Lehrstuhl für Nachrichtentechnik

Frame Synchronization Processes in Gene Expression

Johanna Weindl

Vollständiger Abdruck der von der Fakultät für Elektrotechnik und Informationstechnik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor–Ingenieurs

genehmigten Dissertation.

Vorsitzender:	Univ.–Prof. Dr. rer. nat. habil. B. Wolf
Prüfer der Dissertation:	1. Univ.–Prof. Dr.–Ing., Dr.–Ing. E. h. J. Hagenauer (i. R.)
	2. Univ. Prof. Dr.–Ing. K. Diepold

Die Dissertation wurde am 10.06.2008 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 28.11.2008 angenommen.

Preface

This thesis was written between January 2006 and June 2008 during my time at the Institute for Communications Engineering (LNT) of Technische Universität München. It would not have been possible without the following persons:

First, I would like to thank my supervisor Professor Joachim Hagenauer for his constant support and guidance. The fruitful discussions with him during our frequent *ComInGen*-meetings have doubtlessly shaped this work. Moreover, I also thank Professor Klaus Diepold for acting as co-supervisor despite the interdisciplinary nature of this work.

Some other people contributed significantly to this thesis: Jakob Müller was a constant help regarding biological questions, while Zaher Dawy and my colleagues Bernhard and Janis spent many hours proofreading the thesis and discussing its technical aspects with me. Furthermore, Torsten and Steffi were important and precise proofreaders regarding the linguistics. I strongly appreciate your opinion!

I was very lucky to have some excellent students working under my supervision. Nora Tax, Nabeel Sulieman, Tobias Rehrl and Friedrich Kischkel were courageous enough to face the risk of such an interdisciplinary topic and all have a major share in this thesis.

Last but most importantly, my friends and my family have made this thesis possible by accompanying me through a time of stress and pressure (towards the end), high demands and doubts (most of the time), overload and frustration (fortunately only from time to time). These are above all my parents Jutta and Hugo, my brother Torsten, Wolfgang as well as my close friends Sibylle, Philipp, Robert and Dominique. The importance of my Habibi Bernhard can hardly be verbalized and will therefore be expressed in personal moments instead of in this preface. Without your support, love and understanding, this thesis would not have become what it is now!

München, June 2008 Johanna Weindl

Contents

Zusammenfassung

Diese Arbeit behandelt die Modellierung der Genexpression (Proteinsynthese) durch die Rahmensynchronisation, einem Verfahren der digitalen Datenübertragung. Hierbei detektiert der Empfänger den Beginn einer Nachricht anhand kurzer Signalisierungssequenzen, sogenannter Synchronisationswörter. Analog dazu verwendet die Natur feste Sequenzmotive, um den Beginn von fundamentalen DNA-Regionen zu markieren. Diese Analogie erlaubt es, Methoden der Rahmensynchronisation anzupassen und mittels Simulationen auf verfügbare Genome anzuwenden. Die beiden Hauptschritte der Genexpression, Transkription und Translation, werden als Rahmensynchronisationsprozess modelliert. Zur weiterführenden Untersuchung der DNA-Sequenzen werden klassische informationstheoretische Maße verwendet. Die Ergebnisse dieser Arbeit belegen, dass die Synchronisationswörter der Genexpression und ihre Umgebung im nachrichtentechnischen Sinne nahezu optimal gewählt wurden.

Abstract

This thesis deals with the modeling of gene expression (protein synthesis) using frame synchronization, a method applied in digital data transmission. There, the receiver detects the beginning of a message based on short signaling sequences, so-called synchronization words. Analogously, nature makes use of fixed sequence motifs to mark the beginning of fundamental DNA regions. This analogy allows to adapt techniques from frame synchronization and to apply these to available genomes using simulations. The two main steps of gene expression, transcription and translation, are modeled as a frame synchronization process. For a continuative analysis of the DNA sequences, classical information theoretic measures are applied. The results of this thesis indicate that the synchronization words of gene expression and their surrounding have been chosen nearly optimally in the communications engineering sense.

1

Introduction

In 1940, Claude E. Shannon submitted his doctoral dissertation entitled "An Algebra for Theoretical Genetics" [Sha40] in which he mathematically investigated Mendelian heredity. His results have never appeared in a publication besides his Ph. D. thesis for three main reasons: First, it was a time of personal and professional changes in Shannon's life. Second, genetics was at that time in a crisis since it had been revealed that the Nazis used eugenics to justify their genocide. Last and most importantly, many geneticists lacked appreciation of mathematics without experimental evidence, and mathematicians were not interested in problems related to population genetics. Vannevar Bush, Shannon's thesis supervisor, wrote in a note to Barbara Barks at Cold Spring Harbor that "*few scientists are ever able to apply creatively a new and unconventional method furnished by some one else - at least of their own generation*" and further in a correspondence with Shannon: "*I doubt very much whether your publication will result in further work by others using your method, for there are very few individuals in this general field who would be likely to do so*" (quotations taken from [CLM+01]).

During the following 50 years, information and communication theory on the one hand and molecular biology and genetics on the other hand did not enter substantial cooperations. While the biological community focussed on the structure, transmission and transformation of genetic information as well as on the sequencing of complete genomes, information theory mainly focussed on the optimization of the reliable transmission of digital data. With few exceptions (e.g. Solomon W. Golomb, see [Hay98]), it is only by today that communications engineers and information theorists are beginning to foster the cooperation with biologists. To name but few, Gérard Battail derived speculative indication for the existence of error-correcting codes in genomes [Bat04,Bat06], Elebeoba E. May presented coding theory based models of protein synthesis [MVBR04, MVB06], and Joachim Hagenauer applied mutual information to infer the relationship between positions in the

DNA and genetic diseases like Schizophrenia and Parkinson [HDG+04, DGH+06]. Accordingly, several biologists are beginning to appreciate methods from information theory for the analysis of complex genetic problems (e.g. [Moo08, For81, Yoc92, Sch97, Sch96]).

In the past years, an increasing number of completely sequenced genomes have become available in public databases. Recalling that the typical genome length of higher organisms ranges between 1 and 100 billion bases illustrates impressively the need for methods to computationally store, handle and analyze the huge amount of data – a framework that information and communication theory can provide. Moreover, the processes occurring inside the cell during protein synthesis (e.g. the conservation, readout and transformation of genetic information) bear substantial analogies to processes in digital data transmission. Therefore, modeling cell processes through communication and information theory ideally serves two tasks: give yet unknown insights into the details of biological processes and inspire the design of technical systems based on biological systems that have been optimized over millions of years.

This thesis shall attempt to shed light upon the processes involved in protein synthesis (gene expression) taking place in the two steps transcription and translation. The models are derived from frame synchronization in communication systems, which refers to the detection of a message in a continuous data stream. The common procedure to achieve frame synchronization in digital data transmission is to insert a known pattern into the data stream with a fixed distance to the message. At the receiver side, a correlation measure is applied to detect the pattern. Analogously, short DNA motifs are found at positions where information has to be read out during gene expression. This analogy facilitates and suggests to apply measures and algorithms from frame synchronization in technical systems for the analysis of biological synchronization processes and motifs. As mentioned above for the case of Shannon's Ph. D. thesis, the successful establishment of such interdisciplinary approaches is rendered difficult by the limited number of cooperating research groups and potential reviewers. Therefore, the ultimate objective of this thesis work is to be recognized by and to be published in both communities – biology and communications engineering.

The structure of this thesis is as follows:

Chapter 2 introduces the reader to the basics of frame synchronization in digital transmission systems. It starts with the definition of detection methods and sync word location rules. Thereafter, the synchronization performance is derived based on possible error sources. Due to its importance for later chapters, a strong focus is subsequently laid on the design of sync words. Based on these design criteria, sync word families have emerged for various channel scenarios.

Chapter 3 provides the biological background needed to understand the communication theoretic models in later chapters. It is written such that the reader is not required to have any prior knowledge. Important literature references are supplied that provide a broader introduction to the topic than possible in this thesis. The focus lies on gene expression, the process of protein synthesis vital to all organisms. It is covered separately for bacteria (prokaryotes) and higher organisms (eukaryotes) as well as for transcription

and translation – the two main steps of gene expression.

Chapter 4 follows with the investigation of biological synchronization words in bacteria. While sync words in technical systems are carefully designed to minimize the probability of false synchronizations, the biological community has spent little attention to the analysis of biological signalization sequences. First, the bacterial promoter – the sync word of transcription – is investigated with respect to its synchronization properties. Second, the bacterial Shine-Dalgarno sequence – the sync word of translation – is analyzed using information theoretic measures.

The following four chapters (Chapter 5 to Chapter 8) detail the communication theoretic modeling of transcription and translation in both bacteria (prokaryotes) and higher organisms (eukaryotes).

Chapter 5 marks the beginning with the modeling of transcription in the bacterium *Escherichia coli*. A synchronization algorithm based on a weight matrix is derived that models the detection of the promoter sequence whose synchronization properties were investigated in Chapter 4. The results are interpreted with respect to their impact on reliable and fast detection of the transcription start site. The interpretations are verified using biophysical theory and computer simulations on real promoter sequences.

Chapter 6 deals with eukaryotic transcription. Since the process is far more complex than in bacteria, the focus lies on a general sequence analysis of promoter sequences instead of modeling single interactions. For this purpose, two information theoretic measures are adapted for the application to DNA sequences: mutual information and the related Kullback-Leibler divergence. In order to obtain meaningful results, the datasets of promoter sequences are thereafter subdivided according to promoter characteristics.

Chapter 7 follows with the synchronization modeling of translation in bacteria. Different codebooks are designed to model the detection of the Shine-Dalgarno sequence – the sync word of prokaryotic translation. The codebook models are applied to sequence data of *Escherichia coli* and evaluated based on the achieved detection strength. The best-performing model is subsequently adapted to include a synchronization metric based on binding energies. Thereafter, the effect of mutations is analyzed using codebook changes.

Chapter 8 completes the work on the modeling of gene expression by investigating eukaryotic translation. The emphasis is placed on information theoretic measures for sequence analysis. Kullback-Leibler divergence and mutual information are applied to detect functional positions and dependencies in eukaryotic sequence datasets. Additionally, a codebook model is derived for the detection of the Kozak sequence – the sync word of eukaryotic translation.

Chapter 9 concludes this thesis. The main achievements are detailed along with future research directions.

Parts of this work have been published in the technical conference publication [WH07b], in the technical journal publications [DHW+07] and [HGD+07] as well as in the biological journal publications [WHD+07] and [DMWM09].

2

Frame Synchronization in Continuous Transmission

Frame synchronization is an essential problem of data transmission in all digital communication systems. It refers to localizing the beginning of a message in received data, i.e. to *"the correct association at the receiver of the received symbols to blocks such as words, bytes or data-frames"* [Rob95]. The most common procedure to achieve frame synchronization is to insert a fixed pattern into the data stream which is known to both the transmitter and the receiver. This so-called marker concept was introduced by R. H. Barker in 1953 [Bar53] and further investigated by J. L. Massey in 1972 [Mas72]. Since then, researchers have addressed the design and reliable detection of these sync words for both packet transmission and continuous transmission.

This chapter provides the basics of frame synchronization in continuous transmission. In Section 2.1, the objective of frame synchronization in digital data transmission systems is introduced along with necessary notations. Moreover, two detection methods are introduced, namely threshold detection and maximum selection. Section 2.2 details the optimum synchronization rule for periodically as well as aperiodically inserted sync words as first proposed by J. L. Massey in 1972. Section 2.3 follows with a derivation of the synchronization performance of threshold detection. It is presented separately for periodically and aperiodically inserted sync words as well as two different sequence models (symbols independently and uniformly distributed vs. Markov chain). Subsequently, Section 2.4 presents the main aspects of sync word design. Several measures are detailed to rate the suitability of a sequence as a sync word, e.g. characteristics of its autocorrelation function and the Hamming distance between its suffices and prefixes. Based on these quality measures, well-known sync word families are presented in Section 2.5.

2.1 Problem definition

To achieve frame synchronization, the receiver evaluates the incoming discrete data stream $\mathbf{d} = \{d_1, \ldots, d_{N_d}\}$ of length N_d, where the d_k are elements of the alphabet $\mathcal{A}$. At each position $\mu \in [1; N_d - L + 1]$, it compares the subsequence $\{d_\mu, \ldots, d_{\mu+L-1}\}$ of the data stream with the sync word $\mathbf{s} = \{s_1, \ldots, s_L\}$ of length L to determine the position μ_s at which the possibly altered sync word is most likely located. In the periodic case (synchronous communication), the data is divided into frames of constant length N_f, where each frame contains one sync word at the same position. In the aperiodic case (asynchronous communication), the sync words start at random positions along the data stream. In both cases, detection of the sync word is based on a likelihood function $L(\mu)$, which the receiver evaluates at each position μ to decide about the location of a sync word (see Figure 2.1). $L(\mu)$ is a measure for the probability of the currently considered – possibly erroneous – sequence of symbols to be a sync word. In Section 2.2, the optimal definition of $L(\mu)$ with respect to minimizing synchronization errors is derived.

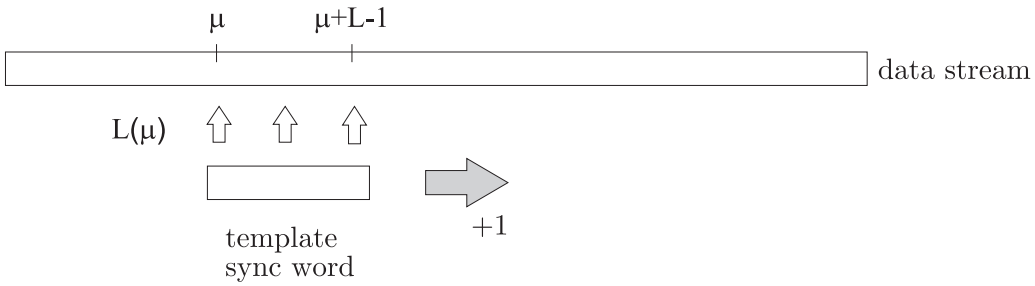


Figure 2.1: For detection of the sync word, the receiver evaluates a likelihood function $L(\mu)$ at each step μ of the incoming data stream.

2.1.1 Threshold detection

Threshold detection is based on defining a threshold for the likelihood function $L(\mu)$, above which the receiver assumes the currently considered sequence $\{d_\mu, \ldots, d_{\mu+L-1}\}$ to be a sync word (see Figure 2.2, left):

$$\mu_{s,l} = \{\mu_l | L(\mu_l) > L_{\text{th}}\}, \tag{2.1}$$

where μ_l refers to the l^{th} position μ_s of the data stream that fulfills the condition. A sync word is assumed to be located at all positions $\mu_{s,l}$, $l \in [1; N_d - L + 1]$. A crucial point for successful detection lies in the optimization of the threshold L_{th} with respect to the false positive and false negative rate. Threshold detection is applied for detecting aperiodically inserted sync words (asynchronous transmission).

2.1.2 Maximum selection

In contrast to threshold detection, maximum selection only considers one frame length N_f of the data stream at a time. That position μ of the l^{th} data frame $\{d_{(l-1)\cdot N_f+1}, \ldots, d_{l\cdot N_f}\}$ with the highest value of $L(\mu)$ is assumed to be the location of the l^{th} sync word (see Figure 2.2, right):

$$\mu_{s,l} = \underset{\mu\in[(l-1)\cdot N_f+1; l\cdot N_f]}{\operatorname{argmax}} L(\mu). \tag{2.2}$$

Of course, this method is only applicable for periodically embedded sync words (synchronous transmission), i.e. if the receiver knows a priori that each frame of length N_f contains exactly one sync word.

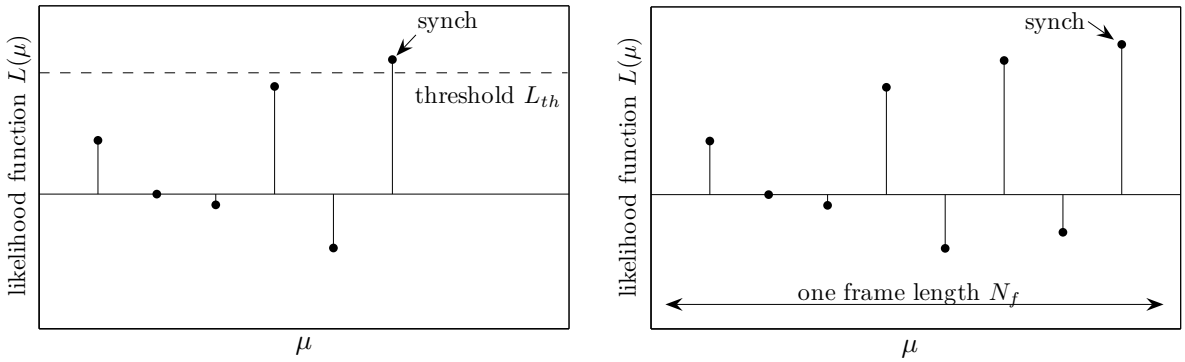


Figure 2.2: Illustration of sync word detection methods. Left: threshold detection. Right: maximum selection.

2.2 Optimum sync word location rule

An intuitive definition of $L(\mu)$ is the cross-correlation function between the incoming data stream $\mathbf{d}$ and the known sync word $\mathbf{s}$, i.e.

$$L(\mu) = \sum_{k=1}^{L} s_k \cdot d_{k+\mu}, \tag{2.3}$$

where L denotes the length of the sync word. This so-called correlation rule was applied in synchronization systems until J. J. Stiffler first recognized that the data surrounding the sync word should be taken into account for the sync word localization [Sti71]. However, this was assumed to be computationally too extensive until J. L. Massey derived an optimal decision rule for periodically inserted sync words transmitted over the additive white Gaussian noise (AWGN) channel with binary phase shift keying (BPSK) [Mas72].

2.2.1 Periodically inserted sync words

In the case of periodically inserted sync words, each frame length N_f contains exactly one sync word (SW) (see Figure 2.3).

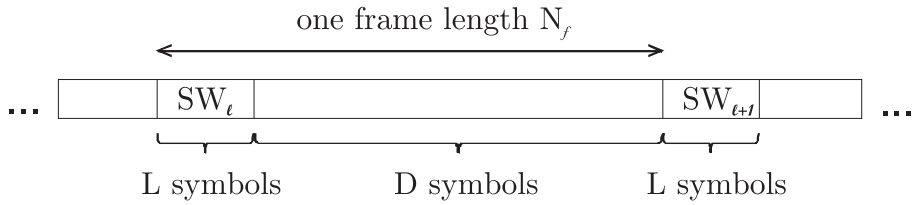


Figure 2.3: Frame structure for periodically inserted sync words (SW).

J. L. Massey found out that the optimum location rule $L_{\text{opt}}(\mu)$ for this case with respect to minimizing the number of erroneously synchronized frames is achieved by adding a correction term to the correlation rule (Eq. (2.3)):

$$L_{\text{opt}}(\mu) = \sum_{k=1}^{L} s_k \cdot d_{k+\mu} - \frac{N_0}{2\sqrt{E_b}} \sum_{k=1}^{L} \ln\left[\cosh\left(\frac{2\sqrt{E_b} d_{k+\mu}}{N_0}\right)\right], \tag{2.4}$$

where N_0 is the one-sided noise spectral density and E_b the bit energy. The d_k are defined to have value either $+\sqrt{E_b}$ or $-\sqrt{E_b}$, while the s_k are either $+1$ or -1. Therefore, the metric $L_{\text{opt}}(\mu)$ has the same dimension as $\sqrt{E_b}$. J. L. Massey moreover presented easily computable approximations $L_{\text{low}}(\mu)$ and $L_{\text{high}}(\mu)$ of the correction term for low SNRs ($E_b/N_0 \ll 1$) and high SNRs ($E_b/N_0 \gg 1$):

$$L_{\text{low}}(\mu) = \sum_{k=1}^{L} s_k \cdot d_{k+\mu} - \frac{\sqrt{E_b}}{N_0} \sum_{k=1}^{L} d_{k+\mu}^2, \tag{2.5}$$

$$L_{\text{high}}(\mu) = \sum_{k=1}^{L} s_k \cdot d_{k+\mu} - \sum_{k=1}^{L} |d_{k+\mu}|. \tag{2.6}$$

These approximations are based on the replacement of the cosh-function in Eq. (2.4) by the first term of its Maclaurin series expansion in case of low SNRs and by an exponential function in case of high SNRs [Mas72]:

$$\text{SNR} \ll 1: \ \ln\left[\cosh\left(\mathrm{x}\right)\right] \approx \frac{1}{2} \cdot \mathrm{x}^2, \quad \text{SNR} \gg 1: \ \cosh\left(\mathrm{x}\right) \approx \frac{1}{2} \cdot \mathrm{e}^{|\mathrm{x}|}. \tag{2.7}$$

Using Monte Carlo simulations, J. L. Massey verified that the optimum location rule provides a 3-dB improvement over the correlation rule in the case of SNRs near one ($E_b/N_0 \approx 1$) for wide ranges of the sync word length and the frame length [Mas72].

2.2.2 Aperiodically inserted sync words

In the case of aperiodically inserted sync words, two successive sync words SW_l and SW_{l+1} are spaced at distance T_l (see Figure 2.4).

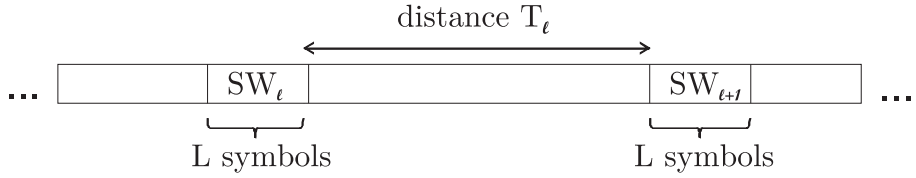


Figure 2.4: Data stream structure for aperiodically inserted sync words (SW).

The optimal decision rule introduced by J. L. Massey was later extended to the case of aperiodically inserted sync words transmitted over the AWGN channel. A. Kopansky and M. Bystrom [KB04] derived for BPSK modulation

$$L'_{\mathrm{opt}}(\mu) = \sum_{k=1}^{L} s_k \cdot d_{k+\mu} - \frac{N_0}{2\sqrt{E_b}} \sum_{k=1}^{L} \ln\left[\cosh\left(\frac{2\sqrt{E_b} d_{k+\mu}}{N_0}\right)\right] + \frac{N_0}{2\sqrt{E_b}} \ln\left[\Pr\{\mu = \mu_{s,l}\}\right], \tag{2.8}$$

where $\Pr(\mu = \mu_{s,l})$ is the probability of the l^{th} synchronization pattern to occur at position μ. It can be seen that $L'_{\mathrm{opt}}(\mu)$ adds a correction term to $L_{\mathrm{opt}}(\mu)$ taking into account the probabilities of synchronization patterns starting in a particular position. For the periodic case (see Section 2.2.1), the probability $\Pr(\mu = \mu_{s,l})$ becomes one, and thus the third term becomes zero in accordance with Eq. (2.4). To evaluate Eq. (2.8), knowledge about the distribution of sync words in the received data is required. An approximation for high SNRs is given by

$$L'_{\mathrm{high}}(\mu) = \sum_{k=1}^{L} s_k \cdot d_{k+\mu} - \sum_{k=1}^{L} |d_{k+\mu}| - \frac{N_0}{2\sqrt{E_b}} L \ln\frac{1}{2} + \frac{N_0}{2\sqrt{E_b}} \ln\left[\Pr(\mu = \mu_{s,l})\right]. \tag{2.9}$$

The third term results from the factor $\frac{1}{2}$ in the approximation of the cosh-function by an exponential function (see Eq. (2.7)) and was left out by Massey in Eq. (2.6) since it is independent of the data stream and, thus, does not add value to the metric. This approximation was shown to result in no performance loss for SNRs of above three ($E_b/N_0 > 3$), which is consistent with the findings for the periodical insertion of sync words considered in Section 2.2.1 [KB04].

2.3 Synchronization performance and error sources

In this section, the synchronization performance of threshold detection is derived. Since the case of maximum selection has no relevance for later application to biological processes,

its synchronization performance is not derived here but can be found in [Nie73b, Rob95]. In general, three error sources have to be taken into account to evaluate the performance of a frame synchronizer:

1. An overlap of a part of the sync word with neighboring symbols yields a valid sync word (shifted detection, see Figure 2.5, (A)).
2. The sync pattern appears in the random data in the surrounding of the correct sync word (false detection, see Figure 2.5, (B)).
3. The sync word is not recognized by the receiver because it was too strongly altered by transmission errors (missed detection).

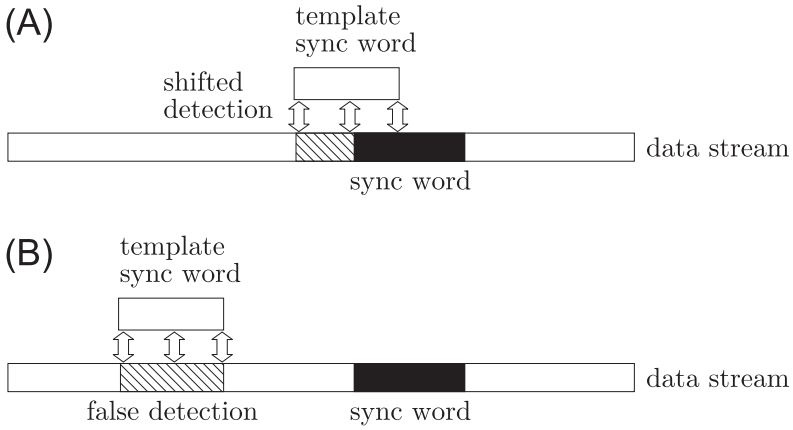


Figure 2.5: Illustration of error sources during frame synchronization.

In the following, the probabilities of these three error sources is presented. In Section 2.3.1, the symbols of the data stream are assumed to be independently and uniformly distributed on an arbitrary discrete alphabet $\mathcal{A}$. In Section 2.3.2, the data stream is assumed to exhibit statistical dependencies according to a Markov model Mm.

2.3.1 Sequence model: symbols independently and uniformly distributed

The symbols are now assumed to be independently and uniformly distributed (i. u. d.) on the alphabet $\mathcal{A}$. They are transmitted over a symmetric channel, i.e. each symbol is mistaken for another (arbitrary) one with probability P_e. More information on the following derivations can be found in [Sch80].

ad 1) A false detection on random data takes place if a correct or slightly altered sync pattern occurs by chance in the surrounding of the sync word. The probability P_{FD} of a false detection can thus be written as

$$P_{FD} = \left(\frac{1}{|\mathcal{A}|}\right)^{L} \cdot \sum_{x=0}^{h} \binom{L}{x}, \tag{2.10}$$

where $|\mathcal{A}|$ denotes the cardinality of the alphabet $\mathcal{A}$, and h is the error tolerance of the synchronization algorithm. The latter refers to the maximum number of acceptable changes in the sync word that still leads to a correct detection. Depending on the application, different thresholds h are practicable.

ad 2) In general, the probability P_{SD} for a shifted detection ($L - v$ symbols too early) of the sync word at position μ is [Sch80]

$$\begin{aligned} P_{SD} = \sum_{x=0}^{\min(v,h)} & \Pr\{d_H[\{d_\mu \dots d_{\mu+v-1}\}; \{s_1 \dots s_v\}] = x\} \cdot \\ & \Pr\{d_H[\{d_{\mu+v} \dots d_{\mu+L-1}\}; \{s_{v+1} \dots s_L\}] \leq h - x\}, \end{aligned} \tag{2.11}$$

where $\{d_\mu \dots d_{\mu+L-1}\}$ denotes the part of $\mathbf{d}$ that is at position μ evaluated with respect to its resemblance to the sync word. d_H denotes the Hamming distance between two sequences of the same length, i.e. $d_H(x, y) = |\{k|x_k \neq y_k\}|$. The first term of Eq. (2.11) defines the probability that the v received marker symbols contribute x units to the Hamming distance and the second term defines the probability that the surrounding data symbols contribute at most the remaining $h - x$ units. Let h_v denote the Hamming distance d_H between the first v and the last v symbols of the sync word:

$$h_v = d_H[\{s_1 \dots s_v\}; \{s_{L-v+1} \dots s_L\}]. \tag{2.12}$$

Then, Eq. (2.11) can be evaluated as [Sch80]

$$\begin{aligned} P_{SD}(h, v, h_v) = \sum_{x=0}^{\min(v,h)} & \left[\sum_{y=\max(0,x-h_v)}^{\min(x,v-h_v)} \binom{v-h_v}{y} \binom{h_v}{x-y} (1-P_e)^v \left(\frac{P_e}{1-P_e}\right)^{h_v-x+2y}\right] \cdot \\ & \left[\left(\frac{1}{|\mathcal{A}|}\right)^{L-v} \sum_{z=0}^{\min(h-x,L-v)} \binom{L-v}{z}\right], \end{aligned} \tag{2.13}$$

where the first term in brackets is the probability that y transmission errors occurred in the $v - h_v$ marker symbols which otherwise would have matched. The second term in brackets refers to the probability of z mismatches occurring in the remaining $L - v$ data symbols.

ad 3) The probability P_{CD} for a correct detection of the sync word at a fixed position is [Sch80]

$$P_{CD} = \sum_{x=0}^{h} \binom{L}{x} (1-P_e)^{L-x} P_e^x, \tag{2.14}$$

Thus, the probability P_{MD} of a missed detection due to transmission errors in the sync word is given by

$$P_{MD} = 1 - P_{CD} = 1 - \sum_{x=0}^{h} \binom{L}{x} (1-P_e)^{L-x} P_e^x. \tag{2.15}$$

2.3.2 Sequence model: Markov chain

Since the assumption of i. u. d. symbols does not apply to all data streams, the symbols are now assumed to be statistically dependent following a Markov chain of order m. In this case, the probability $P_m(\mathbf{r})$ of a sequence $\mathbf{r} = \{r_1, \ldots, r_L\}$ being generated by the Markov chain is

$$P_m(\mathbf{r}) = \Pr\{\mathrm{r_1}\} \cdot \Pr\{\mathrm{r_2}|\mathrm{r_1}\} \cdot \ldots \cdot \Pr\{\mathrm{r_{m+1}}|\{\mathrm{r_1}, \ldots, \mathrm{r_m}\}\} \cdot \ldots \cdot \Pr\{\mathrm{r_L}|\{\mathrm{r_{L-m}}, \ldots, \mathrm{r_{L-1}}\}\}. \tag{2.16}$$

As in Section 2.3.1, the symbols are transmitted over a symmetric channel, i.e. each symbol is mistaken for another (arbitrary) one with probability P_e. In the following, the probabilities of the three error sources introduced in the beginning of Section 2.3 are derived for the case of a Markov chain.

ad 1) If taking the error tolerance h of the synchronization algorithm into account, there exist $u = \sum_{t=0}^{h} \binom{L}{t}$ different sequences that yield a false detection. Thus, the probability P_{FD} for a false detection on random data is given by

$$P_{FD} = \sum_{t=1}^{u} P_m(\mathbf{r}_t), \tag{2.17}$$

where $P_m(\mathbf{r}_t)$ refers to the probability that the t^{th} possible sequence $\mathbf{r}_t$ is generated by the Markov chain (according to Eq. (2.16)).

ad 2) The probability P_{SD} for a shifted synchronization at position μ by $L-v$ positions is given in Eq. (2.13) for the case of an i. u. d. data stream. In the case of statistical dependencies according to Mm, the first term stays the same since it only deals with y transmission errors occurring in the v received marker symbols. The second term refers to the probability that the number of mismatches in the remaining $L-v$ symbols is z. Since there exist $u_z = \sum_{t=0}^{z} \binom{L-v}{t}$ such sequences, the following expression of P_{SD} is obtained:

$$P_{SD}(h, v, h_v) = \sum_{x=0}^{\min(v,h)} \left[\sum_{y=\max(0, x-h_v)}^{\min(x, v-h_v)} \binom{v-h_v}{y} \binom{h_v}{x-y} (1-P_e)^v \left(\frac{P_e}{1-P_e} \right)^{h_v - x + 2y} \right] \cdot \left[P_m(\{s_1, \ldots, s_{L-v}\}) + \sum_{z=1}^{\min(h-x, L-v)} \sum_{t=1}^{u_z} P_m(\mathbf{r}_t) \right], \tag{2.18}$$

where $P_m(\{s_1, \ldots, s_{L-v}\})$ refers to the probability that the first $L - v$ marker symbols are received without a mismatch, and $P_m(\mathbf{r}_t)$ refers to the probability of the t^{th} out of u_z possible sequences containing z mismatches to the marker symbols $\{s_1, \ldots, s_{L-v}\}$.

ad 3) The probability P_{MD} of a missed detection of a sync word is not influenced by the sequence statistics since it only depends on the error probability P_e of the channel. Therefore, it remains the same as derived in Section 2.3.1 for i. u. d. symbols:

$$P_{MD} = 1 - P_{CD} = 1 - \sum_{x=0}^{h} \binom{L}{x} (1 - P_e)^{L-x} P_e^x. \tag{2.19}$$

2.3.3 Threshold detection of periodically inserted sync words

In the case of periodically inserted sync words, each frame length N_f contains exactly one sync word (SW) (see Figure 2.3). To derive a bound on the performance of threshold detection of periodically inserted sync words, the following worst case scenario is considered: The first received symbol is s_2, i.e. the next sync word lies $L + D = N_f$ symbols ahead. Then, the probability $P(N_f)$ for a correct detection at position N_f is given by

$$\begin{aligned} P(N_f) \;=\; & \Pr\{\text{detection at } \mu = N_f \mid \text{no detection for } \mu < N_f\} \cdot \\ & \Pr\{\text{no detection for } \mu < N_f\}. \end{aligned} \tag{2.20}$$

When taking the three error sources into account, this probability can be bounded by (see [Sch80])

$$P(N_f) \geq \left[1 - (D - L + 1) \cdot P_{FD} - 2 \cdot \sum_{v=1}^{L-1} P_{SD}(h, v, h_v)\right] \cdot P_{CD}. \tag{2.21}$$

The factor $(D - L + 1) \cdot P_{FD}$ is the probability that the sync word occurs by chance in the D symbols between SW_l and SW_{l+1}. The term $2 \sum_{v=1}^{L-1} P_{SD}(h, v, h_v)$ is the probability of a shifted synchronization of a part of SW_l with the successive symbols or of a part of SW_{l+1} with the preceding symbols. If the sync word is well designed (see Section 2.4 for details about sync word design), the following assumption holds true [Sch80]:

$$P_{SD}(h, v, h_v) \leq P_{FD}. \tag{2.22}$$

Plugging this into Eq. (2.21) yields

$$P(N_f) \geq [1 - (N_f - 1) \cdot P_{FD}] \cdot P_{CD}. \tag{2.23}$$

R. A. Scholtz simulated the behavior of $P(N_f)$ as a function of L for binary sync words with $P_e = 0.01$, data frames of length $D = 1000$ and different values of h [Sch80]. He found out that a success probability of at least 0.9994 can be achieved if 3.6 % of the transmitted

symbols are spent on markers. P. Robertson later showed that even for transmission over an AWGN-channel ($E_b/N_0 > 3$) with BPSK modulation, a success probability of at least 0.99 can be achieved for a sync word length of $L = 13$ and of at least 0.999 for $L = 18$ ($N_f = 130$ and $N_f = 135$, respectively) [Rob95].

2.3.4 Threshold detection of aperiodically inserted sync words

In the case of aperiodically inserted sync words, two successive sync words SW_l and SW_{l+1} are spaced at distance T_l (see Figure 2.4). The lower bound of the probability of a correct detection of the sync word SW_{l+1} at position T_l is again derived by considering the following worst case scenario: The first received symbol is s_2 of SW_l, i.e. the successive sync word SW_{l+1} lies T_l symbols apart. Then, Eq. (2.21) can be generalized to

$$P(T_l) \geq \left[1 - (T_l - L + 1) \cdot P_{FD} - 2 \cdot \sum_{v=1}^{L-1} P_{SD}(h, v, h_v)\right] \cdot P_{CD} \tag{2.24}$$

$$\overset{\text{Eq. (2.22)}}{\geq} [1 - (T_l + L - 1) \cdot P_{FD}] \cdot P_{CD}. \tag{2.25}$$

2.4 Sync word design

In Section 2.3.1, the performance of frame synchronizers was derived based on the assumption that $P_{SD}(h, v, h_v) \leq P_{FD}$ (see Eq. (2.22)) for well-designed sync words without going into details. In this section, the design of sync words depending on characteristics of the transmission channel is presented. The major criteria are introduced which are applied to search for sequences that yield low probabilities of false detections on random data (Section 2.4.1) as well as of shifted synchronizations (Section 2.4.2).

2.4.1 Random occurrence of the sync word

To avoid false detections of the sync word, it should be chosen such to occur with a small probability P_{FD} in random data. In case of i. u. d. symbols, this probability does not depend on the sequence itself but only on its length L (see Eq. 2.10). In case of inter-symbol dependencies, the probability of an occurrence of the sync word in the random data can be calculated from a Markov chain Mm that describes the data (see Section 2.3.2):

$$P_{FD} = \sum_{t=1}^{u} P_m(\mathbf{r}_t), \quad u = \sum_{t=0}^{h} \binom{L}{t}, \tag{2.26}$$

where the $P_m(\mathbf{r}_t)$ are calculated according to Eq. (2.16). The sync word $\mathbf{s}$ should be chosen in a way to minimize P_{FD}, which corresponds to choosing an unlikely word with respect to the data stream. Let $\widetilde{N}_m(\mathbf{r})$ denote the random variable of the count of sequence $\mathbf{r}$ in

the data stream under Markov model Mm. Then, the expected count $\mathbb{E}\{\widehat{N}_m(\mathbf{r})\}$ of this sequence $\mathbf{r}$ of length L depending on the Markov model is [Sch06, RRS05]

$$\mathbb{E}\{\widehat{N}_m(\mathbf{r})\} = \frac{N(\{r_1,\ldots,r_{m+1}\})\cdot\ldots\cdot N(\{r_{L-m},\ldots,r_L\})}{N(\{r_2,\ldots,r_{m+1}\})\cdot\ldots\cdot N(\{r_{L-m},\ldots,r_{L-1}\})} = \frac{\prod_{x=1}^{L-m} N(\{r_x,\ldots,r_{m+x}\})}{\prod_{x=2}^{L-m} N(\{r_x,\ldots,r_{x+m-1}\})}, \tag{2.27}$$

where $N(\mathbf{r})$ denotes the observed number of occurrences of the sequence $\mathbf{r}$ in the data stream. This is simply given by

$$N(\mathbf{r}) = \sum_{\mu=1}^{N_d-L+1} \mathbf{1}(\{d_\mu,\ldots,d_{\mu+L-1}\} = \{r_1,\ldots,r_L\}), \tag{2.28}$$

with $\mathbf{1}(expr)$ equaling one if *expr* is true and zero otherwise. Thus, in order to avoid random occurrences in the data stream, the word with the minimum expected count $\mathbb{E}\{\widehat{N}_m(\mathbf{r})\}$ should be chosen as the sync word $\mathbf{s}$:

$$\mathbf{s} = \operatorname*{argmin}_{\mathbf{r}} \mathbb{E}\{\widehat{N}_m(\mathbf{r})\}. \tag{2.29}$$

If the sync word has to satisfy additional constraints that preclude it to be chosen according to its expected number of occurrences, it should instead be a word that is avoided in the surrounding data, e.g. a word that is no codeword in the coding scheme underlying the data stream. This corresponds to choosing an under-represented word (occurring exceptionally seldom), i.e. the word that maximizes the following probability:

$$\mathbf{s} = \operatorname*{argmax}_{\mathbf{r}} \Pr\{\widehat{N}_m(\mathbf{r}) \geq N(\mathbf{r})\}. \tag{2.30}$$

2.4.2 Shifted synchronization

In addition to avoiding random occurrences of the sync word, the probability of shifted synchronizations should be minimized, i.e. it should be excluded that a prefix or suffix of the sync word together with the surrounding data yields a valid sync pattern. In the following, the common measures for evaluating the quality of a sync word are presented.

▷ **Example 2.1**

We first consider the binary sequence `+1 +1 +1 +1`. In case of equally probable symbols, the probability of the sequence being preceded by a +1 is 0.5, which may lead to a shifted synchronization (one position too early). In contrast to that, the sequence `+1 -1 +1 +1` needs to be preceded by three specific symbols (`+1 -1 +1`) to yield a shifted sync word.

◁

Aperiodic autocorrelation behavior

One way of determining the suitability of a sequence $\mathbf{s}$ to serve as a sync word is to evaluate its aperiodic autocorrelation function (ACF) $\varphi_{ss}(\tau)$:

$$\varphi_{ss}(\tau) = \sum_{k=1}^{L-|\tau|} s_k \cdot s_{k+\tau}. \tag{2.31}$$

The aperiodic autocorrelation function describes the similarity of a sequence $\mathbf{s}$ surrounded by random data to itself for different shifts $\tau \in [-(L-1); +(L-1)]$. Since we consider the aperiodic ACF (and not the periodic one), the surrounding of the sequence $\mathbf{s}$ is assumed to be random and uniformly distributed over the symbol alphabet. Thus, for $|\tau| > L-1$, the autocorrelation function $\varphi_{ss}(\tau)$ represents the expected value (i.e. usually equals zero for uniformly distributed symbols).

High sidelobes of the aperiodic ACF indicate periodicities in the sequence which may yield a shifted synchronization if noise is present. Therefore, the task in sync word design lies in finding sequences with low sidelobes. In case of expected phase ambiguities (e.g. after BPSK modulation), their absolute value should be as small as possible, whereas otherwise they should be as low (and possibly negative) as possible.

In the 1960s, the peak sidelobe level (PSL) was introduced to rate the autocorrelation sidelobes of a sequence [Boe67, Tur68]. It is also known as minimal peak sidelobe [Gol77] or maximum sidelobe correlation [Lev75] and is defined as

$$\mathrm{PSL}(\mathbf{s}) = \max_{\tau \setminus \{0\}} |\varphi_{ss}(\tau)| \tag{2.32}$$

in case of expected phase ambiguities (e.g. after BPSK modulation) and defined as

$$\mathrm{PSL'}(\mathbf{s}) = \max_{\tau \setminus \{0\}} \varphi_{ss}(\tau) \tag{2.33}$$

if no phase ambiguities are present. Thus, the task in ensuring successful frame synchronization over noisy channels lies in choosing a sync word which minimizes the PSL- and PSL'-value, respectively. For the case of no expected phase ambiguities, the merit factor MF is another measure for the synchronization quality of a sequence [Lue92]:

$$\mathrm{MF} = \frac{\varphi_{ss}^2(0)}{2\sum_{\tau=1}^{L-1} |\varphi_{ss}(\tau)|^2}. \tag{2.34}$$

In contrast to the PSL-value, the merit factor should be maximized to avoid shifted synchronizations. Moreover, while the PSL-value only rates the position with the worst effect on the synchronization rate, the MF-value sums over all sidelobes and, thus, is a measure for the overall effect of non-ideal autocorrelation properties on the synchronization rate.

▷ **Example 2.2**

We again consider the two sequences from Example 2.1. Figure 2.6 (left) shows the aperiodic ACF of the sequence `+1 +1 +1 +1`, Figure 2.6 (right) shows that of the sequence `+1 -1 +1 +1`. It can be seen that the former has higher sidelobes than the latter, which makes it a worse sync word. This is confirmed by their PSL-values (which here correspond to the PSL'-values):

$$\text{PSL}(\texttt{+1 +1 +1 +1}) = \varphi_{ss}(\tau = 1) = 3, \quad \text{PSL}(\texttt{+1 -1 +1 +1}) = \varphi_{ss}(\tau = 3) = 1. \tag{2.35}$$

The merit factor also confirms the poor synchronization properties of the sequence `+1 +1 +1 +1`:

$$\text{MF}(\texttt{+1 +1 +1 +1}) = 0.57, \quad \text{MF}(\texttt{+1 -1 +1 +1}) = \varphi_{ss}(\tau = 3) = 4. \tag{2.36}$$

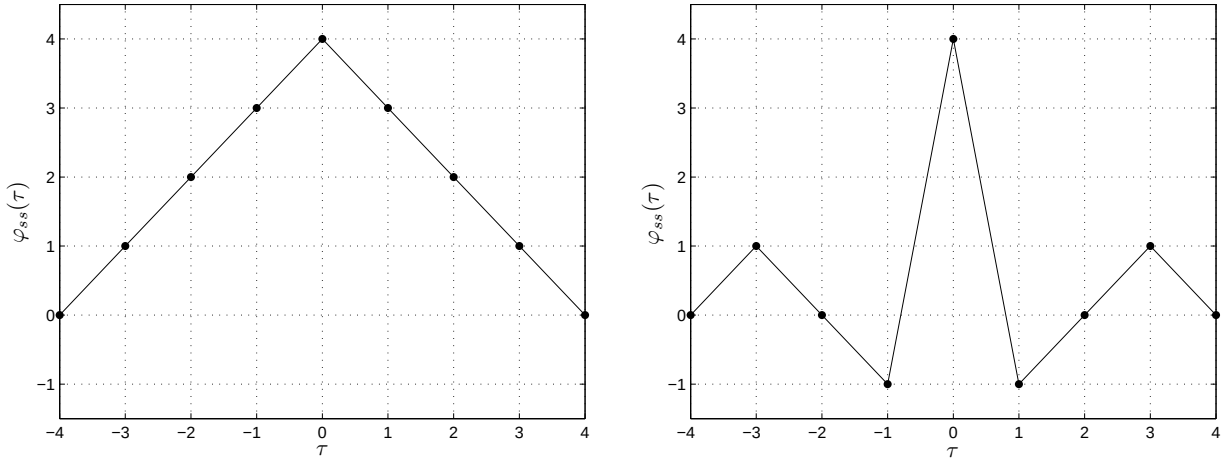


Figure 2.6: Aperiodic autocorrelation functions of the binary sequences `+1 +1 +1 +1` (left) and `+1 -1 +1 +1` (right).

◁

Distribution of correlation values

In 1971, F. Neuman and L. Hofman derived a measure to rate the synchronization quality of a binary sequence transmitted over a BSC channel [NH71]. They focused on sync words for application in systems without phase ambiguities (type I), but extended their results also to the case of expected phase ambiguities (type II). Depending on the surrounding data and the symbol error probability, the correlation between the altered sync word in the data stream and the known sync word $\mathbf{s}$ follows a probability distribution at each shift τ from the complete overlap. To ensure a low synchronization failure probability, the probability distribution at shift $\tau = 0$ should be widely separated from the probability distributions at all shifts $\tau \neq 0$. The resulting distance D_τ between the means of the probability distributions should thus be maximized to minimize the synchronization failure probability:

$$D_\tau = \begin{cases} [L-\varphi_{ss}(\tau)](1-2P_e) - \sqrt{4LP_e(1-P_e)} - \\ \quad \sqrt{4(L-\tau)P_e(1-P_e)+\tau} & \text{for } \tau \leq L \\ L(1-2P_e) - \sqrt{4LP_e(1-P_e)} - \sqrt{L} & \text{for } \tau > L \end{cases}, \tag{2.37}$$

where P_e here refers to the error probability of a BSC channel. In contrast to the peak sidelobe level (PSL), the distance measure D_τ particularly punishes high correlations at high shifts. A special case to consider is the noiseless case, i.e. if the symbol error probability approaches zero and false detections are only caused by random occurrences of the sync word in the data stream:

$$D_\tau = \begin{cases} L - \varphi_{ss}(\tau) - \sqrt{\tau} & \text{for } \tau \leq L \\ L - \sqrt{L} & \text{for } \tau > L \end{cases}. \tag{2.38}$$

It can be seen that $\varphi_{ss}(\tau)$ is the only parameter in Eq. (2.37) and Eq. (2.38) that depends on the sync word design. Using the proposed distance measure D_τ, the following conclusions can be drawn about the quality of sync words in the BSC case:

- ▷ Noisy case: evaluating the elements of Eq. (2.37) that are independent of the structure of the sync word (see [NH71]) shows that for low symbol error probabilities sync words with low autocorrelation sidelobes $\varphi_{ss}(\tau)$ especially for large shifts τ are preferable. For high error probabilities, low and nearly uniform autocorrelation sidelobes are preferable.
- ▷ Noiseless case (Eq. (2.38)): $L - \sqrt{\tau}$ is independent of the sync word structure but punishes correlations at large shifts, thus, the autocorrelation $\varphi_{ss}(\tau)$ should be minimized especially for large shifts.

Therefore, that sequence $\mathbf{r}$ with the lowest PSL'-value should be chosen as sync word $\mathbf{s}$ in noisy systems with high error probabilities (see Section 2.4.2). In the case of low or zero error probabilities, that sequence $\mathbf{r}$ should be chosen that satisfies

$$\mathbf{s} = \underset{\mathbf{r}}{\operatorname{argmax}}\, D_{\tau,\min}(\mathbf{r}), \tag{2.39}$$

where $D_{\tau,\min}(\mathbf{r})$ defines the minimum value of D_τ obtained for the sequence $\mathbf{r}$.

Hamming distance between prefixes and suffixes

Another measure to rate the synchronization properties of a sequence is the Hamming distance h_v between a prefix and suffix of length v. This should be high to preclude shifted synchronizations resulting from concatenations of the prefix or suffix and the surrounding

random data. This results in minimizing $P_{SD}(h, v, h_v)$ for the given value of the error tolerance h of the synchronization algorithm (see Section 2.3). The Hamming distance measure h_v of the pattern and the autocorrelation measure are related by (see [Sch80])

$$\tau = L - v, \tag{2.40}$$

$$\varphi_{ss}(\tau) = \sum_{k=1}^{L-\tau} (-1)^{s_k + s_{k+\tau}} = L - 2h_v. \tag{2.41}$$

▷ **Example 2.3**

For the two sequences from Example 2.1, the following values of h_v are obtained:

sequence	h_1	h_2	h_3	h_4
`+1 +1 +1 +1`	0	0	0	0
`+1 -1 +1 +1`	0	1	2	0

Equivalent to earlier results, the sequence `+1 +1 +1 +1` shows to have the minimum Hamming distance $h_v = 0$ for each v. Thus, it is a poor sync word and will be avoided in most systems.

◁

The optimum case is to use a sync word with $h_v \geq L-1$, a so-called bifix-free sequence. A bifix is a sequence which is both a prefix and a suffix of a longer sequence (see e.g. [Nie73a]). For example, the sequence `+1 -1 +1 +1 +1 -1 +1` is not bifix-free since the first three bits equal the last three bits (i.e. `+1 -1 +1` is a bifix of the sequence).

2.5 Sync word families

2.5.1 Sync words for channels with phase ambiguities

Barker sequences

In 1953, R. H. Barker published his pioneering work on frame synchronization and the design of sync words [Bar53]. He presented sequences with absolute autocorrelation sidelobes smaller than 1, i.e. with PSL ≤ 1 (see Eq. (4.3)). All existing Barker sequences with $2 \leq L \leq 13$ are listed in Table B.1 (Appendix B). It was shown that no Barker sequences exist for $13 < L \leq 12100$ and that the one with $L = 4$ is the only Barker sequence of even length [TS61]. Since Barker sequences minimize the absolute autocorrelation sidelobe (i.e. the PSL), they are only suitable for application in noisy systems where phase ambiguities are expected after demodulation.

Neuman-Hofman sequences (type II)

In 1971, F. Neuman and L. Hofman applied their distance measure D_τ (see Section 2.4.2) to search for sync words with desirable sync word properties [NH71]. They found sequences with $7 \leq L \leq 24$ for BSC channels with phase ambiguities (see Table B.3, Appendix B). A comparison between Neuman-Hofman sequences (type II) and Barker sequences can be found in [Mas72], where the former outperformed the latter on unmodulated binary data streams for the considered SNR range of $0.5 \leq E_b/N_0 \leq 2$.

2.5.2 Sync words for channels without phase ambiguities

Sequences found by Maury and Styles

In 1965, J. L. Maury and F. J. Styles presented their search for sync words for channels without phase ambiguities aimed at usage in PCM telemetry [MS64]. They tried to minimize the probability of shifted synchronizations by minimizing the values of the so-called agreement vector which is inversely related to the Hamming distance h_v between suffices and prefixes of the sync word. The error tolerance was set to $h = 2$ and the symbol error probability was set to $P_e = 0.1$. Table B.2 (Appendix B) lists the found sequences with $7 \leq L \leq 30$.

Neuman-Hofman sequences (type I)

In addition to the Neuman-Hofman sequences suitable for channels with expected phase ambiguities (see Section 2.5.1), F. Neuman and L. Hofman also applied their distance measure D_τ to search for sync words for application in BSC systems without phase ambiguities. They focused on sequences with good properties for high symbol error probabilities. The found sequences with $7 \leq L \leq 24$ are listed in Table B.4 (Appendix B).

