

ulm university universität
uulm

Theoretical Aspects of Overlapping Genes

DISSERTATION

zur Erlangung des akademischen Grades eines

DOKTOR-INGENIEURS
(Dr.-Ing.)

der Fakultät für Ingenieurwissenschaften
und Informatik der Universität Ulm

von

Katharina Schilling, geb. Mir

aus Ulm

Gutachter:	Prof. Dr.-Ing. Martin Bossert
	Prof. Dr. rer. nat. Hans A. Kestler
Amtierende Dekanin:	Prof. Dr. Tina Seufert

Ulm, 13. Januar 2015

To my parents

Preface

This dissertation presents recent results of my research activities at the Institute of Communications Engineering at Ulm University. My work has been supported by the German research council "Deutsche Forschungsgemeinschaft" (DFG) under grants BO 867/23-1, BO 867/23-2 and BO 867/23-3 in the priority program SPP 1395 (InKoMBio). The project entitled "Finding new overlapping genes and their theory (FOG-Theory)" was a joint proposal that combined the research areas of information and communication theory, data analysis and visualisation and microbial ecology. Parts of the results have been published in the following articles in conference proceedings and scientific journals:

[1] Katharina Mir, Klaus Neuhaus, Siegfried Scherer, Martin Bossert, and Steffen Schober. Predicting Statistical Properties of Open Reading Frames in Bacterial Genomes. *PLoS ONE*, 7(9):e45103, September 2012.

[2] Katharina Mir and Steffen Schober. Investigation of genetic code optimality for overlapping protein coding sequences. In *Proc. 8th International Symposium on Turbo Codes and Iterative Information Processing (ISTC)*, Bremen, Germany, August 2014.

[3] Katharina Mir and Steffen Schober. Selection Pressure in Alternative Reading Frames. *PLoS ONE*, 9(10):e108768, October 2014.

Acknowledgments

This dissertation contains part of my work as a research assistant at the Institute of Communications Engineering, Ulm University. Several people contributed to its development and I would like to express my gratitude to them.
First and foremost, I would like to thank my supervisor, Prof. Dr.-Ing. Martin Bossert, for giving me the opportunity to work in the great atmosphere of his institute and to participate in the interdisciplinary research project FOG-Theory. He provided a perfect working environment, combining support and entertainment, that made the last years memorable and enjoyable.
Secondly, I would like to express my gratitude to Prof. Dr. rer. nat. Hans A. Kestler from Friedrich Schiller University Jena for the external review of my thesis and for the interest in my work. Further, I would like to thank Prof. Ph.D. Carl Krill and Jun.-Prof. Dr. Jens Anders for being part of my commission.
Special thanks to the research group of Bioinformation Theory, consisting of Johannes Klotz, David Kracht and Steffen Schober. It was a pleasure to be part of this group. Thank you for all the scientific and private discussions that contributed to this work and made the last years special. Our group leader, Steffen Schober, was always a source of great support, ideas and motivation. Further, I would like to thank all project partners of FOG-Theory for fruitful and interesting discussions.
My sincere thanks goes to Max Riederle, who accompanied me over the last years, supporting and encouraging me in difficult decisions.
Moreover, I would like to thank all current and former members of the institute, especially Antonia Wachter-Zeh, Alexander Zeh, Frederic Knabe, Henning Zörlein and again the bio group for unforgettable discussions during lunch breaks and enjoyable excursions. Further, I thank Max Riederle and David Kracht for proofreading my manuscript.
Finally, special thanks to my parents, my husband and my whole family for all their love and support.

Katharina Schilling
Ulm, January 2015

Abstract

This thesis investigates theoretical aspects of overlapping genes in prokaryotes. Overlapping genes are protein-coding sequences that are encoded in different reading frames of the same DNA region, such that the sequences overlap non-trivially. In contrast to viruses, where overlapping genes are an accepted phenomenon and eukaryotes, where many examples have been found, they were thought to be exotic exceptions in prokaryotes. To shed light on this phenomenon, we study the theory behind overlapping protein-coding sequences in bacterial genomes.

First, an analytical model, based on the codon composition of genes and the length of the genome is developed. It predicts statistical properties of Open Reading Frames (ORFs) in all possible reading frames on the DNA sequence. The model is applied to overlapping ORFs, showing significant deviations in some cases indicating a potential functionality. Detailed comparison of the model predictions with *Escherichia coli* is presented. This bacterium is a pathogen that can lead to severe fooborne diseases when it is transmitted to humans. Additionally, a large scale comparison over 71 bacterial genomes is performed. Further, the model is applied to alternative genetic codes, to examine if the standard genetic code is optimized to encode long overlapping genes.

The second topic of this thesis is dedicated to theoretical aspects of overlapping gene evolution, studying the influence of selection pressure in different reading frames. Since evolution can be described as a communication process, concepts of information and communication theory are applied. Based on a codon evolution model in the protein-coding reading frame, the impact on alternative reading frames, that may contain gene candidates, is investigated. Information theoretic measures are applied to quantify the amount of information that is transmitted over time.

Contents

1. Motivation and Overview **1**

2. Fundamentals **5**
2.1. Basic Notations . . . 5
2.1.1. Probability Theory . . . 6
2.1.2. Hypothesis Testing . . . 8
2.2. Information Transmission . . . 10
2.2.1. Basic Definitions Information Theory . . . 11
2.2.2. Basic Definitions Coding Theory and Channel . . . 13
2.3. Markov Chains and Markov Processes . . . 15
2.3.1. Discrete-time Markov Chains . . . 15
2.3.2. Continous-time Markov Chains . . . 19
2.4. Biology . . . 22
2.4.1. Information Storage . . . 23
2.4.2. Information Processing . . . 25
2.4.3. Biological Notations . . . 29

3. Modeling Statistical Properties on Genome Level **33**
3.1. Analytical Model of Independent Nucleotides . . . 34
3.1.1. Description and Derivation of Model . . . 35
3.1.2. Prediction Limits of Nucleotide Model . . . 39
3.2. Discrete-time Markov Chain of Codons . . . 41
3.2.1. A First-Order Markov Chain Model . . . 41
3.2.2. Simplification to Finite Mixture Model . . . 45
3.3. Predicting Statistical Properties of Prokaryotic Genomes . . . 46
3.3.1. Application to a Particular Organism . . . 46
3.3.2. Application to Multiple Organisms . . . 62
3.4. Optimality Investigations of Alternative Genetic Codes . . . 70
3.5. Summary . . . 75

4. Evolution Channels of Overlapping Genes **77**
4.1. Information Transmission over Time . . . 78
4.1.1. Evolution Channel of Codon Substitutions . . . 79

4.1.2. Evolution Channel in Reading Frames . . . 82
4.1.3. Simulation Results . . . 84
4.2. Selection Pressure During Evolution . . . 87
4.2.1. Estimation of Substitution Rates . . . 88
4.2.2. Estimation of Substitution Rates in Reading Frames . . . 92
4.2.3. Simulation Results . . . 94
4.3. Quantifying Noise During Evolution . . . 97
4.3.1. Information Loss over Time . . . 98
4.3.2. Sequence Similarity over Time . . . 99
4.3.3. Simulation Results . . . 101
4.4. Influence of Genetic Code . . . 105
4.5. Summary . . . 109

5. Concluding Remarks **111**

A. Details regarding Markov Chains and Processes **113**
A.1. Proof of Theorem 2.1 . . . 113
A.2. Proof of Theorem 2.2 . . . 113
A.3. Solve Matrix Exponential . . . 114

B. Details regarding Statistical ORF Properties **117**
B.1. Statistical Parameters of Geometric Distribution . . . 117
B.2. Statistical Parameters of Weighted Geometric Distribution . . . 119
B.3. Transition matrices in Section 3.2.1 . . . 121
B.4. List of Species . . . 123
B.5. Application to *Mycoplasma mycoides* . . . 126
B.6. Application to single start codon *ATG* . . . 128

C. Details regarding Evolution **131**

Bibliography **133**

Nomenclature **151**

CHAPTER 1

Motivation and Overview

The foundations of Bioinformation Theory were laid in the 20th century, when the fields of communication theory and molecular genetics experienced profound changes. Back in 1948, Shannon published his fundamental work [Sha48] that established the field of information theory. He introduced an abstract measure of information, that depends only on the statistical properties of an underlying source, which makes it applicable to various problems. A few years later, in 1953, Watson and Crick [WC53] discovered the double helix structure of the Deoxyribonucleic acid (DNA), which was the start of modern genetics. Their work was so significant, that they were awarded the Nobel Prize in 1962 [Med14].

An early interdisciplinary cooperation of biologists with computer scientists, statisticians and mathematicians dates back to 2001, when under the Human Genome Project the complete human genome sequence was published almost simultaneously by [LLB$^+$01, VAM$^+$01]. Ever since high-throughput technologies reduced the sequencing cost and time, leading to an exponential growth of publicly available genome sequences, the field of *Bioinformatics* is well established. Challenges like sequence assembly, annotation of genes, data storage and handling as well as analysis are covered by that field in order to manage the enormous amount of available data [REHA11]. Combination of biology with mathematics, physics and engineering deals with interactions between biological systems in the research area called *Systems biology*, investigating, e.g., gene regulatory networks to understand protein interactions [CHI10].

All organisms store their genetic information within the DNA, that can be considered as digital signal over an alphabet of four letters. Biological concepts like protein synthesis, inheritance and evolution are molecular communication processes that can be treated as information transmission systems in an abstract way by applying concepts of information theory. In order to understand the underlying principles, the cooperation of communication engineers and information theorists with biologists became necessary and the *Bioinformation Theory* emerged. To name a few contributions to this fast growing field: Battail investigated error correcting codes in order to explain biological evolution [Bat97]. The group of Hagenauer studied the risk of complex diseases based on mutations of the DNA sequence

applying information theory [DGH$^+$06, HDG$^+$04, SDHM05], further compression algorithms for gene classification [DGHM05, HDG$^+$04, HGD$^+$07] and synchronization of the transcription process [WHD$^+$07]. Boolean Networks introduced by Kauffman [Kau69, Kau74] were studied with respect to their stability, e.g., [Sch06, KKBS14]. Milenkovic *et al.* studied the relationship between channel codes and Boolean networks with the goal to understand the DNA proofreading mechanism [MV04]. The minimum description length (MDL) principle was applied to DNA segmentation [SRS03] and inference of gene regulatory networks [DTA08]. Schneider invented sequence logos, a graphical representation of conserved DNA sequences [SS90], a simple but well accepted application of entropy. Even Shannon was interested in biology: He introduced an algebra for theoretical genetics in his PhD thesis in 1940 [Sha40], a work he did not pursue.

Overview The focus of this thesis lies on statistical properties of *overlapping genes* in order to gain knowledge about their theory. Overlapping genes are two protein-coding sequences that share a significant part of the same DNA region, resulting in two different proteins with different functions. Since the DNA sequence consists of two complementary strands and is translated by groups of three nucleotides into amino acids, forming proteins, up to six different reading frames exist, three on the forward and three on the reverse strand. An overlapping gene pair can be either located in different reading frames of the same strand, or the second gene is located on the complementary strand.

But why do overlapping genes exist and why is their existence of interest? The first overlapping genes were discovered in viruses, e.g., $\Phi X174$ [Bar76, SAB$^+$77]. Its genome contains only 5386 nucleotides, which is too short to encode its 11 proteins, resulting in several overlapping genes. However, apart of their functionality, the size limitations of viruses seem to explain the need for the existence of overlapping genes, in contrast to eukaryotes, where also overlapping genes exist, e.g., [KCGC02, VMG$^+$04, KCH$^+$09]. But in viruses and higher organisms they are also known to carry out important tasks. For example, the overlapping gene pair *INK4a/ARF*, that is located in the human genome, plays an important role in two independent tumour suppressor pathways [She01, She06] and according to [KTR$^+$04] the process of ageing. Mutations in those genes, disturbing their functionality are associated with cancer [LS03]. Another interesting pair is known from yeast, where the mitochondrial protein *Tar1p* codes on the opposite strand for ribosomal RNA, which is necessary for regulation of protein synthesis [CCA$^+$02, BCLWS08].

In contrast to eukaryotes and viruses, where many examples of overlapping genes exist, in prokaryotes only a handful of candidates for overlapping gene pairs have been described [BSN02, KSP$^+$09, SRL04, JPF$^+$06]. Especially their functionality is rarely verified [TBCM09, FBW$^+$14]. In order to detect more overlapping genes in prokaryotes, such that biologists gain a better understanding of their role in gene regulation and evolution, it is essential to study their underlying theory. In this thesis, we concentrate on the investigation of prokaryotic DNA sequences. Our working hypothesis is, that in prokaryotes, similar to viruses and eukaryotes, many functional overlapping protein-coding sequences exist, that undertake important tasks.

The thesis is structured as follows:

In **Chapter 2** we introduce necessary fundamentals starting with basic notations and mathematical concepts, followed by information and communication theory. The models presented in this thesis are based on stochastic processes, which are also introduced in this chapter. Further, biological terms and concepts are explained in the depth required to understand the problems that arise from the field of molecular biology.

Chapter 3 is dedicated to the statistical properties of bacterial genomes. We derive an analytical model that is based on a discrete-time Markov chain, which can be used to predict statistical parameters like the length distribution of Open Reading Frames (ORFs). A detailed comparison of the model predictions with the natural genome called *EHEC* is presented as well as a large scale analysis over 71 bacterial genomes. Further we study how the genetic code influences the model predictions. Our investigations give interesting insights in the theory of long ORFs and overlapping genes.

Chapter 4 covers the topic of molecular evolution measuring the intensity of selection acting on an overlapping gene pair. Understanding the evolution of overlapping genes is challenging, since any mutation affects both proteins simultaneously. We derive and analyze evolution channels in different reading frames, based on a continuous-time Markov process. Information theoretic measures are applied to quantify the amount of information that can be transmitted over evolutionary time. Our findings suggest that a simultaneous protection of an overlapping gene pair is theoretically possible.

Finally, we summarize and conclude this thesis in **Chapter 5**.

CHAPTER 2

Fundamentals

NECESSARY PRELIMINARIES needed to investigate aspects of overlapping genes are introduced in this chapter. Within this thesis, we deal with statistical properties of overlapping genes and selection pressure based on evolution channels, therefore we can not avoid to explore biology to a certain extent.
Section 2.1 provides notations and mathematical concepts needed during this thesis. In Section 2.2 we describe the basics of digital communication- and information theory to the extend necessary for the understanding of subsequent chapters. Stochastic processes, that build the basis for our models, are introduced in Section 2.3. Finally, since the whole thesis deals with problems that arise in the field of molecular biology, in Section 2.4 a short introduction to the biological concepts is included.

2.1. Basic Notations

We start with a brief introduction into the mathematical notations used, Section 2.1.1 contains an introduction to probability theory and Section 2.1.2 the basics of hypothesis testing.
We write $\mathbb{R}$ and $\mathbb{R}_0^+$ to represent a set of real numbers and a set of non-negative real number. Whereas $\mathbb{Z}$ and $\mathbb{Z}^+$ are a set of integers and a set of positive integers, respectively. For a vector $\boldsymbol{x}$ bold small characters, and for a matrix $\boldsymbol{X}$ bold large characters are used. We write $\boldsymbol{x} = (x_0 x_1 \dots x_{n-1})$ to denote a vector of length n.

Definition 2.1 (Euclidean Distance). In the n-dimensional space $\mathbb{R}^n$ the Euclidean distance or ℓ^2 norm of $\boldsymbol{x} = (x_0 x_1 \dots x_{n-1})$ and $\boldsymbol{y} = (y_0 y_1 \dots y_{n-1})$ is given by

$$||\boldsymbol{x} - \boldsymbol{y}|| = \sqrt{\sum_{i=0}^{n-1} (x_i - y_i)^2}.$$

□

Definition 2.2 (Hamming Distance). Given two vectors $\boldsymbol{x}, \boldsymbol{y}$ of equal length n. The Hamming distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ is the number of positions, in which the two vectors differ

$$\mathrm{d_H}\left[\boldsymbol{x}, \boldsymbol{y}\right] = \sum_{i=0}^{n-1} \delta(x_i, y_i) \text{ with } \delta(x_i, y_i) = \begin{cases} 0, & x_i = y_i \\ 1, & x_i \neq y_i \end{cases}.$$

□

Definition 2.3 (Empirical distribution function). Given a data set $\boldsymbol{x} = (x_0 x_1 \dots x_{n-1})$ of observed values, the empirical distribution function (EDF) is defined as

$$F_n(x) = \frac{1}{n} \sum_{i=0}^{n-1} \mathbf{1}\{x_i \leq x\},$$

with indicator function

$$\mathbf{1}\{x_i \leq x\} = \begin{cases} 1 & \text{if } x_i \leq x \\ 0 & \text{else.} \end{cases}$$

□

2.1.1. Probability Theory

The concepts and definitions presented here can be found in standard textbooks on probability theory, e.g., [Fel68, BSMM01]. We use the following notations:

- Capital characters X, Y denote random variables.
- Calligraphic characters $\mathcal{X}, \mathcal{Y}$ denote finite sets and alphabets of random variables.
- We denote the cardinality or size of a set with $|\mathcal{X}|$.
- Let small characters $x \in \mathcal{X}$ be a realization of X.
- The probability for the event that the realization of X is x is denoted as $P\left(X = x\right)$.
- If X is a discrete random variable, we refer to $P\left(X = x\right) = p_X(x)$ as probability mass function (PMF). Note, the PMF satisfies $\sum_{x \in \mathcal{X}} p_X(x) = 1$, with $p_X(x) \geq 0, \forall x \in \mathcal{X}$.

Definition 2.4 (Cumulative Distribution Function). The cumulative distribution function (CDF) of X is defined as $F_X(x) = P\left(X \leq x\right)$. For a discrete random variable this is

$$F_X(x) = \sum_{x_i \in \mathcal{X}, x_i \leq x} p_X(x_i).$$

□

Definition 2.5 (Expected Value and Variance). The expected value or mean of a discrete random variable X is defined as

$$\mathbf{E}[X] = \sum_{x \in \mathcal{X}} x \cdot p_X(x).$$

The corresponding variance is defined as

$$\sigma^2 = \mathbf{E}\left[(X - \mathbf{E}[X])^2\right],$$

where σ is the standard deviation. □

The joint probability mass function of two random variables (X, Y) over alphabets $\mathcal{X}, \mathcal{Y}$ is denoted as

$$p_{X,Y}(x, y) = P(X = x, Y = y).$$

In this context, we call $p_X(x) = \sum_{y \in \mathcal{Y}} p_{X,Y}(x, y)$ and $p_Y(y) = \sum_{x \in \mathcal{X}} p_{X,Y}(x, y)$ the marginal probability mass functions. Random variables are independent if $p_{X,Y}(x, y) = p_X(x) \cdot p_Y(y)$ holds, for all $x, y \in \mathcal{X}, \mathcal{Y}$.
If (X, Y) are jointly distributed random variables, the conditional probability mass function of Y given X is defined as

$$p_{Y|X}(y|x) = P(Y = y|X = x) = \frac{p_{X,Y}(x, y)}{p_X(x)} = \frac{p_{X|Y}(x|y)p_Y(y)}{p_X(x)},$$

where the latter follows from Bayes' rule.
For jointly distributed random variables, the parameters of interest are, the expected values $\mathbf{E}[X], \mathbf{E}[Y]$ and the variances σ_X^2, σ_Y^2 calculated according to Definition 2.5. Further, the covariance

$$\sigma_{XY} = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])] \tag{2.1}$$

and correlation coefficient

$$\varrho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}. \tag{2.2}$$

If X and Y are independent random variables, the correlation coefficient is $\varrho_{XY} = 0$.
In the following, we introduce some probability distributions, concepts and terms that are of interest during the thesis. We characterize the probability distributions by the corresponding PMFs.

