})$.

The motivation to introduce generalized entropies stems from the advantage that through the characterizing order q we arrive at a whole-spectrum of entropies. Entropy spectra, in particular the differences between entropies of different order q contain information about the underlying process. In the light of the fact that (i) $H_q^{(\mathrm{R})}$ is additive but neither concave or convex, and (ii) that $H_q^{(\mathrm{T})}$ is convex (concave) but not additive, it is remarkable that yet by Eqn. 6.5 and 6.6 we are able to switch between two types of entropies of order q.

F.4 Bayes entropy estimation

In this section, we will first describe how to preprocess the necessary variables from sequence data and we outline the concept of Bayes estimation.

For a given sequence S_N of N observed data points, we define the vector $\mathbf{F} = \mathbf{N}/N$, which stores the relative frequencies of occurrence. Its ith component contains $f_i = N_i/N$, the ratio of the number of occurrences of symbol i and the total number of symbols in the sequence $N = \sum_{i=1}^{M} N_i$. We derive N_i independently from S_N such that $\mathbf{N}$ is multinomially distributed[1]

$$P(\mathbf{N}|\mathbf{P}) \quad = \quad \frac{1}{Z(\mathbf{N})}\, p_1^{N_1} \cdots p_M^{N_M} \tag{6.7}$$

with the normalization constant

$$Z(\mathbf{N}) \quad = \quad \frac{N_1! \cdots N_M!}{N!}$$

In Bayes estimation, the first quantity one would wish to write down is the posterior distribution for the vector $\mathbf{P}$ given the observed frequencies $\mathbf{N}$. For a given stochastic process under consideration, the individual components p_i of $\mathbf{P}$ always take on a specific value. Yet we can use a prior distribution $P(\mathbf{P}|\lambda\mathbf{U})$ as a conditioner to embody our (subjective) assumptions about $\mathbf{P}$. This allows for the direct inclusion of specific knowledge about the

[1] Note that $\mathbf{N}$ is multinomially distributed only in the case of independent identical-distributed data points. Otherwise, Eqn. 6.7 can only serve as an approximation which, however, is often the case in practical situations.

parameters p_i. We choose the p_i to obey a Dirichlet distribution [MacKay & Peto 1994], namely

$$P(\mathbf{P}|\lambda\mathbf{U}) = \frac{1}{Z(\lambda\mathbf{U})} p_1^{\lambda u_1 - 1} \cdots p_M^{\lambda u_M - 1} \tag{6.8}$$

with the normalization constant

$$Z(\lambda\mathbf{U}) = \frac{\Gamma(\lambda u_1) \cdots \Gamma(\lambda u_M)}{\Gamma(\lambda)}.$$

The vector $\mathbf{U} = (u_1, \ldots, u_M)$ is the expectation value of $P(\mathbf{P}|\lambda\mathbf{U})$, writing $\int \mathrm{d}\mathbf{P}\ P(\mathbf{P}|\lambda\mathbf{U})\mathbf{P} = \mathbf{U}$. It is positive in each component u_i, and it is normalized though the scalar $\lambda > 0$. In case that $\mathbf{U} = \mathbf{1}/\lambda$, the Dirichlet distribution simply reduces to the assumption of a homogeneous probability density on the simplex, $P(\mathbf{P}|\mathbf{1}) = \Gamma(M)$. The role of λ is that it measures how different we expect typical vectors $\mathbf{P}$ to be from the expectation $\langle \mathbf{P} \rangle_{P(\mathbf{P}|\lambda\mathbf{U})} = \mathbf{U}$.

We are now in a position to write down the complete distribution of $\mathbf{P}$, given the data $\mathbf{N}$ and the parameterization $\mathbf{U}$,

$$\begin{aligned} P(\mathbf{P}|\mathbf{N}, \lambda\mathbf{U}) &= \frac{P(\mathbf{N}|\mathbf{P})P(\mathbf{P}|\lambda\mathbf{U})}{P(\mathbf{N}|\lambda\mathbf{U})} \\ &= \frac{\prod_i^M p_i^{(\lambda u_i - 1 + N_i)} / Z(\mathbf{N}) / Z(\lambda\mathbf{U})}{P(\mathbf{N}|\lambda\mathbf{U})} \end{aligned} \tag{6.9}$$

where the normalization constant is obtained by integration over the simplex

$$P(\mathbf{N}|\lambda\mathbf{U}) = \frac{\int \mathrm{d}\mathbf{P} \prod_i^M p^{(\lambda u_i - 1 + N_i)}}{Z(\mathbf{N})Z(\lambda\mathbf{U})}.$$

In the above expression, $P(\mathbf{P}|\lambda\mathbf{U})$ is called the prior, $P(\mathbf{N}|\mathbf{P})$ is called the likelihood, the data-dependent term $P(\mathbf{N}|\lambda\mathbf{U})$ is called the evidence (how much the data $\mathbf{N}$ are in favor of $\lambda\mathbf{U}$), and $P(\mathbf{P}|\mathbf{N}, \lambda\mathbf{U})$ is called the posterior distribution. The latter quantity can be understood as summarizing what we know about $\mathbf{P}$ is what we knew before the experiment $[P(\mathbf{P}|\lambda\mathbf{U})]$ and what the observables conveyed us $P(\mathbf{N}|\mathbf{P})$.

The **Bayes entropy estimator** [Berger 1985, Wolpert & Wolf 1995], which emerges as the expectation value of $H_q^{(\mathrm{T})}$ over the posterior distribution, $\mathcal{H}_q^{(\mathrm{T})} = \langle H_q^{(T)} \rangle_{P(\mathbf{P}|\mathbf{N},\lambda\mathbf{U})}$, minimizes the mean-squared deviations of its estimates from the true value $H_q^{(\mathrm{T})}$, constraint to a prior distribution. Inserting Eqn. 6.8, we have

$$\begin{aligned} \mathcal{H}_q^{(\mathrm{T})}(\mathbf{N}, \lambda\mathbf{U}) &= \int \mathrm{d}\mathbf{P}\ H_q^{(\mathrm{T})}(\mathbf{P}) P(\mathbf{P}|\mathbf{N}, \lambda\mathbf{U}) \\ &= \frac{\int \mathrm{d}\mathbf{P}\ H_q^{(\mathrm{T})}(\mathbf{P}) P(\mathbf{N}|\mathbf{P}) P(\mathbf{P}|\lambda\mathbf{U})}{\int \mathrm{d}\mathbf{P}\ P(\mathbf{N}|\mathbf{P}) P(\mathbf{P}|\lambda\mathbf{U})}. \end{aligned} \tag{6.10}$$

In order to obtain $\mathcal{H}_q^{(\mathrm{T})}$, we note that it suffices to calculate the Bayes estimator of the partition function, $\mathcal{Z}_q$, writing

$$\mathcal{H}_q^{(\mathrm{T})}(\mathbf{N},\lambda\mathbf{U}) \;=\; \frac{1}{\ln 2}\frac{1}{1-q}\Big[\mathcal{Z}_q(\mathbf{N},\lambda\mathbf{U})-1\Big]. \tag{6.11}$$

Evaluating expression 6.11 yields [Holste et al. 1998]

$$\mathcal{Z}_q(\mathbf{N},\lambda\mathbf{U}) \;=\; \sum_i^M \left\{\frac{\Gamma(r_i+1+q)}{\Gamma(r_i+1)}\frac{\Gamma(R+M)}{\Gamma(R+M+q)}\right\}$$

for all $r_i \in \mathcal{R}$, $r_i = \lambda u_i - 1 + N_i$ and $R = \sum_i^M r_i = \lambda - M + N$. Note that the parameter interval for which $\mathcal{Z}_q$ is defined is $q \in (-U_{\min},\infty)$ with $U_{\min} = \min_{\forall i}\{u_i\}$. If all r_i are integers, the Γ-functions simplify to their factorial representation. For instance, for $q = 1$ and $u_i = 1/M$ the Bayes estimation of the probability is given by $(N_i + 1/M)/(N + 1)$. According to Schürmann & Grassberger [Schürmann & Grassberger 1996], this is the estimator found numerically to yield 'best' entropy estimates when inserted as an estimator for the individual probability p_i. Here we realize that it corresponds to choose a homogeneous Dirichlet distribution over the simplex.

As $H_q^{(\mathrm{T})}$ is a generalization of the Shannon entropy, the same is expected to hold for the Bayes entropy estimator of H. For integer r_i, the limit $q \to 1$ yields

$$\widehat{H_1}(\mathbf{N},\lambda\mathbf{U}) \;=\; \frac{1}{\ln 2}\sum_i^M \frac{r_i+1}{R+M}\left(\sum_{n=r_i+2}^{R+M}\frac{1}{n}\right). \tag{6.12}$$

For $\mathbf{U} = \mathbf{1}/\lambda$, this is identical with the result derived in [Grosse 1996, Wolpert & Wolf 1995].

We now turn to the Bayes estimator of the Rényi entropy $\mathcal{H}_q^{(\mathrm{R})}$. Substituting $H_q^{(\mathrm{R})}$ for $H_q^{(\mathrm{T})}$ in Eqn. 6.10, the problem of deriving the Bayes entropy estimator reduces to the calculation of the integral $\mathcal{H}_q^{(\mathrm{R})}(\mathbf{N},\lambda\mathbf{U}) \propto \int \mathrm{d}\mathbf{P}\ \mathrm{ld}Z_q(\mathbf{P})P(\mathbf{N}|\mathbf{P})P(\mathbf{P}|\lambda\mathbf{U})$. Even in the simplest case, namely for a 2-letter alphabet, no analytical expression could be derived. Therefore we seek another strategy which is of practical use, in particular in the multivariate case. We propose to use the closed-form result of $\mathcal{Z}_q$ for the estimation of $\mathcal{H}_q^{(\mathrm{R})}$, by making use of the relationship between $H_q^{(\mathrm{T})}$ and $H_q^{(\mathrm{R})}$. We first estimate $H_q^{(\mathrm{T})}$ from data, and then compute $H_q^{(\mathrm{R})}$ through Eqn. 6.5. After these preliminary remarks, we write down the (indirect) Bayes entropy estimator of $\mathcal{H}_q^{(\mathrm{R})}$

$$\mathcal{H}_q^{(\mathrm{R})}(\mathbf{N},\lambda\mathbf{U}) \;=\; \frac{1}{1-q}\ \mathrm{ld}\mathcal{Z}_q(\mathbf{N},\lambda\mathbf{U}). \tag{6.13}$$

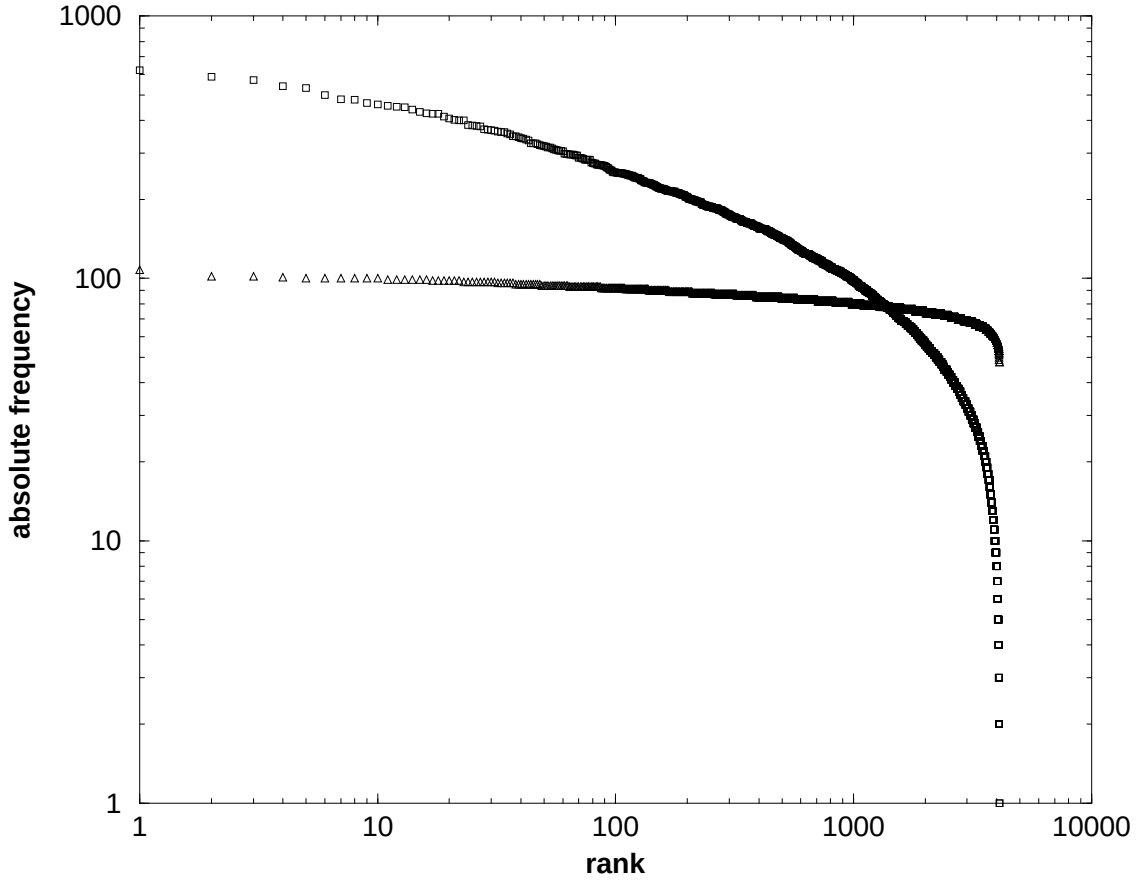


Figure F.1: The rank-ordered hexamer (6-letter) distribution of the complete DNA sequence of *H. influenzae* displayed as a double-logarithmic plot (□). For a comparison, the rank ordered distribution of a corresponding Bernoulli sequence of same length has been included in the figure (△).

F.5 Numerical Tests

In this section, we test the performance of the Bayes entropy estimators. We generate by Monte-Carlo simulations synthetic sequence data and investigate the statistical properties of $H_q^{(\mathrm{T})}$ and $H_q^{(\mathrm{R})}$ for both Bayes entropy estimators and frequency-count estimators.

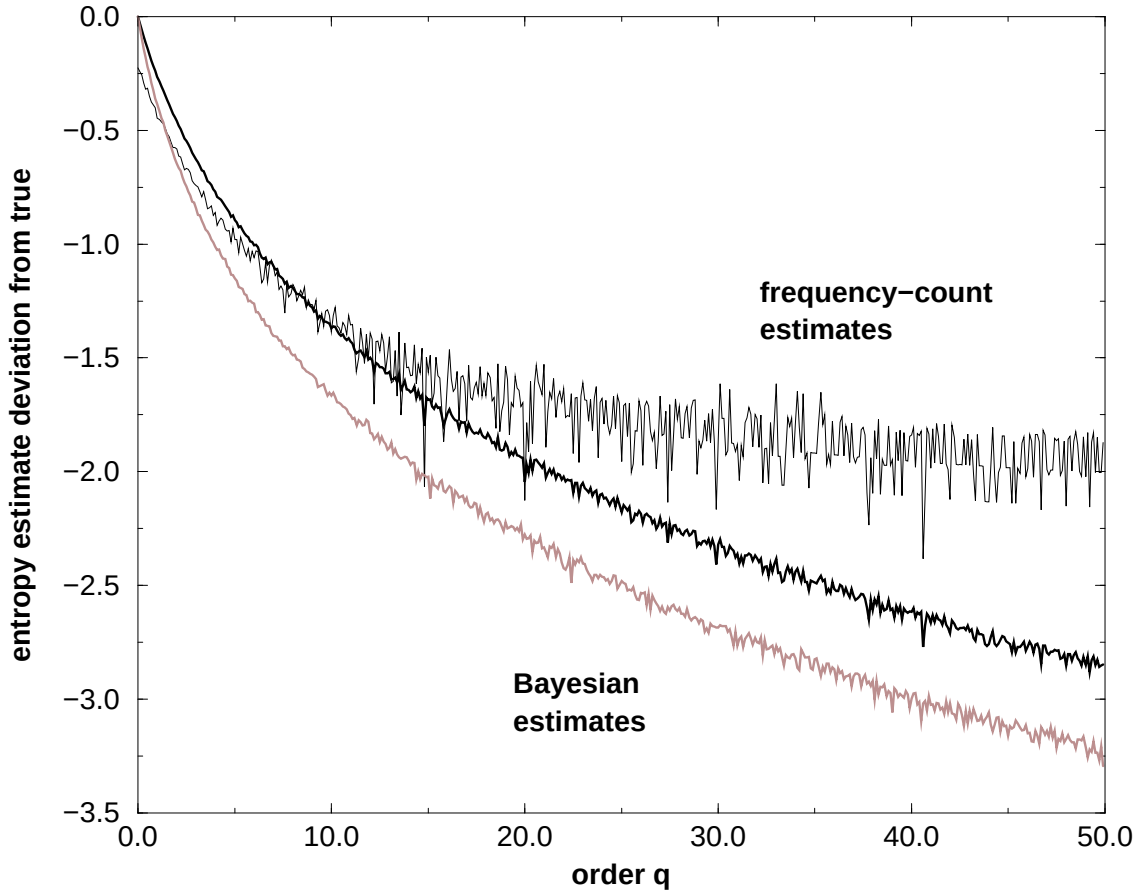


Figure F.2: Comparison of entropy estimators: $H_q^{(\mathrm{R})}$ with parameter set $M = 4096$, $N = 8000$, $\mathbf{P} = \mathbf{1}/M$. We observe the smaller variance of the Bayes entropy estimator (thick line, the black curve corresponds to $\mathbf{U} = \mathbf{2}/\lambda$ and the grey curve corresponds to $\mathbf{U} = \mathbf{1}/\lambda$) as compared with the frequency-count estimator (thin line) for a single realization for each order q. We observe the significant small width of the variance of the Bayes entropy estimator as compared with the frequency-count estimator.

We focus on the case where the size of the alphabet M is in the order of the number

of data points N, and we investigate the sample variance of the entropy estimators[2], by performing the following simulations:

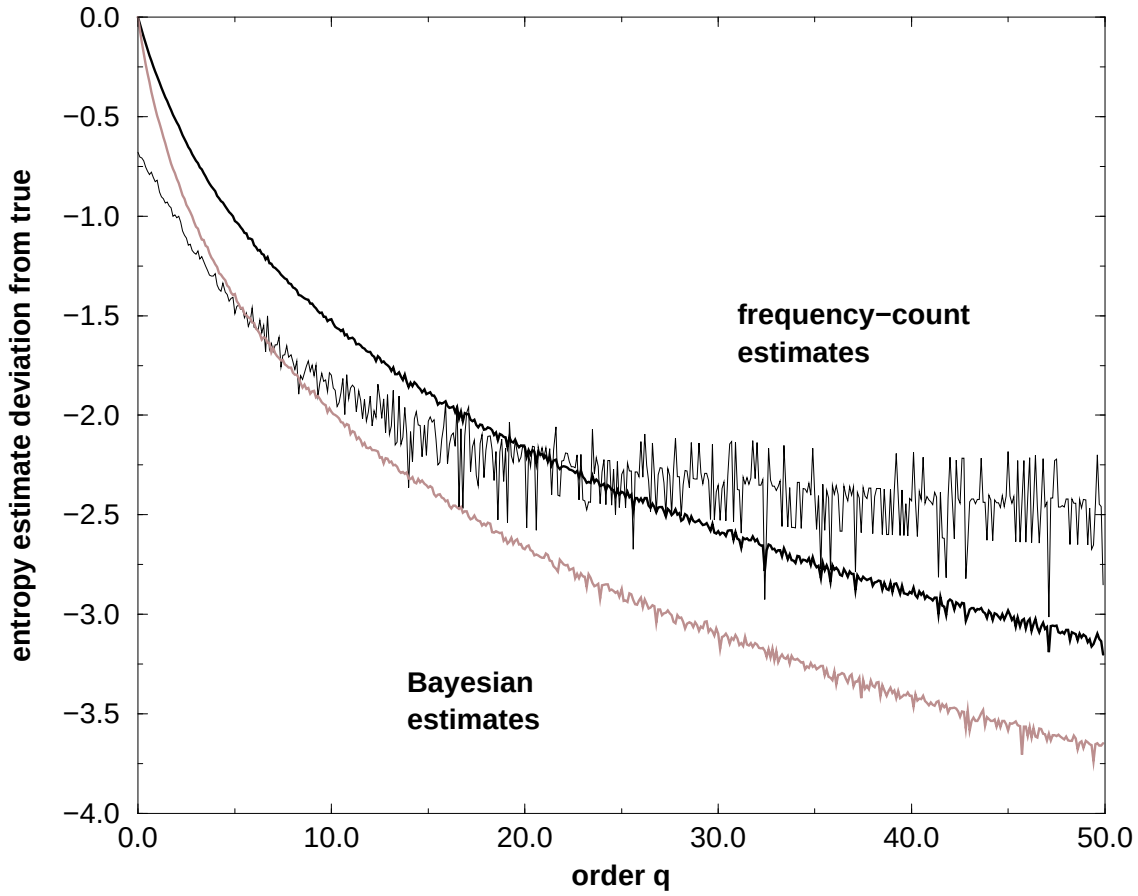


Figure F.3: Comparison of entropy estimates: $H_q^{(\mathrm{R})}$ with parameter set $M = 4096$, $N = 4000$ $\mathbf{P} = \mathbf{1}/M$, and $\mathbf{U} = \mathbf{2}/\lambda$. We observe that fluctuations of the Bayes entropy estimator (thick line, the black curve corresponds to $\mathbf{U} = \mathbf{2}/\lambda$ and the grey curve corresponds to $\mathbf{U} = \mathbf{1}/\lambda$) are strongly suppressed as compared with the frequency-count estimator (thin line) for a single realization for each order q. The significant small width of the variance of the Bayes entropy estimator as compared with the frequency-count estimator is visible.