2.5.3 Bifix-free sequences

As mentioned Section 2.4.2, a bifix is a sequence which is both a prefix and a suffix of a longer sequence. In noise-free systems, the autocorrelation properties play a minor role, instead, the sync word should be bifix-free. In most cases, one will try to find a pattern that is bifix-free but at the same time has preferable autocorrelation properties to ensure independence of the noise in the transmission system [Nie73a]. Table B.5 (Appendix B) lists half of the existing bifix-free sequences for $2 \leq L \leq 6$, the second half is created by exchanging zeros and ones of the listed sequences. Note that the bifix-free sequences $aa \ldots ab \ldots bb$ as well as $aa \ldots ab$ and $ab \ldots bb$ – where a and b denote two letters from the alphabet $\mathcal{A}$ – generally have bad autocorrelation properties and thus will be avoided in most systems.

2.5.4 Distributed sequences

In [dLvW98] and [dLvWW00], A. J. de Lind van Wijngaarden proposed the use of so-called distributed sequences. These are defined as patterns containing constrained and unconstrained symbols, i.e. containing synchronization symbols interspersed with data symbols. One example of a distributed sequence is the pattern 10001**1**1, which has seven constrained symbols (synchronization symbols) and four unconstrained positions (*) that can take on any value. In [dLvWW00], A. J. de Lind van Wijngaarden and T. J. Willink presented a performance comparison between the Barker sequence 1110010 and the distributed sequence 111**0***0*10. For this purpose, they considered not only the autocorrelation function of the sequences but also the maximum correlation that it can yield if taking the surrounding data symbols into account. This analysis brought up that the correlation peak of the Barker sequence is more distinct, however, in case of the distributed sequence the correlation values remain below the peak over more positions. They used these correlation properties to search for distributed sequences that are bifix-free and at the same time fulfil the Barker-criterion PSL ≤ 1 (i.e. distributed sequences suitable for channels with expected phase ambiguities). The found sequences for $5 \leq L \leq 32$ are listed in Table B.6 (Appendix B).

2.6 Summary

In this chapter, frame synchronization in digital data transmission was introduced. This included the synchronization performance, major error sources and the design of sync words depending on the channel characteristics. Four main aspects hereof are particularly important for the gene expression models derived in later chapters:

- ▷ The detection of the sync word at the receiver side is based on a likelihood function $L(\mu)$ which measures the similarity between the incoming data stream and the sync word at each position μ. In technical systems, $L(\mu)$ is generally defined based on the cross-correlation function.
- ▷ The sync word should be designed such that the probability of shifted and false synchronizations on random data is minimized.
- ▷ The probability of shifted synchronizations is minimized if choosing sync words with lowest possible self-similarity. This is rated using e.g. the autocorrelation function of the sync word which should not exhibit high sidelobes.
- ▷ The probability of false synchronizations through random occurrences of the sync word does not depend on the sync word in case the data stream carries no statistical dependencies. If it does, however, the sync word should be chosen such that its random occurrences in the data stream are minimized. This aspect is commonly evaluated using a Markov model of the data stream.

3

Biological Background

This chapter provides the biological background necessary to understand the communication theory models in later chapters. In Section 3.1, the DNA is described as a digital signal. Section 3.2 introduces important historical steps that led to today's knowledge about genetics and molecular biology. Section 3.3 deals with basic terms and definitions. This includes the structure of DNA and RNA, genes, mutations as well as the organization of genomes of both bacteria (prokaryotes) and higher organisms (eukaryotes). In Section 3.4, a detailed description of gene expression, the process of protein synthesis, is presented. The main steps and involved components are presented for prokaryotic and eukaryotic organisms. Section 3.5 introduces interactions between proteins and the DNA that occur during all steps of gene expression. The fundamental interactions are presented together with common theories of target search on the DNA by proteins. Finally, the analogies between gene expression and digital data transmission are discussed in Section 3.6. For an in-depth introduction to molecular biology refer to [Lew07, AJL+02].

3.1 The DNA as a digital signal

The deoxyribonucleic acid (DNA) is the primary carrier of genetic information, which can be seen as a digital signal of the quaternary alphabet $\mathcal{A} = \{A, C, G, T\}$. In digital data transmission, information is processed in numerous steps: it is read out, transformed (modulation, coding), transmitted, possibly altered by transmission errors, corrected and interpreted (see Figure 3.1). The genetic information stored in the DNA comes into effect only after transformation into proteins, molecules that determine many genetic traits of living beings. This process takes place in a series of transformation steps similar to those in digital data transmission: parts of the DNA sequence are read out, transformed into

different alphabets, possibly altered by mutations and corrected. In this chapter, the basic steps underlying protein synthesis are detailed.

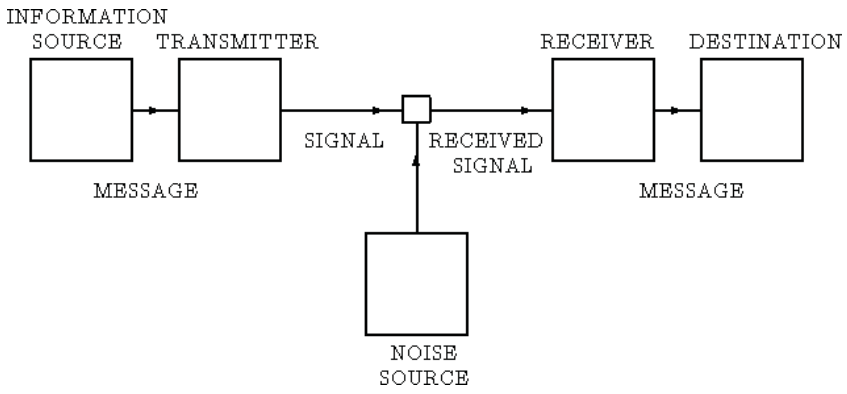


Figure 3.1: Illustration of a digital transmission system as proposed by C. E. Shannon in his famous Fig. 1 [Sha48].

3.2 Historical steps in molecular biology

The first step in the long story of research in the field of genetics was done by monk Gregor Mendel, who investigated inheritance in the pea plant and published his results and the Mendelian laws of heredity in 1866 [SMC08]. In the beginning of the 20$^{\text{th}}$ century, scientist Thomas Morgan conducted experiments with *Drosophila melanogaster*, the fruit fly. In the course of his research, he already identified the genes that are responsible for certain traits of the external appearance [Wat04]. Nonetheless, for a long time, biologists had difficulties accepting the deoxyribonucleic acid (DNA) as the carrier of genetic information due to its apparent chemical simplicity (see Section 3.3.1). The three-dimensional chemical structure of the DNA could first be obtained in 1953, when James Watson and Francis Crick brought up an exact model of the DNA molecule and proved that genes determine heredity. It took however until the 1960s until the transformation of DNA sequences into proteins following the genetic code was widely understood (see Section 3.4.4) [Hay98]. Since then, many ground-breaking discoveries have been made that led to a better understanding of heredity as well as to the sequencing of an increasing number of complete genomes. After years of experiments, the sequencing of the complete human genome along with the identification of the majority of genes was achieved in 2003 [NIoH08a] [DoEOoS08].

3.3 Terms and definitions

3.3.1 DNA and RNA

The DNA is formed by two strands, linked together and twisted in the shape of a double helix. The strands consist of a chain of nucleotides, small molecules built up by a nucle-

obase, a pentose sugar and a phosphate. The nucleobase can be of four types: adenine (A), cytosine (C), guanine (G) and thymine (T). The larger nucleotides adenine and guanine belong to the class of a double-ringed chemical structure called purine. They form hydrogen bonds with their respective complements thymine and cytosine, belonging to the single-ringed pyrimidines (see Figure 3.2, right). Two nucleotides on opposite complementary DNA strands that are connected via hydrogen bonds are called a base pair (in the following abbreviated as bp). Base-pairing in the DNA can exclusively occur between A and T as well as between G and C [Lew07]. Other bonds are unfavorable since the patterns of hydrogen acceptors and donors do not match: While adenine and thymine bind via two hydrogen bonds, cytosine and guanine are connected via three hydrogen bonds (see Figure 3.3). The binding between the two strands is called Watson-Crick base-pairing. A DNA sequence is typically written from its 5'-end to its 3'-end, where the naming originates from the chemical structure of the pentose sugar. With respect to this directionality of the DNA, the relative position of a sequence element is either denoted as upstream (towards the 5'-end) or downstream (towards the 3'-end).

▷ **Example 3.1**

An exemplary nucleotide sequence (upper line) and its Watson-Crick complement (lower line) look as follows:

$$\begin{array}{lcl} 5'\ldots & \texttt{T A A C G C A T G C C T A A G} & \ldots 3' \\ 3'\ldots & \texttt{A T T G C G T A C G G A T T C} & \ldots 5' \end{array}$$

◁

During protein synthesis (gene expression), parts of the DNA are transformed into RNA (ribonucleic acid) in the process of transcription (see Section 3.4). The difference to DNA lies in the chemical structure: RNA is single-stranded and uracil replaces thymine as the base complement to adenine (see Figure 3.2, left). Since RNA is single-stranded, it often contains short sequences of nucleotides that can base-pair with complementary sequences found somewhere else on the same molecule [AJL+02]. These interactions arrange for an RNA molecule to fold into a stable three-dimensional structure, which allows it to play regulatory roles during protein synthesis.

3.3.2 Mutations

Mutations are changes happening to the nucleotide sequence of genetic material (DNA or RNA). They correspond to transmission errors and noise in communication systems that alter the signal during transmission. Mutations occur due to both external influences (like radiation) and errors during cell processes (like replication). Three basic types of small-scale mutations exist [GGML99]:

- ▷ Point mutation: a single nucleotide is exchanged for another one. The most common point mutation exchanges a purine for a purine (A↔G) or a pyrimidine for

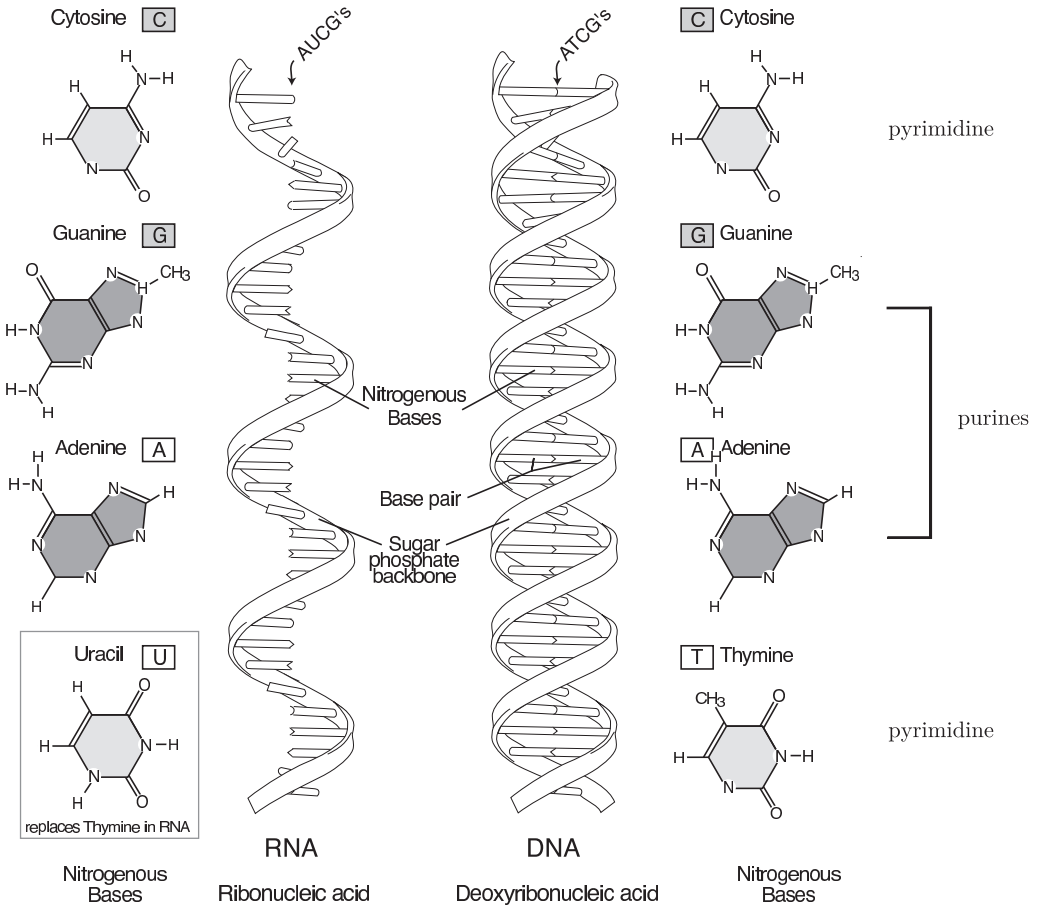


Figure 3.2: Structure of RNA (left) and DNA (right) [NIoH08b].

a pyrimidine (C↔T). This type of mutation is called a transition as opposed to a transversion which exchanges a purine for a pyrimidine or vice versa (A/G↔C/T).

- ▷ Deletion: one or several nucleotides are deleted irreversibly from their position in the DNA.
- ▷ Insertion: one or several nucleotides are inserted at a random position in the DNA.

The effect of mutations depends on their position and on whether they effect the synthesis of a protein. In the majority of cases, the effect is either neutral or can be corrected before bringing harm to the organism. In some cases, mutations are harmful or – in rare cases – beneficial and, thus, lead to an evolutionary change through positive or negative selection.

3.3.3 Genes and proteins

A gene is a sequence of nucleic acids containing the information for a functional product (usually a protein). Proteins (Greek *prota* = 'of primary importance') are large organic

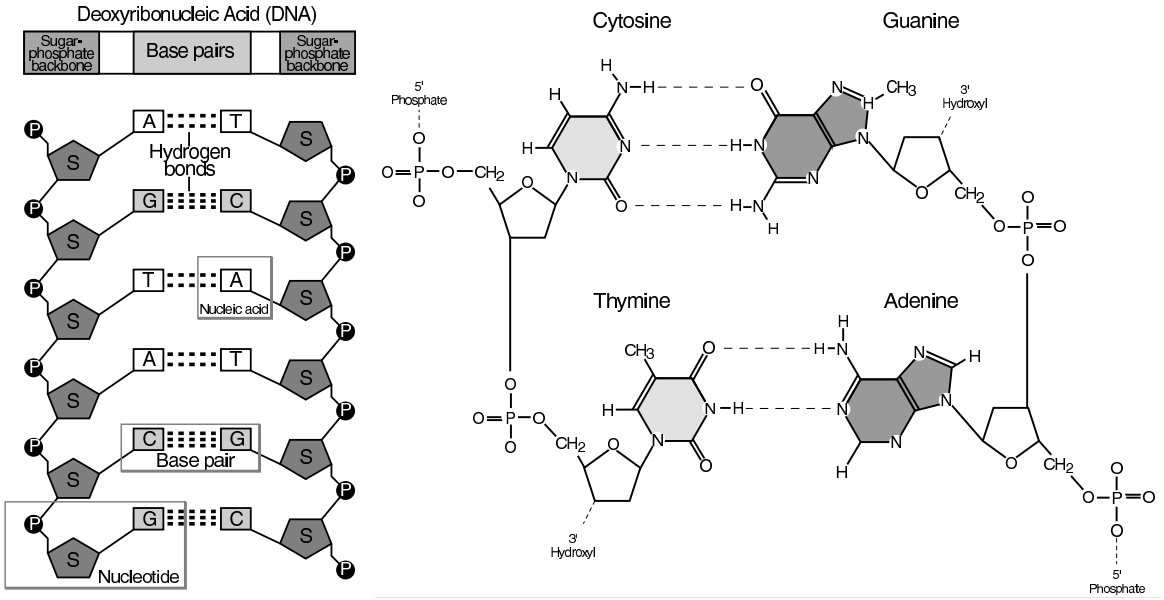


Figure 3.3: Bonding between complementary bases (N: nitrogen, H: hydrogen, O: oxygen, P: phosphate) [NIoH08b].

compounds that constitute an essential part of all living beings [Lew07]. They are responsible for oxygen transport, cell signaling, catalysis of biochemical reactions, immune response as well as maintaining the cell scaffold. While it was long believed that one gene codes for one protein which is itself responsible for one trait, many exceptions to this rule have been found until today [Pea06]. Nonetheless, in the following, the term gene is used in its traditional definition as those parts of the DNA that are copied into mRNA in the process of transcription (see Section 3.4).

3.3.4 Genome

The term genome refers to the entire genetic information or hereditary material possessed by an organism, i.e. the entirety of genes and extra-genic DNA [AJL+02]. The latter refers to those parts of the DNA that are not transformed into mRNA during gene expression. The organization of the genome depends on the organism: While simple organisms carry only a single chromosome organized as a ring, the majority of higher organisms has between 8 and 100 chromosomes organized in an X-shape [Lew07]. In addition to the organization, the length of the genome varies significantly between organisms (bacterium *Carsonella ruddii*: $1.6 \cdot 10^5$ bases, human: $3 \cdot 10^9$ bases).

3.3.5 Prokaryotic and eukaryotic organisms

Organisms are classified into two basic families, namely prokaryotes and eukaryotes (Greek *pro* = 'before', *eu* = 'true', *karyon* = 'kernel'). Prokaryotes comprise all organisms from

the families archaea and bacteria (see Figure 3.4). Prokaryotic organisms are in most cases unicellular, and their cells have no cell nucleus, i.e. the genetic material is not membrane-bound but freely floating in the cytoplasm. The DNA of prokaryotes consists of one single circular chromosome localized in an area called nucleoid. The single chromosome is densely packed with genes (typically several thousand [TPM07]), only few percent are non-coding and serve regulatory purposes. Research on prokaryotes strongly focuses on the bacterium *Escherichia coli* (*E. coli*), which infects the lower intestines of mammals. Eukaryotes comprise all unicellular and multicellular organisms whose cells contain a cell nucleus. The genetic information is stored in chromosomes localized inside the membrane-bound nucleus which is surrounded by the cytoplasm. In contrast to prokaryotes, the chromosomes in eukaryotes contain a high percentage ($> 90\%$) of DNA not coding for proteins. Eukaryotes comprise all higher organisms like plants and animals. The best studied eukaryotes are the human (*Homo sapiens*), the thale cress (*Arabidopsis thaliana*), the fruit fly (*Drosphila melanogaster*) as well as the yeast species *Sacharomyces cerevisiae*.

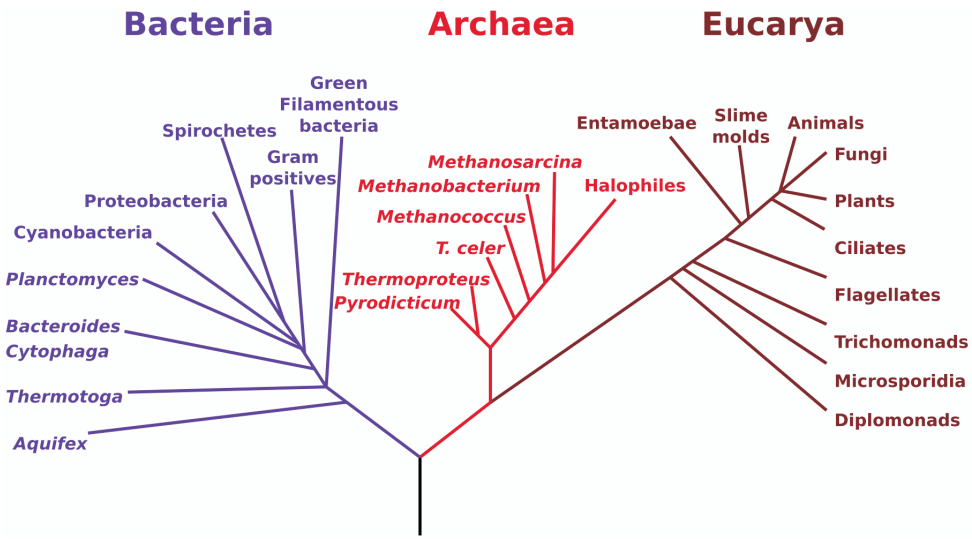


Figure 3.4: The phylogenetic tree of life.

3.4 Gene expression

3.4.1 Overview

Gene expression is the process in which the information stored in the DNA is used to synthesize proteins. It takes place in two basic steps:

$$\text{DNA} \xrightarrow{\text{Transcription}} \text{mRNA} \xrightarrow{\text{Translation}} \text{Protein}$$

During the process of transcription, the double stranded DNA serves as a template to synthesize the single stranded mRNA (messenger RNA, see Figure 3.5, middle). In the second step of gene expression (translation), this mRNA is translated into proteins by chaining amino acids (see Figure 3.5, bottom).

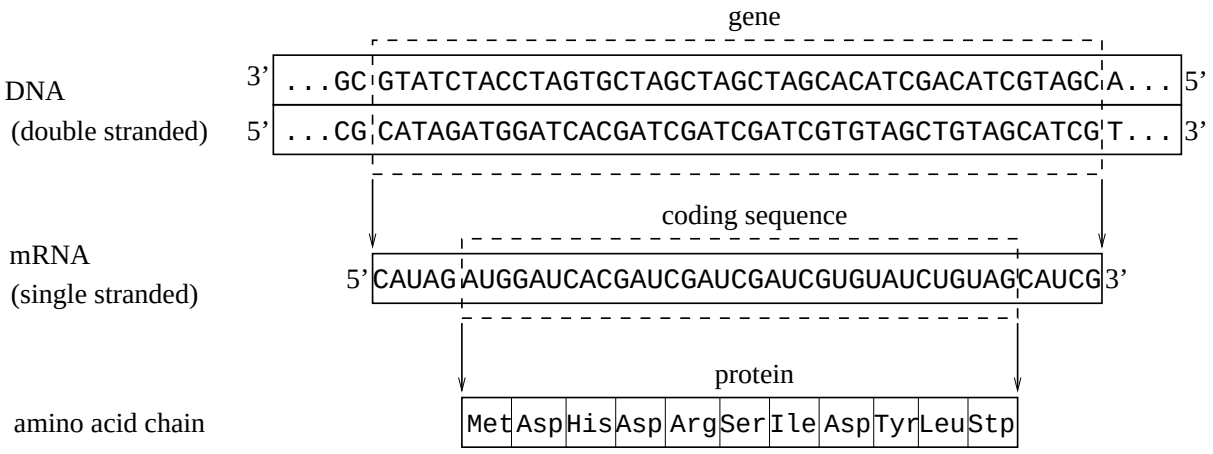


Figure 3.5: Illustration of sequence transformations during gene expression.

The main differences in the gene expression of prokaryotes and eukaryotes are:

- ▷ In eukaryotic organisms, the mRNA only contains the information to translate one protein (monocistronic mRNA), but it may as well contain the information for several proteins in prokaryotes (polycistronic mRNA).
- ▷ After transcription, the mRNA of eukaryotes consists of coding regions (exons) and non-coding regions (introns) separating the exons. In the process of splicing, the introns are cut out of the mRNA yielding the so-called mature mRNA.
- ▷ In contrast to prokaryotes, translation and transcription of eukaryotes are locally separated. Transcription and splicing take place inside the nucleus membrane, whereas translation takes place in the cytoplasm surrounding the nucleus membrane. Therefore, the mature mRNA is exposed to additional radiations and thermal noise during its travel to the less protected cytoplasm.
- ▷ Due to the missing separation of transcription and translation in prokaryotic cells, no intermediate step lies between transcription and translation, which allows simultaneous processing, i.e. the mRNA can already be translated while still being synthesized through transcription.

The process of transcription is described in Section 3.4.2 for prokaryotes and in Section 3.4.3 for eukaryotes. Subsequently, the process of translation is detailed in Section 3.4.4 for prokaryotes and in Section 3.4.5 for eukaryotes.

3.4.2 Prokaryotic transcription

During transcription, a part of the DNA (the gene) is copied into mRNA (see Figure 3.5). This step is performed by the macromolecule RNA polymerase (RNAP) and its sigma subunit which first randomly bind to the DNA and move along it [Lew07]. Equivalent to frame synchronization in continuous transmission, a short DNA motif (the so-called promoter) informs the RNA polymerase about the upcoming gene start. After the sigma factor recognizes the promoter, the RNA polymerase opens and unwinds the DNA (also called DNA melting) on a range of around 12 base pairs to enable the copying of one strand [LBZ+00]. The sigma factor does not play a role in this copying process: in around 30 % of the cases, it dissociates from the RNA polymerase directly after initiation, while it otherwise dissociates at random points during transcription [GvH05]. During transcription elongation, the RNA polymerase moves along the DNA, opens the double helix and copies one strand (the so-called coding strand) by building the complement of the template strand (see Figure 3.6). Termination of transcription is either induced by an RNA-binding protein or based on sequences in the RNA that fold into hairpin structures that mechanically interrupt transcription [AJL+02]. After dissociating from the DNA, the RNA polymerase can bind to another sigma factor and restart the process.

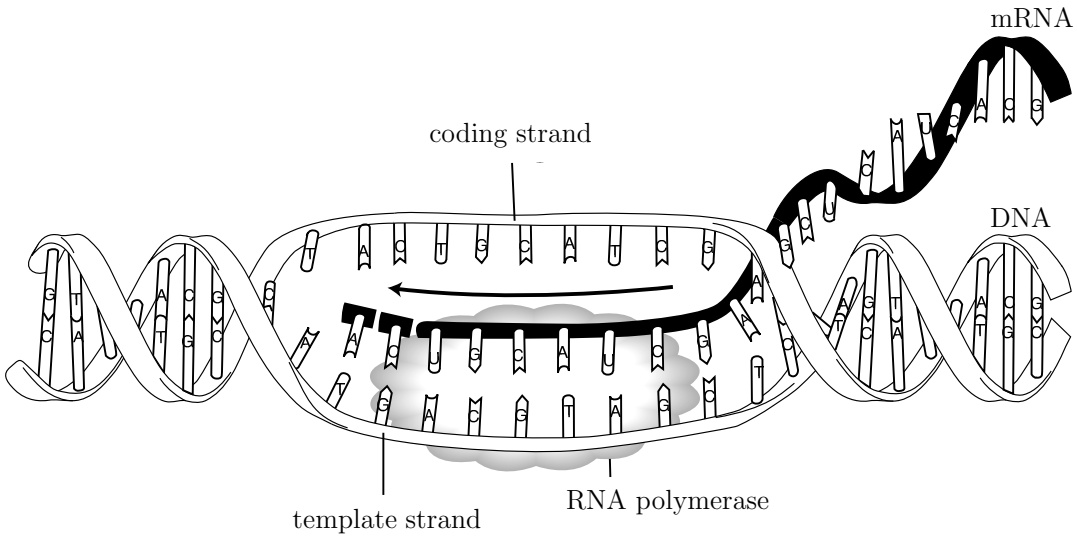


Figure 3.6: Transcription by RNA polymerase [NIoH08b].

Sigma factor

As mentioned before, the sigma factor is a kidney-shaped molecule that attaches to the RNA polymerase to enable promoter detection. In the communications engineering sense, it corresponds to the synchronization unit of the RNA polymerase responsible for detection of the sync word (the promoter, see Figure 3.7). Every prokaryotic organism has one main and several alternative sigma subunits, each of which transcribes its own set of genes. The

main sigma factor is responsible for transcription under exponential cell growth, i.e. for transcription of the so-called housekeeping genes. The alternative ones come into effect to adapt protein synthesis to certain environmental circumstances, e.g. temperature rise or changes in food supply. At the same time, the mentioned external factors activate anti-sigma factors that form complexes with their cognate factors to inhibit their function. This happens for instance after a heat shock, i.e. a steep rise of the temperature: The main sigma factor gets bound and hereby inhibited by its anti-sigma factor, while at the same time one alternative sigma factor becomes chemically more stable, and thus its probability to bind to a RNA polymerase is increased. In *E. coli*, six alternative sigma factors exist in addition to the main sigma factor σ^{70} to ensure expression of specific sets of genes under various environmental conditions (see Table 3.1).

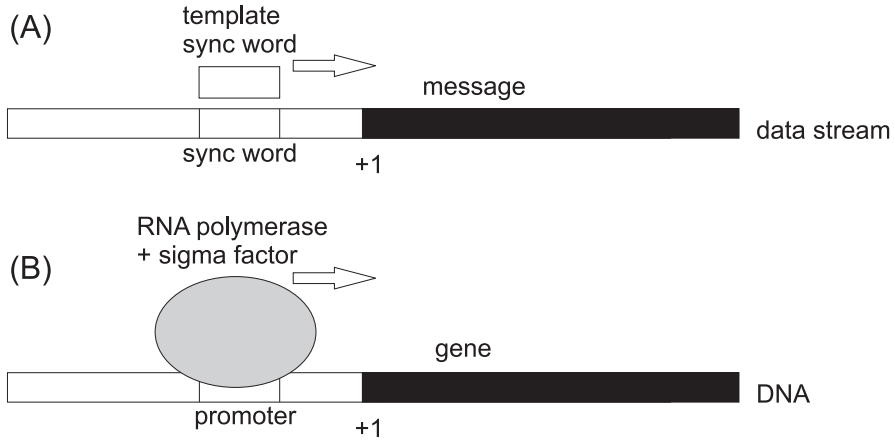


Figure 3.7: Analogy between frame synchronization (A) and promoter detection (B).

Table 3.1: Functions of the seven sigma factors in *E. coli* [GG03]. The exponents refer to the molecular weight in kilo Dalton (kD).

sigma factor	**function**
σ^{70}	exponential cell growth
σ^{54} (σ^{N})	nitrogen starvation
σ^{38} (σ^{S})	general stress conditions (stationary phase)
σ^{32} (σ^{H})	heat shock
σ^{28} (σ^{F})	flagellar development
σ^{24} (σ^{E})	extreme heat stress
σ^{19} (σ^{FecI})	regulation of the iron transport

Promoter

The prokaryotic promoter is the sync word of transcription and consists of two sequences of length six: the -35 region and the -10 region, named after their approximate position before the gene start. Each set of promoters associated with one sigma factor has a consensus sequence, which can be seen as the template sync word the sigma factor uses to search for sync word positions in the DNA.

Definition 3.1

The consensus sequence is a way of representing the results of an alignment of related sequences, e.g. known binding sites of a certain protein. High positional nucleotide biases in the alignment indicate a functional significance of these positions for the underlying protein-DNA interaction. Therefore, the consensus sequence is built up using the most frequently observed nucleotide at each position.

□

The consensus sequence of σ^{70} consists of the -35 region TTGACA, a spacing of 19 arbitrary nucleotides and the -10 region TATAAT. However, only in few cases the actual promoter corresponds in all bases to the consensus. More than 90 % of promoter sequences differ in at least one nucleotide from the consensus. This fact constitutes an important possibility for the regulation of protein synthesis: the degree of divergence determines the rate of synthesis of the corresponding protein, i.e. genes with promoter sequences that are far from the consensus (so-called weak promoters) are less often transcribed than genes with promoters close to the consensus (so-called strong promoters) [AJL+02]. In communication systems, this would correspond to intentionally introducing errors in the sync words (at the transmitter side) to determine their rate of detection by the receiver. While this application of a "soft sync word" hardly makes sense in technical frame synchronization systems, it allows nature a first rough adjustment of the transcription rate according to the cell requirements.

3.4.3 Eukaryotic transcription

Transcription in eukaryotic organisms is far more complex than in prokaryotes. It involves numerous proteins (so-called general transcription factors) that interact with each other and the DNA to detect the promoter and initiate transcription. Promoter detection can still be seen as a process of synchronization, however, it involves several biological sync words and a complex synchronization unit of many interacting proteins (the so-called transcription initiation complex). Additionally, in contrast to prokaryotic cells, the nuclei of eukaryotes contain three RNA polymerases RNAP I, RNAP II and RNAP III, each of which is responsible for a different set of genes. RNAP II is most similar to bacterial RNAP and responsible for the majority of genes [Ebr00]. The complexity of eukaryotic transcription and the involvement of so many factors makes it highly flexible in its response to environmental changes and tissue specific requirements.

Assembly of the transcription initiation complex

Six main transcription factors are involved in transcription by RNAP II (therefore denoted as TFII): In the first step, the transcription factor TFIID binds to the DNA, more precisely to the TATA-box (the main promoter of eukaryotic transcription). During this step, its subunit TBP (TATA-binding protein) is responsible for the recognition of the TATA-box as well as a deformation of the double helix [AJL+02]. While it was long believed that every gene has a TATA-box, many exceptions have been found in the last years. In those cases, the transcription complex assembles on other promoter elements around the transcription start site. Nonetheless, the TBP can roughly be considered as the main synchronization unit of the transcription complex. After binding of TFIID to the TATA-box, two other transcription factors (TFIIA and TFIIB) bind to the complex of TBP and TFIID with the DNA, enabling the binding of the RNAP II and the three remaining transcription factors TFIIE, TFIIF and TFIIH [Lat04]. After DNA melting and the first transcribed nucleotides, all transcription factors except TFIIF dissociate from the DNA, and RNAP II processes elongation. The end of the gene and thus, termination of transcription is again detected by a transcription factor. The exact order of transcription factors binding to each other and the DNA is not known with certainty and depends on the transcribed gene, e.g. some transcription factors may assemble before binding to the DNA. The functions of the transcription factors involved are listed in Table 3.2 in the order of the assembly of the transcription initiation complex.

Table 3.2: Role of transcription factors during transcription by RNA polymerase II.

transcription factor	function
TFIID	recognizes the core promoter
TFIIA	stabilizes TFIID
TFIIB	enables the binding of RNA polymerase II
TFIIE	enables binding of TFIIH
TFIIF	guides the RNA polymerase to the promoter
TFIIH	unwinds the DNA and eases the start of transcription

Promoter

The promoters of eukaryotes can be classified into three subgroups: core, proximal and distal promoter [CH07]. The first corresponds in function and structure to the promoter of prokaryotes and refers to the range of 35 bp before the transcription start site (gene start). It contains the main sync words of eukaryotic transcription that are detected by proteins of the transcription complex. The second class of promoters, the proximal promoters, includes binding sites for transcription factors up to 250 bp before the transcription start

site. The third class, the distal promoter, denotes all binding sites with more than 250 bp distance to the transcription start site. Through binding of additional transcription factors to these sites, the amount of transcription of the respective gene is regulated.

3.4.4 Prokaryotic translation

During translation, the mRNA is transformed into a protein. This step is performed by the ribosome, a complex of two large subunits that are themselves made up of protein subunits as well as rRNAs (ribosomal RNAs). The larger subunit is denoted as 50S subunit, the smaller one as 30S subunit, together building the 70S ribosome[1]. It is important to note that not the complete mRNA is translated into a protein but only the so-called coding sequence, which is delimited by the start codon AUG and one of the stop codons UAA, UAG or UGA (a codon is a nucleotide triplet).

Initiation

In the first step of translation, the 30S subunit binds to the initiator region of the mRNA (the so-called 5' untranslated region or 5'-UTR, see Figure 3.8). The length of the 5'-UTR varies between 0 and 920 bp with the mean length being around 100 bp [BLZ05,SCLS07]. After binding to the 5'-UTR, the 30S subunit moves rapidly along the mRNA until it detects the start codon (AUG, position +1) and the Shine-Dalgarno sequence (SD), a hexamer located shortly before the coding sequence. The 16S rRNA is the part of the 30S subunit of the ribosome that is responsible for the detection of the Shine-Dalgarno sequence via base-pairing [SJ75]. In the communications engineering sense, the Shine-Dalgarno sequence corresponds to the sync word of translation that needs to be detected by the 16S rRNA, the synchronization unit of the ribosome.

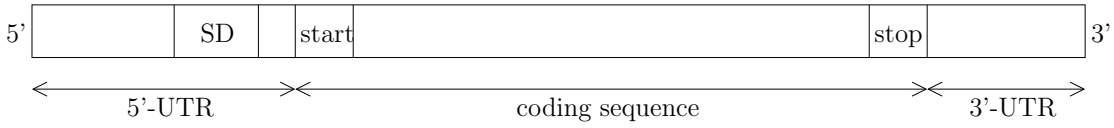


Figure 3.8: mRNA structure in prokaryotes.

Elongation

After the 30S subunit has detected the Shine-Dalgarno sequence, the 50S subunit joins the complex. It then starts translating the coding sequence in steps of three nucleotides (triplets, the so-called codons) beginning with the start codon AUG. At each step, the ribosome serves as a platform for the tRNA (transfer RNA), a functional RNA carrying

[1]S: Svedberg (sedimentation coefficient); dependent on the mass and shape of the molecule as well as the interaction with the medium; not additive.

a specific amino acid (out of 21 possible types of amino acids). On its opposite end, the tRNA has a so-called anticodon, a triplet of nucleotides that is complementary to the currently translated mRNA-triplet (see Figure 3.9, left). The mapping between the anticodon and the amino acid of the tRNA follows the genetic code (see Figure 3.9, right), which defines the relation between the $4^3 = 64$ codons and the 21 amino acids. For example, the codon AGG is mapped to the amino acid Arginine (Arg).

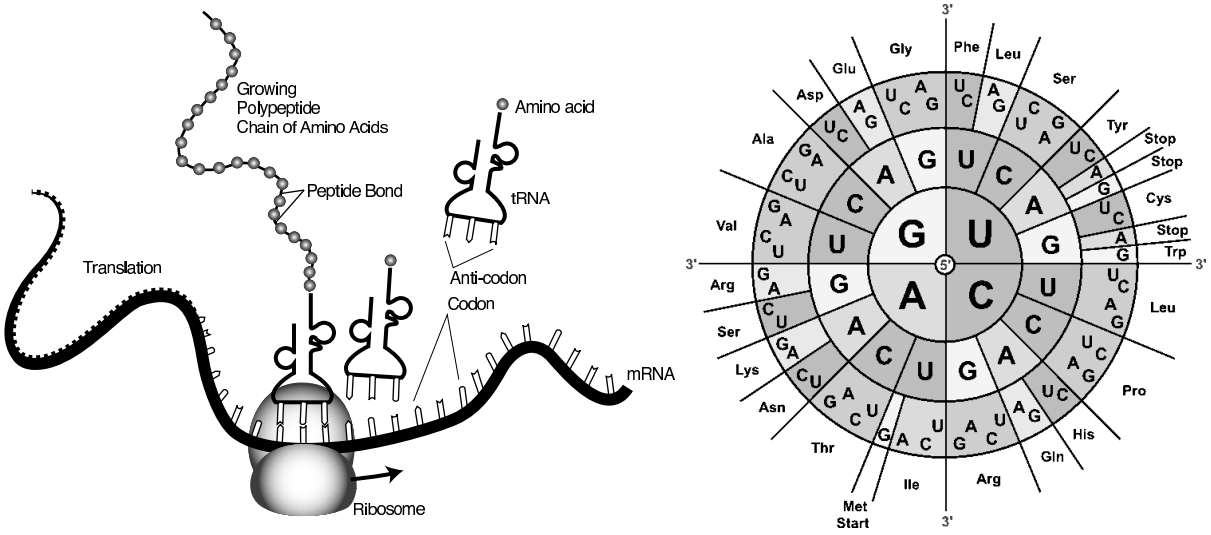


Figure 3.9: The tRNAs map the codons in the mRNA to an amino acid (left) [NIoH08b]. This mapping follows the genetic code (right) [NIoH08b].

3.4.5 Eukaryotic translation

Like it is the case with transcription, translation in eukaryotes involves far more factors than in prokaryotes. The ribosome is larger and comprises a 60S and a 40S subunit, moreover, it is made up of more protein and rRNA subunits. In addition, the mRNA has a cap of methylated guanine bases (denoted as m7G-cap) and carries 100 to 200 adenine bases at its 3'-end (denoted as poly(A)-tail). In the first step of translation, the 40S subunit binds to the 5'-UTR and scans along it until it detects a start codon in a favorable context: This context was described by M. Kozak and is therefore named Kozak sequence [Koz97]. Thus, the Kozak sequence can be seen as the sync word of eukaryotic translation, and the ribosome again corresponds to the receiver in frame synchronization systems. As soon as the 40S subunit has detected the start codon, the 60S subunit joins the complex, and translation starts.

3.5 Protein-DNA interactions

Interactions between proteins and the DNA occur at several steps of cellular processes like gene expression, replication and recombination [Slu05]. They include some of the tightest

and most specific molecular interactions known in molecular biology and are based on bonds of different type and strength. In addition to specific bonds, the interaction involves a deformation of the DNA to achieve the best possible fit of the protein. The protein interacts with the DNA using 10 to 20 weak bonds that add together to ensure that the interaction is both highly specific and very strong [AJL+02]. Although thousands of protein binding sites have already been identified, it is not yet possible to accurately predict contacts between amino acids and base pairs of the DNA [Lew07].

3.5.1 Changes in the DNA geometry

For a long time after the discovery of the DNA structure, it was not clear how proteins read the DNA and thus recognize their specific binding sites without opening the double helix. For 20 years after its discovery, the DNA was thought to have the same monotonous structure with a uniform helical twist. However, in the 1970s, scientists found out that the exact shape of the DNA actually depends on the nucleotide sequence on the inside [AJL+02]: The double helix shows small irregularities in the helical twist angle depending on the nucleotide sequence. Apart from that, the sequence on the inside of the double helix also determines the flexibility for deformations, which is a critical feature for the binding of proteins. In general, AT-rich regions (sequences with a high content of the nucleotides A and T) are more flexible than GC-rich regions, which shows the importance of the TATA-box and the -10 promoter region for transcription initiation (see Section 3.4).

3.5.2 Major and minor groove

In addition to the changes in the overall structure of the DNA depending on the nucleotide sequence, the edges of base pairs constitute an important factor for the recognition by proteins. These edges are exposed on the surface of the helix, presenting a distinctive pattern of bonds [Lew07]. As the two edges of base pairs do not comprise the same angle, the DNA is structured into the major and minor groove (see Figure 3.10). The interactions of proteins with the base pairs inside the double helix occur almost exclusively to the major groove since here – in contrast to the minor groove – the pattern of bonds markedly differs between A-T and G-C base pairs [AJL+02].

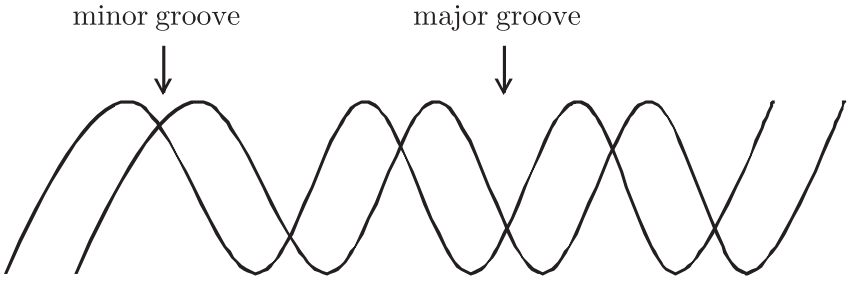


Figure 3.10: Major and minor groove of the DNA.

3.5.3 Fundamental interactions

Binding of proteins to the DNA occurs via different structural motifs containing alpha-helices and beta-sheets, two common folding patterns of proteins [AJL+02]. Both form through hydrogen bonding of amino acids, yielding a regular helix in case of the alpha-helix and a pleated sheet in case of the beta-sheet (see Figure 3.11). The three most common motifs of protein-DNA interactions during gene expression based on these folding patterns are the helix-turn-helix motif, the zinc finger and the leucine zipper motif [AJL+02].

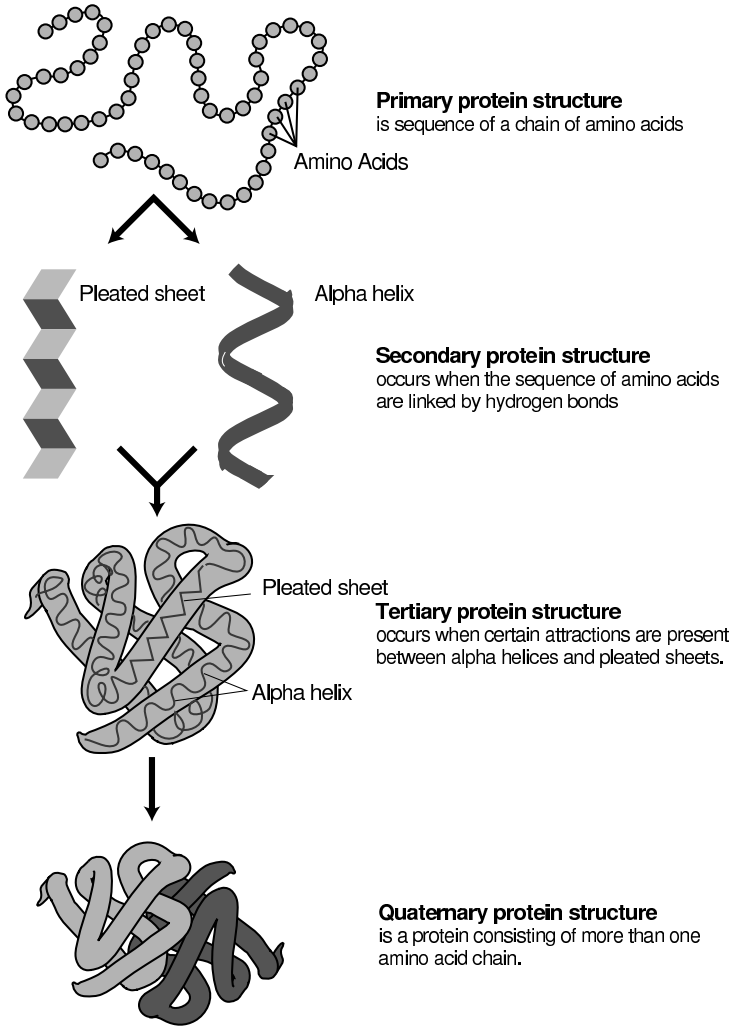


Figure 3.11: Structure of a protein [NIoH08b].

3.5.4 Target search of proteins on the DNA

The exact details of promoter detection by the RNA polymerase (RNAP) or, more generally, how DNA-binding proteins find their cognate site could still not be clarified un-

ambiguously. In the first step, the protein diffuses randomly through the cell (three-dimensional motion) until it associates with the DNA molecule. It was long assumed that if the attached site is not its target site, the protein would dissociate and rebind at another random position. However, Riggs et al. measured in 1970 that the association rate of the *E.coli LacI* repressor and its target sequence on DNA is much higher than the rate achievable by three-dimensional diffusion [RBC70]. In 1981, Berg, Winter and von Hippel published a seminal series of articles presenting a theory for protein-DNA interactions which provided a first explanation for this faster-than-diffusion search [BWvH81, WBvH81]. They conjectured that the dimensionality changes while the protein searches its target site; the protein at first randomly binds to the DNA in a round of three-dimensional diffusion through the cell and subsequently moves along it in a process of one-dimensional diffusion. On short ranges, the one-dimensional search round was assumed to be a sliding process along the double helix. Two additional mechanisms were later suggested to supplement sliding, namely hopping and intersegmental transfer (see [vHB89] and references therein). These three processes could first be visualized in 1999 for the RNA polymerase of *E. coli* over several hundred base pairs [BGZY99, GZR$^+$99].

3.6 Gene expression as a communication system

In this section, the analogies between gene expression in prokaryotes and communications engineering systems are discussed. The following basic analogies exist:

- ▷ The DNA contains all the information necessary for protein synthesis; hence, it can be regarded as the storage medium for the message that is to be transmitted.
- ▷ Transcription initiation corresponds to a process of frame synchronization, where the sigma factor detects the promoter sequences (two biological sync words). Subsequently, the RNA polymerase processes transcription, i.e. reads out the genetic information.
- ▷ Translation can be divided into two steps. At first, the ribosome detects the Shine-Dalgarno sequence (a biological sync word) that marks the beginning of the coding sequence. In the second step, mRNA triplets are mapped to amino acids by the molecule tRNA (transfer RNA). Since three nucleotides are mapped to one amino acid, this step can be seen as a process of decoding during which redundancy is removed.
- ▷ Mutations correspond to transmission errors and noise that is added to the signal during transmission. Mutations occur during all stages of gene expression due to radiation and external influences and can damage the genetic information in the DNA or mRNA.

A channel model of gene expression is depicted in Figure 3.12. Steps (A) - (E) are detailed in the following subsections (Section 3.6.1 - Section 3.6.4). Additionally, the

analogies between frame synchronization and protein-DNA interactions are derived in Subsection 3.6.5 due to their importance for subsequent chapters.

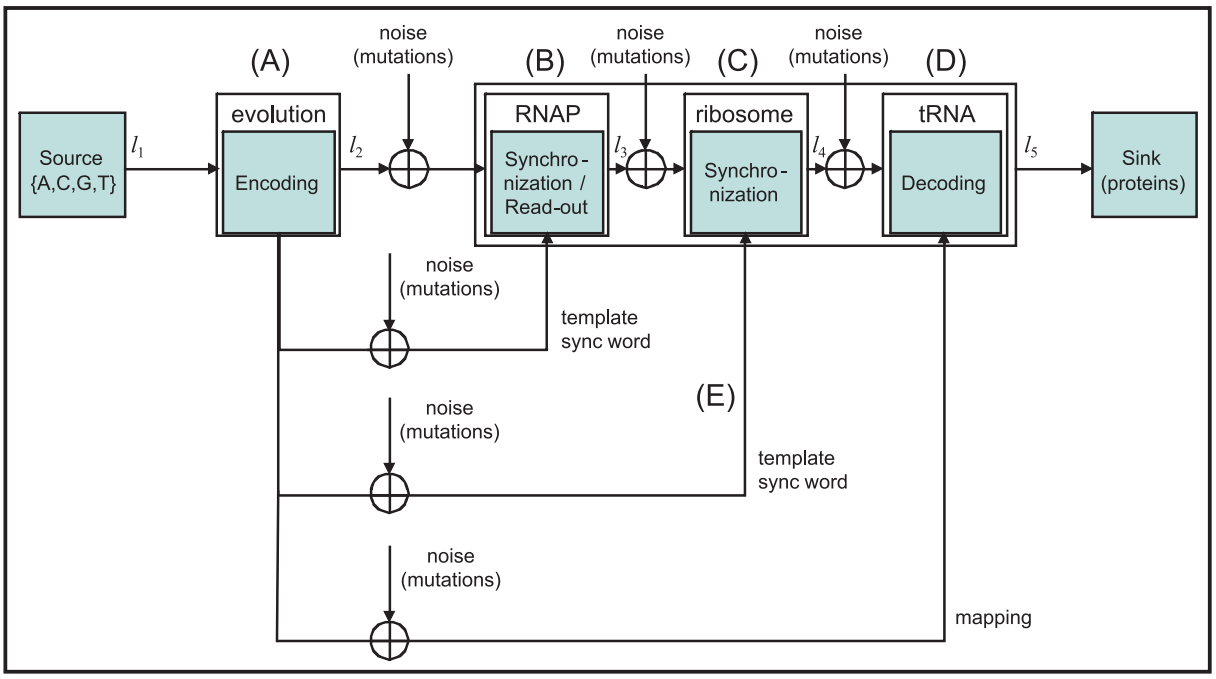


Figure 3.12: Channel model for the gene expression of prokaryotes.

3.6.1 Non-protein-coding DNA

As mentioned in Section 3.3.5, eukaryotic genomes contain a high percentage of DNA that does not code for proteins and whose function is not yet understood. Taft et al. [TPM07] found a relationship between the amount of non-protein-coding DNA and eukaryotic complexity, which might suggest that the former was added during evolution to protect the vital genetic information. This line of thought was later taken up by G. Battail who tried to find theoretical evidence for error-correcting codes in eukaryotic genomes [Bat04,Bat06]. These theories are supported by the fact that the genomes of eukaryotes generally contain by far more redundancy ($> 90\%$ of the genome) than those of prokaryotes (only $< 5\%$ of the genome) which have such short reproduction cycles that errors in single cells have no great impact. Instead, prokaryotes aim at decreasing the energy cost of cell processes like DNA replication by keeping the DNA as short as possible. If this hypothesis holds true, the addition of non-coding DNA would correspond to an encoding process, during which redundancy is added to protect the data. However, no indication has yet been found for specific error-correcting codes embedded in the genetic information – except for a repetition code: many genes exist in numerous copies spread along the genome [Lew07]. In 2005, Lolle et al. [LVYP05] presented a first experimental support for the existence of error-correction capabilities of eukaryotic cells: They inserted a point mutation in a single gene of the plant *Arabidopsis thaliana* and observed that the mutation was reversed within

two generations in about 10 % of the offsprings. In the following, evolution is therefore modeled as a process of encoding, during which extra information is added ($l_2 > l_1$) to protect the genetic information (see Figure 3.12, (A)).

3.6.2 Transcription

Transcription starts as soon as the RNA polymerase has detected the promoter shortly before the gene. Therefore, transcription initiation can be seen as a synchronization process, involving the detection of the two promoter regions (see Figure 3.13) and subsequent extraction of the genetic information ($l_3 < l_2$, see Figure 3.12, (B)). The RNA polymerase and its sigma factor correspond to the receiver and its synchronization unit. Since the distance between any two promoters varies, transcription corresponds to an asynchronous transmission (see Section 2.2.2).