Definition 2.6 (Discrete Uniform Distribution). A random variable X has a discrete uniform distribution, if the probability for each element in the corresponding alphabet is equal. We write $X \sim \text{DiscreteUniform}(\mathcal{X})$ to denote that random variable X has PMF

$$P(X = x) = p_X(x) = \frac{1}{|\mathcal{X}|}, \qquad \forall\, x \in \mathcal{X}.$$

□

Definition 2.7 (Independent and Identically Distributed). Given a vector of n independent discrete random variables $(X_1 X_2 \dots X_n)$ with $X_i \in \mathcal{X}$. They are independent and identically distributed (IID) if

$$p_{X_i}(x) = p_{X_j}(x), \qquad \forall\, x \in \mathcal{X}, i = 1, \dots, n,\ j = 1, \dots, n$$

If $p_{X_i}(x) = \frac{1}{|\mathcal{X}|}, \forall\, x \in \mathcal{X}, i = 1, \dots, n$, the random variables are independent and uniformly distributed (IUD). □

Definition 2.8 (Bernoulli Distribution/Trials). A binary random variable X over $\mathcal{X} = \{0, 1\}$ has a Bernoulli distribution, if $P(X = 1) = p$ and $P(X = 0) = 1 - p$, where p is the probability of "success" and $1 - p$ the probability of "failure". We write $X \sim \text{Bern}(p)$ to denote that the number of "successes" X has PMF

$$P(X = x) = p_X(x) = p^x (1 - p)^{1-x}, \qquad x = 0, 1.$$

The mean is $\mathbf{E}[X] = p$ and the variance $\sigma^2 = p \cdot (1 - p)$. □

Definition 2.9 (Binomial Distribution). For n independent Bernoulli distributed random variables, random variable $K = \sum_{i=1}^{n} X_i$ follows a binomial distribution $K \sim \text{Bin}(n, p)$ with PMF

$$P(K = k) = p_K(k) = \binom{n}{k} p^k (1 - p)^{n-k}, \qquad 0 \leq k \leq n.$$

measuring the number of independent success trials. The mean is $\mathbf{E}[X] = n \cdot p$ and the variance $\sigma^2 = n \cdot p \cdot (1 - p)$. □

Definition 2.10 (Geometric Distribution). The random variable X has a geometric distribution, if it counts the number of Bernoulli trials until the first occurrence of "success". We denote $X \sim \text{Geo}(p)$ with PMF

$$P(X = x) = p_X(x) = (1 - p)^{x-1} p, \qquad x = 1, 2, \dots,$$

where p is the probability of success. The distribution has mean $\mathbf{E}[X] = \frac{1}{p}$ and variance $\sigma^2 = \frac{1-p}{p^2}$. □

Definition 2.11 (Quantile). Let X be a random variable with CDF $F_X(x) = P(X \leq x)$. For $r \in (0, 1)$, the quantile function of r is defined as [Par04]

$$Q(r) = \inf\{x : F_X(x) \geq r\},$$

which corresponds to the minimum value among all x, where the CDF exceeds r. □

2.1.2. Hypothesis Testing

An introduction to hypothesis testing and correlation can be found in [Pri14, Chapter 8] and [Wei14, Tay90], which are the basic for the following informations. Based on the outcome

of measurements, the task of "binary" hypothesis testing is to decide which of two given hypotheses is true. The two hypotheses are usually called, the null hypothesis H_0 and the alternative hypothesis H_1, whereby we know, that one of them is true.
The process of hypothesis testing can be described in four steps

1. Formulate the null hypothesis H_0 and the alternative hypothesis H_1 that contradicts the null hypothesis. Commonly H_1 models the theory one would like to prove.

2. Select a criterion to decide if the null hypothesis is true or not. Therefore identify a random variable T, a so called test statistic, that is able to differentiate random samples from the null hypothesis and determine its realization t from the given data.

3. Convert the test statistic to a conditional probability, that is called the p-value.

$$p\text{-value} := P(T \geq t|H_0)$$

   This parameter determines the probability of observing a value from the test statistic at least as extreme as the observed, under the assumption that the null hypothesis is true.

4. Make a decision by comparison of the p-value to an acceptable significance level α. The observed effect is statistically significant and the null hypothesis is rejected if p-value $\leq \alpha$. In this case, the alternative hypothesis is valid, but the probability that this decision is wrong is α. More practically, we fix a threshold the so called critical value t_C, such that $P(T \geq t_C|H_0) = \alpha$ and reject the null hypothesis, if the test statistic of the sample is $t \geq t_C$. Since we observe only samples and not the complete population, four decision outputs are possible, that are summarized in Table 2.1.

		Decision	
		Reject H_0	Do not reject H_0
Truth	H_0	Type I error (α)	Right decision
	H_1	Right decision	Type II error (β)

Table 2.1.: Decision outcome with corresponding errors (modified version of [Pri14]).

Hypothesis testing is often applied to investigate if there is a linear relationship between two variables. This relationship is measured by correlation. Given two jointly distributed random variables (X, Y), the correlation coefficient ϱ_{XY} is determined from (2.2). To test, if two given features are independent, we follow [BSMM01].
Given a sample of n data points (x_i, y_i) with $i = 0, 1, \ldots, n-1$ based on random variables X, Y. Investigate, if the data comes from a bivariate normal distributed population with $\varrho_{XY} = 0$, meaning that X, Y are independent. The null hypothesis and alternative hypothesis are formulated as: H_0 there is no correlation ($\varrho_{XY} = 0$) and H_1 there is a correlation ($\varrho_{XY} \neq 0$). The corresponding test statistic follows a Student's t-distribution with $m = n - 2$

degrees of freedom. Tables exist, where for fixed values α and m the corresponding critical value $t_{\alpha,m}$ is given. The test value is calculated by

$$t = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$$

with the empirical correlation coefficient

$$r_{xy} = \frac{\sum_{i=0}^{n-1}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=0}^{n-1}(x_i - \bar{x})^2 \sum_{i=0}^{n-1}(y_i - \bar{y})^2}} \tag{2.3}$$

and empirical means $\bar{x} = \frac{1}{n}\sum_{i=0}^{n-1} x_i$ as well as $\bar{y} = \frac{1}{n}\sum_{i=0}^{n-1} y_i$. The null hypothesis is rejected if $|t| \geq t_{\alpha,m}$. The empirical correlation coefficient takes values $-1 \leq r_{xy} \leq +1$, where ± 1 reveals a perfect linear correlation and $r_{xy} = 0$ no correlation at all, therefore no linear association between measured variables exists, e.g., [Tay90]. Note, that a calculation of r_{xy} alone is not useful, since this value can be interpreted only in combination with the corresponding p-value that tells if the relationship between variables is statistically significant. A basic review regarding the correlation coefficient is given in [Tay90], where the following details can be found. In the context of correlation, the so called coefficient of determination r_{xy}^2 is of importance. It defines the percent of variation in the values of the y that can be explained by variation of x, i.e., a statistically significant correlation coefficient $r_{xy} = 0.20$ at $\alpha = 0.05$ explains only 4% of the total variation observed in y by variation in x. This value is more conservative for interpretation than the correlation coefficient, however it is seldom used in practice and many research articles investigating a linear relationship publish only ($r = r_{xy}, p = p$-value).

2.2. Information Transmission

The development of information theory dates back to the seminal paper of Shannon published in 1948 [Sha48]. It provides fundamental limits of information processing in the context of message transmission based on a mathematical framework. The main advantage is the universal representation of information as a stochastic signal which is independent of the underlying medium. Therefore the concepts can be applied to a wide range of different scenarios including biology, e.g., [Yoc05].
The basic blocks of a communication system according to Shannon are presented in Figure 2.1, where a *source* produces messages that are processed for transmission over the *channel* within the *transmitter*. The channel is the medium used to transmit the signal to the *receiver*, where the message is reconstructed, such that it can be delivered to the *sink*.
Shannon proved, that error free data transmission over a noisy channel is possible, if the code rate is less than the channel capacity. In Section 2.2.1 we give some information theoretic definitions needed in this context. Since Shannon did not state how good error-correcting codes should be constructed, the field of coding theory emerged. A brief introduction to coding theory is given in Section 2.2.2.

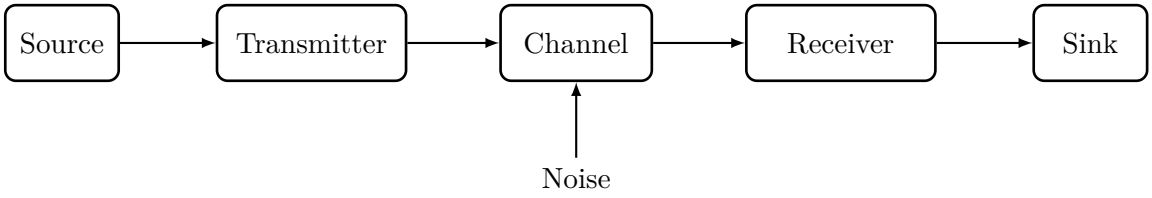


Figure 2.1.: Model of a digital communication system (modified version of [Sha48]).

2.2.1. Basic Definitions Information Theory

Definitions and concepts presented here, can be found in standard text books on information theory, e.g., [CT06, Bos12].
Given two discrete random variables X and Y and their realizations $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. The random variables X, Y have joint PMF $p_{X,Y}(x, y)$ with marginal PMFs $p_X(x)$ and $p_Y(y)$.
A measure of the uncertainty about the realization of random variable X is the *entropy*.

Definition 2.12 (Entropy). The entropy $H(X)$ of a discrete random variable X is defined by

$$H(X) = -\sum_{x \in \mathcal{X}} p_X(x) \log p_X(x).$$

□

It can be shown that $0 \leq H(X) \leq \log |\mathcal{X}|$, where the maximum entropy is reached for a uniformly distributed random variable and the lower bound is reached for a deterministic process that carries no uncertainty about the outcome. The unit of the entropy depends on the chosen logarithm, where for $\log = \log_2$ the unit is bit.
For the *joint entropy* of $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ follows straight forward

$$H(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{X,Y}(x, y) \log p_{X,Y}(x, y).$$

The joint entropy is symmetric $H(X, Y) = H(Y, X)$ and bounded by

$$\max\left(H(X), H(Y)\right) \leq H(X, Y) \leq H(X) + H(Y),$$

where the upper bound is reached for independent random variables.
The *conditional entropy* describes the uncertainty that remains in Y given the outcome of X.

Definition 2.13 (Conditional Entropy). The conditional entropy of a discrete random variable Y given X is defined by

$$\begin{aligned} H(Y|X) &= \sum_{x \in \mathcal{X}} p_X(x) H(Y|X = x) && (2.4) \\ &= -\sum_{x \in \mathcal{X}} p_X(x) \sum_{y \in \mathcal{Y}} p_{Y|X}(y|x) \log p_{Y|X}(y|x) && (2.5) \\ &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{X,Y}(x, y) \log p_{Y|X}(y|x). && (2.6) \end{aligned}$$

□

The conditional entopy is in the range $0 \leq H(Y|X) \leq H(Y)$, reaching zero, if X determines Y completely and is maximized, if the random variables are stochastically independent.
The so called *chain rule* combines the definitions of entropy

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) = H(Y,X) \tag{2.7}$$

The *relative entropy* or *Kullback-Leibler divergence* measures the distance between two probability mass functions

Definition 2.14 (Kullback-Leibler divergence). Let X and Y have probability mass functions $p_X(x)$ and $p_Y(x)$ for $x \in \mathcal{X}$, then the Kullback-Leibler divergence is defined as

$$D(p_X||p_Y) = \sum_{x\in\mathcal{X}} p_X(x) \log \frac{p_X(x)}{p_Y(x)},$$

with convention $0 \log \frac{0}{p_Y(x)} = 0$ and $p_X(x) \log \frac{p_X(x)}{0} = \infty$. □

It holds, that $D(p||q)$ is always non-negative and zero if and only if $p(x) = q(x)$. Note, that this metric is not a real distance, since it is neither symmetric nor does it satisfy the triangle equation.
Next, we introduce the *mutual information*, which is a measure on the amount of information one random variable contains about another random variable.

Definition 2.15 (Mutual Information). The mutual information between two random variables X and Y is defined as

$$I(X;Y) = H(X) - H(X|Y).$$

□

The mutual information is symmetric $I(X;Y) = I(Y;X)$ and takes values in the range $0 \leq I(X;Y) \leq \min(H(X), H(Y))$. The lower bound is reached for independent variables. In terms of probabilities, we get

$$I(X;Y) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p_{X,Y}(x,y) \log \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)} \tag{2.8}$$

$$= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p_X(x)p_{Y|X}(y|x) \log \frac{p_{Y|X}(y|x)}{\sum_{k\in\mathcal{X}} p_X(k)p_{Y|X}(y|k)}. \tag{2.9}$$

The Kullback-Leibler divergence is connected to the mutual information via

$$D(p_{X,Y}(x,y)||p_X(x)p_Y(y)) = I(X;Y).$$

After introducing the necessary definitions, we have a closer look at the communication system in Figure 2.1 on page 11. The source is modelled by random variable $X \sim p_X(x)$, emitting a sequence $\boldsymbol{x} = (x_0 x_1 \dots x_{n-1})$ of elements $x_i \in \mathcal{X}$. The message $\boldsymbol{x}$ is transmitted over a noisy channel, changing into a received vector $\boldsymbol{y} = (y_0 y_1 \dots y_{n-1})$ with elements $y_i \in \mathcal{Y}$. Formally, the channel is defined as follows.

Definition 2.16 (Discrete Memoryless Channel). A *discrete channel* is denoted by $(\mathcal{X}, p(y|x), \mathcal{Y})$ consisting of an input alphabet $\mathcal{X}$ and an output alphabet $\mathcal{Y}$ and a probability mass function $p(y|x)$ specifying the probability to observe $y \in \mathcal{Y}$ given that $x \in \mathcal{X}$ was sent. Transmitting n symbols, the channel is *memoryless* if

$$p(y_i|x_i x_{i-1} \ldots) = p(y_i|x_i), \qquad i = 0, \ldots n-1$$

which results in

$$p(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=0}^{n-1} p(y_i|x_i),$$

a discrete memoryless channel (DMC). □

The conditional probabilities of the channel are often described by the corresponding transition matrix $\boldsymbol{P}$. If the rows as well as the columns of the channel matrix are permutations of each other, the channel is called *symmetric*. The following example for such a channel, is often used in a biological setup.

Example 2.1 (q-ary symmetric channel). Consider the channel $(\mathcal{X}, \boldsymbol{P}_{qsc}, \mathcal{Y})$ with alphabet $\mathcal{X} = \mathcal{Y} = \{1, 2, 3, 4, \ldots, q\}$ of cardinality q given by the transition matrix $\boldsymbol{P}_{qsc}$ with elements $p(y|x)$ is

$$p(y|x) = \begin{cases} 1-p & \text{if } x = y \\ \frac{p}{q-1} & \text{if } x \neq y \end{cases},$$

with $0 \leq p \leq 1 - \frac{1}{q}$. □

The largest possible mutual information determines the *channel capacity*, to measure how much information can be transmitted over a channel.

Definition 2.17 (Channel Capacity). For a discrete memoryless channel the capacity is

$$\mathcal{C} = \sup_{p_X} I(X;Y),$$

where the supremum is taken over all possible choices of p_X. □

2.2.2. Basic Definitions Coding Theory and Channel

The *channel coding theorem* of [Sha48] states, that codes exist, that are able to reach the channel capacity mentioned in Definition 2.17. However, the prove of the theorem is non-constructive, such that it describes neither code constructions nor decoding procedures for practical realization of error free data transmission. Since the publication of Shannon's work, the main goal of coding theorists was the construction of practically useful codes. Nowadays, this goal is reached. For a introduction to this field, standard text books on coding theory are recommended, e.g., [Bos99, Rot06, MS77]. We start with some necessary definitions, without going into detail.

Definition 2.18 (Prime field). Let q be a prime number. A prime field of order q is a finite set $\mathbb{F}_q$ of elements $\{0, 1, \dots q-1\}$ on which the following properties and operations on additions and multiplications are defined:

- Closure: $\forall\, i, j \in \mathbb{F}_q :\ (i+j) \in \mathbb{F}_q,\ (i \cdot j) \in \mathbb{F}_q$
- Identity: there exist neutral elements for addition and multiplication
- Inverse: there exists an additive inverse for all numbers as well as a multiplicative inverse for all numbers except zero
- Commutative law: $\forall\, i, j \in \mathbb{F}_q :\ i + j = j + i,\ i \cdot j = j \cdot i$
- Associative law: $\forall\, i, j, k \in \mathbb{F}_q :\ (i+j)+k = i+(j+k),\ (i \cdot j) \cdot k = i \cdot (j \cdot k)$
- Distributive law: $\forall\, i, j, k \in \mathbb{F}_q :\ (i+j) \cdot k = i \cdot k + j \cdot k$

Additions ad multiplications are modulo q. □

Multiplication of a vector $\boldsymbol{a} \in \mathbb{F}_q^n$ with a skalar $u \in \mathbb{F}_q$ is done element wise for each component of the vector, i.e., $u \cdot a_i$ with $i = 0, \dots, n-1$.

Definition 2.19 (Linearity of a Code). A code is called linear, if the linear combination of two arbitrary codewords $\boldsymbol{a}, \boldsymbol{b} \in C \subseteq \mathbb{F}_q^n$ is again a codeword

$$\forall\, \boldsymbol{a}, \boldsymbol{b} \in C : u\boldsymbol{a} + v\boldsymbol{b} \in C, \quad u, v \in \mathbb{F}_q.$$

□

Definition 2.20 (Linear Block Code). A linear block code $C(n,k) \subseteq \mathbb{F}_q^n$ over alphabet $\mathbb{F}_q$ has *code length* n and *dimension* k, i.e., $|C| = q^k$. The *rate* of C is defined by $R = \frac{k}{n}$. □

As long as the code rate is smaller than the channel capacity of Definition 2.17, i.e., $R < \mathcal{C}$, an error free transmission is theoretically possible for $n \to \infty$. For finite code lengths Gallager [Gal68] investigated how fast the error probability drops with n.

We consider the model of a digital communication system presented in Figure 2.2, where a *source* outputs an *information* vector $\boldsymbol{x} = (x_0 x_1 \dots x_{k-1})$ over $\mathbb{F}_q^k$. The information symbols are encoded into a *codeword* $\boldsymbol{c} = (c_0 c_1 \dots c_{n-1})$ over $\mathbb{F}_q^n$ of length $n \geq k$ adding specific redundancy. Codeword $\boldsymbol{c}$ is transmitted over the *channel*, resulting in a noisy *received* vector $\boldsymbol{r} = (r_0 r_1 \dots r_{n-1})$ of length n. The *decoder* decides from $\boldsymbol{r}$ which message $\boldsymbol{x}$ was originally transmitted, based on the estimated codeword $\hat{\boldsymbol{c}}$ or information vector $\hat{\boldsymbol{x}}$, respectively.



Figure 2.2.: Model of a digital communication system (modified version of [Bos12]).

After this overview, we have a more detailed look at the different blocks. The encoder maps each of the q^k possible messages $\boldsymbol{x}$ uniquely to a codeword $\boldsymbol{c}$ of a code C, e.g., a block code.

$$\mathrm{enc}(\boldsymbol{x}) : \mathbb{F}_q^k \to C, \quad \boldsymbol{x} \mapsto \boldsymbol{c}$$

During transmission over the discrete channel, the codeword $\boldsymbol{c}$ may be distorted by noise, leading to a received vector $\boldsymbol{r}$ that can be different from $\boldsymbol{c}$. The goal is to reconstruct the originally transmitted message $\boldsymbol{x}$ from $\boldsymbol{r}$ without further knowledge of the channel. In contrast to the encoder, the decoder is an estimator that determines a codeword $\hat{\boldsymbol{c}}$ by searching for the codeword $\boldsymbol{c}$, that has been transmitted most likely knowing $\boldsymbol{r}$. Different decoding principles are possible for a certain channel model. It may happen that the decision of the decoder is not unambiguous, in this case a list of possible codewords $\mathcal{L}(C)$ is returned by the decoder.

$$\mathrm{dec}(\boldsymbol{r}) : \mathbb{F}_q^n \to C \cup \mathcal{L}(C), \quad \boldsymbol{r} \mapsto \hat{\boldsymbol{c}}.$$

Once the estimate $\hat{\boldsymbol{c}}$ is known, the codeword is associated to the unique information vector $\hat{\boldsymbol{x}}$.

2.3. Markov Chains and Markov Processes

Throughout this thesis, we consider sequences of random variables to investigate properties of biological sequences. Such a collection of random variables is called a stochastic process, where the actual outcome depends on some probability distribution. Markov chains are a fundamental tool which introduces some kind of dependency in order to model stochastic processes. The advantage of Markov chains lies in the fact that future development of a random variable depends only on its current value and not on the intermediate steps that led to this value. An introduction to Markov chains can be found in standard literature, e.g., [Nor98, Fel68, GS01, GS97, Cin75, Ros10], where the content of this chapter can be found. In Section 2.3.1 we introduce the basic theory of discrete-time Markov chains which are applied in Chapter 3 for modelling random DNA sequences. An extension to the continuous-time case is given in Section 2.3.2 applied in Chapter 4 to model sequence evolution of biological sequences.