1. We simulate homogeneous 0-step Markov processes. Considering the interval $q \in (0, 50)$ and incrementing q by $\Delta q = 0.1$, we monitor the estimates for a single re-

[2]Frequency-counts are directly obtained by sampling the relative frequencies f_i from sequence data. Replacing the individual probability p_i by their sampled relative frequency f_i, the frequency-count entropy estimator is respectively given by $\mathcal{H}_q^{(\mathrm{T})} = \big[Z_q(\mathbf{F}) - 1\big]/(1-q)$ and $\mathcal{H}_q^{(\mathrm{R})} = \big[\mathrm{ld} Z_q(\mathbf{F})\big]/(1-q)$.

alization for each q (and hence, allowing fluctuations to become visible rather them averaging them out), we study the observed variance of the Rényi entropy estimates and hence compare the Bayes entropy estimator with the frequency-count estimator. The outcome of the numerics is shown in Figure 2 and 3.

2. We simulate 0-step homogeneous and 5-step inhomogeneous memory Markov process. In the latter case, the transition probabilities are taken from the complete $1,830,240$ nucleotides long DNA sequence of the prokaryotic genome of *H. influenzae* [Fleischman et al. 1995]. The probability vector $\mathbf{P}_{H.influenzae}$, derived from the above DNA sequence is used to test the performance of the Bayes entropy estimator versus the frequency-counts estimator. In either case, $\mathbf{P} = \mathbf{1}/M$ or $\mathbf{P}_{H.influenzae}$, a sequence S_N is randomly generated from which we estimate the entropy values. In this test, we can also compute the 'true' hexamer entropies (taking the relative frequencies as probabilities by definition). The difference between the estimated and the theoretical values defines a random variable, which we define as the *entropy estimate deviation from true*. Generating an ensemble of sequences $\{S_N^{(i)}\}_{i=1}^{10,000}$ and estimating the order-2 entropies $H_2^{(\mathrm{R})}$ and $H_2^{(\mathrm{T})}$, we calculate the histograms of the entropy estimate distribution and compare the sample variance of both estimators. The outcome of the numerics is shown in Figure 4 and 5.

Figure 1 shows the rank-ordered distribution obtained from both the DNA sequence of *H. influenzae* and from a sequence of same length derived from a homogeneous 0-step Markov process. It can be seen that the DNA sequence is far more inhomogeneous than the realization of the latter process. The derived frequencies can be regarded as a typical example representing hexamer distributions in (prokaryotic) DNA.

These studies demonstrate the merit of the Bayes entropy estimator as compared with the frequency-count estimator in deriving reliable estimates from small samples: the variances of the Bayes estimates are significantly smaller than the variances of the frequency-count estimates for both homogeneous 0-step and inhomogeneous 5-step Markov processes.

F.6 Conclusions

We derived the Bayes estimator for generalized entropies, $H_q^{(\mathrm{T})}$ and $H_q^{(\mathrm{R})}$, for a discrete set of data points. Our approach of deriving the estimators $H_q^{(\mathrm{T})}$ and $H_q^{(\mathrm{R})}$ is motivated by

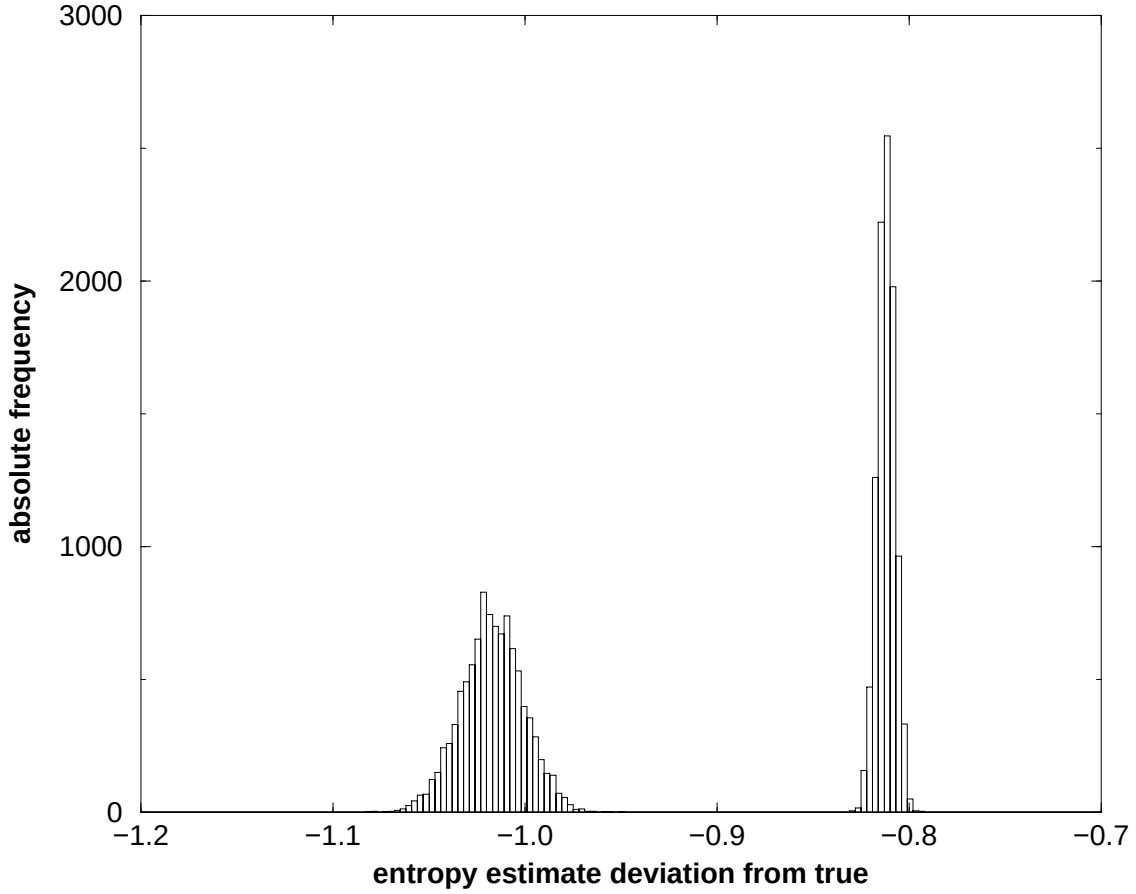


Figure F.4: Histograms of entropy estimates: $H_q^{(\mathrm{R})}$ with parameter set $M = 4096$, $N = 4000$, $\mathbf{P} = \mathbf{1}/M$, $\mathbf{U} = \mathbf{1}/\lambda$, and $q = 2$. We observe the smaller variance of the Bayes entropy estimator as compared with the frequency-count estimator.

the requirement to estimate generalized entropies from realizations where the total sample size N is of the order of magnitude of the size of the state space M. Both entropy estimators are obtained from the estimator of the partition function, Z_q, which could be used to estimate further related quantities like, *e.g.*, generalized dimensions. A comparative study

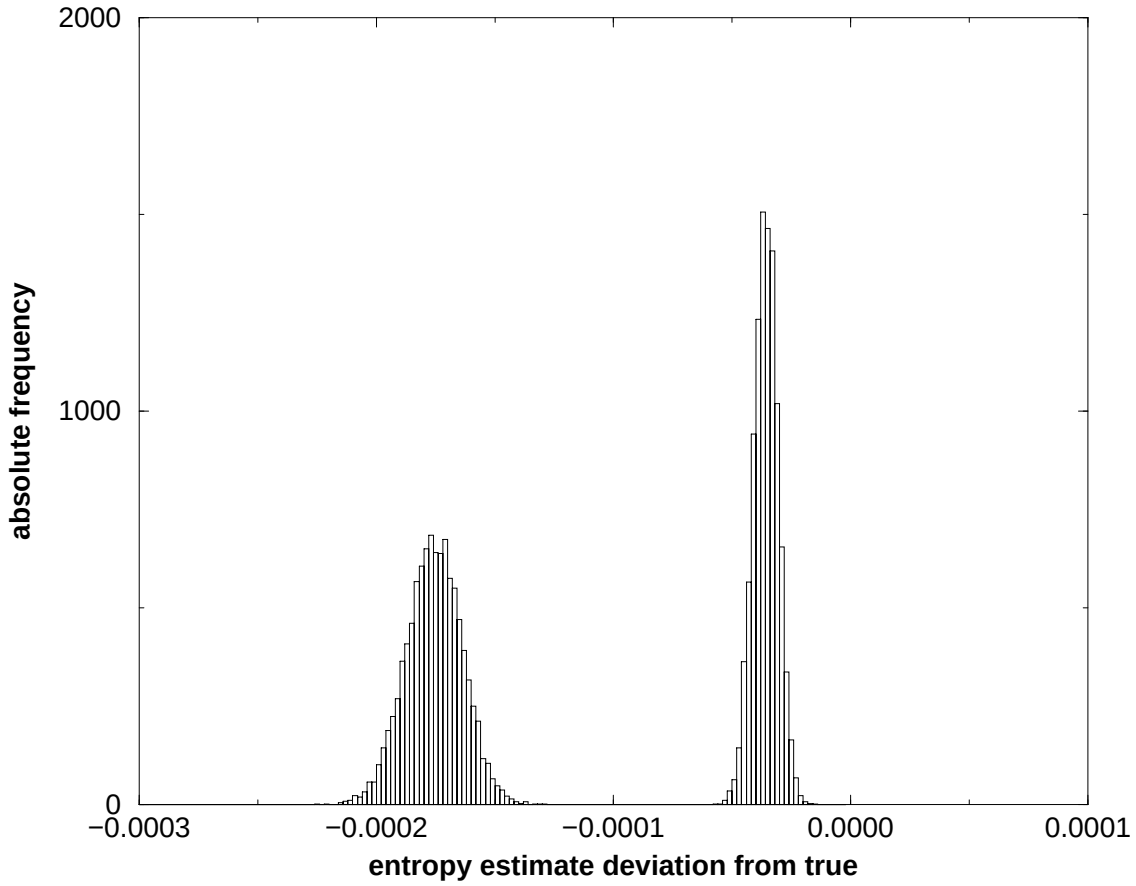


Figure F.5: Histograms of entropy estimates: $H_q^{(\mathrm{T})}$ with parameter set $M = 4096$, $N = 8000$, $\mathbf{P}_{H.influanzae}$, $\mathbf{U} = \mathbf{1}/\lambda$, and $q = 2$. We observe the smaller variance of the Bayes entropy estimator as compared with the frequency-count estimator.

of the accuracy by which the Bayes entropy estimator and the frequency-count estimators estimate $H_q^{(\mathrm{R})}$ and $H_2^{(\mathrm{T})}$ associated to 0-step homogeneous and 5-step inhomogeneous Markov models demonstrates the strength of the Bayes entropy estimator. Over the whole range of order q considered, the Bayes entropy estimator outperforms the frequency-count estimator by a significantly smaller variance of its estimates. This makes the Bayes entropy estimator appropriate to be applied in situations where the sample size N is of the order

of the size of the state space M.

The numerical simulations demonstrate that for the $\mathbf{P}$-vectors considered is this work, the Bayes entropy estimator leads to variances which are significantly smaller as compared to the variances of the frequency-counts estimators of generalized entropies extracted from small samples.

The Bayes entropy estimators have been derived under the assumption of a Dirichlet prior distribution. The specific parameterization ($\mathbf{U}$) is application-dependent. Given no further information about $\mathbf{P}$, a homogeneous distribution constitutes the least biased guess; this does not mean that the individual probabilities p_i are equidistributed, but rather that all vectors $\mathbf{P}$ are equiprobable. The central result in Eqn. 6.12 allows for the direct implementation of individual situation-tailored *a priori*-information.

Appendix G

Approximation of the Shannon Entropy $H(p)$ for all $p \in [0,1]$

In this section, we will present an alternative to the Taylor expansion of the Shannon entropy, and thus outline alternative approaches to calculating the expectation value, the variance, or even higher moments of the distribution of observed Shannon entropies.

The weak point of all Taylor approximations is, as we have seen in chapter 9, the divergence of the power series $\tilde{H}(\vec{x}) = \sum_{k=0}^{\infty} \sum_{i=1}^{M} a_k^{(i)} \cdot (x_i - p_i)^k$ for $x_i > 2 \cdot p_i$.

One possibility of circumventing these divergences is to compute the power series expansion of $H(\vec{x})$ about the point $\vec{x} = (\frac{1}{2}, \frac{1}{2}, ..., \frac{1}{2})$ or about points with components even greater than one half. However, this approximation is fairly bad, i.e., it converges only slowly to $H(\vec{x})$.

Hence, we present an approximation derived from the theory of linear regression that combines the following two features:

- it converges for all $x_i \in [0,1]$ and all $i = 1, 2, ..., M$

and

- it converges quickly to the limit function $H(\vec{x})$.

Imagine we are given a cloud of n points in the x-y-plane and are to determine a function

$$f(x; \theta_0, \theta_1, ..., \theta_m) = \sum_{k=0}^{m} \theta_k \cdot x^k \tag{7.1}$$

that optimally fits the points. Here, 'optimally' means that we search for those θ_k ($k = 0, 1, ..., m$) for which the sum of the squared deviations between the $f(x_i)$ and the y_i becomes minimal, i.e.,

$$\sum_{i=1}^{n} \left(f(x_i; \theta_0, \theta_1, ..., \theta_m) - y_i\right)^2 \quad \Rightarrow \quad min. \tag{7.2}$$

Now it is our goal to approximate the cloud consisting of an infinite number of points given by $y = -x \cdot \ln(x)$ with a uniform density on the interval $[0, 1]$ by a finite power series $\sum_{k=0}^{m} \theta_k \cdot x^k$.

We substitute the finite sums over all points by definite integrals and thus end up with the following analytic problem:

$$\int_0^1 \left(f(x; \theta_0, \theta_1, ..., \theta_m) - y(x)\right)^2 \, dx \quad \Rightarrow \quad min, \tag{7.3}$$

with

$$f(x; \theta_0, \theta_1, ..., \theta_m) = \sum_{k=0}^{m} \theta_k \cdot x^k \tag{7.4}$$

and

$$y(x) = -x \cdot \ln(x). \tag{7.5}$$

We can either try to solve this problem numerically, which would however shift our analytic problem to a numerical one by back-substituting the above integral by a finite sum, or to solve the integrals analytically, which would then reduce our analytic problem to an algebraic one.

The latter goal is indeed achievable, because closed form expressions for all appearing integrals can be derived.

$$\begin{aligned}
& \int_0^1 \left(f(x; \theta_0, \theta_1, ..., \theta_m) - y(x)\right)^2 \, dx \\
= \; & \int_0^1 f^2(x; \vec{\theta}) \, dx + 2 \int_0^1 f(x; \vec{\theta}) \cdot x \cdot \ln(x) \, dx + \int_0^1 x^2 \cdot \ln^2(x) \, dx \\
= \; & \sum_{i,j=0}^{m} \int_0^1 \theta_i \cdot \theta_j \cdot x^{i+j} \, dx + 2 \sum_{i=0}^{m} \int_0^1 \theta_i \cdot x^{i+1} \cdot \ln(x) \, dx \\
+ \; & \int_0^1 x^2 \cdot \ln^2(x) \, dx
\end{aligned}$$

$$= \sum_{i,j=0}^{m} A_{ij} \cdot \theta_i \cdot \theta_j + 2 \sum_{i=0}^{m} B_i \cdot \theta_i + C \tag{7.6}$$

by defining

$$A_{ij} \equiv \int_0^1 x^{i+j}\, dx = \frac{1}{i+j+1}, \tag{7.7}$$

$$B_i \equiv \int_0^1 x^{i+1} \cdot \ln(x)\, dx = J(i+1,0) = -\left(\frac{1}{i+2}\right)^2, \tag{7.8}$$

and

$$C \equiv \int_0^1 x^2 \cdot \ln^2(x)\, dx. \tag{7.9}$$

Minimizing

$$\sum_{i,j=0}^{m} A_{ij} \cdot \theta_i \cdot \theta_j + 2 \sum_{i=0}^{m} B_i \cdot \theta_i + C \tag{7.10}$$

leads to the following $m+1$ equations for all $k = 0, 1, ..., m$:

$$2 \cdot A_{kk} \cdot \theta_k + \sum_{i=0}^{m} (1 - \delta_{ik}) \cdot A_{ik} \cdot \theta_i + \sum_{j=0}^{m} (1 - \delta_{kj}) \cdot A_{kj} \cdot \theta_j + 2 \cdot B_k = 0 \tag{7.11}$$

Since $A_{ij} = A_{ji}$, we obtain

$$\sum_{i=0}^{m} A_{ik} \cdot \theta_i = -B_k \tag{7.12}$$

for all $k = 0, 1, ..., m$, which can be displayed in the following matrix representation of a system of linear equations:

$$\begin{pmatrix} \frac{1}{1} & \frac{1}{2} & \cdots & \frac{1}{m+1} \\ \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{m+2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{m+1} & \frac{1}{m+2} & \cdots & \frac{1}{2 \cdot m+1} \end{pmatrix} \cdot \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_m \end{pmatrix} = \begin{pmatrix} \frac{1}{2^2} \\ \frac{1}{3^2} \\ \vdots \\ \frac{1}{(m+2)^2} \end{pmatrix}. \tag{7.13}$$

Let $\hat{A}$ be the $(m+1) \times (m+1)$ matrix containing the elements $A_{ij} = \frac{1}{i+j+1}$, $\vec{\theta} = (\theta_0, \theta_1, ..., \theta_m)$ be the vector containing our regression parameters, and $\vec{B} = (B_0, B_1, ..., B_m)$ be the inhomogeneity vector of our system of linear equations. Then, our system

$$\hat{A} \cdot \vec{\theta}^T = \vec{B}^T \tag{7.14}$$

is uniquely solvable by

$$\vec{\theta}^T = \hat{A}^{-1} \cdot \vec{B}^T \tag{7.15}$$

if $\hat{A}$ is regular, i.e., if $\hat{A}^{-1}$ exists.

Our next task will thus be the inversion of the matrix $\hat{A}$ given by the elements $A_{ij} = \frac{1}{i+j+1}$ for $i, j = 0, 1, ..., m$.

Doing it by hand is no fun for $m \geq 5$. Hence, a numerical inversion of the matrix $\hat{A}$ seems to be the only alternative. But unfortunately, almost all programs fail to derive proper results for increasing m. The reason for this dilemma is fairly simple: the matrix $\hat{A}$, which is called the *Hilbert matrix*, is ill conditioned. This means that this matrix looks as if it were singular although it is indeed regular. In other words, all eigenvalues of this matrix are nonzero, but extremely small.

The question that we have not yet answered is whether the Hilbert matrix is indeed regular. In the following paragraph, we will prove the regularity by displaying the inverse matrix explicitly.

Theorem G.1 *Let $\hat{A}$ be the $n \times n$ Hilbert matrix defined by its elements*

$$A_{ij} = \frac{1}{i+j+1}. \tag{7.16}$$

Then the inverse matrix $\hat{A}^{-1}$ is given by the elements

$$\tilde{A}_{ij} = \frac{(-1)^{i+j} \cdot i^2 \cdot j^2}{(i+j-1) \cdot (i!)^2 \cdot (j!)^2 \cdot n^2} \cdot \prod_{k=1}^{i} (n^2 - (k-1)^2) \cdot \prod_{k=1}^{j} (n^2 - (k-1)^2). \tag{7.17}$$

Multiplying the matrix $\hat{A}^{-1}$ by the inhomogeneity vector $\vec{B}$ yields our desired vector $\vec{\theta}$ containing the coefficients of our power series approximating $y(x) = -x \cdot \ln(x)$.

This power series expansion can eventually be used to derive approximations (in any order) for the expectation value, the variance, or even higher moments of the Shannon entropy estimates distribution, since all moments of the multinomial distribution are known.

Appendix H

Useful Integrals and Sums

This section is devoted to the derivation of some analytic expressions for the following integrals, which we are frequently using throughout our work:

$$I(p;s,t) \equiv \int p^s \cdot (1-p)^t \, dp \tag{8.1}$$

and

$$J(s,t) \equiv \int_0^1 p^s \cdot (1-p)^t \cdot \ln(p) \, dp \tag{8.2}$$

for any $s \in \mathcal{N}$ and $t \in \mathcal{N}$ as well as

$$\begin{aligned} & K(p_1;k_j,M,N) \\ & \equiv \int_{p_2=0}^{1-p_1} \int_{p_3=0}^{1-p_1-p_2} \cdots \int_{p_{M-1}=0}^{1-p_1-p_2-\ldots-p_{M-2}} \binom{N}{\vec{k}} \cdot \prod_{i=2}^{M} p_i^{k_i} \, dp_{M-1} \cdots dp_3 \, dp_2. \end{aligned} \tag{8.3}$$

In section H.4, we will derive the marginal distribution of the component k_i ($i = 1,2,...,M$) of a multinomially distributed vector $\vec{k} = (k_1,k_2,...,k_M)$, which confronts us with the finite sum

$$S(p_1;k_1,M,N) \equiv \sum_{k_2=0}^{N-k_1} \sum_{k_3=0}^{N-k_1-k_2} \cdots \sum_{k_{M-1}=0}^{1-k_1-k_2-\ldots-k_{M-2}} \binom{N}{\vec{k}} \cdot \prod_{i=2}^{M} p_i^{k_i} \tag{8.4}$$

for $i = 1$.

H.1 The Indefinite Integral $I(p;s,t)$

The indefinite integral $I(p;s,t) = \int p^s \cdot (1-p)^t \, dp$ is obtainable by partial integration, which yields

$$\begin{aligned}
\int p^s \cdot (1-p)^t \, dp &= \frac{1}{s+1} \cdot p^{s+1} \cdot (1-p)^t + \frac{t}{s+1} \cdot \int p^{s+1} \cdot (1-p)^{t-1} \, dp \\
&= \frac{1}{s+1} \cdot p^{s+1} \cdot (1-p)^t + \frac{t}{(s+1)\cdot(s+2)} \cdot p^{s+2} \cdot (1-p)^{t-1} \\
&+ \frac{t \cdot (t-1)}{(s+1)\cdot(s+2)} \cdot \int p^{s+2} \cdot (1-p)^{t-2} \, dp \\
&= \ldots \\
&= \sum_{i=0}^{k-1} \frac{s! \cdot t!}{(s+i+1)! \cdot (t-i)!} \cdot p^{s+i+1} \cdot (1-p)^{t-i} \\
&+ \frac{s! \cdot t!}{(s+k)! \cdot (t-k)!} \int p^{s+k} \cdot (1-p)^{t-k} \, dp \qquad (8.5)
\end{aligned}$$

after k times partial integration for integer $k \leq t$.