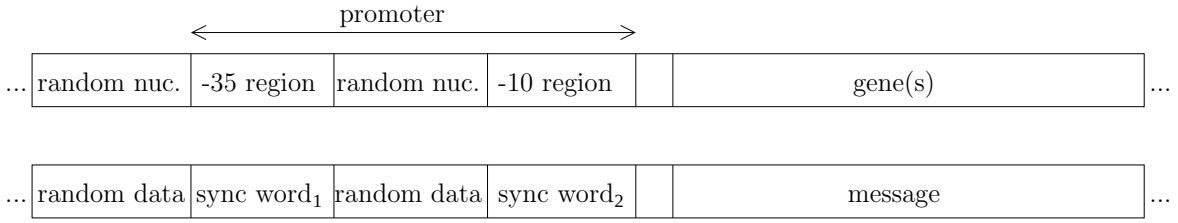


Figure 3.13: Analogy between promoter regions and sync words.

3.6.3 Translation

Similar to transcription, translation is initiated when the ribosome has detected the Shine-Dalgarno sequence, the sync word of translation ($l_4 < l_3$, see Figure 3.12, (C)). The mapping of codons to amino acids corresponds to a process of decoding, during which redundancy is removed by mapping three nucleotides to one amino acid ($l_5 < l_4$, see Figure 3.12, (D)). Since mRNAs generally contain only one Shine-Dalgarno sequence (even those with more than one coding sequence), translation can be considered as a synchronous transmission where data frames contain exactly one sync word (see Section 2.2.1).

3.6.4 Mutations

The DNA, though protected by the cell membrane, encounters different types of radiations that may lead to mutations and damage the genetic material. Furthermore, errors and mutations may occur during transcription and translation as well as during the short lifetime of the mRNA. In addition to these errors mutating the genetic information that is to be synthesized, mutations may also occur in the genes encoding for the proteins involved as well as the proteins themselves, i.e. the RNA polymerase for transcription, the ribosome and the tRNA for translation. These mutations may result in the failure

to properly detect synchronization sequences or to correctly map codons to amino acids. Thus, they can be modeled as errors that occur during the transmission of the template synchronization word or the mapping from the encoder to the receiver (see Figure 3.12, (E)). Hence, synchronization and decoding would be performed using an erroneous template synchronization word and wrong mapping, respectively.

3.6.5 Protein-DNA interactions

Interactions between proteins and the DNA constitute the crucial first step of important cell processes, as for example the interaction between the RNA polymerase and the DNA initiates transcription (see Section 3.4). During protein-DNA interactions, a protein binds to the DNA double-helix and searches for its cognate site (see Section 3.5.4). This can be compared to the receiver in technical systems which evaluates the incoming data stream symbol by symbol. However, while the likelihood function $L(\mu)$ in frame synchronization is usually defined based on the cross-correlation function between the sync word and the data stream (see Section 2.2), in biological synchronization processes it is based on the binding energy between protein and DNA [PG02]. The binding region of the protein that is in contact with the DNA is highly specific due to the geometry of possible bonds to the nucleotides (see Section 3.5). Therefore, one certain sequence – the consensus sequence – is bound tightly, while strong variations of this sequence may not allow the formation of hydrogen bonds with the protein. The binding region can therefore be considered as the template sync word used for comparison with the data stream. If the protein encounters its target site, it is strongly attached to it based on the concordant pattern of bonds and hereby halted in its movement along the DNA. This enables the initiation of its regulatory process. The analogies are summarized in Table 3.3.

Table 3.3: Comparison between synchronization in communication systems and during protein-DNA interactions.

	communication systems	**protein-DNA interactions**
data	received data stream	DNA
alphabet	arbitrary (mostly binary)	quaternary
template sync word	stored in the memory of the receiver	binding region of the protein
correlator	receiver	protein
L(μ)	cross-correlation	binding energy

As mentioned in Section 3.4, not only the optimal target site sequence is detected (the consensus sequence) but many variations of it. As mentioned in Section 3.4.2, it is ob-

tained by using the most frequently observed nucleotide at each position of the target site. Therefore, the target sites of protein-DNA interactions can be considered as a "soft sync word" whose homology to the consensus determines the rate of detection (see also Section 3.4.2). It should be mentioned that the analogies derived in this section apply in the same way to interactions between proteins and the mRNA as in the case of translation initiation where the ribosome binds to the mRNA initiator region.

3.7 Summary

This chapter aimed at providing the basics of molecular biology. After necessary terms and definitions, the process of gene expression (protein synthesis) was detailed for bacteria (prokaryotes) and higher organisms (eukaryotes). Subsequently, interactions between proteins and the DNA were elaborated due to their importance for later synchronization models. Furthermore, the analogies between gene expression and digital data transmission were derived. The following points should be taken along to subsequent chapters:

- ▷ Organisms are divided into prokaryotes and eukaryotes. The former comprise the mostly unicellular organisms without a cell nucleus (especially bacteria), the latter refer to all higher organisms, which have a cell nucleus and are usually multi-cellular.
- ▷ The DNA is organized in a double-helix and can be considered as a digital signal of the alphabet $\mathcal{A} = \{A, C, G, T\}$. The entire hereditary information encoded in the DNA is referred to as the genome.
- ▷ During gene expression, parts of the genome – the genes – are transformed into proteins in the two steps transcription and translation. During the former, the genes are copied into a template molecule, the mRNA (messenger RNA). During the latter, a part of the mRNA (the coding sequence) is then transformed into a protein.
- ▷ Short DNA motifs mark the beginning of the gene and the coding sequence: For transcription, the promoter sequence is located shortly before the gene. For translation, the Shine-Dalgarno sequence (prokaryotes) and the Kozak sequence (eukaryotes), respectively, are located shortly before the coding sequence. These are detected by the proteins RNA polymerase (transcription) and the ribosome (translation) to initiate the respective process.
- ▷ It could be demonstrated for the RNA polymerase and is assumed to be the case for all DNA-binding proteins that they attach to the double-helix and slide along it until they find the short DNA motif before the regulatory regions (e.g. the genes).
- ▷ Substantial analogies exist between gene expression and communication systems, especially between the protein-DNA / protein-mRNA interactions underlying transcription and translation on the one hand and frame synchronization in digital data transmission on the other hand.

4

Analysis of Biological Synchronization Words in Bacteria

As outlined in Chapter 2, the receiver in digital data transmission evaluates each position of the incoming data stream with respect to the similarity to the sync word that indicates the beginning of a message. Analogously, DNA-binding proteins randomly bind to the double-helix and move along it to find their target site, a short sequence of nucleotides. This short DNA motif marks the beginning of a regulatory region (e.g. the gene in case of transcription or the coding sequence in case of translation) (see Chapter 3). In technical systems, the sync pattern is chosen from all possible patterns such that the probability of false synchronizations is minimized. In this chapter, two sync patterns underlying protein-DNA / protein-mRNA interactions are investigated with respect to their synchronization properties: the bacterial promoter (the sync word of transcription) and the Shine-Dalgarno sequence (the sync word of bacterial translation).

In Section 4.1, the promoter sequences (the -35 region and the -10 region) of the bacterium *Escherichia coli* are investigated. Their synchronization properties are evaluated using an adapted autocorrelation function and a Markov analysis of the genome. Moreover, the promoter is modeled as a distributed synchronization sequence, where the spacing between the -35 promoter region and the -10 promoter region corresponds to unconstrained nucleotides not used for synchronization purposes. In Section 4.2, the prokaryotic Shine-Dalgarno sequence is investigated using the information theoretic measures Kullback-Leibler divergence and mutual information. The results are detailed in terms of their impact on translational frameshifts resulting from shifted synchronizations.

4.1 Promoter in *Escherichia coli*

In this section, the synchronization properties of the promoter sequences in the bacterium *Escherichia coli* (*E. coli*) are investigated. The promoters act as synchronization patterns that need to be detected by the RNA polymerase to indicate the beginning of the gene, i.e. that part of the DNA that is copied in the process of transcription. In the DNA of prokaryotes, these promoters consist of two regions of six nucleotides each (so-called hexamers): The first one is located around 35 base pairs before the start site, the second one around 10 base pairs before the start site (see Figure 4.1). Due to their position, the two regions are called -35 region and -10 region (or also Pribnow-box), respectively. Position +1 refers to the transcription start site (TSS).

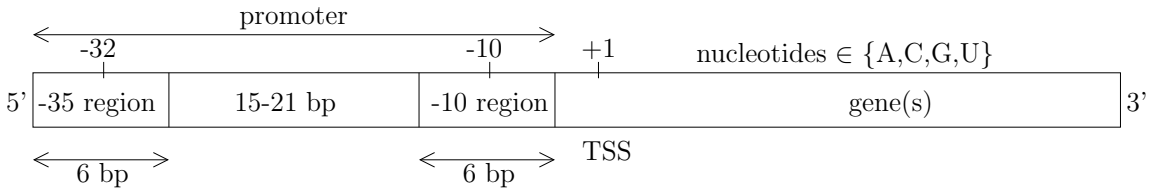
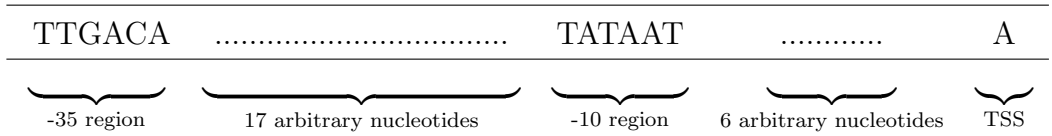


Figure 4.1: Structure of the promoter region in prokaryotes.

The consensus sequence (i.e. the optimal sequence) for the detection by the main sigma factor σ^{70} is given by [Lew07]:



In addition to this consensus sequence, many variations of it occur in the genome and are successfully detected by the RNA polymerase. However, the homology to the consensus decides about the frequency of detection: sequences far from the consensus result in a weaker binding energy which does not always suffice to halt the movement of the RNA polymerase along the DNA (see Section 3.4.2).

4.1.1 Autocorrelation properties

One measure to rate the synchronization properties of a sequence $\mathbf{s}$ is its aperiodic autocorrelation function (see Section 2.4.2):

$$\varphi_{ss}(\tau) = \sum_{k=1}^{L-|\tau|} s_k \cdot s_{k+|\tau|}. \tag{4.1}$$

For reasons of clarifying the application to promoter sequences in subsequent sections, Eq. (4.1) can be rewritten in a generalized form

$$\varphi_{ss}(\tau) = \sum_{k=1}^{L-|\tau|} \mathbf{D}(s_k, s_{k+|\tau|}), \tag{4.2}$$

where $\mathbf{D}$ denotes a matrix defining the multiplication of the elements s_k and $s_{k+|\tau|}$ with s_k indexing the rows of $\mathbf{D}$ and $s_{k+|\tau|}$ indexing the columns of $\mathbf{D}$. In the binary, antipodal case, i.e. for $s_k \in \{-1; +1\}$, this results in

$$\begin{array}{lll} s_{k+|\tau|} \rightarrow & +1 \quad -1 & s_k \downarrow \\ \mathbf{D}_{\text{bin}} = & \begin{pmatrix} +1 & -1 \\ -1 & +1 \end{pmatrix} & \begin{matrix} +1 \\ -1 \end{matrix} \end{array}$$

In order to avoid shifted synchronizations, the autocorrelation function of the sync word should have a sharp peak at $\tau = 0$ and

- ▷ smallest possible values for $\tau \neq 0$ if unambiguous phase recovery after demodulation is guaranteed.
- ▷ smallest possible absolute values for $\tau \neq 0$ if phase ambiguities are expected after demodulation (e.g. for BPSK modulated data streams), i.e. the autocorrelation function should be as similar as possible to the Dirac delta function $\delta(t)$ [Lev75].

As introduced in Section 2.4.2, the peak sidelobe level PSL is a measure of the synchronization properties of a sequence:

$$\text{PSL} = \max_{\tau \setminus \{0\}} |\varphi_{ss}(\tau)|. \tag{4.3}$$

If correct phase recovery is guaranteed, the absolute values in Eq. (4.3) are omitted since negative values indicate strong dissimilarity and therefore minimize the probability of false synchronizations:

$$\text{PSL'} = \max_{\tau \setminus \{0\}} \varphi_{ss}(\tau). \tag{4.4}$$

4.1.2 Adapted autocorrelation function

To rate the synchronization properties of the promoter sequences, the autocorrelation function has to be adapted to the quaternary alphabet of nucleotides $n \in \mathcal{A} = \{A, C, G, T\}$ and the product in Eq. (4.1) has to be redefined with respect to its biological meaning. Apparently, nature does not use a multiplication as in the field of real or complex numbers,

thus, the product has to be redefined such that it rates the effect of nucleotide matches and mismatches on the synchronization quality of the sequence. In order not to violate the properties of aperiodic autocorrelation functions, which are

$$\varphi_{ss}(0) = L, \tag{4.5}$$
$$\varphi_{ss}(\tau) = 0 \quad \forall \quad |\tau| > (L-1), \tag{4.6}$$

an accordance of nucleotides is rated by $+1$ and a divergence of nucleotides with a negative value such that mismatches are punished with an overall weight of -1. As mentioned before, the binding energy decides about detection of the promoter regions (i.e. correct synchronization). This implies that if during autocorrelation shifted versions of the investigated DNA sequence yield low (i.e. strong) binding energies, they might cause shifted synchronizations. Thus, the adapted autocorrelation has to be related to the binding energy of the shifted sequences. Therefore, the individual values rating mismatches are derived from the binding energy between sigma factor and DNA. In [KOA05], H. Kiryu et al. calculated the effect of the nucleotides on the binding energy depending on their position in the promoter of *E. coli*. Figure 4.2 shows the average effect of the four nucleotides on the binding energy. It is important to note that negative energies reflect a strong binding, whereas positive energies imply a weak binding.

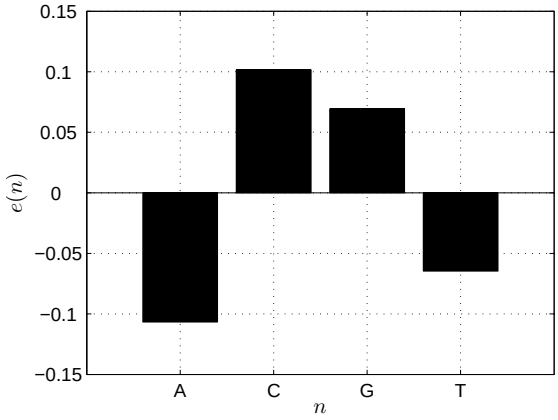


Figure 4.2: Average contribution of nucleotides in the promoter sequence to the binding energy between the sigma factor and the promoter in *E. coli*.

It can be seen that the nucleotides A and T in the promoter have on average a strengthening effect on the contact between sigma factor and DNA sequence ($e = -0.11$ and $e = -0.06$, respectively), whereas the nucleotides C ($e = +0.10$) and G ($e = +0.07$) make the contact loose. A mismatch during autocorrelation is rated by the absolute difference $|e(n_x) - e(n_y)|$ between the binding energies of the two nucleotides n_x and n_y, where $n_x, n_y \in \mathcal{A} = \{A, C, G, T\}$:

$$d(n_x, n_y) = \begin{cases} 1 & \text{for} \quad n_x = n_y \\ c \cdot |e(n_x) - e(n_y)| & \text{for} \quad n_x \neq n_y \end{cases} . \tag{4.7}$$

The constant c is still to be determined since in addition to reflecting the differences of binding energies, the values of $d(n_x, n_y)$ have to satisfy Eq. (4.6), i.e. the expected value $\mathbb{E}\{d(n_x, n_y)\}$ has to be zero if assuming independently and uniformly distributed nucleotides:

$$\mathbb{E}\{d(n_x, n_y)\} = 0, \tag{4.8}$$

which corresponds to

$$\begin{aligned} \sum_{n_x, n_y} d(n_x, n_y) &= 0, \\ \Rightarrow \quad \underbrace{\sum_{\substack{n_x, n_y \\ n_x = n_y}} d(n_x, n_y)}_{\overset{(4.7)}{=} 4} + \sum_{\substack{n_x, n_y \\ n_x \neq n_y}} d(n_x, n_y) &= 0, \\ \Rightarrow \quad \sum_{\substack{n_x, n_y \\ n_x = n_y}} d(n_x, n_y) \overset{!}{=} - \sum_{\substack{n_x, n_y \\ n_x \neq n_y}} d(n_x, n_y) &= -4. \end{aligned} \tag{4.9}$$

Eq. (4.9) is fulfilled if scaling the individual energy differences in Eq. (4.7) by the value

$$c = \frac{-4}{\sum\limits_{\substack{n_x, n_y \\ n_x \neq n_y}} |e(n_x) - e(n_y)|} = \frac{-4}{1.56} = -2.56. \tag{4.10}$$

In order to adapt the autocorrelation function to the quaternary alphabet of nucleotides and detection by the RNA polymerase and its sigma factor, Eq. (4.2) is used with a matrix $\mathbf{D}_{\text{nuc}}$ containing the values of $d(n_x, n_y)$, i.e. $\mathbf{D}_{\text{nuc}}(s_k, s_{k+|\tau|}) = d(n_x = s_k, n_y = s_{k+|\tau|})$, which results for the presented case of *E. coli* promoters in

$$\mathbf{D}_{\text{nuc}} = \begin{array}{c} s_{k+|\tau|} \rightarrow \\ \\ \\ \\ \\ s_k \downarrow \end{array} \begin{array}{c} \begin{array}{cccc} \text{A} & \text{C} & \text{G} & \text{T} \end{array} \\ \begin{pmatrix} 1 & -0.55 & -0.46 & -0.11 \\ -0.55 & 1 & -0.08 & -0.44 \\ -0.46 & -0.08 & 1 & -0.36 \\ -0.11 & -0.44 & -0.36 & 1 \end{pmatrix} \end{array} \begin{array}{c} \\ \\ \\ \\ \\ \begin{array}{c} \text{A} \\ \text{C} \\ \text{G} \\ \text{T} \end{array} \end{array}$$

Therefore, the autocorrelation function $\widetilde{\varphi}_{ss}(\tau)$ of *E. coli* promoter sequences is given by

$$\widetilde{\varphi}_{ss}(\tau) = \sum_{k=1}^{L-|\tau|} \mathbf{D}_{\text{nuc}}(s_k, s_{k+|\tau|}). \tag{4.11}$$

This adapted autocorrelation function allows to evaluate the synchronization properties of promoter sequences. It considers a linear algebraic addition of binding energies, which

is reported in biological literature to be a valid assumption (see e.g. [DSS03, SF98]). It has to be mentioned that the matrix values of $\mathbf{D}_{\text{nuc}}$ are calculated based on the data from [KOA05], i.e. the adapted autocorrelation is based on the interaction between sigma factor and promoter regions in *E. coli* and is, therefore, specific for this biological synchronization process.

4.1.3 Results

Figure 4.3 shows the autocorrelation functions of the consensus sequences of the -35 region (left) and the -10 region (right) calculated using Eq. (4.11). As mentioned before, the autocorrelation function of sync words should have small and possibly negative sidelobes to minimize the probability of false synchronizations. This criteria seems to be well satisfied for the -35 region (Figure 4.3, left), whereas the autocorrelation function of the -10 region (Figure 4.3, right) has relatively high sidelobes for $|\tau| = 2$ and $|\tau| = 3$, which indicates periodicities in the sync word that may lead to shifted synchronizations. Calculation of the peak sidelobe level for both promoter regions according to Eq. (4.4) confirms this observation:

$$\text{PSL'}_{-35} = \widetilde{\varphi}_{-35}(|\tau| = 2) = 0.45, \quad \text{PSL'}_{-10} = \widetilde{\varphi}_{-10}(|\tau| = 3) = 1.89.$$

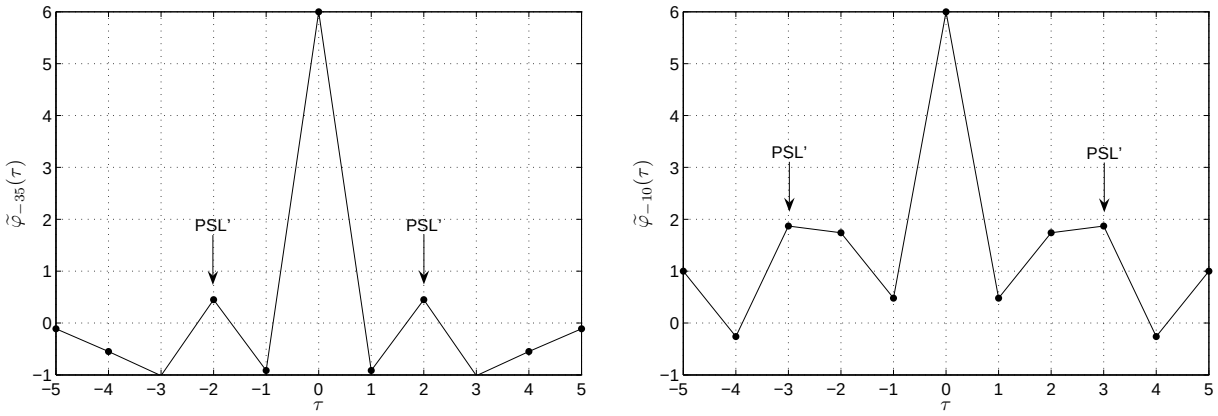


Figure 4.3: Autocorrelation function of the consensus sequences of the -35 promoter region (left, TTGACA, $L = 6$) and the -10 promoter region (right, TATAAT, $L = 6$).

To rate the autocorrelation properties of the promoter sequences, the values of PSL' are additionally calculated for all $4^6 = 4096$ possible nucleotide sequences of length $L = 6$. The mean value and the standard deviation of the resulting values are listed in Table 4.1. Figure 4.4 shows the histogram of PSL' with the values of the -35 and -10 region highlighted by vertical lines. It can be seen that the value of the -35 promoter sequence is well below average, whereas those of the -10 promoter sequence lies above the mean value. In fact, only 11.1 % of all possible sequences of length $L = 6$ have better autocorrelation properties with respect to the peak sidelobe level than the -35 region. Opposed to that, 75.4 % of all sequences have lower values of PSL' compared to the -10 region.

Table 4.1: Mean and standard deviation of PSL' for all possible sequences of $\mathcal{A} = \{A, C, G, T\}$ with length $L = 6$.

	mean	**std. deviation**
PSL'	1.32	0.77

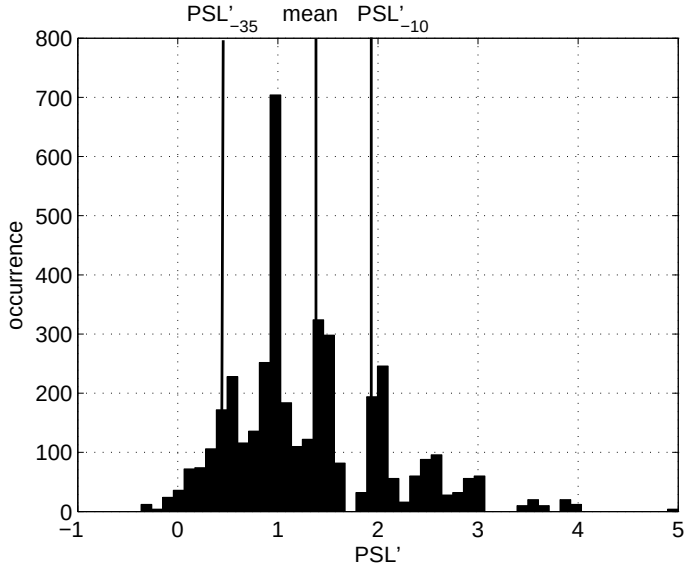


Figure 4.4: Histogram of the peak sidelobe level PSL' for all possible sequences of $\mathcal{A} = \{A, C, G, T\}$ with length $L = 6$.

4.1.4 Interpretation

The outstanding PSL' value of the -35 region compared to those of the -10 region suggests that the synchronization takes place in two steps: First, the -35 region has to be detected out of all possible sequences with high accuracy to enable localization of the transcription start site (see Figure 4.5, (A)). In the second step, the -10 region is detected, however, due to the synchronization conducted before, the sigma factor only needs to detect the -10 region out of around seven sequences based on the shape and limited deformability of the sigma factor that allow a variable spacing of 15 to 21 base pairs between the two promoter regions (see Figure 4.5, (B)). Therefore, the sequence of the -10 promoter region is less important for synchronization. This brings up the conclusion that the two promoters evolved to serve two tasks with different priorities and during different steps of transcription initiation: While the -35 region is indispensable for indicating the close-by transcription start site and, thus, needs to have excellent synchronization properties, the sequence and structure of the -10 region seems to play a more important role during later steps of transcription initiation. These steps may e.g. impose stronger constraints on the AT-richness (i.e. a high content of the nucleotides A and T) than on the sequence's

detectability: The DNA double helix is easily opened and unwound in AT-rich regions which is necessary during transcription initiation [SY99]. Therefore, the AT-richness of the -10 region is assumed to have had a stronger impact on promoter evolution than its synchronization properties and, thus, the latter evolved with lower priority.

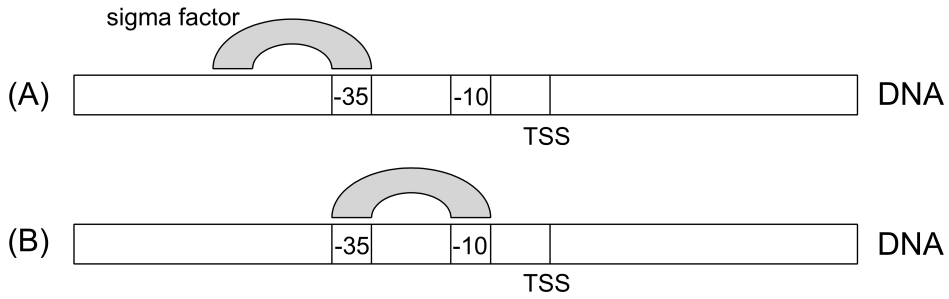


Figure 4.5: Detection of promoters by the sigma subunit.

To corroborate the assumption about the stronger importance of the AT-richness compared to the synchronization properties, the PSL' values for all $2^6 = 64$ possible nucleotide sequences of length $L = 6$ made up of only A and T are calculated. The mean value and the standard deviation of the resulting values are listed in Table 4.2. Figure 4.6 shows the histogram of PSL' for the considered sequences. Recalling the calculated value of the -10 region (PSL' = 1.89) shows clearly that it belongs to the sequences with highly below-average values if restricting the alphabet to $\mathcal{A}' = \{A, T\}$. In fact, no other sequence of the 64 ones considered has a better PSL' value. This result strongly supports the conclusion that the bacterial promoter sequences evolved with respect to their synchronization properties: While the -35 region is an excellent synchronization pattern, the -10 region seems to constitute a good trade-off between the AT-richness required for DNA opening / unwinding and the sequence's detectability.

Table 4.2: Mean and standard deviation of PSL' for all possible sequences of $\mathcal{A}' = \{A, T\}$ with length $L = 6$.

	mean	**std. deviation**
PSL'	2.89	0.76

4.1.5 The promoter as a distributed synchronization sequence

As detailed in Section 2.5.4, a distributed synchronization sequence is a sync word containing unconstrained bits, i.e. the synchronization bits are interspersed with data bits (denoted by *). The receiver knows the unconstrained positions a priori and thus ignores them for synchronization. Due to the 15 to 21 arbitrary nucleotides between the -35 and

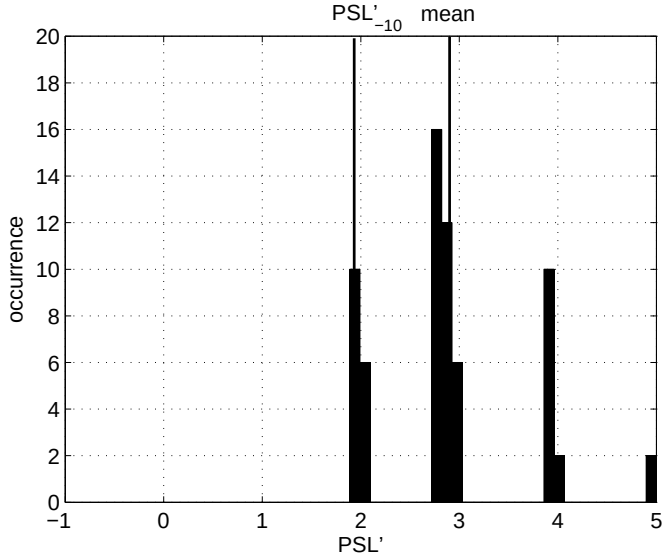


Figure 4.6: Histogram of the peak sidelobe level SPL' for all possible sequences of $\mathcal{A}' = \{A, T\}$ with length $L = 6$.

the -10 region, the promoter of *E. coli* can be considered as a distributed sync word with 12 synchronization bits separated midway by a sequence of 17 unconstrained bits. To rate the synchronization properties of this pattern, the adapted autocorrelation function from Section 4.1.1 is extended by the following definition:

$$\mathbf{D}(s_k, *) = \mathbf{D}(*, s_{k+|\tau|}) = \mathbf{D}(*, *) = 0.$$

Including this definition into the derivation of the adapted autocorrelation function (Eq. (4.11)) yields an extended matrix $\mathbf{D'}_{\text{nuc}}$:

$$\mathbf{D'}_{\text{nuc}} = \begin{array}{c} s_{k+|\tau|} \rightarrow \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccc} \text{A} & \text{C} & \text{G} & \text{T} & * \end{array} \\ \begin{pmatrix} 1 & -0.55 & -0.46 & -0.11 & 0 \\ -0.55 & 1 & -0.08 & -0.44 & 0 \\ -0.46 & -0.08 & 1 & -0.36 & 0 \\ -0.11 & -0.44 & -0.36 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array} \begin{array}{c} s_k \downarrow \\ \text{A} \\ \text{C} \\ \text{G} \\ \text{T} \\ * \end{array}$$

Figure 4.7 shows the autocorrelation function of the promoter consensus sequence as a distributed sequence. The PSL'-value is $\varphi_{ss}(|\tau| = 2) = 2.23$.

The left-most and right-most part represent the cross-correlation between the -35 and the -10 region, while the middle part stems from the autocorrelation of the -35 region as well as the -10 region. To rate the quality of the autocorrelation properties of the promoter sequence, the PSL'-values of all possible distributed sequences made up of two hexamers

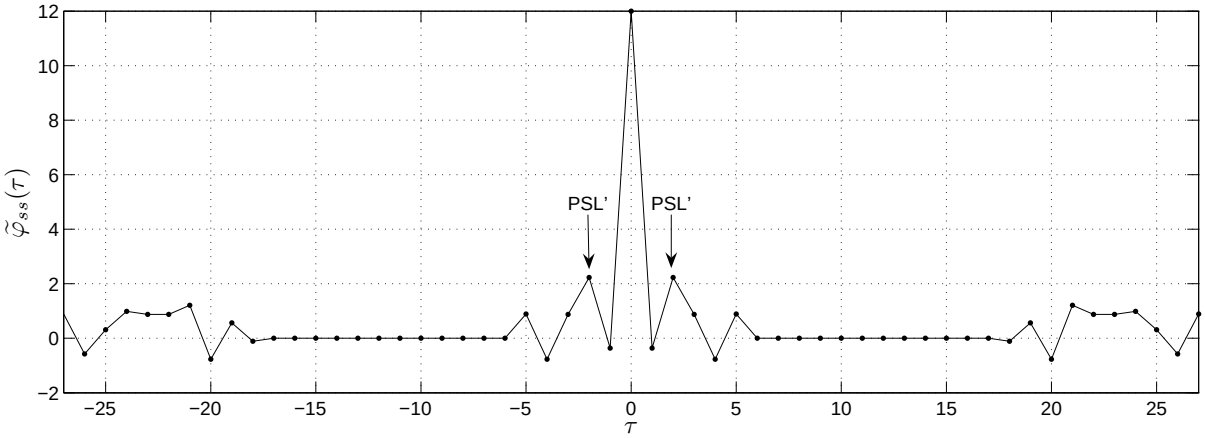


Figure 4.7: Autocorrelation function of the promoter sequence as a distributed sequence.

separated by 17 unconstrained bits are calculated. Due to the symmetry of the cross-correlation function ($\varphi_{xy}(\tau) = \varphi_{yx}(\tau)$), only $2^{23} + 2048$ of the 4^{12} possible sequences from the alphabet $\mathcal{A} = \{A, C, G, T\}$ yield different values. This analysis reveals that 49.6 % of the sequences exhibit a lower PSL'-value compared to the promoter (PSL'=2.23). If again restricting the -10 promoter region to the alphabet $\mathcal{A}' = \{A, T\}$, still 33.0 % of the sequences have a lower PSL'-value. This poor performance of the promoter as a distributed sequence strengthen the conclusion drawn in Section 4.1.4 that the two promoter regions act as separate synchronization sequences with the -35 sequence being the more important detection signal and the -10 region primarily serving DNA opening.

4.1.6 Markov analysis

As detailed in Section 2.4, the sync word should be chosen such that it minimizes the probability of shifted synchronization as well as of random occurrences. The latter is independent of the sync word in case of i. i. d. symbols but needs to be taken into account if the data stream exhibits statistical dependencies. In that case, the sync word should be a pattern that occurs with the smallest possible probability by chance. If the sync word has to satisfy additional constraints that preclude it being chosen as the most unlikely sequence, it should instead be avoided in the surrounding data stream, i.e. it should be an under-represented word – occurring exceptionally rare – with respect to the Markov model Mm of the data stream. The exceptionality of a pattern $\mathbf{r}$ depends on the relation between expected and observed occurrences in the data stream. The expected number $\mathbb{E}\{\widehat{N}_m(\mathbf{r})\}$ of occurrences of a pattern $\mathbf{r}$ depending on Mm is given by (see Section 2.4.1)

$$\mathbb{E}\{\widehat{N}_m(\mathbf{r})\} = \frac{N(\{r_1, \ldots, r_{m+1}\}) \cdot \ldots \cdot N(\{r_{L-m}, \ldots, r_L\})}{N(\{r_2, \ldots, r_{m+1}\}) \cdot \ldots \cdot N(\{r_{L-m}, \ldots, r_{L-1}\})} = \frac{\prod_{x=1}^{L-m} N(\{r_x, \ldots, r_{m+x}\})}{\prod_{x=2}^{L-m} N(\{r_x, \ldots, r_{x+m-1}\})}, \tag{4.12}$$

the observed number of occurrences of the word $\mathbf{r}$ is simply given by

$$N(\mathbf{r}) = \sum_{\mu=1}^{N_d-L+1} \mathbf{1}(\{d_\mu, \ldots, d_{\mu+L-1}\} = \{r_1, \ldots, r_L\}), \tag{4.13}$$

where N_d refers to the length of the analyzed sequence $\{d_1, \ldots, d_{N_d}\}$ (here: the *E. coli* genome).

Exceptionality score

The exceptionality of words in the data stream is measured using the following probability, called p-value [RRS05]:

$$p(\mathbf{r}) = \Pr\{\widehat{N}_m(\mathbf{r}) \geq N(\mathbf{r})\}, \tag{4.14}$$

If the p-value is close to zero, the word is exceptionally frequent since there is almost no chance of observing it so many times in random sequences. In contrast to that, if the p-value is close to one, the probability $\Pr\{\widehat{N}_m(\mathbf{r}) < N(\mathbf{r})\}$ is close to zero. Thus, the word is exceptionally rare under the model since there is almost no chance that it occurs so rarely in random sequences [RRS05].

For calculation of the p-values, the statistical distribution of the count $\widehat{N}_m(\mathbf{r})$ is required. Since it is computationally extensive to derive the exact distribution – especially for long sequences and orders $m \geq 2$ – two approximations are frequently used: Gaussian or compound Poisson distribution. The former yields accurate results for short word lengths L, the latter is applicable for long words. Both approximations were shown to be highly accurate for sequence lengths $N_d \geq 10000$ [RRS05].

The software R'MES – aimed at finding exceptional sequence motifs in DNA sequences – is used to evaluate the representation of sequences in the *E. coli* genome (available at [SfSB07], see [RRS05, RSV07] for more information). It calculates an exceptionality score derived from the p-value: For reasons of better resolution of very low and very high values, the interval $p \in [0; 1]$ is mapped to $\mathbb{R}$ such that positive values indicate exceptionally frequent words and negative values indicate exceptionally rare words. To ensure accurate results, the complete sequence length N_d should not be lower than $3000 \cdot 4^m$ [Sch06], and the order of the Markov model should be chosen as $m = L - 2$ (see e.g. [RRS05]).

Promoter analysis

To gain a more detailed insight into the representation of the promoter sequences, the genome is divided into two categories: promoter regions and non-promoter regions. The former are taken as the region ranging from 300 bp before the TSS to 200 bp after the TSS

of known promoters downloaded from RegulonDB [SGCPG+06] (see Appendix C.1.1 for more information on the used dataset). The non-promoter regions simply comprise the rest of the genome, i.e. those regions that do not fall into the promoter regions of either strand. Subsequently, a Markov analysis is conducted on the two categories as presented above using the software R'MES. The Markov model is derived from the whole genome for the maximum order $m = 4$ (since $L = 6$ for the promoter regions), and the Gaussian approximation is used to derive the exact distribution of $\widehat{N}_m(\mathbf{r})$ (since the promoters constitute short words). The order of the Markov model underlying the DNA is actually unknown, however, previous investigations of the *E. coli* genome (conducted until order $m = 6$) showed that models of low orders overlooked exceptional motifs [RRS05].

Results and interpretation

The resulting scores for the occurrence of the two promoter sequences on the forward strand are listed in Table 4.3.

Table 4.3: Scores of the -35 promoter region (TTGACA) and the -10 region (TATAAT) on the forward strand of the *E. coli* genome.

	promoter regions		**non-promoter regions**	
sequence	TTGACA	TATAAT	TTGACA	TATAAT
observed count	26	32	507	473
expected count	30	36	624	705
p-score	-1.07	-0.80	-5.74	-10.85

It can be seen that both promoter sequences are occurring almost as often as expected in the promoter regions (scores close to zero). In contrast to that, both are clearly under-represented in the non-promoter regions (negative scores), which indicates that they were evolutionary avoided in those regions. For reasons of comparison, Figure 4.8 shows the histogram of scores in the promoter regions (left) and in the non-promoter regions (right) for all 4096 possible sequences of length $L = 6$. The under-representation of both promoter sequences is in accordance with the objective of frame synchronization to avoid the sync pattern outside the header of transmitted messages to prevent synchronizations on random data. Among all 4096 possible sequences, only 5.9 % have lower scores than the -35 region, however, only 15 of these sequences (0.36 %) also have better synchronization properties (regarding the PSL'-value). In case of the -10 region, none of the sequences with a lower score and made up of only A and T has better synchronization properties. This fact again indicates that the promoter sequences evolved with respect to their synchronization properties. Moreover, it strengthens the conclusion from Section 4.1.5 that both promoter regions act as separate synchronization signals.

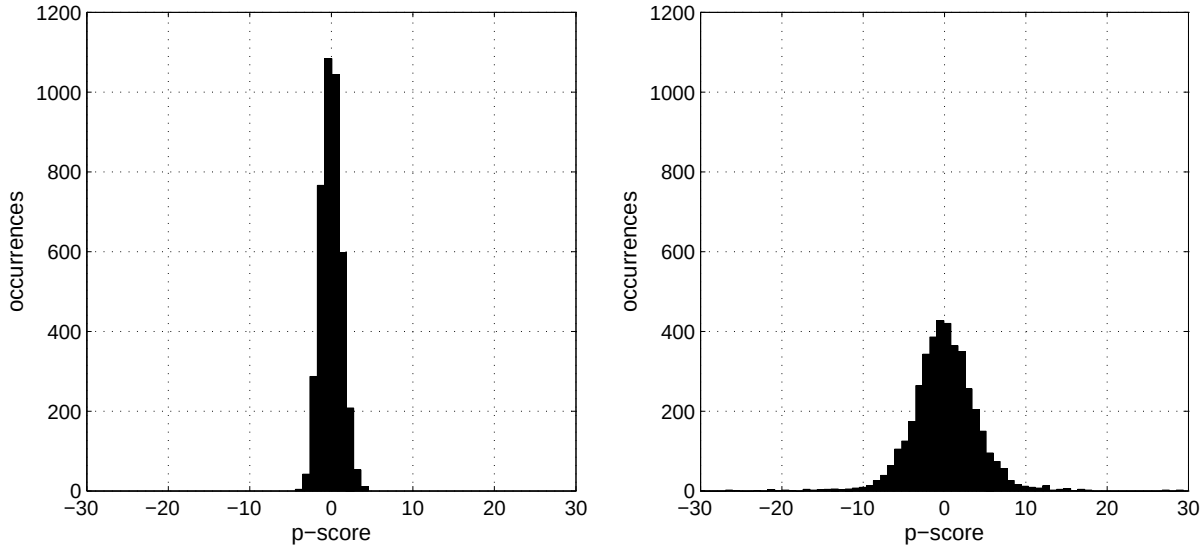


Figure 4.8: Histogram of p-scores in the promoter regions (left) and in the non-promoter regions (right).

4.2 Translation initiator region in *Escherichia coli*

After the RNA polymerase has copied a gene into mRNA in the process of transcription, the ribosome initiates translation of the mRNA into a protein (see Section 3.4.4). After binding to the 5'-UTR, the 30S ribosomal subunit moves rapidly along the mRNA until it detects the start codon (AUG, position +1) and the Shine-Dalgarno sequence (SD), a hexamer located shortly before the coding sequence (see Figure 7.1).

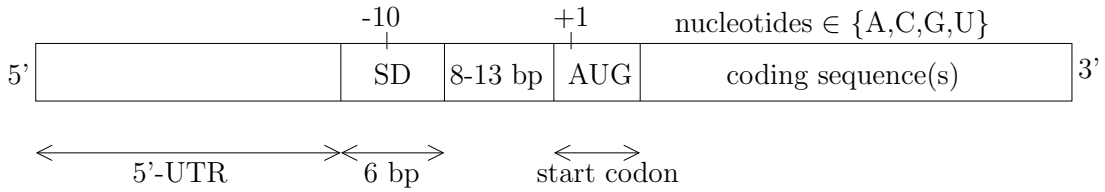
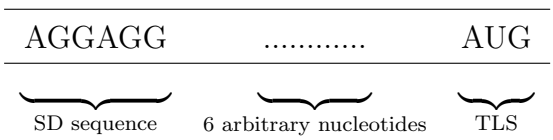


Figure 4.9: Structure of the initiator region of prokaryotic mRNA.

The Shine-Dalgarno (SD) sequence hereby acts as the sync word to ensure reliable detection of the close-by start codon. Its consensus sequence (i.e. its optimal sequence for detection by the ribosome) is given by [CBKJ94]:



Similarly to the promoter sequences, numerous variations of this consensus sequence are also detected by the ribosome. However, the homology to the consensus decides about the frequency of detection. In the following, the initiator region and the coding sequence of *Escherichia coli* are investigated using information theoretic measures.

4.2.1 Sequence data

A set $\mathcal{S}$ of 3194 *E. coli* mRNA sequences is downloaded from the NCBI data base [NCfBI08] (see Appendix C.1.3 for more information on sequence extraction). Thereafter, the mRNA sequences are aligned (centered) to the start as well as to the stop codon. Since the coding sequences (CDS) and the untranslated regions (UTRs) do not all have the same length, this alignment implies the need to cut them to a fixed length: the UTRs are truncated to 200 bp each and the middle part of the coding sequence is cut out leaving the first and the last 300 bp. The sequence layout is presented in Figure 4.10.

5' UTR	start codon	CDS	CDS (ctd.)	stop codon	3' UTR
...	AUG	...	...	UAA	...
...	AUG	...	...	UAG	...
⋮		⋮	⋮		⋮
...	AUG	...	...	UGA	...
...	AUG	...	...	UAG	...
200 bp		300 bp	300 bp		200bp

Figure 4.10: Sequence layout of aligned mRNA sequences.

4.2.2 Kullback-Leibler divergence

The relative entropy – or Kullback-Leibler divergence – is a measure for the dissimilarity of two probability distributions $p_X(x)$ and $q_X(x)$. It is defined as [CT91]

$$D(p_X \parallel q_X) = \sum_{x \in \mathcal{X}} p_X(x) \mathrm{ld} \frac{p_X(x)}{q_X(x)}. \tag{4.15}$$

An intuitive interpretation of $D(p_X \parallel q_X)$ based on Shannon's compression theorem is that it refers to the additional number of bits required for transmission of symbols that are distributed according to $p_X(x)$ if they are coded according to a wrongly estimated distribution $q_X(x)$. Note that $D(p_X \parallel q_X)$ is always non-negative (i.e. $D(p_X \parallel q_X) \geq 0$), in general not symmetric (i.e. $D(p_X \parallel q_X) \neq D(q_X \parallel p_X)$) and zero for $p_X(x) = q_X(x)$.

Positions in the mRNA or DNA whose nucleotide distribution strongly differ from the background distribution (in prokaryotes usually close to a uniform distribution) are expected to be of functional significance to cell processes, since non-essential genetic features tend to be degraded over the course of evolution. To investigate the Shine-Dalgarno sequence, the actual distribution $p_X(x)$ is estimated by the nucleotide distribution at each position of the alignment $\mathcal{S}$ of mRNA sequences. The general nucleotide distribution observed in the genome is used as the assumed probability $q_X(x)$. Then, the Kullback-Leibler divergence $D(\hat{p}_\mathcal{S}(n,i) \parallel \hat{p}_\mathcal{S}(n))$ at position i of the dataset is calculated as

$$D(\hat{p}_\mathcal{S}(n,i) \parallel \hat{p}_\mathcal{S}(n)) = \sum_{n \in \mathcal{A}} \hat{p}_\mathcal{S}(n,i) \mathrm{ld} \frac{\hat{p}_\mathcal{S}(n,i)}{\hat{p}_\mathcal{S}(n)}, \tag{4.16}$$

where n denotes a nucleotide from the alphabet $\mathcal{A} = \{A, C, G, T\}$, $\hat{p}_\mathcal{S}(n,i)$ the relative occurrence of base n at position i of the aligned dataset $\mathcal{S}$ and $\hat{p}_\mathcal{S}(n)$ the overall occurrence frequency of nucleotide n in the dataset. It is important to note that $D(\hat{p}_\mathcal{S}(n,i) \parallel \hat{p}_\mathcal{S}(n))$ is only an estimate of the Kullback-Leibler divergence and that the accuracy of this estimate strongly depends on the size of the dataset.

For detection by the ribosome, the Shine-Dalgarno sequence can be seen as a first synchronization signal indicating the close-by start codon, which serves as a second synchronization signal. Figure 4.11 shows the Kullback-Leibler divergence around the Shine-Dalgarno sequence (position -13 to -7) and the start codon (position +1 to +3). Since the sequences are aligned to the position +1, the conservation of the start codon is higher (almost 2) than that of the Shine-Dalgarno sequence whose position varies between 5 and 13 bp before the start codon. The Kullback-Leibler divergence of the first base in the start codon is below those of the second and third base due to alternative start codons: GUG and UUG are reported to occur in 10 % of bacterial mRNAs [Koz99].

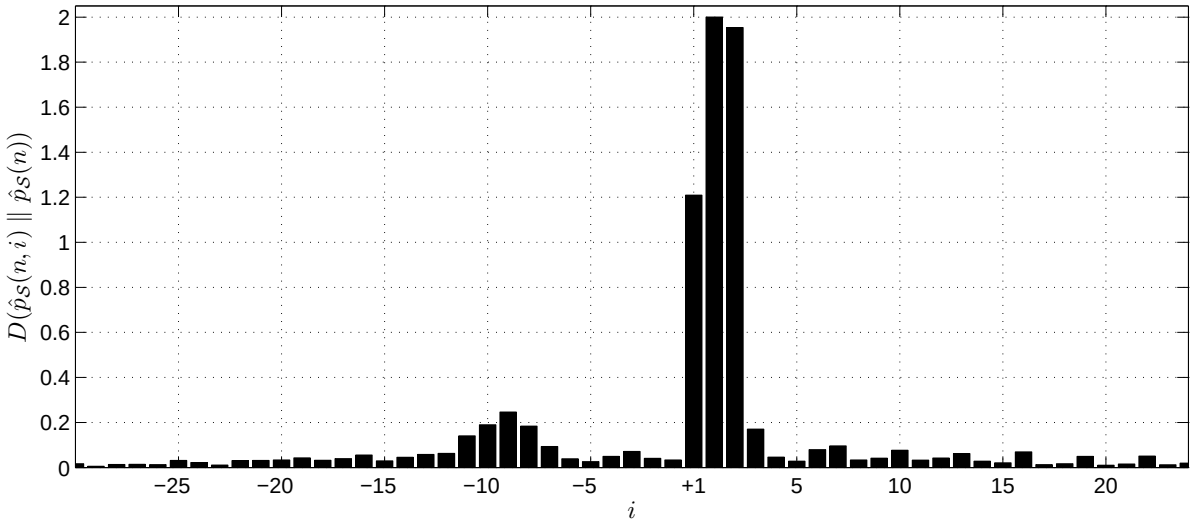


Figure 4.11: Kullback-Leibler divergence around the Shine-Dalgarno sequence (position -13 to -7) and the start codon (position +1 to +3).

4.2.3 Mutual information

The mutual information is a measure for the mutual statistical dependence between two random variables X and Y. It is defined as [CT91]

$$I(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p_{X,Y}(x,y)\mathrm{ld}\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}, \tag{4.17}$$

where $p_{X,Y}(x,y)$ denotes the joint probability mass function and $p_X(x)$ and $p_Y(y)$ the marginal probability mass functions of X and Y. Mutual information is the Kullback-Leibler divergence between the joint distribution and the product distribution $p_X(x)p_Y(y)$, thus, it becomes zero only in the case of independence between X and Y (i.e. for the case that $p_{X,Y}(x,y) = p_X(x)p_Y(y)$). Note that mutual information is always non-negative (i.e. $I(X;Y) \geq 0$) and symmetric (i.e. $I(X;Y) = I(Y;X)$).

An important aspect of biological sequence analysis is to investigate which positions in a binding site (e.g. the SD sequence) are mutually dependent. These are expected to have a conjoint functional significance [OST06]. The mutual information between two positions i_x and i_y is given by

$$I(i_x;i_y) = \sum_{n_x\in\mathcal{A}}\sum_{n_y\in\mathcal{A}} \hat{p}_{\mathcal{S}}(n_x,n_y,i_x,i_y)\mathrm{ld}\frac{\hat{p}_{\mathcal{S}}(n_x,n_y,i_x,i_y)}{\hat{p}_{\mathcal{S}}(n_x,i_x)\hat{p}_{\mathcal{S}}(n_y,i_y)}. \tag{4.18}$$

where the relative occurrence $\hat{p}_{\mathcal{S}}(n_x,n_y,i_x,i_y)$ is again calculated from the set $\mathcal{S}$ of aligned mRNA sequences and refers to the count of conjointly observing nucleotide n_x at position i_x and nucleotide n_y at position i_y.

In the first step, the region around the start codon is investigated in terms of nucleotide dependencies. Figure 4.12 depicts the mutual information $I(i;i+1)$ between neighboring nucleotides. It exhibits high values of dependence between the nucleotides in the Shine-Dalgarno sequence (position -13 to -8) and directly after the start codon (beginning at position +4). Contrary, the mutual information between neighboring bases in the start codon (positions +1 to +3) are nearly zero due to the strongly limited nucleotide variation. Y. Osada reported that the nucleotides at positions $i = -2$ and $i = -1$ are strongly correlated in several prokaryotes [OST06]. This seems not to be the case for *E. coli*, where $I(-2;-1)$ exhibits a value of around 0.03 which is below many other values.

To further investigate the observed dependence between distant positions beginning directly after the start codon (see Figure 4.12), the whole range of the dataset is analyzed in the next step. Figure 4.13 shows the values of the mutual information between positions with a distance $d = i_x - i_y$, $d \in [1;3]$ for the 500 base pairs around the start codon (left part of the plots) and for the 500 base pairs around the stop codon (right part of the plots). Interestingly, the mutual information between neighboring nucleotides (top) exhibits constantly high values throughout the coding sequence. This fact indicates strong dependencies in that region that are likely to stem from the codon bias: First, several

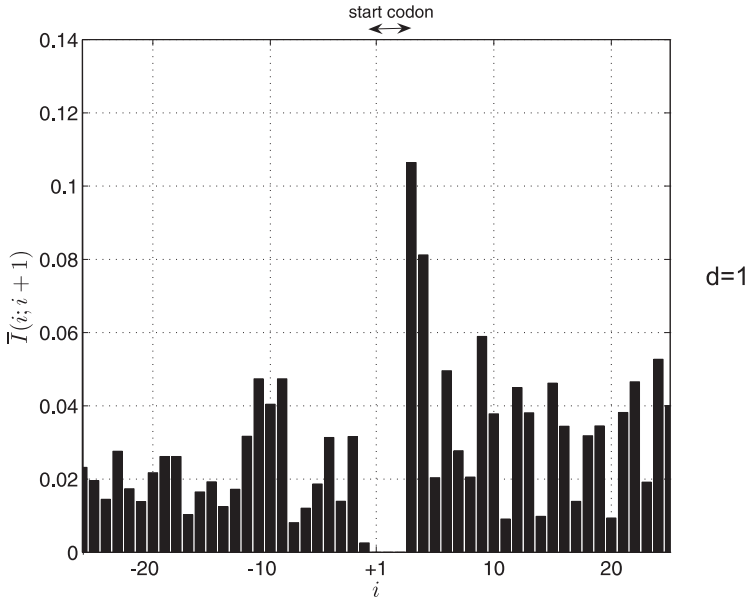


Figure 4.12: Detailed view of the mutual information between neighboring nucleotides ($d = 1$) around the start codon.

codons may code for the same amino acid but are not uniformly employed in the coding sequence. Second, the sequence structure of the codons is not random, i.e. they usually have nucleotide preferences (e.g. GC- or purine richness) [BBDI+06]. This interpretation is supported by the fact that the dependence decreases if increasing the distance between the two considered nucleotides to $d = 2$ (Figure 4.13, middle) and $d = 3$ (Figure 4.13, bottom). The results observed in Figure 4.13 for the first 200 bp (5'-UTR) and the last 200 bp (3'-UTR) can be seen as an indication for the mutual information estimates on random sequences. Their small range of values thus enhances the significance of the variations inside the coding sequence.