2.3.1. Discrete-time Markov Chains

In this section we consider only discrete-time Markov chains (DTMC) using time indices $n, k \in \mathbb{Z}_0^+ = \{0, 1, 2, \ldots\}$. We denote random variables with capital letters X_n and their realizations with the corresponding lower-case letters x_n. For a discrete-time stochastic process, which is a sequence of random variables, we write $\{X_n\}_{n\geq 0}$. The values, the stochastic process can attain, belong to a countable set $\mathcal{S}$, called the state space. Each element $i \in \mathcal{S}$ is called a state. Further we need the following definitions:

Definition 2.21 (Transition matrix). A transition matrix $\boldsymbol{P} = \{P_{ij}\}_{i,j\in\mathcal{S}}$ is a stochastic matrix, where each row satisfies the following conditions:

1. $0 \leq P_{ij} \leq 1 \qquad \forall\, i, j \in \mathcal{S}$,

2. $\sum_j P_{ij} = 1 \qquad \forall\, i \in \mathcal{S}.$

The elements of P_{ij} are the transition probabilities moving from state i to state j in one time step

$$P_{ij} = \mathrm{P}\left(X_1 = j | X_0 = i\right) \qquad \forall\, i, j \in \mathcal{S}.$$

□

Let $\boldsymbol{p}^{(0)}$ be a distribution over $\mathcal{S}$, where

$$p_i^{(0)} = \mathrm{P}\left(X = i\right) \qquad i \in \mathcal{S},$$

with $0 \leq p_i^{(0)} \leq 1$ for $i \in \mathcal{S}$ and $\sum_{i \in \mathcal{S}} p_i^{(0)} = 1$.

Definition 2.22 (Discrete-time Markov chain). A stochastic process $\{X_n\}_{n \geq 0}$ is called a Markov chain with initial distribution $\boldsymbol{p}^{(0)}$ and transition matrix $\boldsymbol{P}$ if X_0 has distribution $\boldsymbol{p}^{(0)}$, i.e.,

$$p_i^{(0)} = \mathrm{P}\left(X_0 = i\right) \qquad i \in \mathcal{S}$$

and fulfills the Markov property, i.e., for $n \geq 0$ and $x_0, \ldots, x_{n+1} \in \mathcal{S}$

$$\mathrm{P}\left(X_{n+1} = x_{n+1} | X_n = x_n, \ldots, X_0 = x_0\right) = \mathrm{P}\left(X_{n+1} = x_{n+1} | X_n = x_n\right).$$

As the Markov chain is completely determined by $\boldsymbol{p}^{(0)}$ and $\boldsymbol{P}$, we denote it in short as $\mathcal{MC}\left(\boldsymbol{p}^{(0)}, \boldsymbol{P}\right)$. □

The probability that a process $\{X_n\}_{0 \leq n \leq N}$ results in a sequence $x_0 x_1 \ldots x_N$ with $x_i \in \mathcal{S}$ is given by

$$\begin{aligned}\mathrm{P}\left(X_0 = x_0, X_1 = x_1, \ldots X_N = x_N\right) &= \mathrm{P}\left(X_0 = x_0\right) \prod_{n=1}^{N} \mathrm{P}\left(X_n = x_n | X_{n-1} = x_{n-1}\right) \\ &= p_{x_0} P_{x_0 x_1} P_{x_1 x_2} \cdots P_{x_{N-1} x_N}.\end{aligned}$$

Definition 2.23 (Homogeneous Markov chain). A Markov chain is time-homogeneous if the transition probabilities are independent of the time index n, i.e.,

$$P_{ij} = \mathrm{P}\left(X_{n+1} = j | X_n = i\right) = \mathrm{P}\left(X_1 = j | X_0 = i\right) \quad \forall\, i, j \in \mathcal{S}, \forall\, n \geq 0.$$

□

As already mentioned, P_{ij} is the probability for a transition from state i to state j in one time step. The probability to reach state j starting in state i in two time steps is denoted as $P_{ij}^{(2)} = \sum_{l \in \mathcal{S}} P_{il} P_{lj}$, where the summation runs over all possible intermediate states l. The generalization that the homogeneous Markov chain is in state j after k transitions, given it started in state i is.

$$P_{ij}^{(k)} = \mathrm{P}\left(X_{k+n} = j | X_n = i\right)$$

A method to compute these k-step transition probabilities, assuming a homogeneous Markov chain, are the Chapman-Kolmogorov equations, where we determine the probability that starting in state i, the process will go to state j in $k+m$ steps through intermediate state ℓ at the kth transition

$$\begin{aligned} P_{ij}^{(k+m)} &= \mathrm{P}\left(X_{k+m} = j | X_0 = i\right) \\ &= \sum_{\ell \in \mathcal{S}} \mathrm{P}\left(X_{k+m} = j | X_k = \ell, X_0 = i\right) \mathrm{P}\left(X_k = \ell | X_0 = i\right) \\ &= \sum_{\ell \in \mathcal{S}} \mathrm{P}\left(X_{k+m} = j | X_k = \ell\right) \mathrm{P}\left(X_k = \ell | X_0 = i\right) \\ &= \sum_{\ell \in \mathcal{S}} P_{\ell j}^{(m)} P_{i\ell}^{(k)}. \end{aligned} \tag{2.10}$$

Starting from the initial distribution $\boldsymbol{p}^{(0)}$, the state distribution after one time step is determined by

$$p_j^{(1)} = \sum_{i \in \mathcal{S}} p_i^{(0)} P_{ij} \qquad \forall\, j \in \mathcal{S},$$

or in matrix notation $\boldsymbol{p}^{(1)} = \boldsymbol{p}^{(0)}\boldsymbol{P}$. After a second time step $\boldsymbol{p}^{(2)} = \boldsymbol{p}^{(1)}\boldsymbol{P} = \boldsymbol{p}^{(0)}\boldsymbol{P}^2$, and correspondingly for any time step k

$$\boldsymbol{p}^{(k)} = \boldsymbol{p}^{(0)}\boldsymbol{P}^k. \tag{2.11}$$

Definition 2.24 (Stationary distribution). A distribution $\boldsymbol{\pi}$ is called stationary distribution of a Markov chain if

$$\boldsymbol{\pi} = \boldsymbol{\pi}\boldsymbol{P}$$

holds. □

This is equivalent to $\boldsymbol{\pi}(\boldsymbol{P} - \boldsymbol{I}) = 0$, where $\boldsymbol{I}$ is the identity matrix. Solving this equation for $\boldsymbol{\pi}$ corresponds to a left eigenvector of $\boldsymbol{P}$ that belongs to an eigenvalue equal to one. The questions arise, under which circumstances a non trivial stationary distribution exists and if the distribution is unique. The answer is given in Theorem 2.2 after some necessary definitions about classifications of Markov chains.

Definition 2.25 (Irreducible Markov chains). A Markov chain is said to be irreducible if a state i is reachable from any other state j, i.e., if there exists a $k \geq 0$ such that $P_{ij}^{(k)} > 0$. □

Definition 2.26 (Periodicity of Markov chains). The period of a state in a Markov chain is defined as

$$d_i = gcd\{n : \boldsymbol{P}_{ii}^{(n)} > 0\},$$

where *gcd* is the greatest common divisor.

1. State i is said to be periodic if $d_i > 1$.

2. State i is said to be aperiodic if $d_i = 1$.

3. A Markov chain is aperiodic, if all states are aperiodic.

□

Theorem 2.1 (Period of irreducible Markov chains). Every state in an irreducible Markov chain has the same period. Thus in an irreducible Markov chain, if one state is aperiodic, then the chain is aperiodic. □

Proof. See Appendix A.1. □

Definition 2.27 (Recurrent and transient states). A state $i \in \mathcal{S}$ of a Markov chain is called recurrent if

$$\mathrm{P}\left(X_n = i \text{ for some } n \geq 1 | X_0 = i\right) = 1,$$

which is the probability that the Markov chain, starting from i, will return to i. If this probability is less than one, state i is called transient. A chain is recurrent if all states are recurrent. More formally: State i is recurrent if

$$\sum_{n=1}^{\infty} P_{ii}^{(n)} = \infty$$

and transient if $\sum_{n=1}^{\infty} P_{ii}^{(n)} < \infty$. □

Definition 2.27 is directly related to the question how long it takes to visit a certain state.

Definition 2.28 (Mean recurrence time). Let

$$T_i = \min\{n \geq 1 : X_n = i\}$$

be the time until state i is visited for the first time and define $T_i = \infty$ if the state is never visited. The expected value of the return time to state i is called the mean recurrence time of state i

$$M_i = \mathbf{E}\left[T_i | X_0 = i\right].$$

A recurrent state is called positive recurrent if $M_i < \infty$. □

The following theorem shows that the distribution of an irreducible and aperiodic Markov chain converges to the stationary distribution starting from any initial distribution.

Theorem 2.2 (Convergence). Given a finite state space irreducible Markov chain. If all states of the chain are positive recurrent, then there exists a unique stationary distribution $\boldsymbol{\pi}$ with $\pi_i = \frac{1}{M_i}$ for each $i \in \mathcal{S}$, where M_i is the mean recurrence time of state i.
If the chain is aperiodic, then

$$\lim_{k \to \infty} \boldsymbol{p}^{(k)} = \boldsymbol{\pi}.$$

□

Proof. See Appendix A.2. □

Definition 2.29 (Ergodic Markov chains). An irreducible and aperiodic Markov chain that is recurrent is called ergodic. □

In the following we introduce the concept of absorbing Markov chains.

Definition 2.30 (Absorbing Markov chains). A special case of a recurrent state is called absorbing state, where the probability to stay in state i is $P_{ii} = 1$. An irreducible Markov chain is absorbing if it has at least one absorbing state. □

Given an absorbing Markov chain with r absorbing states and t transient states, the states can be sorted in the transition matrix such that the transient states come first

$$\boldsymbol{P} = \begin{pmatrix} \boldsymbol{Q} & \boldsymbol{R} \\ \boldsymbol{0} & \boldsymbol{I}_r \end{pmatrix},$$

where $\boldsymbol{Q}$ it a $t \times t$ matrix of the transient states, $\boldsymbol{I}_r$ the $r \times r$ identity matrix of the absorbing states and $\boldsymbol{R}$ a $t \times r$ matrix for transitions of transient states to absorbing states.

Definition 2.31 (Fundamental matrix). Given an absorbing Markov chain with transition matrix $\boldsymbol{P}$, the matrix

$$\boldsymbol{N} = (\boldsymbol{I}_t - \boldsymbol{Q})^{-1}$$

is called the fundamental matrix, where $\boldsymbol{I}_t$ is the $t \times t$ identity matrix. The expected number of times, the Markov chain is in transient state j given that it started in the transient state i are the entries N_{ij} of $\boldsymbol{N}$. □

Concept 2.1 (Time to absorption). Given an absorbing Markov chain starts in state i, then t_i is the expected number of steps before the chain is absorbed. Let $\boldsymbol{t}$ be a column vector, where t_i is the ith entry, then

$$\boldsymbol{t} = \boldsymbol{N}\boldsymbol{1},$$

where $\boldsymbol{1}$ is the all one column vector. □

2.3.2. Continous-time Markov Chains

Based on the theory of discrete-time Markov chains, we consider now stochastic processes $\{X_t\}_{t\geq 0}$ taking elements from a finite state space $\mathcal{S}$ over continuous time $t \in \mathbb{R}_0^+ = [0, \infty)$. Note, that only those aspects of continuous-time Markov chains are introduced, that are relevant for the thesis. The continuous-time Markov chain (CTMC) is defined as follows:

Definition 2.32 (Continuous-time Markov chain). A stochastic process $\{X_t\}_{t\geq 0}$ is called a continuous-time Markov chain or continuous-time Markov process with initial distribution $\boldsymbol{p}^{(0)}$ if

1. each value X_t is a random variables on state space $\mathcal{S}$,

2. X_0 has distribution $\boldsymbol{p}^{(0)}$
3. and if for all $t_0 < t_1 < \ldots < t_{n+1}$ and $x_0, \ldots, x_{n+1} \in \mathcal{S}$ the Markov property

$$\mathrm{P}\left(X_{t_{n+1}} = x_{n+1} | X_{t_n} = x_n, \ldots, X_{t_0} = x_0\right) = \mathrm{P}\left(X_{t_{n+1}} = x_{n+1} | X_{t_n} = x_n\right)$$

is fulfilled.

□

In a CTMC, the amount of time, spent in each state is exponentially distributed. The Markov property holds as the exponential distribution is memoryless. Similar to the transition matrix in discrete-time Markov chains, we define

Definition 2.33 (Continuous-time transition matrix). A transition matrix $\boldsymbol{P}(t) = \{P_{ij}(t)\}_{i,j\in\mathcal{S}}$ is a stochastic matrix, where each row satisfies the following conditions:

1. $0 \leq P_{ij}(t) \leq 1 \qquad \forall\, i, j \in \mathcal{S}$,
2. $\sum_j P_{ij}(t) = 1 \qquad \forall\, i \in \mathcal{S}$.

The elements $P_{ij}(t)$ with $t, h \geq 0$ are the transition probabilities moving from state i at time h to state j at time $t + h$

$$P_{ij}(t) = \mathrm{P}\left(X_{t+h} = j | X_h = i\right) \qquad \forall\, i, j \in \mathcal{S},$$

□

Definition 2.34 (Homogeneous Continuous-time Markov chain). A Markov chain is time-homogeneous if the transition probabilities depend only on the time difference, i.e., for all $t, h \in \mathbb{R}_0^+$, we have

$$\mathrm{P}\left(X_{t+h} = j | X_h = i\right) = \mathrm{P}\left(X_t = j | X_0 = i\right) = P_{ij}(t) \quad \forall\, i, j \in \mathcal{S}.$$

□

Like in discrete time, we can compute transition probabilities starting from the Chapman-Kolmogorov equation of (2.10). An analogue derivation leads in the continuous case to

$$P_{ij}(t+h) = \mathrm{P}\left(X_{t+h} = j | X_0 = i\right) = \sum_{\ell\in\mathcal{S}} P_{\ell j}^{(h)} P_{i\ell}^{(t)}. \tag{2.12}$$

In matrix notation, this is the matrix multiplication of

$$\boldsymbol{P}(t+h) = \boldsymbol{P}(t)\boldsymbol{P}(h) \tag{2.13}$$

As fundamental parameter, we define a rate matrix $\boldsymbol{Q}$ to specify continuous-time Markov chains.

Definition 2.35 (Rate matrix). A rate matrix $\boldsymbol{Q} = \{Q_{ij}\}_{i,j\in\mathcal{S}}$ satisfies the following conditions:

1. $0 \leq -Q_{ii} < \infty \qquad \forall\, i \in \mathcal{S}$,
2. $Q_{ij} \geq 0 \qquad \forall\, i, j \in \mathcal{S}$ such that $i \neq j$,
3. $\sum_j Q_{ij} = 0 \qquad \forall\, i \in \mathcal{S}$.

Thus each off-diagonal entry can take values $Q_{ij} \in \mathbb{R}_0^+$, subject only to the constraint that the off-diagonal row sum is finite. The diagonal entries are chosen by

$$Q_{ii} = -\sum_{i \neq j} Q_{ij}.$$

□

We interpret the off-diagonal entries Q_{ij} as the rate of going from state i to j and the diagonal elements Q_{ii} as the rate of leaving i. We define a connection between the transition matrix and the rate matrix as $\boldsymbol{Q} = \frac{\mathrm{d}\boldsymbol{P}(0)}{\mathrm{d}t}$. Differentiation of (2.13) with respect to t and setting $t = 0$ leads to the so called *backward equation*

$$\frac{\mathrm{d}\boldsymbol{P}(h)}{\mathrm{d}t} = \boldsymbol{Q}\boldsymbol{P}(h).$$

Similarly, if we differentiate (2.13) with respect to h and set $h = 0$ we get the *forward equation*

$$\frac{\mathrm{d}\boldsymbol{P}(t)}{\mathrm{d}h} = \boldsymbol{P}(t)\boldsymbol{Q}.$$

In both cases, the initial condition is $\boldsymbol{P}(0) = \boldsymbol{I}$, the identity matrix. The solution for both differential equations is exactly the transition probability matrix at each time t

$$\boldsymbol{P}(t) = e^{\boldsymbol{Q}t} = \boldsymbol{U}\mathrm{diag}\{e^{\lambda_1 t}, e^{\lambda_2 t}, \ldots, e^{\lambda_{|\mathcal{S}|} t}\}\boldsymbol{U}^{-1}, \tag{2.14}$$

where λ_i are the eigenvalues of $\boldsymbol{Q}$ and columns of $\boldsymbol{U}$ are the corresponding right eigenvectors of $\boldsymbol{Q}$. Derivation of this matrix exponential can be found in Appendix A.3.
Let $\boldsymbol{p}^{(t)}$ be the state distribution over $\mathcal{S}$, with components

$$p_i^{(t)} = \mathrm{P}\left(X_t = i\right) \qquad i \in \mathcal{S},$$

with $0 \leq p_i^{(t)} \leq 1$ for $i \in \mathcal{S}$ and $\sum_{i \in \mathcal{S}} p_i^{(t)} = 1$. This is the probability that the Markov chain is at time t in state $i \in \mathcal{S}$. The initial distribution, is $\boldsymbol{p}^{(0)}$. If the chain $\{X_t\}_{t \geq 0}$ has initial distribution $\boldsymbol{p}^{(0)}$, then time t later the distribution $\boldsymbol{p}^{(t)}$ will be given by

$$\boldsymbol{p}^{(t)} = \boldsymbol{p}^{(0)}\boldsymbol{P}(t).$$

If the initial and target distributions are the same, the chain is said to be stationary. In the following, we assume that the Markov chain allows any state to change into any other state in finite time, i.e., $P_{ij}(t) > 0$. As already discussed for the DTMC, a chain, where all states are positive recurrent has unique stationary distribution for $t \to \infty$. We extend the stationarity to CTMC.

Concept 2.2 (Stationary Distribution CTMC). Let $\{X_t\}_{t\geq 0}$, be a continuous-time Markov process, the stationary distribution is given by

$$\boldsymbol{\pi} = \boldsymbol{\pi}\boldsymbol{P}(t)$$

which is equivalent to

$$\boldsymbol{\pi}\boldsymbol{Q} = 0.$$

□

As in the discrete case, $\boldsymbol{\pi}$ is the left eigenvector of $\boldsymbol{P}(t)$ that belongs to an eigenvalue equal to one. From above, we know that $\boldsymbol{Q}$ has eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_{|\mathcal{S}|}$ and $\boldsymbol{P}$ has eigenvalues $e^{\lambda_1 t}, e^{\lambda_2 t}, \ldots, e^{\lambda_{|\mathcal{S}|} t}$. If $\boldsymbol{P}$ has $e^{\lambda_i t} = 1$, the eigenvalue of $\boldsymbol{Q}$ must be zero, i.e., $\lambda_i = 0$. As the CTMC is completely determined by $\boldsymbol{\pi}$ and $\boldsymbol{Q}$, we denote it in short as $\mathcal{MC}_{CT}(\boldsymbol{\pi}, \boldsymbol{Q})$.

Concept 2.3 (Reversibility). An ergodic $\{X_t\}_{t\geq 0}$ is said to be time-reversible, iff the detailed balance equations

$$\pi_i Q_{ij} = \pi_j Q_{ji} \qquad \forall\, i, j \in \mathcal{S} \text{ such that } i \neq j$$

are fulfilled, which is equivalent to

$$\pi_i P_{ij}(t) = \pi_j P_{ji}(t) \qquad \forall t \geq 0,\ \forall\, i, j \in \mathcal{S} \text{ such that } i \neq j.$$

□

This means that the flow from state i to j is the same as the opposite direction j to i.

2.4. Biology

This part provides the biological background necessary for understanding the problems covered within this thesis. For a detailed introduction into the concepts presented here, see [BAJ+11], where large parts of this section stem from.
In Figure 2.3 the basic communication system of molecular biology is shown. Starting with a simplified representation, we will explain the terms and abbreviations in more detail within the following sections. The *DNA* contains the genetic information stored in the *genome*, which is the data that should be transmitted. The DNA message is encoded to the *mRNA*, an operation that is called *transcription*. The mRNA is the channel of the system and *point mutations* that distort the mRNA message correspond to the noise. During *translation* the mRNA is repaired and decoded into a *protein*.
In Section 2.4.1 the basic storage media for genetic information are introduced and in Section 2.4.2 the blocks of information processing are explained in more detail. Definitions of terms and notations referred to within the thesis are presented in Section 2.4.3.