For $k = t$, we obtain

$$\begin{aligned}
I(p;s,t) &= \int p^s \cdot (1-p)^t \, dp \\
&= \sum_{i=0}^{t-1} \frac{s! \cdot t!}{(s+i+1)! \cdot (t-i)!} \cdot p^{s+i+1} \cdot (1-p)^{t-i} \\
&+ \frac{s! \cdot t!}{(s+t+1)!} \cdot p^{s+t+1} \\
&= \sum_{i=0}^{t} \frac{s! \cdot t!}{(s+i+1)! \cdot (t-i)!} \cdot p^{s+i+1} \cdot (1-p)^{t-i} \qquad (8.6)
\end{aligned}$$

by defining

$$0^0 \equiv \lim_{p \to 1} (1-p)^0 = 1. \qquad (8.7)$$

Hence, the definite integral with limits 0 and 1 becomes

$$\int_0^1 p^s \cdot (1-p)^t \, dp = \frac{s! \cdot t!}{(s+t+1)!}. \qquad (8.8)$$

H.2 The Definite Integral $J(s,t)$

Since we have obtained an analytic expression for $I(p;s,t) = \int p^s \cdot (1-p)^t \, dt$, we can easily derive $J(s,t) = \int_0^1 p^s \cdot (1-p)^t \cdot \ln(p) \, dp$ by partial integration:

$$\int_0^1 p^s \cdot (1-p)^t \cdot \ln(p)\, dp = [I(p;s,t) \cdot \ln(p)]_0^1 - \int_0^1 I(p;s,t) \cdot \frac{1}{p}\, dp. \tag{8.9}$$

Since

$$\lim_{p \to 0} p^k \cdot \ln(p) = \lim_{p \to 1} p^k \cdot \ln(p) = 0 \tag{8.10}$$

for all $k \in \mathcal{N} \backslash \{0\}$,

$$[I(p;s,t) \cdot \ln(p)]_0^1 = 0. \tag{8.11}$$

The remaining term, $\int_0^1 I(p;s,t) \cdot \frac{1}{p}\, dp$, yields:

$$\begin{aligned}
\int_0^1 I(p;s,t) \cdot \frac{1}{p}\, dp &= \int_0^1 \sum_{i=0}^{t} \frac{s! \cdot t!}{(s+i+1)! \cdot (t-i)!} \cdot p^{s+i} \cdot (1-p)^{t-i}\, dp \\
&= \sum_{i=0}^{t} \frac{s! \cdot t!}{(s+i+1)! \cdot (t-i)!} \int_0^1 p^{s+i} \cdot (1-p)^{t-i}\, dp \\
&= \sum_{i=0}^{t} \frac{s! \cdot t!}{(s+i+1)! \cdot (t-i)!} \cdot \frac{(s+i)! \cdot (t-i)!}{(s+t+1)!} \\
&= \frac{s! \cdot t!}{(s+t+1)!} \sum_{i=0}^{t} \frac{1}{s+i+1}.
\end{aligned} \tag{8.12}$$

Hence,

$$J(s,t) = \int_0^1 p^s \cdot (1-p)^t \cdot \ln(p)\, dp = -\frac{s! \cdot t!}{(s+t+1)!} \sum_{i=0}^{t} \frac{1}{s+i+1}. \tag{8.13}$$

H.3 The Definite Integral $K(p_1; k_1, M, N)$

In this subsection, we will display an analytic expression for the definite integral

$$\begin{aligned}
&K(p_1; k_1, M, N) \equiv \\
&\int_{p_2=0}^{1-p_1} \int_{p_3=0}^{1-p_1-p_2} \cdots \int_{p_{M-1}=0}^{1-p_1-p_2-\ldots-p_{M-2}} \binom{N}{\vec{k}} \cdot \prod_{i=2}^{M} p_i^{k_i}\, dp_{M-1} \cdots dp_3\, dp_2,
\end{aligned} \tag{8.14}$$

which almost always appears if Bayes estimators of multinomially distributed likelihoods $P(\vec{k}|\vec{p})$ are to be derived.

Since a complete mathematical proof is too long to get displayed in this work, we just present the main ideas and recommend all interested readers to exercise the induction by themselves.

First, the following variable substitutions seem to be recommendable:

$$\alpha_i = 1 - p_1 - p_2 - ... - p_i \tag{8.15}$$

and

$$N_i = N - k_1 - k_2 - ... - k_i \tag{8.16}$$

for all $i = 1, 2, ..., M$.

Then, we can rewrite

$$\begin{aligned} & K(p_1; k_1, M, N) \equiv \\ & \binom{N}{k_1} \cdot p_1^{k_1} \cdot \int\limits_{p_2=0}^{\alpha_1} \binom{N_1}{k_2} \cdot p_2^{k_2} \cdot \int\limits_{p_3=0}^{\alpha_2} \binom{N_2}{k_3} \cdot p_3^{k_3} \cdots \int\limits_{p_{M-1}=0}^{\alpha_{M-2}} \binom{N_{M-2}}{k_{M-1}} \cdot \\ & p_{M-1}^{k_{M-1}} \cdot (\alpha_{M-2} - p_{M-1})^{N_{M-2}-k_{M-1}} \, dp_{M-1} \cdots dp_3 \, dp_2, \end{aligned} \tag{8.17}$$

since

$$\alpha_i - \alpha_{i+1} = p_{i+1} \tag{8.18}$$

and

$$N_i - N_{i+1} = k_{i+1} \tag{8.19}$$

for all $i = 1, 2, ..., M-1$.

Finally, calculating the integral

$$\begin{aligned} & \int\limits_0^{\alpha} \binom{N}{k} \cdot p^k \cdot (\alpha - p)^{N-k+l} \, dp \\ = \; & \binom{N}{k} \cdot \alpha^{N+l} \cdot \int\limits_0^{\alpha} \left(\frac{p}{\alpha}\right)^k \cdot \left(1 - \frac{p}{\alpha}\right)^{N-k+l} \, dp \\ = \; & \binom{N}{k} \cdot \alpha^{N+l+1} \cdot \int\limits_0^1 x^k \cdot (1-x)^{N-k+l} \, dx \\ = \; & \binom{N}{k} \cdot \alpha^{N+l+1} \cdot \frac{k! \cdot (N-k+l)!}{(N+l+1)!} \\ = \; & \frac{N! \cdot (N-k+l)!}{(N-k)! \cdot (N+l+1)!} \cdot \alpha^{N+l+1} \end{aligned} \tag{8.20}$$

for all positive real α, integer N, and integer $k \leq N$, starting the induction, and calculating the occurring integrals as outlined above yields

$$K(p_1; k_1, M, N) = \frac{N!}{k_1! \cdot (N - k_1 + M - 2)!} \cdot p_1^{k_1} \cdot (1 - p_1)^{N-k_1+M-2}. \tag{8.21}$$

This result — the proportionality to $p_1^{k_1} \cdot (1-p_1)^{N-k_1+M-2}$ — can easily be interpreted in the following way. The quotient of the two integrals

$$\frac{\int_0^1 f(p) \cdot p^k \cdot (1-p)^{N-k}\, dp}{\int_0^1 p^k \cdot (1-p)^{N-k}\, dp} \tag{8.22}$$

can be read as the expectation value of the posterior density of the function $f(p)$ under the prior assumption of uniformly distributed p, which is identical to the Bayes estimator for $f(p)$ under the Bayes hypothesis for p:

$$\hat{f}(k) = \frac{\int_0^1 f(p) \cdot P(k|p) \cdot P(p)\, dp}{\int_0^1 P(k|p) \cdot P(p)\, dp}. \tag{8.23}$$

Let us now consider the multinomial case and assume f be a function that only depends on one component of p. Without loss of generality, let p_1 be the only component that f depends on, i.e. $f = f(p_1)$.

Then, under the assumption of a uniform prior distribution of p, the Bayes estimator for $f(p_1)$ becomes

$$\begin{aligned} \hat{f}(\vec{k}) &= \frac{\int f(p_1) \cdot P(\vec{k}|\vec{p}) \cdot P(\vec{p})\, d\vec{p}}{\int P(\vec{k}|\vec{p}) \cdot P(\vec{p})\, d\vec{p}} \\ &= \frac{\int_0^1 f(p_1) \cdot p_1^{k_1} \cdot (1-p_1)^{N-k_1} \cdot (1-p_1)^{M-2}\, dp_1}{\int_0^1 p_1^{k_1} \cdot (1-p_1)^{N-k_1} \cdot (1-p_1)^{M-2}\, dp_1} \end{aligned} \tag{8.24}$$

This, however, is exactly the one-dimensional Bayes estimator of the function $f(p_1)$ with the prior density

$$P(p_1) \sim (1-p_1)^{M-2}, \tag{8.25}$$

where the proportionality constant can be derived from the normalization constraint, which eventually yields

$$P(p_1) = (M-1) \cdot (1-p_1)^{M-2}. \tag{8.26}$$

H.4 The Finite Sum $S(p_1; k_1, M, N)$

In this section, we will show that the marginal distribution of the component k_i of a multinomially distributed vector $\vec{k}$ is binomially distributed with parameters p_i and N,

if the parameters of the underlying multinomial distribution are $\vec{p} = (p_1, p_2, ..., p_i, ..., p_M)$ and N.

Without loss of generality, we consider the marginal distribution of the component k_1 of the vector $\vec{k}$, which is multinomially distributed according to

$$P(\vec{k}; \vec{p}, N) = \binom{N}{\vec{k}} \cdot \prod_{i=1}^{M} p_i^{k_i} \tag{8.27}$$

with p_i being the probability that an M-sided die shows up its face i and k_i being the absolute frequency of observing the i-th face in a sample of size N.

Let us first introduce a generalized binomial formula by rewriting the N_1-th power of a sum over $M-1$ positive real numbers q_j as

$$\begin{aligned}
\left(\sum_{j=2}^{M} q_j\right)^{N_1} &= \left(q_2 + \sum_{j=3}^{M} q_j\right)^{N_1} = \sum_{k_2=0}^{N_1} \binom{N_1}{k_2} \cdot q_2^{k_2} \cdot \left(\sum_{j=3}^{M} q_j\right)^{N_1-k_2} \\
&= \sum_{k_2=0}^{N_1} \binom{N_1}{k_2} \cdot q_2^{k_2} \cdot \left(\sum_{k_3=0}^{N_1-k_2} \binom{N_1-k_2}{k_3} \cdot q_3^{k_3} \cdot \left(\sum_{j=4}^{M} q_j\right)^{N_1-k_2-k_3} \right) \\
&= \ ... \\
&= \sum_{k_2=0}^{N_1} \sum_{k_3=0}^{N_1-k_1} \cdots \sum_{k_{M-1}=0}^{N_1-\sum_{j=2}^{M-2} k_j} \binom{N_1}{k_2} \cdot \binom{N_1-k_2}{k_3} \\
&\cdots \binom{N_1-k_2-k_3-...-k_{M-2}}{k_{M-1}} \cdot q_2^{k_2} \cdot q_3^{k_3} \cdots q_M^{k_M} \\
&= \sum_{\{\vec{k_1}\}} \binom{N_1}{\vec{k_1}} \prod_{j=2}^{M} q_j^{k_j},
\end{aligned} \tag{8.28}$$

where the sum is to be taken over all vectors $\vec{k_1} = (k_2, k_3, ..., k_M)$ of positive integers k_j for which the equality

$$\sum_{j=2}^{M} k_j = N_1 \tag{8.29}$$

holds.

If we now set $N_1 \equiv N - k$, $p_j = q_j$ for $j = 2, 3, ..., M$, and exploit the normalization constraint

$$\sum_{j=2}^{M} p_j = 1 - p_1, \tag{8.30}$$

we immediately realize

$$\sum_{\{\vec{k_1}\}} \binom{N}{\vec{k}} \prod_{j=1}^{M} q_j^{k_j} = \binom{N}{k_1} \cdot p_1^{k_1} \cdot \sum_{\{\vec{k_1}\}} \binom{N_1}{\vec{k_1}} \prod_{j=2}^{M} q_j^{k_j}$$
$$= \binom{N}{k_1} \cdot p_1^{k_1} \cdot (1-p_1)^{N-k_1}. \tag{8.31}$$

Since this derivation is possible for all $i = 1, 2, ..., M$, we obtain the general result

$$\sum_{\{\vec{k_i}\}} \binom{N}{\vec{k}} \prod_{j=1}^{M} q_j^{k_j} = \binom{N}{k_i} \cdot p_i^{k_i} \cdot (1-p_i)^{N-k_i} \tag{8.32}$$

for all N, $k_i \leq N$, and $\vec{k_i} = (k_1, k_2, ..., k_{i-1}, k_{i+1}, ..., k_M)$, which states that all components of a multinomially distributed vector are binomially distributed. This result, however, must not mislead us to the wrong conclusion that all components are statistically independent. The reverse is right, i.e., all components of a multinomially distributed vector are mutually dependent.

H.5 The normalization constant W

Under the assumption of a stationary, independent distributed sample of data points, the conditional probability density to observe a sample with occupation numbers $\{N_1, \ldots, N_M\}$ is given by the multinomial distribution $P(\vec{N}|\vec{p}) = C_{\vec{N}} \prod_{i=1}^{M} p_i^{N_i}$. Here the multinomial coefficient reads as $C_{\vec{N}} = N!/\prod_{i=1}^{M} N_i!$, and the size of the sample is $N = \sum_{i=1}^{M} N_i$.

We define $W'(\vec{N}) = W(\vec{N})/C_{\vec{N}}$. Then, with a uniform prior probability density, the reduced normalization constant reads as

$$W'(\vec{N}) = \frac{1}{C_{\vec{N}}} \int_{\mathcal{S}} d\vec{p}\; P(\vec{N}|\vec{p})\; Q(\vec{p}) = \int_{\mathcal{S}} \prod_{j=1}^{M} dp_j\; p_j^{N_j}. \tag{8.33}$$

Introducing the auxiliary variable $k_j = 1 - \sum_{l=1}^{j} p_l$, the explicit integral takes on the form

$$W'(\vec{N}) = \int_{p_1=0}^{1} dp_1\; p_1^{N_1} \int_{p_2=0}^{k_1} dp_2\; p_2^{N_2} \cdots \int_{p_{M-1}=0}^{k_{M-2}} dp_{M-1}\; p_{M-1}^{N_{M-1}} \left(k_{M-2} - p_{M-1}\right)^{N_M}. \tag{8.34}$$

In the above expression, all integrals are of the type $\int du\; u^a(\xi - u)^b$. Changing co-ordinates $u = \xi v$, these integrals can be rewritten in terms of ordinary Beta-functions

$$\int_0^{\xi} du\; u^a(\xi - u)^b = \xi^{a+b+1}\; \mathrm{B}(a+1, b+1) \tag{8.35}$$

for all positive real numbers a and b, and $\mathrm{B}(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$. The relation $\Gamma(n+1) = n!$ holds for $n \in \mathcal{N}$.

Using relation (8.35), we may integrate equation (8.34) over p_{M-1} to get

$$W'(\vec{N}) = \mathrm{B}\Big(N_{M-1}+1, N_M+1\Big) \times \Bigg\{ \int_{p_1=0}^{1} dp_1\; p_1^{N_1} \int_{p_2=0}^{k_1} dp_2\; p_2^{N_2} \cdots \\ \cdots \int_{p_{M-2}=0}^{k_{M-3}} dp_{M-2}\; p_{M-2}^{N_{M-2}}\; (k_{M-3} - p_{M-2})^{(N_{M-1}+N_M+1)} \Bigg\}.$$

Completing the iteration for all but the integration over p_1, this yields

$$W'(\vec{N}) = \prod_{m=2}^{M-1} \mathrm{B}\Big(N_m+1, \sum_{j=m}^{M-1} N_{j+1} + (M-m)\Big) \times \\ \Bigg\{ \int_{p_1=0}^{1} dp_1\; p_1^{N_1}\; (1-p_1)^{\left(\sum_{j=2}^{M} N_j + (M-2)\right)} \Bigg\}.$$

Expressing the Beta-functions in terms of Gamma-functions, we obtain $W'(p_1;\vec{N})$ in the form

$$W'(p_1;\vec{N}) = p_1^{N_1}\; \frac{\prod_{j=2}^{M} \Gamma(N_j+1)}{\Gamma\left(\sum_{j=2}^{M} N_j + M - 1\right)}\; (1-p_1)^{\left(\sum_{j=2}^{M} N_j + (M-2)\right)}. \tag{8.36}$$

Inspecting the above expression, we realize that equation (8.36) can, in fact, be readily written down for a general ith component:

$$W'(p_i;\vec{N}) = p_i^{N_i}\; \frac{\prod_{\substack{j=1 \\ j\neq i}}^{M} \Gamma(N_j+1)}{\Gamma\left(\sum_{j=1}^{M}(1-\delta_{ij})N_j + M - 1\right)}\; (1-p_i)^{\left(\sum_{j=1}^{M}(1-\delta_{ij})N_j + (M-2)\right)}. \tag{8.37}$$

Integrating (8.37) over p_i, we arrive at the normalization constant

$$W(\vec{N}) \;=\; C_{\vec{N}} \int_{p_i=0}^{1} dp_i\; W'(p_i;\vec{N}) \;=\; \frac{\Gamma(N+1)}{\Gamma(N+M)}. \tag{8.38}$$

Appendix I

Characteristic Functions

In this appendix, we will briefly introduce characteristic functions $f_{\vec{k}}(\vec{t})$ of a discrete random vector $\vec{k}$ with the probability distribution $P\left(\vec{k}\right)$, which we will then calculate to display the first moments of the multinomial distribution given by

$$P\left(\vec{k}\right) = \binom{N}{\vec{k}} \prod_{i=1}^{M} p_i^{k_i}, \tag{9.1}$$

where $\vec{k} = (k_1, k_2, ..., k_M)$ denotes the M-dimensional vector containing the nonnegative integers k_i constrained by

$$\sum_{i=1}^{M} k_i = N \tag{9.2}$$

and $\vec{p} = (p_1, p_2, ..., p_M)$ denotes the M-dimensional probability vector containing the nonnegative reals p_i obeying the constraint

$$\sum_{i=1}^{M} p_i = 1. \tag{9.3}$$

More detailed introductions of characteristic functions containing a set of valuable theorems can be found, e.g., in [Bronstein & Semendjajew 1989], [Rényi 1982], or [Fisz 1989].

The table of the central multinomial moments displayed at the bottom of this appendix can also be found in [Harris 1975].

Definition I.1 *Let k be a random variable with its distribution function $F(k)$. Then we define the* **characteristic function** *$f_k(t)$ as the expectation value of the numbers* $\exp(i\,k\,t)$, *i.e.,*

$$f_k(t) \equiv \int \exp(i\,k\,t)\, dF(k). \tag{9.4}$$

By expanding $\exp(i\,k\,t)$, we obtain

$$f_k(t) = E\left(1 + i\,k\,t + \frac{(i\,k\,t)^2}{2!} + \frac{(i\,k\,t)^3}{3!} + \cdots\right) = \sum_{n=0}^{\infty} \frac{i^n\,\eta_n\,t^n}{n!}, \tag{9.5}$$

where η_n denotes the n-th central moment of k, i.e.,

$$\eta_n = E\,(k^n) = \int k^n\,dF(k). \tag{9.6}$$

Differentiating n times and then setting $t = 0$ gives

$$i^n \cdot \eta_n = \left[\frac{\partial^n f_k(t)}{\partial t^n}\right]_{t=0}, \tag{9.7}$$

i.e., the n-th derivative of the characteristic function $f_k(t)$ yields the n-th moment of k.

Considering the binomial distribution, we obtain

$$f_k(t) = \sum_{k=0}^{N} \exp(i\,k\,t) \cdot \binom{N}{k} \cdot p^k \cdot (1-p)^{N-k} = (p \cdot \exp(i\,t) + 1 - p)^N \tag{9.8}$$

and thus all moments by differentiating this function.

Let us now consider the characteristic function of multivariate distributions.

Definition I.2 *Let $\vec{k}$ be a random vector and $F(\vec{k})$ its distribution function. Then we define the* **characteristic function** *$f_{\vec{k}}(\vec{t})$ as the expectation value of the numbers $\exp(i\,\vec{k}\,\vec{t})$, i.e.,*

$$f_{\vec{k}}(\vec{t}) \equiv \int \exp(i\,\vec{k}\,\vec{t})\,dF(\vec{k}). \tag{9.9}$$

Hence, the characteristic function of a multinomially distributed vector $\vec{k}$ is given by

$$\begin{aligned} f_{\vec{k}}(\vec{t}) &= \sum_{\{\vec{k}\}} \binom{N}{\vec{k}} \cdot \prod_{i=1}^{M} p_i^{k_i} \cdot \prod_{i=1}^{M} \exp(i\,k_i\,t_i) \\ &= \sum_{\{\vec{k}\}} \binom{N}{\vec{k}} \cdot \prod_{i=1}^{M} (p_i \cdot \exp(i\,t_i))^{k_i} \\ &= \left(\sum_{i=1}^{M} p_i \cdot \exp(i\,t_i)\right)^N \end{aligned} \tag{9.10}$$

according to appendix H.4.

In analogy to the univariate case, we obtain all higher moments of the multinomial distribution by differentiating the characteristic function $f_{\vec{k}}(\vec{t})$.