4.2.4 Synchronization properties

Frame synchronization in communication systems is most often a hit-or-miss problem, where shifted synchronizations by as few as one bit can lead to a decoding failure of the respective message. Synchronization during translation is similarly important for the synthesis of a correct protein. As mentioned in Section 3.4.4, the mRNA is translated in steps of three nucleotides (codons) with the start codon AUG being the first translated codon. If a frameshift occurs, i.e. if translation starts with a wrong phase, the whole mRNA sequence is translated into an erroneous and possibly shortened or – seldom – lengthened protein.

▷ **Example 4.1**

The correct translation of the following nucleotide sequence according to the genetic code (see Figure 3.9) would be

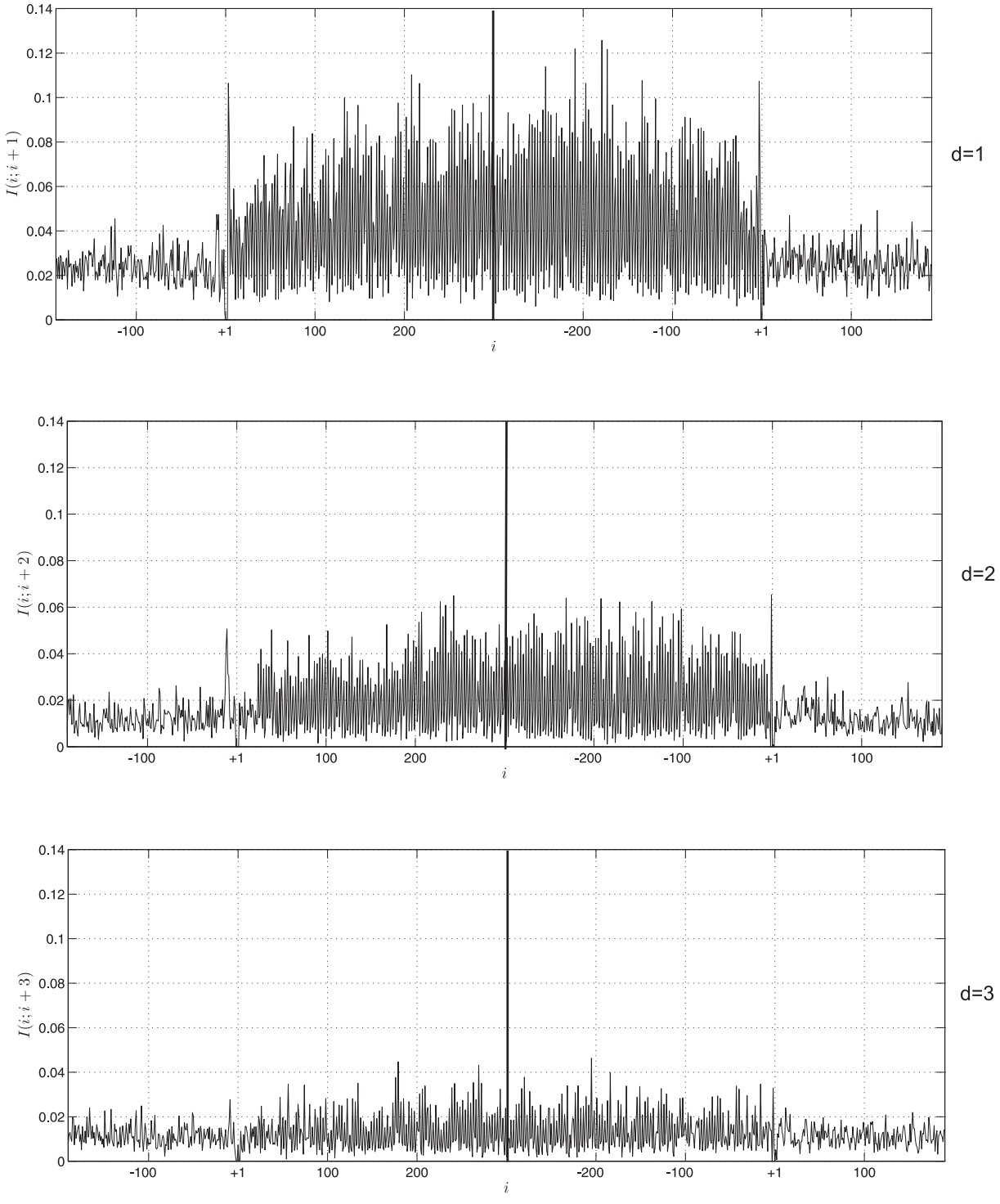


Figure 4.13: Mutual information between bases at distance $d = 1$ (top), $d = 2$ (middle) and $d = 3$ (bottom) for the *E. coli* mRNAs (left part: start codon at position $i = +1$, right part: stop codon at position $i = +1$).

AGA	UGU	CCG	UAC	CUC	AUC	GCU	UGG	Axx
Arg	Cys	Pro	Tyr	Leu	Ile	Ala	Trp	

If a frameshift by one base (+1) occurred, this amino acid chain would change to

GAU	GUC	CGU	ACC	UCA	UCG	CUU	GGA
Asp	Val	Arg	Thr	Ser	Ser	Leu	Gly

It can be seen that even a small frameshift can have dramatic effects on the resulting amino acid chain and, thus, the synthesized protein.

◁

The example illustrates the importance of synchronization in enabling the correct in-frame translation. As mentioned before, this synchronization is achieved by detection of the Shine-Dalgarno sequence shortly upstream of the start codon. In Section 2.4, the design of sync words was detailed: To avoid shifted synchronizations, the sync word should not exhibit periodicities. The Shine-Dalgarno sequence (AGGAGG) is periodic to $\tau = 3$ and thus – at first sight – appears to be a poor choice for such an important task as maintaining the reading frame of translation. This interpretation is, however, misleading since the mapping of codons to amino acids occurs in steps of three, i.e. a frame shift of +3 yields the same amino acid chain. If taking this into account, the Shine-Dalgarno sequence even appears to be a smart choice: Due to the periodicity, the ribosome has two chances to synchronize to the correct phase of translation. The two sequences AGG constitute rather short, but bifix-free synchronization sequences, whose concatenation strongly diminishes the probability of missed detections in the correct phase.

4.3 Summary

During vital cell processes, proteins bind to short DNA motifs which serve as biological synchronization words that mark the beginning of a regulatory sequence. In this chapter, two types of biological sync words were investigated with respect to their synchronization properties: the bacterial promoter and the Shine-Dalgarno sequence. The former is the sync word of transcription which is detected by the RNA polymerase, the latter is the sync word of bacterial translation which is detected by the ribosome. The main results of the investigations in this chapter are:

- ▷ An adapted autocorrelation function was derived based on binding energies between the synchronizing protein (the sigma factor) and the nucleotides. It was applied to rate the synchronization properties of the two promoter regions (the -35 and the -10 region) in *E. coli*. This brought up that the -35 promoter region is an excellent sync word in terms of minimizing the probability of shifted synchronizations. The -10 region showed to be the best possible sync word if taking other constraints imposed by transcription into account. The results suggest that the promoter regions evolved with respect to avoiding shifted synchronizations.

- ▷ The promoter was subsequently modeled as a distributed sync word, in which synchronization symbols are dispersed with arbitrary symbols. Its autocorrelation properties showed to be only average compared to all other possible sequences, which indicates that the promoter regions do not jointly serve synchronization.

- ▷ The probability of false synchronizations on random data is minimized by avoiding the sync pattern in the surrounding data stream, i.e. by choosing an under-represented word with respect to the Markov model underlying the data stream. Therefore, a Markov analysis of the *E. coli* genome was conducted, and the promoter regions were investigated in terms of their exceptionality. Both regions showed to be under-represented in the genome, which indicates that they were evolutionary avoided outside the promoter regions to minimize synchronization errors.

- ▷ The Shine-Dalgarno sequence was thereafter analyzed using the Kullback-Leibler divergence. For this purpose, a dataset of mRNA sequences was aligned to the translation start site. A high Kullback-Leibler divergence then indicates nucleotide biases at a fixed distance to the start site and thus might suggest a functional role in translation. The Shine-Dalgarno sequence was successfully detected by the relative entropy measure, however, no additional signal appeared.

- ▷ The Shine-Dalgarno sequence and the coding sequence were investigated using mutual information between two nucleotides at short distances from each other ($d \in [1;3]$). It exposed a strong dependency of neighboring nucleotides ($d = 1$) in the Shine-Dalgarno sequence. Moreover, strong dependencies for neighboring nucleotides were detected in the whole coding sequence, which is likely to stem from the triplet structure (3 nucleotides $\rightarrow$ 1 amino acid).

- ▷ Finally, the synchronization properties of the Shine-Dalgarno sequence were discussed. It exhibits a strong periodicity of three, which is generally an unfavored characteristic of sync words. In this case, however, a shifted synchronization caused by the periodicity would not have a strong impact on translation due to the 3-periodicity of the coding sequence itself. This fact supports the conclusion that the synchronization properties influenced the evolution of biological sync words.

5

Modeling Transcription Initiation in Prokaryotes

In Chapter 4, the analogies between transcription initiation and frame synchronization were outlined. While synchronization in technical systems is usually based on the cross-correlation between the sync word and the received data stream, the RNA polymerase detects the promoter based on the binding energy between its sigma subunit and the DNA sequence. Therefore, in order to obtain a valid synchronization model of transcription, a measure for the binding energy has to be derived. This measure is afterwards applied to known promoter sequences of *Escherichia coli* and their surrounding to investigate the synchronization signals the RNA polymerase encounters during its target search.

In Section 5.1, a derivation of the binding energy from published experimental data is presented. The energy measure is thereafter fit into a matrix to ease its application to DNA sequences. In contrast to known methods from bioinformatics, the matrix is applied to a large set of available promoter sequences to extract the common energy landscape, that is the behavior of binding energies in the surrounding of the transcription start site. The results of applying the algorithm to a wide surrounding of the promoters are presented and thoroughly interpreted in Section 5.2. Moreover, the promoter dataset is subdivided into smaller datasets to lay open correlations between the energy landscape and promoter characteristics. In Section 5.3, the process of promoter search by the RNA polymerase is analytically modeled as a random walk along the DNA with sequence-dependent transition probabilities. This enables the estimation of biophysical parameters with the aim of explaining the efficiency of promoter detection in the over-abundance of non-promoter sites present in the large genome of *Escherichia coli* ($4.6 \cdot 10^6$ bp).

5.1 Promoter detection in *Escherichia coli*

In general, the binding energy $E_l(s)$ between sigma factor and a promoter sequence l can be written as the sum of three energy terms:

$$E_l(s) = \underbrace{\sum_{k=1}^{6} e(n_k, k)}_{\text{-35 region}} + \underbrace{\sum_{k=7}^{12} e(n_{k+s}, k)}_{\text{-10 region}} + \underbrace{\epsilon(s)}_{\text{Spacing}}, \tag{5.1}$$

where $e(n, k)$ denotes the partial binding energy between the nucleotide $n \in \{A, C, G, T\}$ and the binding site of the sigma factor associated with promoter position k (see Figure 5.1). Thus, n_k and n_{k+s} refer to the nucleotides at positions k and $k+s$, respectively, of a given DNA sequence. The term $\epsilon(s)$ is the contribution of the spacing s between the promoter regions to the binding energy. This adds the energy the sigma factor needs to stretch or to squeeze in order to detect promoters with non-ideal spacing (i.e. other than 17 bp) [DJLG96, MBM85]. In Eq. (5.1), the contribution of nucleotides are assumed to be independent of their neighboring nucleotides, which is in most cases a reasonable approximation [DSS03, SF98].

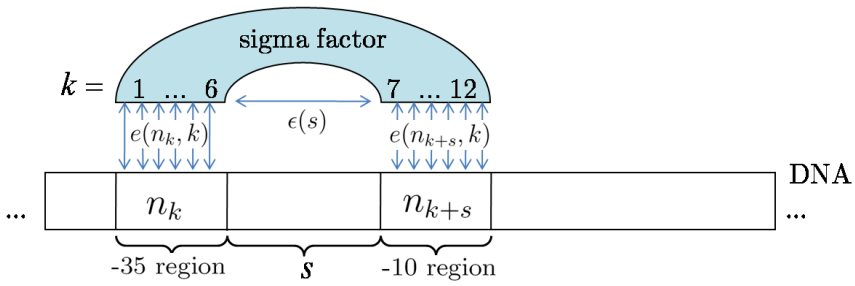


Figure 5.1: Components of the binding energy between sigma factor and the promoter.

5.1.1 Weight matrix model of σ^{70}

The values of $e(n, k)$ are extracted from [KOA05], where H. Kiryu et al. derived a measure for the nucleotide-dependent contribution to the binding energy by applying vector regression on gene expression data. The values $e(n, k)$ are used to build up a [4 x 12] weight matrix $\mathbf{W}(n, k)$ containing the contribution of the 12 nucleotides to the binding energy:

$$\mathbf{W}(n,k) = \begin{bmatrix} e(A,1) & e(A,2) & \dots & e(A,11) & e(A,12) \\ e(C,1) & e(C,2) & \dots & e(C,11) & e(C,12) \\ e(G,1) & e(G,2) & \dots & e(G,11) & e(G,12) \\ e(T,1) & e(T,2) & \dots & e(T,11) & e(T,12) \end{bmatrix}.$$

The values $k \in [1;6]$ reference the positions in the -35 region and $k \in [7;12]$ reference those in the -10 region. In [KOA05], the partial binding energies were defined such that positive values indicate a strengthening effect on the overall binding energy, whereas negative values imply a weakening effect. However, since in chemistry binding energies underlying stable interactions are generally given by negative values and, thus, high negative overall energies should indicate candidate target sites [SDS02], all values are multiplied by -1. Figure 5.2 shows the obtained values for each position $k \in [1;12]$ in the two promoter regions (left) as well as for each spacing $s \in [15;19]$ (right). It has to be mentioned that the values were obtained after various normalizations and hence have no physical unit. Nevertheless, in the following, the obtained measure is denoted by the term binding energy given without unit. It can be seen in Figure 5.2 and is reported in [KOA05] that the -35 sequence yielding the strongest (i.e. lowest) binding energy (AAGAAT) differs from the generally reported -35 consensus sequence (TTGACA, [LM93]).

▷ **Example 5.1**

According to Figure 5.2, a promoter with the -35 region TTCTAT, the -10 region TAAACC and a spacing of $s = 17$ would yield a binding energy of

$$\begin{aligned} E_{-35} &= -0.04 - 0.17 + 0.07 + 0.08 - 0.11 - 0.11 = -0.28, \\ E_{-10} &= -0.21 - 0.32 - 0.10 - 0.18 + 0.03 + 0.04 = -0.74, \\ E_s &= -0.30, \\ \Rightarrow \quad E &= E_{-35} + E_{-10} + E_s = -1.32. \end{aligned}$$

This highly negative energy indicates a strong binding of the sigma factor to the given sequence. Since this is the prerequisite for detection by the RNA polymerase, the given sequence would constitute a frequently detected promoter.

◁

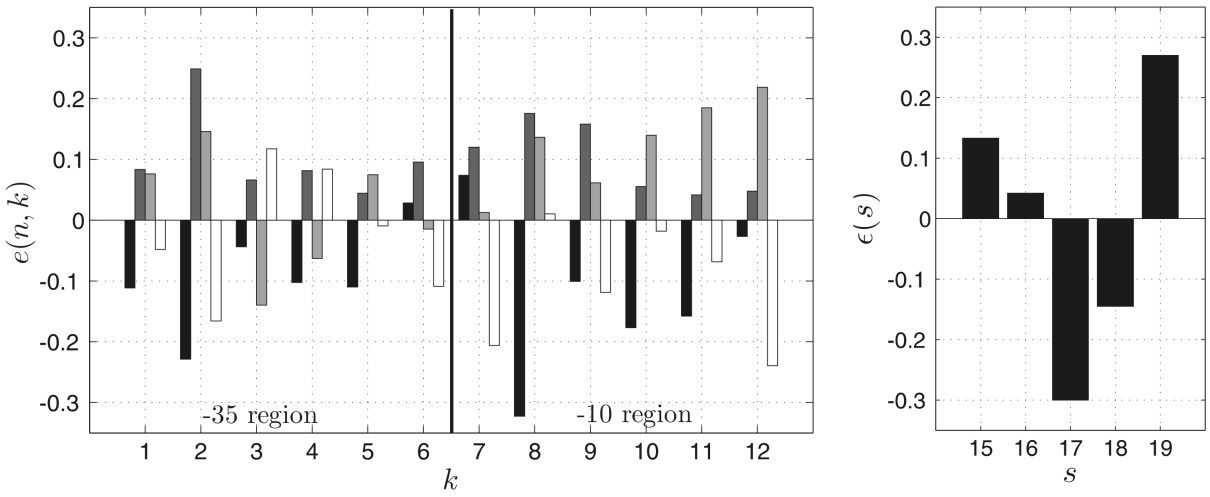


Figure 5.2: Partial binding energy contributions as extracted from [KOA05] and modified. Left: Promoter regions, color scheme: Black = A, dark gray = C, light gray = G, white = T. Right: Spacing between the promoter regions.

5.1.2 Synchronization algorithm

The matrix $\mathbf{W}(n,k)$ is applied using a sliding window that is shifted in single steps over the DNA. As mentioned before, the sigma factor can expand or compress and hereby adapt to different promoter spacings s in order to bind to the energetically most favorable site. That is, the sigma factor minimizes the binding energy over the possible spacings at each position. Therefore, the binding energy $E(i)$ at position i is obtained by minimizing the energy score $E(s,i)$ calculated according to Eq. (5.1) over the spacing s:

$$E(i) = \min_{s\in[15;19]}[E(s,i)] = \min_{s\in[15;19]}\left[\sum_{k=1}^{6} e(n_{i+k-1},k) + \sum_{k=7}^{12} e(n_{i+k+s-1},k) + \epsilon(s)\right], \quad (5.2)$$

where n_{i+k-1} and $n_{i+k+s-1}$ reference the nucleotides at positions k and $k+s$, respectively, of the sliding window, which is situated at position i with respect to the transcription start site (TSS, position 0, see illustration in Figure 5.3). The spacing is limited to $s \in [15;19]$ since most of the promoters fall in this range [LM93]. The binding energy $E(i)$ inversely measures the similarity between the template promoter sequence and the currently considered window. Therefore, the likelihood function defined in Section 2.1 is here given by:

$$L(\mu) = -E(i). \quad (5.3)$$

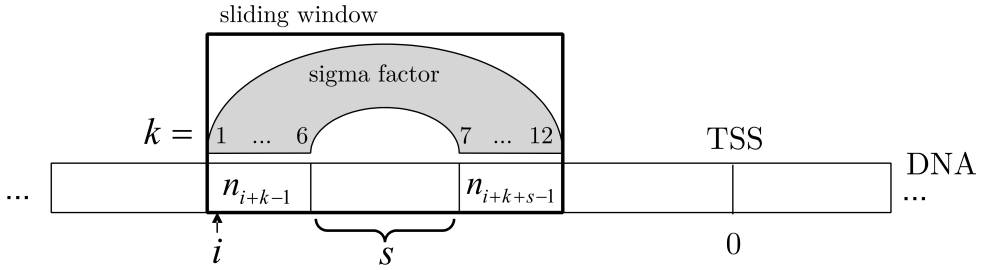


Figure 5.3: Graphical illustration of the parameters k, s and i from Eq. (5.2).

5.1.3 Average consideration

Applying the described algorithm to single sequences exhibits a noisy output with many false-positive signals, i.e. positions with energies as low or even lower compared to the actual promoter site. Therefore, the algorithm is applied on average, i.e. not for promoter detection of individual sequences but to a set of N known promoters of σ^{70} aligned to the transcription start site. Afterwards, the arithmetic mean (average) of the resulting values $E_l(i)$ is calculated for each position i, where the index l references the l-th promoter ($l \in [1;N]$). Additionally, the algorithm is applied to 10000 random sequences of length

31 (i.e. enabling all possible spacings) considering $p(A) = p(T) = 0.246$ and $p(C) = p(G) = 0.254$ as estimated from the entire *E. coli* genome, which yields a mean value of $\overline{E}_{\text{ran}} = -0.4423$ (calculated according to Eq. (5.2)). Since it is convenient to set the average energy as 0 (see e.g. [SDS02]), the energy measure obtained through averaging over all $E_l(i)$ is normalized accordingly:

$$\overline{E}(i) = \frac{1}{N} \sum_{l=1}^{N} E_l(i) - \overline{E}_{\text{ran}}. \tag{5.4}$$

In considering average values, the noise of individual sequences can be eliminated in order to extract the common energy landscape of all σ^{70}-promoters, i.e. the characteristic behavior of binding energies around the transcription start sites. Due to the normalization by $\overline{E}_{\text{ran}}$ in Eq. (5.4), positive values of $\overline{E}(i)$ indicate a below-average binding strength, while negative values refer to an above-average binding strength between sigma factor and DNA sequence.

5.2 Results and interpretation

The algorithm is applied to a set of *E.coli* σ^{70}-promoters from the NCBI database (see Appendix C.1.1) that classifies promoters into computationally predicted and experimentally documented. Figure 5.4 shows the modeled average binding energy $\overline{E}(i)$ calculated according to Eq. (5.2) and Eq. (5.4) (see Section 5.1.2) for all 3765 predicted and documented promoters in a range of 200 bp around the promoters aligned to the transcription start site (TSS, $i = 0$). $\overline{E}(i) = 0$ corresponds to the energy $\overline{E}_{\text{ran}}$ of random sequences.

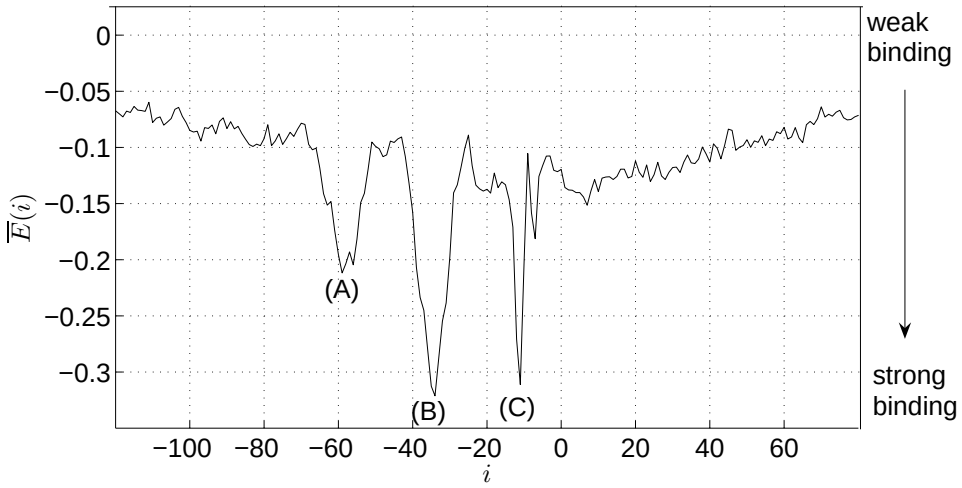


Figure 5.4: Average binding energy $\overline{E}(i)$ of 3765 known σ^{70}-promoters aligned to the transcription start site (TSS).

5.2.1 Additional synchronization signals

Figure 5.4 shows three significant synchronization signals at positions -58 (see (A)), -35 (see (B)) and -12 (see (C)) compared to the surrounding and to the average binding energy $\overline{E}(i) = 0$ of random sequences. The most significant minimum at around -35 reflects the actual recognition of both promoter regions, whereas those at -58 and -12 occur due to correlation between the -35 sequence and the -10 sequence; At position -58, the -10 part of the weight matrix (modeling the sigma factor) is overlapping the -35 promoter region (see Figure 5.5, (A)). The same applies for the minimum at -12, which occurs due to correlation between the -35 matrix part and the -10 promoter region (see Figure 5.5, (C)). The absolute minimum at -35 indicates the actual detection site and hereby the appropriate modeling through the weight matrix (see Figure 5.5, (B)).

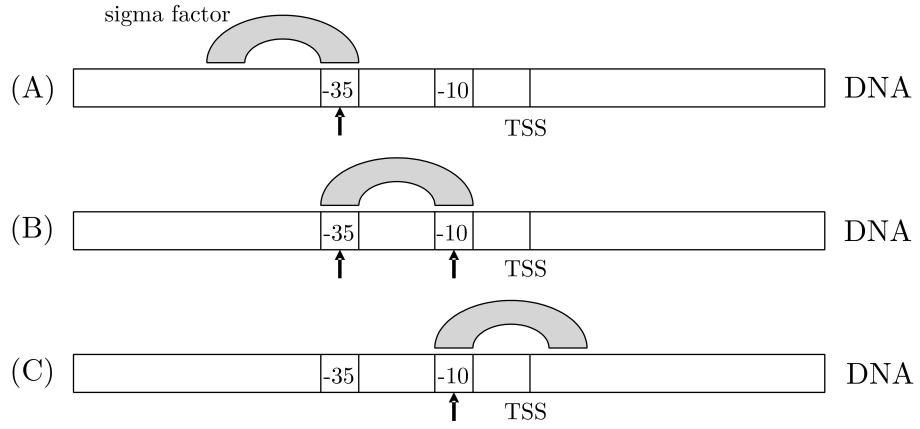


Figure 5.5: Illustration of the process of promoter detection in three steps.

The additional synchronization signals before and after the promoter suggest an interesting approach to ensure reliable detection of the target site: Due to correlation between the two promoter region, the sigma factor encounters a pre-synchronization signal from either side which may prepare the RNA polymerase for the close-by target site.

It should be mentioned that the strength of the binding is not only reflected by the depth of the minima but by their area due to the flexibility of the promoter's position with respect to the TSS. While the position of the -35 region varies by around ± 3, the position of the -10 region is restricted within around ± 1. Therefore, since Figure 5.4 shows the average binding energies of 3765 promoters, the minimum at -58 is broad but flat, whereas that at -12 is deep but narrow. The minimum at -35 is broad and deep at the same time and, thus, reflects – as expected – the strongest binding.

5.2.2 Energy landscape in the wider surrounding

In order to investigate the energy landscape in the wider surrounding of the promoters, the presented method is applied to a range of 800 bp around the transcription start site.

Figure 5.6 shows the average binding energy $\overline{E}(i)$ of all 3765 promoters aligned to the TSS (position 0). Remarkable are the decline beginning 300 bp before the promoter and the constant incline of the binding energy in the 200 bp after the promoter. It is obvious in comparison with the energy $\overline{E}(i) = 0$ of random sequences that the average binding energy deviates significantly from the random case in a range of about 500 bp around the promoter. (A), (B), and (C) in Figure 5.6 show the three minima at -58, -35, and -12 that were observed in Figure 5.4. The wide range of non-random binding energies around the promoters implies that the movement of the RNA polymerase is influenced long before the target site is encountered. This suggests that not only the promoter site itself but a range of 500 bp around the latter is involved in the synchronization process underlying transcription initiation. The underlying mechanisms is biophysically investigated in Section 5.3.

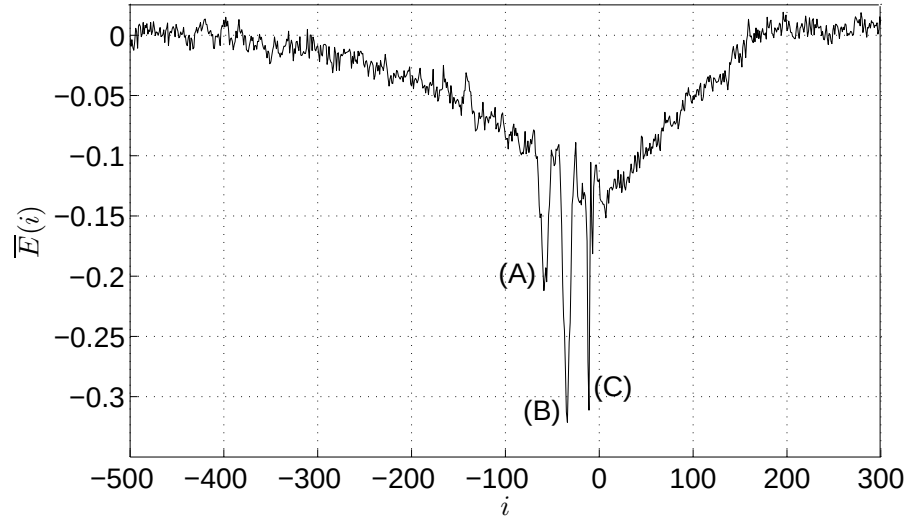


Figure 5.6: Average energy landscape $\overline{E}(i)$ in a wider range around the 3765 known σ^{70}-promoters aligned to the transcription start site.

5.2.3 Clustering of promoters

In the next step, it is investigated whether the energy landscape observed in Figure 5.6 is present in all σ^{70}-promoters or only occurs in certain sets with specific promoter strength. Therefore, the 3765 promoters are split into subgroups according to their strength as measured by the sigma factor's binding energy $E_l(i)$ to the l-th promoter sequence calculated according to Eq. (5.2) (see Section 5.1.2). Here, $i \in [-38; -32]$ depending on the position of the l-th promoter with respect to the TSS. In order to maintain a sufficient statistical basis, the promoters are divided into four groups with approximately 900 promoters each. Figure 5.7 shows the resulting plots sorted from weak (top) to strong promoters (bottom). It can be seen that the characteristic energy landscape is not observed for weak promoters, whereas it becomes distinct for stronger promoters. It is generally assumed that the binding energy at the promoter itself determines the rate of detection and hereby the

expression rate of the respective gene [KNI90, SDS02, KOA05]. However, a dependence of the promoter strength on the wider surrounding has not been reported before.

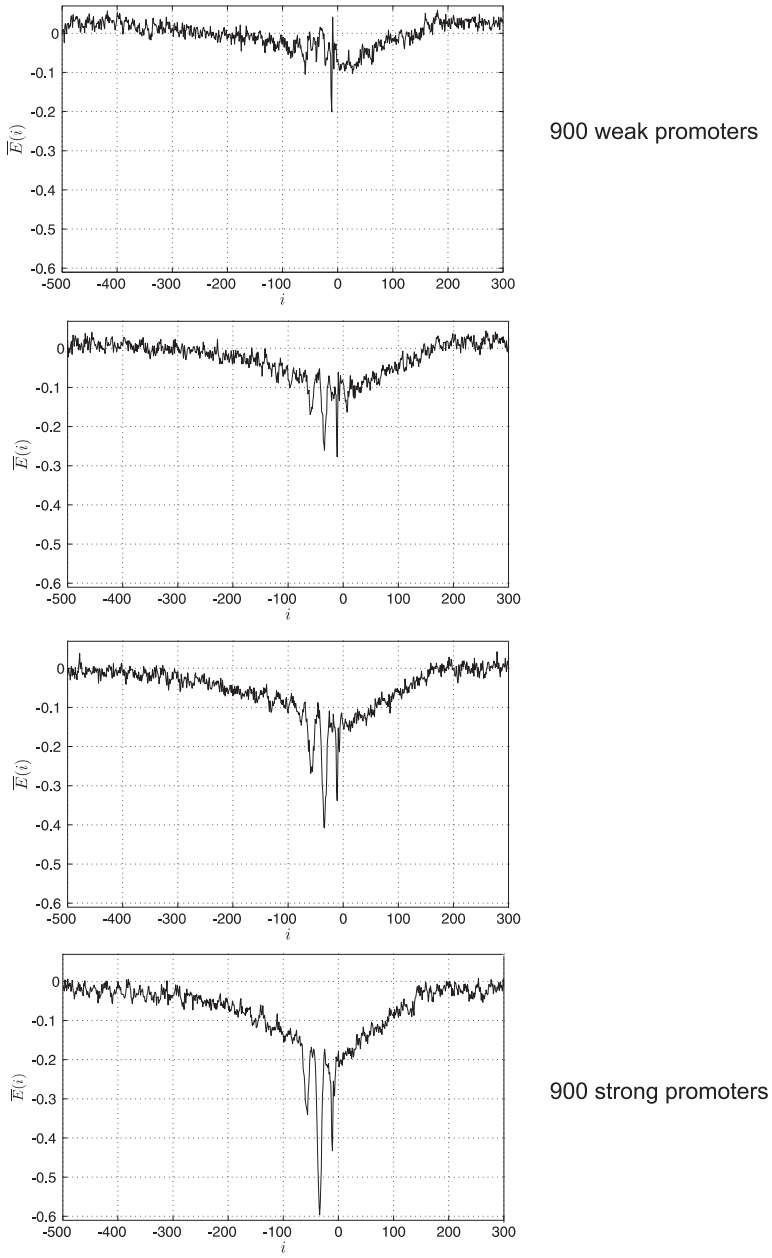


Figure 5.7: Average energy landscapes $\overline{E}(i)$ of 4 groups with approximately 900 sequences each (top: weak promoters, bottom: strong promoters).

5.3 Kinetic analysis of promoter search by σ^{70}

In 1999, it was impressively visualized that the RNA polymerase searches its target site (the promoter) by randomly binding the DNA and subsequently sliding along several

hundred base pairs before dissociating or moving to another site through hopping or intersegmental transfer [BGZY99, GZR+99]. The sliding process was assumed to be a 1D Brownian motion along the DNA – a random movement with same probabilities for moving right and left. In contrast to that, the following considerations are based on the assumption that the sliding process has a sequence-dependent component, i.e., that the RNA polymerase encounters a specific energy landscape that depends on the bound DNA sequence (sliding model). This assumption is supported by results from L. Mirny et al. [SM04], M. Barbi et al. [BPPS04] and others.

5.3.1 Arrhenius equation

According to the assumptions underlying the sliding model, the RNA polymerase does not perform a random walk with equal probabilities of stepping forward or backward, but the sliding is influenced by the binding energy at each position. In this case, the transition rates $w_{i,i+1}$ and $w_{i,i-1}$ from site i to site $i+1$ and site $i-1$, respectively, depend on the binding energies E between RNA polymerase and the DNA at these sites through

$$w_{i,i\pm1} = \nu \cdot \begin{cases} e^{-\beta[E(i\pm1)-E(i)]} & \text{if } E(i\pm1) > E(i) \\ 1 & \text{otherwise} \end{cases}, \tag{5.5}$$

with $\beta = (k_B T)^{-1}$, where ν denotes the affective attempt frequency, k_B the Boltzmann constant, and T the ambient temperature in Kelvin [SM04]. The affective attempt frequency can be considered as the rate at which thermal fluctuations try to push the protein away from site i. According to Eq. (5.5), known as the regular activated transport form or Arrhenius equation, the transition rate $w_{i,i\pm1}$ is under constant conditions solely dependent on the difference between $E(i\pm1)$ and $E(i)$, which corresponds to the gradient g of the function $E(i)$. The constant rate ν of $w_{i,i\pm1}$ in the case $E(i\pm1) \leq E(i)$ is used since any thermal fluctuation will push the RNA polymerase independently of the value $\Delta E = E(i\pm1) - E(i)$. Hence, it steps uphill (i.e., to a site with higher energy) with a rate smaller than ν (depending on ΔE) and steps downhill (i.e., to a site with lower energy) with constant rate ν. Note that $\beta = N_0^{-1}$, i.e. it corresponds to the thermal noise present in all communications engineering systems.

5.3.2 Linear approximation of the energy landscape

As mentioned before, the movement of the RNA polymerase depends on the gradient of the binding energy function. In order to ease the calculation of the transition rates $w_{i,i\pm1}$ and further kinetic parameters based on the gradient of the energy landscape, the latter is approximated by four straight lines with different gradients g (see Figure 5.8). Region 1 and region 4 correspond to random sequences, since $g = 0$ and $\overline{E}(i) = 0$. In region 2 and region 3, however, a negative and positive gradient g, respectively, is observed. Since g is constant in each of the four regions of the approximation, it can easily be deduced that

$$g = \overline{E}(i+1) - \overline{E}(i) = -\big[\overline{E}(i-1) - \overline{E}(i)\big]. \tag{5.6}$$

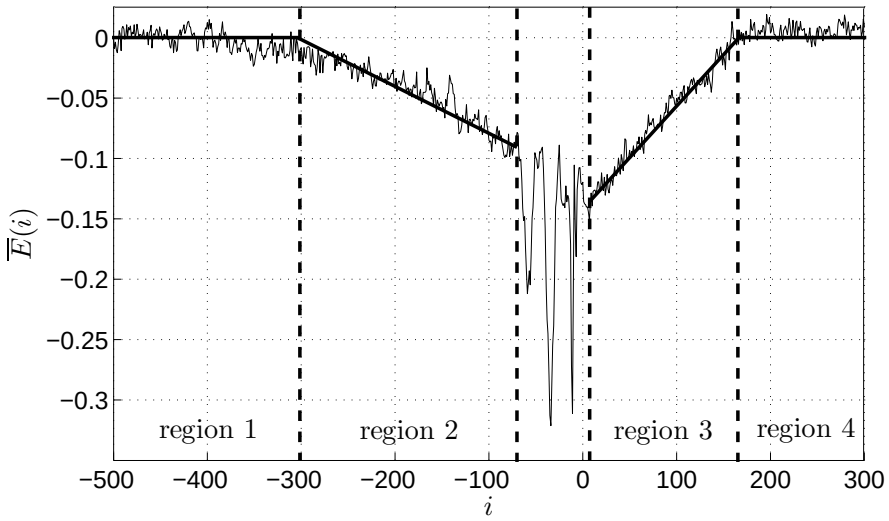


Figure 5.8: Linear approximation of the characteristic energy landscape.

5.3.3 Speed

In the sliding model, the protein is assumed to slide in single nucleotide steps along the DNA. The escape rate of the protein at site i to one of the neighboring sites and, therefore, the speed of the sliding process is given by (see Appendix D.1)

$$\rho_i = \frac{1}{\tau_i} = w_{i,i+1} + w_{i,i-1} = \nu(e^{-\beta|g|} + 1), \tag{5.7}$$

where τ_i denotes the time the protein spends bound to site i [SM04]. Applying Eq. (5.7) to the four regions of the linear approximation yields the following estimations for ρ_i:

$$\begin{aligned}
&\text{Region 1:} \quad |g| = 0 \quad \rightarrow \quad \rho_i = 2\nu. \\
&\text{Region 2:} \quad |g| > 0 \quad \rightarrow \quad \rho_i < 2\nu. \\
&\text{Region 3:} \quad |g| > 0 \quad \rightarrow \quad \rho_i < 2\nu. \\
&\text{Region 4:} \quad |g| = 0 \quad \rightarrow \quad \rho_i = 2\nu.
\end{aligned}$$

It can be seen that the escape rate ρ_i is lower in the direct surrounding of the promoters than on random sequences ($|g| = 0$). Consequently, the on-site time τ_i of the protein is higher and thus the speed of the sliding process is lower than on random sequences.

5.3.4 Direction

At each site i, the protein eventually escapes to the site $i+1$ with probability p_i and to the site $i-1$ with probability $q_i = 1 - p_i$. This probability p_i depends on the transition rates $w_{i,i+1}$ and $w_{i,i-1}$ through [SM04]

$$p_i = \frac{w_{i,i+1}}{w_{i,i+1} + w_{i,i-1}} = \frac{w_{i,i+1}}{\rho_i}. \tag{5.8}$$

Applying Eq. (5.8) to the approximation of $\overline{E}(i)$ yields the following estimations of the transition probability p_i:

$$\begin{aligned}
&\text{Region 1:} \quad w_{i,i+1} = \nu \text{ and } w_{i,i-1} = \nu \quad \rightarrow \quad p_i = 0.5, \quad q_i = 0.5. \\
&\text{Region 2:} \quad w_{i,i+1} = \nu \text{ and } w_{i,i-1} < \nu \quad \rightarrow \quad p_i > 0.5, \quad q_i < 0.5. \\
&\text{Region 3:} \quad w_{i,i+1} < \nu \text{ and } w_{i,i-1} = \nu \quad \rightarrow \quad p_i < 0.5, \quad q_i > 0.5. \\
&\text{Region 4:} \quad w_{i,i+1} = \nu \text{ and } w_{i,i-1} = \nu \quad \rightarrow \quad p_i = 0.5, \quad q_i = 0.5.
\end{aligned}$$

Note that the transition probability p_i from site i to site $i+1$ increases when approaching the promoter regions from upstream and decreases when leaving the promoter regions. In case the RNA polymerase approaches the promoter from downstream, the probability q_i increases upon entering region 3. Thus, in both cases the surrounding of the transcription start site seems to direct the sliding towards the target site. Guthold et al. [GZR$^+$99] showed that the RNA polymerase has no defined direction of sliding when binding to promoter-less DNA fragments, which is consistent with the calculated values of p_i and q_i for region 1 and region 4.

5.3.5 Efficiency

It was shown in previous studies [BGZY99, GZR$^+$99] that the RNAP slides forward and backward several times during its search for the promoters. The efficiency of the sliding process can be regarded as the number of steps the protein needs to reach from one site to another. Therefore, the mean first-passage time (MFPT) is used, which is defined as the mean number of steps the protein will make to slide from site $i = 0$ to site $i = x$ assuming a certain set of transition probabilities $\{p_i\}$ (see [MK89] for more information). The mean first-passage time $\overline{t}_{0,x}$ is given by

$$\overline{t}_{0,x} = x + \sum_{k=0}^{x-1} \alpha_k + \sum_{k=0}^{x-2} \sum_{i=k+1}^{x-1} (1 + \alpha_k) \prod_{j=k+1}^{i} \alpha_j, \tag{5.9}$$

where $\alpha_i = q_i/p_i$ (see [SM04] and references therein). For negative values of x (i.e., a sliding backwards), the definition of α changes to $\alpha_i = p_i/q_i$ and x in Eq. (5.9) has to be exchanged for its absolute value $|x|$. Considering the linear approximation in Figure 5.8,

the values of α_i are constant over wide ranges. Therefore, $\alpha_0 = \alpha_1 = \cdots = \alpha_x := \alpha$ is assumed in each region, which simplifies Eq. (5.9) to (see Appendix D.2 for the derivation)

$$\bar{t}_{0,x}^{\,\prime} = (1+\alpha)\sum_{k=0}^{x-1}(x-k)\alpha^k, \tag{5.10}$$

i.e., the mean first-passage time $\bar{t}_{0,x}^{\,\prime}$ for the approximation is given by a polynomial in α of degree x. Figure 5.9 shows the MFPT for $x = 50$. It can be seen that even small decreases of α from the random walk case with equal transition probabilities ($\alpha = 1$, i.e., $p_i = q_i = 0.5$) significantly decrease the mean first-passage time. At the same time, small increases of α lead to dramatic increases of $\bar{t}_{0,x}^{\,\prime}$. For the four regions of the approximation and for $x = 50$ ($\alpha_i = q_i/p_i$), this can be summarized by:

$$\begin{aligned}
&\text{Region 1:} \quad p_i = 0.5, \quad \text{i.e., } \alpha = 1 \quad \rightarrow \quad \bar{t}_{0,x}^{\,\prime} = 2550.\\
&\text{Region 2:} \quad p_i > 0.5, \quad \text{i.e., } \alpha < 1 \quad \rightarrow \quad \bar{t}_{0,x}^{\,\prime} \downarrow\\
&\text{Region 3:} \quad p_i < 0.5, \quad \text{i.e., } \alpha > 1 \quad \rightarrow \quad \bar{t}_{0,x}^{\,\prime} \Uparrow\\
&\text{Region 4:} \quad p_i = 0.5, \quad \text{i.e., } \alpha = 1 \quad \rightarrow \quad \bar{t}_{0,x}^{\,\prime} = 2550.
\end{aligned}$$

The symbols $\downarrow$ and $\Uparrow$ denote a slight decrease and a strong increase, respectively. For $x = -50$ ($\alpha_i = p_i/q_i$), i.e., reaching 50 positions backward of position i, the values in region 2 and region 3 are exchanged. The value $\bar{t}_{0,x}^{\,\prime} = 2550$ shows clearly the inefficiency of the sliding process on random sequences: According to this calculation, the RNA polymerase needs on average 2550 steps to bridge a distance of only 50 bases. Apparently, the decrease of binding energies the RNA polymerase faces when approaching the promoters strongly influences the efficiency of promoter search. In case the RNA polymerase approaches the promoter from downstream, region 3 increases the efficiency in direction of the promoter. If approaching the promoter from upstream, the search becomes more efficient upon entering region 2. Hence, the RNA polymerase is directed towards the promoter from either side. Opposed to that, it seems to be nearly impossible for the RNA polymerase to move further downstream if the promoter has been missed (entering region 3 from upstream), i.e., if no transcription initiation has taken place. Since it is known that the RNA polymerase is able to slide backward, this suggests that the RNA polymerase is guided back to the promoter in case it missed the detection.

5.3.6 Verification

In Sections 5.3.3 - 5.3.5, it was hypothesized that the observed characteristic behavior of binding energies around the promoter guides the RNA polymerase to the transcription start site. As mentioned before, the binding energies are assumed to depend on the underlying sequence, thus, the surrounding seems to exhibit sequence features that are

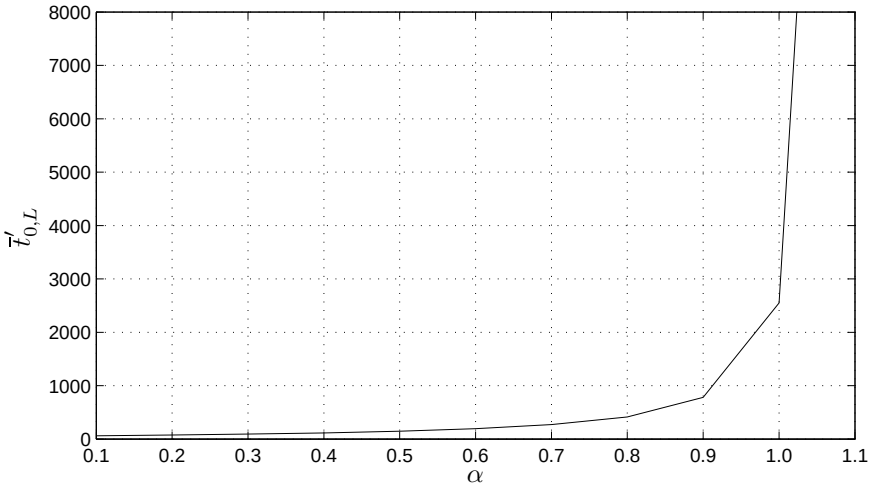


Figure 5.9: Mean first-passage time for $x = 50$ ($\alpha_i = q_i/p_i$) or $x = -50$ ($\alpha_i = p_i/q_i$) if assuming constant values of α_i.

unlikely to continue into the coding sequence since that has to serve other sequence constraints and can only code further signals through the variability of the third codon position. Therefore, one would expect a positive correlation between the length of the 5'-UTR (the range between the transcription start site, TSS, and the translation start site, TLS) and the promoter strength, since a long 5'-UTR constitutes a larger range that can guide the RNA polymerase during promoter detection. Figure 5.10 shows the average correlation between the length of the 5'-UTR and the promoter strength measured by the transcription rate. A sliding window of length 150 is used to smooth the results. The transcription rates are extracted from the ASAP database [UoWM07] (see Appendix C.1.1 for more information). As expected, it exhibits a positive trend and thus supports the hypotheses presented in Section 5.3.

5.4 Summary

In this chapter, the process of transcription initiation in *E. coli* was modeled using a synchronization algorithm. It was built upon previously published binding energies between the sigma factor (the synchronizing protein) and the promoter sequence (the sync word of transcription). This binding energy corresponds to the likelihood function that is used in technical systems to measure the similarity between sync word and data stream. It was subsequently applied to an aligned set of known promoters to extract the energy landscape, i.e. the behavior of binding energies around the transcription start sites. Minima of the energy landscape indicate signaling sequences related to transcription initiation. The following main results were achieved:

- ▷ In addition to a minimum at the exact promoter site, the energy landscape exhibited two minima shortly after and before that site. These showed to occur due to

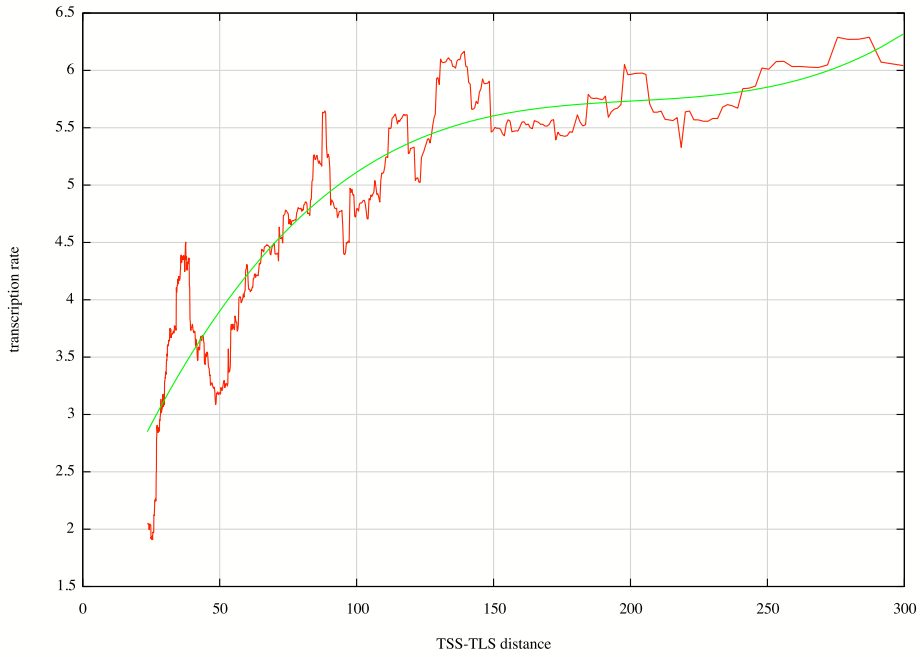


Figure 5.10: Correlation between the length of the 5'-UTR (distance between TSS and TLS) and the transcription rate.

correlation between the two promoter regions (the -35 and the -10 region). Due to this correlation, the sigma factor encounters pre-synchronization signals when approaching the promoter from either side.

- ▷ A characteristic behavior of binding energies in the wider surrounding of the promoter (500 base pairs) was observed. This allows the assumption that not only the promoter site itself aids the detection by the sigma factor but that in addition the wider surrounding guides the sigma factor during its search for the promoter.

- ▷ In the next step, the promoters were subdivided according to their strength as measured by the binding energy at their location. Applying the synchronization algorithm to these subsets brought up that only the strong promoters exhibit the characteristic energy landscape. This fact strengthens the hypothesis that the latter is related to guiding the sigma factor towards the promoter site.

- ▷ The characteristic energy landscape was thereafter theoretically analyzed in terms of the underlying biophysical properties of the movement of the sigma factor along the DNA during promoter search. The results indicate that the movement is slowed down, guided towards the promoter and made efficient through the observed energy landscape.

- ▷ In summary, all listed results imply that the behavior of binding energies aids the synchronization process underlying transcription initiation: While the wider surrounding guides the sigma factor to the transcription start site of highly expressed genes, pre-synchronization signals ensure that the exact promoter site is not missed.

6

Modeling Transcription Initiation in Eukaryotes

The process of promoter search and transcription by bacterial RNA polymerase has been visualized using scanning force microscopy (see e.g. [GZR+99, HFM+99, BGZY99]). In contrast to that, the process is far less understood in higher organisms (eukaryotes) and no according direct observations exist to date. Not even the exact order of proteins binding to the DNA is known yet. For this reason, this chapter focusses on a thorough analysis of the DNA sequences in the surrounding of annotated transcription start sites rather than on modeling single interactions.

In Section 6.1, the major differences between transcription in prokaryotes and in eukaryotes are specified. Section 6.2 follows with an information theoretic analysis of two promoter datasets that aim at detecting the TATA-box – the main promoter element in eukaryotes. First, weight matrices are introduced, a standard bioinformatics tool for the detection of protein binding sites. Subsequently, two alternative methods are derived based on mutual information and the Kullback-Leibler divergence. The results of all three methods are thereafter interpreted in Section 6.3. In Section 6.4, the promoter sequences are subdivided into smaller groups to obtain more meaningful results.