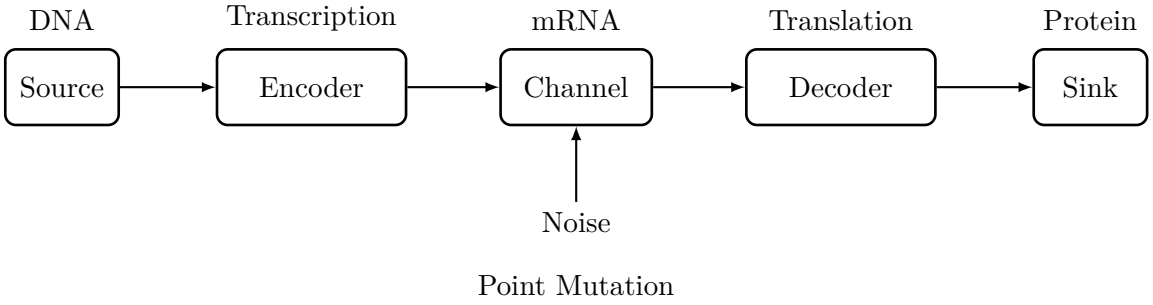


Figure 2.3.: Model of a information transmission in molecular biology (modified version of [Yoc92]).

2.4.1. Information Storage

Organisms are classified into eukaryotes and prokaryotes, where the main difference is, that eukaryotic cells contain a cell nucleus and prokaryotic cells have no nucleus. Eukaryotes comprise all higher multicellular organisms, e.g., plants, animals, fungi or humans, whereas prokaryotes are simpler organisms, mostly unicellular, like bacteria and archaea. Within this thesis, we concentrate on bacteria. In the following paragraphs, the terms of Figure 2.3, are introduced.

DNA and RNA All organisms store their genetic information within the Deoxyribonucleic acid (DNA). The backbone of DNA is build from blocks consisting of phosphate and sugar that are connected to a chain. From each sugar-phosphate block one of four possible nucleobases protrudes, building the so called nucleotides. The four bases Adenine (A), Guanine (G), Cytosine (C) and Thymine (T) have different chemical structures that are indicated in Figure 2.4 on page 24. Bases (A, G) consist of two carbon-nitrogen rings and belong to the group of purines, whereas (T, C) that consist of only one carbon-nitrogen ring are pyrimidines. DNA is double stranded, meaning that it consists of two strands, where (A, T) and (G, C) bind together forming base pairs. We say that A is the complementary base of T and vice versa and the same holds for G and C. We see also from the figure, that A and T are connected by only two hydrogen bonds, whereas G and C have three hydrogen bonds, a fact that will be of interest during the transcription process. DNA has also a direction; the forward strand or sense strand goes from the five prime end $5'$ to the three prime end $3'$, where the numbers are based on the position of the carbon within the sugar ring of the DNA backbone. The direction from $3'$ to $5'$ is the reverse strand or antisense strand. The strands of DNA are twisted, forming a double helix (indicated on the right panel of Figure 2.4).

During protein synthesis, specified parts of DNA are transcribed into single stranded Ribonucleic acid (RNA), by copying one strand of DNA (detailed description in Section 2.4.2). During this process base thymine is replaced by Uracil (U).

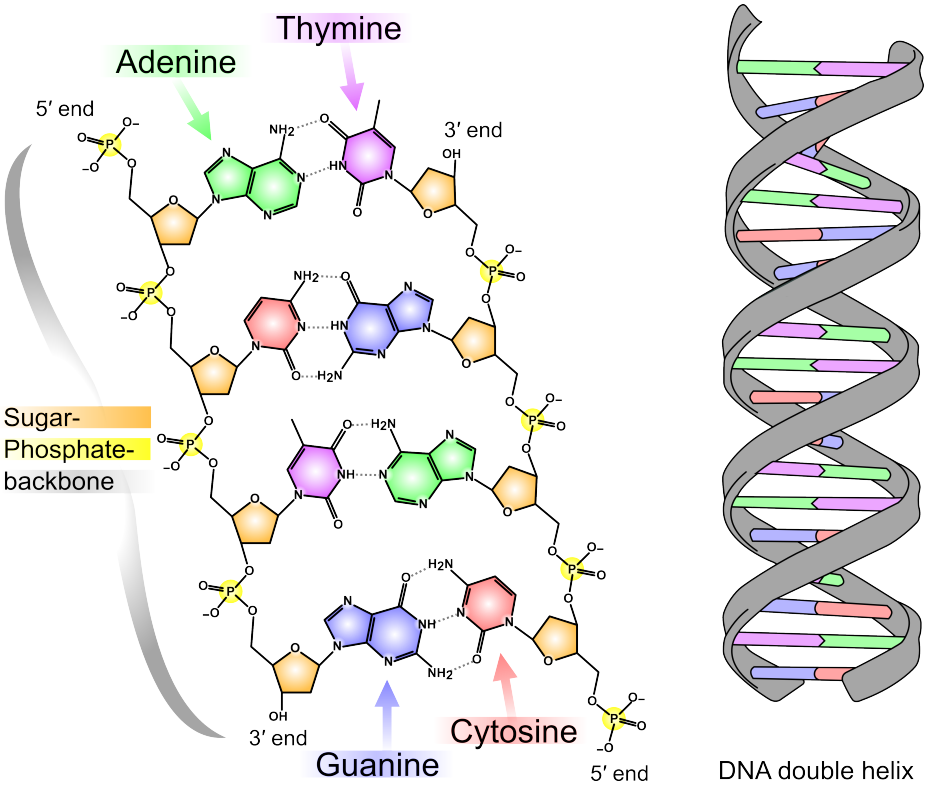


Figure 2.4.: Building blocks and structure of DNA (modified and merged version of [Bal13, Spo10]). The chemical structures consist of nitrogen (N), hydrogen (H), oxygen (O), carbon (empty) and phosphate (P).The complementary base pairs are connected by hydrogen bonds.

Genes and Proteins Proteins are molecules that turn the genetic information, stored in the DNA, into action. Proteins consist of long chains of amino acids. There exist 20 different amino acids that are connected based on the sequence of RNA molecules during translation (detailed description in Section 2.4.2). Each protein folds into an individual three-dimensional structure. In this form proteins carry out essential functions needed for living, e.g., as enzymes they catalyze reactions. DNA sequences contain the templates for thousands of different proteins. Those parts of DNA that are copied into RNA to synthesize functional proteins are called genes. The speed of protein production is controlled by regulatory entities. This means that genes are transcribed and translated according to the individual needs of the cell. This regulation is also stored within the DNA.

Genome The total amount of DNA of a cell or an organism is called the genome, which contains one or several chromosomes. The organisms studied within this thesis carry a single chromosome containing the DNA in form of a ring, where the genes on the DNA are

densely packed. The length of the genome varies significantly between organisms, where one of the shortest known bacterial genomes has $112 \cdot 10^3$ base pairs [BM13]. For comparison human beings have a genome of length $3.2 \cdot 10^9$ base pairs. Also the number of genes varies extremely: Human beings have approximately $21 \cdot 10^3$ genes, whereas organisms with only 137 genes exist [MB14]. One of the best studied model organisms belonging to the group of prokaryotes is *Escherichia coli*, where laboratory strain *K-12* has around $4.6 \cdot 10^6$ base pairs with approximately 4300 genes.

2.4.2. Information Processing

The process of protein synthesis based on a gene is called gene expression and consists of two steps, namely transcription and translation. From Figure 2.3 on page 23, we already know that basically DNA is transcribed into mRNA and translated into a protein. The processing of genetic information, is called the central dogma of molecular biology, a term coined by Crick in 1958 [Cri58].

Transcription Genes, the coding regions of double stranded DNA, are synthesized into single stranded messanger RNA (mRNA) by an enzyme called RNA polymerase. The polymerase consists of two subunits, where the so called σ-factor searches for a synchronization signal on the DNA, called promoter. The promoter shows the polymerase the upcoming transcription start site that marks the start of mRNA synthesis. Now the polymerase unwinds the DNA double helix and copies the antisense strand into the mRNA transcript by adding the complement nucleotides (see Figure 2.5). After the process started, the σ-factor is relieved and the polymerase copies around 50 nucleotides per second making one error per 10^4 copied nucleotides until the enzyme reaches a termination signal. This termination signal builds a hairpin structure, that interrupts transcription mechanically and releases the DNA and the synthesized mRNA. Afterwards, the polymerase binds to another σ-factor and restarts the process of transcription.

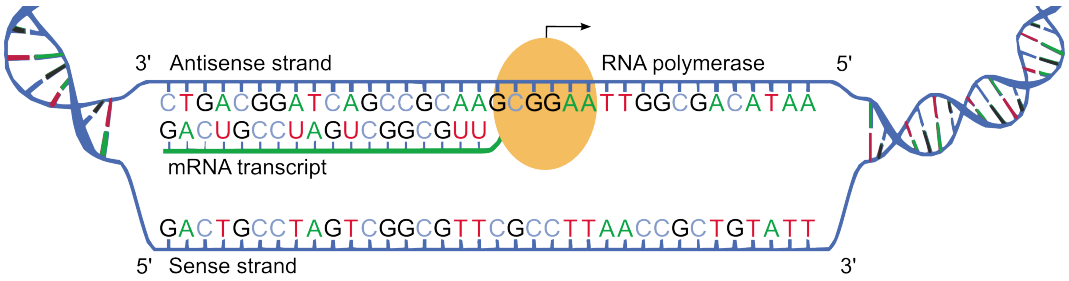


Figure 2.5.: Transcription process consists of separation of DNA strands and synthesis of mRNA by RNA polymerase based on one DNA strand (modified from [Sul09]).

The promoter consists of two regions, that are located shortly before the transcription start site counted as +1. One possible promoter is presented in Figure 2.6 on page 26, where the

first six nucleotides $TTGACA$ are followed by arbitrary $15 - 19$ nucleotides and by further six nucleotides $TATAAT$. The name of the regions express roughly the location with respect to the transcription start site. This promoter occurs in *Escherichia coli*. The corresponding σ-factor is kidney-shaped and searches for both regions. The rate of mRNA synthesis is influenced by the promoter. The better a promoter matches to a σ-factor, the higher is the probability that it binds to the DNA sequence which results in a higher frequency of gene transcription.

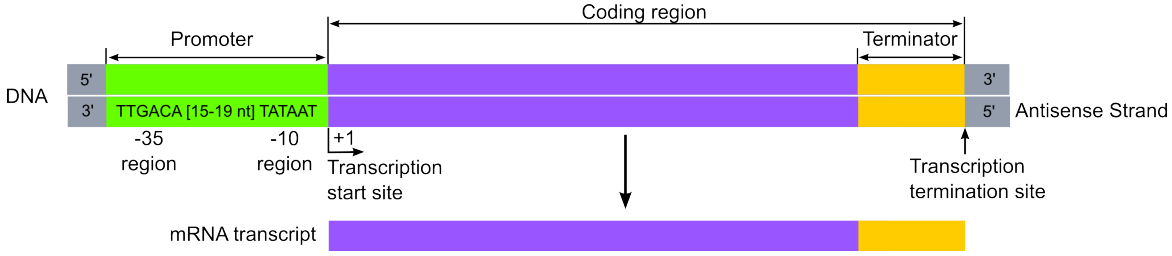
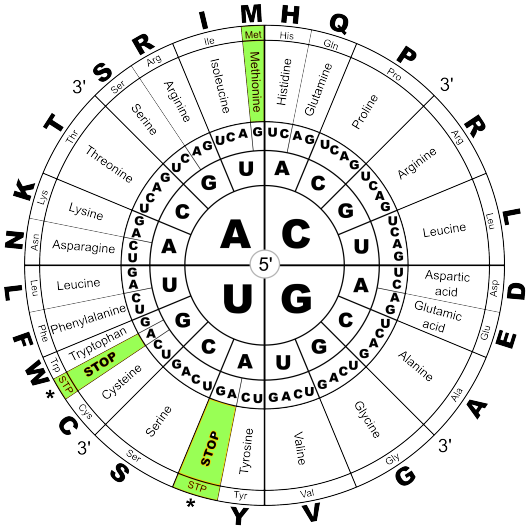


Figure 2.6.: Promoter detection by σ-factor. Transcription start and termination site specifies the boundaries that are copied into the mRNA trancript (based on [Zac]).

Translation The coding sequence of mRNA is translated into proteins by chaining amino acids by an enzyme called ribosome. The ribosome consists of two subunits, where the smaller subunit binds to the mRNA at the 5' end and searches for a synchronization signal on the mRNA, the Shine-Dalgarno sequence, that marks the upcoming translation start. Afterwards the larger subunit joins the complex and the translation begins. Since the mRNA consists of the nucleotide alphabet $\{A, C, G, U\}$, whereas there are 20 different amino acids, a mapping of nucleotides to amino acids is necessary, which is called the genetic code. The mRNA is translated by the ribosome in steps of three nucleotides, the so called codons. There are $4^3 = 64$ codons. Some of them code for the same amino acid according to Figure 2.7. On the left side, the genetic code is read from the centre of the circle to the outer circle, such that, e.g., AUG codes for the amino acid Methionine, which is abbreviated by the three letter code Met or the one letter code M. Codon AUG is of special interest, since the translation process starts with this so called start codon and ends with one of the three stop codons UAA, UAG, UGA. The stop codon maps to the stop label $STOP$ or $*$, but does not code for an amino acid. Figure 2.7(b) shows an alternative representation of the genetic code in block form. Note, that different organisms can have alternative genetic code mappings.
As already mentioned, not the complete mRNA sequence is translated into a protein, only the coding sequence that lies between a start and a stop codon preceeded by the Shine-Dalgarno sequence, e.g., ($AGGAGGU$) shown in Figure 2.8.



(a)

Second Base

First Base	U	C	A	G	Third Base
U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr STOP STOP	Cys Cys STOP Trp	U C A G
C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	U C A G
A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	U C A G
G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	U C A G

(b)

Figure 2.7.: (a) Genetic code sun shows mapping of codons to amino acids, where the first codon position is located in the middle circle and the third codon position at the outer circle (modified and merged from [art]). (b) shows an alternative representation is table form (modified version from [IA07]).

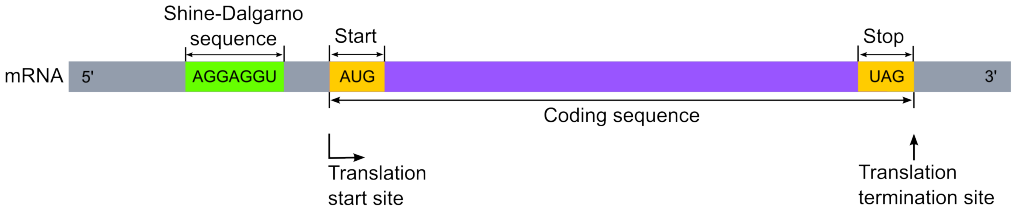


Figure 2.8.: Start and stop codons specify the boundaries of the coding sequence that is translated into a protein. The upcoming start codon is preceeded by the Shine-Dalgarno sequence.

Without the Shine-Dalgarno sequence, random amino acid sequences that build no functional protein could arise. Depending on the position, where the transcription process starts, there are three reading frames possible, each shifted by one nucleotide, where only one of them results in a protein. Figure 2.9 on 28 shows the possible amino acid chains, where the different reading frames are denoted as $\{+1, +2, +3\}$.
The process of protein synthesis is presented in Figure 2.10 on page 28. The ribosome slides along the mRNA codon-wise and comprises three codons at the same time. It serves as a platform for the so called transfer RNA (tRNA). Each tRNA is an adapter that consists

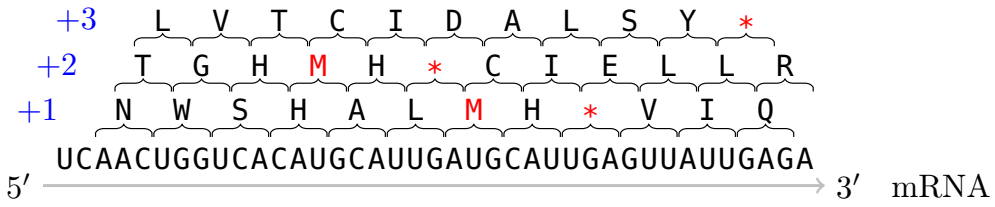


Figure 2.9.: Reading frames on the mRNA that result in different chains of amino acids.

of an amino acid and of an anticodon. The tRNA, that has a matching anticodon enters the ribosome. In the next step the amino acid of the tRNA is released and appended at the growing chain of amino acids, the final protein. In the last step, the empty tRNA is removed from the mRNA. In this way, the ribosome attaches around 20 amino acids per second making one error per 10^3 amino acids.

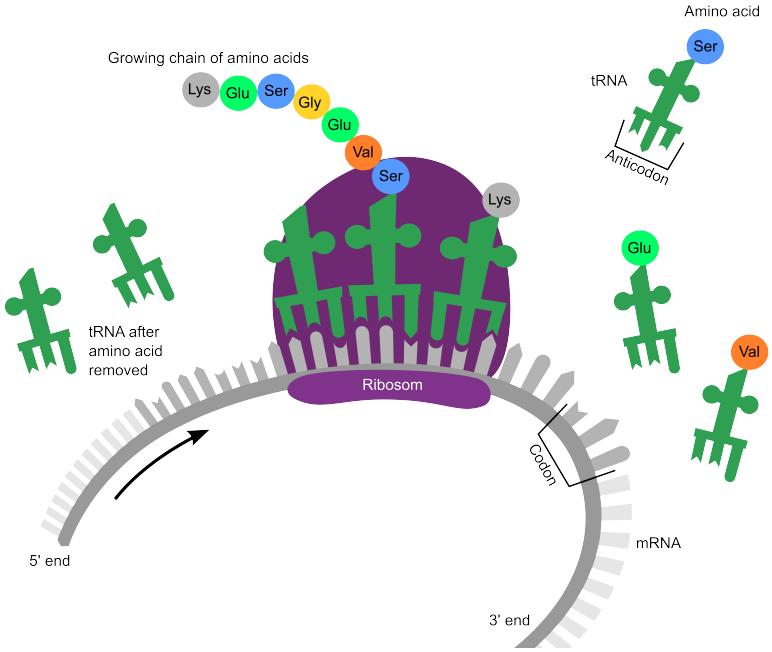


Figure 2.10.: Translation process synthesizes chain of amino acids from codons on mRNA following the mapping of the genetic code (modified from [Kel12]).

The resulting amino acid chain is only the primary protein structure. In another step, the protein builds secondary structures like pleated sheet or alpha helix and finally the protein folds into a tertiary protein structure. In prokaryotes simultaneous processing of transcription and translation is possible.

Mutations Noise in an information theoretic system describes a perturbation of the useful information that should be transmitted over the channel. Within a genetic system, noise describes anything that changes the result of protein synthesis in an unplanned way by external influences like radiation or internal influences during transcription, translation or folding [Yoc92].
In Figure 2.3 on page 23 noise is concentrated at the mRNA level, although there are several sources of noise. On a small scale, point mutations, deletion and insertions are distinguished. Point mutations are single changes of nucleotides, deletions remove single nucleotides and insertions introduce extra nucleotides. In general mutations can be harmful, changing the sequence in a way that the resulting protein loses its function, but without mutations evolution would not be possible. A point mutation, that occurs at the third codon position is often neutral, i.e., it does not affect the corresponding amino acid due to the degeneracy of the genetic code Figure 2.7 on page 27, whereas insertion and deletion errors change the reading frame and therefore the resulting protein sequence. Those mutations are called frame-shift errors.

2.4.3. Biological Notations

Basic Notations and Definition of Terms The process of protein synthesis presented above reveals, that all information that is necessary to generate a protein, is already contained in the DNA sequence. Within this thesis, we present models based on the DNA level. In this regard, DNA is a sequence of nucleotides, where the nucleotide or base alphabet is denoted as $\mathcal{B} = \{A, C, G, T\}$. The corresponding codon alphabet is $\mathcal{C}_{64} = \mathcal{B}^3 = \{A, C, G, T\}^3$. Since the chemical structure of nucleotides U and T are similar and DNA sequences are the basis of our considerations, T and U are not distinguished within this thesis.
We deal exclusively with bacterial genomes, that have the genetic code presented in Figure 2.7(b), where U is replaced by T, leading to stop codons $\{TAG, TAA, TGA\}$ and start codon ATG. According to [oM13, BIB$^+$97] additional start codons TTG, GTG and CTG and even ATT are observed, e.g., in *Escherichia coli*. But since CTG and ATT are extremely rare [MGR93, SFD$^+$82], we restrict the start codons considered within this thesis to $\{ATG, GTG, TTG\}$. We refer to this mapping as the standard genetic code $\mathcal{G}$ with translation table 11 in [oM13].
The amino acid alphabet is denoted as $\mathcal{A}_{21}$ including the stop label, whereas we denote the amino acid alphabet without stop label as $\mathcal{A}_{20}$. The genetic code is presented as a mapping

$$f : \mathcal{C}_{64} \to \mathcal{A}_{21}, c \mapsto a$$

or shortly written as $\mathcal{A}_{21} = \mathcal{G}(\mathcal{C}_{64})$. If we concentrate on amino acids without the stop label, we denote the corresponding codons as $\mathcal{C}_{61} = \mathcal{C}_{64} \setminus \{TAG, TGA, TAA\}$ with a mapping $\mathcal{A}_{20} = \mathcal{G}(\mathcal{C}_{61})$.
All sequences starting with a start codon $\{ATG, GTG, TTG\}$ and ending with a stop codon $\{TAG, TAA, TGA\}$ are called open reading frames (ORFs). Only ORFs that code for a functional protein are genes. Every gene, that appears in databases is an annotated ORF (aORF) or annotated gene, while the remaining ORFs are not-annotated ORFs (naORFs).