Let us, in the remainder of this appendix, display the first centered moments $\mu_{n_1\, n_2\, \ldots\, n_l}$ defined by

$$\mu_{n_1\, n_2\, \ldots\, n_l} \equiv E\left((k_1 - N\, p_1)^{n_1} \cdot (k_2 - N\, p_2)^{n_2} \cdots (k_l - N\, p_l)^{n_l} \right), \tag{9.11}$$

which can be obtained by completely elementary methods from the moments $\eta_{n_1\, n_2\, \ldots\, n_l}$.

$$\begin{aligned}
\mu_1 &= 0 && (9.12)\\
\mu_2 &= N \cdot (p_1 - p_1^2) && (9.13)\\
\mu_3 &= N \cdot (p_1 - 3 \cdot p_1^2 + 2 \cdot p_1^3) && (9.14)\\
\mu_4 &= 3\, N^2 \cdot (p_1^2 - 2 \cdot p_1^3 + p_1^4) && \\
&+ N \cdot (p_1 - 7 \cdot p_1^2 + 12 \cdot p_1^3 - 6 \cdot p_1^4) && (9.15)\\
\mu_{1\,1} &= -N \cdot p_1 \cdot p_2 && (9.16)\\
\mu_{2\,1} &= N \cdot (2 \cdot p_1^2 \cdot p_2 - p_1 \cdot p_2) && (9.17)\\
\mu_{3\,1} &= 3\, N^2 \cdot (p_1^3 \cdot p_2 - p_1^2 \cdot p_2) && \\
&+ N \cdot (-6 \cdot p_1^3 \cdot p_2 + 6 \cdot p_1^2 \cdot p_2 - p_1 \cdot p_2) && (9.18)\\
\mu_{2\,2} &= N^2 \cdot (3 \cdot p_1^2 \cdot p_2^2 - p_1^2 \cdot p_2 - p_1 \cdot p_2^2 + p_1 \cdot p_2) && \\
&+ N \cdot (-6 \cdot p_1^2 \cdot p_2^2 + 2 \cdot p_1^2 \cdot p_2 + 2 \cdot p_1 \cdot p_2^2 - p_1 \cdot p_2) && (9.19)\\
\mu_{1\,1\,1} &= 2\, N \cdot p_1 \cdot p_2 \cdot p_3 && (9.20)\\
\mu_{2\,1\,1} &= N^2 \cdot (3 \cdot p_1^2 \cdot p_2 \cdot p_3 - p_1 \cdot p_2 \cdot p_3) && \\
&+ N \cdot (-6 \cdot p_1^2 \cdot p_2 \cdot p_3 + 2 \cdot p_1 \cdot p_2 \cdot p_3) && (9.21)\\
\mu_{1\,1\,1\,1} &= 3\, N^2 \cdot p_1 \cdot p_2 \cdot p_3 \cdot p_4 - 6\, N \cdot p_1 \cdot p_2 \cdot p_3 \cdot p_4 && (9.22)
\end{aligned}$$

Appendix J

Mean and Variance of $\ln \chi^2$

In this appendix, we derive the mean and variance of the logarithm of a χ^2-distributed random variable.

Let X_i be i. i. d. continuous random variables with outcomes x_i and probability density functions

$$P(x_i) \equiv \frac{1}{\sqrt{2\pi}} e^{-\frac{x_i^2}{2}} \tag{10.1}$$

for $i = 1, 2, 3, ..., \infty$. Let Y_m be the continuous random variable with outcomes $y_m \equiv \sum_{i=1}^{m} x_i^2$. Then the probability density function of Y_m is

$$Q_m(y) = \frac{y^{\frac{m}{2}-1} e^{-\frac{y}{2}}}{2^{\frac{m}{2}} \Gamma(\frac{m}{2})}. \tag{10.2}$$

The distribution of X_i is called *normal distribution* with mean 0 and variance 1, and the distribution of Y_m is called χ^2 *distribution* with m degrees of freedom. Eq. (10.2) can be derived in (at least) two ways:

1. Complete Induction over m

   - Compute by brute force (but easily)

     $$Q_1(y) = \frac{e^{-\frac{y}{2}}}{\sqrt{2\pi}\sqrt{y}} \tag{10.3}$$

     and

     $$Q_2(y) = \int_0^y Q_1(x) Q_1(y-x) dx = \frac{e^{-\frac{y}{2}}}{2}. \tag{10.4}$$

   - Show that

     $$Q_{m+1}(y) = \int_0^y Q_1(x) Q_m(y-x) dx \tag{10.5}$$

     for all $m = 1, 2, ..., \infty$.

- This implies that the sum of k independent χ^2-distributed random variables with $m_1, m_2, ..., m_k$ degrees of freedom is χ^2-distributed with $m = m_1 + m_2 + ... + m_k$ degrees of freedom, i. e.

$$Q_{m_1+m_2}(y) = \int_0^y Q_{m_1}(x) Q_{m_2}(y-x) dx. \tag{10.6}$$

2. Characteristic Functions

- Define the characteristic function

$$f_m(t) \equiv \int_0^\infty Q_m(x) e^{xt} dx \tag{10.7}$$

and compute

$$f_1(t) = (1-2t)^{-\frac{1}{2}}. \tag{10.8}$$

- Use the convolution theorem, which states that the characteristic function of the sum of independent random variables is equal to the product of their characteristic functions, to obtain

$$f_m(t) = f_1^m(t) = (1-2t)^{-\frac{m}{2}}. \tag{10.9}$$

- Perform the inverse Laplace transform on $f_m(t)$ to obtain eq. (10.2).
- Obviously

$$\begin{aligned} f_{m_1+m_2}(t) &\equiv \int_0^\infty Q_{m_1+m_2}(x) e^{xt} dx \\ &= f_{m_1}(t) f_{m_2}(t). \end{aligned} \tag{10.10}$$

The n-th moment, $\langle y^n \rangle$, of the χ^2 distribution with m degrees of freedom can be obtained by brute force integration,

$$\langle y^n \rangle \equiv \int_0^\infty Q_m(y) y^n dy = 2^n \frac{\Gamma(\frac{m}{2}+n)}{\Gamma(\frac{m}{2})}, \tag{10.11}$$

or, more elegantly, by Taylor-expanding the generating function,

$$\begin{aligned} f_m(t) &\equiv \int_0^\infty Q_m(y) e^{yt} dy \\ &= \int_0^\infty Q_m(y) (1 + yt + \frac{(yt)^2}{2} + \frac{(yt)^3}{3!} + ...) dy \\ &= \langle 1 \rangle + \langle y \rangle t + \langle y^2 \rangle \frac{t^2}{2} + \langle y^3 \rangle \frac{t^3}{3!} + ..., \end{aligned} \tag{10.12}$$

yielding

$$\langle y^n \rangle = \frac{d^n f_m(t)}{dt^n}|_{t=0} = 2^n \frac{\Gamma(\frac{m}{2}+n)}{\Gamma(\frac{m}{2})}, \tag{10.13}$$

in agreement with eq. (10.11). From

$$\langle y^0 \rangle = 1, \tag{10.14}$$

$$\langle y^1 \rangle = m, \tag{10.15}$$

$$\langle y^2 \rangle = m(m+2), \tag{10.16}$$

we obtain

$$\sigma^2(y) \equiv \langle y^2 \rangle - \langle y \rangle^2 = 2m. \tag{10.17}$$

The first two moments of $\ln y$ can be obtained by brute force integration,

$$\begin{aligned} \langle \ln y \rangle &\equiv \int_0^\infty Q_m(y) \ln y \, dy \\ &= \ln 2 + \psi_0(\frac{m}{2}), \end{aligned} \tag{10.18}$$

$$\begin{aligned} \langle \ln^2 y \rangle &\equiv \int_0^\infty Q_m(y) \ln^2 y \, dy \\ &= \ln^2 2 + \ln 4 \, \psi_0(\frac{m}{2}) + \psi_0^2(\frac{m}{2}) + \psi_1(\frac{m}{2}), \end{aligned} \tag{10.19}$$

or by performing the variable transformation $z = \ln y$ and computing the moments $\langle z \rangle$ and $\langle z^2 \rangle$ of the density

$$R_m(z) \equiv Q_m(y)\frac{dy}{dz} = \frac{e^{\frac{mz-e^z}{2}}}{2^{\frac{m}{2}}\Gamma(\frac{m}{2})}. \tag{10.20}$$

The characteristic function of $R_m(z)$ is

$$g_m(t) \equiv \int_{-\infty}^\infty R_m(z) e^{zt} dz = 2^t \frac{\Gamma(\frac{m}{2}+t)}{\Gamma(\frac{m}{2})}, \tag{10.21}$$

and the first and second derivatives of $g_m(t)$ evaluated at $t = 0$ yield eqs. (10.18) and (10.19), where

$$\psi_n(x) \equiv \frac{d^{n+1} \ln \Gamma(x)}{dx^{n+1}} \tag{10.22}$$

is the *polygamma function*. Hence, we obtain

$$\sigma^2(\ln y) \equiv \langle \ln^2 y \rangle - \langle \ln y \rangle^2 = \psi_1(\frac{m}{2}). \tag{10.23}$$

With

$$\gamma \equiv \lim_{n\to\infty} \sum_{k=1}^n \frac{1}{k} - \ln n \tag{10.24}$$

being *Euler's constant*, we obtain for $m = 6$

$$\langle \ln y \rangle = \ln 2 + \frac{3}{2} - \gamma, \tag{10.25}$$

and

$$\sigma^2(\ln y) = \frac{\pi^2}{6} - \frac{5}{4}. \tag{10.26}$$

Appendix K

Definitions of C and U

In this appendix, we detail the calculation of the correlation coefficient, $C(X,Y)$, and the uncertainty coefficient, $U(X,Y)$.

K.1 Rank-Ordered Correlation Coefficient C

Linear statistical dependences can be quantified by calculating the correlation coefficient $C(X,Y)$ of two random variables X and Y. $C(X,Y)$ ranges from -1 to 1, where $(-)1$ corresponds to perfect sample (anti-)correlation, and 0 is the value for linearly statistically independent samples. Denote the average over the data set by $\langle\ \rangle_S$, and define the covariance $\sigma^2_{XY} = \langle (x-\bar{x})(y-\bar{y})\rangle_S$ and the mean values $\bar{x} = \frac{1}{S}\cdot\sum_{s=1}^{S} x_s$ and $\bar{y} = \frac{1}{S}\cdot\sum_{s=1}^{S} y_s$. $C(X,Y)$ is defined as (Sachs 1984)

$$C(X,Y) \quad = \quad \frac{\sigma^2_{XY}}{\sigma_{XX}\cdot\sigma_{YY}}.$$

In this study, we use rank numbers rather than direct measurements, since then $C(X,Y)$ becomes independent on monotonic scaling. We obtain rank numbers through $X' = \{r_i(x)\}$, where $r_i(x)$ is the rank of the ith element of the original data sample permuted according to $r_i(x) = \#\{j | x_j \leq x_i, 1 \leq j \leq S\}$.

K.2 Uncertainty Coefficient U

We use the uncertainty coefficient $U(X,Y)$ to quantify non-linear statistical dependences. $U(X,Y)$ is defined as (Press *et al.* 1992)

$$U(X,Y) \quad = \quad 2 \cdot \frac{H(X) + H(Y) - H(X,Y)}{H(X) + H(Y)},$$

where $H(X) = -\sum_{m=1}^{M} p_m \cdot \log_2 p_m$. Eqn (11.1) is the (normalized) mutual information in X about Y (Shannon 1948). The marginal entropies, $H(X)$ and $H(Y)$, and the joint entropy, $H(X,Y)$ are computed as follows:

1. Distribute X and Y on an array consisting of $M \times M$ bins such that the marginal probabilities $\Pr(X = x_m) = p_m$ and $\Pr(Y = y_n) = p_n$ are uniform, that is $p_m = p_n = 1/M$. Consequently, we have $H(X) = H(Y) = \log_2 M$.
2. Determine the joint probabilities, $\Pr\{X = x_m, Y = y_n\} = p_{m,n}$ from the distribution of X and Y on the array of M^2 bins..
3. Calculate $U(X,Y)$ according to eqn (11.1). Since the marginal probabilities obey *a priori* a uniform distribution, we compute $U(X,Y) = 2 - \frac{H(X,Y)}{\log_2 M}$.

$U(X,Y)$ ranges from 0 to 1, where 0 corresponds to the case in which X and Y are statistically independent, and 1 to the case in which X and Y are interdependent.

Appendix L

Scale Invariance and Non-Self-Averaging Behavior in a Simple Fragmentation Process

In this chapter we investigate statistical properties of a simple recursive fragmentation processes. We show that the fragment length distribution is purely algebraic, and that the fragmentation process is non-self-averaging. Additionally, extremal properties, e.g., the distribution of the largest fragment, exhibit an infinite number of singularities. In d-dimensions, the volume distribution is given by a sum of d power-laws, and consequently, the small-size tail diverges algebraically.

L.1 Introduction

Numerous physical phenomena are characterized by a set of variables, say $\{x_j\}$, which evolves according to a random process, and are subject to the conservation law $\sum_j x_j =$ const. An important example of such a stochastic process is fragmentation, with applications ranging from geology [Turcotte 1996] and fracture [Lawn & Wilshaw 1975] to the breakup of liquid droplets [Shinnar 1961] and atomic nuclei [Chase et al. 1998, Redner 1990]. Other examples include spin glasses [Mezard et al. 1987], where x_j represents the equilibrium probability of of finding the system in the j^{th} valley, genetic populations, where x_j is the frequency of the j^{th} allele [Higgs 1995, Derrida & Jung-Muller 1999], and random Boolean networks [Kauffman 1993, Flyvbjerg & Kjaer].

Our primary motivation arises from applications of a DNA segmentation algorithm [Bernaola-Galván et al. 1996, Román-Roldán et al. 1998, Bernaola-Galván et al. 1999, Li et al. 1998, Bernaola-Galván et al. 1999, Oliver et al. 1999], which is used to decompose a heterogeneous DNA sequence into homogeneous sub-sequences. The segmentation algorithm attempts to divide recursively a given heterogeneous sequence into two sub-sequences. For each possible "break point" one computes a heterogeneity measure, for example, the Jensen-Shannon divergence [Lin 1991], and chooses that "break point" at which the heterogeneity measure is maximal. If that maximal value is greater than some predefined confidence level, the sequence is divided into two sub-sequences, and the outlined segmentation procedure is repeated recursively for both sub-sequences. Otherwise, the sequence does not undergo further segmentation, and it is considered homogeneous with the given confidence level. To evaluate whether the length distributions of the resulting fragments from DNA sequences are significantly different from the length distribution of fragments resulting from random, uncorrelated sequence, we derive the latter distribution here.

L.2 Recursive Fragmentation Process

Specifically, we investigate the following *recursive fragmentation process*. We start with the unit interval and choose a break point l in $[0, 1]$ with a uniform probability density. Then, with probability p, the interval is divided into two fragments of lengths l and $1 - l$, while with probability $q = 1 - p$, the interval becomes "frozen" and is never fragmented any further. If the interval is fragmented, we recursively apply the above fragmentation procedure to both of the resulting fragments.

First, let us examine the average total number of fragments, N. Since with probability q a single fragment is produced, and with probability p the process is repeated with two fragments, N satisfies $N = q + 2pN$, yielding

$$N = \begin{cases} q/(1-2p), & \text{if } p < 1/2; \\ \infty, & \text{if } p \geq 1/2. \end{cases} \tag{12.1}$$

The average total number of fragments becomes infinite at the critical point $p_c = 1/2$, reflecting the critical nature of the underlying branching process [Harris 1989].

L.3 Fragment Length Distribution

Next, we study $P(x)$, the density of fragments of length x. The recursive nature of the process can be used to obtain the fragment length density

$$P(x) = q\delta(x-1) + 2p\int_x^1 \frac{dy}{y}\, P\left(\frac{x}{y}\right). \tag{12.2}$$

The gain term indicates that a fragment can be created only from a larger fragment, and the y^{-1} kernel reflects the uniform fragmentation density. Eq. (12.2) can be solved by introducing the Mellin transform

$$M(s) = \int dx\, x^{s-1} P(x). \tag{12.3}$$

Eqs. (12.2) and (12.3) yield $M(s) = q + 2ps^{-1}M(s)$, which implies

$$M(s) = q + \frac{2pq}{s-2p}. \tag{12.4}$$

The average total number $M(1) = N$ is consistent with Eq. (12.1), and the total fragment length $M(2) = 1$ is conserved in accord with $1 = \int dx\, xP(x)$. (Here and in the following all integrals with unspecified limits are taken over the interval $0 < x < 1$.) The inverse Mellin transform of Eq. (12.4) gives

$$P(x) = q\delta(x-1) + 2pq\, x^{-2p}. \tag{12.5}$$

Apart from the obvious delta function, the length density is a purely algebraic function. In particular, the fragment distribution diverges algebraically in the limit of small fragments. Interestingly, given such an algebraic divergence near the origin $P(x) \sim x^{-\gamma}$, length conservation restricts the exponent range to $\gamma < 2$. In our case $\gamma = 2p$, and since $0 < p < 1$, the entire range of acceptable exponents emerges by tuning the only control parameter p.

Interestingly, at the critical point $p_c = \frac{1}{2}$, the fragment length distribution becomes independent of the initial interval length. Starting from an interval of length L, Eq. (12.5) can be generalized to yield

$$P(x) = q\delta(x-L) + 2pqL^{1-2p}x^{-2p}.$$

Thus, the critical point may be detected by observing that point at which the segment distribution becomes independent of the original interval length.

L.4 Multidimensional Generalization

The recursive fragmentation process can be generalized to d dimensions. For instance, in two dimensions we start with the unit square, choose a point (x_1, x_2) with a uniform probability density, and divide, with probability p, the original square into four rectangles of sizes $x_1 \times x_2$, $x_1 \times (1-x_2)$, $(1-x_1) \times x_2$, and $(1-x_1) \times (1-x_2)$. With probability q, the square becomes frozen and we never again attempt to fragment it. The process is repeated recursively for each fragment produced in the previous step.

Let $P(x_1, \ldots, x_d)$ be the probability density of fragments of size $x_1 \times \cdots \times x_d$. $P(x_1, \ldots, x_d)$ satisfies

$$\begin{aligned} P(x_1, \ldots, x_d) &= q \prod_{i=1}^{d} \delta(x_i - 1) \\ &+ 2^d p \int \prod_{i=1}^{d} \frac{dy_i}{y_i} P\left(\frac{x_1}{y_1}, \ldots, \frac{x_d}{y_d}\right). \end{aligned} \tag{12.6}$$

Following the steps leading to Eq. (12.4), we find that the d-dimensional Mellin transform, defined by $M(s_1, \ldots, s_d) = \int \prod_{i=1}^{d} dx_i \, x_i^{s_i - 1} P(x_1, \ldots, x_d)$, satisfies

$$M(s_1, \ldots, s_d) = q \left[1 + \frac{\alpha^d}{\prod_{i=1}^{d} s_i - \alpha^d} \right], \tag{12.7}$$

with $\alpha = 2p^{1/d}$.

The overall number of fragments is $N = M(1, \ldots, 1) = q/(1 - 2^d p)$ if $p < 2^{-d}$, and $N \to \infty$ if $p \geq 2^{-d}$. One can verify that the total volume $M(2, \ldots, 2) = 1$ is conserved. Interestingly, there is an additional infinite set of conserved quantities: all moments whose indices belong to the hyper-surface

$$\prod_{i=1}^{d} s_i^* = 2^d \tag{12.8}$$

satisfy $M(s_1^*, \ldots, s_d^*) = 1$. In a continuous time formulation of this process, which is similar (but not identical) to the special case $p = 1$, the same moments were found to be integrals of motion [Krapivsky & Ben-Naim 1996]. The existence of an infinite number of conservation laws is surprising, because only the volume conservation has a clear physical justification.

L.5 Volume Distribution in d Dimensions

Next, we study the volume density $P(V)$, defined by

$$P(V) = \int dx_1 \cdots dx_d \, P(x_1, \ldots, x_d) \delta(V - \prod x_i). \tag{12.9}$$

The Mellin transform $M(s) = \int dV V^{s-1} P(V)$ can be obtained from Eq. (12.7) by setting $s_i = s$,

$$M(s) = q\left[1 + \frac{\alpha^d}{s^d - \alpha^d}\right], \tag{12.10}$$

with $\alpha = 2p^{1/d}$. Using the d^{th} root of unity, $\zeta = e^{2\pi i/d}$, and the identity $\frac{1}{s^d-1} = \frac{1}{d}\sum_{k=0}^{d-1} \frac{\zeta^k}{s-\zeta^k}$, $M(s)$ can be expressed as a sum over simple poles at $\alpha\zeta^k$. Consequently, the inverse Mellin transform is given by a linear combination of d power laws

$$P(V) = q\left[\delta(V-1) + \frac{\alpha}{d}\sum_{k=0}^{d-1} \zeta^k V^{-\alpha\zeta^k}\right]. \tag{12.11}$$

One can verify that $P(V)$ equals its complex conjugate, and hence $P(V)$ is real. Additionally, the one-dimensional case (12.5) is recovered by setting $d = 1$.

The small-volume tail of the distribution can be obtained by noting that the sum in Eq. (12.11) is dominated by the first term in the series, which leads to

$$P(V) \simeq A_d V^{-2p^{1/d}} \qquad \text{as} \qquad V \to 0, \tag{12.12}$$

with $A_d = \alpha q/d$. Although the value of the exponent changes, the possible range of exponents for this process remains the same since $0 < 2p^{1/d} < 1$ when $0 < p < 1$. In the infinite dimension limit, $P(V)$ becomes universal: $P(V) \sim V^{-2}$.

The leading behavior of $P(V)$ in the large size limit can be derived by using the Taylor expansion and the identity $\sum_{k=0}^{d-1} \zeta^{kn} = \delta_{n,0}$ for $n = 0, \ldots, d-1$. One finds that in higher dimensions the volume distribution vanishes algebraically near its maximum value,

$$P(V) \simeq B_d (1-V)^{d-1} \qquad \text{as} \qquad V \to 1, \tag{12.13}$$

with $B_d = \alpha^d/(d-1)!$.