6.1 Differences to bacterial transcription initiation

In Section 3.4, the process of transcription was detailed for bacteria (prokaryotes) as well as for higher organisms (eukaryotes). It could be seen that while the process is based on

relatively few interactions in the former, it involves numerous components and steps in the latter. Therefore, modeling transcription initiation in eukaryotes, especially in more complex ones like animals or humans, yields no complete picture if focusing only on single interactions. For this reason, the results presented in the following should be considered as a general analysis that, however, most likely does not apply to all conditions of the cell and all tissues.

6.1.1 Protein-DNA interaction of the RNA polymerase

In Chapter 5, a model of the bacterial RNA polymerase during its search for the promoter site was presented. Since prokaryotic transcription can take place without the additional binding of other proteins, the interaction between the RNA polymerase and the DNA constituted the clear focus of communication theoretic modeling. In eukaryotes, however, this interaction is only made possible through the binding of transcription factors, proteins that bind the DNA at specific positions and later guide the RNA polymerase to the promoter. Moreover, the RNA polymerase in *E. coli* has been shown to slide along the DNA, thus enabling a scanning process of the underlying nucleotide sequence (equivalent to the receiving process in frame synchronization) [GZR+99, HFM+99]. In contrast to that, the process of promoter binding in eukaryotes is believed to occur based on three-dimensional looping of the DNA combined with the mentioned aiding proteins [Nog00].

6.1.2 Promoter elements

The most important promoter for eukaryotic transcription initiation is the TATA-box, an AT-rich region located around 25 bp before the transcription start site (see Figure 6.1, consensus sequence TATAAAAG). In contrast to prokaryotes, the promoters of eukaryotes do not contain a -35 box. Instead, an initiator region (Inr) overlapping with the transcription start site serves as a second, though weaker, signal for transcription initiation [RH05].

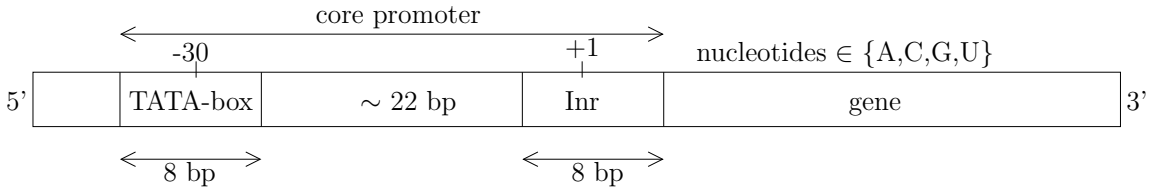


Figure 6.1: Structure of the core promoter region in eukaryotes.

In addition to the TATA-box, the equivalent to the bacterial -10 region (see Section 3.4.2), many promoters also exhibit a pyrimidine-rich region around the transcription start site (the initiator region), a GC-rich sequence immediately upstream of the TATA-box (the TFIIB recognition element, BRE) and the downstream promoter element (DPE, located around 30 bp after the transcription start site). In the following, the focus lies on the

TATA-box since this is the most conserved promoter element between different species which indicates its functional importance for transcription initiation. It is reported to be present in a high number of promoters and located around position $i = -30$.

6.1.3 Transcription factor binding sites

Transcription factors are proteins that bind to the DNA to regulate transcription. The core promoter region in prokaryotes is confined to approximately 60 bp before the transcription start site (TSS, see Section 3.4.2), and transcription factor binding sites cumulate in the few hundred base pairs around the TSS. In eukaryotes, the core promoter region comprises a similar region as in bacteria, however, transcription factor binding sites play a more important role and are scattered over more than 1000 bp around the TSS.

6.1.4 CpG islands

A prominent sequence feature of eukaryotic DNA are CpG islands, i.e. regions with high concentration of the dinucleotide CG (5'-CG-3'). The "p" refers to the phosphodiester bond between the C and the G. CpG islands accumulate in promoter regions and constitute regulatory targets during transcription initiation. The human genome is expected to carry 29 000 CpG islands [Lew07], especially in the promoter regions of constitutively expressed genes [SBB06]. Methylation of CpG islands – the addition of a methyl group to the cytosine ring – can both prevent and cause proteins to bind to their target DNA sequence. The strongest deviation in the CpG-content is observed between insects and mammals [JB04].

6.1.5 Chromatin

In most prokaryotes, the genome is a single circular molecule of easily accessible DNA. In contrast to that, the DNA in eukaryotic organisms is organized in a complex structure called chromatin that makes up the chromosomes. One function of chromatin is to compress the DNA by wrapping it tightly around histone proteins and hereby reaching a high packing ratio (see Figure 6.2). Another effect of chromatin lies in the resulting limited accessability of the DNA: proteins required for gene expression to occur can only bind to the DNA at stretches not occupied by proteins of the chromatin complex [SFMC$^+$06].

6.2 Information theoretic analysis

In the first step of transcription initiation, the transcription factor TFIID binds to the TATA-box (see Section 3.4.3). As soon as this step has taken place, the other transcription factors join the complex to build the initiation complex that starts transcription. The

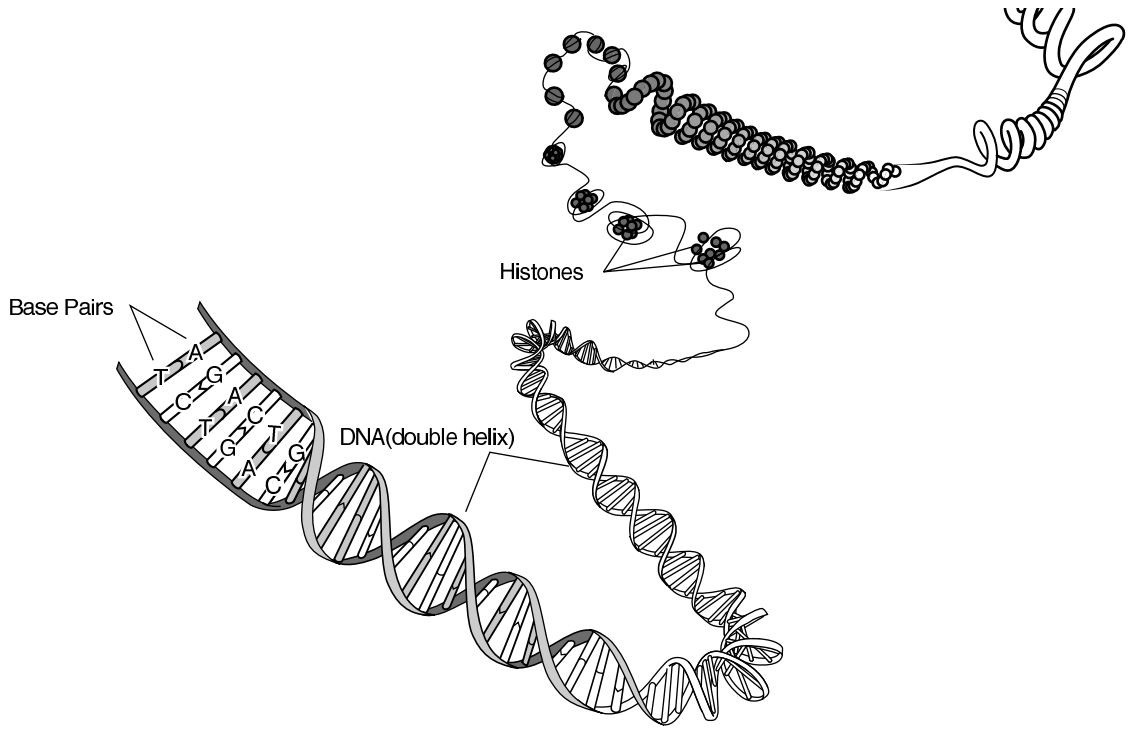


Figure 6.2: Chromatin structure of eukaryotic DNA [NIoH08b].

detection of the TATA-box thus constitutes an essential step that corresponds to a frame synchronization process where TFIID acts as the receiver. In this section, information theoretic measures are adapted for their application to promoter datasets. They are derived such that they constitute a similarity measure between the TATA-box and the dataset, i.e. they mimic the likelihood function applied for technical frame synchronization (see Section 2.1). Additionally, a weight matrix – a standard method from bioinformatics for the detection of sequence motifs – is applied for reasons of comparison. The approaches are applied to two datasets of DNA sequences: human and arthropod promoter sequences downloaded from the EPD database (see Appendix C.1.2 for more information on the datasets). The human dataset comprises $N = 1871$ sequences, the arthropod dataset comprises $N = 1996$ sequences. Both are aligned (i.e. centered) to the transcription start site as done in Section 5.1.3 with prokaryotic promoters.

6.2.1 Weight matrix model

Weight matrices constitute a method frequently applied in bioinformatics for the detection of sequence motifs (or words) in the DNA. They are trained on known motifs and then assign a score to a DNA sequence that reflects its similarity to the sought sequence motif (see e.g. [Sto00]). In Chapter 5, a weight matrix based on partial binding energies was presented for the bacterial promoter.

Derivation

The more common way of constructing a weight matrix – due to the limited availability of binding energy data – is based on the nucleotide distribution at each position of the motif. Table 6.1 lists the probabilities of bases at each of the eight positions of the TATA-box (taken from [Buc90]). In position $k = 1$ for example, the nucleotide T was observed in 80 % of 389 analyzed promoter sequences with a TATA-box, C was observed in 12 % of the sequences etc. In most applications, those probability values are subsequently normalized to the background distribution of bases. Moreover, the values are transformed such that a value of zero refers to a maximum over-representation of a base ($p(n, k) = 1$), while high negative values indicate an under-representation of a base at a certain position:

$$\mathbf{W}(n, k) = \ln\left(\frac{\hat{p}_{\mathbf{s}}(n, k)}{\hat{p}(n, k)} + c_1\right) + c_2, \tag{6.1}$$

where $\hat{p}_{\mathbf{s}}(n, k)$ denotes the observed probability of nucleotide n at position k of the motif $\mathbf{s}$ and $\hat{p}(n, k)$ the expected frequency of nucleotide n at position k. The latter is usually derived from the dinucleotide composition of the genome, i.e. the occurrence frequencies of bases under Markov model $M2$ (order $m = 2$, see Section 2.3.2). Moreover, c_1 is a smoothing parameter that prevents zero-terms in the logarithm, and c_2 is a constant commonly chosen such that the maximum resulting value is zero (see [Buc90] for more details on the transformation and normalization). The normalized and transformed weight matrix of the TATA-box is given in Table 6.2 [SIoB07].

Table 6.1: Probability matrix of the eukaryotic TATA-box.

position k	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
$\mathbf{W}(A, k)$	0.04	0.91	0.01	0.91	0.69	0.93	0.57	0.40
$\mathbf{W}(C, k)$	0.12	0.00	0.03	0.00	0.00	0.01	0.01	0.11
$\mathbf{W}(G, k)$	0.04	0.00	0.00	0.01	0.00	0.05	0.11	0.40
$\mathbf{W}(T, k)$	0.80	0.09	0.96	0.08	0.31	0.01	0.31	0.09
consensus	T	A	T	A	A	A	A	A/G

Homology score

The homology score $S(\mathbf{n}, \mathbf{s})$ between a DNA sequence $\mathbf{n} = \{n_1, \ldots, n_L\}$, $n_k \in \mathcal{A} = \{A, C, G, T\}$, and a sought sequence motif $\mathbf{s} = \{s_1, \ldots, s_L\}$, $s_k \in \mathcal{A}$, is given by

$$S(\mathbf{n}, \mathbf{s}) = \sum_{k=1}^{L} \mathbf{W}(n_k, k), \tag{6.2}$$

Table 6.2: Normalized and transformed weight matrix of the eukaryotic TATA-box.

position k	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
$\mathbf{W}(A,k)$	-3.05	0.00	-4.61	0.00	0.00	0.00	0.00	-0.01
$\mathbf{W}(C,k)$	-2.06	-5.22	-3.49	-5.17	-4.63	-4.12	-3.74	-1.13
$\mathbf{W}(G,k)$	-2.74	-4.28	-4.61	-3.77	-4.73	-2.65	-1.50	0.00
$\mathbf{W}(T,k)$	0.00	-2.28	0.00	-2.34	-0.52	-3.65	-0.37	-1.40

where L denotes the length of the motif. High scores $S(\mathbf{n},\mathbf{s})$ indicate a high similarity between the two sequences.

▷ **Example 6.1**

The homology score between the DNA sequence $\mathbf{n}$ = CAAATAAA and the TATA-box consensus sequence $\mathbf{s}$ = TATAAAAG is given by

$$S(\mathbf{n},\mathbf{s}) = -2.06 + 0.00 - 4.61 + 0.00 - 0.52 + 0.00 + 0.00 - 0.01 = -7.2.$$

Since this constitutes a rather high score, the sequence CAAATAAA has – as expected – a high similarity to the TATA-box consensus sequence TATAAAAG.

◁

In the past, weight matrices have been applied for the detection of protein binding sites, e.g. yet unknown promoters or transcription factor binding sites (see e.g. [Sto00, SH89]). For this purpose, the weight matrix is shifted in single-nucleotide steps over the DNA and the matrix score is calculated between the current subsequence $\mathbf{n}(i) = \{n_i, \ldots, n_{i+L-1}\}$ and the motif for each position i:

$$S(\mathbf{n}(i),\mathbf{s}) = \sum_{k=1}^{L} \mathbf{W}(n_{i+k-1},k). \tag{6.3}$$

Positions with high scores constitute candidate sites for the existence of binding sites. This, however, showed to have a highly limited specificity [Sto00]. For this reason, the weight matrix of the TATA-box is in the following not applied for detection of individual promoters but to an aligned set $\mathcal{S}$ of known promoters (see also Section 5.1.3). Then, the average matrix score over all N aligned sequences is calculated for each position:

$$\overline{S}(i) = \frac{1}{N}\sum_{l=1}^{N} S(\mathbf{n}_l(i),\mathbf{s}), \tag{6.4}$$

where $S(\mathbf{n}_l(i),\mathbf{s})$ denotes the matrix score at position i of the l^{th} sequence. High scores indicate sliding windows of the dataset that possess a high sequence similarity to the TATA-box consensus sequence.

Results

Figure 6.3 depicts the resulting average TATA-box matrix score obtained for the two sets of aligned promoters (human and arthropod) from the EPD database [SIoB07] ($N = 1871$ and $N = 1996$, respectively). It shows a slow decrease of the matrix scores towards the human promoter site and two peaks shortly before the transcription start site (position $+1$). Since the TATA-box is highly AT-rich (consensus sequence: $p_{\mathbf{s}}(A) = 0.625$, $p_{\mathbf{s}}(C) = 0$, $p_{\mathbf{s}}(G) = 0.125$, $p_{\mathbf{s}}(T) = 0.25$), the matrix score is related to the AT-content. Hence, the low values around the human promoters indicate a high GC-content. A different behavior of matrix scores is observed around the arthropod promoters, where the surrounding indicates a high AT-content. The characteristics are interpreted in Section 6.3.

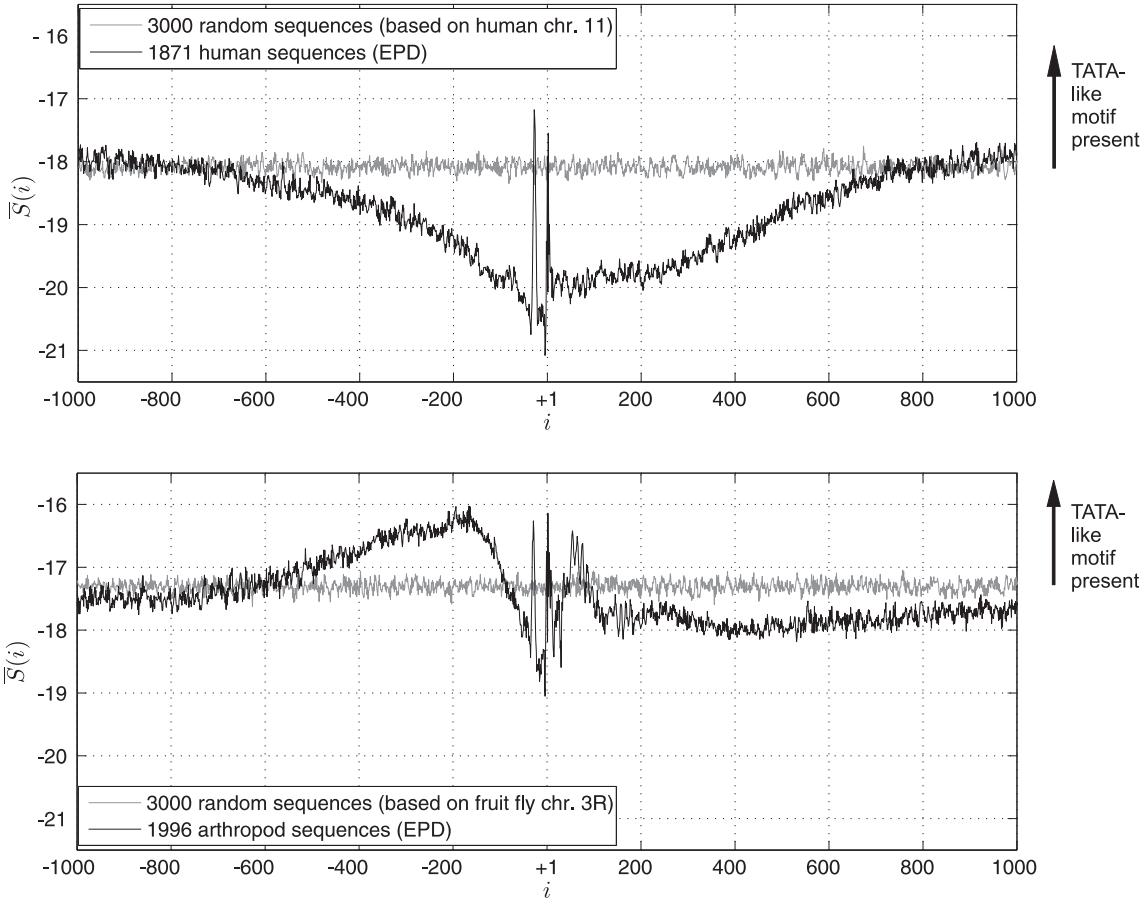


Figure 6.3: Homology score to the TATA-box around human EPD promoters (top) and arthropod EPD promoters (bottom).

6.2.2 Mutual information

In the previous section, the surrounding of the transcription start site was analyzed using a weight matrix. In this section, an alternative measure based on the mutual information

between the TATA-box and the aligned dataset of promoter sequences is derived. The mutual information $I(X;Y)$ between two variables X and Y is defined by Eq. (4.17) (see Section 4.2.3). The following convention applies for calculating the mutual information:

$$0 \log \frac{0}{0} = 0. \tag{6.5}$$

Objective

Mutual information is a measure of the statistical dependence between two random variables. In its application to DNA sequences, it has a major advantage over weight matrices: While those only detect stretches exhibiting the same structure and nucleotide sequence, mutual information does not depend on the nucleotide sequence. For example, the DNA sequences AAGAAG and TTCTTC yield a low homology score of the weight matrix, but they yield high values of mutual information due to their highly similar sequence structure. For this reason, mutual information has already been applied for detecting dependencies between different parts of eukaryotic genomes (see [AKL+07]). In the following, mutual information is adapted for its application to a large set of promoter sequences aligned to the transcription start site. It is derived such that it detects positions of the dataset with statistical dependencies to the TATA-box consensus sequence.

Application for promoter analysis

In the following, an estimation of the mutual information $I(\mathbf{s};\mathcal{S}(i))$ between a short sequence motif $\mathbf{s}$ and sliding windows of a set of N aligned DNA sequences $\mathcal{S}$ of length N_d is derived. At each position i of the dataset, the average mutual information between the sought motif and the sequence subset in the sliding window is given by

$$I(\mathbf{s};\mathcal{S}(i)) = \sum_{n_x \in \mathcal{A}} \sum_{n_y \in \mathcal{A}} \hat{p}_{\mathbf{s},\mathcal{S}(i)}(n_x, n_y) \mathrm{ld} \frac{\hat{p}_{\mathbf{s},\mathcal{S}(i)}(n_x, n_y)}{\hat{p}_{\mathbf{s}}(n_x)\hat{p}_{\mathcal{S}(i)}(n_y)}, \tag{6.6}$$

where $\mathcal{S}(i)$ references a sliding window of the dataset, which begins at position i and is as long as the sought motif (i.e. L nucleotides). The empirical probability mass function $\hat{p}_{\mathbf{s}}(n_x)$ is derived from the consensus sequence of the short motif, while $\hat{p}_{\mathcal{S}(i)}(n_y)$ and $\hat{p}_{\mathbf{s},\mathcal{S}(i)}(n_x, n_y)$ are calculated for each position i from the given sequence dataset (see Example 6.2). The former refers to the occurrence of nucleotide n_y in the sliding window, the latter denotes the joint occurrence of nucleotide n_x in the sequence motif and nucleotide n_y in the sliding window at position i. Since $\hat{p}_{\mathcal{S}(i)}(n_y)$ and $\hat{p}_{\mathbf{s},\mathcal{S}(i)}(n_x, n_y)$ are derived from the given dataset, the calculation of $I(\mathbf{s};\mathcal{S}(i))$ constitutes an estimate to the exact mutual information, whose accuracy depends on the availability of sequence data.

▷ Example 6.2

Consider the following set $\mathcal{S}$ of $N = 5$ DNA sequences of length $N_d = 20$:

C	G	T	A	A	C	C	G	G	T	T	A	A	G	A	C	C	A	G	T
T	G	G	A	C	A	C	A	T	A	C	T	G	C	A	T	C	A	T	G
T	A	C	G	A	T	A	C	A	T	A	T	T	A	A	A	C	G	A	T
C	A	G	T	A	C	G	A	T	A	C	A	G	C	A	T	A	C	G	A
A	A	G	G	T	A	C	T	A	C	T	A	T	G	C	G	G	A	T	G

The probability mass function $\hat{p}_{\mathbf{s}}(n_x)$ is given by the base composition of the consensus sequence of the TATA-box ($L = 8$, sequence TATAAAAG) and does not depend on the position i:

$$\hat{p}_{\mathbf{s}}(n_x = A) = \frac{5}{8},\ \ \hat{p}_{\mathbf{s}}(n_x = C) = \frac{0}{8},\ \ \hat{p}_{\mathbf{s}}(n_x = G) = \frac{1}{8},\ \ \hat{p}_{\mathbf{s}}(n_x = T) = \frac{2}{8},$$

For calculating the mutual information between the TATA-box consensus sequence and the first position of the dataset, the sliding window $\mathcal{S}(1)$ of the dataset is considered, which is $L = 8$ positions long and starts at position $i = 1$:

C	G	T	A	A	C	C	G
T	G	G	A	C	A	C	A
T	A	C	G	A	T	A	C
C	A	G	T	A	C	G	A
A	A	G	G	T	A	C	T

The probability mass function $\hat{p}_{\mathcal{S}(1)}(n_y)$ of the sliding window is then given by (e.g. 14 A's in 40 positions)

$$\hat{p}_{\mathcal{S}(1)}(n_y = A) = \frac{14}{40},\ \ \hat{p}_{\mathcal{S}(1)}(n_y = C) = \frac{10}{40},\ \ \hat{p}_{\mathcal{S}(1)}(n_y = G) = \frac{9}{40},\ \ \hat{p}_{\mathcal{S}(1)}(n_y = T) = \frac{7}{40},$$

Finally, the joint probability mass function $\hat{p}_{\mathbf{s},\mathcal{S}(1)}(n_x, n_y)$ is determined for the alignment of the TATA-box consensus sequence and the sliding window $\mathcal{S}(1)$ from the number of positions where nucleotide n_x in the motif co-occurs with nucleotide n_y in the sliding window:

C	G	T	A	A	C	C	G
T	G	G	A	C	A	C	A
T	A	C	G	A	T	A	C
C	A	G	T	A	C	G	A
A	A	G	G	T	A	C	T
T	A	T	A	A	A	A	G

This yields the following values of $\hat{p}_{\mathbf{s},\mathcal{S}(1)}(n_x, n_y)$:

$$\begin{aligned}
&\hat{p}_{\mathbf{s},\mathcal{S}(1)}(n_x = A, n_y = A) = \frac{11}{40}, && \ldots && \hat{p}_{\mathbf{s},\mathcal{S}(1)}(n_x = A, n_y = T) = \frac{3}{40},\\
&\hat{p}_{\mathbf{s},\mathcal{S}(1)}(n_x = C, n_y = A) = \frac{0}{40}, && \ldots && \hat{p}_{\mathbf{s},\mathcal{S}(1)}(n_x = C, n_y = T) = \frac{0}{40},\\
&\hat{p}_{\mathbf{s},\mathcal{S}(1)}(n_x = G, n_y = A) = \frac{2}{40}, && \ldots && \hat{p}_{\mathbf{s},\mathcal{S}(1)}(n_x = G, n_y = T) = \frac{1}{40},\\
&\hat{p}_{\mathbf{s},\mathcal{S}(1)}(n_x = T, n_y = A) = \frac{1}{40}, && \ldots && \hat{p}_{\mathbf{s},\mathcal{S}(1)}(n_x = T, n_y = T) = \frac{3}{40}.
\end{aligned}$$

The mutual information $I(\mathbf{s}; \mathcal{S}(1))$ is then calculated using Eq. (6.6) and yields 0.086 bits. ◁

Figure 6.4 depicts the mutual information between the TATA-box consensus sequence and the two aligned datasets (human, arthropod) downloaded from the EPD database. High values of mutual information indicate a statistical dependence between the currently considered sliding window $\mathcal{S}(i)$ and the TATA-box consensus sequence. For both datasets, a maximum is observed at the position of the TATA-box ($i = -30$), however, the output is generally rather noisy, which is due to the short length of the TATA-box: $\hat{p}_{\mathbf{s}}(n_x)$ and $\hat{p}_{\mathbf{s},\mathcal{S}(i)}(n_x, n_y)$ are calculated from a sequence / alignment of length $L = 8$, which does not yield an accurate estimate of the probability mass functions. Since the most prominent sequence feature of the TATA-box is its alternations between the nucleotides A and T (which is a required feature for detection by TFIID [JLB+96]), the mutual information values can only be considered as a noisy measure of short alternating sequences of two bases. A thorough interpretation is presented in Section 6.3. More detailed views of the exact promoter site are depicted in Section 6.3.3.

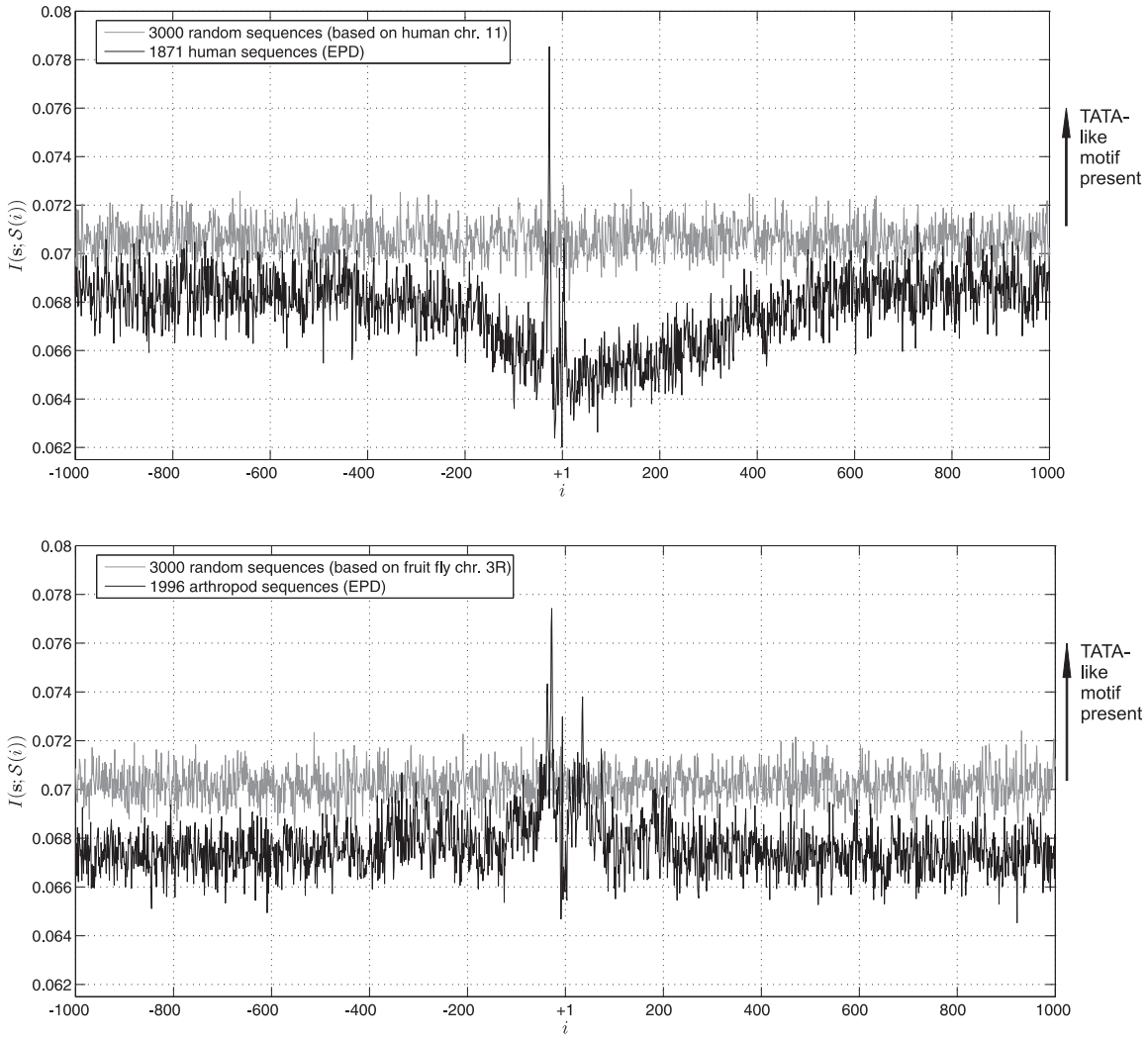


Figure 6.4: Mutual information between two EPD datasets (top: human promoters, bottom: arthropod promoters) and the consensus sequence of the TATA-box.

6.2.3 Kullback-Leibler divergence

The Kullback-Leibler divergence (or relative entropy) between two probability mass functions $p_X(x)$ and $q_X(x)$ is defined by Eq. (4.15) (see Section 4.2.2). The following conventions apply for the calculation of the Kullback-Leibler divergence:

$$0 \log \frac{0}{q_X(x)} = 0 \quad \forall\, q_X(x), \tag{6.7}$$

$$p_X(x) \log \frac{p_X(x)}{0} = \infty \quad \forall\, p_X(x) \neq 0. \tag{6.8}$$

Note that the Kullback-Leibler divergence is non-negative and zero for $p_X(x) = q_X(x)$.

Objective

The Kullback-Leibler divergence is a measure for the similarity of two probability distributions. Since the positional nucleotide distribution of the TATA-box is given from the weight matrix in Table 6.1, it can be used as the background distribution $q_X(x)$, and the Kullback-Leibler divergence is investigated between this background distribution and the observed distribution in the dataset of promoter sequences. A low Kullback-Leibler divergence then indicates the presence of stretches of the aligned data set with a similar positional nucleotide distribution as the TATA-box.

Application for promoter analysis

In the case of promoter search, the objective is to detect ranges of the given set of aligned DNA sequences that have a similar distribution and arrangement of nucleotides as the TATA-box. For this case, the empirical Kullback-Leibler divergence is given by

$$D(\hat{p}_{\mathbf{s}}(n) \parallel \hat{p}_{\mathcal{S}(i)}(n)) = \sum_{k=1}^{L} \sum_{n \in \mathcal{A}} \hat{p}_{\mathbf{s}}(n,k) \mathrm{ld} \frac{\hat{p}_{\mathbf{s}}(n,k)}{\hat{p}_{\mathcal{S}(i)}(n,k)}, \tag{6.9}$$

where $\hat{p}_{\mathbf{s}}(n,k)$ refers to the nucleotide occurrences at position k of the sequence motif as given from the weight matrix and $\hat{p}_{\mathcal{S}(i)}(n,k)$ to the nucleotide distribution at position k of the sliding window $\mathcal{S}(i)$ which is situated at position i of the aligned dataset.

▷ **Example 6.3**

Consider again the set of five DNA sequences of length $N_d = 20$ presented in Example 6.2.

In contrast to the approach based on mutual information, the probability mass function $\hat{p}_{\mathbf{s}}(n,k)$ for calculation of the Kullback-Leibler divergence is given from the unnormalized weight matrix of the TATA-box (see Table 6.1):

$$\hat{p}_{\mathbf{s}}(n=A,k=1)=0.04,\quad \hat{p}_{\mathbf{s}}(n=A,k=2)=0.91,\quad \ldots\quad \hat{p}_{\mathbf{s}}(n=A,k=8)=0.40,$$
$$\hat{p}_{\mathbf{s}}(n=C,k=1)=0.12,\quad \hat{p}_{\mathbf{s}}(n=C,k=2)=0.00,\quad \ldots\quad \hat{p}_{\mathbf{s}}(n=C,k=8)=0.11,$$
$$\hat{p}_{\mathbf{s}}(n=G,k=1)=0.04,\quad \hat{p}_{\mathbf{s}}(n=G,k=2)=0.00,\quad \ldots\quad \hat{p}_{\mathbf{s}}(n=G,k=8)=0.40,$$
$$\hat{p}_{\mathbf{s}}(n=T,k=1)=0.80,\quad \hat{p}_{\mathbf{s}}(n=T,k=2)=0.09,\quad \ldots\quad \hat{p}_{\mathbf{s}}(n=T,k=8)=0.09.$$

This defines the background distribution that is searched for in the given dataset to detect the TATA-box. The actual distribution $\hat{p}_{\mathcal{S}(1)}(n,k)$ at position $i=1$ in the dataset of aligned sequences is exemplarily assumed to be:

$$\hat{p}_{\mathcal{S}(1)}(n=A,k=1)=0.19,\quad \hat{p}_{\mathcal{S}(1)}(n=A,k=2)=0.08,\quad \ldots\quad \hat{p}_{\mathcal{S}(1)}(n=A,k=8)=0.19,$$
$$\hat{p}_{\mathcal{S}(1)}(n=C,k=1)=0.35,\quad \hat{p}_{\mathcal{S}(1)}(n=C,k=2)=0.47,\quad \ldots\quad \hat{p}_{\mathcal{S}(1)}(n=C,k=8)=0.14,$$
$$\hat{p}_{\mathcal{S}(1)}(n=G,k=1)=0.43,\quad \hat{p}_{\mathcal{S}(1)}(n=G,k=2)=0.23,\quad \ldots\quad \hat{p}_{\mathcal{S}(1)}(n=G,k=8)=0.45,$$
$$\hat{p}_{\mathcal{S}(1)}(n=T,k=1)=0.03,\quad \hat{p}_{\mathcal{S}(1)}(n=T,k=2)=0.22,\quad \ldots\quad \hat{p}_{\mathcal{S}(1)}(n=T,k=8)=0.22.$$

The Kullback-Leibler divergence is subsequently calculated for each position i based on the probability mass functions $\hat{p}_{\mathbf{s}}(n,k)$ and $\hat{p}_{\mathcal{S}(i)}(n,k)$ using Eq. (6.9). For position $i=1$, this yields

$$\begin{aligned}
D(\hat{p}_{\mathbf{s}}(n,k=1)\parallel\hat{p}_{\mathcal{S}(1)}(n,k=1)) &= \underbrace{(-0.09)}_{x=\mathrm{A}}+\underbrace{(-0.19)}_{x=\mathrm{C}}+\underbrace{(-0.14)}_{x=\mathrm{G}}+\underbrace{(+3.79)}_{x=\mathrm{T}}=3.37\\
&\vdots\\
D(\hat{p}_{\mathbf{s}}(n,k=8)\parallel\hat{p}_{\mathcal{S}(1)}(n,k=8)) &= \underbrace{(+0.43)}_{x=\mathrm{A}}+\underbrace{(-0.04)}_{x=\mathrm{C}}+\underbrace{(-0.07)}_{x=\mathrm{G}}+\underbrace{(-0.12)}_{x=\mathrm{T}}=0.20\\
\Rightarrow D(\hat{p}_{\mathbf{s}}(n)\parallel\hat{p}_{\mathcal{S}(i)}(n)) &= \sum_{k=1}^{L} D(\hat{p}_{\mathbf{s}}(n,k)\parallel\hat{p}_{\mathcal{S}(i)}(n,k))
\end{aligned}$$

◁

Figure 6.5 shows the result of applying the approach based on the Kullback-Leibler divergence to the two promoter datasets from the EPD database. In contrast to the weight matrix score and the mutual information, low values of the Kullback-Leibler divergence indicate a strong similarity between the probability mass function of the current sliding window $\mathcal{S}(i)$ and that of the TATA-box. Likewise, high values indicate a strong dissimilarity to the TATA-box (even more dissimilar than random sequences). A clear trend is observed for both the human promoters (top) and the arthropod promoters (bottom): While the scores steadily increase around the former, they are slightly below the surrounding around the latter. A minimum at the TATA-box (around position $i=-30$) is observed for both datasets. A thorough interpretation is presented in Section 6.3.

6.3 Results and interpretation

In the last section, mutual information and the Kullback-Leibler divergence were modified for the application to promoter sequences and compared to a weight matrix of the TATA-

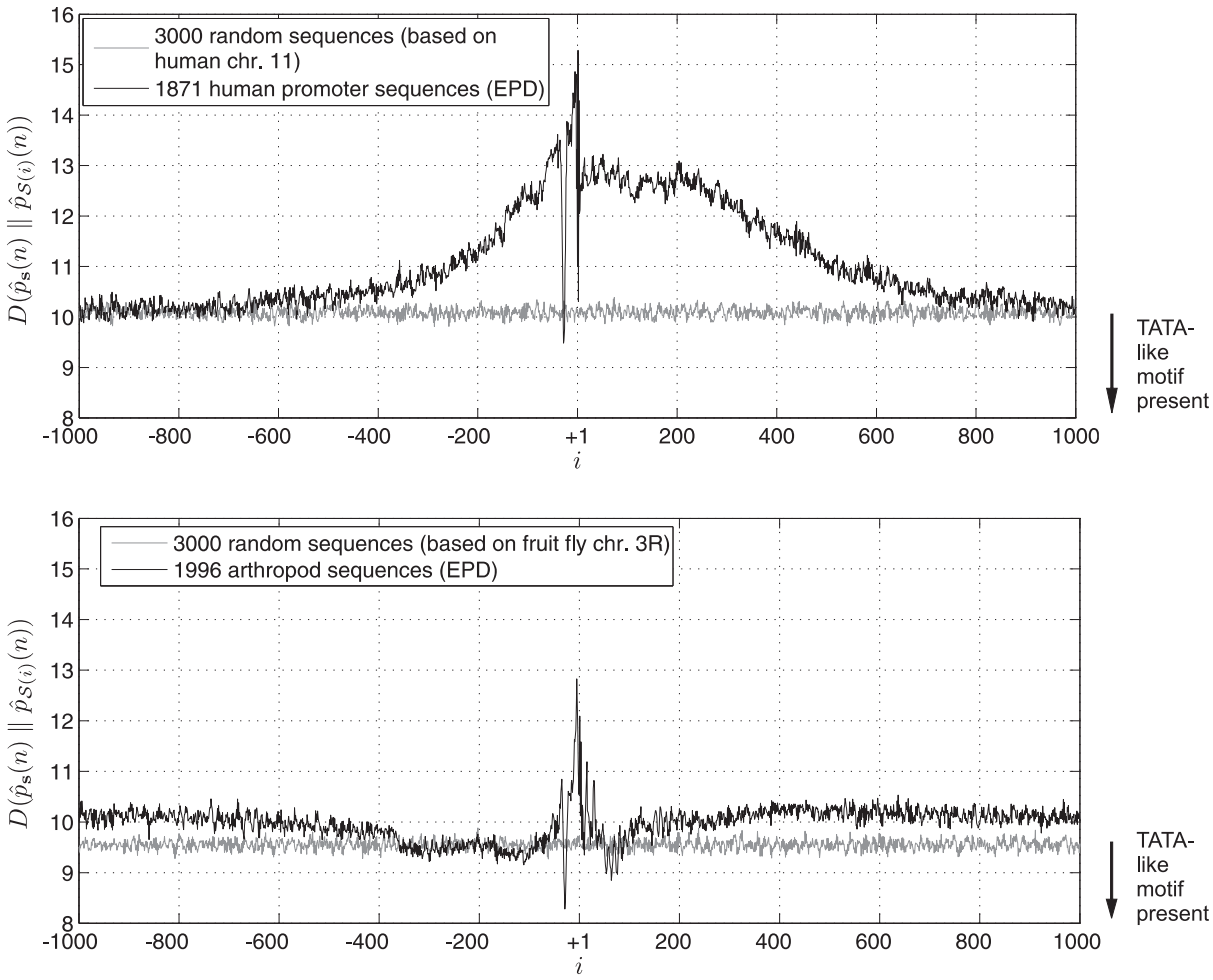


Figure 6.5: Kullback-Leibler divergence between the nucleotide distribution of EPD datasets (top: human promoters, bottom: arthropod promoters) and the nucleotide distribution of the TATA-box.

box. The results showed a characteristic behavior of the weight matrix scores and the Kullback-Leibler scores in the 3000 bp surrounding the transcription start site. In contrast to that, the mutual information scores were too noisy to expose a characteristic behavior. In the following sections, these results are interpreted with respect to their influence on transcription initiation. An important background information for the interpretation of the results is the nucleotide composition in the promoter surrounding. This can be found in Appendix C.3.1 for the human EPD dataset and in Appendix C.3.2 for the arthropod EPD dataset (see Appendix C.1.2 for more information about the datasets).

6.3.1 Comparison of the information theoretic measures

The application of mutual information yielded highly noisy results. This is due to the short length of the TATA-box ($L = 8$) that is used to estimate the sought nucleotide

distribution: the probability mass function of a random variable with $|\mathcal{A}| = 4$ is estimated over only $L = 8$ samples. Since the main sequence characteristic of the TATA-box is its alternations between A and T, the mutual information score simply detects short alternations between two arbitrary bases. In [AKL+07], H. M. Aktulga applied a similar mutual information estimate to search for mutually dependent sequences in a maize gene. They derived the probability mass function of a sequence of length 369 and still obtained noisy results. In the approach presented in Section 6.2.2, the noise is partially diminished by estimating the joint probability mass function over all sequences in the aligned dataset $\mathcal{S}$, which however can by far not account for the insufficient motif length. B. Goebel showed in [GDHM05] that that even for very large samples sizes, e.g. N=10000, mutual information estimates exhibit a considerable error range. It is thus preferable to solely test the results for significance. For a sample size of $N = 8$, the significance level is 1.5256 bits[1], i.e. all values in Figure 6.4 are – as expected – far from being significant.

Contrary to that, the Kullback-Leibler score showed a clear trend around the transcription start site – although its nucleotide distribution is also obtained from the TATA-box. However, it is not estimated over only 8 positions of one sequence but derived for each position over a large set of aligned sequences: The probability mass function of the random variable with $|\mathcal{A}| = 4$ is here estimated over $N = 1871$ samples in case of the human dataset and over $N = 1996$ samples in case of the arthropod dataset. The nucleotide distribution of the TATA-box is given by the weight matrix, which was derived from $N = 502$ sequences [Buc90]. The estimation of a probability mass function is in essence a Bernoulli trial, and the resulting estimates (the probability masses) follow a scaled multinomial distribution. The variance of each estimate is thus given by $\sigma^2 = p(n) \cdot (1 - p(n))/N$, where N is the total number of samples and $p(n)$ is the (true) probability of the n^{th} event (here: $n \in \{A, C, G, T\}$) [Fel68]. In the worst case of having to estimate a true probability of $p(n) = 0.5$, a variance of only $\sigma^2 = 5.0 \cdot 10^{-4}$ is reached for $N = 502$ and of only $\sigma^2 = 1.3 \cdot 10^{-4}$ for $N = 1871$.

6.3.2 Promoter surrounding

Impact of CpG islands around human promoters

Figure 6.3 (top) shows a continuous decrease of weight matrix scores around the transcription start sites of the human EPD dataset. Since the consensus sequence of the TATA-box is highly AT-rich (7 of the 8 positions are A or T), the weight matrix score correlates inversely with the GC-content (low scores indicate GC-rich sequences). It was mentioned in Section 6.1.4 that many eukaryotic promoters are surrounded by CpG islands, accumulations of GC-dinucleotides. Especially the human genome is reported to be rich in CpG islands [Lew07]. Accordingly, Figure C.1 (Appendix C) shows a high GC-content around the promoters of the human EPD dataset. For this reason, the continuous decrease of weight matrix scores observed in Figure 6.3 (top) is expected to stem from the

[1]Calculated as the inverse of the gamma cumulative distribution function, see [GDHM05]

CpG islands. While the approach based on mutual information does not clearly detect the CpG islands (Figure 6.4, top), the approach based on Kullback-Leibler divergence exhibits a significant increase around the human promoters (Figure 6.5, top). The latter is expected if recalling that the Kullback-Leibler divergence measures the dissimilarity between the background nucleotide distribution and the nucleotide distribution of the TATA-box, which is highly AT-rich.

Impact of AT accumulation around arthropod promoters

In contrast to human promoters, the weight matrix scores around arthropod promoters increase before the transcription start site (see Figure 6.3 (bottom)). It can be seen from Figure C.2 (Appendix C) that the according region exhibits a high AT-content, which explains the observed increase of scores. At the same time, the low GC-content implies the absence of CpG-islands. In fact, the arthropod EPD dataset mainly consists of sequences of the fruitfly (*Drosophila melanogaster*), whose non-genic regions are reported to be generally AT-rich [DBSH07]. The approach based on mutual information shows no significant deviation from its values on random sequences (see Figure 6.4, bottom). This indicates that – though the nucleotide composition changes before the promoter – the sequence structure with respect to nucleotide alternations does not. The approach based on Kullback-Leibler divergence exhibits a weak but noticeable decrease before the promoter (see Figure 6.5, bottom), indicating that the nucleotide composition slowly approaches that of the TATA consensus sequence.

6.3.3 Promoter site

Figure 6.6 shows the values of the approaches based on a weight matrix (top), on mutual information (middle) and on the Kullback-Leibler divergence (bottom) in the direct surrounding of the transcription start site (position $i = +1$). The results are depicted for human promoters (left) and arthropod promoters (right). It can be seen from Figure 6.6 (top) that the weight matrix yields a clear peak at position $i = -30$ (the position of the TATA-box), both on human and on arthropod sequences. Moreover, it exhibits a maximum at the transcription start site, which occurs due to the high AT-content at that position (see Figure C.1 and Figure C.2, Appendix C). The approach based on mutual information (Figure 6.6, middle) yields a less clear distinction of the TATA-box and the transcription start site, especially for arthropod promoters. However, since the output is generally noisy due to the inaccurate estimation of the probability mass functions (see Section 6.2.2), these results are expected. The approach based on the Kullback-Leibler divergence exhibits the most distinct picture (see Figure 6.6, bottom). In case of human promoters (left), a clear minimum is visible at the TATA-box and at the transcription start site. In contrast to that, the arthropod promoters (right) exhibit a minimum at the TATA-box but a maximum at the transcription start site. This fact indicates that the position-specific nucleotide content is dissimilar to that of the TATA-box – despite the high overall AT-content of both sequences.

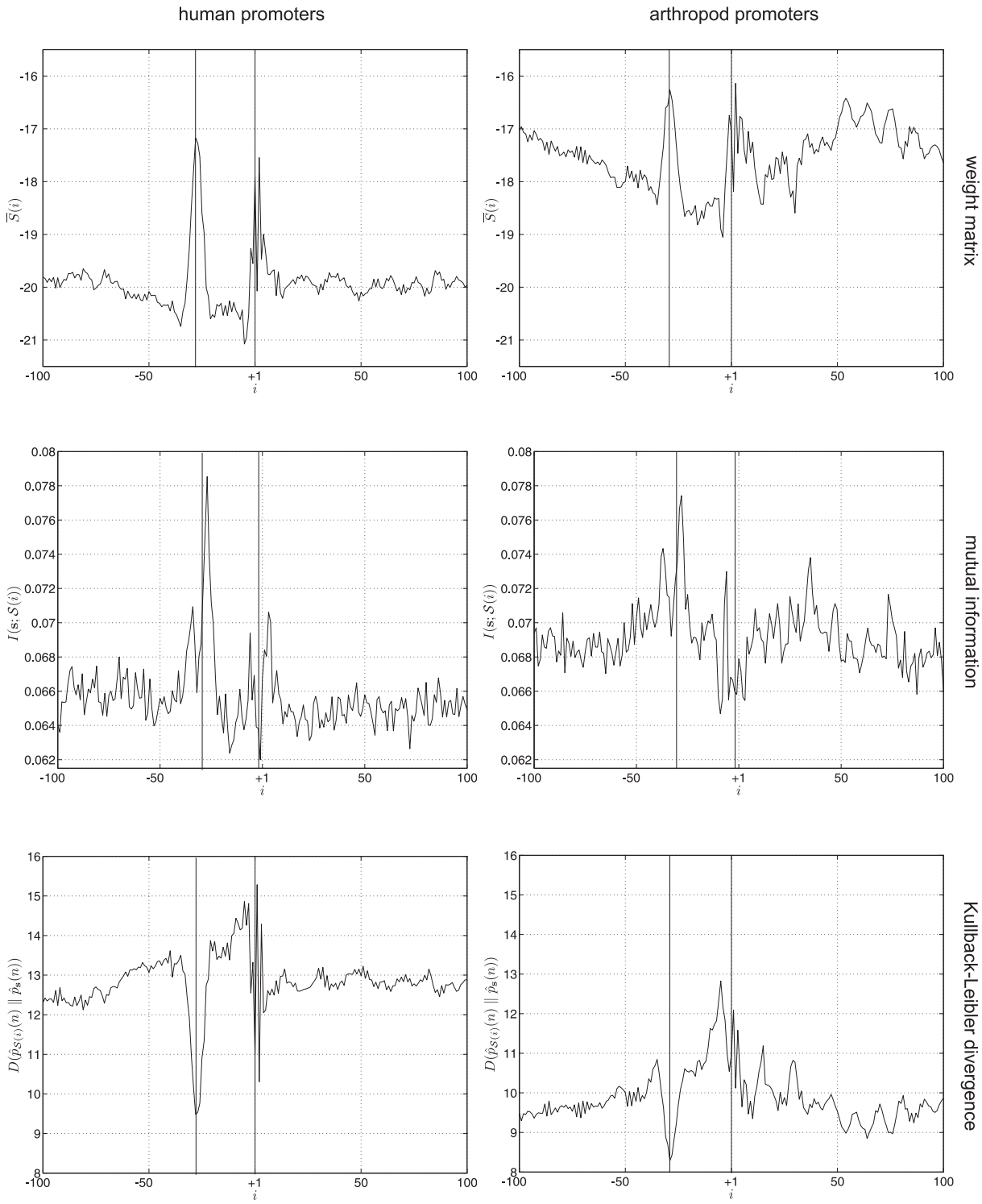


Figure 6.6: Detailed view of scores around the promoter site in human (left) and arthropod (right) dataset. Top: weight matrix, middle: mutual information, bottom: Kullback-Leibler divergence. The TATA-box (around position $i = -30$) and the transcription start site (position $i = +1$) are marked by vertical lines.

6.4 Clustering of promoters

It was shown in Section 6.3.2 that the weight matrix score is best capable of detecting CpG islands and AT-rich sequences. Therefore, it is now applied to specific sets of promoters to gain a more detailed understanding of the promoter surrounding. The promoter strength – i.e. the promoter's ability to initiate transcription – depends among other factors on the sequence homology to the TATA consensus sequence, which is in turn measured by the weight matrix score. Since promoter strength is assumed to influence the initiation mechanism, the human promoters are subdivided into three groups based on their maximum weight matrix score in the region $i \in [-40; -23]$, where the TATA-box is frequently located. Figure 6.7 depicts the results for the three datasets: 450 promoters with a weak TATA-box (top), 900 promoters with an average TATA-box (middle) and 450 promoters with a strong TATA-box (bottom). The two maxima at the TATA-box and at the transcription start site are visible in all three datasets, however, the strong and the weak promoters additionally exhibit an interesting minimum / maximum around 450 nucleotides after the transcription start site. In the following subsections, three possible explanations for the additional peak are elucidated.

6.4.1 Transcription-factor binding site

A possible explanation for the additional minimum / maximum is that the binding site of a transcription factor is located at the respective position. Due to the dependency on the promoter strength, it is likely to constitute a way to compensate for promoters with a weak TATA-box or without a TATA-box. However, finding a maximum in weak promoters and a minimum in strong promoters implies that a transcription factor with a highly AT-rich binding site regulates weak promoters and that a transcription factor with a highly GC-rich sequence regulates strong promoters. Moreover, the peaks are rather broad, while transcription factor binding sites usually cover ranges of only 6 to 20 bp.