From Figure 2.9 on page 28, we already know that different reading frames exist. In case of DNA, the genetic triplet code provides six different reading frames Figure 2.11, where three reading frames are located on the forward strand $\{+1, +2, +3\}$ and three on the reverse strand $\{-1, -2, -3\}$.
We are interested in the theory behind overlapping genes, i.e., genes that share the same DNA locus with an already annotated protein coding gene (aORF). We define the reading frame, where this annotated gene is located as $\{+1\}$ reading frame, and denote the remaining frames accordingly with respect to the protein-coding reading frame. The ORFs that overlap with an aORF and form a potential overlapping gene pair are referred to as shadow ORFs (sORFs), a term introduced by [YSR+07]. In Figure 2.11 a potential overlapping gene pair is indicated, we see in reading frame $+2$ sequence ATG, CAT, TGA on the forward strand and on the reverse strand in reading frame -1 sequence $ATG, CAT, CAA, TGC, ATG, TGA$ that arises, if we read the bases in $5'$ to $3'$ direction. Note, another overlapping ORF (oORF) is located in -3. Assume that the sequence on the forward strand is an annotated gene, the corresponding frame is defined to be located in $+1$ and the labelled sORF lies in frame -2. In general sense and antisense overlapping genes exist, making a distinction if both ORFs of the overlapping gene pair are located on the same strand or if they are located on different strand.



Figure 2.11.: Six reading frames on the DNA, three on the forward strand and three on the reverse strand.

Complement and Reverse Complement We know that nucleotide A is the complement of T and vice versa, while G is the complement of C. Given a nucleotide sequence $(x_0 x_1 \dots x_{n-1})$ with $x_i \in \mathcal{B}$ on the forward strand, we obtain the complement sequence by replacing each element by its complement version

$$\bar{x}_i = \begin{cases} A & \text{if } x_i = T \\ T & \text{if } x_i = A \\ G & \text{if } x_i = C \\ C & \text{if } x_i = G \end{cases}.$$

The sequence that is encoded on the reverse strand is called the reverse complement sequence. To obtain it, we have to turn around the complement sequence resulting in $(\bar{x}_{n-1} \dots \bar{x}_1 \bar{x}_0)$.

Sequence length, GC content and codon usage Important parameters of the DNA of a bacterial genome are the length of the sequence, the GC content and the codon usage of the organism. Consider $\boldsymbol{x} \in \mathcal{B}^{n_g}$ as a DNA sequence. The GC content counts the number of bases G and C and relates it to the length of the sequence n_g,

$$GC(\boldsymbol{x}) = \frac{1}{n_g} \sum_{i=0}^{n_g-1} \mathbf{1}_{GC}(x_i), \quad \text{with} \quad \mathbf{1}_{GC}(x_i) = \begin{cases} 1 & \text{if } x_i \in \{G, C\} \\ 0 & \text{if } x_i \in \{A, T\}, \end{cases}$$

that leads to a range $GC(\boldsymbol{x}) \in [0, 1]$. The GC content can be multiplied by 100 to obtain a representation in percent. The codon usage of an organism is extracted from the sequences of the aORFs by counting the relative frequency of occurrences of all codons in $\mathcal{C}_{64}$. Each codon usage has a corresponding amino acid distribution.

Transition and Transversion Errors In Chapter 4, we concentrate on point mutations as genetic noise. On the DNA level, we distinguish two types of point mutations: Transitions on the one hand and transversions on the other hand. In Section 2.4.1 and Figure 2.4 on page 24, we introduced purines T, C and pyrimidines A, G that can be distinguished from their structure. Interchanges within those groups, i.e., when the chemical structure is maintained, are called transitions t_i. The remaining substitutions, i.e., interchanges between purine and pyrimidines, called transversions t_v, involve an replacement of chemical structures. An overview is given in Figure 2.12.

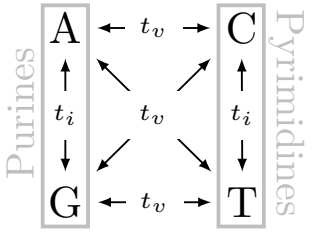


Figure 2.12.: Transitions t_i are interchanges within purines or pyrimidines. Transversions t_v are interchanges of purine for pyrimidine bases.

It is obvious that transitions occur at higher frequency than transversions. If we have a look at the genetic code sun Figure 2.7(a) on page 27, transitions (replacements of $A \leftrightarrow G$ and $C \leftrightarrow T$ in the third codon position lead with high probability to the same amino acid. Without mutations sequences can not evolve and adapt to changing environmental conditions. But mutations are not the only source of the evolution of new genes. New genes can develop from existing genes by gene duplication, shuffling or horizontal gene transfer. Two genes that descend from the same ancestor, are called homologous genes, they are

further distinguished as follows: If the developed sequences are found in different species, they are called orthologous genes and if both sequences are found within a single genome, they are called paralogous genes.

CHAPTER 3

Modeling Statistical Properties on Genome Level

DNA DOUBLE HELIX is the physical basis for inheritance. Some Open Reading Frames (ORFs) delimited by a start and a stop codon are genes, i.e., protein-coding sequences that are transcribed into mRNA and later translated into polypeptide chains. Those genes act as basic organizational units of prokaryotes at the genome level, since the coding density of bacterial genomes is high in comparison with eukaryotes [Pat99]. The genome of a typical bacterium is around 10^6 to 10^7 base pairs long, containing in the order of 10^4 to 10^5 ORFs, where only about 10^3 to 10^4 are annotated genes. Even though the number and typical length of ORFs may vary, bacteria share some characteristics of their ORF length distributions. In fact the length distribution as well as the overall codon usage is correlated with the GC content of the considered genome, which can be simply explained by AT-rich stop codons [OM96, GF95]. As most parts of a bacterial genome are covered by genes, the statistical behaviour of ORFs is expected to be determined by the codon usage. The interest of this chapter lies in the prediction of statistical ORF properties in order to investigate the potential existence of overlapping genes. Overlapping genes are two protein-coding sequences sharing a significant part of the same DNA locus in different reading frames. We know that an amino acid is encoded by a codon, that is comprised by three nucleotides. As DNA consists of two separate strands, up to six different reading frames exist, where three frames $\{+1, +2, +3\}$ are located on the forward strand in $5'$ to $3'$ direction and three on the reverse strand $\{-1, -2, -3\}$. For details on overlapping genes see Section 2.4. Overlapping genes are a well known phenomenon in viruses [Bar76], but the existence in bacterial genomes was for a long time highly controversial. Sanger *et al.* [SAB$^+$77] published the first completely sequenced genome, bacteriophage $\Phi X174$, that displays a number of such overlapping genes. However, because of space limitation of the capsid volume, viruses were thought to be a special case, as evolutionary pressure restricts genome size [CVB10, NBE$^+$83, Kra00, SG00]. In most cases, overlapping ORFs in alternative reading frames were omitted in bacterial genome annotations [YSR$^+$07, DBPS07, WAFS10] due to information content constraints [JC04] that should limit evolution since two genes are interlocked. Another explanation for the development of overlapping genes, were regulatory functions [NBE$^+$83, IHS$^+$00]. Later overlaps have been discovered in higher organisms [SGH86, WF86, Gil94, SP80, MGA83, MLM05]

and the number of non-trivially overlapping genes described in prokaryotes is increasing lately [MFS+00, BSN02, SL08, TBCM09, FBW+14]. For example, a series of publications about *Pseudomonas fluorescens* identified several overlapping genes, e.g., [SRL04, SL08], where for some of them the protein products have been identified using mass spectrometry [KSP+09]. Another pair was published by Tunca *et al.* [TBCM09], who used strand specific knock-out mutants to demonstrate a phenotype for both genes of the overlapping gene pair *dmdR1/adm*. Thus evidence increases that overlapping genes are no rare statistical artefact. We claim that, if two protein coding sequences in overlapping reading frames exist, the statistical parameters in those frame should deviate from random expectation, since bacterial genomes should be highly non-random. This chapter deals with statistical properties of ORFs and overlapping genes based on an analytical model.

First, in Section 3.1 we introduce the basic idea behind the model as well as the prediction limits of the model. This approach is overly simplistic and does not obtain good results for comparison with observations in real biological data. This is the reason, why in Section 3.2 a more realistic model is derived that can be applied to natural genomic data, which is done finally, in Section 3.3. In this section, we present a detailed comparison with an organism called *EHEC* and give a detailed explanation how our model is a useful instrument for prediction of global statistical parameters like, e.g., the length distribution of ORFs. Furthermore, our model can be applied to extract ORFs that are special from a statistical point of view and might therefore be candidates for overlapping gene pairs. Afterwards we present that our model predictions are very robust, by performing a large scale analysis over 71 genomes with differing GC contents. Several interesting aspects like influence of the GC content and genome sequence lengths are attended.
During the analysis of overlapping genes the question arises, if the statistical findings presented, are a consequence of the standard genetic code mapping. We investigate, if the standard genetic code is somehow optimized for the existence of long overlapping gene pairs in Section 3.4. Finally, Section 3.5 concludes the chapter. The results of this chapter, were partly published in [1, 2].

3.1. Analytical Model of Independent Nucleotides

Distributions of ORF lengths in bacterial genomes have already been studied by, e.g., Li [Li99], who proposed a piece-wise exponential model, where the exponential length distribution decays faster in the region of shorter ORFs and slower for longer ORFs. Interestingly, the author concluded that the transition point between both distributions is independent of the organism studied. However, his approach is empirical only and gives no mathematical explanation why such a transition point exists. McCoy *et al.* [MAG09] proposed a different model describing the natural ORF lengths as a mixture of two distributions, consisting of an exponential and, depending on the organism investigated, either a lognormal or a gamma distribution. They noted that the exponential part is connected to a random IID nucleotide model whereas their choice of the second part was motivated by the fact that the

length distribution of annotated proteins is empirically described by either a lognormal or a gamma distribution [Zha00]. We follow in parts Oliver *et al.* [OM96], who model the length of a random sequence between successive occurrences of stop codons as geometric distribution.

In this section we study the properties of ORFs based on a discrete-time random process $\{C_n\}_{0\leq n<N_C}$ of codons $C_n \in \mathcal{B}^3$, $\mathcal{B} = \{A, C, G, T\}$, where each codon consists of three independent and identically distributed nucleotides (X_1, X_2, X_3) with $X_j \in \mathcal{B}$, $j = \{1, 2, 3\}$ and common probability mass function $p_X(b) = p_{X_j}(b)$. The distribution depends on the GC content of an organism $GC \in [0, 1]$ and is given by

$$p_X(b) = \begin{cases} \frac{GC}{2} & \text{for } b \in \{G, C\} \\ \frac{1-GC}{2} & \text{for } b \in \{A, T\} \end{cases}.$$

We denote the length of the sequence in base pairs with $N_{BP}= 3N_C$.
In Section 3.1.1 the basic idea behind the model is derived and the limits of the model are presented in Section 3.1.2. The methods obtained are adapted from [1].

3.1.1. Description and Derivation of Model

We determine analytically the length distribution of the ORFs generated by the stochastic process $\{C_n\}_{0\leq n<N_C}$ as well as the average number of ORFs expected for the random sequence. Waiting time in a discrete process follows the geometric distribution introduced in Definition 2.10 as $\mathrm{Geo}\,(p_{stop})$. This distribution determines the probability of a success in a repeated experiment of independent trials. The following relations were in parts already published by, e.g., [OM96].

Concept 3.1 (Length distribution of ORFs). The length ℓ of an ORF measured in codons is given by the probability p_{stop} that a stop codon occurs, after $\ell - 1$ non-stop codons

$$P\,(L = \ell) = (1 - p_{stop})^{\ell-1} \cdot p_{stop} \qquad \text{for } \ell \in \{1, 2, 3, \ldots\}, \tag{3.1}$$

for $0 \leq p_{stop} \leq 1$. □

The average ORF length is given by the expected value of the geometric distribution

$$\mathbf{E}\,[L] = \sum_{\ell=1}^{\infty} \ell \cdot (1 - p_{stop})^{\ell-1}\, p_{stop} = \frac{1}{p_{stop}} \tag{3.2}$$

with standard deviation

$$\sigma = \frac{\sqrt{1 - p_{stop}}}{p_{stop}}.$$

The proof of (3.2) can be found in Appendix B.1. Given the standard genetic code, these equations can be evaluated as follows: The stop codon probability is given by the sum over

all three stop codon frequencies TAA, TAG and TGA. We follow the notation of [OM96] and denote the nucleotide frequencies

$$p_X(b) = \begin{cases} p & \text{for } b \in \{G, C\} \\ q & \text{for } b \in \{A, T\} \end{cases}. \tag{3.3}$$

From $\sum_{b \in \mathcal{B}} p_X(b) = 1$ the correspondence $2q + 2p = 1 \rightarrow p = \frac{1-2q}{2}$ follows.

$$p_{stop} = p_X(T)p_X(A)^2 + 2p_X(T)p_X(A)p_X(G) = q^3 + 2q^2p = q^3 + 2q^2\frac{1-2q}{2} = q^2 - q^3. \tag{3.4}$$

If we plug in $q = \frac{1-GC}{2}$ to (3.4) and (3.1), we obtain that the probability for an ORF of length ℓ codons depending only on the GC content of the nucleotide distribution. This relation is presented in Figure 3.1 and accordingly the average ORF length in Figure 3.2.

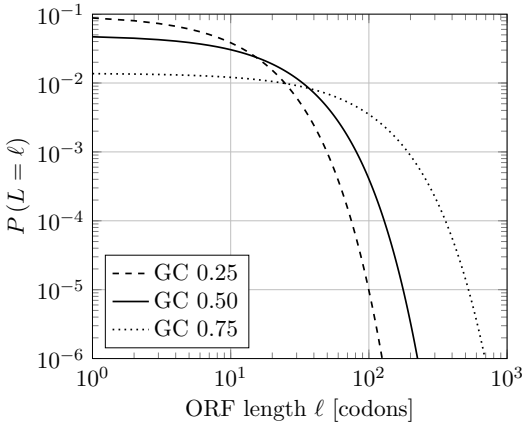


Figure 3.1.: Probability distribution of ORF length in random nucleotide sequences for three different GC contents (modified version of [OM96]).

The interpretation of Figure 3.1 and Figure 3.2 is straight forward: The higher the GC content of the nucleotide sequence, the lower is the corresponding probability for a stop codon which results in a longer waiting time for a stop codon and therefore longer ORFs. Since we are able to determine the length distribution of an ORF, we ask for the expected number of ORFs observed within a random sequence of arbitrary length.

Concept 3.2 (Expected number of ORFs). The number of ORFs expected from a stochastic process $\{C_n\}_{0 \leq n < N_C}$ of length N_C codons is given by

$$n_{ORF} = N_C \cdot p_{stop} \cdot \frac{p_{start}}{p_{start} + p_{stop}} = N_C \cdot n_{ORF}^{rel}, \tag{3.5}$$

with stop codon probability $0 \leq p_{stop} \leq 1$ and start codon probability $0 \leq p_{start} \leq 1$. The term n_{ORF}^{rel} is the relative average number of ORFs (independent of the sequence length). □

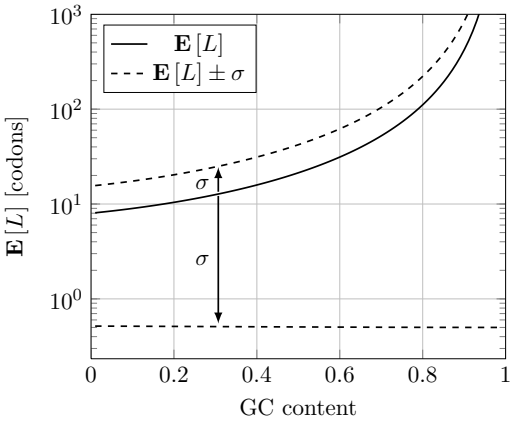


Figure 3.2.: Expected average ORF length in random nucleotide sequences of different GC contents (modified version of [OM96]).

The derivation of Concept 3.2 is an adapted version from [1]. As the stochastic process $\{C_n\}_{0 \leq n < N_C}$ is a random genome sequence, two regions have to be distinguished. The fraction of the sequence that is covered by an ORF is denoted with π_{ORF}, whereas the fraction between two consecutive ORFs is $1 - \pi_{ORF}$. This scenario is modelled using a discrete-time Markov chain as introduced in Section 2.3.1. The state space of the chain consists of only two states $\mathcal{S} = \{\text{Not ORF}, \text{ORF}\}$ and the transition matrix is completely determined by probabilities for a start codon p_{start} and a stop codon p_{stop}

$$\boldsymbol{P} = \begin{pmatrix} 1 - p_{start} & p_{start} \\ p_{stop} & 1 - p_{stop} \end{pmatrix}.$$

A diagram of the Markov chain is presented in Figure 3.3.

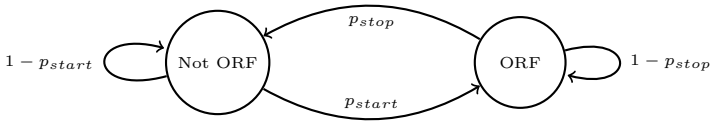


Figure 3.3.: Discrete-time Markov chain with only two states.

The fraction π_{ORF} corresponds directly to the stationary distribution (Definition 2.24) of the two state Markov chain. Solving $\boldsymbol{\pi} = \boldsymbol{\pi}\boldsymbol{P}$ leads to

$$\boldsymbol{\pi} = (1 - \pi_{ORF}, \pi_{ORF}) = \left(\frac{p_{stop}}{p_{start} + p_{stop}}, \frac{p_{start}}{p_{start} + p_{stop}} \right).$$

The random genome sequence of length N_C covers on average a region of

$$\ell_{ORF} = N_C \cdot \pi_{ORF}$$

with ORFs. Another way to determine the length of the part that is covered with ORFs, multiplies the average ORF length (3.2) with the number of ORFs denoted with n_{ORF}

$$\ell_{ORF} = n_{ORF} \cdot \mathbf{E}[L].$$

We obtain the number of ORFs expected from the random genome of N_C codons by

$$n_{ORF} = \frac{N_C \cdot \pi_{ORF}}{\mathbf{E}[L]} = N_C \cdot p_{stop} \cdot \frac{p_{start}}{p_{start} + p_{stop}} = N_C \cdot n_{ORF}^{rel}.$$

The relative number of ORFs (independent of the sequence length) is denoted with n_{ORF}^{rel}. Similar to the derivation of the stop codon probability (3.4), the probability for a start codon can be determined. We chose here the simplest case where the only start codon is ATG. Based on the nucleotide frequencies defined in (3.3) and $p = \frac{1-2q}{2}$ we obtain

$$p_{start} = p_X(A)p_X(T)p_X(G) = q^2\frac{1-2q}{2} = \frac{1}{2}q^2 - q^3 = q^2\left(\frac{1}{2} - q\right). \tag{3.6}$$

This can be adapted to any start codon combination easily, e.g., for start codons ATG, TTG and GTG we obtain $p^*_{start} = q\left(\frac{1}{4} - q^2\right)$.
Figure 3.4 shows the relative number of ORFs with the corresponding start and stop codon probabilities over different GC content values.

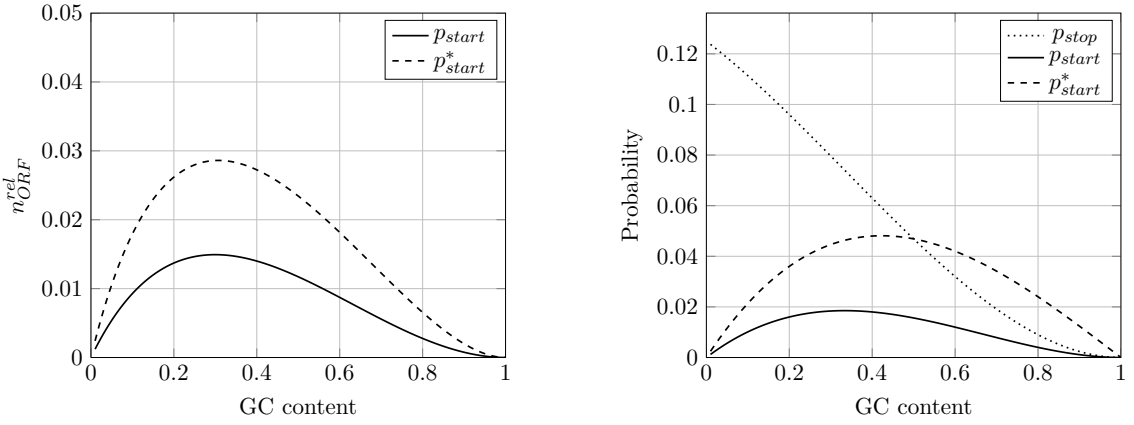


Figure 3.4.: Relative number of ORFs for a random nucleotide sequence with start codon ATG denoted with label p_{start} and p^*_{start} for additional start codons TTG and GTG together with the corresponding start and stop codon probabilities.