L.6 Fragment Length Distribution in d Dimensions

The entire multivariate fragment length density can be derived by performing the inverse Mellin transform of Eq. (12.7). We expand the geometric series $\frac{\alpha^d}{\prod_{i=1}^d s_i - \alpha^d} = \sum_{n\geq 0} \prod_{i=1}^d \left(\frac{\alpha}{s_i}\right)^{n+1}$, and perform the inverse Mellin transform for each variable separately by using the identity $\int dx\, x^{s-1} \left[\ln \frac{1}{x}\right]^n = n! s^{-n-1}$. Hence, we obtain

$$P(x_1, \ldots, x_d) = q\left[\prod_{i=1}^d \delta(x_i - 1) + \alpha^d F_d(z)\right], \tag{12.14}$$

with the shorthand notations

$$F_d(z) = \sum_{n=0}^{\infty} \left(\frac{z^n}{n!}\right)^d, \qquad z = \alpha \left(\prod_{i=1}^{d} \ln \frac{1}{x_i}\right)^{1/d}. \tag{12.15}$$

The small size behavior of $P(x_1, \ldots, x_d)$ can be obtained by using the steepest decent method. The leading tail behavior, $F_d(z) \simeq (2\pi z)^{\frac{1-d}{2}} e^{zd}$ for $z \gg 1$, corresponds to the case when at least one of the lengths is small, i. e. $x_i \ll 1$. Hence, we find an unusual "log-stretched-exponential" expression

$$P(x_1, \ldots, x_d) \sim \sqrt{z/\alpha}^{1-d} e^{zd} \tag{12.16}$$

L.7 Non-Self-Averaging Behavior

The fragment length density represents an average over infinitely many realizations of the fragmentation process, and hence does not capture sample to sample fluctuations. These fluctuations are particularly important in non-self-averaging systems, where sample to sample fluctuations do not vanish in the thermodynamic limit. Useful quantities to characterize disordered systems, which are often non-self-averaging [Mezard et al. 1987, Derrida 1997], are the moments Y_α defined by

$$Y_\alpha = \sum_i x_i^\alpha, \tag{12.17}$$

where the sum runs over all fragments.

We are interested in the average values $\langle Y_\alpha \rangle$ and $\langle Y_\alpha Y_\beta \rangle$. For integer α, $\langle Y_\alpha \rangle$ is the probability that α points randomly chosen in the unit interval belong to the same fragment. The expected value of Y_α satisfies

$$\langle Y_\alpha \rangle = q + p \langle Y_\alpha \rangle \int dy \, [y^\alpha + (1-y)^\alpha]. \tag{12.18}$$

The first term corresponds to the case where the unit interval is not fragmented, and the second term corresponds to the case where fragmentation occurs. Eq. (12.18) gives

$$\langle Y_\alpha \rangle = q \left[1 + \frac{2p}{\alpha + 1 - 2p}\right] \tag{12.19}$$

if $\alpha > 2p - 1$, and $\langle Y_\alpha \rangle \to \infty$ if $\alpha \leq 2p - 1$. As expected, Eq. (12.19) agrees with the moments of $P(x)$ (12.5), $\langle Y_\alpha \rangle = \int dx \, x^\alpha P(x)$.

Higher order averages do not follow directly from the fragment density. For example, consider for example $\langle Y_\alpha Y_\beta \rangle$. For integer α and β, $\langle Y_\alpha Y_\beta \rangle$ is the probability that, if $\alpha + \beta$ points are chosen at random, the first α points all lie on the same fragment, and the last β points all lie on another (possibly the same) fragment. $\langle Y_\alpha Y_\beta \rangle$ satisfies

$$\begin{aligned} \langle Y_\alpha Y_\beta \rangle &= q + p\langle Y_\alpha Y_\beta \rangle \int dy \left[y^{\alpha+\beta} + (1-y)^{\alpha+\beta} \right] \\ &+ p\langle Y_\alpha \rangle \langle Y_\beta \rangle \int dy \left[y^\alpha (1-y)^\beta + (1-y)^\alpha y^\beta \right], \end{aligned}$$

yielding

$$\begin{aligned} \langle Y_\alpha Y_\beta \rangle &= q + \frac{2pq}{\alpha+\beta+1-2p} \\ &+ 2p \, \frac{\Gamma(\alpha+1)\Gamma(\beta+1)}{\Gamma(\alpha+\beta+1)} \, \frac{\langle Y_\alpha \rangle \langle Y_\beta \rangle}{\alpha+\beta+1-2p}, \end{aligned}$$

if $\alpha, \beta > 2p-1$ and $\alpha + \beta > 2p-1$, and $\langle Y_\alpha Y_\beta \rangle \to \infty$ otherwise.

Note that $\langle Y_\alpha Y_\beta \rangle \neq \langle Y_\alpha \rangle \langle Y_\beta \rangle$, and in particular $\langle Y_\alpha^2 \rangle \neq \langle Y_\alpha^2 \rangle$. Hence, fluctuations in Y_α do not vanish in the thermodynamic limit, which states that the recursive fragmentation process is non-self-averaging. While for $p < 1/2$ non-self-averaging behavior is expected because the total number of fragments is finite, the emergence of non-self-averaging quantities for $p > 1/2$ is surprising. Hence, statistical properties obtained by averaging over all realizations are insufficient to probe sample to sample fluctuations. In principle, higher order averages such as $\langle Y_\alpha^n \rangle$ can be calculated recursively by the procedure outlined above.

L.8 Length Distribution of the Largest Fragment

Extremal properties provide an additional probe of sample to sample fluctuations. Specifically, let us consider $\mathcal{L}(x)$, the length density of the largest fragment. For a self-averaging fragmentation processes with an infinite number of fragments one expects $\mathcal{L}(x) \to \delta(x)$ in the thermodynamic limit. To see that $\mathcal{L}(x)$ is non-trivial for any p, let us first determine $\mathcal{L}(x)$ for $x \geq 1/2$. In this region,

$$\mathcal{L}(x) = q\delta(x-1) + p \int_x^1 \frac{dy}{y} \, \mathcal{L}\left(\frac{x}{y}\right). \tag{12.20}$$

If the original unit interval is not fragmented, the largest fragment is obviously the unit interval. If the first fragmentation is performed, only one of the two resulting fragments can be larger than $x > 1/2$. Therefore, only subsequent breaking of this fragment (of length $y > x$) can contribute to $\mathcal{L}(x)$, which explains Eq. (12.20).

Eq. (12.20) is similar to Eq. (12.2), and can be solved by the same technique to give

$$\mathcal{L}(x) = q\delta(x-1) + pq\, x^{-p} \qquad \text{for} \quad x \geq 1/2. \tag{12.21}$$

In the complementary case of $x < 1/2$, $\mathcal{L}(x)$ satisfies

$$\begin{aligned} \mathcal{L}(x) &= p\int_{1-x}^{1} \frac{dy}{y}\,\mathcal{L}\left(\frac{x}{y}\right) \\ &+ p\int_{1/2}^{1-x} \frac{dy}{y}\,\mathcal{L}\left(\frac{x}{y}\right)\mathcal{L}_-\left(\frac{x}{1-y}\right) \\ &+ p\int_{1/2}^{1-x} \frac{dy}{y}\,\mathcal{L}\left(\frac{x}{1-y}\right)\mathcal{L}_-\left(\frac{x}{y}\right). \end{aligned}$$

The first term on the right-hand side of this equation is constructed as in Eq. (12.20): if we first break the unit interval into two fragments of lengths $y > 1/2$ and $1-y$, then for $1-y < x$ the longest fragment is produced by breaking the fragment of length y. The next two terms describe the situation when $1-y > x$, so the longest fragment can arise out of breaking any of the two fragments. The factors $\mathcal{L}_-(u) = \int_0^u dv\,\mathcal{L}(v)$ guarantee that the longest fragment of length x comes from the fragment of length v in the first generation.

Since we already know $\mathcal{L}(x)$ for $x \geq \frac{1}{2}$, we can compute $\mathcal{L}(x)$ for $\frac{1}{3} \leq x \leq \frac{1}{2}$ by substituting the expression for $\mathcal{L}(x)$ into the above equation. Similarly, we can compute $\mathcal{L}(x)$ for $\frac{1}{k+1} \leq x \leq \frac{1}{k}$. Clearly, $\mathcal{L}(x)$ is analytic in the intervals $\left(\frac{1}{k+1}, \frac{1}{k}\right)$, but singular at the boundaries of these intervals, namely at $x = 1/k$. These singularities (we loosely use the term singularity to denote the existence of only a finite number of derivatives at $x = 1/k$) become weaker as k increases. Singularities at $x = 1/k$ appear to be a generic property of several disordered systems, including random walks, spin glasses, random maps, and random trees [Chase et al. 1998, Higgs 1995, Derrida & Jung-Muller 1999, Derrida 1997, Derrida & Flyvbjerg 1987, Frachebourg et al. 1995, Derrida & Flyvbjerg 1988].

L.9 Summary

We have found that the recursive fragmentation process is scale free, i.e., the fragment length distribution is purely algebraic. In higher dimensions, the volume distribution is a linear combination of d power laws, and consequently, an algebraic divergence characterizes the small-fragment tail of the distribution. A number of experimental fragmentation studies, where solid objects impact a hard surface, report algebraic mass (or equivalently volume) distributions with exponents ranging from 1 to 2

[Oddershede et al. 1993, Kadono 1997]. As in such situations a fragment may impact the surface more than once, the recursive fragmentation process may serve as a starting point to model those experimental fragmentation processes. We have found that the recursive fragmentation process exhibits features that are typical of complex and disordered systems, such as non-self-averaging behavior and the existence of an infinite number of singularities in the distribution of the largest fragment. These features indicate that even in the thermodynamic limit sample to sample fluctuations remain, and that knowledge of first order averages may not be sufficient for characterizing the system. This implies that experimental results obtained from DNA sequences must be handled with great care, as large sample to sample fluctuations (from DNA of different organisms) are expected and do not necessarily reflect organism-specific biological features.

Appendix M

How random are random numbers?

Random numbers were needed to perform most simulations in this work and they are generally required in many areas of statistical physics, e.g. stochastic optimization, Monte-Carlo methods or stochastic simulation [Ebeling & Feistel 1982, Allen & Tildesley 90, Schnakenberg 1995]. Random numbers are usually generated by a pseudo random number generator (PRNG). A problem with all PRNG is that pseudo random sequences generated in this way contain weak correlations that may lead to spurious simulation results [Ferrenberg et al. 1992].

In this appendix a simple and efficient method [Beule & Grosse 1996] for testing *randomness* is introduced and applied to several widely used PRNGs in order to detect weak correlations in pseudo random sequences and helps to determine which PRNGs are suitable for sampling large discrete spaces.

M.1 Pseudo Random Numbers

A PRNG is an iterative map F of a number (or a set of numbers) x_j onto a new number $x_{j+1} = F(x_j)$. The map is chosen in such a way that for suitable initial values x_0 the sequence $\{x_j\}$ (or parts of it) appear randomly distributed in a certain interval [Marsaglia 1992]. The pseudo random numbers generated in this way have several desirable features: they are reproducible (e.g. for counter checking results) and produced efficiently

without any special equipment[1]. An overview of the many possibilities for choosing the map $F(x_j)$ and the corresponding initial values is given in [Knuth 1981, Marsaglia 1992].

Because of (binary) coding in computers the number of different values x_j is limited and PRNGs are periodic. For many PRNGs the length of this period can be determined analytically or at least estimated [Marsaglia 1992, Knuth 1981]. Besides periodicity the sequences x_j can contain further correlations. For good generators these correlations are quite small and therefore difficult to detect especially because any correlation measure has finite-size effects, that simulate correlations even in truly random finite sequences [Herzel et al. 1994a, Grosse 1995]. Nevertheless these small correlations - especially if sampling high dimensional or fine structured spaces - may lead to spurious simulation results [Ferrenberg et al. 1992].

M.2 What is a random sequence?

Before proceeding to actual tests for PRNGs a more precise definition of *random* has to be given. Even for infinite sequences it is difficult to define *random* in such a way that on the one hand there are *random* sequences and on the other hand no contradiction to the intuitive understanding of *random* arises cf. [Knuth 1981]. When performing empirical tests of PRNGs one is always restricted to finite sequences. It seems impossible to give a proper definition of *random* for a finite sequence because any sequence of given length has equal probability. However, everybody will agree that the decimal sequence 897932384626 appears to be more *random* then 919191919191, 1234567890123 or 000000000000. A *typical* finite sequence of length N is characterized by the fact that all finite subsequences of length k for any $k \leq \log_\lambda N$ will appear with equal probability $p(k)$. Here λ is the size of the alphabet i.e. the number of different letters in the sequence. There are $M = \lambda^k$ different sequences of length k and therefore $p(k) = 1/M$. Sequences that have the desired distribution for a given k are called k-distributed [Knuth 1981].

M.3 An example

Let us consider a PRNG of the widely used class of linear-congruential generators:

$$x_{j+1} = F(x_j) = (a \cdot x_j + c) \bmod m \,, \qquad j > 0 \,, \tag{13.1}$$

[1] True random numbers can be generated e.g. from thermal electron noise [Knuth 1981].

where a, c, m and x_j are integer. The quality of a linear-congruential generator (LCG) depends on the proper choice of the parameter a, c, m, x_0. As an instructive example we consider $a = 137$, $c = 187$, and $m = 256$. For any initial value (13.1) will generate equidistributed (1-distributed) pseudo random numbers of period 256 [Knuth 1981].

In order to randomly sample the fields of a chess board with 8×8 fields one will need 2-distributed coordinates (y_{2j}, y_{2j+1}). These y_j might be generated from the 3 leading bits of x_j i.e. $y_j = x_j >> 5$, where $>>$ denotes the bit-shift operator. It will turn out that the pairs of coordinates (y_{2j}, y_{2j+1}) are not chosen with equal probability. If the coordinates are generated from the last 3 bits of x_j i.e. $y_j = x_j \bmod 8$, one will always select the same field while never reaching any other. The reason for this non-uniform distribution are correlations between x_j and x_{j+1} that cause the pseudo random numbers (x_{2j}, x_{2j+1}) to fill only a fraction[2] of the possible $[0, 255] \times [0, 255]$ space. Therefore the coordinates (y_{2j}, y_{2j+1}) do not properly sample the fields of the chess board. The sequence of y_j coordinates are not 2-distributed and simulations on a 8×8 grid may lead to spurious results.

A better choice of the parameter a, c, m, x_0 (especially larger values for a and m) will improve the situation, but if larger chess boards or higher dimensional spaces are considered the problem will appear again. The effect that was just described is well known for linear-congruential PRNGs [Marsaglia 1968] and it is even possible to determine the maximal distance between neighboring d-tuples $(x_{j+1}, \ldots, x_{j+d})$ with the help of the semi-empirical spectral-test [Knuth 1981]. For PRNGs that are not of the linear-congruential type one can not perform this test.

M.4 Bit-Decay

In order to test whether a sequence of length N is k-distributed one can determine the probability of any subsequence (serial test) or some subsequences (poker test) of length k and compare it with the expectation value $p(k)$ [Knuth 1981]. Another possibility is to determine the number of sequences that did not appear up to length t [Stauffer 1996, Beule & Grosse 1996]. In the chess-board example ($\lambda = 8$, $k = 2$) this means to ask: how many fields have not been selected after the coordinates have been generated t times.

[2]The reason for this is the reconstruction of the attractor of the underlying nonlinear map by means of delay coordinates.

For a large number of fields $M = \lambda^d$ the nature of this test can be understood easily by considering the analogy to the radioactive decay. Radioactive nuclei decay randomly, independently of their position and with a fixed rate. Hence the number of remaining (not decayed) nuclei follows the well known exponential decay law. For a random sequence (using the definition given above) all M cells are selected with the same fixed probability $1/M$. For $M \gg 1$ one has a Poisson process and the expectation value $\langle f(t) \rangle$ for the number of cells, that were not selected after t trials decays exponentially:

$$\langle f(t) \rangle \approx M \exp(-t/M) \qquad \text{for} \qquad 1 \ll M = \lambda^k \; . \tag{13.2}$$

Not k-distributed sequences lead in general to deviations from the exponential decay law (13.2), cf. [Beule & Grosse 1996]. Expectation value and standard deviation of $f(t)$ can be given exactly even for finite M, see section M.7.

The information whether a cell has been selected can be stored by a single bit. Therefore the *bit-decay* can be used as an efficient test of the quality of the PRNG generating the sequence. It allows to test whether a given sequence of pseudo random numbers is k-distributed.

M.5 Expectation Value

In order to calculate the expectation value for $f(t)$ after t trials consider the binary variable $X_i(t)$ $(i = 1, 2, ..., M)$. Let $X_i(t) = 1$ if the i-th cells is empty after t trials and $X_i(t) = 0$ otherwise. The number of empty cells is given by [Feller 1968]

$$f(t) = \sum_{i=1}^{M} X_i(t) \; . \tag{13.3}$$

Therefore the expectation value $\langle f(t) \rangle$ of empty cells after t trials is given by

$$\langle f(t) \rangle = \left\langle \sum_{i=1}^{M} X_i(t) \right\rangle = \sum_{i=1}^{M} \langle X_i(t) \rangle \; . \tag{13.4}$$

The probability of selecting a single fixed cell i in a single trial is $p = 1/M$. Therefore the probability that the cell is empty after t trials is: $A(t) \equiv \langle X_i(t) \rangle = (1-p)^t$. This gives the exponential decay law

$$\langle f(t) \rangle = M \; (1-p)^t = M \; \exp(-b \cdot t) \quad \text{with} \quad b = \ln(M/(M-1)) \; . \tag{13.5}$$

M.6 Variance

The variance of the number of empty cells $f(t)$ is defined by

$$\begin{aligned}
\sigma^2(f(t)) &\equiv \langle f(t)^2\rangle - \langle f(t)\rangle^2 \\
&= \sum_i \langle X_i(t)^2\rangle + \sum_{i\neq j}\langle X_i(t)\cdot X_j(t)\rangle - \sum_{i,j}\langle X_i(t)\rangle\cdot\langle X_j(t)\rangle \qquad (13.6)\\
&= \sum_i (\langle X_i(t)^2\rangle - \langle X_i(t)\rangle^2 + \sum_{i\neq j}(\langle X_i(t)\cdot X_j(t)\rangle - \langle X_i(t)\rangle\cdot\langle X_j(t)\rangle)\\
&= \sum_i (\langle X_i(t)^2\rangle - \langle X_i(t)\rangle\cdot\langle X_j(t)\rangle) + \sum_{i\neq j} cov(X_i(t), X_j(t))\ .
\end{aligned}$$

In each sum every term gives the same contribution because all cells are equal. Therefore one finds for the variance

$$\begin{aligned}
\sigma^2(f(t)) &= M\cdot(\langle X_1(t)^2\rangle - \langle X_1(t)\rangle^2) + M\cdot(M-1)\cdot cov(X_1(t), X_2(t))\\
&= M\cdot(A(t) - A(t)^2) + M\cdot(M-1)\cdot[(1-2p)^t - A(t)^2] \qquad (13.7)\\
&= M\cdot A(t) - (M\cdot A(t))^2 + M\cdot(M-1)\cdot(1-2p)^t.
\end{aligned}$$

The statistics of the *bit-decay* can be related to the finite-size effects of the topological entropy[3] , see [Beule & Grosse 1996] for details.

M.7 Distribution

The probability distribution $P(f,t,M)$ for occupying $M-f$ of the $M=\lambda^k$ cells in t trials is given by [Feller 1968]:

$$P(f,t,M) = \binom{M}{f}\cdot\sum_{j=0}^{M-f}(-1)^j\cdot\binom{M-f}{j}\cdot\left(1-\frac{f+j}{M}\right)^t\ . \qquad (13.8)$$

In the limit $M\to\infty$ the number of occupied cells is given by a Poisson distribution.

M.8 Tests

The behavior of the *bit-decay* for a variety of sequences with different correlations has been discussed in [Beule & Grosse 1996]. Here only the advantages and limitations of two selected PRNGs will be presented:

[3] The limit $q\to 0$ of the Rényi entropies [Rényi 1970].

(i) LCG16807: The linear-congruential PRNG (13.1) with the parameter $a = 16807$, $c = 0$, $m = 2^{31} - 1$, and $x_0 = 1$ defined as the *minimal standard* PRNG in [Press et al. 1992]. The period length of this generator is $2.1 \cdot 10^9$ [Marsaglia 1992].

(ii) RAND55: A so called lagged-Fibonacci PRNG with the iteration

$$x_j = (x_{j-55} - x_{j-24}) \bmod m \,, \tag{13.9}$$

initialized with 55 positive integers and with $m = 2^{32}$ [Knuth 1981]. The length of the period is $7.7 \cdot 10^{25}$ [Marsaglia 1992].

For the stochastic simulations of 2 and 3-dimensional systems one needs random coordinates that are at least 1-distributed, 2-distributed, and 3-distributed. As all modern PRNGs easily pass tests for 1-distribution *bit-decay* tests for $k = 2$ and $k = 3$ for different λ are performed here. The cell coordinates y_j are again generated from subsequent pseudo random integers x_j. It is important to generate y_j from the leading bits of the x_j as this results in a much better distribution of the y_j, cf. section M.3 and [Knuth 1981]. First the sequences LCG16807 and RAND55 are used to sample a chess board of 256×256 fields, i.e. $\lambda = 256$ and $k = 2$. The number of cells that are empty after t trials is shown in Figure M.1 together with the expectation value $\langle f(t) \rangle$ given by the exponential decay (13.5).

In order to assess whether the observed deviations are are significant one has to compare these deviations with the standard deviation $\sigma(f(t))$ obtained from (13.6). This comparison is shown in Figure M.2 For the chosen parameter $\lambda = 256$ and $k = 2$ the deviations are within the expected range.

Thus the pseudo random coordinates generated from LCG16807 and RAND55 for the 256×256 grid can be considered 2-distributed according to this test. The situation changes when one considers finer grids (Figure M.3) or higher dimensions (Figure M.4). In these two figures the deviations $f(t) - \langle f(t) \rangle$ are plotted in units of the standard deviation $\sigma(f(t))$.

For LCG16807 one finds significant deviations from the exponential decay up to $+14\,\sigma$ and $-52\,\sigma$. This shows that this generator is not suitable for properly sampling the spaces under consideration. Similar deviations are already found for smaller λ and may lead to spurious results in simulations [Beule & Grosse 1996]. Lattices where λ is close to a power of 2 are especially prone to this problem. The lagged-Fibonacci generator RAND55 shows a significantly better performance even for these large high dimensional lattices. Therefore RAND55 was chosen as the standard random number generator for all simulation in this work.