6.4.2 Nucleosome positioning

As detailed in Section 6.1.5, eukaryotic DNA is packaged into chromatin, in which the DNA is wrapped around histone proteins. The compact units of DNA and four histone proteins is referred to as a nucleosome. It is known that the positioning of nucleosomes is a sequence-specific process that depends on initiating sequences in the DNA [SFMC+06]. The positions of the nucleosomes determine the accessability of the promoter and can thus be related to promoter strength. G. C. Yuan et al. [YL08] derived the nucleosome occupancy along the genomes of yeast and human. They found a relationship between promoter strength and nucleosome occupancy. Evidence for a connection between the detected maximum / minimum and nucleosome positioning are the facts that it is broader than expected for transcription factor binding sites and that nucleosome positioning is reported to strongly depend on the AT-richness of the underlying sequence [YL08].

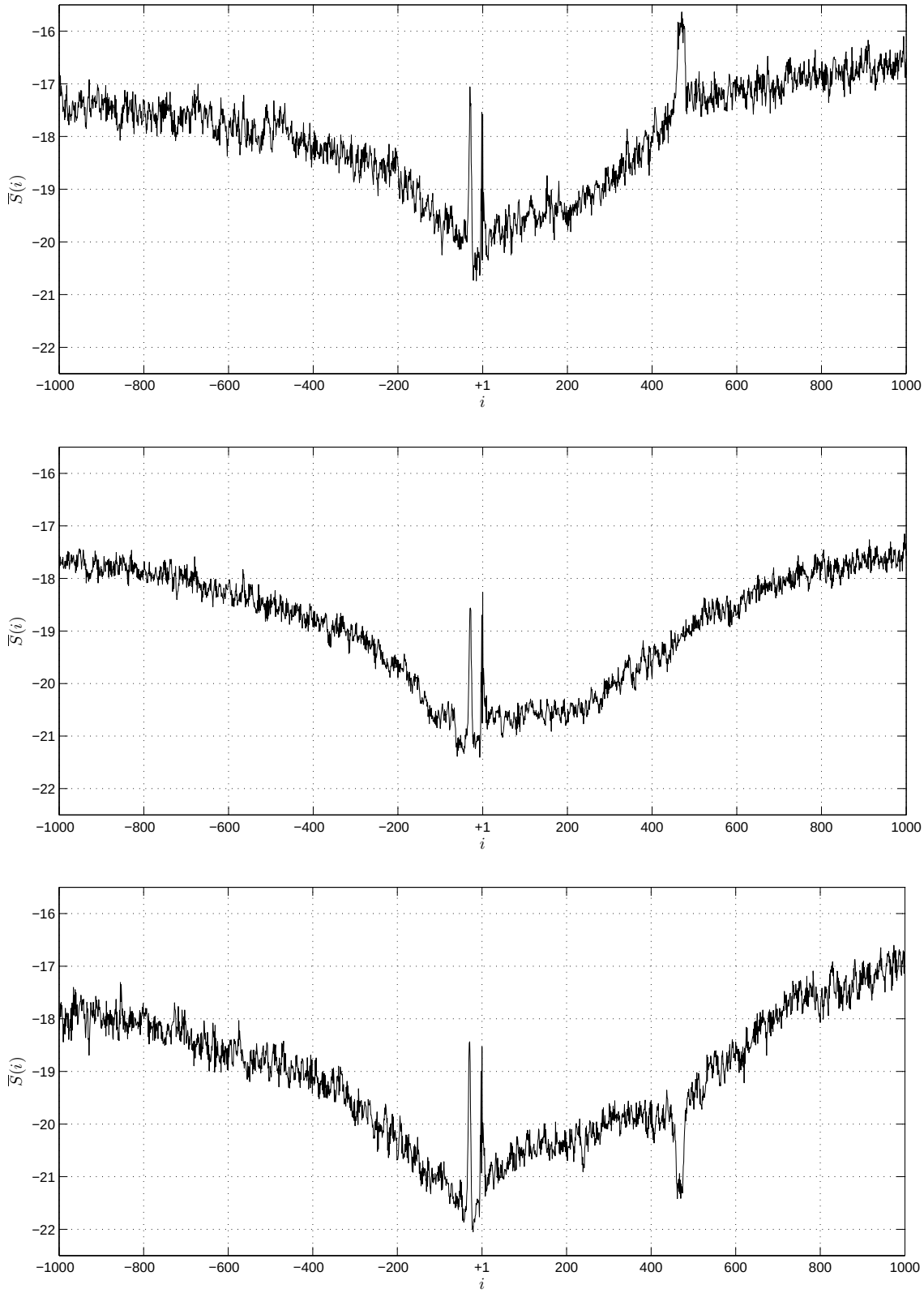


Figure 6.7: Weight matrix score for three subsets of human promoters. Top: weak TATA-box, middle: TATA-box of average strength, bottom: strong TATA-box.

6.4.3 DNA bendability

It is known that the bendability of DNA strongly depends on the underlying DNA sequence: While AT-rich sequences are easily bent (especially sequences with an alternation between A and T), GC-rich sequences are rigid [PKS+99, JLB+96]. During transcription initiation, the DNA is strongly bent by the RNA polymerase in order to form loops and thus attach to distant transcription factors. Since the detected maximum / minimum indicates an AT-rich region in weak promoters and a GC-rich sequences in strong promoters, it might constitute a way to allow or disallow the bending of DNA in order to position it correctly in the transcription initiation machinery and hereby compensate for a weak TATA-box.

6.5 Summary

This chapter dealt with the analysis of transcription initiation sites in higher organisms (eukaryotes) using measures from information theory. Mutual information and the Kullback-Leibler divergence were adapted for their application to large sets of aligned promoter sequences, more precisely as a similarity measure to the TATA-box. Additionally, a weight matrix – a standard tool from bioinformatics that assigns a score to a short sequence depending on its homology to a template motif – was applied for comparison. The following main results could be achieved:

- ▷ The application of the weight matrix to promoter datasets exposed huge differences between human sequences and arthropod sequences. This fact suggests that the mechanisms underlying transcription initiation in these species vary significantly. Moreover, a characteristic behavior of the score in a region of 2000 base pairs around the transcription start site was observed that indicates an influence of a wide surrounding on the detection of the promoter.

- ▷ Mutual information was adapted to measure the dependence between a sliding window of the dataset and the consensus sequence. The results were rather inconclusive and not even exhibited a strong detection signal at the TATA-box. This is due to the insufficient length of the TATA-box for accurately estimating the probability mass functions required for calculation of the mutual information.

- ▷ Subsequently, the Kullback-Leibler divergence was applied to the datasets to expose positions with very similar and very dissimilar nucleotide distributions compared to the TATA-box. Since its probability mass functions were not estimated over the positions of the TATA-box but over all sequences of the dataset, it showed to be a promising approach that exposed clear characteristics around the promoters. It revealed that the nucleotide distribution around human transcription start sites strongly differs from that at the TATA-box, which might serve synchronization since it amplifies the signal at the promoter site.

▷ Finally, the promoter sequences were subdivided according to their strength measured as the detection strength of the TATA-box. The promoter set with a very weak TATA-box exhibited an unexpected maximum, that with a very strong TATA-box exhibited an unexpected minimum, both at around 500 bp after the transcription start site. These showed to be most likely related to nucleosome positioning – a process involved in packaging the DNA into the compact form that builds the chromosomes.

7

Modeling Translation Initiation in Prokaryotes

Translation is the second step of gene expression and refers to the transformation of an mRNA into a protein. Not the complete mRNA is hereby translated but only the so-called coding sequence (see Chapter 3), which makes a process of frame synchronization necessary to detect the beginning of the start point marked by the Shine-Dalgarno sequence. In this chapter, this process is modeled using different codebooks derived from the underlying interactions between the ribosome and the mRNA.

In Section 7.1, four codebook models for the detection of the Shine-Dalgarno sequence are derived, the biological sync word of translation. Moreover, it includes the synchronization algorithm as well as a performance measure to rate the presented codebook models. Subsequently, three energy metrics based on binding energies between nucleotides are introduced in Section 7.2 which make the model output more biologically meaningful. Finally, Section 7.3 presents a mutational analysis where nucleotide mutations in the ribosome are simulated through changes of the codebook. This allows to rate the effect of mutations without conducting expensive biological experiments.

7.1 Detection of the Shine-Dalgarno sequence in *Escherichia coli*

The process of translation initiation corresponds to a frame synchronization where the ribosome – more precisely the 3'-end of its 16S rRNA subunit – detects the Shine-Dalgarno

sequence (SD) located shortly before the translation start site (TLS) (see Figure 7.1), which is marked by the start codon AUG.

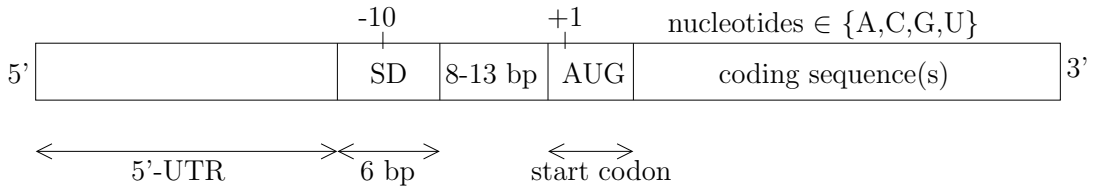


Figure 7.1: Structure of the initiator region of prokaryotic mRNA.

In this chapter, synchronization models based on codebooks are presented. The codebooks contain codewords of length L and can be seen as a list of template sync words that mimics allowed variations of the detected sequence. Several different codebooks are derived, and the detection strength of the Shine-Dalgarno sequence is measured as an indicator for the significance of the models. Since it is known that the detection of the Shine-Dalgarno sequence is based on an interaction between the mRNA and the 3'-end (13 bases) of the 16S rRNA (see Section 3.4.4), the codebooks are developed from the complement of these last 13 bases:

$$5' \quad \text{U A } \underbrace{\text{A G G A G G}}_{\text{SD sequence}} \text{ U G A U C} \quad 3'.$$

It can be seen that the Shine-Dalgarno sequence (AGGAGG) is part of these 13 bases. Figure 7.2 shows the 16S rRNA of *Escherichia coli* with its exposed 3'-end.

7.1.1 Synchronization algorithm

The detection of the Shine-Dalgarno sequence by the 16S rRNA is modeled as a process of frame synchronization. The synchronization process is executed in single nucleotide steps along the mRNA. At each step i, every codeword $\mathbf{s}_j, j \in [1; J]$, with $\mathbf{s}_j = \{s_{1_j}, \ldots, s_{L_j}\}$ of the codebook is compared to an mRNA sequence of length L [DGHM05]. That codeword with the minimum distance $d_{\min}$ with respect to a defined distance metric δ is chosen:

$$d_{\min}(i) = \min_{j \in [1;J]} \delta(\mathbf{d}(i), \mathbf{s}_j), \tag{7.1}$$

where $\mathbf{d}(i) = \{d_i, \ldots, d_{i+L-1}\}$, and $\mathbf{d}$ is the mRNA sequence which corresponds to the received data stream in technical synchronization processes. In the following sections, the distance δ is defined to be the Hamming distance $d_H(\mathbf{d}_{i,L}, \mathbf{s}_j)$. In Section 7.2, it is then extended to an energy metric based on the binding energies between 16S rRNA and mRNA to make the results biologically more meaningful.

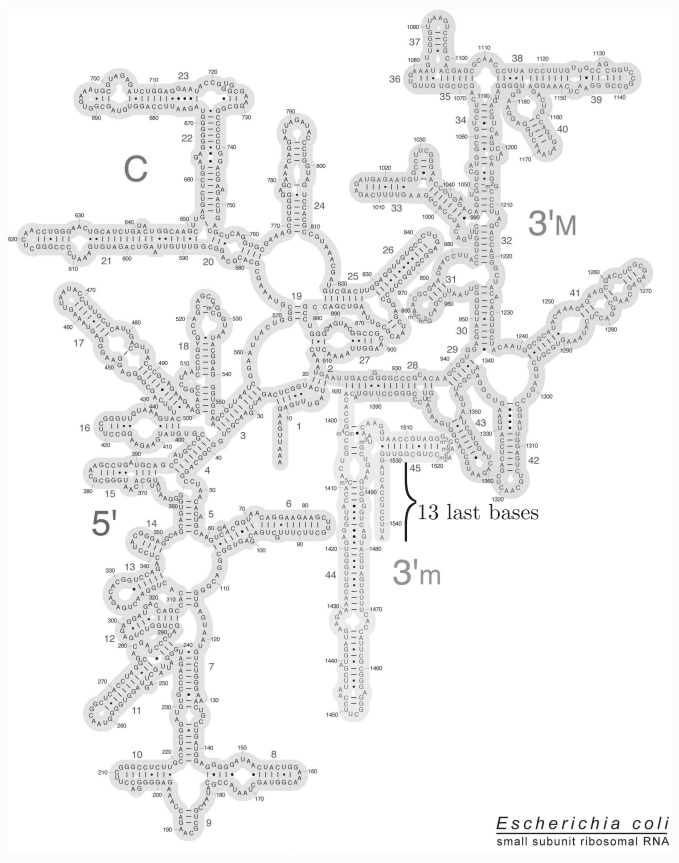


Figure 7.2: The 16S rRNA subunit of *Escherichia coli* [CfMBoR08].

▷ **Example 7.1**

Consider the mRNA subsequence $d_{i,L}$=AUGUCA and a codebook containing the 3 codewords $\mathbf{s}_1$ =AAGAAG, $\mathbf{s}_2$ =AACAAC and $\mathbf{s}_3$ =AUGAAA. In case of the Hamming distance, this results in the following value of the minimum distance $d_{\min}(i)$:

$$d_H(\mathbf{d}_{i,L}, \mathbf{s}_1) = 4, \quad d_H(\mathbf{d}_{i,L}, \mathbf{s}_2) = 5, \quad d_H(\mathbf{d}_{i,L}, \mathbf{s}_3) = 2 \quad \Rightarrow \quad d_{\min}(i) = d_h(\mathbf{d}_{i,L}, \mathbf{s}_3) = 2.$$

◁

7.1.2 Sequence data

The decoding algorithm is applied to the same set of 3194 *E. coli* mRNA sequences from the NCBI data base used in Section 4.2 [NCfBI08] (see Appendix C.1.3 for more information on sequence extraction). Similar to the handling of promoter sequences presented in Section 5.1.3, the sequences are not treated individually but average values are calculated over all available sequences. The mRNA sequences are again aligned to the start as well as to the stop codon and cut to a fixed length: the UTRs are truncated to 100 bp each and the middle part of the coding sequence is cut out leaving the first and the last 150 bp.

The sequence layout is presented in Figure 7.3.

5' UTR	start codon	CDS	CDS (ctd.)	stop codon	3' UTR
...	AUG	...	...	UAA	...
...	AUG	...	...	UAG	...
...	AUG	...	...	UAA	...
⋮		⋮	⋮		⋮
...	AUG	...	...	UGA	...
...	AUG	...	...	UAA	...
...	AUG	...	...	UAG	...
100 bp		150 bp	150 bp		100bp

Figure 7.3: Sequence layout of aligned mRNA sequences.

In addition to the 3194 mRNA sequences, non-translated sequences are used as control of the results. These are sequences from the complete genome of *E. coli* that contain a start codon (AUG) and a stop codon (UAA, UAG, UGA) but that are not part of the 3194 mRNAs or part of a gene.

7.1.3 Performance measure

To rate the quality of detection of the Shine-Dalgarno sequence obtained using models with different codeword length, the following performance ratio is proposed:

$$R = \frac{\Delta_{\text{act}}}{\Delta_{\text{max}}}, \quad 0 \leq R \leq 1, \tag{7.2}$$

where Δ_{act} refers to the actual depth of the minimum at the Shine-Dalgarno sequence (with respect to the average surrounding level), and Δ_{max} refers to the maximum possible depth of the minimum (i.e. a perfect detection of the Shine-Dalgarno sequence). Figure 7.4 depicts an arbitrary scenario illustrating the parameters needed for calculation of R for the case that the Hamming distance is used as the distance measure in Eq. (7.1), i.e. if the minimum achievable distance is zero.

7.1.4 13 bases complement model

In the first model, the 13 last bases of the 16S rRNA are used as a single codeword ($J = 1$) of length $L = 13$. The resulting codebook – which here consists of only one codeword – is shown in Table 7.1.

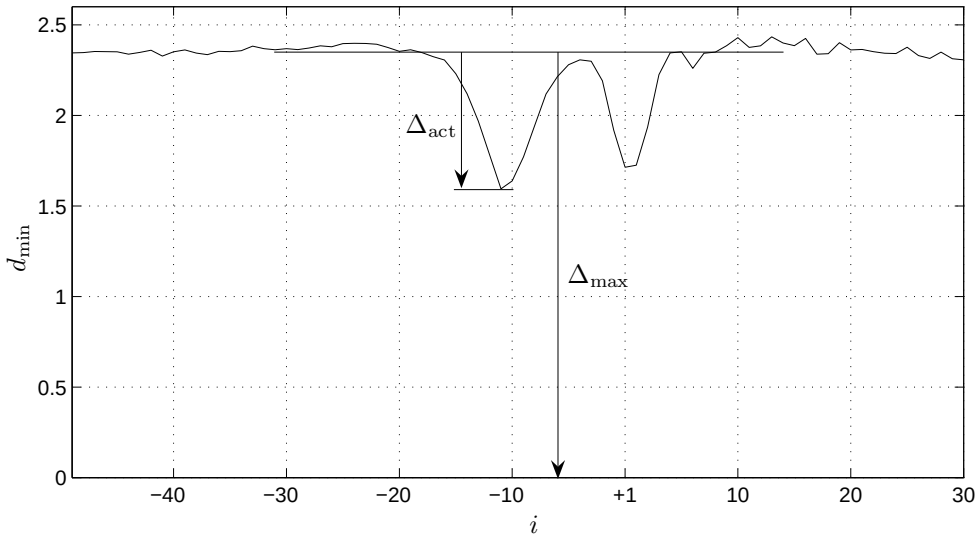


Figure 7.4: Illustration of the parameters used in the performance ratio.

Table 7.1: Codebook of the 13 bases complement model.

$\mathbf{s}_1$	U A A G G A G G U G A U C

Application of the 13 bases complement model to the above described mRNA sequences as well as the non-translated sequences yields the output depicted in Figure 7.5 [Gon04]. It can be seen that the Shine-Dalgarno sequence is detected, however, several side-peaks are exhibited for both types of sequences. The performance ratio yields a value of $R = 0.08$, i.e. only 8 % of the possible detection strength in terms of the depth of the minimum is achieved by the 13 bases complement model. This is due to the inflexibility of the model, since it contains only one codeword that does not allow any variations in the detected sequence. One interesting aspect though is the detection minimum at the start codon (position +1): it stems from the sequence similarity between subsequences of the 13 last bases and the start codon AUG.

7.1.5 Shine-Dalgarno sequence based model

The second model is constructed from the variability of the Shine-Dalgarno sequence. It is known that not only the consensus sequence AGGAGG is detected but also certain variations of it. In [SBRS01], R. Y. Shultzaberger et al. conducted an information theoretic analysis of Shine-Dalgarno sequences in *E. coli*. For this purpose, they aligned all available sequences to the first base and calculated the conservation of bases similar to the way presented in Section 4.2.2. They found out that bases 2 to 4 of the Shine-Dalgarno sequence are most conserved and that base 3 must be G. Surprisingly, the last

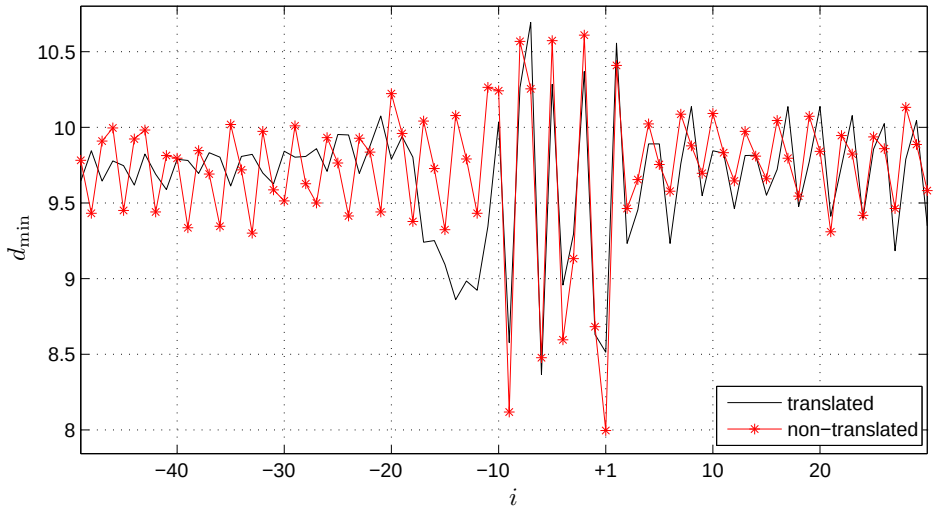


Figure 7.5: Decoding output obtained for the 13 bases complement model.

base neither showed a strong conservation nor a strong preference for one particular base, which indicates that it is not of great importance for the detection by the 16S rRNA. This raises the question why the Shine-Dalgarno sequence has been reported to comprise six nucleotides and not just five. Based on the preferences for bases at the positions in the Shine-Dalgarno sequence, a codebook with codeword length $L = 6$ and $J = 32$ codewords is created (see Table 7.2). For example, A and U are the most probable bases at position 1 of the Shine-Dalgarno sequence, thus, codewords with these two bases at the first position are created. For the last position - the one not showing strong preferences for a particular base - A and G are selected since they occur slightly more frequently than C and U.

Table 7.2: Codebook of the Shine-Dalgarno based model.

$\mathbf{s}_1$	A G G A G G	$\mathbf{s}_9$	A G G A A G	$\mathbf{s}_{17}$	A G G A G A	$\mathbf{s}_{25}$	A G G A A A
$\mathbf{s}_2$	U G G A G G	$\mathbf{s}_{10}$	U G G A A G	$\mathbf{s}_{18}$	U G G A G A	$\mathbf{s}_{26}$	U G G A A A
$\mathbf{s}_3$	A A G A G G	$\mathbf{s}_{11}$	A A G A A G	$\mathbf{s}_{19}$	A A G A G A	$\mathbf{s}_{27}$	A A G A A A
$\mathbf{s}_4$	U A G A G G	$\mathbf{s}_{12}$	U A G A A G	$\mathbf{s}_{20}$	U A G A G A	$\mathbf{s}_{28}$	U A G A A A
$\mathbf{s}_5$	A G G U G G	$\mathbf{s}_{13}$	A G G U A G	$\mathbf{s}_{21}$	A G G U G A	$\mathbf{s}_{29}$	A G G U A A
$\mathbf{s}_6$	U G G U G G	$\mathbf{s}_{14}$	U G G U A G	$\mathbf{s}_{22}$	U G G U G A	$\mathbf{s}_{30}$	U G G U A A
$\mathbf{s}_7$	A A G U G G	$\mathbf{s}_{15}$	A A G U A G	$\mathbf{s}_{23}$	A A G U G A	$\mathbf{s}_{31}$	A A G U A A
$\mathbf{s}_8$	U A G U G G	$\mathbf{s}_{16}$	U A G U A G	$\mathbf{s}_{24}$	U A G U G A	$\mathbf{s}_{32}$	U A G U A A

Figure 7.6 shows the decoding output of the Shine-Dalgarno sequence based model [Gon04]. Compared to the non-translated sequences, the mRNA sequences ex-

hibit a clear detection of the Shine-Dalgarno sequence. The performance ratio yields a value of $R = 0.25$, i.e. 25 % of the possible detection strength is achieved using the Shine-Dalgarno sequence based model.

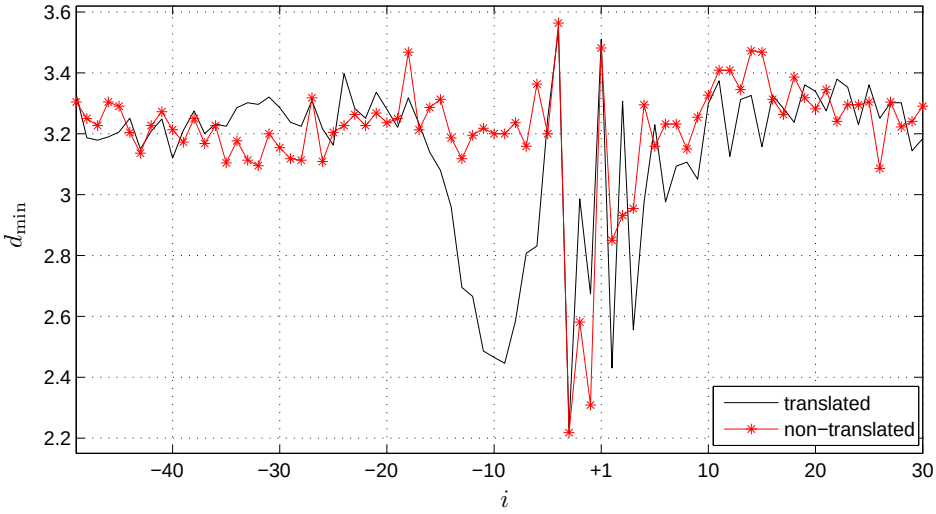


Figure 7.6: Decoding output obtained for the Shine-Dalgarno sequence based model.

7.1.6 May's parity check model

In 2004, E. E. May presented a coding theory model of translation initiation in *E. coli* based on the hypothesis that the mRNA is a block-encoded sequence [MVBR04, MVB06]. Consequently, the ribosome acts as a decoder that decodes the mRNA-sequence to detect the Shine-Dalgarno sequence. The block code was (arbitrarily) assumed to be a (5,2) parity check code, where the parity vectors are obtained from the subsequences of length 3 of the 13 bases complement (see Table 7.3). The nucleotides were mapped to I=0[1], A=1, G=2, C=3 and U=4 and operations are calculated modulo 5. The mapping was chosen in a way so that the modulo 5 sum of nucleotides that form hydrogen bonds is zero (i.e. A and T, C and G). The codewords were then chosen as all possible combinations of two information bits and three parity bits that satisfy the following parity condition:

$$\sum_{k=1}^{5} s_k = 0 \pmod 5, \tag{7.3}$$

If several combinations of an information vector and a parity vector satisfy Eq. (7.3), all of these are valid codewords. The resulting codebook is listed in Table 7.4.

[1]Inosine, a nucleotide rarely found in the mRNA but commonly occurring in tRNAs. Base pairs with A, C and U.

Table 7.3: Parity vectors of May's parity check code.

parity vector	sum	parity vector	sum
U A A	1	G G U	8 = 3 (mod 5)
A A G	4	G U G	8 = 3 (mod 5)
A G G	5 = 0 (mod 5)	U G A	7 = 2 (mod 5)
G G A	5 = 0 (mod 5)	G A U	7 = 2 (mod 5)
G A G	5 = 0 (mod 5)	A U C	8 = 3 (mod 5)

▷ **Example 7.2**

If the information sequence is given as (G U) = (2 4), the modulo 5 sum of the information vector is 1. To satisfy Eq. (7.3), the modulo 5 sum of the parity vector needs to be 4. According to Table 7.3, the according parity vector is (A A G). Thus, the resulting codeword is (2 4 1 1 2) = (G U A A G).

◁

Table 7.4: Codebook of May's parity check model.

$\mathbf{s}_1$-$\mathbf{s}_3$	I I	A G G - G G A - G A G	$\mathbf{s}_{25}$-$\mathbf{s}_{27}$	C G	A G G - G G A - G A G
$\mathbf{s}_4$	I A	A A G	$\mathbf{s}_{28}$-$\mathbf{s}_{30}$	C U	G G U - G U G - A U C
$\mathbf{s}_5$-$\mathbf{s}_6$	I C	U G A - G A U	$\mathbf{s}_{31}$-$\mathbf{s}_{33}$	G I	G G U - G U G - A U C
$\mathbf{s}_7$-$\mathbf{s}_9$	I G	G G U - G U G - A U C	$\mathbf{s}_{34}$-$\mathbf{s}_{35}$	G A	U G A - G A U
$\mathbf{s}_{10}$	I U	U A A	$\mathbf{s}_{36}$-$\mathbf{s}_{38}$	G C	A G G - G G A - G A G
$\mathbf{s}_{11}$	A I	A A G	$\mathbf{s}_{39}$	G G	U A A
$\mathbf{s}_{12}$-$\mathbf{s}_{14}$	A A	G G U - G U G - A U C	$\mathbf{s}_{41}$	U I	U A A
$\mathbf{s}_{15}$	A C	U A A	$\mathbf{s}_{40}$	G U	A A G
$\mathbf{s}_{16}$-$\mathbf{s}_{17}$	A G	U G A - G A U	$\mathbf{s}_{42}$-$\mathbf{s}_{44}$	U A	A G G - G G A - G A G
$\mathbf{s}_{18}$-$\mathbf{s}_{20}$	A U	A G G - G G A - G A G	$\mathbf{s}_{45}$-$\mathbf{s}_{47}$	U C	G G U - G U G - A U C
$\mathbf{s}_{21}$-$\mathbf{s}_{22}$	C I	U G A - G A U	$\mathbf{s}_{48}$	U G	A A G
$\mathbf{s}_{23}$	C A	U A A	$\mathbf{s}_{49}$-$\mathbf{s}_{50}$	U I	U G A - G A U
$\mathbf{s}_{24}$	C C	A A G			

Figure 7.7 shows the output of May's parity check model applied to the 3194 mRNA sequences [Gon04]. The performance ratio yields a value of $R = 0.22$. This fairly good

performance is due to the construction of the codebook based on the 13 last bases, i.e. such that it includes subsequences of the Shine-Dalgarno sequence. Nonetheless, there is no indication that the DNA has been block encoded during evolution and – even more in-questioned – if the 16S rRNA can perform such complicated decoding tasks with modulo 5 operations. In [MVBR04], May et al. also developed an (8,2) code analog to the presented (5,2) code which yielded very similar results. In summary, May based her models on highly speculative hypotheses, and the results should thus not be over-interpreted.

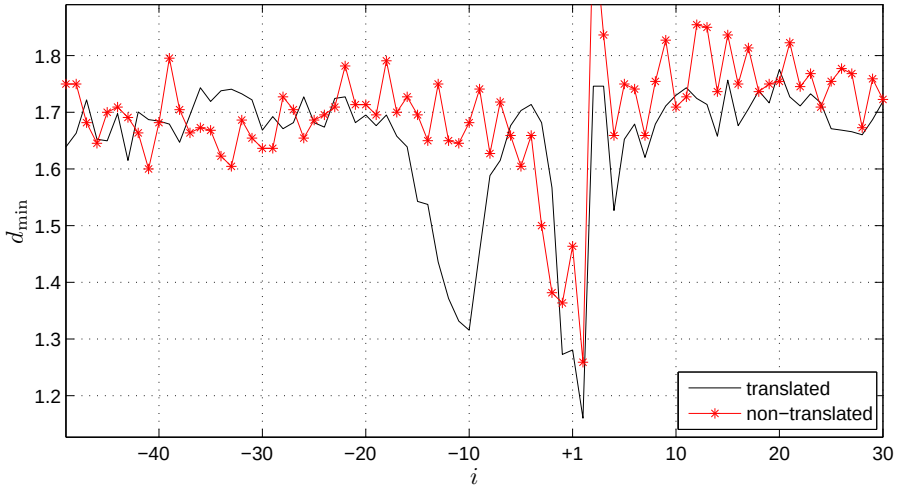


Figure 7.7: Decoding output obtained for May's parity check model.

7.1.7 16S rRNA based model

The three models presented above all showed to detect the Shine-Dalgarno sequence, however, in some cases with a very poor performance (13 bases complement model) or based on speculative hypotheses (May's parity check model). Therefore, a new model is proposed which is based on a codebook created from the 13 last bases of the 16S rRNA. The derivation follows closely that in [DGHM05]. It can be seen that the Shine-Dalgarno (length 6) is part of the complement of these 13 bases which suggests that detection takes place via a successive comparison of subsequences of the last 13 bases with the mRNA. For this reason, the codebook is obtained by taking all subsequences of length L of the complement of these bases. Table 7.5 lists the 9 resulting codewords for $L = 5$.

The output of the 16S rRNA model with $L = 5$ is depicted in Figure 7.8 [Gon04]. The performance ratio yields a value of $R = 0.32$, the highest value achieved so far. Compared to the 13 bases complement model, the performance improved four-fold. The performance of other codeword lengths ($2 \leq L \leq 7$) showed to be slightly inferior. This is surprising since the Shine-Dalgarno sequence is reported to have length 6 and is part of the 13 bases complement, which would suggest $L = 6$ to show the best performance. The fact that the codebook with $L = 5$ performed better implies that only 5 bases of the Shine-Dalgarno sequence are important for detection by the 16S rRNA. This is confirmed by the results

Table 7.5: Codebook of the 16S rRNA model for $L = 5$.

$\mathbf{s}_1$	U A A G G
$\mathbf{s}_2$	A A G G A
$\mathbf{s}_3$	A G G A G
$\mathbf{s}_4$	G G A G G
$\mathbf{s}_5$	G A G G U
$\mathbf{s}_6$	A G G U G
$\mathbf{s}_7$	G G U G A
$\mathbf{s}_8$	G U G A U
$\mathbf{s}_9$	U G A U C

in [SBRS01] (see also Section 7.1.5), where the last base of the Shine-Dalgarno sequence showed a low conservation and no strong preference for a particular base. The results and the inferior performance obtained with the 16S rRNA model indicate that it simulates the behavior of the ribosome during translation initiation best.

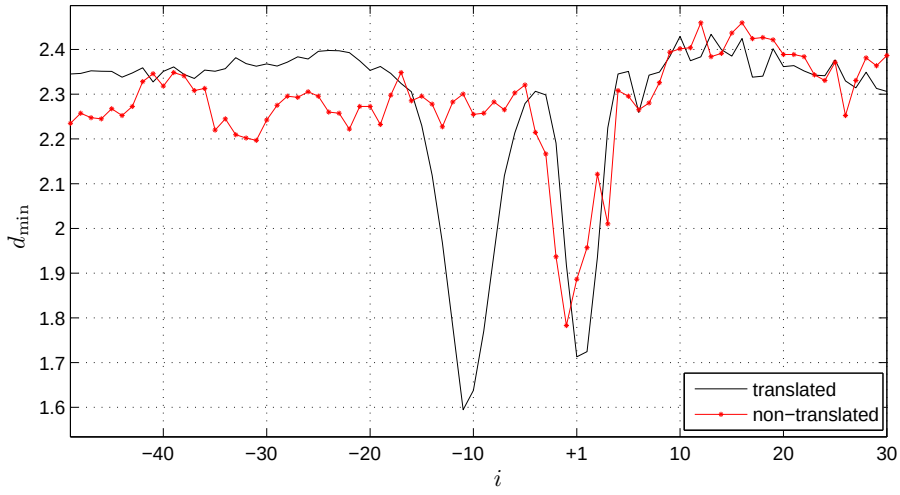


Figure 7.8: Decoding output obtained for the 16S rRNA based model.

7.1.8 Detection signals

Different models for the detection of the translation initiation signals were presented in the last sections. The 16S rRNA model showed the best performance and is less based on speculative hypotheses. In the next step, the model is applied to the full range of the

dataset as depicted in Figure 7.3, i.e. to the region around the start codon as well as that around the stop codon. Figure 7.9 shows the resulting output, the horizontal line depicts the position of the dataset where the sequences were cut to a fixed length. Interestingly, it exhibits not only the minimum at the Shine-Dalgarno sequence and at the start codon but also a significant minimum at the stop codon. This fact suggests that the 16S rRNA is involved in detection of the termination signal, which was already proposed in the early work by J. Shine and L. Dalgarno [SD74] but neither substantiated nor disproved since then. It was proposed in [DGHM05] that the 16S rRNA formerly used to be responsible for detection of both initiation and termination signals but was later replaced by more sophisticated systems involving more interactions. Nonetheless, the evolutionary role of the 16S rRNA for translation termination is still observable in the results. The results are supported by publications reporting that mutations in the 16S rRNA have a negative effect on translation termination [GHMD91, PG90].

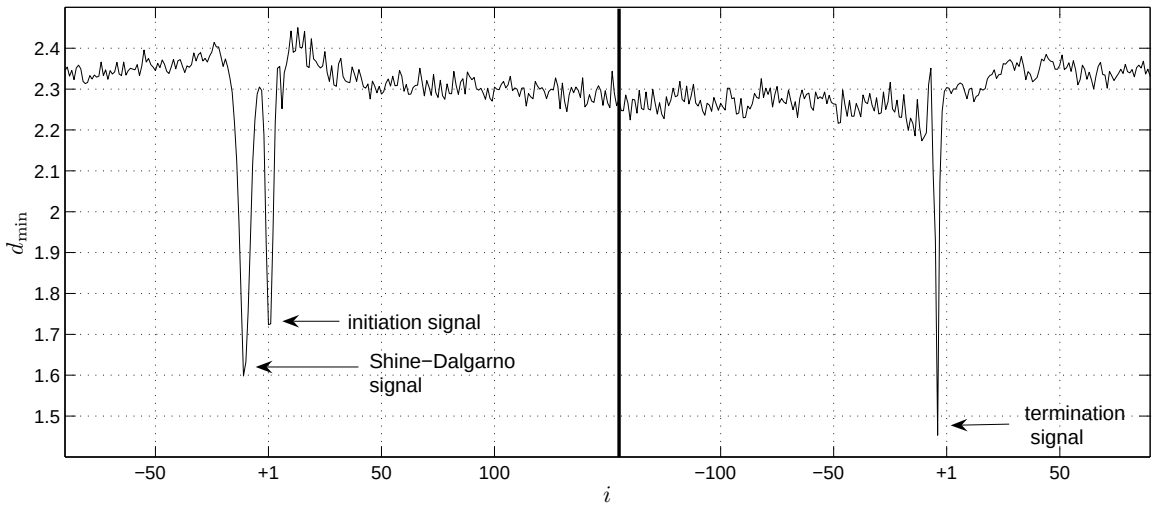


Figure 7.9: Decoding output obtained for the 16S rRNA model ($L = 5$) applied to the whole range of the mRNA dataset.

7.2 Energy metric

In the proposed models, the Hamming distance was used for calculation of the minimum distance (Eq. (7.1)). Since the actual interaction between the 16S rRNA and the mRNA is based on hydrogen bonds between the bases, a free energy metric is now proposed to replace the Hamming distance. The free energy ΔG is the energy that is released if a chemical reaction takes place, e.g. the hydrogen bonding between nucleotides. The more negative the free energy, the stronger is the bond between the nucleotides. The values presented in the following are downloaded from the website of D. H. Turner [UoR08], since these are publicly available and well accepted among biologists. The free energy values are always given for two neighboring nucleotides, i.e. they refer to the free binding energy of two nucleotides in the mRNA to two nucleotides in the 16S rRNA (see Figure 7.10). The

derivations and more information on the calculation of the free energy values can be found in [FKJ+86, JTZ89]. The energy metrics presented in the following are solely calculated from the binding energies of nucleotides (the so-called nearest neighbor parameters) and do not account for loops that may occur in RNA sequences. The nearest neighbor model is generally considered as a valid approach for the calculation of RNA-RNA binding energies (see e.g. [KMC+06]). As mentioned in Section 4.1 and Section 5.1 for bacterial transcription, low energies indicate a strong binding.

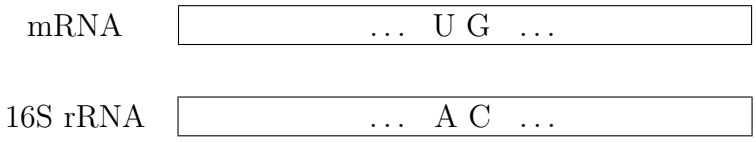


Figure 7.10: Illustration of the calculation of the free binding energy ΔG.

▷ **Example 7.3**

To illustrate the calculation of the free energy, the mRNA sequence AGGAG and the codeword UCCUC are considered exemplarily. In the first step, the first two nucleotides of both sequences are considered (energy between the dinucleotide AG and the dinucleotide UC). The subsequent energies are calculated in single-nucleotide steps, i.e. the next energy term is that between the dinucleotide GG and the dinucleotide CC. This ends up with

$$\Delta G_{seq} = \Delta G_{\text{AG-UC}} + \Delta G_{\text{GG-CC}} + \Delta G_{\text{GA-CU}} + \Delta G_{\text{AG-UC}}. \qquad \triangleleft$$

7.2.1 Watson-Crick base pairing

The first proposed energy metric only takes into account the Watson-Crick base pairing, i.e. bonds between A and U as well as between C and G. The resulting values are listed in Table 7.6.

Table 7.6: Turner's free energy values based on Watson-Crick base pairing [kcal/mol].

mRNA	AA	AC	AG	AU	CA	CC	CG	CU
16S rRNA	UU	UG	UC	UA	GU	GG	GC	GA
energy	-0.9	-2.2	-2.1	-1.1	-2.1	-3.3	-2.4	-2.1
mRNA	GA	GC	GG	GU	UA	UC	UG	UU
16S rRNA	CU	CG	CC	CA	AU	AG	AC	AA
energy	-2.4	-3.4	-3.3	-2.2	-1.3	-2.4	-2.1	-0.9

7.2.2 Including wobble base pairs

In addition to the Watson-Crick base pairing, a so-called wobble base pair between G and U can also occur and is a fundamental part of all RNAs [VM00]. It is even reported to have the same thermodynamic stability as Watson-Crick base pairs. Table 7.7 lists the resulting energy values between the 16S rRNA and the mRNA if including wobble base pairs.

Table 7.7: Turner's free energy values including wobble base pairs [kcal/mol].

mRNA	AA	AC	AG	AU	AG	AU	CA	CC	CG	CU	CG	CU
16S rRNA	UU	UG	UC	UA	UU	UG	GU	GG	GC	GA	GU	GG
energy	-0.9	-2.2	-2.1	-1.1	-0.6	-1.4	-2.1	-3.3	-2.4	-2.1	-1.4	-2.1
mRNA	GA	GC	GG	GU	GG	GU	UA	UC	UG	UU	UG	UU
16S rRNA	CU	CG	CC	CA	CU	CG	AU	AG	AC	AA	AU	AG
energy	-2.4	-3.4	-3.3	-2.2	-1.5	-2.5	-1.3	-2.4	-2.1	-0.9	-1.0	-1.3
mRNA	GA	GC	GG	GU	GG	GU	UA	UC	UG	UU	UG	UU
16S rRNA	UU	UG	UC	UA	UU	UG	GU	GG	GC	GA	GU	UG
energy	-1.3	-2.5	-2.1	-1.4	-0.5	+1.3	-1.0	-1.5	-1.4	-0.6	+0.3	-0.5

7.2.3 Including terminal mismatches

The third presented energy metric considers a base doublet with a Watson-Crick or a wobble base pair at the first position and any other combination of base X in the mRNA and base Y in the 16S rRNA at the second position. Table 7.8 lists the resulting energy values including terminal mismatches.

7.2.4 Comparison

In case of usage of an energy metric, the codewords are constructed from the 13 last bases themselves instead of their complement. Figure 7.11 shows a comparison between the decoding output obtained with the three presented energy metrics combined with the 16S rRNA model for $L = 5$, since this codeword length showed the best performance in Section 7.1.7. The results are plotted with respect to the optimum detection: In case of the Hamming distance, the optimum detection yields a value of zero, whereas in case of the energy metrics the optimum detection gives a high negative energy value. For the

Table 7.8: Turner's free energy values with terminal mismatches [kcal/mol].

AX		Y				**CX**		Y			
UY		A	C	G	U	**GY**		A	C	G	U
	A	-0.8	-1.0	-0.8	-1.0		A	-1.5	-1.5	-1.4	-1.5
X	C	-0.6	-0.7	-0.6	-0.7	X	C	-1.0	-1.1	-1.0	-0.8
	G	-0.8	-1.0	-0.8	-1.0		G	-1.4	-1.5	-1.6	-1.5
	U	-0.6	-0.8	-0.6	-0.8		U	-1.0	-1.4	-1.0	-1.2
GX		Y				**UX**		Y			
CY		A	C	G	U	**AY**		A	C	G	U
	A	-1.1	-1.5	-1.3	-1.5		A	-1.0	-0.8	-1.1	-0.8
X	C	-1.1	-0.7	-1.1	-0.5	X	C	-0.7	-0.6	-0.7	-0.5
	G	-1.6	-1.5	-1.4	-1.5		G	-1.1	-0.8	-1.2	-0.8
	U	-1.1	-1.0	-1.1	-0.7		U	-0.7	-0.6	-0.7	-0.5
GX		Y				**UX**		Y			
UY		A	C	G	U	**GY**		A	C	G	U
	A	-0.3	-1.0	-0.8	-1.0		A	-1.0	-0.8	-1.1	-0.8
X	C	-0.6	-0.7	-0.6	-0.7	X	C	-0.7	-0.6	-0.7	-0.5
	G	-0.6	-1.0	-0.8	-1.0		G	-0.5	-0.8	-0.8	-0.8
	U	-0.6	-0.8	-0.6	-0.6		U	-0.7	-0.6	-0.7	-0.5

16S rRNA model with $L = 5$, optimum detection is achieved if the currently considered mRNA sequence is AGGAG (the first 5 bases of the Shine-Dalgarno sequence) yielding an energy value of -9.9 for the energy metrics without terminal mismatches and an energy value of -5.0 for the energy metric including terminal mismatches. With respect to these values, the energy metric including terminal mismatches seems to perform best, which is confirmed by evaluating Eq. (7.2): The energy metric based on Watson-Crick base pairs has a performance ratio of $R = 0.32$, that including wobble base pairs $R = 0.36$ and that including terminal mismatches $R = 0.43$. Therefore, the last metric seems to provide the best model of nature due to its flexibility in base pairings at the second position. Recalling the performance ratio obtained for the 16S rRNA model based on the Hamming distance ($R = 0.32$) shows that the energy metrics are able to improve the detection of the Shine-Dalgarno sequence due to their appropriate modeling of the chemical interaction.

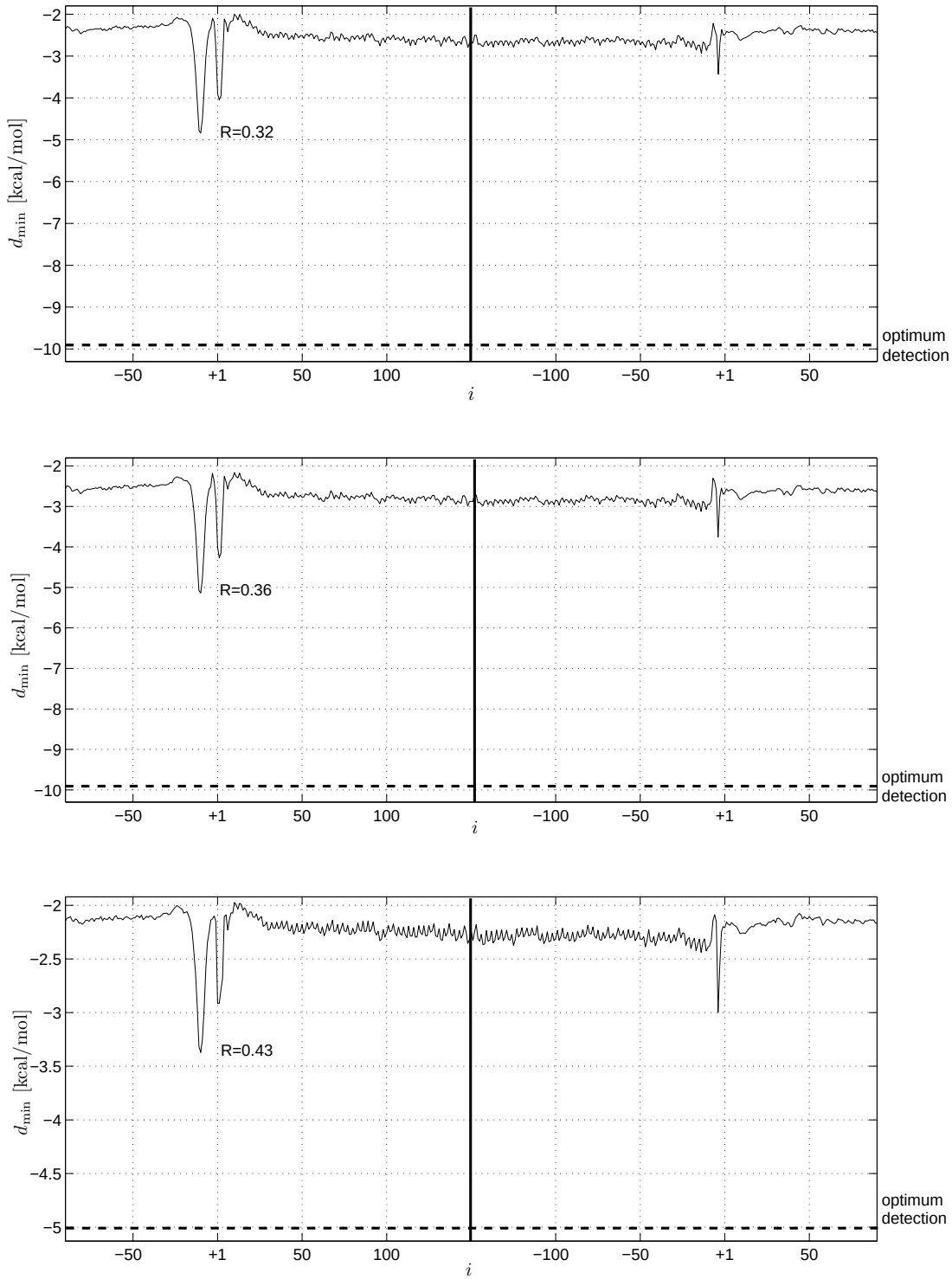


Figure 7.11: Decoding output obtained for the 16S rRNA model with the three energy metrics (top: Watson-Crick base pairing, middle: including wobble base pairs, bottom: including terminal mismatches). Note: y-axes range is adjusted with respect to the value of optimum detection.

7.3 Mutational analysis

In the last section, three models were presented together with several metrics of calculating the distance between mRNA sequences and the codewords. The 16S rRNA model combined with Turner's free energy including terminal mismatches exhibited the best performance since they simulate the flexibility in nature best. In the next step, this model is used to predict the behavior of point mutations in the 16S rRNA. Biological laboratory experiments in this direction are time and cost extensive, which can be bypassed through simulations based on the presented model [DGHM05].

7.3.1 Verification

To prove that simulations on the effect of mutations are valid, three mutations whose effect on translation is known are inserted into the codebook and the effect in terms of detection of the Shine-Dalgarno sequence is measured. The positions of the mutations are referenced in the following way:

position		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	
base	3'	A	U	U	C	C	U	C	C	A	C	U	A	G	5'
				|	|	|	|	|	|					
SD sequence	5'			A	G	G	A	G	G						3'

Mutations by Hui and de Boer

In 1987, A. Hui and H. A. de Boer investigated the effect of the following mutation in a single mRNA species [HdB87]: They changed bases 2 to 6 of the Shine-Dalgarno sequence (normally complementary to positions 4 - 8 of the 16S rRNA) from GGAGG to CCUCC, which corresponds to changing bases 4 - 8 of the codebook from CCUCC to GGAGG. The mutation led to a strong decrease of the synthesis of the respective protein. Thereafter, they mutated the 16S rRNA at positions 4 to 8 such that the complementarity to the mutant Shine-Dalgarno sequence was restored, which led to a full resumption of protein synthesis. Figure 7.12 depicts the effect of the mentioned mutation on the simulation output. It shows clearly that the detection of the Shine-Dalgarno sequence is strongly diminished by the changes in the codebook (the performance decreases from $R = 0.43$ to $R = 0.21$), while the detection of the start codon remains the same.

Mutations by Firpo et al.

In 1996, M. A. Firpo et al. mutagenized positions 12 and 13 of the 16S rRNA from A and G to G and A [FCGD96]. They observed a dramatic effect on the vitality of the organism,

which they showed to be due to a hindered binding of the 16S rRNA to initiation factors and tRNAs. However, an influence on binding to the mRNA during translation initiation could be experimentally excluded. Figure 7.12 depicts the effect of the mutation on the detection of the Shine-Dalgarno sequence and the start codon based on the 16S rRNA model. It can be seen that the detection of both regions is not altered by the mutations, which is in accordance with the experimental results of Firpo et al.