A numerical evaluation of n_{ORF}^{rel} results in a maximal ORF number at a GC content of 0.300 for p_{start} over ATG and a GC content of 0.307 for p^*_{start} over ATG, TTG and GTG. The absolute number of expected ORFs (3.5) for different values of GC content and sequence length is presented in Figure 3.5.

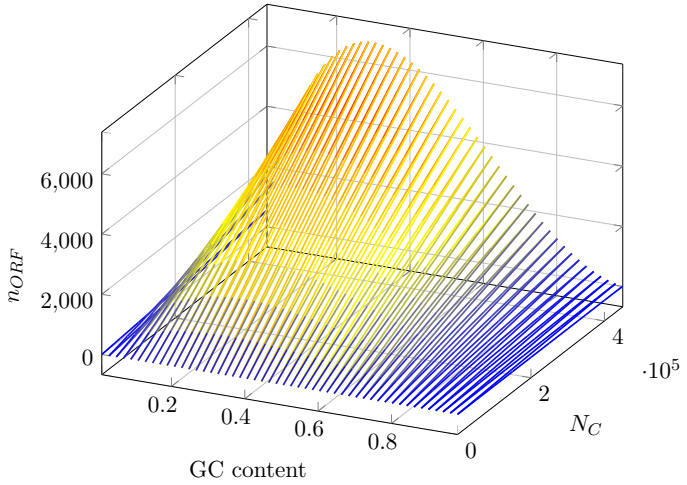


Figure 3.5.: Average number of ORFs for a random nucleotides sequence of $N_{BP}= 3N_C$ codons and different GC content values for start codon ATG.

3.1.2. Prediction Limits of Nucleotide Model

To determine the prediction limits of the nucleotide model, we apply the basic principles of hypothesis testing, introduced in Section 2.1.2. Assume, the length distribution of a natural genome is given. The question we want to consider is, if this distribution can be explained by chance. Therefore we formulate as null hypothesis H_0 for a statistical test: In a natural genome, the observation of an ORF of length ℓ is a result of pure chance. As test statistic, we introduce the ORF lengths emerged from the stochastic process $\{C_n\}_{0\leq n<N_C}$. Now we proceed as follows. We use the so called survival probability Concept 3.3 as p-value. This p-value has to be compared with a significance level α which answers the question, if the hypothesis can be rejected or not.
The probability to draw an ORF of length $L \geq \ell'$ is given by

$$\begin{aligned} P\left(L \geq \ell'\right) &= \sum_{\ell=\ell'}^{\infty} P\left(L = \ell\right) = \sum_{\ell=\ell'}^{\infty} \left(1 - p_{stop}\right)^{\ell-1} \cdot p_{stop} \\ &= p_{stop} \sum_{\ell=\ell'}^{\infty} \left(1 - p_{stop}\right)^{\ell-1} \\ &= \left(1 - p_{stop}\right)^{\ell'-1}, \end{aligned} \tag{3.7}$$

applying (3.1) and the geometric series for the last step (details Appendix B.1). Figure 3.6(a) on page 40 shows the corresponding probabilities at the example of three GC content values. For n_{ORF} trials, the probability to observe k ORFs of length $L \geq \ell'$ follows a Binomial distribution introduced in Definition 2.9

$$\mathrm{Bin}\left(n_{ORF}, p\right)(k) = \binom{n_{ORF}}{k} p^k (1-p)^{n_{ORF}-k}, \text{ with } p = \left(1 - p_{stop}\right)^{\ell'-1}.$$

Concept 3.3 (Survival Probability). The probability to observe at least one ORF with length $L \geq \ell'$ in n_{ORF} trials is given by

$$P_{SV}\left(\ell'\right) = 1 - \mathrm{Bin}\left(n_{ORF}, p\right)(0) = 1 - (1-p)^{n_{ORF}}, \text{ with } p = (1 - p_{stop})^{\ell'-1}. \tag{3.8}$$

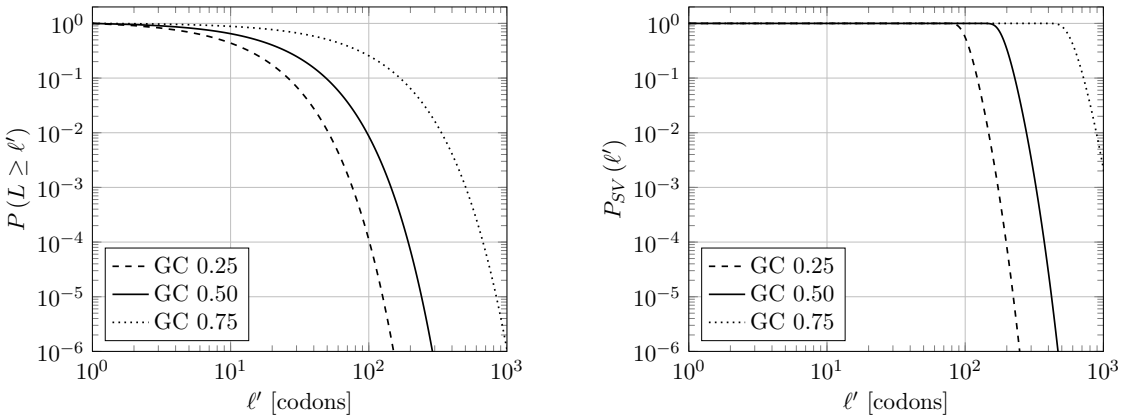
The survival probability can be interpreted as p-value against the hypothesis to observe an ORF exceeding a certain length ℓ'. □

Concept 3.4 (Survival Bound). The survival bound B_{SV}^{GC} for a fixed GC content, is the point, where the survival probability falls below the significance value α, i.e., $P_{SV}\left(\ell'\right) < \alpha$. If the ORF length ℓ' exceeds the survival bound, the hypothesis of having an ORF of length ℓ' produced from random process $\{C_n\}_{0 \leq n < N_C}$ is rejected. □

We set the significance value $\alpha = 0.01$ and denote the critical value as survival bound. At the example of $N_C = 5 \cdot 10^5$ random codons, Figure 3.6(b) shows the survival probability for three different GC content values. The corresponding number of ORFs as well as the survival bounds are given in Table 3.1.

GC	n_{ORF}	B_{SV}^{GC} [codons]
0.25	7324.2	148
0.50	5859.4	278
0.75	2050.8	890

Table 3.1.: Number of ORFs and Survival bound at the example of $N_C = 5 \cdot 10^5$ codons for different values of GC content and significance value $\alpha = 0.01$.



(a) (b)

Figure 3.6.: (a) Probability to draw an ORF of length $L \geq \ell'$. (b) Survival probability to observe at least one ORF with length $L \geq \ell'$ in n_{ORF} trials, which is calculated at the example of $N_C = 5 \cdot 10^5$ codons.

3.2. Discrete-time Markov Chain of Codons

Assuming independent nucleotides is definitely overly simplistic, and the results obtained in the previous section can not be applied to explain the observed behaviour in natural genomic data. Therefore, we extend the model in Section 3.2.1 assuming a discrete time first-order Markov chain. Afterwards, we present in Section 3.2.2 how this extended model, can be drawn back again to a model of geometric distributions. The results of Section 3.2.1 and Section 3.2.2 were partly published in [1, 2].

3.2.1. A First-Order Markov Chain Model

In this section we derive an analytical model based on the statistical properties of a sequence of random codons of length N_C given the following input process. In reading frame $\{+1\}$ we assume an ergodic discrete-time Markov chain $\{C_n\}_{0\leq n<N_C}$ of codons C_n over alphabet $\mathcal{C}_{64} = \mathcal{B}^3 = \{A, C, G, T\}^3$, with common probability mass function $p_{C_n}(c) = p_C(c), \forall\, c \in \mathcal{C}_{64}$ as initial distribution (Definition 2.22). Where the transitions between the codons are based on a given input transition matrix $\boldsymbol{P}$. We denote the individual nucleotides within each codon with $C_n = (N_n^{(1)}, N_n^{(2)}, N_n^{(3)})$. The Markov chain is ergodic (Definition 2.29), which is ensured by the transition matrix $\boldsymbol{P}$. We define that states $C_j, C_{j-1} \in \mathcal{C}_{64}$ are connected with $P_{C_{j-1},C_j} > 0$ and that all states have a period of one. A scheme of the considered Markov chain is presented in Figure 3.7.

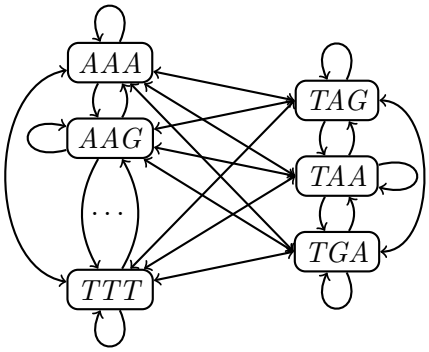


Figure 3.7.: Scheme of ergodic discrete-time Markov chain connecting all codons.

Our first goal is to determine the length distribution of the ORFs implicated by the model. In contrast to Section 3.1, where we assume independent codons, the expected distribution of the ORF lengths does not follow a geometric distribution at first hand. In Definition 2.30 we introduced the concept of absorbing Markov chains, that is applied here.
Given a transition matrix $\boldsymbol{P}$ of an ergodic finite state Markov chain with 64 codons as states, determines the corresponding transition matrix of the absorbing Markov chain $\boldsymbol{P}_{abs}$ by setting $P_{cc} = 1\, \forall\, c \in \{TAG, TGA, TAA\}$. This results in $t = 61$ transient states and $r = 3$

absorbing states. The states are sorted, such that

$$\boldsymbol{P}_{abs} = \begin{pmatrix} \boldsymbol{Q} & \boldsymbol{R} \\ \boldsymbol{0} & \boldsymbol{I}_3 \end{pmatrix}, \tag{3.9}$$

where $\boldsymbol{Q}$ is a (61×61) matrix of the transient states, $\boldsymbol{I}_3$ the (3×3) identity matrix of the absorbing states and $\boldsymbol{R}$ a (61×3) matrix for transitions of transient states to absorbing states.
The probability that the length of an ORF is ℓ, i.e., ends in an absorbing state $\{TAG, TGA, TAA\}$ started in a start codon $\{ATG, GTG, TTG\}$ ℓ steps ago is

$$P\left(L = \ell\right) = \boldsymbol{p}^{(0)} \cdot \boldsymbol{Q}^{\ell-2} \cdot \boldsymbol{R} \qquad \ell \in \{2, 3, \ldots\}, \tag{3.10}$$

where $\boldsymbol{p}^{(0)}$ is the input distribution of start codons. This equation results in a (1×3) vector, where each component corresponds to one of the absorbing states. For our problem, there is no difference in which of the three stop codons we end up, such that we sum over all three possibilities.

The relation of (3.10) holds for the alternative reading frames $\{-1, \pm 2 \pm 3\}$ in a similar manner, but for that we have to determine the corresponding transition matrices for each frame based on the matrix $\boldsymbol{P} = \{P_{C_{j-1},C_j}\}$ in reading frame $\{+1\}$. Note, a transition from C_{j-1} to C_j with realizations c_{j-1} and c_j is written as $P_{C_j|C_{j-1}}\left(c_j|c_{j-1}\right)$. In Figure 3.8 the connection between frame $\{+1\}$ and frames $\{+2, +3, -1\}$ is indicated. The complement version of a nucleotide $N_n^{(i)}$, $i \in \{1, 2, 3\}$ on the reverse reading frame is denoted with $\bar{N}_n^{(i)}$. Note that the reverse direction is read from right to left. As only codons in frame $\{+1\}$ are given, all transition matrices have to be determined based on three consecutive codons $C_{j-1}^{+1} C_j^{+1} C_{j+1}^{+1}$.
The correspondence in frame $\{-1\}$ is simple. We are interested in the transition probability from codon C_j^{+1} to C_{j-1}^{+1}, given the transition probabilities $P_{C_j^{+1}|C_{j-1}^{+1}}$. Bayes' rule leads us via

$$P_{C_{j-1}|C_j}\left(c_{j-1}|c_j\right) = \frac{P_{C_j|C_{j-1}}\left(c_j|c_{j-1}\right) P_{C_{j-1}}\left(c_{j-1}\right)}{P_{C_j}\left(c_j\right)}.$$

to the transition probability of interest

$$\begin{aligned} P_{C_i^{-1}|C_{i-1}^{-1}}\left(\bar{n}_{j-1}^{(3)}\bar{n}_{j-1}^{(2)}\bar{n}_{j-1}^{(1)}|\bar{n}_j^{(3)}\bar{n}_j^{(2)}\bar{n}_j^{(1)}\right) &= P_{C_{j-1}^{+1}|C_j^{+1}}\left(n_{j-1}^{(1)}n_{j-1}^{(2)}n_{j-1}^{(3)}|n_j^{(1)}n_j^{(2)}n_j^{(3)}\right) \\ &= \frac{P_{C_j^{+1}|C_{j-1}^{+1}}\left(n_j^{(1)}n_j^{(2)}n_j^{(3)}|n_{j-1}^{(1)}n_{j-1}^{(2)}n_{j-1}^{(3)}\right) P_{C_{j-1}^{+1}}\left(n_{j-1}^{(1)}n_{j-1}^{(2)}n_{j-1}^{(3)}\right)}{P_{C_j^{+1}}\left(n_j^{(1)}n_j^{(2)}n_j^{(3)}\right)}^{(+1)} . \end{aligned}$$

This transition matrix is denoted as $\boldsymbol{P}^{(-1)}$. For the other frames, we apply the conditional probability

$$P_{C_i|C_{i-1}}\left(c_i|c_{i-1}\right) = \frac{P_{C_i,C_{i-1}}\left(c_i, c_{i-1}\right)}{P_{C_{i-1}}\left(c_{i-1}\right)},$$

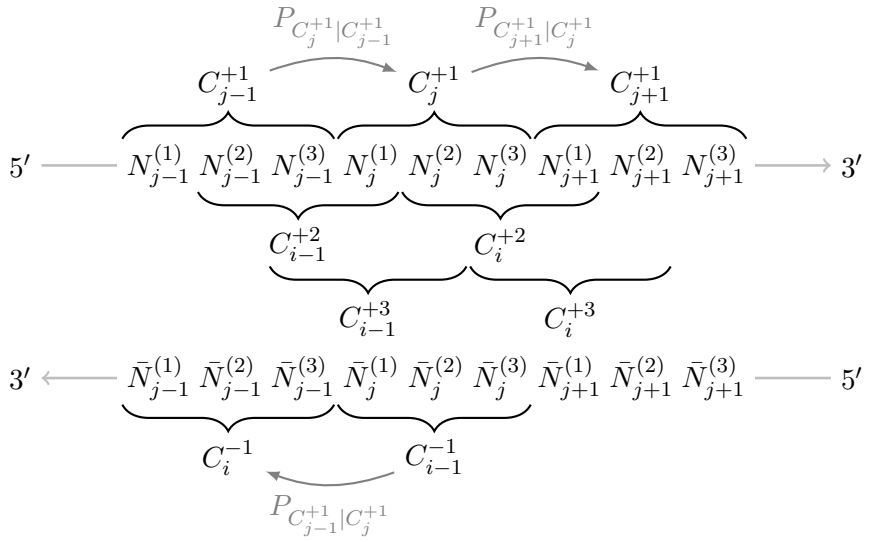


Figure 3.8.: Scheme of transitions within alternative reading frames in sequence direction.

which leads for example to transition matrix $\boldsymbol{P}^{(+2)}$ of reading frame $\{+2\}$

$$P_{C_i^{+2}|C_{i-1}^{+2}}\left(n_j^{(2)}n_j^{(3)}n_{j+1}^{(1)}|n_{j-1}^{(2)}n_{j-1}^{(3)}n_j^{(1)}\right) = \frac{\sum_{n_{j-1}^{(1)},n_{j+1}^{(2)},n_{j+1}^{(3)}\in\mathcal{B}} P_{C_j^{+1}|C_{j-1}^{+1}}\left(n_j^{(1)}n_j^{(2)}n_j^{(3)}|n_{j-1}^{(1)}n_{j-1}^{(2)}n_{j-1}^{(3)}\right) P_{C_{j+1}^{+1}|C_j^{+1}}\left(n_{j+1}^{(1)}n_{j+1}^{(2)}n_{j+1}^{(3)}|n_j^{(1)}n_j^{(2)}n_j^{(3)}\right)}{\sum_{n_{j-1}^{(1)},n_j^{(2)},n_j^{(3)}\in\mathcal{B}} P_{C_j^{+1}|C_{j-1}^{+1}}\left(n_j^{(1)}n_j^{(2)}n_j^{(3)}|n_{j-1}^{(1)}n_{j-1}^{(2)}n_{j-1}^{(3)}\right)}. \tag{3.11}$$

The proof of (3.11) can be found in Appendix B.3 together with the transition matrix of reading frame $\{+3\}$ in (B.4). The transitions in frames $\{-2,-3\}$ can be determined similarly. The difference is that the input transition matrix is $\boldsymbol{P}^{(-1)}$ which results in transition matrix $\boldsymbol{P}^{(-2)}$ and $\boldsymbol{P}^{(-3)}$ given in (B.5) and (B.6) in Appendix B.3.
For each frame $f \in \{-1, \pm 2 \pm 3\}$ we determine the unique stationary distributions $\boldsymbol{\pi}^{(f)}$ following Theorem 2.2 given the transition matrix $\boldsymbol{P}^{(f)}$ of an ergodic finite state Markov chain $\mathcal{MC}\left(\boldsymbol{\pi}^{(f)}, \boldsymbol{P}^{(f)}\right)$. Further, we determine the transition matrix $\boldsymbol{P}_{abs}^{(f)}$ of the corresponding absorbing Markov chain and derive the matrices $\boldsymbol{Q}_f$ and $\boldsymbol{R}_f$. The input distribution of start codons is based on the stationary distribution per reading frame

$$\boldsymbol{p}_f^{(0)} = \begin{cases} \pi_c^{(f)} & \text{if } c \in \{ATG, GTG, TTG\} \\ 0 & \text{else ,} \end{cases} \tag{3.12}$$

normalized to sum to one. For each reading frame $f \in \mathcal{F} = \{\pm 1, \pm 2 \pm 3\}$ we are now able to determine the individual ORF length distribution by application of (3.10).

Concept 3.5 (Length distribution of ORFs). The length ℓ of an ORF measured in codons in any of the six reading frames $f \in \mathcal{F}$ is

$$P(L = \ell) = \sum_{STOP} \boldsymbol{p}_f^{(0)} \cdot \boldsymbol{Q}_f^{\ell-2} \cdot \boldsymbol{R}_f \qquad \ell \in \{2, 3, \ldots\},\ STOP \in \{TAG, TGA, TAA\}, \quad (3.13)$$

where $\boldsymbol{Q}_f$ and $\boldsymbol{R}_f$ depend on $\boldsymbol{P}_{abs}^{(f)}$ applying (3.9) and $\boldsymbol{p}_f^{(0)}$ is based on $\boldsymbol{\pi}^{(f)}$ following (3.12). □

The average waiting time for a stop codon is given by the absorption time of the Markov chain. The expected number of steps starting in transient state c before being absorbed is the cth entry of the (61×1) vector $\boldsymbol{t}$ introduced in Concept 2.1

$$\boldsymbol{t} = \boldsymbol{N}\mathbf{1}, \quad (3.14)$$

where $\mathbf{1}$ is the all one column vector and $\boldsymbol{N}$ the fundamental matrix of Definition 2.31

$$\boldsymbol{N} = (\boldsymbol{I}_{61} - \boldsymbol{Q}_f)^{-1},$$

where $\boldsymbol{I}_{61}$ is the 61×61 identity matrix.