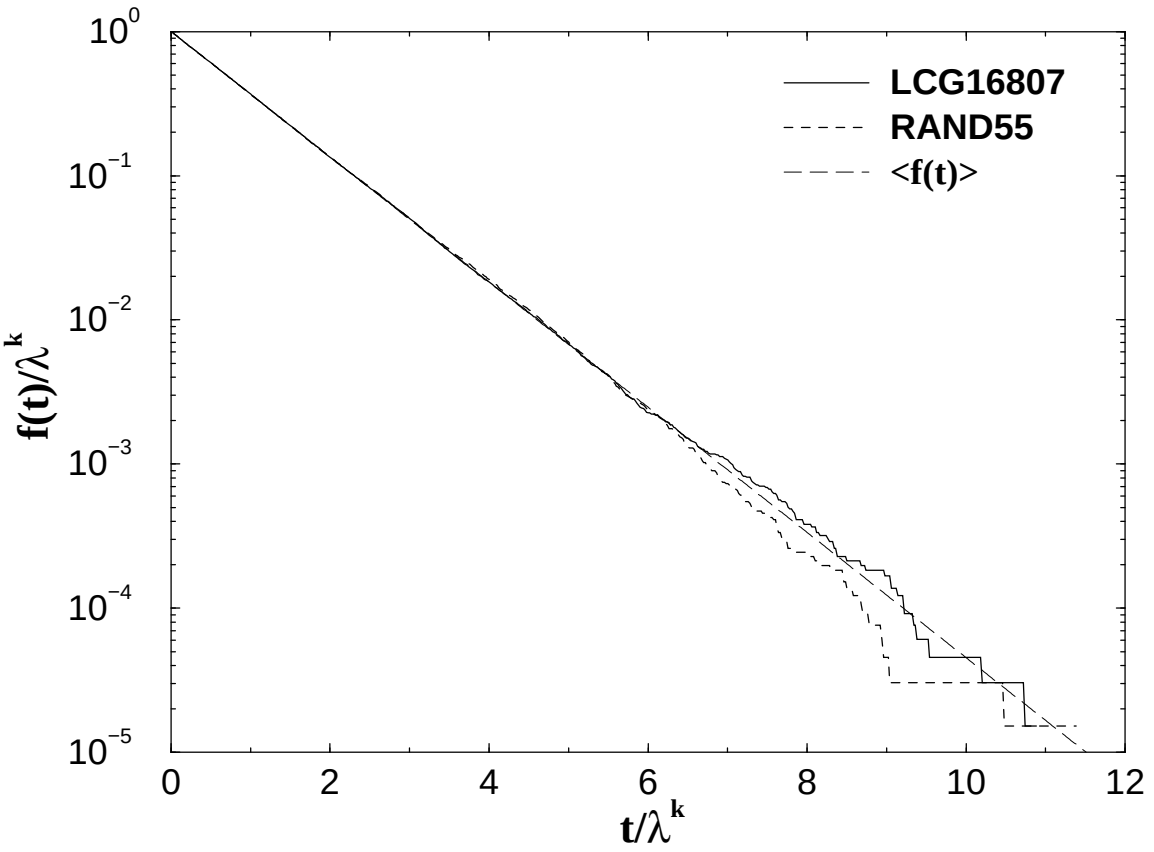


Figure M.1: Bit-decay $f(t)$ for $k = 2$ and $\lambda = 256$ compared with the expected exponential decay (13.5) for the pseudo random sequences RAND55 and LCG16807.

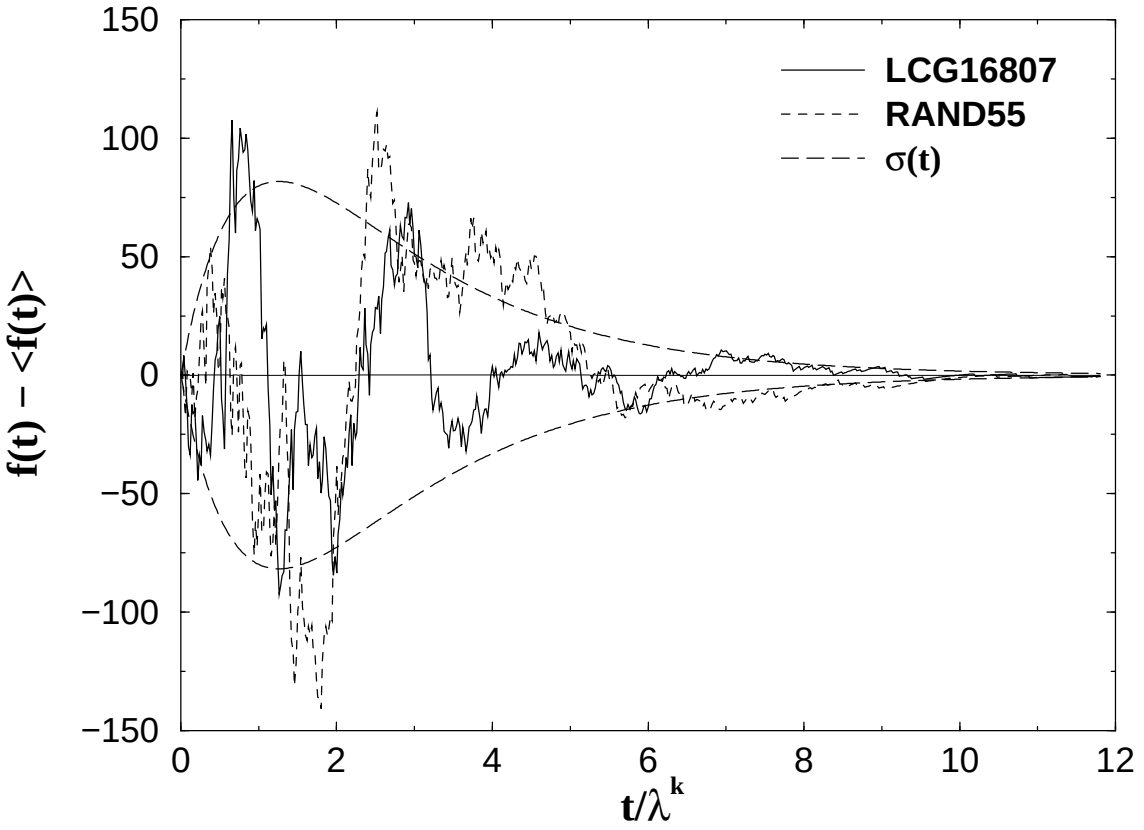


Figure M.2: Bit-decay for $k = 2$ and $\lambda = 256$: deviations of $f(t)$ from the mean $\langle f(t) \rangle$ for the pseudo random sequences RAND55 and LCG16807 compared with the standard deviation $\sigma(f(t))$.

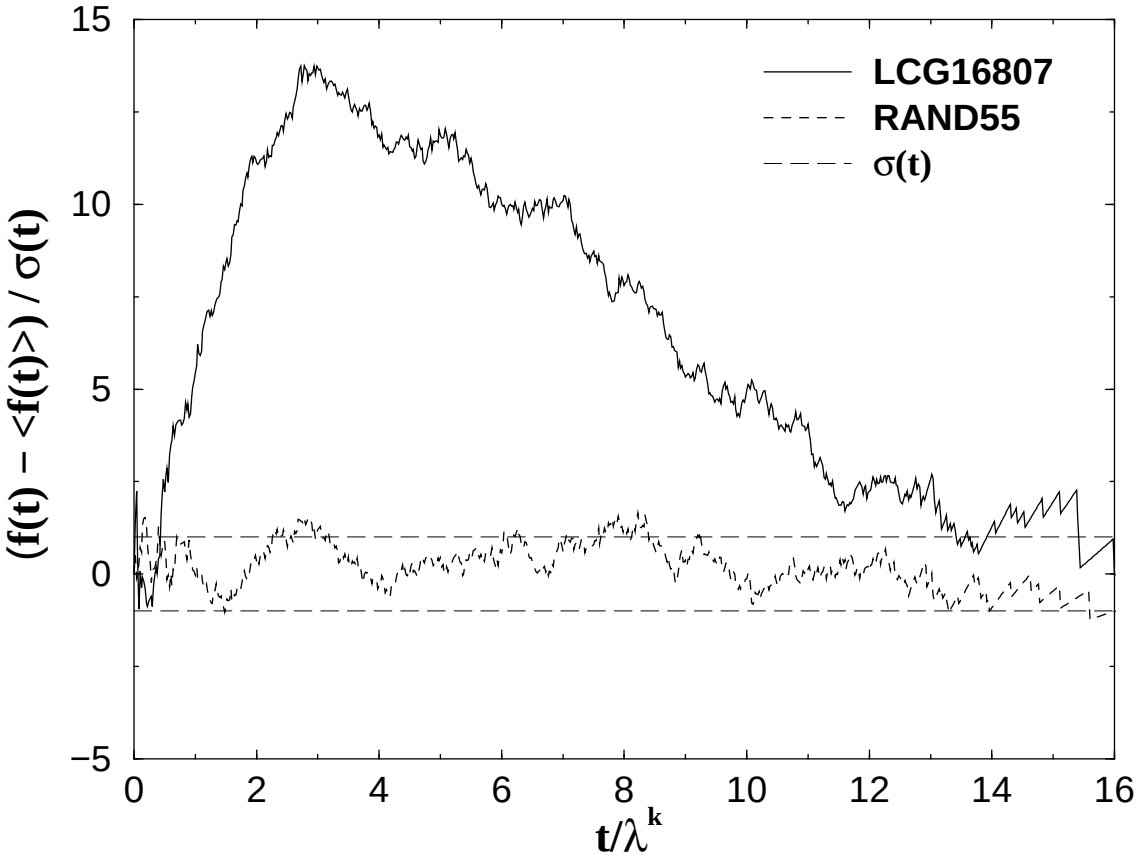


Figure M.3: Bit-decay for $k = 2$ and $\lambda = 3000$: deviations $f(t) - \langle f(t) \rangle$ in units of the standard deviation $\sigma(f(t))$ for RAND55 and LCG16807. Deviations from the exponential decay up to $+14\sigma$ are found for LCG16807. RAND55 only shows deviations in the expected range.

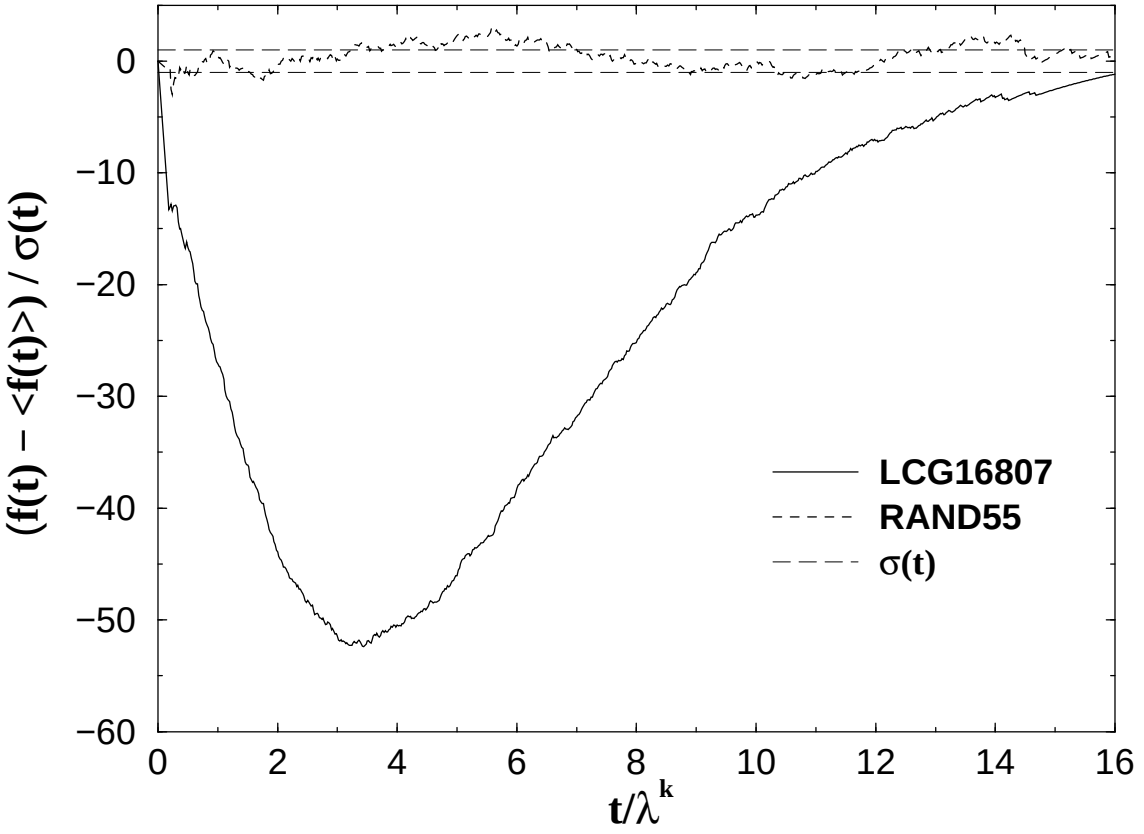


Figure M.4: Bit-decay for $k = 3$ and $\lambda = 300$ deviations $f(t) - \langle f(t) \rangle$ in units of the standard deviation $\sigma(f(t))$ for RAND55 and LCG16807. Deviations from the exponential decay up to -52σ are found for LCG16807. RAND55 only shows deviations in the expected range.

M.9 Conclusions

Weak correlations are always present between subsequent pseudo random numbers due to the deterministic character of PRNGs. When sampling fine structured or high dimensional spaces these correlations may lead to spurious results because the sampling becomes non-uniform. The *bit-decay* method allows to detects weak correlations and thus reveals the limitations of popular PRNGs. It can help in selecting proper PRNGs suitable for the planed stochastic simulations. In conclusion, we agree with Donald Knuth: *"Random number generators should not be chosen at random."*

Bibliography

[Alberts et al. 1994] B. Alberts et al., *Molecular Biology of the Cell* (Garland Publishing, New York, 1994).

[Allen & Tildesley 90] M.P. Allen and D.J. Tildesley, *Computer simulation of liquids*, Clarendon Press, Oxford, 1990.

[Altschul 1991] Altschul, S. F. (1991) *J. Mol. Biol.* **219**, 555–565.

[Amalgor 1985] Amalgor, H., "Nucleotide distribution and the recognition of coding regions in DNA sequences: an information theory approach", *J. theor. Biol.* **117** (1985), 127–136.

[Anderson et al. 1998] S. G. E. Anderson *et al.* , Nature **396**, 133 (1998).

[Arneodo 1995] A. Arneodo, E. Bacry, P. V. Graves, and J. F. Muzy, Phys. Rev. Lett. 74, 3293 (1995).

[Basharin 1959] Basharin, G. P. *Theory Prob. Appl.* **4** 333 (1959).

[Beck & Schlögl 1993] Beck, C. & F. Schlögl, *Thermodynamics of chaotic systems*, Cambridge University Press (1993).

[Bell 1992] Bell, G. I. *Computers Chem.* **16** 135 (1992).

[Bell & Marr 1990] G. Bell and T. Marr (eds.), *Computers and DNA* (Addison-Wesley, Reading, 1990).

[Berger 1985] Berger, O. *Statistical Decision Theory and Bayes Analysis* (New York, Springer, 1985).

[Berg & Singer 1992] Berg, P. & M. Singer, *Dealing with Genes*, University Science Books (1992).

[Bernaola-Galván et al. 1996] P. Bernaola-Galván, R. Román-Roldán, and J. L. Oliver, Phys. Rev. E **53**, 5181 (1996).

[Bernaola-Galván et al. 1999] P. Bernaola-Galván, J. L. Oliver and R. Román-Roldán, Phys. Rev. Lett. (to appear).

[Bernardi 1989] G. Bernardi, *Ann. Rev. Genet.* **23**, 637–661 (1989) ;

[Bernardi et al. 1985] G. Bernardi, B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Meunier-Rotival, and F. Rodier, Science **228** (1985) 953.

[Bernasconi 1987] Bernasconi, J., *J. Physique* **48** 559 (1987).

[Bettecken et al. 1992] T. Bettecken, R. Aussani, C. R. Müller, and G. Bernardi, Gene **122** (1992) 329.

[Beule & Grosse 1996] D. Beule and I. Grosse, Dynamik, Evolution, Strukturen J.A. Freund, ed., Verlag Dr. Köster, Berlin, 1996, p. 161.

[Bibb et al. 1984] Bibb, M.L., Findlay, P.R. & Johnson, M.W. (1984). The relationship between base composition and codon usage in bacterial genes and its use for the simple and reliable identification of protein-coding sequences. *Gene* **30,** 157–166.

[Boghosian 1996] Boghosian B M 1996 *Phys. Rev.* E **53** p 475

[Bolshoy et al. 1996] Bolshoy, A., Ioshikhes, I. and Trifonov, E. N., CABIOS **12** (1996) 383.

[Borodovsky & McIninch 1993] M. Borodovsky and J. McIninch, *J. Mol. Biol.* **268**, 1–17 (1993);

[Borštnik et al. 1993] B. Borštnik, D. Pumpernik, and D. Lukman, Europhys. Lett. 23, 389 (1993).

[Bronstein & Semendjajew 1989] Bronstein, I. N. & K. A. Semendjajew, *Taschenbuch der Mathematik*, Verlag Nauka, Moskau (1989).

[Buldyrev et al. 1993] S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. E 47, 4514 (1993).

[Buldyrev et al. 1993] Buldyrev, S. V., A. L. Goldberger, S. Havlin, C.-K. Peng, H. E. Stanley & M. Simons, *Biophys. J.*, **65** 2673 (1993).

[Bulmer 1987] Bulmer, M. (1987) *Nature* **325**, 728–730.

[Burge & Karlin 1997] C. Burge and S. Karlin, *J. Mol. Biol.* **268**, 78–94 (1997).

[Burge & Karlin 1998] Burge, C. & Karlin, S. (1998) *Curr. Opin. Struct. Biol.* **8**, 346–354.

[Burset & Guigó 1996] M. Burset and R. Guigó, *Genomics* **34**, 353–367 (1996).

[Carpena & Bernaola-Galván 1999] P. Carpena and P. Bernaola-Galván, Phys. Rev. B **60**, 201 (1999).

[Chame & Mello 1994] Chame A and de Mello E V L 1994 *J. Phys.* A **27** p 3363.

[Chame & Mello 1997] Chame A and de Mello E V L 1997 *Phys. Lett.* **A 228** p 159.

[Chase et al. 1998] K. C. Chase, P. Bhattacharyya, and A. Z. Mekjian, Phys. Rev. C **57**, 822 (1998).

[Chechetkin & Turygin 1994] V. R. Chechetkin and A. Yu. Turygin, J. Phys. A: Math. Gen. 27, 4875 (1994).

[Churchill 1989] G. A. Churchill, Bull. Math. Biol. 51, 79 (1989).

[Claverie 1997] J.-M. Claverie, *Hum. Mol. Gen.* **6**, 1735–1744 (1997).

[Clay et al. 1995] O. Clay, W. Schaffner, and K. Matsuo, Somatic Cell and Mol. Gen. **21** (1995) 91.

[Cover & Thomas 1989] Cover, T. M. & Thomas, J. A. (1989) *Elements of Theory* (Wiley, New York).

[Cramer 1946] The mathematical proof can be found in: H. Cramer, *Mathematical Methods of Statistics* (Princeton Univ. Press, Princeton, 1946). An intuitive heuristic argument of why the number of degrees of freedom is equal to 6 is that there are $4+3-1$ independent linear constraints that the $4 \times 3 = 12$ numbers $p_i^{(m)} - p_i$ must satisfy. Hence, the number of degrees of freedom is $4 \times 3 - (4+3-1) = (4-1) \times (3-1) = 6$.

[Creighton 1993] T. E. Creighton, *Proteins: structure and molecular properties* 2nd ed.: (W. H. Freeman, 1993).

[Crick 1966] Crick, F. H. C., *J. Mol. Biol.* **19** 548 (1966).

[Curado & Tsallis 1991] Curado, E. M. F. and C. Tsallis, "Generalized statistical mechanics: connection with thermodynamics", *J. Phys. A* **24** (1991), L69-L72.

[Denisov et al. 1997] D. A. Denisov, E. S. Shpigelman, and E. N. Trifonov, *Protective nucleosome centering at splice sites as suggested by sequence-directed mapping of the nucleosomes*, Gene, submitted.

[Derrida 1997] For a review of non-self-averaging phenomena, see B. Derrida, Physica D **107**, 186 (1997).

[Derrida & Flyvbjerg 1987] B. Derrida and H. Flyvbjerg, J. Phys. A **20**, 5273 (1987).

[Derrida & Flyvbjerg 1988] B. Derrida and H. Flyvbjerg, J. Physique **48**, 971 (1987); B. Derrida and D. Bessis, J. Phys. A **21**, L509 (1988).

[Derrida & Jung-Muller 1999] B. Derrida and B. Jung-Muller, J. Stat. Phys. **94**, 277 (1999).

[Deutsch & Long 1999] Deutsch, M. & Long, M. (1999). Intron–exon structures of eukaryotic model organisms. *Nucl. Acids Res.* **27,** 3219–3228.

[Dewey 1997] Dewey, T. G. (1997) *Fractals in Molecular Biophysics* (Oxford University Press, Oxford).

[Dong & Searls 1994] Dong, S. & Searls, D. B. (1994) *Genomics* **23**, 540–551.

[Dreismann & Larhammer 1993] C. A. C. Dreismann and D. Larhammer, Nature 361, 212 (1993).

[Dujon et al. 1994] Dujon, B. et al., *Nature* **369** 371 (1994).

[Ebeling 1993] Ebeling, W., Chaos, Entropie und Sequenzanalyse, in: R. Hofestädt, F. Krückeberg & T. Lengauer (Eds.), *Informatik in den Biowissenschaften*, Springer-Verlag, Berlin (1993).

[Ebeling et al. 1987] W. Ebeling, R. Feistel, and H. Herzel, Physica Scripta **35** (1987) 761.

[Ebeling et al. 1995] Ebeling, W., T. Pöschel, and K.-F. Albrecht, "Entropy, transinformation and word distribution of information-carrying sequences", *Int. J. Bif. & Chaos* **5** (1995), 51–61.

[Ebeling & Feistel 1982] W. Ebeling and R. Feistel, Physik der Selbstorganisation und Evolution (Akademie Verlag, Berlin, 1982).

[Ebeling & Neiman 1994] Ebeling, W. & A. Neiman, On the origin of long-range correlations in human writings: computer experiments with mixing on word and sentence level, submitted to *Phys. Rev. Lett.* (1994a).

[Ebeling & Nicolis 1991] W. Ebeling and G. Nicolis, Europhys. Lett. 14 (1991) 191.