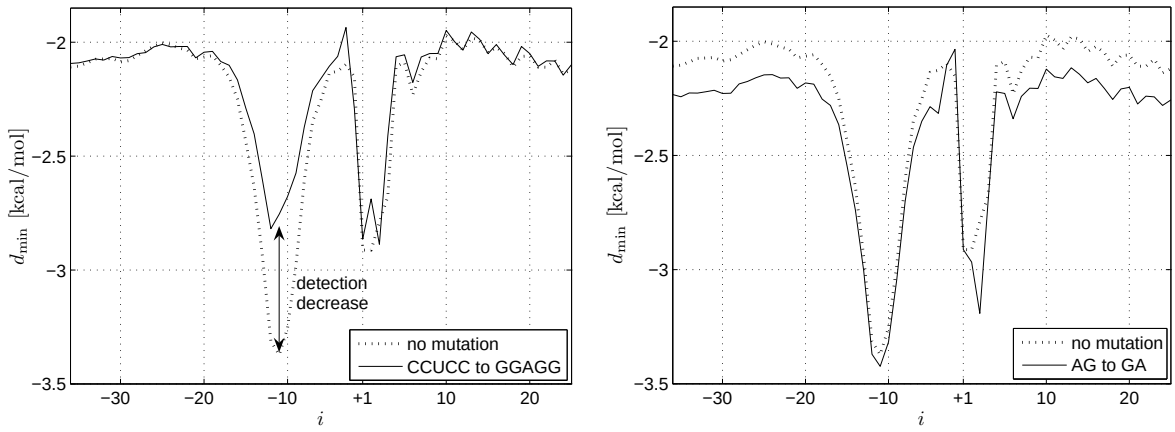


Figure 7.12: Effect of mutations by Hui and de Boer (left) and by Firpo (right).

7.3.2 Generalization to all bases

In the last sections, the validity of simulations for measuring the effect of mutations was supported. Therefore, the effect can now be generalized to all bases by inserting single-base changes in the codebook and classifying the increase or decrease on the detection. Table 7.9 lists the effect on the detection of the start codon and the stop codon, Table 7.10 lists the obtained effect on the detection of the Shine-Dalgarno sequence (SD). $\downarrow$ and $\Downarrow$ refer to a slight and strong decrease, respectively, in the detection strength as measured by the ratio R.

Recalling that positions 3 to 8 of the 13 last bases are complementary to the Shine-Dalgarno sequence explains the strong effect of bases 4 to 7 on the detection. In Section 7.1.5 and Section 7.1.7, it was reported that the last base of the Shine-Dalgarno sequence seems not to be of great importance for the detection by the 16S rRNA. This is confirmed by the mutational analysis, where a mutation in position 8 showed to have no influence. Positions 5 and 6 are complementary to the third and fourth base in the Shine-Dalgarno sequence, the most conserved bases that are thus expected to have the dramatic effect observed after inserting changes in the codebook at these positions. In case of the start codon, many bases seem to be involved in its detection. The effect on the detection of the stop codon is especially strong for the first few bases, a region that otherwise contains the complement of the stop codon UAA.

Table 7.9: Qualitative, worst effect of mutations in the last 13 bases of the 16S rRNA on the detection of the start and the stop codon.

position	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
start codon	→	→	→	↓	↓	↓	→	↓	↓	→	→	→	→
mutation	–	–	–	G	G	G	–	A	G	–	–	–	–
position	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
stop codon	↓	⇓	⇓	↓	⇓	↓	↓	↓	↓	↓	↓	→	→
mutation	U	G	G	G	G	G	G	G	G	G	G	–	–

Table 7.10: Effect of mutations in the last 13 bases of the 16S rRNA on the detection of the Shine-Dalgarno sequence (performance loss or gain in percent of R).

position	**1**	**2**	**3**	**4**	**5**
A	–	± 0 %	**-15.9 %**	**-30.2 %**	-30.3 %
C	± 0 %	**-3.2 %**	-9.0 %	–	–
G	+8.3 %	+3.2 %	-13.9 %	-26.8 %	**-31.1 %**
U	± 0 %	–	–	-2.4 %	-2.4 %
average	→ (+2.8 %)	→ (± 0 %)	↓ (-12.9 %)	↓ (-19.8 %)	⇓ (-21.3 %)
position	**6**	**7**	**8**	**9**	**10**
A	**-24.4 %**	-17.3 %	-2.4 %	–	± 0 %
C	-15.9 %	–	–	+8.0 %	–
G	-23.8 %	**-18.7 %**	**-3.4 %**	+2.4 %	**-2.4 %**
U	–	-2.4 %	± 0 %	+8.0 %	± 0 %
average	⇓ (-21.4 %)	↓ (-12.8 %)	→ (-1.9 %)	→ (+6.1 %)	→ (-0.8 %)
position	**11**	**12**	**13**	-	-
A	**-2.4 %**	–	± 0 %		
C	± 0 %	± 0 %	± 0 %		
G	**-2.4 %**	± 0 %	–		
U	–	± 0 %	± 0 %		
average	→ (-0.8 %)	→ (± 0 %)	→ (± 0 %)		

7.4 Summary

In this chapter, the process of prokaryotic translation initiation was modeled using frame synchronization and codebook models. Biological experiments have shown that the detection of the Shine-Dalgarno sequence – the sync word of prokaryotic translation – is based on the interaction between the 16S rRNA (a subunit of the ribosome) and the mRNA. Therefore, codebook models were derived from this interaction in the bacterium *Escherichia coli*, which led to the following results:

- ▷ Four codebook models were designed focusing on different aspects of the interaction and on different underlying hypotheses. The detection strength of the Shine-Dalgarno sequence was compared to rate the quality of the models. Interestingly, a simplistic model with few underlying hypotheses showed to perform best, which allows conclusions on the nature of the interaction: The last 13 bases of the 16S rRNA seem to serve as a codebook for detection of the Shine-Dalgarno sequence.

- ▷ Subsequently, the best-performing codebook model was applied to a wide range of an aligned dataset of mRNA sequences. An astonishing result was that not only a detection signal at the Shine-Dalgarno sequence was observed but also one at the stop codon. This suggests that the 16S rRNA is also responsible for detecting the end of the coding sequence which strengthens an old hypothesis that could not be substantiated until today.

- ▷ The codebook model was thereafter extended by an energy metric, i.e. the binding energies between the 16S rRNA and the mRNA were included in the distance calculation to the codewords. Three different energy metrics were introduced.

- ▷ The best-performing codebook model combined with the best-performing energy metric was then used to model mutations in the 16S rRNA. Conducting mutational analyses in biological experiments is highly time- and cost-extensive and can therefore only be performed for few individual mutations. The correct prediction of the effect of mutation through the model was tested using known mutations. Subsequently, the model was applied to predict the effect of all possible mutations on the detection of the Shine-Dalgarno sequence, the start codon and the stop codon. The mutational analysis supported the involvement of the 16S rRNA in the detection of the start and stop codon and exposed the varying importance of its 13 last bases for the detection of the Shine-Dalgarno sequence.

8

Modeling Translation Initiation in Eukaryotes

In Chapter 7, codebook models for the detection of the Shine-Dalgarno sequence – the sync word of bacterial (prokaryotic) translation – were derived. In higher organisms (eukaryotes), the process involves far more interactions and factors. Moreover, the details of many interactions are not yet fully understood, which renders the development of a simple model nearly impossible. Instead, the focus of this chapter is to provide a general sequence analysis of the coding sequence and the translation initiator region based on methods derived from information theory.

In Section 8.1, the main differences between translation initiation in prokaryotes and eukaryotes are detailed. Subsequently, Section 8.2 investigates the coding sequence of eukaryotic mRNAs using Kullback-Leibler divergence and mutual information. Results are interpreted with respect to their influence on the synchronization process underlying translation initiation. Section 8.3 follows with a codebook model for the detection of the Kozak sequence – the biological sync word of eukaryotic translation.

8.1 Differences to prokaryotic translation initiation

It was stated in Section 6.1 that transcription in eukaryotes is highly complex and involves many factors. The same holds true for translation: While a model for translation initiation in prokaryotes could be derived by focusing on the interaction between the 16S rRNA and the mRNA (see Chapter 7), it is difficult to simply extend this approach to eukaryotes.

8.1.1 Initiator region

The major differences between translation in prokaryotes and eukaryotes are related to the initiation, while the process of elongation does not differ significantly. In the first step of translation, the 40S subunit binds to the 5'-UTR and scans along it until it detects a start codon and the Kozak sequence. The latter ranges from -7 to +4 with respect to the first position of the start codon (+1, see Figure 8.1).

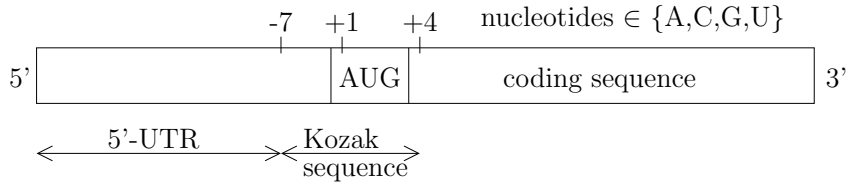


Figure 8.1: Structure of the initiator region of eukaryotic mRNA.

The most important difference to prokaryotes lies in the interaction of the small ribosomal subunit with the mRNA. As detailed in Section 3.4.4, the prokaryotic 16S rRNA base-pairs to the mRNA to detect the Shine-Dalgarno sequence. In contrast to that, it was long assumed that no base-pairing between the 18S rRNA (the eukaryotic equivalent to the 16S rRNA) and the mRNA takes place. It was only in 2006 that evidence for this mechanism could be found for a specific mRNA [DCZM06].

8.1.2 mRNA modification for protection

Transcription and translation in eukaryotes occur locally separated (see Section 3.4.1). After transcription, the mRNA exits the nucleus to the surrounding cytoplasm. It is thus strongly exposed to radiation and degradation, which could alter the mRNA before being translated if it was not specifically protected. This protection is achieved by a methylated guanine cap at the 5'-end and a poly(A)-tail at the 3'-end of the mRNA. The former seals the sensitive 5'-end and hereby prevents degradation in the 5'→3'-direction. The latter is a sequence of 100 - 200 adenine nucleotides which the poly(A)-binding protein (PABP) binds to in order to prevent degradation of the mRNA in the 3'→5'-direction. Both mRNA modifications seem to be important for translation, since it was shown that initiation is strongly inhibited if one modification is reversed experimentally [Lew07].

8.1.3 Translation initiation factors

Translation in eukaryotes involves far more proteins than in prokaryotes. While only three initiation factors (IF) are required for bacterial translation initiation, 12 factors are currently known to be involved in the equivalent eukaryotic steps (denoted by eIF) [Lew07]. These are required for all stages of translation initiation including formation of an ini-

tiation complex at the 5'-end, movement of the ribosome along the mRNA as well as detection of the start codon.

8.1.4 mRNA ring structure

During translation initiation, the initiation factor eIF4G binds to the guanine cap at the 5'-end of the mRNA. The poly(A)-binding protein (PABP) can thereafter bind to eIF4G to create a loop where the 5'-end and the 3'-end find themselves held in the same protein complex [Lew07]. The function of this ring structure is not fully understood yet, however, it is assumed to enable re-initiation of translation, i.e. the ribosome can directly start a new round of translation after having completed one.

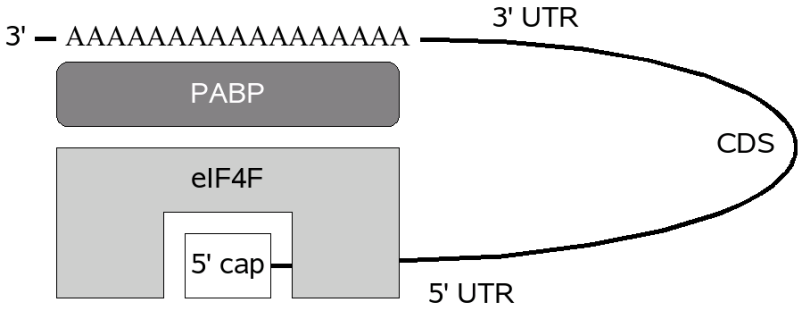


Figure 8.2: Ring structure of eukaryotic mRNAs.

8.1.5 Protein interactions during initiation

Translation initiation in prokaryotes is based on the interaction between the 16S rRNA and the mRNA (see Section 3.4.4). In eukaryotes however, it could not yet been proven that an according rRNA-mRNA interaction takes place. The 18S rRNA subunit of the eukaryotic ribosome – the evolutionary equivalent of the 16S rRNA in prokaryotes – was shown to affect translational efficiency of individual mRNAs [DCZM06], but generalizing experimental results stay inconclusive [Koz01] This fact makes the extension of the codebook models derived in Chapter 7 for prokaryotes particularly difficult.

8.2 Information theoretic analysis

In the following, two information theoretic measures – Kullback-Leibler divergence and mutual information – are applied to investigate the coding sequence and the initiator region of eukaryotic mRNAs. The latter refers to the 5'-UTR of the mRNA, i.e. the non-coding part ranging from the 5'-end to the start codon AUG. It is responsible for

initiation of transcription based on the recruitment of the translation machinery: The eukaryotic ribosome binds to the 5'-UTR and slides along it until it detects the Kozak sequence – the sync word of eukaryotic translation – and the start codon [Koz02, Koz99] (see also Section 3.4.5). A set of mRNAs of the house mouse *Mus musculus* is used for the analysis (see Appendix C.1.4), the data structure is as presented in Section 4.2.

8.2.1 Kullback-Leibler divergence

The relative entropy (or Kullback-Leibler divergence) at position i of a given dataset $\mathcal{S}$ of aligned mRNA sequences is given by (see also Section 4.2.2)

$$D(\hat{p}_{\mathcal{S}}(n,i) \parallel \hat{p}_{\mathcal{S}}(n)) = \sum_{n \in \mathcal{A}} \hat{p}_{\mathcal{S}}(n,i) \mathrm{ld} \frac{\hat{p}_{\mathcal{S}}(n,i)}{\hat{p}_{\mathcal{S}}(n)}, \tag{8.1}$$

where $\hat{p}_{\mathcal{S}}(n,i)$ refers to the distribution of the nucleotide $n \in \{A, C, G, U\}$ at position i of the dataset $\mathcal{S}$ and $\hat{p}_{\mathcal{S}}(n)$ the background nucleotide distribution in the whole dataset. In case of a uniform background distribution, Eq. (8.1) can be simplified to

$$D(\hat{p}_{\mathcal{S}}(n,i) \parallel \hat{p}_{\mathcal{S}}(n)) = 2 - H(n,i), \tag{8.2}$$

where $H(n,i)$ refers to the entropy at position i. In case of a perfect nucleotide conservation at position i, i.e. if one specific nucleotide occurs in all sequences at that position, the Kullback-Leibler divergence to the uniform distribution reaches a value of $D(\hat{p}_{\mathcal{S}}(n,i) \parallel \hat{p}_{\mathcal{S}}(n)) = 2$, since there is no uncertainty about the nucleotide ($H(n,i) = 0$).

Objective

The Kullback-Leibler divergence is a measure of the evolutionary conservation of a base at a certain position. In this context, position-specific conservation is measured as opposed to inter-species conservation. The first expresses the conservation of a nucleotide at a specific position on an mRNA over the whole mRNA population of an organism. The latter compares genes or mRNAs that code proteins of functional or chemical similarity or identity in different organisms. A number of metrics exist to research inter-species conservation. If a sufficiently large dataset of mRNA sequences is available, the Kullback-Leibler divergence constitutes a promising approach for measuring position-specific conservation. Highly conserved nucleotides in this sense are generally expected to be related to important cellular functions.

Results

The presented adaptation of the Kullback-Leibler divergence is applied to the set of mRNAs of the house mouse (*Mus musculus*). Figure 8.3 (top) depicts the Kullback-

Leibler divergence between the nucleotide distribution at position i of the aligned dataset and the background nucleotide distribution. Figure 8.3 (middle, bottom) show detailed views of the Kullback-Leibler divergence around the start codon and around the stop codon, respectively. Note that the peaks at the exact position of the start and stop codon are cut to enhance the surrounding details. As can be seen from Figure 8.3 (top), the Kullback-Leibler divergence at those positions is around $D(\hat{p}_{\mathcal{S}}(n,i) \parallel \hat{p}_{\mathcal{S}}(n)) = 2$, which implies an almost perfect base conservation if the background distribution is close to a uniform distribution. No significant values are observed in the 5'-UTR (see Figure 8.3, middle), except at the positions directly upstream of the start codon, which are known to be occupied by the Kozak sequence. Equivalently, low values are observed in the 3'-UTR (see Figure 8.3, bottom). Interestingly, the pattern of the Kullback-Leibler divergence in the coding sequence changes between start and stop codon: While most codons after the start codon exhibit the strongest conservation at their third position (third position marked by arrows), most codons before the stop codon exhibit the strongest conservation at their second position (second position marked by arrows).

Interpretation

As can be seen from the genetic code (see Figure 3.9), several codons may code for the same amino acid (the so-called synonymous codons), e.g. the four triplets GCA, GCC, GCG and GCU are translated into the amino acids alanine. These codons are however not used with equal frequency, but one codon (named the major codon) is generally preferred over the others, which is known as codon usage bias. In most cases of synonymous codons, only the third base varies, while the first and – above all – the second base stay the same. This explains well that the second base exhibits the strongest conservation at the end of the coding sequence (see Figure 8.3, bottom). The fact that the second base is normally the least varied one between synonymous codons makes the high conservation at the second position that is observed directly after the start codon (see Figure 8.3, middle) even more astonishing. Several explanations might be conceivable for this:

First, certain codons might be favored at the beginning of the coding sequence to achieve a charge of the protein end. Amino acids are categorized into hydrophobic (repelled from water) and hydrophilic (soluble in water). While the former are non-polar (uncharged), the latter are further categorized into positively and negatively charged. Several classes of proteins require a charged terminus for specific functions or processes, e.g. secretory proteins need a positively charged end for exiting the plasma membrane [ZPJ07]. Due to the preference of the major codon over the other synonymous codons, the preference for certain amino acids also introduces nucleotide biases and thus a higher conservation at the respective positions.

Second, the strong conservation of the third position could stem from nucleotide biases serving the interaction with the ribosome and initiation factors: The first codons of the coding sequence are reported to have a strong impact on translational efficiency. Specifically, adenine-rich codons promote translation initiation in prokaryotes [ZPJ07, CPCB$^+$94]. However, it could not yet be conclusively inferred whether the same holds

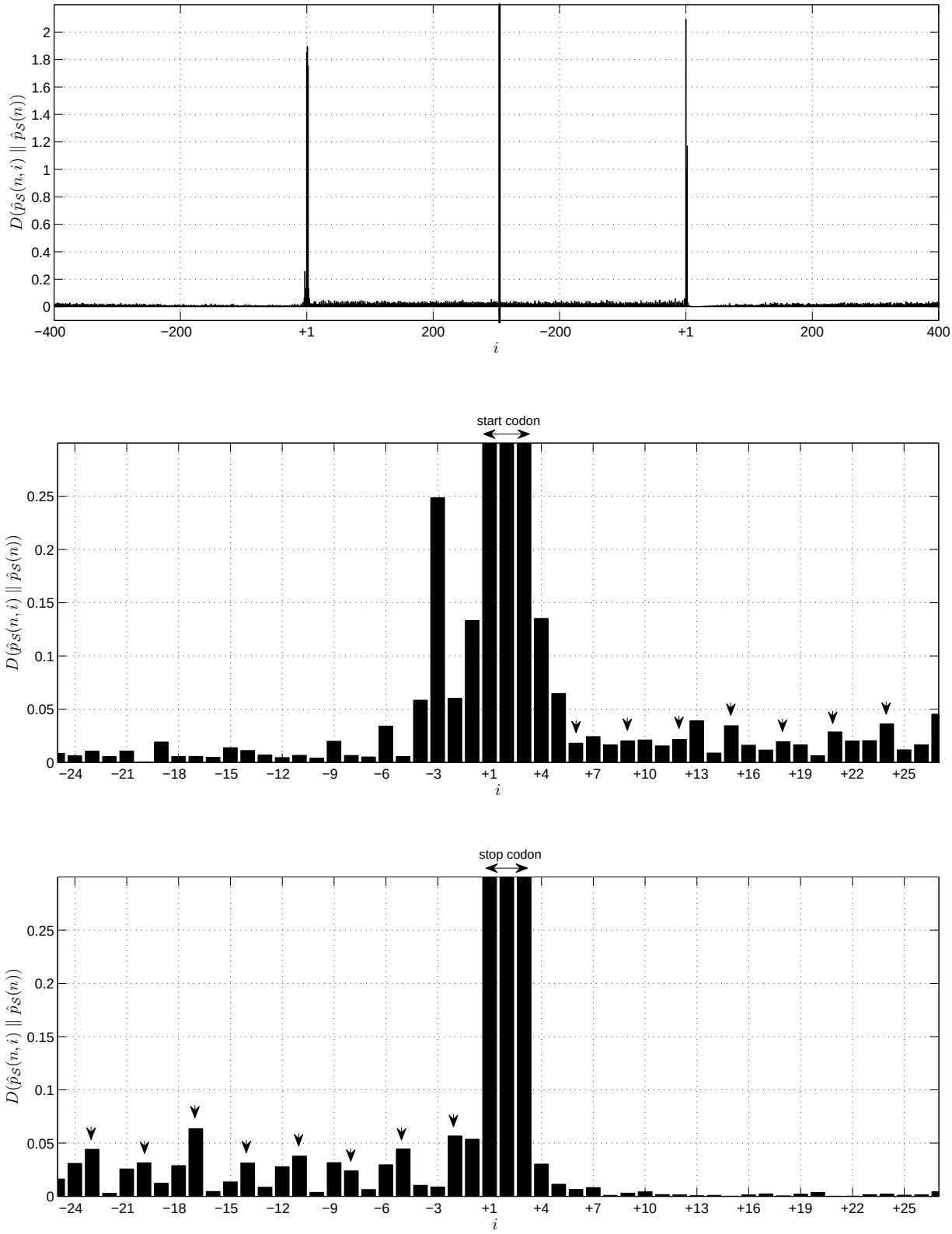


Figure 8.3: Kullback-Leibler divergence of the *Mus musculus* mRNAs (top: wide range, middle: start codon, bottom: stop codon).

true for eukaryotic organisms. Hence, an influence of adenine-rich codons on the high conservation at the third codon position can neither be supported nor disproved.

Third, protein binding sites in the coding sequence could influence the codon usage and, thus, the nucleotide conservation. As exemplarily listed for the amino acid arginine above, many amino acids are coded by fixed bases at the first two positions combined with any arbitrary nucleotide at the third position. Therefore, the third position can be used to encode additional signals, e.g. binding sites of so-called exonic splice enhancers (ESE) [WH07a]. Those are proteins that bind to the exonic regions of the pre-mRNA to induce splicing, the process during which the introns are removed to create the mature mRNA that is thereafter translated (see Section 3.4.1). The binding sites of these ESEs are typically 6 to 20 nucleotides long and purine-rich (i.e. AG-rich) [CB07]. Possibly existent ESE binding sites shortly after the start codon may thus introduce a nucleotide bias leading to the observed conservation at the third codon position. This hypothesis is supported by the fact that a low cytosine content co-occurs with high entropy (exposed through correlation analysis, data not shown).

8.2.2 Mutual information

The mutual information between two positions i_x and i_y of a given dataset of aligned mRNAs can be empirically estimated by (see also Section 4.2.3)

$$I(i_x; i_y) = \sum_{n_x \in \mathcal{A}} \sum_{n_y \in \mathcal{A}} \hat{p}_{\mathcal{S}}(n_x, n_y, i_x, i_y) \mathrm{ld} \frac{\hat{p}_{\mathcal{S}}(n_x, n_y, i_x, i_y)}{\hat{p}_{\mathcal{S}}(n_x, i_x)\hat{p}_{\mathcal{S}}(n_y, i_y)}, \tag{8.3}$$

where $\hat{p}_{\mathcal{S}}(n_x, n_y, i_x, i_y)$, $\hat{p}_{\mathcal{S}}(n_x, i_x)$, and $\hat{p}_{\mathcal{S}}(n_y, i_y)$ are again estimated from the dataset $\mathcal{S}$. The first refers to the counted joint occurrences of base n_x at position i_x and base n_y at position i_x. $I(i_x; i_y)$ quantifies the information that is obtained about nucleotide n_x by observing nucleotide n_y at a distance $d = i_x - i_y$.

Objective

Application of mutual information to mRNA sequences allows conclusions on the relations between two nucleotides that are neighboring each other or separated by short distances d. High mutual information between two nucleotides at distance d is a strong indicator for a functional relationship, since non-essential genetic features tend to be degraded over the course of evolution. Large-sized datasets are hereby necessary to produce meaningful results, for more information about the required sample size see [GDHM05].

Results

The presented mutual information estimate between distant positions is again applied to the dataset of mRNAs of *Mus musculus* (see Appendix C.1.4). Figure 8.4 shows the

average mutual information between nucleotides at distance d in the 5'-UTR (top), the coding sequence (middle) and the 3'-UTR (bottom). The values of the mutual information obtained for the whole dataset are depicted by the dashed lines in all three plots for reasons of comparison. As expected, a 3-periodicity is observed in the coding sequence (CDS), which stems from the triplet structure and has been detected before using signal processing techniques (see e.g. [Vai04]). Also expected is the fact that no periodicity is exhibited in the 3'-UTR. However interesting is the 3-periodicity in the 5'-UTR, which most investigations overlooked since they analyzed all non-coding DNA as one dataset instead of separating it into 3'-UTR, 5'-UTR etc. (see e.g. [GHBS00]).

Interpretation

As detailed in Section 4.2.4, translation is a highly sensitive process with respect to frame shifts, i.e. if shifted mRNA triplets are transformed into amino acids, which may yield a completely different protein. It was therefore concluded that the 3-periodicity in the coding sequence ensures the maintenance of the reading frame [XBA+06]. The observed 3-periodicity in the 5'-UTR strongly suggests that the ribosome is already synchronized to the triplet structure during its scanning process for the start codon. This hypothesis is backed by the fact that in *E. coli* not only the 5'-UTR but also the 3'-UTR exhibit the 3-periodicity [AM90]. Since bacterial mRNAs are polycistronic – i.e. they generally carry more than one coding sequence – many 3'-UTR constitute the 5'-UTRs of the subsequent coding sequence and, thus, should also carry the periodicity if it was related to reading frame maintenance.

Another possible explanation for the 3-periodicity in the 5'-UTR lies in evolutionary remainders of former coding sequences. Mutations occur at all times and during all stages of gene expression. These can either alter existing start codons, change similar codons into a start codon or modify the surrounding of randomly occurring AUG-codons such that it can initiate translation. This explanation however appears to be unlikely to be the only cause of the 3-periodicity, since single remainders of coding sequences would not have such a major effect on a huge dataset of mRNA sequences. Furthermore, if the periodicity was due to former coding sequences, a similar effect would be expected for the 3'-UTR, where mutations can alter the stop codon. Since this is not the case there, the 3-periodicity is more likely an aspect of synchronizing the ribosome to the codon structure.

8.3 Detection of the Kozak sequence

In this section, a codebook model for the detection of the Kozak consensus sequence is derived. The biological relevance of the Kozak consensus sequence is explained in Section 3.4.5. As mentioned in Section 8.1, the exact details of the interaction between the ribosome and the Kozak sequences are still unknown. Hence, the results of the

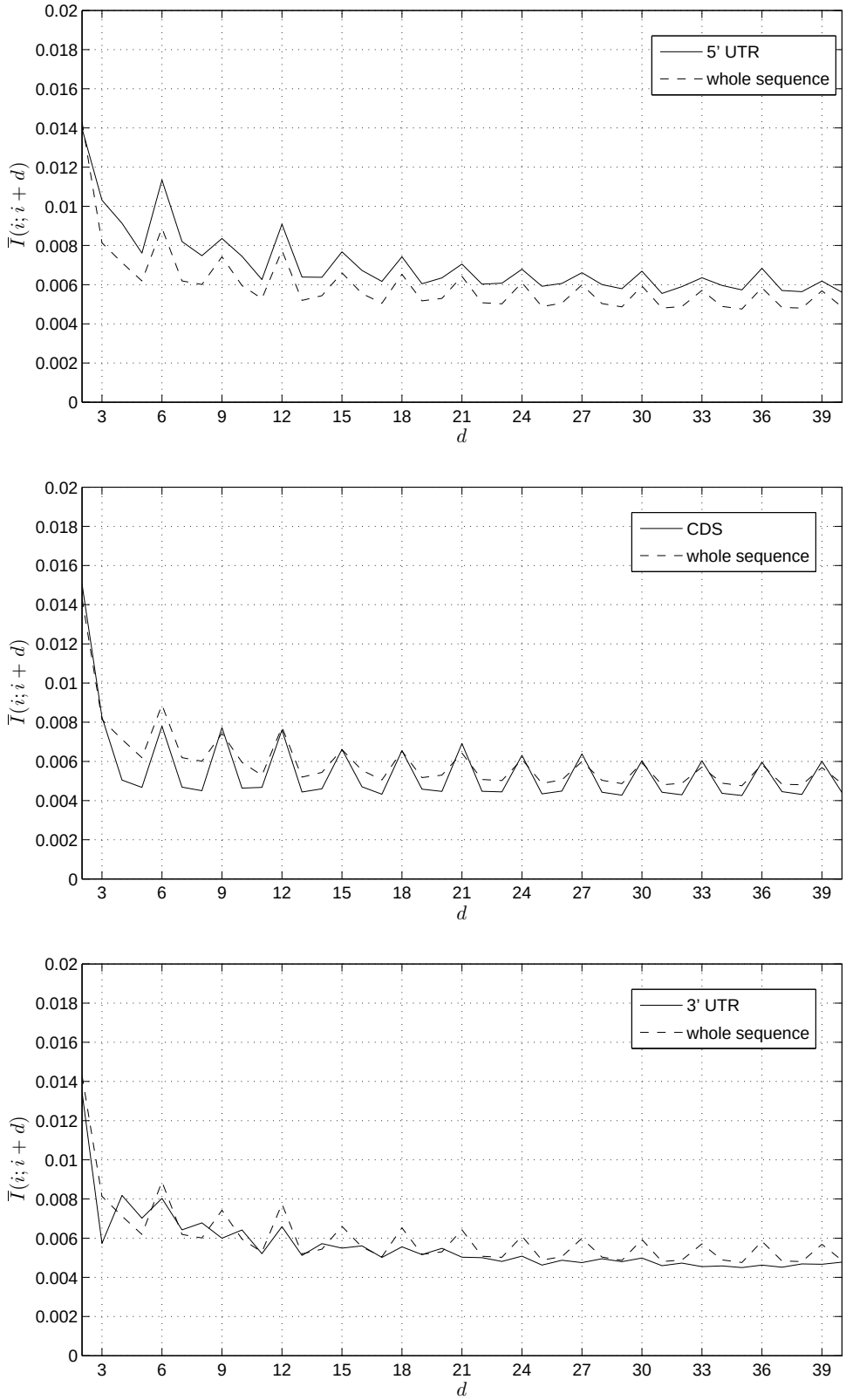


Figure 8.4: Mutual information between distant positions of the *Mus musculus* mRNAs (top: 5'-UTR, middle: CDS, bottom: 3'-UTR).

codebook modeling should be seen as a further aspect of sequence analysis rather than conclusions on a specific protein.

8.3.1 Codebook model

The codebook is created using the results presented in Section 8.2.1. As detailed there, the Kullback-Leibler divergence hints at how well conserved a nucleotide at a specific position is and thus at how important it is for translation. At each highly conserved position, the most probable base is chosen for the respective position in the codeword. When two bases have almost equal occurrence rates, one codeword is created for each. Taking into consideration all possible combinations, the codebook shown in Table 8.1 is obtained. Note that nucleotides remaining constant over all codewords are printed in lowercase. The exact location of the sequence used to build up the codebook is depicted in Figure 8.5.

Table 8.1: Kozak sequence codebook.

$\mathbf{s}_1$	g c c g C c A C c
$\mathbf{s}_2$	g c c g G c A C c
$\mathbf{s}_3$	g c c g C c A A c
$\mathbf{s}_4$	g c c g G c A A c
$\mathbf{s}_5$	g c c g C c G C c
$\mathbf{s}_6$	g c c g G c G C c
$\mathbf{s}_7$	g c c g C c G A c
$\mathbf{s}_8$	g c c g G c G A c

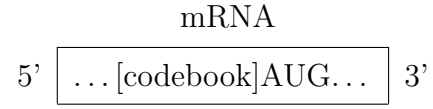


Figure 8.5: Position of the Kozak sequence codebook relative to the start codon.

8.3.2 Results and interpretation

The resulting output of applying the codebook to the aligned mRNA dataset of *Mus musculus* using Turner's free energy metric (see Section 7.2) is plotted in Figure 8.6 (top), while an enlarged view of the start codon region is given in Figure 8.6 (bottom).

It exhibits a relatively strong negative peak just upstream of the start codon followed by a positive one. A similar yet weaker behavior can be seen upstream of the stop codon. The peak before the start codon (marked by an arrow) indicates a good match between one of the codewords and the sequence there. Since the codebook is created from analysis of this very region, this match is not surprising. Much more surprising are the positive peaks that precede and follow and that stand for poor binding properties. Furthermore, a minimum is exhibited at the stop codon.

In order to ensure initiation at the exact location of the start codon, it seems plausible that its surroundings should match as badly as possible to relatively enhance it. As derived in Chapter 2, sync words should not exhibit a strong self-similarity for any overlap v. Fulfillment of this criteria minimizes the probability for shifted synchronization with the surrounding random data. If, however, the surrounding data is not random but can be chosen by the transmitter, it should be designed such that it is as dissimilar as possible to the sync word. The dissimilarity should include the symbol distribution as well as the structure. The positions with very poor binding properties directly before and after the detection of the Kozak sequence (see Figure 8.6, bottom) thus suggest that the surrounding nucleotides aid synchronization. This enables an exact detection of the start codon as hereby an in-frame start of translation. As observed for prokaryotic translation in Section 7.1.8, another position of strong binding is exhibited at the stop codon. This indicates that the ribosome subunits involved in start codon detection have also a share in detection of the stop codon.

8.4 Summary

This chapter dealt with transcription initiation in higher organisms (eukaryotes). Due to the high number of interactions between protein subunits and the mRNA, a strong focus was laid on sequence analysis using information theoretic measures rather than on deriving models for specific interactions. All sequence data used in the analyses belong to the house mouse (*Mus musculus*). The following results have been achieved:

- ▷ First, the Kullback-Leibler divergence (relative entropy) was adapted to its application to large sets of aligned mRNA sequences. Positions with high relative entropy are generally assumed to be of functional significance for translation initiation. Due to the fact that numerous amino acids are coded by codons with a varying third base, the lowest relative entropy is expected at that position. Surprisingly, this showed not to be true at the beginning of the coding sequence. This suggests that signals for other cell functions are hidden at the beginning of the coding sequence and hereby decrease the variability of the bases.

- ▷ Subsequently, the mutual information between nucleotides at short distances from each other was calculated. This analysis revealed a 3-periodicity in the coding sequence and the 5'-UTR. While the former has already been detected by methods

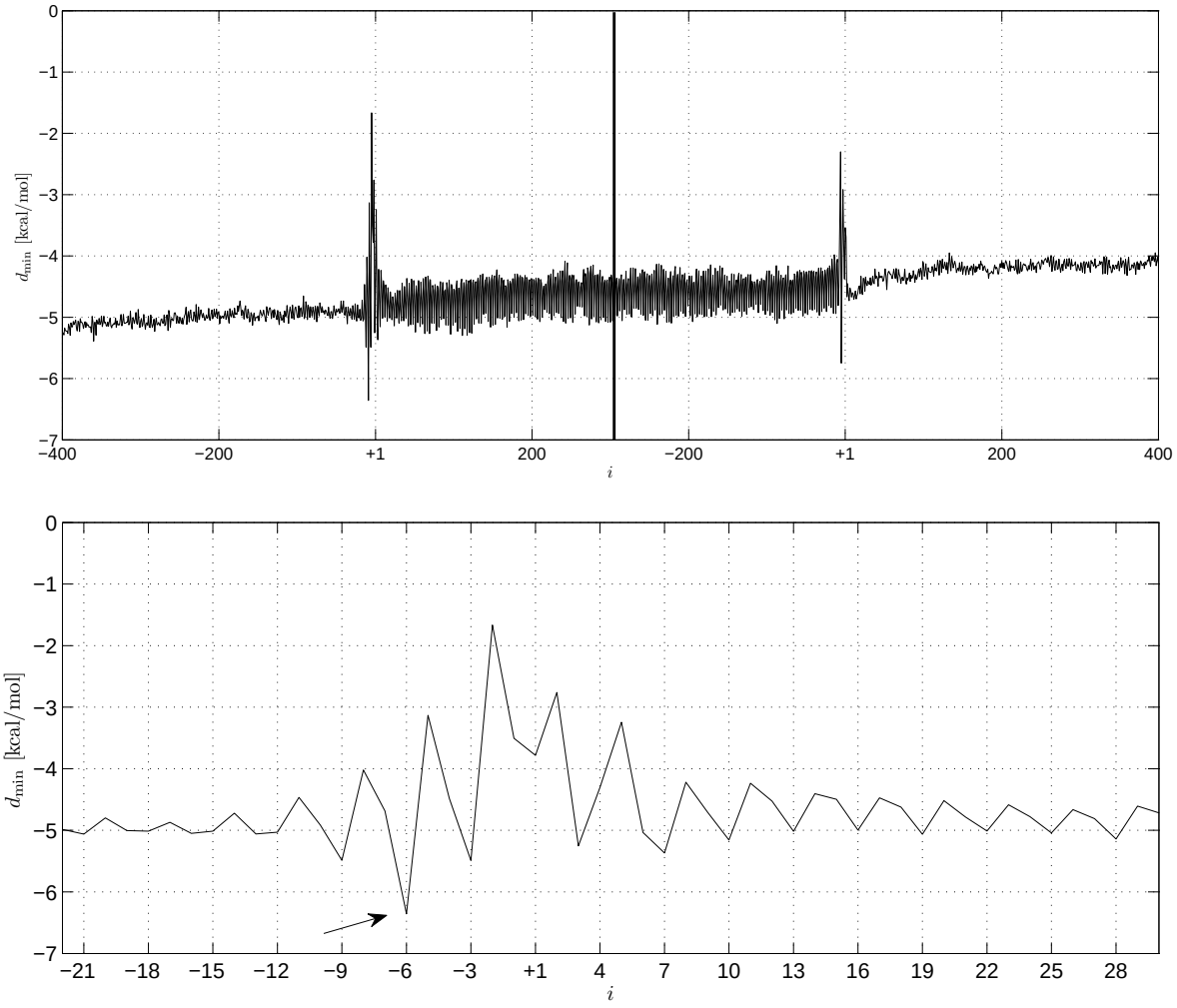


Figure 8.6: Detection of the Kozak sequence of *Mus musculus*. Top: complete mRNAs, bottom: start codon (the arrow marks the first position of the Kozak sequence).

from signal processing, the latter has never been reported. It suggests that the ribosome is synchronized to the triplet structure of the coding sequence during its search for the start codon.

▷ Thereafter, a codebook was derived from the Kozak sequence – the sync word of eukaryotic translation. Positions with very bad binding properties were observed on both sides of the detection signal. This indicates that the surrounding of the Kozak sequence is chosen such that the probability of shifted synchronizations is minimized. These may have a severe effect on the resulting protein due to the codon structure of the coding sequence.

9

Conclusions

This thesis dealt with the application of frame synchronization techniques combined with methods from coding and information theory to gene expression. With few exceptions, the two fields communications engineering and molecular biology have only in the last years started a cautious rapprochement, 50 years after Claude E. Shannon's "An Algebra for Theoretical Genetics" [Sha40]. This thesis aimed at fostering the cooperation between the two fields and to achieve the appreciation of this interdisciplinary topic in both communities. In the following sections, the main aspects and achievements of this thesis are summarized. Moreover, possible future research directions are presented.

9.1 Summary

After two introductory chapters detailing the technical and biological background, five chapters investigated the processes of transcription and translation for both bacteria (prokaryotes) and higher organisms (eukaryotes).

Chapter 2 introduced the basics of frame synchronization in technical systems. Its main objective was to provide the criteria underlying sync word detection and design.

Chapter 3 followed with the biological background needed for subsequent chapters. A strong focus was laid on gene expression, the vital process of protein synthesis taking place in the two steps transcription and translation. The two steps were detailed separately for prokaryotes and eukaryotes.

In Chapter 4, biological sync words were investigated. These are located shortly before specific regions of the DNA and need to be detected by proteins to initiate regulatory

processes. First, the prokaryotic promoter was analyzed with respect to its synchronization properties. Second, the Shine-Dalgarno sequence was examined using information theoretic measures.

In Chapter 5, a synchronization model of prokaryotic transcription was presented. It was derived based on binding energies and applied to an exhaustive set of known promoter sequences. The results were thereafter interpreted in terms of their biophysical impact on promoter detection by the RNA polymerase.

In Chapter 6, a sequence analysis of the promoter region in eukaryotes was conducted. For this purpose, mutual information and the Kullback-Leibler divergence were adapted and compared to weight matrices, a standard approach from bioinformatics.

Chapter 7 followed with a codebook model approach to modeling translation in prokaryotes. Four codebook models were derived based on the interaction between the 16S rRNA (a subunit of the ribosome) and the Shine-Dalgarno sequence. The best-performing model was extended by the base pairing energies to make the results more biologically meaningful. It was thereafter applied for a mutational analysis of the 16 rRNA.

Chapter 8 detailed translation initiation in eukaryotes. It focussed on a sequence analysis of large mRNA datasets based on mutual information and Kullback-Leibler divergence. Additionally, a codebook was derived for the detection of the Kozak sequence by the eukaryotic ribosome.

9.2 Achievements

The following main results were achieved for the bacterium *Escherichia coli* as the model organism of prokaryotes:

- ▷ The promoter and the Shine-Dalgarno sequence were analyzed with respect to their synchronization properties. The question that was sought to be answered was: Are biological sync words designed to serve synchronization by proteins? Interestingly, the autocorrelation properties and a Markov analysis of the whole genome suggest that the promoter is among the best possible sequences to ensure reliable detection by the RNA polymerase. The Shine-Dalgarno has unfavorable autocorrelation properties, which, however, help to synchronize the ribosome to the triplet structure of the coding sequence. These results give new perspectives on sequence evolution (see Chapter 4).
- ▷ A synchronization model was derived for transcription initiation. It was applied to a large set of known promoters and exposed a characteristic behavior of binding energies between the RNA polymerase and the DNA. Since the movement of the RNA polymerase along the double helix during promoter search depends on these binding energies, the characteristic behavior allowed a biophysical interpretation that yielded new insights into the interaction. The reliable synchronization to the

transcription start site seems to not only depend on the sync word (the promoter) but also on a wide surrounding (500 nucleotides). This fact should give strong indications for the future design of technical synchronization systems (see Chapter 5).

- ▷ Four codebooks were derived from the 16S rRNA (a subunit of the ribosome) involved in translation initiation. The models showed strong differences in their detection strength of the Shine-Dalgarno sequence. The results of the best-performing model suggest that the 13 last bases of the 16S rRNA serve as a codebook that enables a reliable detection (see Chapter 7).
- ▷ The successful development of an appropriate codebook model allowed to conduct a mutational analysis of the last 13 bases of the 16S rRNA. Biological experiments on the effect of single mutations are highly time and cost extensive but important to gain a sophisticated understanding on the important positions of an interaction. The conducted mutational analysis supported the importance of five nucleotides for detection of the Shine-Dalgarno sequence. Additionally, several nucleotides were found to have a strong impact on the detection of the start and stop codon.

The results achieved for eukaryotes are more generic, since less information is available about the underlying interactions. Moreover, the processes involve numerous factors that disallow simple models. Three model organisms were used for the investigations: the human (*Homo sapiens*), the house mouse (*Mus musculus*) and the family of arthropods (including the fruit fly *Drosophila melanogaster*).

- ▷ The promoter and the Kozak sequence were analyzed in terms of their synchronization properties. An important criterion in technical systems is to minimize shifted synchronizations, i.e. if an overlap of a part of the sync word with neighboring bits yields a valid sync word. In addition to an appropriate design of the sync word, the surrounding data should be as dissimilar as possible to the sync word. In the case of the human promoter, an extensive surrounding (2000 nucleotides) differs strongly in terms of nucleotide distributions and sequence structure. In the case of the Kozak sequence, the neighboring positions showed to be particularly dissimilar. These results indicate that – as seen for prokaryotes – biological synchronization processes incorporate both the design of the sync word and the choice of the wide surrounding (see Chapter 6 and Chapter 8).
- ▷ Strong differences were detected between the promoter surrounding in human and arthropods. This fact implies differences in the process of transcription initiation. While only minor differences were detected between gene expression processes in bacterial species, the results obtained for the eukaryotic species indicate a strong organism differentiation that took place during eukaryotic evolution (see Chapter 6).
- ▷ Mutual information and the Kullback-Leibler divergence were adapted for the analysis of promoter datasets. While the former showed no superiority to weight matrices from bioinformatics, the latter turned out to be a powerful tool for laying open the properties of the promoter surrounding (see Chapter 6).

- Both information theoretic measures were additionally adapted for the analysis of eukaryotic mRNA datasets. The Kullback-Leibler divergence exposed a yet unknown pattern of nucleotide biases in the coding sequence of *Mus musculus*. Due to synonymous codons that yield the same amino acid and generally differ only in the third position, the lowest nucleotide conservation would be expected at that position. However, this showed to be not the case at the beginning of the coding sequence, which suggests that additional regulatory signals are hidden in that region. Mutual information revealed that the 3-periodicity that has been reported for the coding sequence is also exhibited in the 5'-UTR (untranslated region). This suggests that the ribosome gets already synchronized to the triplet structure while searching for the Kozak sequence and the start codon (see Chapter 8).

9.3 Future research directions

The most interesting focus of future research would be a detailed analysis of gene expression in eukaryotes. Due to the highly limited understanding of underlying interactions, only a general sequence analysis was possible in this thesis. As soon as a better understanding will have been gained, single interactions could be modeled using techniques from frame synchronization. In this respect as well as regarding the analysis of prokaryotic processes, the following aspects could constitute interesting extensions of this work:

- In this thesis, the promoters of the main sigma factor of *Escherichia coli* were analyzed. Due to the limited availability of promoters, the analysis could not be extended to the six alternative sigma factors. Each sigma factor detects an own set of promoter sequences, which should therefore be as dissimilar as possible to avoid false detections by inappropriate sigma factors. In technical systems, orthogonal sequences and signals are employed to separate information. For example, orthogonal spreading sequences separate the users in code division multiple access (CDMA). The promoters of different sigma factors would be recognized with maximum reliability if they were orthogonal in a biological sense, i.e. if their structure and nucleotide content yields unique binding energy patterns. Therefore, as soon as enough data is available, promoter detection could be modeled as a multi user system where the sigma factors act as the receivers that extract their information from the data stream (the DNA) based on biologically orthogonal sync words (the promoters).

- The analysis of eukaryotic promoter sequences exposed major differences between the human and arthropod species. Therefore, the consideration of other eukaryotic organisms might reveal further differences that yield a better understanding of eukaryotic transcription initiation. For this purpose, the investigated organisms should be evolutionary distant, since those are more likely to exhibit strong differences. For example, the yeast species *Saccharomyces cerevisae* – a unicellular representative of

the kingdom Fungi – and the thale cress (*Arabidopsis thaliana*) – a well annotated species of the kingdom Plantae – could serve as model organisms.

- ▷ Protein-DNA interactions constitute a basic step of all cellular processes. For example, transcription factors bind to positions close to the gene start site. The binding sites of these proteins could be investigated with respect to their synchronization properties as presented in this thesis for prokaryotic and eukaryotic promoter sequences. Transcription factor binding sites are available in databases like Transfac [BIO08].

B. Hayes speculates in [Hay98] about possible reactions of researchers if they had known the rather simple mapping of nucleotide triplets to amino acids (known as the genetic code) before its discovery in 1961: "*My guess is that Nature*[1] *would have rejected the paper. 'This notion of the ribosome ratcheting along the messenger RNA three bases at a time – it sounds like a computer reading a data tape. Biological systems don't work that way. In biochemistry we have templates, where all the reactants come together simultaneously, not assembly lines where machines are built step by step.'* ". This thesis demonstrated that biological systems sometimes do work that way and that the analogies between data transmission in communications engineering and molecular biology are not even confined to the genetic code – but rather comprise the complete process leading from DNA sequences to proteins.

[1] B. Hayes here refers to the prominent scientific journal *Nature*, see [NPG08].

A

Notation and Symbols

A.1 Abbreviations

A	adenine
aa	amino acid
ACF	autocorrelation function
Arg	arginine
Asp	aspartic acid
AWGN	additive white Gaussian noise
bp	base pair
BPSK	binary phase shift keying
BRE	TFIIB recognition element
BSC	binary symmetric channel
C	cytosine
CDS	coding sequence
DNA	deoxyribonucleic acid
DPE	downstream promoter element
E. coli	*Escherichia coli*
ESE	exonic splice enhancer

G	guanine
His	histidine
HTH	helix-turn-helix motif
I	inosine
Ile	isoleucine
Inr	initiator region
i. u. d.	independently and uniformly distributed
kD	kilo Dalton
ld	logarithm to the base 2
Leu	leucine
Met	methionine
MF	merit factor
MFPT	mean first-passage time
mRNA	messenger RNA
PSL	peak sidelobe level
RNA	ribonucleic acid
RNAP	ribonucleic acid polymerase (RNA polymerase)
rRNA	ribosomal RNA
S	Svedberg (unit), sedimentation coefficient
SD	Shine-Dalgarno sequence
Ser	serine
SNR	signal-to-noise ratio
Stp	stop codon
SW	sync word
T	thymine
TBP	TATA-binding protein
TFIIx	transcription factor x for eukaryotic RNA polymerase II
TLS	translation start site
tRNA	transfer RNA
TSS	transcription start site
Tyr	tyrosine
UTR	untranslated region

A.2 Symbols

c	constant	page 45
d	distance between positions i_x and i_y ($d = i_y - i_x$)	page 56
$\mathbf{d}$	received data stream $\{d_1, \ldots, d_{N_d}\}$	page 5
D	length of the data part of a frame (synchronous transmission)	page 12
$\mathbf{D}$	matrix defining the multiplication of two arbitrary symbols	page 43
$\mathbf{D}_{\text{bin}}$	matrix defining the multiplication of two binary symbols	page 43
d_H	Hamming distance	page 10
$d_{\min}$	minimum distance measure	page 96
$d(n_x, n_y)$	distance measure between nucleotide n_x and nucleotide n_y	page 44
$\mathbf{D}_{\text{nuc}}$	matrix defining the multiplication of two nucleotide symbols	page 45
$\mathbf{D'}_{\text{nuc}}$	matrix defining the multiplication of nucleotides and non-constrained symbols (in distributed sync words)	page 49
$D(\hat{p}_{\mathcal{S}}(n,i) \parallel \hat{p}_{\mathcal{S}}(n))$	Kullback-Leibler divergence between the background nucleotide distribution and the actual nucleotide distribution at position i of a given set of aligned sequences	page 55
$D(\hat{p}_{\mathbf{s}}(n) \parallel \hat{p}_{\mathcal{S}(i)}(n))$	Kullback-Leibler divergence between the nucleotide distribution in a sliding window of the dataset $\mathcal{S}$ and in a short sequence motif $\mathbf{s}$	page 85
$D(p_X \parallel q_X)$	Kullback-Leibler divergence between the probability distributions $p_X(x)$ and $q_X(x)$	page 54
D_τ	distance measure used to rate the synchronization properties of a sync word depending on the shift τ	page 17
E_b	bit energy	page 7
$E(i)$	minimum binding energy over all possible spacings $s \in [15; 19]$ between the sigma factor and a DNA sequence beginning at position i of an aligned dataset	page 64
$\overline{E}(i)$	average $E(i)$ over all N sequences of the aligned dataset	page 65
$e(n)$	average binding energy between the sigma factor and nucleotide n	page 44
$e(n,k)$	partial binding energy between the nucleotide $n \in \{A, C, G, T\}$ and the binding site of the sigma factor associated with promoter position k	page 62

$E_l(s)$	binding energy between the sigma factor and a given promoter sequence l with spacing s between the -35 and the -10 region	page 62
$E(s,i)$	binding energy between the sigma factor and a DNA sequence beginning at position i of an aligned dataset for one spacing s	page 64
E_{ran}	average binding energy for a large set of random sequences	page 65
g	gradient of a function	page 69
ΔG	free energy released by a chemical reaction	page 105
h	error tolerance of the frame synchronizer	page 10
$H(n,i)$	entropy of nucleotide n at position i of a dataset	page 117
h_v	Hamming distance H between the first and the last n bits of a sequence	page 10
i	position of a dataset of aligned sequences	page 55
$\overline{I}(i;i+d)$	average mutual information between two positions at distance d	page 121
$I(\mathbf{s};\mathcal{S}(i))$	mutual information estimate between a sought motif and a sliding window of an aligned dataset $\mathcal{S}$	page 82
$I(i_x;i_x)$	mutual information between position i_x and i_y of a given set $\mathcal{S}$ of aligned sequences	page 56
$I(X;Y)$	mutual information between the random variables X and Y	page 56
J	number of codewords in a codebook	page 96
k	position in a short motif	page 5
k_B	Boltzmann constant	page 69
L	length of the sync word	page 5
$L_{\text{high}}(\mu)$	approximation of $L_{opt}(\mu)$ for high SNRs (synchronous transmission)	page 7
$L'_{\text{high}}(\mu)$	approximation of $L'_{opt}(\mu)$ for high SNRs (asynchronous transmission)	page 8
$L_{\text{low}}(\mu)$	approximation of $L_{opt}(\mu)$ for low SNRs (synchronous transmission)	page 7
$L_{\text{opt}}(\mu)$	optimum sync word location rule (synchronous transmission)	page 7
$L'_{\text{opt}}(\mu)$	optimum sync word location rule (asynchronous transmission)	page 8
L_{th}	value of the likelihood function used for threshold detection	page 5
$L(\mu)$	likelihood function to be evaluated at each position μ of the incoming data stream	page 5
Mm	Markov chain of order m	page 11
n	nucleotide	page 43
$\mathbf{n}$	DNA sequence	page 79
N	number of sequences in a dataset	page 64

N_d	length of the received or analyzed data stream	page 5
N_f	frame length (synchronous transmission)	page 6
$\widehat{N}_m(\mathbf{r})$	random variable of the count of sequence $\mathbf{r}$ based on the Markov chain Mm	page 50
$N(\mathbf{r})$	observed number of occurrences of sequence $\mathbf{r}$ in the data stream	page 50
N_0	one-sided noise spectral density	page 7
P_{CD}	probability of a correct detection of the sync word	page 10
P_e	bit error probability	page 9
P_{FD}	probability of a false detection of the sync word	page 10
p_i	probability that a protein at position i moves to the right	page 71
P_{MD}	probability of a missed detection of the sync word	page 11
$\Pr\{expr\}$	probability of the event *expr*	page 8
PSL	peak sidelobe level of sync words for application in systems with expected phase ambiguities	page 43
PSL'	peak sidelobe level of sync words for application in systems without expected phase ambiguities	page 43
$\hat{p}_{\mathcal{S}}(n)$	occurrence of nucleotide n in the whole dataset $\mathcal{S}$	page 55
$\hat{p}_{\mathcal{S}}(n,i)$	occurrence of nucleotide n at position i of a given set $\mathcal{S}$ of aligned sequences	page 55
$\hat{p}_{\mathcal{S}}(n_x,n_y,i_x,i_y)$	joint occurrence of nucleotide n_x at position i_x and nucleotide n_y at position i_y of a given set $\mathcal{S}$ of aligned sequences	page 56
q_i	probability that a protein at position i moves to the left ($q_i = 1 - p_i$)	page 71
R	performance measure for codebook models of translation initiation	page 98
s	spacing between the -35 promoter region and the -10 promoter region	page 62
$\mathbf{s}$	sync word $\{s_1,\ldots,s_L\}$	page 5
$\overline{S}(i)$	average homology score over all sequences $\mathbf{n}$ of an aligned dataset	page 80
$S(\mathbf{n})$	homology score between a given consensus sequence and a nucleotide sequence $\mathbf{n}$	page 79
$S(\mathbf{n}(i),\mathbf{s})$	homology score between a given consensus sequence $\mathbf{s}$ and a subsequence of $\mathbf{n}$ starting at position i	page 80
t	control variable	
T	temperature	page 69
T_l	distance between the current sync word SW_l and the successive sync word SW_{l+1}	page 8

$\bar{t}_{0,x}$	mean first-passage time from position $i = 0$ to position $i = x$	page 71
$\bar{t}'_{0,x}$	simplified mean first-passage time from position $i = 0$ to position $i = x$	page 72
u	number of sequences that may yield a false synchronization under error tolerance h	page 11
v	self-overlap of a sequence	page 10
$w_{i,i\pm1}$	transition rate from position i to position $i \pm 1$ (of a protein during its movement along the DNA)	page 69
$\mathbf{W}(n,k)$	weight matrix score of nucleotide n at position k of a short motif	page 79
x	control variable	
y	control variable	
z	control variable	

α_i	ratio of p_i and q_i	page 71
β	inverse product of k_B and T	page 69
δ	distance measure	page 96
$\Delta_{\max}$	maximum possible strength of a detection signal	page 98
Δ_{act}	actual strength of a detection signal	page 98
$\mathbb{E}\{.\}$	expected value	page 50
μ_s	position that the receiver evaluates to be the position of the sync word	page 5
ν	affective attempt rate of a protein moving to one of its neighboring sites	page 69
ρ_i	escape rate at position i (of a protein during its movement along the DNA)	page 70
τ_i	time a protein spends bound to position i	page 70
$\varphi_{ss}(\tau)$	aperiodic autocorrelation function of the sync word $\mathbf{s}$ (τ denotes the shift of $\mathbf{s}$ against itself)	page 15
$\widetilde{\varphi}_{ss}(\tau)$	adapted autocorrelation function for bacterial promoter sequences	page 45
$\mathcal{A}$	alphabet	page 9
$\lvert\mathcal{A}\rvert$	cardinality of the alphabet $\mathcal{A}$	page 9
$\mathbb{R}$	set of all real numbers	page 51
$\mathcal{S}$	given set of aligned DNA or mRNA sequences	page 54
$\mathcal{S}(i)$	sliding window (length L) of the aligned dataset $\mathcal{S}$ starting at position i	page 82
$\mathbf{1}\{expr\}$	equals one if *expr* is true and zero otherwise	page 14
$*$	unconstrained symbol of a distributed sync word	page 20

B

Sync Word Families

B.1 Barker sequences

Table B.1: Barker sequences [Lue92].