Next we determine the average ORF length, which corresponds to the mean recurrence time in Definition 2.28 together with Theorem 2.2. Given an unique stationary distribution of codons per reading frame $\boldsymbol{\pi}^{(f)}$, we interpret $\pi_c^{(f)}$ as the average proportion of time spent by the Markov chain in state c. Let T_c be the time until state c is visited for the first time. The mean recurrence time of state c is the expected value of the return time to state c

$$\mathbf{E}[T_c | X_0 = c] = \frac{1}{\pi_c^{(f)}},$$

that is the average first time the chain comes back to state c. Applied to our problem, we are interested in the average waiting time to reach a stop codon after we started in a stop codon. As each stop codon ends in an ORF and a typical ORF starts with a start codon and ends with a stop codon, the mean recurrence time is an approximation for the average ORF length.

Concept 3.6 (Mean Recurrence Time to Stop). The average waiting time from stop codon to stop codon is the mean recurrence time

$$t_M^{(f)} = \frac{1}{\sum_{c \in STOP} \pi_c^{(f)}}, \qquad STOP = \{TAG, TGA, TAA\}.$$

□

3.2.2. Simplification to Finite Mixture Model

Closer inspection of the model prediction reveals that the absorption times (3.14) of all transient state are approximately the same and therefore roughly the same as the recurrence time of Concept 3.6. This means that each state has on average the same probability to end in any of the three stop codons. Therefore it is straight forward to replace the first-order Markov chain Model $\mathcal{MC}\left(\boldsymbol{\pi}^{(f)}, \boldsymbol{P}^{(f)}\right)$ with its 64 states by a model with only one parameter. This leads directly to the geometric distribution we already know from (3.1), where the ORF length depends on the stop codon probability p_{stop} only. The corresponding stochastic process of IID codons will be denoted as $\mathcal{MC}_0\left(\boldsymbol{\pi}^{(f)}\right)$ which is a Markov chain without memory and initial distribution $\boldsymbol{\pi}^{(f)}$.
Given the stationary codon distribution $\boldsymbol{\pi}^{(f)}$, we determine the corresponding probability per reading frame by $p_{stop}^{(f)} = \pi_{TAG}^{(f)} + \pi_{TAA}^{(f)} + \pi_{TGA}^{(f)}$.

Concept 3.7 (Length distribution of ORFs). The length ℓ of an ORF measured in codons in any of the six reading frames $f \in \mathcal{F}$ is the weighted sum over six geometric distributions

$$P\left(L = \ell\right) = \sum_{f \in \mathcal{F}} p^{(f)} \left(1 - p_{stop}^{(f)}\right)^{\ell - 1} \cdot p_{stop}^{(f)} \qquad \ell \in \{1, 2, 3, \ldots\}, \tag{3.15}$$

where $0 \leq p_{stop}^{(f)} \leq 1$ is the stop codon probability in reading frame f and $p^{(f)}$ is a weighting parameter. □

We use the average case analysis described in Section 3.1 to determine the distribution of the ORFs in each reading frame. The number of ORFs expected from a sequence of N_C codons in reading frame f follows (3.5)

$$n_{ORF}^{(f)} = N_C \cdot p_{stop}^{(f)} \cdot \frac{p_{start}^{(f)}}{p_{start}^{(f)} + p_{stop}^{(f)}}, \tag{3.16}$$

where $p_{start}^{(f)}$ is the start codon probability in reading frame f determined from $p_{start}^{(f)} = \pi_{ATG}^{(f)} + \pi_{TTG}^{(f)} + \pi_{GTG}^{(f)}$. The absolute number of ORFs is accordingly

$$n_{ORF} = \sum_{f \in \mathcal{F}} n_{ORF}^{(f)}. \tag{3.17}$$

Then the weighting parameter for an individual reading frame $f \in \mathcal{F}$ is given by

$$p^{(f)} = \frac{n_{ORF}^{(f)}}{\sum_{j \in \mathcal{F}} n_{ORF}^{(j)}}. \tag{3.18}$$

The average ORF length extended to the weighted case, is the expected value of the weighted geometric distribution. Applying (3.16) to (3.18) we obtain

$$\mathbf{E}\left[L\right] = \sum_{f \in \mathcal{F}} \frac{p_{start}^{(f)}}{p_{start}^{(f)} + p_{stop}^{(f)}} \left(\sum_{j \in \mathcal{F}} p_{stop}^{(j)} \cdot \frac{p_{start}^{(j)}}{p_{start}^{(j)} + p_{stop}^{(j)}} \right)^{-1}. \tag{3.19}$$

A detailed derivation of the expected value is presented in Appendix B.2.
To answer the question, which ORF length is significant, if a natural genome is compared with the presented model, we obtain the survival probability introduced in (3.8) adapted to the weighted case (Appendix B.2).

Concept 3.8 (Survival Probability). The probability to observe at least one ORF with minimum length ℓ in any frame $f \in \mathcal{F}$ in $n_{ORF} = \sum_{j \in \mathcal{F}} n_{ORF}^{(j)}$ trials is called the survival probability

$$P_{SV}(\ell) = 1 - \left(1 - \left(\sum_{f \in \mathcal{F}} p^{(f)} \left(1 - p_{stop}^{(f)}\right)^{\ell - 1}\right)\right)^{n_{ORF}}, \tag{3.20}$$

where $0 \leq p_{stop}^{(f)} \leq 1$ is the stop codon probability in reading frame f and $p^{(f)}$ is the weighting parameter. □

Concept 3.9 (Survival Bound Model). The survival bound B_{SV} for a concrete input distribution is the point, where the survival probability falls below the significance value, i.e., $P_{SV}(\ell) < \alpha$. If the ORF length ℓ exceeds the survival bound, the hypothesis of having an ORF of length ℓ produced from our random sequence following $\mathcal{MC}_0\left(\boldsymbol{\pi}^{(f)}\right)$ is rejected. □

3.3. Predicting Statistical Properties of Prokaryotic Genomes

In the first place, we will show in Section 3.3.1 a detailed comparison of the model presented in Section 3.2 with a model organism called *EHEC*. In this section, we give a detailed explanation how concepts of the previous sections are applied to a prokaryotic genome. Further, we present how close the model predictions, e.g., codon usage, ORF length distribution and average ORF lengths are to the corresponding parameters observed in nature. It is astonishing, that such a simple concept, as a first-order Markov chain, is able to reproduce statistical properties of a genome that is subject to complex biological mechanisms. After analysis of some global parameters, we apply our model to alternative reading frames, to predict individual ORF length distributions, that lead to overlapping genes candidates. We will show that they are, according to our model, no mere artefacts.
In a second step in Section 3.3.2 we present a large scale comparison with 71 genomes of completely sequenced bacteria over a large range of different GC contents in order to illustrate that our model is not restricted to *EHEC*. Additionally, we discuss the robustness of our model regarding typical problems of overlapping genes discussed in literature. The results, where in parts already published in [1].

3.3.1. Application to a Particular Organism

At this point we compare the results determined from the discrete-time Markov Model $\mathcal{MC}\left(\boldsymbol{\pi}^{(f)}, \boldsymbol{P}^{(f)}\right)$ introduced in Section 3.2.1 and its simplification $\mathcal{MC}_0\left(\boldsymbol{\pi}^{(f)}\right)$ presented in Section 3.2.2 to a natural prokaryotic genome.

For this detailed analysis, we chose *Escherichia coli* O157:H7 EDL933 (*EHEC*, Accession number NC_002655), a bacterium that causes human disease and was sequenced in 2001 by [PPB+01]. This strain is well studied in hundreds of scientific contributions such that a large amount of reliable data is available. The organism has a GC content of 50.4% and a genome of length 5528445 base pairs. For the analysis, we have to process the genomic data that are publicly available as follows.

Data preparation of prokaryotic genome data We use Entrez to search and access the DNA sequence information from GenBank, which is provided by the National Center for Biotechnology Information (NCBI) [oM14a, oM14b] on behalf of the United States National Library of Medicine (NLM). From the GenBank files we extract the genome sequence as well as the gene locations, the annotated ORFs. Some of the DNA sequences contain symbols $\notin \{A, C, G, T\}$. Those symbols are substituted according to the Table 3.2, where the possible nucleotides are replaced randomly based on the GC content of the organism. Additionally we ignore genes if their length is no multiple of three or if they have a start codon $\notin \{ATG, GTG, TTG\}$ or stop codon $\notin \{TAA, TAG, TGA\}$. We skip the rare start codon CTG in our considerations. For details on start and stop codons see Section 2.4.2. The DNA sequence of *EHEC*[1] contained 6641 symbols $\notin \{A, C, G, T\}$, that were substituted. Overall 53 genes were deleted as they did not match the constraints defined above and 5233 genes remained.

Symbols	Possible substitutions	Origin and designation of symbols
R	A,G	puRine
Y	C,T	pYrimidine
M	A,C	aMino
K	G,T	Ketone
S	C,G	Strong interaction
W	A,T	Weak interaction
B	C,G,T	not-A, B follows A in alphabet
D	A,G,T	not-C, D follows C in alphabet
H	A,C,T	not-G, H follows G in alphabet
V	A,C,G	not-T/U, V follows U in alphabet
N	A,G,C,T	aNy

Table 3.2.: Undetermined nucleotides and their substitutions [CB85].

We denote the final set of genes as $\mathcal{G}$. The input transition matrix $\boldsymbol{P}^{(+1)}$ is determined from the codon transitions within the set of genes $\mathcal{G}$ of the prokaryotic genome. Note that each gene is adjusted such that it starts with the start codon, therefore we create the reverse complement sequence for genes on the negative strand. The number of codon transitions within each gene is counted and normalized to the row sum. As this procedure

[1]Date and time of the download were 04/09/14 at 16:43:24.

underestimates transitions to start codons as well as transitions from stop codons, latter will never be observed, we add pseudo-counts. We set the transition probabilities P_{C_{j-1},C_j} from stop codons to any codon according to the GC content (GC) of the organism. Each codon consists of nucleotides $c_j = \left(n_j^{(1)}, n_j^{(2)}, n_j^{(3)}\right)$. We define the sets

$$N_{AT}(c_j) = \{i \in \{1,2,3\} : n_j^{(i)} = A \vee n_j^{(i)} = T\}, \quad c_j = \left(n_j^{(1)}, n_j^{(2)}, n_j^{(3)}\right), c_j \in \mathcal{C}_{64}$$

and

$$N_{GC}(c_j) = \{i \in \{1,2,3\} : n_j^{(i)} = G \vee n_j^{(i)} = C\}, \quad c_j = \left(n_j^{(1)}, n_j^{(2)}, n_j^{(3)}\right), c_j \in \mathcal{C}_{64}$$

to count the occurrences of $\{A,T\}$ and $\{G,C\}$ and set the probability of transitions from stop codon $c_{j-1} \in \{TAA, TAG, TGA\}$ to any codon $c_j \in \mathcal{C}_{64}$ to

$$P_{C_j|C_{j-1}}\left(c_j|c_{j-1}\right) = \left(\frac{1-GC}{2}\right)^{|N_{AT}(c_j)|} \cdot \left(\frac{GC}{2}\right)^{|N_{GC}(c_j)|},$$

where $|\cdot|$ is the cardinality of the set. As transitions to start codons can appear within genes, we add a pseudo count to the existing number of transitions $N\left(c_j|c_{j-1}\right)$ before the row is normalized. We consider each transition from any codon $c_{j-1} \in \mathcal{C}_{64}$ to start codons $c_j \in \{ATG, TTG, GTG\}$ by adding the expected number of occurrences

$$N\left(c_j|c_{j-1}\right) = N\left(c_j|c_{j-1}\right) + \left\lfloor \left(\frac{1-GC}{2}\right)^{|N_{AT}(c_{j-1})|} \cdot \left(\frac{GC}{2}\right)^{|N_{GC}(c_{j-1})|} \cdot N(c_j) \right\rfloor, \tag{3.21}$$

where $N(c_j)$ is the number of occurrences of start codon c_j at the beginning of all genes $\mathcal{G}$.

For verification of the model prediction, we also investigate the codon usage of the genes. Each gene $g \in \mathcal{G}$ consists of a DNA sequence $\boldsymbol{x}_0^{n_g-1} = (x_0 x_1 \ldots x_{n_g-1})$ of n_g nucleotides or $\frac{n_g}{3}$ codons. To determine the codon usage of the model, we calculate the stationary distribution $\boldsymbol{\pi}^{(f)}$ of transition matrix $\boldsymbol{P}^{(+1)}$. The empirical codon usage $\hat{\boldsymbol{\pi}}^{(f)}$ is determined from the genes as follows.
For frame $\{+1\}$ we determine the empirical distribution $\hat{\boldsymbol{\pi}}^{(+1)} = \{\hat{\pi}_c^{(+1)}\}_{c\in\mathcal{C}_{64}}$ of blocks of length three in sequence $\boldsymbol{x}_0^{n_g-1}$ by counting the number of occurrences of codon c

$$N_{n_g}^{(+1)}(c) = |\{i \in [0, n_g-3] : \boldsymbol{x}_i^{i+2} = c \text{ if } i \equiv 0 \bmod 3\}|, \qquad \forall\, c \in \mathcal{C}_{64}$$

and calculating the sum over all genes divided by the overall number of codons considered

$$\hat{\pi}_c^{(+1)} = \frac{\sum_{g\in\mathcal{G}} N_{n_g}^{(+1)}(c)}{\sum_{g\in\mathcal{G}} n_g/3}, \qquad \forall\, c \in \mathcal{C}_{64}. \tag{3.22}$$

For comparison with the model, we need the empirical codon distributions within each reading frame. For example in frame $\{+2\}$ we consider the sequences shifted by one nucleotide

$$N_{n_g}^{(+2)}(c) = |\{i \in [1, n_g-5] : \boldsymbol{x}_i^{i+2} = c \text{ if } i \equiv 1 \bmod 3\}|,$$

$$\hat{\pi}_c^{(+2)} = \frac{\sum_{g\in\mathcal{G}} N_{n_g}^{(+2)}(c)}{\sum_{g\in\mathcal{G}} (n_g/3) - 1}, \quad \forall\, c \in \mathcal{C}_{64}.$$

and for frame $\{+3\}$ we consider the sequences shifted by two nucleotide

$$N_{n_g}^{(+3)}(c) = |\{i \in [2, n_g - 4] : \boldsymbol{x}_i^{i+2} = c \text{ if } i \equiv 2 \bmod 3\}|,$$
$$\hat{\pi}_c^{(+3)} = \frac{\sum_{g\in\mathcal{G}} N_{n_g}^{(+3)}(c)}{\sum_{g\in\mathcal{G}}(n_g/3) - 1}, \quad \forall\, c \in \mathcal{C}_{64}.$$

For the negative frames, we take the reverse complement versions of sequences $\boldsymbol{x}_0^{n_g-1}$.
For the comparison of the model with *EHEC* we follow the line of arguments presented in [1]. Deviations in the resulting parameters are caused by replacement of the *EHEC* stem investigated and omission of start codon *CTG*. In a first step we compare the resulting codon usage predicted by the model in comparison with the empirical codon usage of the natural genome. To determine the properties of the ORFs predicted by the model, we need start- and stop codon probabilities that will be compared in a second step.

Global Statistical Parameters

Prediction of Codon Usage and Amino Acid Distribution In this paragraph we compare in each reading frame $f \in \{\pm 1, \pm 2 \pm 3\}$ the empirical codon usage $\hat{\boldsymbol{\pi}}^{(f)}$ with the prediction of the model $\boldsymbol{\pi}^{(f)}$ extracted from the transition matrices $\boldsymbol{P}^{(f)}$. Figure 3.9 on page 50 and Figure 3.10 on page 51 show the empirical codon usage in comparison with the stationary codon distributions of the model on the forward and the reverse strand. Given the codon distributions per reading frame, it is straight forward to determine the corresponding amino acid distribution based on the genetic code (Figure 3.11 on page 52 and Figure 3.12 on page 53). For the amino acid distribution $\boldsymbol{\pi}_{\mathcal{A}_{21}}^{(f)}$ we just have to sum up all codon probabilities that belong to the same amino acid. The distributions given by the model are very close to the empirical distributions of *EHEC*.
We compare the predicted with the empirical amino acid distribution by application of the Kullback-Leibler divergence $D(\cdot||\cdot)$, which was introduced in Definition 2.14, over the amino acid alphabet $\mathcal{A}_{21}$. Further, we determine the ℓ^2 norm $||\cdot - \cdot||$ given in Definition 2.1 over $\mathcal{A}_{21}$. The distances of the amino acid distribution of *EHEC* compared with the predictions of the model are shown in Table 3.3 on page 54. Additionally, the amino acid distribution of *EHEC* against a random amino acid distribution (RND), where the probabilities of all amino acids are determined according to the GC content of the organism is compared with the amino acid distribution of *EHEC* in frame $\{+1\}$.

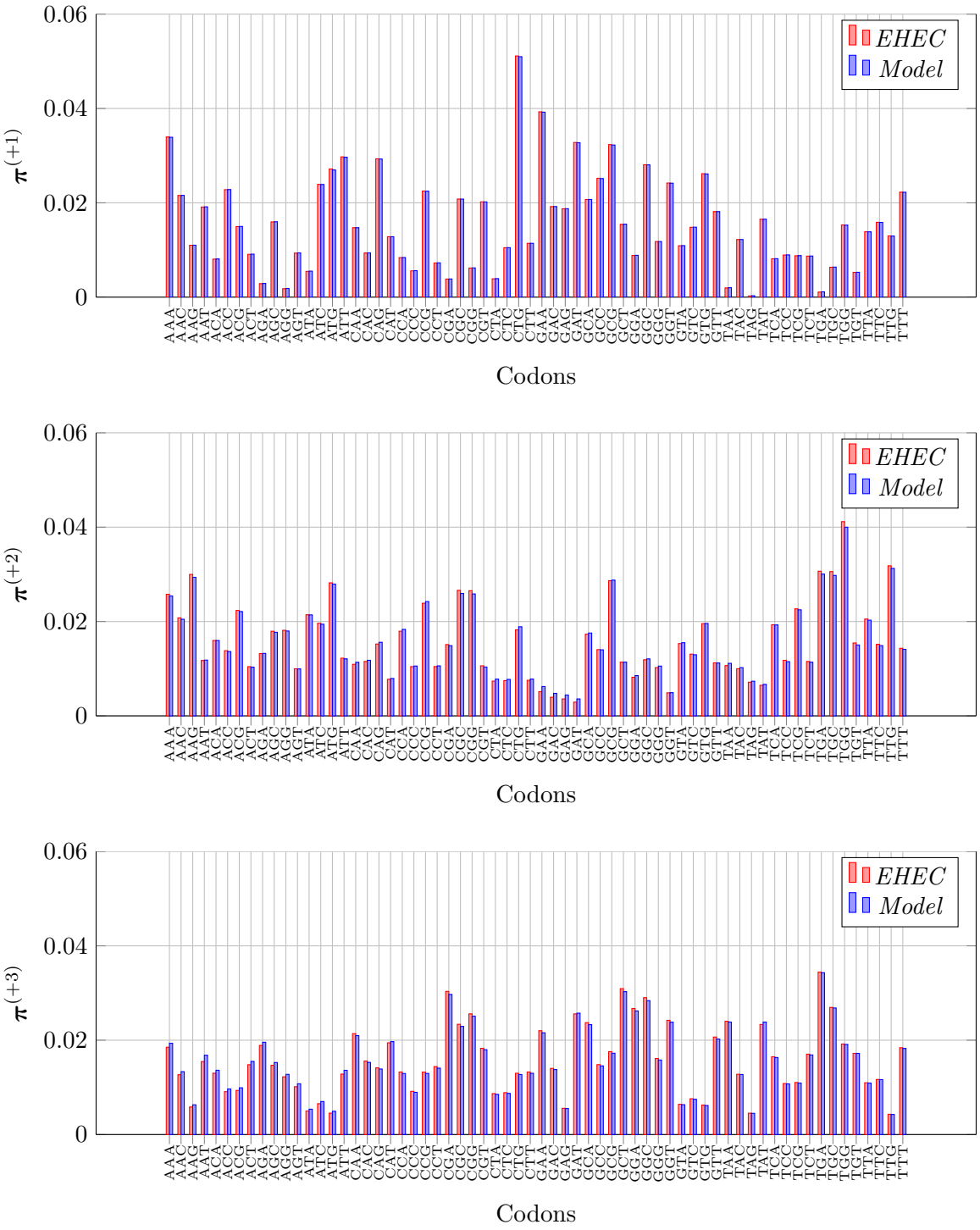


Figure 3.9.: Empirical codon usages on the forward strand of *EHEC* in comparison with the stationary distributions of the model prediction.