[Ebeling & Nicolis 1992] W. Ebeling and G. Nicolis, Chaos, Solitons & Fractals 2 (1992) 635.

[Ebeling & Pöschel 1994] Ebeling, W. & T. Pöschel, *Europhys. Lett.* **26** 241 (1994c).

[Eckmann & Ruelle 1985] J. P. Eckmann and D. Ruelle, Rev. Mod. Phys. 57 (1985) 617.

[Elton 1974] R. A. Elton, J. Theor. Biol. **45** (1974) 533.

[Farber et al. 1992] Farber, R. , A. Lapedes & K. Sirotkin, *J. Mol. Biol.* **226** 471 (1992).

[Farmer 1982] J. D. Farmer, Z. Naturforsch. 37a (1982) 1304.

[Feldmann et al. 1994] H. Feldmann et al., EMBO J. 13, 5795 (1994).

[Feller 1968] W. Feller, *An Introduction to Probability Theory and its Applications*, 3rd ed., John Wiley & Sons, New York, 1968.

[Ferrenberg et al. 1992] A.M. Ferrenberg, D.P. Landau, and Y.J. Wong, Phys. Rev. Lett. **69** (1992), 3382.

[Fickett 1982] J. W. Fickett, *Nucl. Acids Res.* **10**, 5303–5318 (1982).

[Fickett 1992] Fickett, J.W. & Tung, C.-S. (1992). Assessment of protein coding measures. *Nucleic Acids Res.* **20,** 6441-6450.

[Fickett 1996] J. W. Fickett, *Comput. Chem.* **20**, 103–118 (1996);

[Fickett 1998] J. W. Fickett, in *Bioinformatics*, ed. by A. D. Baxevanis and B. F. Francis-Oullette (John Wiley and Sons, New York, 1998).

[Fickett et al. 1992] J. W. Fickett, D. C. Torney, and D. R. Wolf, Genomics 13, 1056 (1992).

[Fickett & Tung 1992] Fickett, J. W. & Tung C.-S., *Nucleic Acids Research* **20** 6441 (1992).

[Fiers & Grosjean 1979] Fiers, W. & Grosjean, H. (1979) *Nature* **277**, 328–328.

[Fisz 1989] Fisz, M., *Wahrscheinlichkeitsrechnung und mathematische Statistik*, VEB Deutscher Verlag der Wissenschaft, Berlin (1989).

[Fleischman et al. 1995] Fleischman, R. *et al.*, "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd", *Science* **269** (1995), 496–512.

[Flyvbjerg & Kjaer] H. Flyvbjerg and N. J. Kjaer, J. Phys. A **21**, 1695 (1988).

[Frachebourg et al. 1995] L. Frachebourg, I. Ispolatov, and P. L. Krapivsky, Phys. Rev. E **52**, R5727 (1995).

[Fraser & Swinney 1986] A. M. Fraser and H. L. Swinney, Phys. Rev. A 33 (1986) 1134.

[Gatlin 1972] L. L. Gatlin, *Information Theory and the Living System* (Columbia Univ. Press, New York, 1972).

[Gelfand 1995] Gelfand, M. S. (1995) *J. Comp. Biol.* **2**, 87–115.

[Gelfand & Roytberg 1993] M. S. Gelfand and M. A. Roytberg, *BioSystems* **30**, 173–182 (1993).

[GenBank 1999] GenBank, release 111, 15 April 1999.

[Gnedenko 1981] Gnedenko, B. W., *Einführung in die Wahrscheinlichkeitsrechnung*, Akademie-Verlag, Berlin (1981).

[Grantham et al. 1981] R. Grantham, C. Gautier, M. Gouy, M. Jacobzone and R. Mercier, Nucleic Acids Res. **9**, R43 (1981).

[Grassberger 1986] Grassberger, P., *Int. J. Theor. Phys.* **25** 907 (1986).

[Grassberger 1988] Grassberger, P., *Phys. Lett. A* **128** 369 (1988)

[Grassberger 1989] Grassberger, P., preprint WU B 89-26, Universität Wuppertal (1989).

[Grassberger 1994] Grassberger, P., private communication (1994).

[Grassberger et al. 1991] Grassberger, P, T. Schreiber, and C. Schaffrath "Nonlinear time sequence analysis", *Int. J. Bif. & Chaos* **1** (1991), 521–543.

[Grassberger & Procaccia 1983] Grassberger P and Procaccia I 1983 *Physica* D **9** p 189

[Griffiths et al. 1993] Griffiths, A. J. F., J. H. Miller, D. T. Suzuki, R. C. Lewontin, W. C. Lewontin & W. M. Gelbart, *An Introduction to Genetic Analysis*, W. H. Freeman, New York (1993).

[Grosberg 1993] A. Yu. Grosberg, Y. Rabin, S. Havlin, and A. Nir, Europhys. Lett. 23 (1993) 373.

[Grosse 1995] Grosse I 1995 *Statistical Analysis of Biosequences* (Humboldt-University Berlin: Thesis)

[Grosse 1996] Grosse, I., *Estimating Entropies from Finite Samples*, in *Dynamik, Evolution, Strukturen*, ed. J. A. Freund, Berlin: Verlag Köster (1996).

[Grosse 1999] I. Grosse, H. Herzel, S. V. Buldyrev, and H. E. Stanley, *Mutual Information of Coding and Noncoding DNA*, submitted.

[Grosse et al. 1999] Grosse, I., Holste, D., Buldyrev, S.V., Stanley, H.E. & Herzel, H. (1999a). Universality of positional information content in coding and non-coding DNA. Submitted.

[Grosse et al. 1999] Grosse, I., Buldyrev, S.V., Stanley, H.E., Holste, D. & Herzel, H. (1999b). Average mutual information of coding and non-coding DNA. *Pacific Symposium on Biocomputing 2000* **5,** 611–620.

[Grosse et al. 1999] I. Grosse, H. Herzel, S. V. Buldyrev, and H. E. Stanley, *Species Independence of Mutual Information in Coding and Noncoding DNA*, Phys. Rev. E, in press.

[Grossmann & Thomae 1977] S. Grossmann and S. Thomae, Z. Naturforsch. 32a (1977) 1353.

[Guigó 1999] Guigó, R. (1999). DNA composition, codon usage and exon prediction. *Nucleic Acid and Protein Databases*, M. Bishop (Ed.), Academic Press.

[Guigó et al. 1992] Guigó, R., Knudsen, S., Drake, N. & Smith T. (1992). Prediction of gene structure. *J. Mol. Biol.* **226,** 141–157.

[Guigó & Fickett 1995] Guigó, R. & Fickett, J.W. (1995). Distinctive sequence features in protein coding, genic non-coding, and intergenic human DNA. *J. Mol. Biol.* **253,** 51–60.

[Halsey et al. 86] Halsey T C, Jensen M H, Kadanoff L P, Procaccia I and Shraiman B I 1986 *Phys. Rev.* A **33** p 1141

[Hamming 1980] Hamming, R. W., *Coding and Information Theory*, Prentice-Hall, Englewood Cliffs, N. J. (1980).

[Harris 1975] Harris, B., *Colloquia Mathematica Societatis Janos Bolyai*, Keszthely **16** 323 (1975).

[Harris 1989] T. E. Harris, *The theory of branching processes* (Dover, New York, 1989).

[Herzel 1988] Herzel, H., *Sys. Anal. Mod. Sim.* **5** 435 (1988).

[Herzel 1989] Herzel, H., *Biomathematik und Bioinformatik I*, Mathematik-Naturwissenschaften-Manuskripte, Humboldt-Universität, Berlin (1989).

[Herzel et al. 1994] H. Herzel, W. Ebeling, and A. O. Schmitt, Phys. Rev. E 50 (1994) 5061.

[Herzel & Ebeling 1985] H. Herzel and W. Ebeling, Phys. Lett. A 111, 1 (1985).

[Herzel et al. 1994a] Herzel, H., A. O. Schmitt & W. Ebeling, *Chaos, Solitons & Fractals* **4** 97 (1994a).

[Herzel et al. 1994b] Herzel, H., W. Ebeling & A. O. Schmitt, *Phys. Rev. E.* **50** 5061 (1994b).

[Herzel et al. 1995] H. Herzel, W. Ebeling, A. O. Schmitt, and M. A. Jiménez-Montaño, in *From Simplicity to Complexity in Chemistry and Beyond*, eds. A. Müller, A. Dress, and F. Vögtle (Vieweg, Braunschweig, 1995).

[Herzel & Grosse 1995] H. Herzel and I. Grosse, Physica A, **216**, 518 (1995).

[Herzel & Grosse 1997] H. Herzel and I. Grosse, Phys. Rev. E **55** (1997) 800.

[Higgs 1995] P. G. Higgs, Phys. Rev. E **51**, 95 (1995).

[Holde 1988] K. E. van Holde, *Chromatin* (Springer Verlag, Town, 1988).

[Holste 1997] Holste, D. (1997) *Entropy Estimators and their Limitations: The Statistical Analysis of Symbol Sets* (Humboldt-University Berlin: Thesis)

[Holste et al. 1998] Holste, D., I. Grosse, and H. Herzel, "Bayes estimators of generalized entropies", *J. Phys. A* **31** (1998), 2551–2566.

[Hutchinson 1992] Hutchinson, G. B. & Hayden, M. R. (1992) *Nucl. Acids Res.* **20**, 3453–62.

[Ikemura 1981] T. Ikemura, *J. Mol. Biol.* **146**, 1–21 (1981);

[Ikemura et al. 1990] T. Ikemura, K.-N. Wada, and S.-I. Aota, Genomics 8, 207 (1990).

[Ioshikhes et al. 1992] I. Ioshikhes, A. Bolshoy, and E. N. Trifonov, J. Biomol. Struct. Dyn. 9, 1111-1117 (1992).

[Ioshikhes et al. 1996] I. Ioshikhes, A. Bolshoy, K. Derenshteyn, M. Borodovsky, and E. N. Trifonov, J.Molec. Biol. **262** (1996) 129.

[Jaglom & Jaglom 1965] Jaglom, A. M. & I. M. Jaglom, *Wahrscheinlichkeit und Information*, Verlag der Wissenschaften, Berlin (1965).

[Jaynes 1983] Jaynes, E. T., Where do we stand on maximum entropy?, in: R. D. Levine & M. Tribus (Eds.), *The Maximum Entropy Formalism*, M. I. T. Press, Cambridge (1983).

[Johnson & Kotz 1970] N. L. Johnson and S. Kotz, *Distributions in Statistics: Continuous Univariate Distributions* (Boston, Houghton-Mifflin-Company, 1970).

[Justice 1986] Justice, J. H. (ed.), *Maximum Entropy and Bayes Methods in Applied Statistics*, Cambridge University Press, Cambridge (1986).

[Kadono 1997] T. Kadono, Phys. Rev. Lett. **78**, 1444 (1997).

[Kanehisa & Tsong 1980] M. I. Kanehisa and T. Y. Tsong, Biopolymers **19** (1980) 1617.

[Karlin & Brendel 1992] S. Karlin and V. Brendel, Science 257 (1992) 39.

[Karlin & Brendel 1993] S. Karlin and V. Brendel, Science 259, 677 (1993).

[Karlin & Mrazek 1996] Karlin, S. & Mrazek, J. (1996) *J. Mol. Biol.* **262**, 459–472.

[Karlin & Mrazek 1997] Karlin, S. & Mrazek, J. (1997) *Proc. Natl. Acad. Sci. USA* **94**, 10227–10232.

[Kauffman 1993] S. A. Kauffman, *The Origin of Order: Self-Organization and Selection in Evolution* (Oxford University Press, London, 1993).

[Khinchin 1957a] Khinchin, A. I., *Mathematical Foundations of Information Theory*, Dover Publ., New York (1957a).

[Khinchin 1957b] Khinchin, A. I., Der Begriff der Entropie in der Wahrscheinlichkeitsrechnung, in: H. Grell (Ed.), *Arbeiten zur Informationstheorie*, VEB Deutscher Verlag der Wissenschaften, Berlin (1957b).

[Kleffe et al. 1998] J. Kleffe et al., *Bioinformatics* **14**, 232–243 (1998).

[Knippers et al. 1990] Knippers, R., P. Philippsen, K. P. Schäfer, E. Fanning, *Molekulare Genetik*, Georg Thieme Verlag, Stuttgart (1990).

[Knuth 1981] D. Knuth, *The Art of Computer Programming, Vol. 2 Seminumerical Algorithms*, 2nd edition ed., Addison Wesley, Reading, 1981.

[Kolchanov & Lim 1994] N. A. Kolchanov and H. A. Lim, *Computer Analysis of Genetic Macromolecules: Structure, Function and Evolution* (World Scientific Publ., Singapore, 1994).

[Kolmogorov 1958] , Kolmogorov, A. N., *Dokl. Akad. Sci. USSR* **119** 861 (1958).

[Konopka & Smythers 1987] A. K. Konopka and G. M. Smythers, CABIOS 3, 193 (1987).

[Korenberg & Rykowski 1988] J. R. Korenberg and M. C. Rykowski, Cell **53** (1988) 391.

[Krapivsky & Ben-Naim 1996] P. L. Krapivsky and E. Ben-Naim, Phys. Rev. E **50**, 3502 (1994); P. L. Krapivsky and E. Ben-Naim, Phys. Lett. A **196**, 168 (1994); E. Ben-Naim and P. L. Krapivsky, Phys. Rev. Lett. **76**, 3234 (1996).

[Krauth & Mezard 1995] W. Krauth and M. Mezard, Zeitschr. Phys., in press.

[Kullback 1959] S. Kullback, *Information Theory and Statistics* (Wiley, New York, 1959).

[Kullback 1991] Kullback, S., *Annals Mathem. Statistics* **22** 79 (1951).

[Kulp et al. 1997] Kulp, D., Haussler, D., Reese, M.G. & Eeckman, F.H. (1997). Integrating database homology in a probabilistic gene structure model. *Pacific Symposium on Biocomputing 1997*, Altman, R.B., Dunker, A.K., Hunter, L, Klein, T.E. (Eds.), Hawaii, World Scientific.

[Kurths & Herzel 1987] Kurths J and Herzel H 1987 *Physica* D **25** p 167

[Lapedes et al. 1990] A. Lapedes, C. Barnes, C. Burks, R. Farber, and K. Sirotkin, in *Computers and DNA*, eds. G. Bell and T. Marr (Addison-Wesley, Reading, 1990).

[Larhammer & Chatzidimitriou-Dreisman 1993]
D. Larhammer and A. A. Chatzidimitriou-Dreismann Nucl. Acids Res. **21** (1993) 5167.

[Lawn & Wilshaw 1975] B. R. Lawn and T. R. Wilshaw, *Fracture of Brittle Solids* (Cambridge University Press, Cambridge, 1975).

[Leven et al. 1989] Leven, R. W., B.-P. Koch & B. Pompe, *Chaos in dissipativen Systemen*, Akademie-Verlag, Berlin (1989).

[Levitin & Feingold 1994] L. B. Levitin and Z. Feingold, Chaos, Solitons & Fractals 4 (1994) 709.

[Levitin & Reingold 1978] Levitin L B and Reingold R 1978 *An Improved Estimate for the Entropy of a Discrete Random Variable*, *Annual Conference of the Israel Statistical Association*, Tel Aviv

[Lewin 1997] B. Lewin, *Genes VI* (Oxford Univ. Press, Oxford, 1997).

[Li 1989] Li, W., Santa Fe Institute preprint 89-008 (1989).

[Li 1990] W. Li, J. Stat. Phys. 60, 823 (1990).

[Li 1991] Li, W., *Complex Systems* **5** 381 (1991).

[Li 1992] Li, W., *Int. J. of Bif. and Chaos* **2** 1 (1992).

[Li 1997] Li, W., "The study of correlation structures of DNA sequences: a critical review", *Computer & Chem.* **21** (1997), 257–272.

[Li et al. 1994] W. Li, T. G. Marr, and K. Kaneko, Physica D 75, 392 (1994).

[Li et al. 1998] W. Li, G. Stolovitzky, P. Bernaola-Galván, and J. L. Oliver, Genome Research **8**, 916 (1998).

[Li & Kaneko 1992] Li, W. & K. Kaneko, *Europhys. Lett.* **17** 655 (1992).

[Lin 1991] J. Lin, IEEE Trans. Inf. Theor. **37**, 145 (1991).

[Liu et al. 1995] K. Liu, E. P. Sandgren, R. D. Palmiter, and A. Stein, Proc. Natl. Acad. Sci. USA **92** (1995) 7724.

[Lodish et al. 1995] H. Lodish et al., *Molecular Cell Biology* (Freeman, New York, 1995).

[Lucena et al. 1995] Lucena L S, da Silva L R and Tsallis C 1995 *Phys. Rev.* E **51** p 6247

[Lukashin & Borodovsky 1998] Lukashin, A. V. & Borodovsky, M. (1998) *Nucl. Acids Res.* **26**, 1107–1115.

[Luo & Li 1991] L. Luo and H. Li, Bull. Math. Biol. 53, 345 (1991).

[MacKay & Peto 1994] MacKay, D. J. C. and L. C. B. Peto, "A hierarchical Dirichlet language model", *Natural Lang. Eng.* **1** (1994).

[Mackey 1989] M. C. Mackey, Rev. Mod. Phys. 61 (1989) 981.

[Mantegna et al. 1994] Mantegna, R. N., S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons & H. E. Stanley, *Phys. Rev. Lett.* **73** 3169 (1994).

[Mantegna et al. 1995] Mantegna, R. N., S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons & H. E. Stanley, Linguistic Analysis of Noncoding DNA, submitted to *Phys. Rev. E.* (1995).

[Marini et al. 1982] J. C. Marini, S. D. Levene, D. M. Grothers, and P. T. England, PNAS **79** (1982) 7664.

[Marsaglia 1968] G. Marsaglia, Proceedings of the National Academy of Science **61** (1968), 25.

[Marsaglia 1992] G. Marsaglia, Proceedings of Symposia in Applied Mathematics Vol. 46 G. Andrews S. Burr, ed., American Mathematical Society, Providence, 1992, p. 73.

[Martin 1971] Martin, B. R., *Statistics for Physicists*, Academic Press, London (1971).

[McEliece 1977] McEliece, R. J., *The theory of information and coding*, Addison-Wesley, New York (1977).

[McKusick 1998] V. A. McKusick, Genomics **45**, 244 (1997); Science (Special Genome Issue) **282** (1998).

[McMillan 1953] McMillan, B., *Ann. Math. Statist.* **24** 196 (1953).

[Mezard et al. 1987] M. Mézard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).

[Munson 1992] P. J. Munson, R. C. Taylor, and G. S. Michaels, Nature 360, 636 (1992).

[Murakami & Takagi 1998] Murakami, K. & Takagi, T. (1998). Gene recognition by combination of several gene-finding programs. *Bioinformatics* **14,** 665–675.

[Nakamura et al. 1996] Y. Nakamura et al., *Nucl. Acids Res.* **24**, 214–215 (1996).

[Nelson et al. 1999] Nelson, K.E., Clayton, R.A., Gill, S.R., Gwinn, M.L. *et al.* (1999). Evidence for lateral gene transfer between archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature* **399,** 323–329.

[Nesti et al. 1995] Nesti, C., Poli, G., Chicca, M., Ambrosino, P., Scapoli, C. & Barrai, (1995) *Comput. Appl. Biosci.* **11**, 167–171.

[Nowak 1994] Nowak, R., *Science* **263** 608 (1994).

[Oddershede et al. 1993] L. Oddershede, P. Dimon, and J. Bohr, Phys. Rev. Lett **71**, 3107 (1993).

[Oliver et al. 1992] Oliver, S. G. *et al.*, *Nature* **357** 38 (1992).

[Oliver et al. 1999] J. L. Oliver, R. Román-Roldán, J. Pirez, and P. Bernaola-Galván, Bioinformatics (in press).

[Ossadnik et al. 1994] Ossadnik, S. M., S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, C.-K. Peng, M. Simons & H. E. Stanley, *Biophys. Journal* **67** 64 (1994).

[Pawelzik & Schuster 1987] Pawelzik, K. and H. G. Schuster, "Generalized dimensions and entropies from measured time series", *Phys. Rev. A* **35** (1987), 481–484.

[Peng et al. 1992] C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortina, M. Simons, and H. E. Stanley, Nature **356** (1992) 186.

[Peng et al. 1993] Peng, C.-K., S. V. Buldyrev, A. L. Goldberger, S. Havlin, M. Simons & H. E. Stanley, *Phys. Rev. E.* **47** (5) 3730 (1993).

[Peng et al. 1994] Peng, C.-K., S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley & A. L. Goldberger, *Phys. Rev. E.* **49** 2 (1994).

[Penna 1995] Penna T J P 1995 *Phys. Rev.* E **51** p 34.

[Plastino 1994] Plastino A R, Plastino A and Tsallis C 1994 *J. Phys.* A **27** p 5707.

[Pompe 1993] Pompe, B., "Measuring statistical dependences in a time series", *J. Stat. Phys.* **73** (1993), 587–610.

[Pompe 1994] Pompe, B., *Chaos, Solitons & Fractals*, **4** 83 (1994).

[Pompe et al. 1986] Pompe, B., J. Kruscha & and R. W. Leven, *Z. Naturforsch.* **41a** 801 (1986).

[Pöschel et al. 1994] Pöschel, T., W. Ebeling & H. Rosé, Guessing probability distributions from small samples, submitted to *Phys. Rev. E.* (1994).

[Press et al. 1992] Press, W.H., Teukolsky, S.A., Vetterling, W.T., & Flannery, B.P. (1992). *The Art of Scientific Computing: Numerical Recipes in C.* Second Edition. Cambridge University Press, U.S.A.

[Ramshaw 1995] Ramshaw J D 1995 *Phys. Lett.* **A 198** pp 119 & 122

[Redner 1990] For a general review of fragmentation, see e.g. S. Redner, in: *Statistical Models for the Fracture of Disordered Media*, ed. H. J. Herrmann and S. Roux (Elsevier Science, New York, 1990).

[Rényi 1970] Rényi, A., *Probability Theory*, Amsterdam: North Holland (1970).

[Rényi 1980] Rényi, A., *Tagebuch über die Informationstheorie*, VEB Deutscher Verlag der Wissenschaften, Berlin (1982).