L	**sync word**
2	10
3	110
4	1101 and 1110
5	11101
7	1110010
11	11100010010
13	1111100110101

B.2 Sequences found by Maury and Styles

Table B.2: Binary sync words for channels without phase ambiguities taken from [Rob95], results of search by Maury and Styles [MS64].

L	**sync word**	L	**sync word**
7	1011000	19	1111100110010100000
8	10111000	20	11101101111000100000
9	101110000	21	111011101001011000000
10	1101110000	22	1111001101101010000000
11	10110111000	23	10110101101011010000000
12	110101100000	24	111110101111001100100000
13	1110101100000	25	111110010110111000100000
14	11100110100000	26	111110100110100110010000000
15	111011001010000	27	111110101101001100110000000
16	1110101110010000	28	111101011110010110011000000
17	11110011010100000	29	1111010111100110011010000000
18	111100110101000000	30	11111010111100110011010000000

B.3 Sequences found by Neuman and Hofman

Table B.3: Binary sync words for channels with expected phase ambiguities (type II), results of search by Neuman and Hofman [NH71].

L	**sync word**	L	**sync word**
7	0001101	16	0000011001101011
8	00011101	17	00001011001110101
9	000011101	18	000010101101100111
10	0000110101	19	0001110111011011010
11	00011101101	20	00010001111100101101
12	000111101101	21	000000101110100111001
13	0000010110011	22	000100011111011011010
14	00001100110101	23	00000010101100110100111
15	001111100110101	24	000001110011101010110110

Table B.4: Binary sync words for channels without phase ambiguities (type I), results of search by Neuman and Hofman [NH71].

L	**sync word**	L	**sync word**
7	0001101	16	0000011001101011
8	00011101	17	00001011001110101
9	000011101	18	000010101101100111
10	0000110101	19	0001110111011011010
11	00011101101	20	00010001111100101101
12	000111101101	21	000000101110100111001
13	0000010110011	22	000100011111011011010
14	00001100110101	23	00000010101100110100111
15	001111100110101	24	000001110011101010110110

B.4 Bifix-free sequences

Table B.5: Bifix-free sequences [Lue92].

L	**sync word**
2	10
3	100 and 110
4	1000 and 1100 and 1110
5	10000 and 10100 and 11000 and 11100 and 11010 and 11110
6	100000 and 101000 and 101100 and 110000 and 110100 and 111000 and 111100 and 110010 and 111010 and 111110

B.5 Distributed sequences

Table B.6: Bifix-free distributed sequences with PSL ≤ 1 [dLvW98].

L	**sync word**	L	**sync word**
5	110*0	24	11****110**0**1****0*0*0
7	110*0*0		110*11******0*0**1***0*0
	1110**0		111**0***0***0***0*10**0
10	1110**0**0		11**11*******0*01**1*0*0
13	111**0***0*10		11**11*******0*0*1**10*0
	11****110*0*0	28	111**0*0****0****0******0110
17	111**0***0***0*10		11**110*****0***10*0*****0*0
	11**110*0*****0*0	32	111**0*0*0****0****0********0110
20	111**0***0***0*10**0		111**0*0****0***0****0******0110
	110*11***0*0*****0*0		11**0*10*1******1****1***0*0*0*0
	110***1**1****10*0*0		

C

Sequence Data and Implementation Details

C.1 Datasets

Throughout this thesis, large datasets of DNA / mRNA sequences are used to prove the applicability of the derived models. This section aims at detailing these datasets, their accessability and their statistical properties.

C.1.1 Promoters of *Escherichia coli*

Two data bases are used for the extraction of *E. coli* promoters: GenBank of the National Institute for Biotechnology Information (NCBI, [NCfBI08]) and RegulonDB of the Centro de Ciencas Genómicas [CdCG07]. Throughout the thesis, it is clearly stated which of the data bases is currently used. Both offer a high number of promoters for the main sigma factor σ^{70} but only few promoters of alternative sigma factors. Generally, the datasets of RegulonDB are frequently updated and expanded and thus are the preferable choice.

RegulonDB

The data base RegulonDB offers a text-file containing experimentally verified promoters of *E. coli* including their respective sigma factor. The promoter sequences are given together with their position, the strand (forward or reverse), the gene they belong to as well as the

type of experimental verification that resulted in their annotation. The mentioned dataset currently contains 651 σ^{70}-promoters and a total of 230 promoters for the six alternative sigma factors (version 5.7). Moreover, RegulonDB offers a large set of computationally predicted promoters. For more information on RegulonDB see [SGCPG+06].

National Center for Biotechnology Information (NCBI)

The whole genome of *E. coli* (strain K-12) is available under GenBank entry u00096. This file contains information on genes but not on promoters. These can be extracted from the 400 subfiles AE000x.1 with $x \in [111; 510]$ (accessible via 400 links in the file u00096) using a parser. The 400 files each contain more detailed information on a part of around 11500 bp of the genome, e.g. promoters together with their respective sigma factor and protein binding sites. The promoters are classified into experimentally documented and computationally predicted. Their position with respect to the current part of the genome, the strand (forward or reverse) and the gene they belong to are given. The 400 files contain 3765 promoters in total for σ^{70} of which only a small number are experimentally documented. Only few promoters are annotated for alternative sigma factors (i.e. other than σ^{70}). Since the 400 files are not frequently updated, this promoter set should only be preferred over the RegulonDB promoters if a very large dataset is required.

Transcription rates

The transcription rate refers to the number of transcripts obtained from a gene under certain circumstances, i.e. how often has the gene been transcribed in the measured time period. These rates are determined using microarray experiments, where single RNAs are attached to the surface of a chip that is exposed to fluoridated cell mRNAs. Those bind to their counterpart on the chip and yield a fluorescence signal after readout [BH02]. An extensive database for microarray experiments is ASAP (A Systematic Annotation Package for Community Analysis of Genomes) [UoWM07] (see [GLP+03] for more information). It offers microarray data for the *E. coli* genome (strain K-12) under different growth conditions and using different chips. Throughout this thesis, the focus lies on standard growth conditions, i.e. the conditions under which σ^{70} is active (see Section 3.4.2). Therefore, the results under standard growth conditions obtained with an Affymetrix chip are used.

C.1.2 Eukaryotic promoters

The Eukaryotic Promoter Database (EPD, [SIoB07]) currently offers 4809 experimentally verified promoter sequences of different eukaryotic species. Most of these belong to *Drosophila melanogaster* (the fruit fly, 1926 sequences) and the human (1871 sequences). Throughout the thesis, the human promoters are used as one dataset and the arthropod promoters as a second one. Arthropods are the largest phylum of animals and include insects like *Drosophila melanogaster*. They can be downloaded as a text file together with

their strand (forward or reverse) and the gene they belong to. See [SPPB06] for more information on EPD.

C.1.3 mRNAs of *Escherichia coli*

The mRNA sequences of *E. coli* are extracted from the same 400 subfiles of the GenBank entry u00096 as the promoters (see Section C.1.1, [NCfBI08]). A parser is used to obtain the position and strand (forward or reverse) of the translated sequences. A total of 3194 mRNA sequences are available.

C.1.4 Eukaryotic mRNAs

The eukaryotic mRNA sequences are downloaded from the UniGene database [NCfBI07], hosted by the National Center for Biotechnology Information (NCBI). It offers mRNA sequences of almost 100 eukaryotic species that can be handily downloaded from the ftp-site (ftp://ftp.ncbi.nih.gov/repository/UniGene/). Among those are 122000 sequence entries of different types for *Homo sapiens* (human), 77000 entries for *Mus musculus* (house mouse) and 17000 entries for *Drosophila melanogaster* (fruit fly).

C.2 Data access and processing

All implementations were done in Matlab using the Bioinformatics Toolbox. The latter offers functions for downloading, converting and analyzing DNA sequences. Most databases provide their sequences in the FASTA format, a text-based format for representing nucleotide or amino acid sequences. These files can be easily imported into MATLAB using the function `fastaread`.

▷ **Example C.1**

A FASTA entry typically looks like this:

> gi | 48994873 | gb | U00096.2 | Escherichia coli K12 MG1655, complete genome
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAG
TTGTCTGATAGCAGCTTCTGAAC...

The first row identifies the sequence in terms of accession numbers (here: 48994873, U00096.2), the organism (here: *Escherichia coli*), the strain (here: K12 MG1655) and a short description (here: complete genome). Using the function `fastaread`, the FASTA file is converted to a Matlab structure with the fields `Header` (containing the identifier) and `Sequence`.

◁

C.3 Nucleotide composition of the eukaryotic promoter datasets

C.3.1 Human promoter surrounding

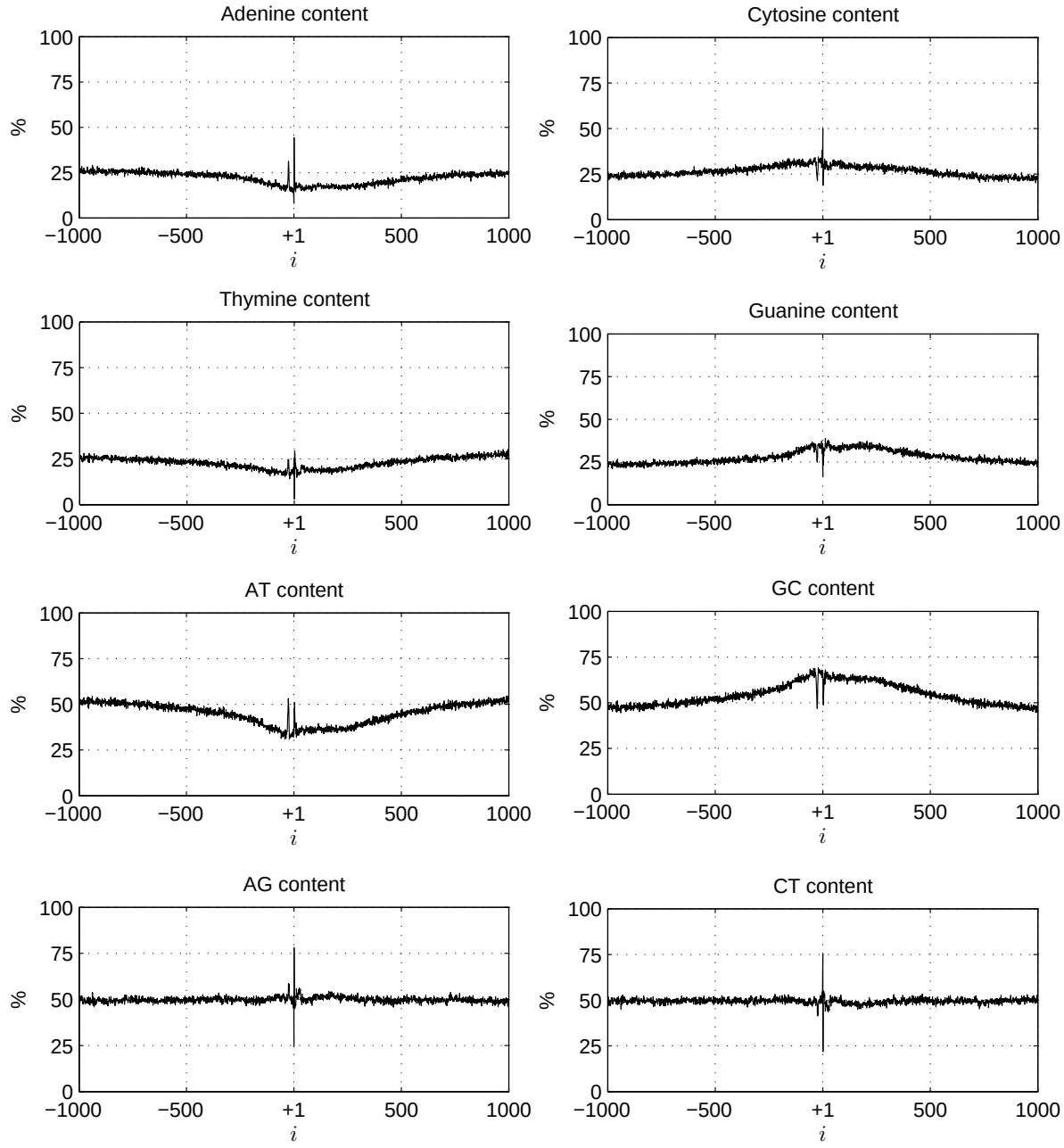


Figure C.1: Nucleotide composition around promoters of the human EPD dataset.

C.3.2 Arthropod promoter surrounding

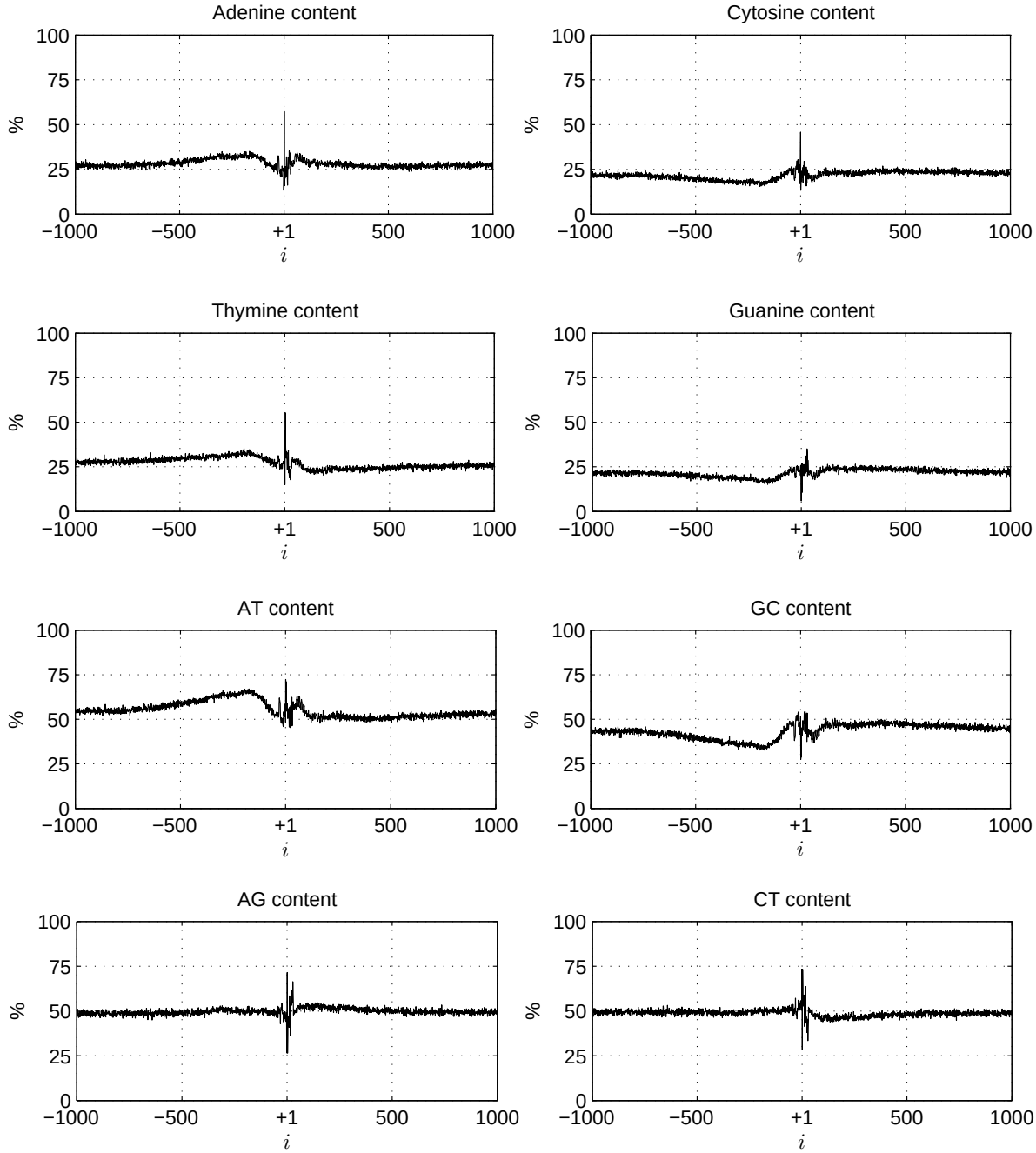


Figure C.2: Nucleotide composition around promoters of the arthropod EPD dataset.

D

Derivations

D.1 Escape rate

During the sliding process of a protein along the DNA, it is assumed to move in single nucleotide steps. The escape rate of the protein at site i to one of the neighboring sites is given by [SM04]

$$\rho_i = \frac{1}{\tau_i} = w_{i,i+1} + w_{i,i-1}, \tag{D.1}$$

where $w_{i,i+1}$ and $w_{i,i-1}$ are defined by the Arrhenius equation as

$$w_{i,i\pm1} = \nu \cdot \begin{cases} e^{-\beta[E(i\pm1)-E(i)]} & \text{if } E(i\pm1) > E(i) \\ 1 & \text{otherwise} \end{cases}. \tag{D.2}$$

A case differentiation for the gradient g is now conducted in order to derive a closed expression for the escape rate ρ_i:

Case 1: $E(i+1) > E(i)$, i.e., $g = E(i+1) - E(i) > 0$

$$\begin{aligned} \rho_i &= w_{i,i+1} + w_{i,i-1} \\ &= \nu \cdot e^{-\beta[E(i+1)-E(i)]} + \nu \\ &= \nu \cdot e^{-\beta g} + \nu \\ &= \nu(e^{-\beta g} + 1). \end{aligned} \tag{D.3}$$

Case 2: $E(i+1) < E(i)$, i.e., $g = E(i+1) - E(i) = -[E(i-1) - E(i)] < 0$

$$\begin{aligned} \rho_i &= w_{i,i+1} + w_{i,i-1} \\ &= \nu + \nu \cdot e^{-\beta[E(i-1)-E(i)]} \end{aligned}$$

Since g is constant in each of the four regions of the approximation (see Figure 5.8), it can easily be deduced that

$$g = \overline{E}(i+1) - \overline{E}(i) = -\big[\overline{E}(i-1) - \overline{E}(i)\big], \tag{D.4}$$

which leads to

$$\begin{aligned} \rho_i &= \nu \cdot e^{-\beta(-g)} + \nu \\ &= \nu(e^{\beta g} + 1). \end{aligned} \tag{D.5}$$

Case 3: $E(i+1) = E(i)$, i.e., $g = 0$

$$\begin{aligned} \rho_i &= w_{i,i+1} + w_{i,i-1} \\ &= \nu + \nu \\ &= 2\nu. \end{aligned} \tag{D.6}$$

Straightforwardly, the three cases (Eq. (D.3), Eq. (D.5), Eq. (D.6)) combine to

$$\rho_i = \nu(e^{-\beta|g|} + 1). \tag{D.7}$$

D.2 Mean first-passage time

The mean-first passage time (MFPT) is defined as the number of steps the protein makes to reach from site $i = 0$ to site $i = x$ if assuming a certain set of transition probabilities α_i. Then, the MFPT is given by [SM04] is

$$\bar{t}_{0,x} = x + \sum_{k=0}^{x-1} \alpha_k + \sum_{k=0}^{x-2} \sum_{i=k+1}^{x-1} (1 + \alpha_k) \prod_{j=k+1}^{i} \alpha_j. \tag{D.8}$$

Since the values of α_i are constant over sufficiently wide ranges, $\alpha_0 = \alpha_1 = \cdots = \alpha_x := \alpha$ is assumed. Plugging this into Eq. (D.8) results in

$$\bar{t}_{0,x}^{'}=x+x\cdot\alpha+(1+\alpha)\sum_{k=0}^{x-2}\sum_{i=k+1}^{x-1}\alpha^{i-k}, \tag{D.9}$$

which leads with

$$\begin{aligned}\sum_{k=0}^{x-2}\sum_{i=k+1}^{x-1}\alpha^{i-k} &= (\alpha^1+\ldots+\alpha^{x-1})+(\alpha^1+\ldots+\alpha^{x-2})+\ldots+(\alpha^1)\\ &= (x-1)\cdot\alpha^1+\ldots+1\cdot\alpha^{x-1}\\ &= \sum_{k=1}^{x-1}(x-k)\alpha^k\end{aligned}$$

to

$$\bar{t}_{0,x}^{'} = x+x\cdot\alpha+(1+\alpha)\sum_{k=1}^{x-1}(x-k)\alpha^k \tag{D.10}$$

$$= (1+\alpha)\sum_{k=0}^{x-1}(x-k)\alpha^k. \tag{D.11}$$

Bibliography

[AJL+02] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland Science, New York, fourth edition, 2002.

[AKL+07] H. M. Aktulga, I. Kontoyiannis, L. A. Lyznik, L. Szpankowski, A. Y. Grama, and W. Szpankowski. Identifying statistical dependence in genomic sequences via mutual information estimates. *EURASIP J Bioinform Syst Biol*, 2007:14741, 2007.

[AM90] D. G. Arquès and C. J. Michel. Periodicities in coding and noncoding region of the genes. *J Theor Biol*, 143(3):307–318, April 1990.

[Bar53] R. H. Barker. Group synchronization of binary digital systems. *Commun Theory*, pages 273–287, 1953.

[Bat04] G. Battail. An engineer's view on genetic information and biological evolution. *Biosystems*, 76(1-3):279–290, August-October 2004.

[Bat06] G. Battail. Should genetics get an information-theoretic education? Genomes as error-correcting codes. *IEEE Eng Med Biol Mag.*, 25(1):34–45, January-February 2006.

[BBDI+06] M. Bailly-Bechet, A. Danchin, M. Igbal, M. Marsili, and M. Vergassola. Codon usage domains over bacterial chromosomes. *PLoS Comput Biol*, 2(4):e37, April 2006.

[BGZY99] C. Bustamante, M. Guthold, X. Zhu, and G. Yang. Facilitated target location on DNA by individual *Escherichia coli* RNA polymerase molecules observed with the scanning force microscope operating in liquid. *J Biol Chem*, 274(24):16665–16668, June 1999.

[BH02] P. Baldi and G. W. Hatfield. *DNA microarrays and gene expression*. Cambridge University Press, Cambridge, 2002.

[BIO08] BIOBASE. BIOBASE Biological Databases: TRANSFAC Gene Transcription Factor Database. http://www.biobase-international.com/pages/index.php?id=transfac, May 2008.

[BLZ05] S. Burden, Y.-X. Lin, and R. Zhang. Improving promoter prediction for the NNPP2.2 algorithm: a case study using *Escherichia coli* DNA sequences. *Bioinformatics*, 21(5):601–607, March 2005.

[Boe67] A. B. Boehmer. Binary pulse compression codes. *IEEE Trans Inform Theory*, 13:156–167, April 1967.

[BPPS04] M. Barbi, C. Place, V. Popkov, and M. Salerno. Base-sequence-dependent sliding of proteins on DNA. *Phys Rev E Stat Nonlin Soft Matter Phys*, 70(4 Pt 1):041901, October 2004.

[Buc90] P. Bucher. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J Mol Biol*, 212(4):563–578, April 1990.

[BWvH81] O. G. Berg, R. B. Winter, and P. H. von Hippel. Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. *Biochemistry*, 20(24):6929–6948, November 1981.

[CB07] H. Chen and M. Blanchette. Detecting non-coding selective pressure in coding regions. *BMC Evol Biol*, 7(Suppl 1):S9, February 2007.

[CBKJ94] H. Chen, M. Bjerknes, R. Kumar, and E. Jay. Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of *Escherichia coli* mRNAs. *Nucleic Acids Res*, 22(23):4953–4957, November 1994.

[CdCG07] Centro de Ciencas Genómicas, Universidad Nacional Autonóma de México. RegulonDB 5.7. http://regulondb.ccg.unam.mx/, August 2007.

[CfMBoR08] Center for Molecular Biology of RNA, University of California Santa Cruz. RNA Center : Ribosome Images. http://rna.ucsc.edu/rnacenter/ribosome_images.html, April 2008.

[CH07] G. M. Cooper and R. E. Hausman. *The Cell: A Molecular Approach*. Sinauer Associates, fourth edition, 2007.

[CLM+01] E. Chiu, J. Lin, B. Mcferron, N. Petigara, and S. Seshasai. Mathematical theory of Claude Shannon. Report of a student project conducted in the framework of the course 'The Structure of Engineering Revolutions', http://users.ece.utexas.edu/∼adnan/syn-07/shannon1.pdf, Massachusetts Institute of Technology, December 2001.

[CPCB+94] H. Chen, L. Pomeroy-Cloney, M. Bjerknes, J. Tam, and E. Jay. The influence of adenine-rich motifs in the 3' portion of the ribosome binding site on human IFN-gamma gene expression in Escherichia coli. *J Mol Biol*, 240(1):20–27, July 1994.

[CT91] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications. John Wiley & Sons, Inc., 1991.

[DBSH07] T. A. Down, C. M. Bergman, J. Su, and T. J. Hubbard. Large-scale discovery of promoter motifs in *Drosophila melanogaster*. *PLoS Comput Biol*, 3(1):e7, January 2007.

[DCZM06] J. Dresios, S. A. Chappell, W. Zhou, and V. P. Mauro. An mRNA-rRNA base-pairing mechanism for translation initiation in eukaryotes. *Nat Struct Mol Biol*, 13(1):30–34, January 2006.

[DGH+06] Z. Dawy, B. Goebel, J. Hagenauer, C. Andreoli, T. Meitinger, and J. C. Mueller. Gene mapping and marker clustering using Shannon's mutual information. *IEEE/ACM Trans Comput Biol Bioinform*, 3(1):47–56, January-March 2006.

[DGHM05] Z. Dawy, F. M. Gonzalez, J. Hagenauer, and J. C. Mueller. Modeling and analysis of gene expression mechanisms: a communication theory approach. *Proc. IEEE Internat Conf Commun (ICC)*, 2:815–819, May 2005.

[DHW+07] Z. Dawy, P. Hanus, J. Weindl, J. Dingel, and F. Morcos. On genomic coding theory. *Europ Trans Telecommun*, 18(8):873–879, November 2007.

[DJLG96] A. J. Dombroski, B. D. Johnson, M. Lonetto, and C. A. Gross. The sigma subunit of *Escherichia coli* RNA polymerase senses promoter spacing. *Proc Natl Acad Sci U S A*, 93:8858–8862, August 1996.

[dLvW98] A. J. de Lind van Wijngaarden. *Frame synchronization techniques*. PhD thesis, Universität GH Essen, Germany, March 1998.

[dLvWW00] A. J. de Lind van Wijngaarden and T. J. Willink. Frame synchronization using distributed sequences. *IEEE Trans Commun*, 48(12):2127–2138, December 2000.

[DMWM09] Z. Dawy, F. Morcos, J. Weindl, and J. C. Mueller. Translation initiation modeling and mutational analysis based on the 3'-end of the *Escherichia coli* 16S rRNA sequence. *Biosystems*, in press, 2009.

[DoEOoS08] U. S. Department of Energy Office of Science. Human genome project information. http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml, May 2008.

[DSS03] M. Djordjevic, A. M. Sengupta, and B. I. Shraiman. A biophysical approach to transcription factor binding site discovery. *Genome Res*, 13:2381–2390, November 2003.

[Ebr00] R. H. Ebright. RNA polymerase: structural similarities between bacterial RNA polymerase and eukaryotic RNA polymerase II. *J Mol Biol*, 304:687–698, 2000.

[FCGD96] M. A. Firpo, M. B. Connelly, D. J. Goss, and A. E. Dahlberg. Mutations at two invariant nucleotides in the 3'-minor domain of *Escherichia coli* 16S rRNA affecting translational initiation and initiation factor 3 function. *J Biol Chem*, 271(9):4693–4698, March 1996.

[Fel68] W. Feller. *An introduction to probability theory and its applications*, volume 1 of *Wiley series in probability and mathematical statistics*. John Wiley & Sons, New York, third edition, 1968.

[FKJ+86] S. M. Freier, R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Neilson, and D. H. Turner. Improved free-energy parameters for predictions of RNA duplex stability. *Proc Natl Acad Sci U S A*, 83(24):9373–9377, December 1986.

[For81] D. R. Forsdyke. Are introns in-series error-detecting sequences? *J Theor Biol*, 93(4):861–866, December 1981.

[GDHM05] B. Goebel, Z. Dawy, J. Hagenauer, and J. C. Mueller. An approximation to the distribution of finite sample size mutual information estimates. *Proc IEEE Internat Conf Commun (ICC)*, 2:1102–1106, May 2005.

[GG03] T. M. Gruber and C. A. Gross. Multiple sigma subunits and the partitioning of bacterial transcription space. *Annu Rev Microbiol*, 57:441–466, October 2003.

[GGML99] A. J. F. Griffith, W. M. Gelbart, J. H. Miller, and R. C. Lewontin. *Modern Genetic Analysis*. W. H. Freeman Publishers, New York, 1999.

[GHBS00] I. Grosse, H. Herzel, S. V. Buldyrev, and H. E. Stanley. Species independence of mutual information in coding and noncoding DNA. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics*, 61(5 Pt B):5624–5629, May 2000.

[GHMD91] H. U. Göringer, K. A. Hijazi, E. J. Murgola, and A. E. Dahlberg. Mutations in 16S rRNA that affect UGA (stop codon)-directed translation termination. *Proc Natl Acad Sci U S A*, 88(15):6603–6607, August 1991.

[GLP+03] J. D. Glasner, P. Liss, G. III. Plunkett, A. Darling, T. Prasad, M. Rusch, A. Byrnes, M. Gilson, B. Biehl, F. R. Blattner, and N. T. Perna. ASAP, a systematic annotation package for community analysis of genomes. *Nucleic Acids Res*, 31(1):147–151, January 2003.

[Gol77] M. J. E. Golay. Sieves for low autocorrelation binary sequences. *IEEE Trans Inf Theory*, 23(1):43–51, January 1977.

[Gon04] Faruck Morcos Gonzáles. On the existence and relevance of coding theory models for genetic regulatory systems. Master's thesis, Technische Universität München, München, September 2004.

[GvH05] S. J. Greive and P. H. von Hippel. Thinking quantitatively about transcriptional regulation. *Nat Rev Mol Cell Biol*, 6(3):221–232, March 2005.

[GZR+99] M. Guthold, X. Zhu, C. Rivetti, G. Yang, N. H. Thomson, S. Kasas, H. G. Hansma, and B. Smith. Direct observation of one-dimensional diffusion and transcription by *Escherichia coli* RNA polymerase. *Biophys J*, 77(4):2284–2294, October 1999.

[Hay98] B. Hayes. The invention of the genetic code. *Comput Sci*, 86(1):8–14, January-February 1998.

[HdB87] A. Hui and H. A. de Boer. Specialized ribosome system: Preferential translation of a single mRNA species by a subpopulation of mutated ribosomes in *Escherichia coli*. *Proc Natl Acad Sci U S A*, 84(14):4762–4766, July 1987.

[HDG+04] J. Hagenauer, Z. Dawy, B. Goebel, P. Hanus, and J. Mueller. Genomic analysis using methods from information theory. *Proc IEEE Inf Theory Workshop (ITW)*, pages 55–59, October 2004.

[HFM+99] Y. Harada, T. Funatsu, K. Murakami, Y. Nonoyama, A. Ishihama, and T. Yanagida. Single-molecule imaging of RNA polymerase-DNA interactions in real time. *Biophys J*, 76(2):709–715, February 1999.

[HGD+07] P. Hanus, B. Goebel, J. Dingel, J. Weindl, J. Zech, Z. Dawy, J. Hagenauer, and J. C. Mueller. Information and communication theory in molecular biology. *Electr Eng (Archiv für Elektrotechnik)*, 90(2):161–173, December 2007.

[JB04] K. Jabbari and G. Bernardi. Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. *Gene*, 333:143–149, May 2004.

[JLB+96] Z. S. Juo, P. M. Leiberman, I. Baikalov, A. J. Berk, and R. E. Dickerson. How proteins recognize the TATA box. *J Mol Biol*, 261(2):239–254, August 1996.

[JTZ89] J. A. Jaeger, D. H. Turner, and M. Zuker. Improved prediction of secondary strucure for RNA. *Proc Natl Acad Sci U S A*, 86(20):7706–7710, October 1989.

[KB04] A. Kopansky and M. Bystrom. Detection of aperiodically embedded synchronization patterns. *IEEE Trans Wireless Commun*, 3(5):1386–1392, September 2004.

[KMC+06] E. Kierzek, D. H. Mathews, A. Ciesielska, D. H. Turner, and R. Kierzek. Nearest neighbor parameters for Watson-Crick complementary heteroduplexes formed between 2'-*O*-methyl RNA and RNA oligonucleotides. *Nucleic Acids Res*, 34(13):3609–3614, July 2006.

[KNI90] M. Kobayashi, K. Nagata, and A. Ishihama. Promoter selectivity of *Escherichia coli* RNA polymerase: effect of base substitutions in the promoter -35 region on promoter strength. *Nucleic Acids Res*, 18(24):7367–7372, December 1990.

[KOA05] H. Kiryu, T. Oshima, and K. Asai. Extracting relations between promoter sequences and their strengths from microarray data. *Bioinformatics*, 21(7):1062–1068, October 2005.

[Koz97] M. Kozak. Recognition of AUG and alternative initiator codons is augmented by G in position +4 but is not generally affected by the nucleotides in positions +5 and +6. *EMBO J*, 16(9):2482–2492, May 1997.

[Koz99] M. Kozak. Initiation of translation in prokaryotes and eukaryotes. *Gene*, 234(2):187–208, July 1999.

[Koz01] M. Kozak. New ways of initiating translation in eukaryotes? *Mol Cell Biol*, 21(6):1899–1907, March 2001.

[Koz02] M. Kozak. Pushing the limits of the scanning mechanism for initiation of translation. *Gene*, 299(1-2):1–34, October 2002.

[Lat04] D. S. Latchman. *Eukaryotic Transcription Factors*. Academic Press, fourth edition, 2004.

[LBZ+00] H. Lodish, A. Berk, L. Zipursky, P. Matsudaira, D. Baltimore, and J. E. Darnell. *Molecular Cell Biology*. W. H. Freeman Publishers, New York, fourth edition, 2000.

[Lev75] B. K. Levitt. Long frame sync words for binary PSK telemetry. *IEEE Trans Commun*, 23:1365–1367, November 1975.

[Lew07] B. Lewin. *Genes IX*. Jones and Bartlett Publishers, Sudbury, Massachusetts, 2007.

[LM93] S. Lisser and H. Margalit. Compilation of *E. coli* mRNA promoter sequences. *Nucleic Acids Res*, 21(7):1507–1516, March 1993.

[Lue92] H. D. Lueke. *Korrelationssignale*. Springer-Verlag, Berlin, 1992.

[LVYP05] S. J. Lolle, J. L. Victor, J. M. Yound, and R. E. Pruitt. Genome-wide non-mendelian inheritance of extra-genomic information in *Arabidopsis*. *Nature*, 434(7032):505–509, March 2005.

[Mas72] J. L. Massey. Optimum frame synchronization. *IEEE Trans Commun*, 20(2):115–119, April 1972.

[MBM85] M. E. Mulligan, J. Brosius, and W. R. McClure. Characterization *in vitro* of the effect of spacer length on the activity of *Escherichia coli* RNA polymerase at the TAC promoter. *J Biol Chem*, 260(6):3529–3538, March 1985.

[MK89] K. P. N. Murthy and K. W. Kehr. Mean first-passage time of random walks on a random lattice. *Phys Rev A*, 40(4):2082–2087, August 1989.

[Moo08] J. H. Moore. Bases, bits and disease: a mathematical theory of human genetics. *Eur J Human Genet*, 16(2):143–144, February 2008.

[MS64] J. L. Maury and F. J. Styles. Development of optimum frame synchronization codes for Goddard Space Flight Center PCM telemetry standards. *Proc Nat Telemet Conf*, pages 889–899, June 1964.

[MVB06] E. E. May, M. A. Vouk, and D. L. Blitzer. Classification of *Escherichia coli* K-12 ribosome binding sites - an error-control coding model. *IEEE Eng Med Biol Mag*, 25(1):90–97, January-February 2006.

[MVBR04] E. E. May, M. A. Vouk, D. L. Blitzer, and D. I. Rosnick. Coding theory based models for protein translation initiation in prokaryotic organisms. *Biosystems*, 76(1-3):249–260, August-October 2004.

[NCfBI07] National Center for Biotechnology Information. UniGene Home. http://www.ncbi.nlm.nih.gov/sites/entrez?db=unigene, July 2007.

[NCfBI08] National Center for Biotechnology Information. NCBI HomePage. http://www.ncbi.nlm.nih.gov/, May 2008.

[NH71] F. Neuman and L. Hofman. New pulse sequences with desirable correlation properties. *Proc Nat Telemet Conf*, pages 277–282, April 1971.

[Nie73a] P. T. Nielsen. A note on bifix-free sequences. *IEEE Trans Inf Theory*, 19(5):704– 706, September 1973.

[Nie73b] P. T. Nielsen. Some optimum and suboptimum frame synchronizers for binary data in Gaussian noise. *IEEE Trans Commun*, 21(6):770–772, June 1973.

[NIoH08a] National Institutes of Health. genome.gov | National Human Genome Research Institute. http://www.genome.gov/, May 2008.

[NIoH08b] National Institutes of Health. genome.gov | Talking Glossary. http://www.genome.gov/10002096, March 2008.

[Nog00] E. Nogales. Recent structural insights into transcription preinitiation complexes. *J Cell Sci*, 113(24):4391–4397, December 2000.

[NPG08] Nature Publishing Group. Journal home : Nature. http://www.nature.com/nature/index.html, March 2008.

[OST06] Y. Osada, R. Saito, and M. Tomita. Comparative analysis of base correlations in 5' untranslated regions of various species. *Gene*, 375:80–86, June 2006.

[Pea06] H. Pearson. Genetics: What is a gene? *Nature*, (441):398–401, May 2006.

[PG90] C. D. Prescott and H. U. Göringer. A single mutation in 16S rRNA that affects mRNA binding and translation-termination. *Nucleic Acids Res*, 18(18):5381–5386, September 1990.

[PG02] M. Ptashne and A. Gann. *Genes & Signals*. Cold Spring Harbor Laboratory Press, New York, 2002.

[PKS+99] G. A. Patikoglou, J. L. Kin, L. Sun, S.-H. Yang, T. Kodadek, and S. K. Burley. TATA element recognition by the TATA box-binding protein has been conserved through evolution. *Genes Dev*, 13(24):3217–3230, December 1999.

[RBC70] A. D. Riggs, S. Bourgeois, and M. Cohn. The lac repressor-operator interaction. 3. Kinetic studies. *J Mol Biol*, 53(3):401–417, November 1970.

[RH05] J. C. Rajapakse and L. S. Ho. Markov encoding for detecting signals in genomic sequences. *IEEE/ACM Trans Comput Biol Bioinform*, 2(2):131–142, April-June 2005.

[Rob95] P. Robertson. *Optimal frame synchronization for continuous and packet data transmission, Fortschritt-Berichte VDI.* 10, no. 376. VDI-Verlag, Düsseldorf, 1995.

[RRS05] S. Robin, F. Rodolphe, and S. Schbath. *DNA, Words and Models*. Cambridge University Press, Cambridge, English edition, 2005.

[RSV07] S. Robin, S. Schbath, and V. Vandewalle. Statistical tests to compare motif count exceptionalities. *BMC Bioinformatics*, 8:84, March 2007.

[SBB06] S. Saxono, P. Berg, and D. L. Brutlag. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Sci U S A*, 103(5):1412–1417, January 2006.

[SBRS01] R. K. Shultzaberger, R. E. Bucheimer, K. E. Rudd, and T. D. Schneider. Anatomy of *Escherichia coli* ribosome binding sites. *J Mol Biol*, 313(1):215–228, October 2001.

[Sch80] R. A. Scholtz. Frame synchronization techniques. *IEEE Trans Commun*, COM-28(8):1204–1213, August 1980.

[Sch96] T. D. Schneider. Reading of DNA sequence logos: prediction of major groove binding by information theory. *Methods Enzymol*, 274:445–455, 1996.

[Sch97] T. D. Schneider. Information content of individual genetic sequences. *J Theor Biol*, 189(4):427–441, December 1997.

[Sch06] S. Schbath. Statistics of motifs. In *Atelier de formation*, volume 1502. Institute national de la santé et de la recherche médicale, April 2006.

[SCLS07] R. K. Shultzaberger, Z. Chen, K. A. Lewis, and T. D. Schneider. Anatomy of *Escherichia coli* σ^{70} promoters. *Nucleic Acids Res*, 35(3):771–788, February 2007.

[SD74] J. Shine and L. Dalgarno. The 3'-terminal sequence of *Escherichia coli* 16S ribosomal RNA: Complementarity to nonsense triplets and ribosome binding sites. *Proc Natl Acad Sci U S A*, 71(4):1342–1346, April 1974.

[SDS02] A. M. Sengupta, M. Djordjevic, and B. I. Shraiman. Specificity and robustness in transcription control networks. *Proc Natl Acad Sci U S A*, 99(4):2072–2077, February 2002.

[SF98] G. D. Stormo and D. S. Fields. Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem Sci*, 23(3):109–113, March 1998.

[SFMC+06] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thaström, Y. Field, I. K. Moore, J. P. Wang, and J. Widom. A genomic code for nucleosome positioning. *Nature*, 442(7104):772–778, August 2006.

[SfSB07] Statistics for Systems Biology, L'Institut National de la Recherche Agronomique. R'MES. http://genome.jouy.inra.fr/ssb/rmes/, August 2007.

[SGCPG+06] H. Salgado, S. Gama-Castro, M. Peralta-Gil, E. Diaz-Peredo, F. Sanchez-Solano, A. Santos-Zavaleta, I. Matrinez-Flores, V. Jimenez-Jacinto, C. Bonavides-Martinez, J. Sequra-Salazar, A. Martinez-Antonio, and J. Collado-Vides. RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res*, 34(Database issue):D394–D397, January 2006.

[SH89] G. D. Stormo and G. W. Hartzell III. Identifying protein-binding sites from unaligned DNA fragments. *Proc Natl Acad Sci U S A*, 86(4):1183–1187, February 1989.

[Sha40] C. E. Shannon. *An algebra for theoretical genetics*. PhD thesis, Massachusetts Institute of Technology, 1940.

[Sha48] C. E. Shannon. A mathematical theory of communication. *Bell Sys Tech J*, 27:379–423, 623–656, July, October 1948.

[SIoB07] Swiss Institute of Bioinformatics. SIB-EPD. http://www.epd.isb-sib.ch/, July 2007.

[SJ75] J. A. Steitz and K. Jakes. How ribosomes select initiator regions in mRNA: Base pair formation between the 3' terminus of 16S rRNA and the mRNA during initiation of protein synthesis in *Escherichia coli*. *Proc Natl Acad Sci U S A*, 72(12):4734–4738, December 1975.

[Slu05] M. Slutsky. *Protein-DNA Interactions, Random Walks and Polymer Statistics*. PhD thesis, Massachusetts Institute of Technology (MIT), May 2005.

[SM04] M. Slutsky and L. A. Mirny. Kinetics of protein-DNA interaction: Facilitated target location in sequence-dependent potential. *Biophys J*, 87(6):4021–4035, December 2004.

[SMC08] Santa Monica College. On The Human Genome Project. http://homepage.smc.edu/hgp/home.htm, May 2008.

[SPPB06] C. D. Schmid, R. Perier, V. Praz, and P. Bucher. EPD in its twentieth year: towards complete promoter coverage of selected model organisms. *Nucleic Acids Res*, 34(Database Issue):D82–D85, January 2006.

[Sti71] J. J. Stiffler. *Theory of Synchronous Communications*. Prentice-Hall, Englewood Cliffs, N.J., 1971.

[Sto00] G. D. Stormo. DNA binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, 2000.

[SY99] B. Shomer and G. Yagil. Long W tracts are over-represented in the *Escherichia coli* and *Haemophilus influenza* genomes. *Nucleic Acids Res*, 27(22):4491–4500, November 1999.

[TPM07] R. J. Taft, M. Pheasant, and J. S. Mattick. The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays*, 29(3):288–299, March 2007.

[TS61] R. Turyn and J. Storer. On binary sequences. *Proc. American Mathematical Society*, 12(3):394–399, June 1961.

[Tur68] R. Turyn. *Error correcting codes*, chapter Sequences with small correlation. Wiley, New York, 1968.

[UoR08] University of Rochester. Turner Group RNA Biophysical Chemistry, Turner Group Home Page. http://rna.chem.rochester.edu, January 2008.

[UoWM07] University of Wisconsin Madison. ASAP Home. https://asap.ahabs.wisc.edu/asap/home.php, August 2007.

[Vai04] P. P. Vaidyanathan. Genomics and proteomics: A signal processor's tour. *IEEE Circuits Syst Mag*, 4(4):6–29, Fourth quarter 2004.

[vHB89] P. H. von Hippel and O. G. Berg. Facilitated target location in biological systems. *J Biol Chem*, 264(2):675–678, January 1989.

[VM00] G. Varani and W. H. McClain. The G x U wobble base pair. a fundamental building block of RNA structure crucial to RNA function in diverse biological systems. *EMBO Rep*, 1(1):18–23, July 2000.

[Wat04] J. Watson. *DNA: The Secret of Life*. Arrow Books, London, 2004.

[WBvH81] R. B. Winter, O. G. Berg, and P. H. von Hippel. Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. The *Escherichia coli lac* repressor-operator interaction: kinetic measurements and conclusions. *Biochemistry*, 20(24):6961–6977, November 1981.

[WH07a] T. Warnecke and L. D. Hurst. Evidence for a trade-off between translational efficiency and splicing regulation in determining synonymous codon usage in *Drosophila melanogaster*. *Mol Biol Evol*, 24(12):2755–2762, December 2007.

[WH07b] J. Weindl and J. Hagenauer. Applying techniques from frame synchronization for biological sequence analysis. *Proc IEEE Internat Conf Commun (ICC)*, pages 833–838, June 2007.

[WHD+07] J. Weindl, P. Hanus, Z. Dawy, J. Zech, J. Hagenauer, and J. C. Mueller. Modeling DNA-binding by *Escherichia coli* σ^{70} exhibits a characteristic energy landscape around strong promoters. *Nucleic Acids Res*, 35(20):7003–7010, November 2007.

[XBA+06] C. Xing, D. L. Blitzer, W. E. Alexander, A.-M. Stomp, and M. A. Vouk. Free energy analysis on the coding region of the individual genes of *Saccharomyces cerevisae*. *Conf Proc IEEE Eng Med Biol Soc*, 1:4225–4228, August-September 2006.

[YL08] G. C. Yuan and J. S. Liu. Genomic sequence is highly predictive of local nucleosome depletion. *PLoS Comput Biol*, 4(1):e13, January 2008.

[Yoc92] H. P. Yockey. *Information theory and molecular biology*. Cambridge University Press, New York, 1992.

[ZPJ07] Y. M. Zalucki, P. M. Power, and M. P. Jennings. Selection for efficient translation initiation biases codon usage at second amino acid position in secretory proteins. *Nucleic Acids Res*, 35(17):5748–5754, September 2007.

Supervised Theses:

[Ber08] C. Bertram. Statistical analysis of DNA-sequences using Markov models. Studienarbeit, Technische Universität München, München, October 2008.

[Gon07] Q. Gong. Anwendung der Transinformation zur Genkartierung von Quantitative Trait Loci (QTL). Studienarbeit, Technische Universität München, München, March 2007.

[Kir08] B. Kirca. Coding theoretic modeling of translation in eukaryotes. Master's thesis, Technische Universität München, München, November 2008.

[Kis07] F. Kischkel. Analysis of translation initiation in higher organisms using information and coding theory. Diplomarbeit, Technische Universität München, München, November 2007.

[Reh07] T. Rehrl. Analysis of transcription initiation in higher organisms using information theoretic measures. Diplomarbeit, Technische Universität München, München, October 2007.

[Sul07] N. Sulieman. Analyzing transcriptional regulation in bacteria using frame synchronization and probabilistic modeling. Master's thesis, Technische Universität München, München, October 2007.

[Tax07] N. Tax. Communication theoretic considerations of translation in bacteria. Master's thesis, Technische Universität München, München, August 2007.