[Román-Roldán et al. 1998] R. Román-Roldán, P. Bernaola-Galván and J. L. Oliver, Phys. Rev. Lett. **80**, 1344 (1998).

[Ruelle 1989] Ruelle D 1989 *Chaotic Evolution and Strange Attractors* (Cambridge: Cambridge University Press)

[Sachs 1984] Sachs, L. (1984). *Applied Statistics.* Springer-Verlag, New York.

[Salzberg et al. 1997] Salzberg, S., Delcher, A., Fasman, K. & Henderson, J. (1997) *Technical Report 1997-03* (Department of Computer Science, Johns Hopkins University).

[Salzberg et al. 1998] Salzberg, S.L., Delcher, A.L., Kasif, S, & White, O. (1998). Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.* **15,** 544–548.

[Schenkel 1993] A. Schenkel, J. Zhang, Y.-C. Zhang, Fractals 1 (1993) 47.

[Schmitt et al. 1993] Schmitt, A. O., H. Herzel & W. Ebeling, *Europhys. Lett.* **23** 303 (1993).

[Schmitt et al. 1996] A. O. Schmitt, W. Ebeling, and H. Herzel, BioSystems, 37, 199 (1996).

[Schmitt et al. 1997] A. O. Schmitt, E. Kolker, and E. N. Trifonov, *Hidden Structure of Protein Sequences,* in preparation.

[Schnakenberg 1995] J. Schnakenberg, *Algorithmen in der Quantentheorie und Statistischen Physik*, Verlag Zimmermann-Neufang, Ulmen, 1995.

[Schneider 1997] Schneider, T.D. (1997). Information content of individual genetic sequences. *J. theor. Biol.* **189,** 427–441.

[Schneider et al. 1986] Schneider, T. D., G. D. Stormo, and L. Gold, "Information content of binding sites on nucleotide sequences", *J. Mol. Biol.* **188** (1986), 415–431.

[Schürmann & Grassberger 1996] Schürmann T. and P. Grassberger, "Entropy estimation of symbol sequences", *CHAOS* **6** (1996), 414–427.

[Searls 1997] Searls, D.B. (1998). Grand challenges in computational biology. *Computational Methods in Molecular Biology*, Salzberg, S.L., Searls, D.B., Kasif, S. (Eds.), Elsevier Science B.V., Amsterdam, The Netherlands.

[Shannon 1948] C. E. Shannon, *Bell Syst. Techn. J.* **27**, 379–423, 623–656 (1948).

[Shannon 1951] Shannon, C. E., *Bell Syst. tech. J.* **30**, 50 (1951).

[Sharp & Li 1987] P. M. Sharp and W.-H. Li, Nucl. Acid Res. 15, 1281 (1987).

[Shepherd 1981] J. W. C. Shepherd, J. Mol. Evol. 17, 94 (1981).

[Shepherd et al. 1981] J. C. W. Shepherd and J. C. W., Proc. Natl. Acad. Sci. USA **78**, 1596 (1981).

[Shinnar 1961] R. Shinnar, J. Fluid Mech. **10**, 259 (1961).

[Silverman et al. 1986] Silverman, B.D. & Linsker, R. (1986). A measure of DNA periodicity. *J. theor. Biol.* **118,** 295–300.

[Snyder & Stormo 1993] Snyder, E. E. & Stormo, G. D. (1993) *Nucl. Acids Res.* **21**, 1107–1115.

[Snyder & Stormo 1995] E. E. Snyder and G. D. Stormo, *J. Mol. Biol.* **248**, 1–18 (1995).

[Solovyev et al. 1994] V. V. Solovyev, A. A. Salomov, and C. B. Lawrence, *Nucl. Acids Res.* **22**, 5156–5163 (1994).

[Staden 1982] Staden, R. & McLachlan, A. D. (1982) *Nucleic Acids Res.* **10**, 141–156.

[Staden 1984] Staden R., *Nucl. Acid Res.* **12** 505 (1984).

[Staden & McLachlan 1982] R. Staden and A. D. McLachlan, *Nucl. Acids Res.* **10**, 141–156 (1982).

[Stanley et al. 1994] Stanley, H. E., S. V. Buldyrev, A. L. Goldberger, Z. D. Goldberger, S. Havlin, R. N. Mantegna, S. M. Ossadnik, C.-K. Peng & M. Simons, *Physica A* **205** 214 (1994).

[Stariolo & Tsallis 1995] Stariolo D A and Tsallis C 1995, appeared in *Annual Reviews of Computational Physics* vol 1, ed Stauffer D (Singapore: World Scientific)

[Stauffer 1996] D. Stauffer, Computational Physics M. Schreiber K.H. Hoffmann, ed., Springer, Berlin, Heidelberg, 1996, p. 1.

[Steveninck et al. 1997] de Ryter van Steveninck, R., G. D. Lewen, S. P. Strong, R. Koberle, and W. Bialek, "Reproducibility and variability in neural spike trains", *Science* **275** (1997), 1805–1807.

[Strait & Dewey 1996] Strait, B. J. and T. G. Dewey, "The Shannon information entropy of protein sequences", *Biophys. J.* **71** (1996), 148–155.

[Sze 1998] Sze, S.-H., Roytberg, M.A., Gelfand, M.S., Mironov, A.A., Astakhova, T.V. & Pevzner, P.A. (1998). Algorithms and software for support of gene identification experiments. *Bioinformatics* **14,** 14–19.

[Thomas & Skolnick 1994] A. Thomas and M. H. Skolnick, *IMA J. Math. Appl. Med. Biol.* **11**, 149–160 (1994).

[Tiwari et al. 1997] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, *Comput. Appl. Biosci.* **13**, 263–270 (1997).

[Trifonov & Brendel 1986] E. N. Trifonov and V. Brendel, *Gnomic - A Dictionary of Genetic Codes* (Balaban Publ., Rehovot, 1986).

[Trifonov 1989] E. N. Trifonov, Bull. Math. Biol. 51, 417 (1989).

[Trifonov 1997] Trifonov, E.N. (1987). Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16S rRNA nucleotide sequences. *J. Mol. Biol.* **194,** 643–652.

[Trifonov & Sussman 1980] E. N. Trifonov and J. L. Sussman, Proc. Natl. Acad. Sci. USA **77** (1980) 3816.

[Tsallis 1988] Tsallis, C., "Possible Generalization of Boltzmann-Gibbs Statistics", *J. Stat. Phys.* **52** (1988), 479–487.

[Tsallis et al. 1995] Tsallis C, Levy S V F, Souza A M C and Maynard R 1995 *Phys. Rev. Lett.* **75** p 3589

[Tsallis Entropies] Tsallis, C., S. V. F. Levy, A. M. C. Souza, and R. Maynard "Statistical-mechanical Foundation of the Ubiquity of Lèvy Distributions in Nature", *Phys. Rev.*

Lett. **75** (1995), 3589–3593; Boghosian, B. M., "Thermodynamic description of the relaxation of two-dimensional turbulence using Tsallis statistics", *Phys. Rev.* E **53** (1996), 4755–4763; Tsallis, C. and D. A. Stariolo, "Generalized simulated annealing", *Physica D* **233** (1996), 395–406. Penna, T. J. P., "Traveling salesman problem and tsallis statistics", *Phys. Rev. E* **51** (1995), R1–R3; regularly updated bibliography on generalized thermostatistics can be found under http://tsallis.cat.cbpf.br/~biblio.htm.

[Turcotte 1996] D. L. Turcotte, J. Geophys. Res. **91**, 1921 (1986).

[Uberbacher & Mural 1991] Uberbacher, E. C. & R. Mural, *Proc. Natl. Acad. Sci. USA* **88** 11261 (1991).

[Ulanovsky & Trifonov 1983] L. E. Ulanovsky and E. N. Trifonov, Cell Biophysics **5** (1983) 281.

[Vologodskii 1992] A. Vologodskii, *Topology and Physics of Circular DNA*, (CRC Press, Boca Raton, Fl, 1992).

[Völz 1982] Völz, H., *Information I & II*, Akademie-Verlag, Berlin (1982, 1983).

[Völz 1990] Völz, H., *Computer und Kunst*, Urania-Verlag, Leipzig (1990).

[von Heijne 1987] G. von Heijne, *Sequence Analysis in Molecular Biology - Treasure Trove or Trivial Pursuit* (Academic Press, San Diego, 1987).

[Voss 1992] Voss, R. F., *Phys. Rev. Lett.* **68** 3805 (1992).

[Watson 1990] Watson, J. D., *Science* **248** 44 (1990).

[Watson & Crick 1953] Watson, J. D. & F. H. C. Crick, *Nature* **171** 737 (1953).

[Watson et al. 1992] J. D. Watson, M. Gilman, J. Witkowski, and H. Zoller, *Recombinant DNA* (W. H. Freeman, New York, 1992).

[Weiss & Herzel 1997] O. Weiss and H. Herzel, *Correlations in Protein Sequences and Property Codes*, J. Theor. Biol., submitted.

[White 1994] S. H. White, Annu. Rev. Biophys. Biomol. Struct. 23, 407 (1994).

[Wickmann 1990] Wickmann, D., *Bayes-Statistik*, BI Wissenschaftsverlag, Mannheim (1990).

[Wolkenstein 1983] Wolkenstein, M. W., *Biophysics*, Mir publishers, Moscow (1983).

[Wolpert & Wolf 1993] D. Wolpert and D. Wolf, Estimating functions of probability distributions from a finite set of samples, preprint Santa Fe Inst. TR-93-07-046.

[Wolpert & Wolf 1995] Wolpert D. H. and D. R. Wolf "Estimating functions of probability distributions from finite samples", *Phys. Rev. E* **52** (1995), 6841.

[Xu & Uberbacher 1997] Y. Xu and E. C. Uberbacher, *J. Comput. Biol.* **4**, 325–338 (1997).

[Xu et al. 1994] Xu, Y., Einstein, J.R., Mural, R.J., Saha, M & Uberbacher, E.C. (1994). An improved system for exon recognition and gene modeling in human DNA sequences. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, Menlo Park, CA.

[Yockey 1992] H. P. Yockey, *Information Theory and Molecular Biology* (Cambridge University Press, Cambridge, 1992).

[Zhang 1997] M. Q. Zhang, *Proc. Natl. Acad. Sci. USA* **94**, 565–568 (1997).

[Zhurkin 1981] V. B. Zhurkin, Nucl. Acids Res. **9** (1981) 1963.

[Comment 1] A *coding measure* is function f that maps a statistical pattern $\vec{x}$ to a real number $y \equiv f(\vec{x})$ such that the probability distribution functions of y are different in coding and noncoding DNA. Typically, $\vec{x}$ is high dimensional, and f depends on many empirical parameters. Typically, these parameters vary significantly from species to species. Hence, these parameters must be fitted by empirical analyses of species-specific data sets. The process of fitting the parameters is called *training* of the coding measure.

[Comment 2] The mutual information function is similar to, but different from, autocorrelation functions [Li 1989, Herzel & Grosse 1995, Herzel & Grosse 1997]. Its main advantage over correlation functions is that it does not require any mapping of symbols to numbers, which affects the analysis of symbolic sequences by correlation functions, because correlation functions are not invariant under changes of the map. Moreover, the mutual information function is capable of detecting any deviation from statistical independence, whereas—by definition—correlation functions measure only

linear dependences. Hence, we use the mutual information function in our analysis of DNA sequences.

[Comment 3] We use all eukaryotic DNA sequences from GenBank release 111 (D. A. Benson , M. S. Boguski, D. J. Lipman, J. Ostell, B. F. Ouellette, B. A. Rapp, and D. L. Wheeler. *Nucl. Acids Res.* **27**, 12–17 (1999), `ftp://ncbi.nlm.nih.gov/genbank/`).

[Comment 4] There are $4^3 = 64$ codons, 3 of which are stop codons, and 61 of which encode 20 amino acids. Hence, the genetic code is *degenerate*, i. e. there are (many) amino acids that are encoded by more than one codon. All codons that encode the same amino acid are called *synonymous codons*.

[Comment 5] Mathematically, $p_i^{(m)}$ can be defined in terms of Q_{XYZ} as follows: $p_i^{(1)} \equiv \sum_{Y,Z} Q_{n_i YZ}$, $p_i^{(2)} \equiv \sum_{X,Z} Q_{X n_i Z}$, and $p_i^{(3)} \equiv \sum_{X,Y} Q_{XY n_i}$.

[Comment 6] Since the genetic code is a non-overlapping triplet code, there are three frames in which a DNA sequence can be translated into an amino acid sequence. In the cell, only one of the three *reading frames* encodes the proper amino acid, but in our statistical analysis the choice of the reading frame is *arbitrary* in the sense that $P_{ij}(k)$ is *invariant* under shifts of the reading frame.

[Comment 7] We choose the length to be 54 bp in order to allow a comparison with the standard data set created in Ref. [Fickett 1992], which consists of sequences of length 54 bp.

[Comment 8] Here, true positives (true negatives) refer to correctly-predicted coding (noncoding) sequences, and positives (negatives) refer to all coding (noncoding) sequences. Hence, T_p (T_n) denotes the fraction of correctly predicted coding (noncoding) sequences over all coding (noncoding) sequences. Mathematically, T_p and T_n are defined by $T_p \equiv \int \theta(\rho_c(\overline{I}) - \rho_n(\overline{I}))d\overline{I}$ and $T_n \equiv \int \theta(\rho_n(\overline{I}) - \rho_c(\overline{I}))d\overline{I}$, where θ denotes the Heavyside function, i. e. $\theta(x) \equiv 1$ for $x \geq 0$ and $\theta(x) \equiv 0$ for $x < 0$.

[Comment 9] If $\rho_c(\overline{I})$ and $\rho_n(\overline{I})$ were identical, $\mathcal{O}(\overline{I})$ would be equal to 1. If $\mathcal{P}_c(\overline{I})$ and $\rho_n(\overline{I})$ were completely disjoint (non-overlapping), $\mathcal{O}(\overline{I})$ would be equal to 2.

[Comment 10] For the probabilities $p_i^{(m)}$ we choose the total number of nucleotides n_i in position m of the biological reading frame divided by the total number of nucleotides

from exactly the same set of coding human sequences to which the model sequences are compared.

[Comment 11] By *correlations or inhomogeneities* we mean that the probability distributions $p_i^{(m)}$ are not constant, but vary along the DNA sequence from gene to gene and also within a gene. These variations of the probability distributions $p_i^{(m)}$ seem to be a typical feature of coding DNA of any living organism.

[Comment 12] By means of numerical simulation, we find that to a good approximation $s_{\max}(x) = \text{Prob}\{\text{JS}_{\max} \leq x\} = [\text{F}_9(\beta \cdot 2N\ln 2 \cdot x)]^\alpha$, where $\alpha = 2.45\ \ln(N) - 9.87$, $\beta = 0.84$, N is the size of the sequence or subsequence to be split and and $\text{F}_\nu(x)$ is the χ^2-distribution function with ν degrees of freedom. Here $\nu = 9$ because there are 3 constraints: $\sum_\ell f_{\ell,j} = 1/3$ for $j = 0, 1, 2$, i.e. the number of nucleotides in each phase is 1/3 of the total.

[Comment 13] A discriminant function is any function that can be computed on a region of a DNA sequence which reaches different values for coding and non coding DNA. In particular we use here $\sum_{\ell,j} |f_\ell - 3f_{\ell,j}|$, where $f_{\ell,j}$ is the relative number of nucleotides of type ℓ with phase j, and f_ℓ is the relative number of nucleotides of type ℓ in any of the three phases. See R. Staden Nucl. Acid Res. **21**, 551 (1984).

Curriculum Vitae

Address:

Ivo Grosse
Boston University
Center for Polymer Studies and Department of Physics
Boston, MA 02215

Phone: 617-353-8936
Fax: 617-353-8936
E-mail: ivo@bu.edu

Education:

- **June 1995 - present**, Boston University, Department of Physics, Boston, Massachusetts, USA. Ph. D. student. Advisor: H. Eugene Stanley. Courses taken in addition to the graduate-school requirements: *Bioinformatics I + II, Molecular Biology I + II, Biophysics, Information Theory, Advanced Parallel Computing I + II, Statistical Mechanics I + II.*

- **November 1994 - June 1995**, Humboldt University, Department of Physics, Berlin, Germany. Advisors: Werner Ebeling and Hanspeter Herzel. Diplom thesis *Statistical Analysis of Biosequences.* Graduated with honors.

Experience:

- **January 1997 - present**, Boston University, Center for Polymer Studies and Department of Physics, Boston, Massachusetts, USA. Research assistant with H. Eugene Stanley and Sergey V. Buldyrev on *Applications of Statistical Physics and Information Theory to the Analysis of DNA Sequences.*

  Investigate statistical differences between coding and noncoding DNA and search for those differences which are species independent with the goal to develop a gene finding program that does not require prior training; in collaboration with Hanspeter Herzel, Humboldt University, Berlin and James W. Fickett, SmithKline Beecham, Philadelphia.

  Combine concepts of information theory with data-mining and visualization techniques for gene identification and protein structure prediction; in collaboration with Kenneth A. Marx and Georges G. Grinstein, University of Massachusetts at Lowell.

- **June 1995 - December 1996**, Boston University, Department of Physics, Boston, Massachusetts, USA. Teaching assistant for courses in *Introductory Physics, Electromagnetism, Quantum Mechanics.* Guest lectures in *Bioinformatics, Biophysics.*

- **September 1992 - June 1995**, Humboldt University, Department of Physics, Berlin, Germany. Research assistant with Werner Ebeling and Hanspeter Herzel.

  Investigated whether information theoretical quantities are different in coding and noncoding DNA with the goal of developing fast and efficient computer algorithms to identify genes in unannotated DNA.

  Studied finite size effects with the goal to derive analytic length correction formulae that allow an unbiased estimation of information theoretical quantities, such as the *Shannon entropy* or the *mutual information*, from DNA or protein sequences.

- **July 1994 - October 1994**, Oak Ridge National Laboratory, Bioinformatics Group, Oak Ridge, Tennessee, USA. Research assistant with Richard J. Mural and Edward C. Uberbacher.

  Participated in the development of the gene recognition module GRAIL.

  Investigated the species dependence of the performance of GRAIL.

  Analyzed pair correlations between nucleotides in acceptor and donor sites with the goal to improve the recognition of splice junctions by GRAIL.

- **January 1994 - June 1994**, City University of New York, Department of Biology, New York, New York, USA. Exchange Student. Graduate studies in molecular biology and genetics.

- **September 1993**, Studienstiftung des Deutschen Volkes, Summer School, St. Johann, Italy. Course on *Introduction to Neural Networks and Neural Computation.*

- **August 1993**, Deutsche Physikalische Gesellschaft, Summer School, Rostock, Germany. Course on *Many-Particle Theory and Quantum Statistics.*

- **September 1992**, Studienstiftung des Deutschen Volkes, Summer School, Völs, Italy. Course on *Physiological Adaptation and Design of Higher Organisms.*
- **September 1991 - August 1992**, Humboldt University, Department of Mathematics, Berlin, Germany. Teaching assistant for courses in *Mathematical Physics.*
- **March 1991 - August 1992**, Humboldt University, Departments of Physics and Electronics, Berlin, Germany. Programmer. Worked in a group operating computer networks in the Physics and Electronics Departments at Humboldt University.
- **August 1991**, University of Reading, Department of English, Reading, Great Britain. English Language Course. Preparation for studies at British universities.
- **November 1988 - July 1989**, National Service.
- **September 1988 - October 1988**, Humboldt University, Department of Electronics, Berlin, Germany. Programmer. Developed data base applications for departmental management.

Awards:

- **January 1997 - present**, NIH Graduate Traineeship Award.
- **June 1995 - December 1996**, Boston University Teaching Fellowship.
- **April 1992 - June 1995**, Studienstiftung des Deutschen Volkes Scholarship.
- **July 1988**, Lessingmedaille in Gold.
- **July 1987**, Heinrich-Hertz-Preis. Created an educational computer program on mathematical integration, which won the first prize in a Berlin-wide competition on scientific topics.

Publications:

1. H. Herzel and I. Große. Measuring Correlations in Symbolic Sequences. *Physica A*, 216, 518–542 (1995).

2. H. Herzel, W. Ebeling, I. Große, and A. O. Schmitt. Statistical Analysis of DNA Sequences. In *Bioinformatics: From Nucleic Acids and Proteins to Cell Metabolism*, VCH-Verlag, Weinheim, 1995, pp. 29–43.

3. I. Große. Estimating Entropies from Finite Samples. In *Dynamik, Evolution, Strukturen: Nichtlineare Dynamik und Statistik komplexer Strukturen*, Köster-Verlag, Berlin, 1996, pp. 181–190.

4. D. Beule and I. Große. Wie zufällig sind Zufallszahlen? In *Dynamik, Evolution, Strukturen: Nichtlineare Dynamik und Statistik komplexer Strukturen*, Köster-Verlag, Berlin, 1996, pp. 161–170.

5. H. Herzel and I. Große. Correlations in DNA Sequences – the Role of Protein Coding Segments. *Phys. Rev. E*, 55, 800–810 (1997).

6. P. Hoffman, G. G. Grinstein, K. A. Marx, I. Große, and H. E. Stanley. DNA Visual and Analytic Data Mining. In *IEEE Visualization '97 Proceedings*, ACM Press, New York, 1997, pp. 437–441.

7. H. Herzel, E. N. Trifonov, O. Weiss, and I. Große. Interpreting Correlations in Biosequences. *Physica A*, 248, 449–459 (1998).

8. D. Holste, I. Große, and H. Herzel. Bayes Estimators of Generalized Entropies. *J. Phys. A*, 31, 2551–2566 (1998).

9. I. Große, H. Herzel, S. V. Buldyrev, and H. E. Stanley. A Species-Independent Measure for Distinguishing Coding and Noncoding DNA. submitted.

10. I. Große, S. V. Buldyrev, H. Herzel, D. Holste, and H. E. Stanley. Tsallis Entropies of Coding and Noncoding DNA Sequences. in preparation.

Memberships:

- International Society for Computational Biology
- American Physical Society

- Deutsche Physikalische Gesellschaft

Special skills:

- Programming languages: C, Fortran, Pascal, Basic.
- Foreign languages: Fluent in written and oral English and German, working knowledge of Russian.