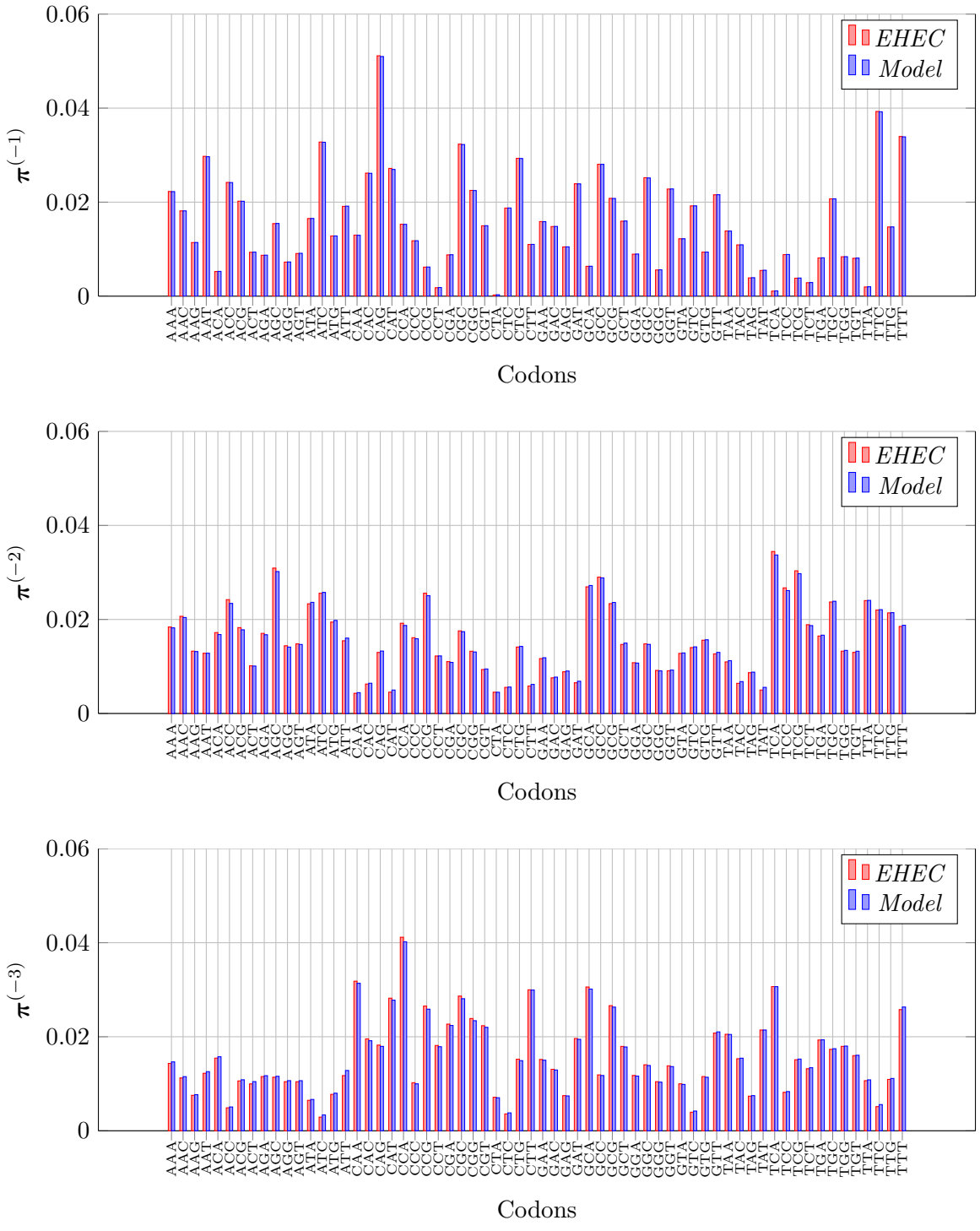


Figure 3.10.: Empirical codon usages on the reverse strand of *EHEC* in comparison with the stationary distributions of the model prediction.

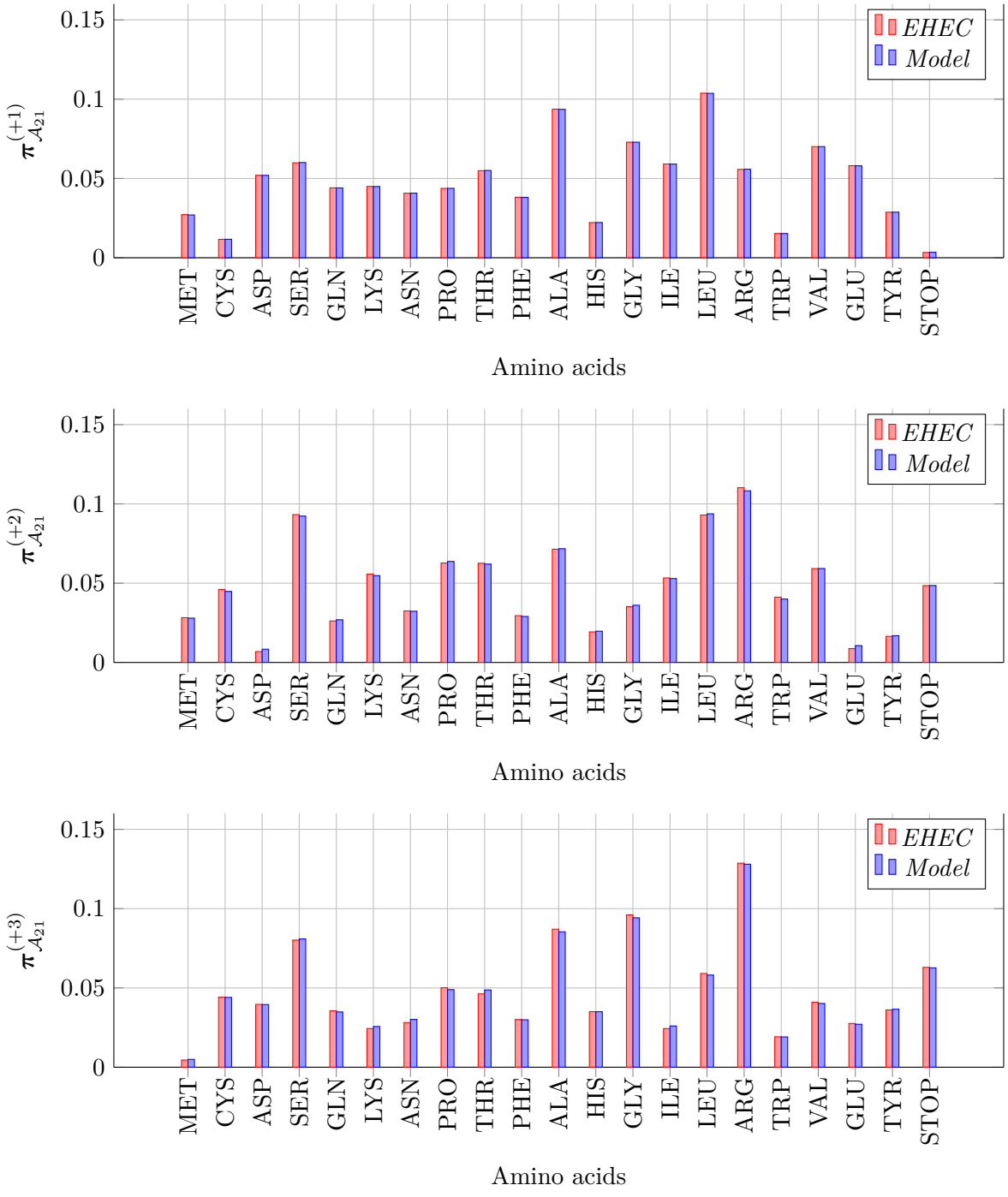


Figure 3.11.: Empirical amino acid distribution on the forward strand of *EHEC* in comparison with the stationary distributions of the model prediction.

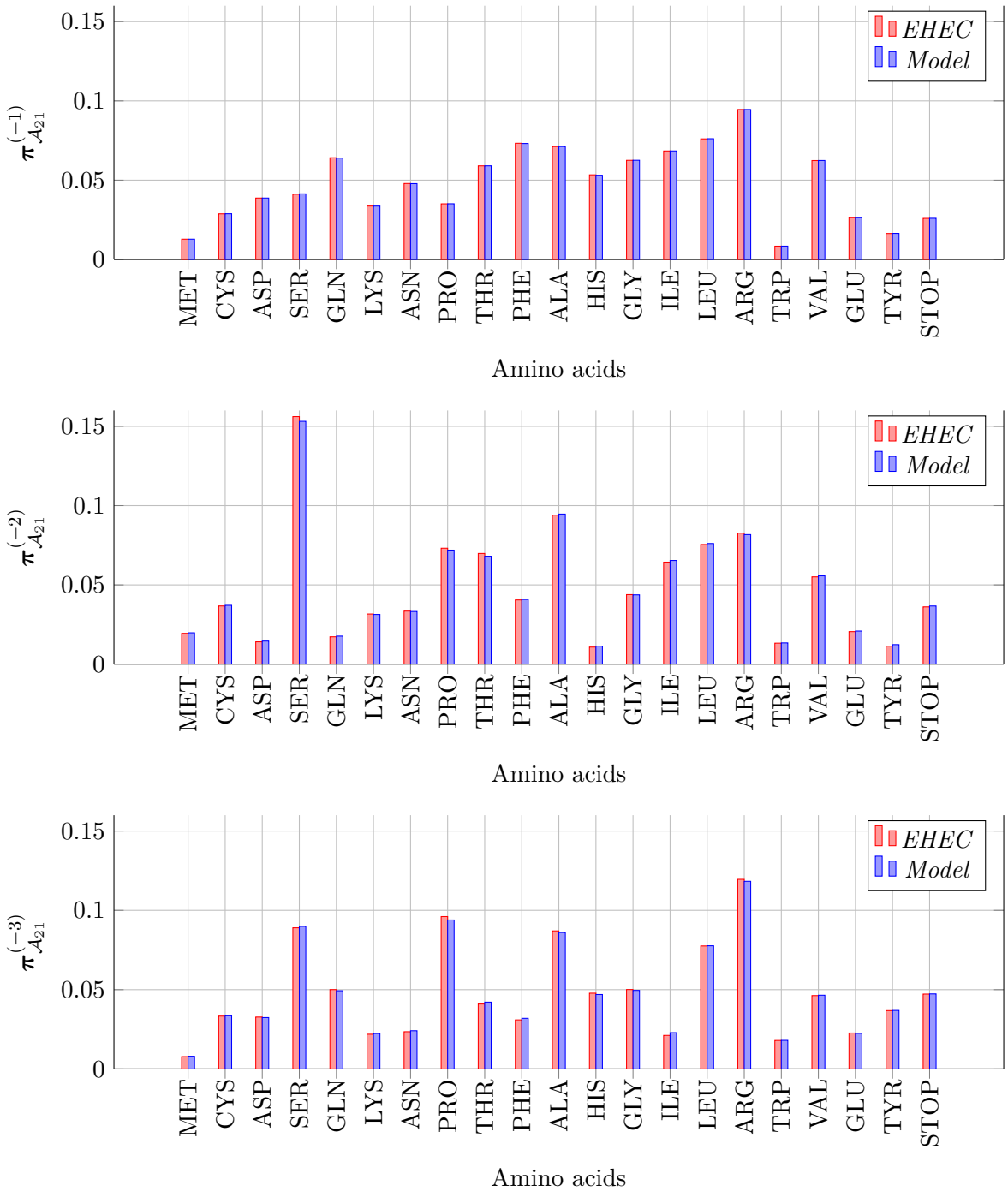


Figure 3.12.: Empirical amino acid distribution on the reverse strand of *EHEC* in comparison with the stationary distributions of the model prediction.

Frame	$\|\|Model - EHEC\|\|$	$D(EHEC\|\|Model) + D(Model\|\|EHEC)$
$+1$	$3.68 \cdot 10^{-4}$	$1.39 \cdot 10^{-5}$
$+2$	$4.33 \cdot 10^{-3}$	$1.32 \cdot 10^{-3}$
$+3$	$5.05 \cdot 10^{-3}$	$9.25 \cdot 10^{-4}$
-1	$4.15 \cdot 10^{-4}$	$5.30 \cdot 10^{-6}$
-2	$4.37 \cdot 10^{-3}$	$4.87 \cdot 10^{-4}$
-3	$3.98 \cdot 10^{-3}$	$5.24 \cdot 10^{-4}$
Frame	$\|\|RND - EHEC\|\|$	$D(EHEC\|\|RND) + D(RND\|\|EHEC)$
$+1$	$9.27 \cdot 10^{-2}$	$3.56 \cdot 10^{-1}$

Table 3.3.: Kullback-Leibler divergence and Euclidean distance of the amino acid distributions of *EHEC* and the predictions of the model. Random amino acid distribution (RND) based on the GC content of the organism.

Start- and Stop codon probabilities Given the transition matrix per reading frame $\boldsymbol{P}^{(f)}$, $f \in \{\pm 1, \pm 2 \pm 3\}$ with the corresponding stationary distribution $\boldsymbol{\pi}^{(f)}$ we determine the probability for a stop codon by

$$p_{stop}^{(f)} = \pi_{TAG}^{(f)} + \pi_{TAA}^{(f)} + \pi_{TGA}^{(f)}$$

and the probability for a start codon by

$$p_{start}^{(f)} = \pi_{ATG}^{(f)} + \pi_{GTG}^{(f)} + \pi_{TTG}^{(f)}.$$

In Table 3.4 we present a comparison with the corresponding probabilities for *EHEC* that can be obtained from the empirical codon distribution $\hat{\boldsymbol{\pi}}^{(f)}$ accordingly. There is a small deviation of $10^{-4} - 10^{-5}$ between the model prediction and the empirical probabilities, the absolute deviations are denoted as Δ_{start} and Δ_{stop} in the Table.

	p_{start}			p_{stop}		
Frame	*EHEC*	*Model*	Δ_{start}	*EHEC*	*Model*	Δ_{stop}
$+1$	0.0663	0.0661	$2.8 \cdot 10^{-4}$	0.0033	0.0035	$1.5 \cdot 10^{-4}$
$+2$	0.0796	0.0788	$8.0 \cdot 10^{-4}$	0.0485	0.0485	$3.2 \cdot 10^{-5}$
$+3$	0.0151	0.0154	$2.8 \cdot 10^{-4}$	0.0630	0.0627	$3.9 \cdot 10^{-4}$
-1	0.0369	0.0369	$4.3 \cdot 10^{-5}$	0.0259	0.0260	$7.6 \cdot 10^{-5}$
-2	0.0565	0.0569	$4.3 \cdot 10^{-4}$	0.0362	0.0367	$5.4 \cdot 10^{-4}$
-3	0.0302	0.0305	$2.9 \cdot 10^{-4}$	0.0472	0.0473	$7.5 \cdot 10^{-5}$

Table 3.4.: Comparison of start and stop codon probabilities of *EHEC* (empirical) with the model predictions. Additionally, the absolute deviation is presented.

Waiting Time From the stop codon probability, we determine the average waiting time to return to a stop codon (Concept 3.6), which is always larger than the real average ORF length expected. We compare it with the results of the natural stop codon probabilities of *EHEC*,

$$\hat{t}_M^{(f)} = \frac{1}{p_{stop}^{(f)}}.$$

For the absorption time defined in (3.14) we end up in a (61×1) vector $\boldsymbol{t}$ based on the absorbing Markov chain $\boldsymbol{P}_{abs}^{(f)}$, where each element t_c gives the expected number of steps before being absorbed starting in transient codon state c. It turns out that for each reading frame, all 61 transient states have approximately the same absorption time. Therefore we determine the arithmetic mean of all transient states

$$t_A = \frac{1}{61} \sum_{c \in \mathcal{C}_{61}} t_c \tag{3.23}$$

as well as the sample standard deviation

$$s_A = \sqrt{\frac{1}{60} \sum_{c \in \mathcal{C}_{61}} (t_c - t_A)^2}.$$

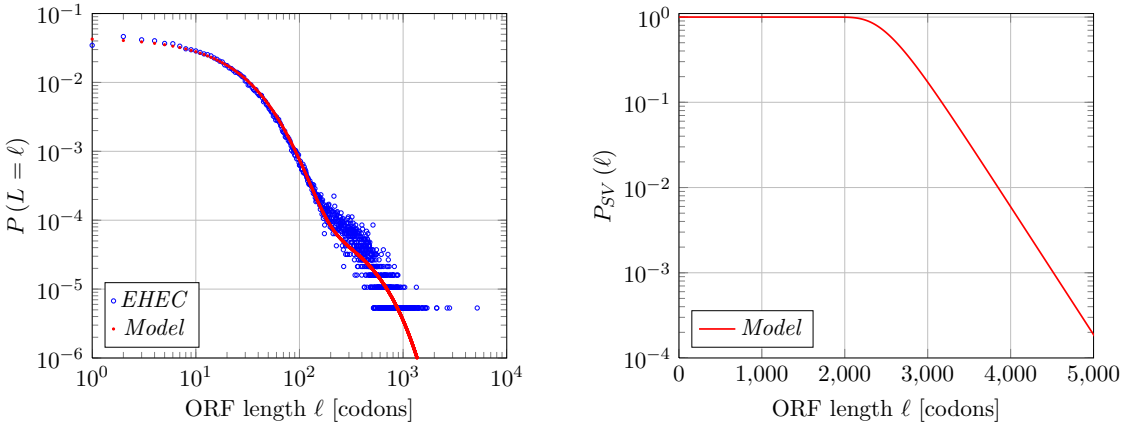
The results are presented in Table 3.5. For further investigations, we replace the Markov chain model by the one parameter Finite Mixture Model presented in Section 3.2.2. We show in the following paragraphs that the model gives insights in several interesting properties of the natural genome.

	EHEC	*Model*		
Frame	$\hat{t}_M$	t_M	t_A	s_A
+1	302.77	289.68	302.66	$7.6 \cdot 10^{-1}$
+2	20.62	20.61	20.48	$5.1 \cdot 10^{-1}$
+3	15.86	15.96	16.0	$3.8 \cdot 10^{-1}$
−1	38.64	38.53	39.31	$4.6 \cdot 10^{-1}$
−2	27.66	27.25	28.08	$4.7 \cdot 10^{-1}$
−3	21.17	21.14	21.67	$5.5 \cdot 10^{-1}$

Table 3.5.: Mean recurrence time t_M compared with mean absorbing time t_A over all transient states together with standard deviation s_A.

ORF length distribution and Survival Probability Knowing the start and stop codon probabilities in the reading frames $p_{start}^{(f)}$ and $p_{stop}^{(f)}$ respectively, as well as the length of the *EHEC* genome in codons N_C, we obtain the length distribution of ORFs applying Concept 3.7 and the survival probability using Concept 3.8. A comparison with the empirical ORF lengths of *EHEC* is presented in Figure 3.13(a) on page 56 together with the survival probability

Figure 3.13(b). For the empirical ORF length, we extract all start codons $\{ATG, GTG, TTG\}$ and stop codons $\{TAA, TAG, TGA\}$ and count the longest run from start to stop codon except an annotated ORF ends at the same stop codon position, in this case the annotated ORF length is counted.



(a) Probability distribution of ORF lengths.

(b) Survival probability.

Figure 3.13.: (a) Probability distribution of ORF lengths of *EHEC* in comparison with the prediction of the model. (b) Survival probability to observe at least one ORF with length $L \geq \ell$.

In Figure 3.13(a) we observe that the largest ORF in *EHEC* consists of 5188 codons, which corresponds in Figure 3.13(b) to a survival probability of $P_{SV}(5188) = 9.81 \cdot 10^{-5}$. But the survival probability for the second largest ORF with 2806 codons increased already to $P_{SV}(2806) = 0.310$. Summarized, the model cannot explain the exceptionally largest ORF in EHEC, but all others. This particular ORF is already annotated as a gene and belongs to a rare class of giant genes [RT08].

More generally we obtain the limits of the model by setting the significance value $\alpha = 10^{-2}$ using the survival bound (Concept 3.9) which results in $B_{SV} = 3850$ codons as smallest ORF length that can not be explained from the model.

To further assess the functionality of the model, we investigate some statistical parameters of interest that are important for the global description of the genome [Li99, MAG09, LSBGO98, OM96]. The number of ORFs predicted by the model is given by the sum over all reading frames applying (3.16)

$$n_{ORF} = \sum_{f \in \mathcal{F}} n_{ORF}^{(f)}.$$

The empirical ORF length of *EHEC* is simply counted from the observed ORF lengths $\mathbf{x} = (x_1 x_2 \dots x_{n_{ORF}})$. We determine the arithmetic mean of ORF lengths $\mathbf{x}$ in *EHEC* to

compare with the average ORF length of the finite model given in (3.19)

$$\bar{\ell}_{ORF} = \mathbf{E}\left[L\right] = \frac{1}{n_{ORF}} \sum_{i=1}^{n_{ORF}} x_i.$$

Other parameters of interest are quantiles introduced in Definition 2.11. We determine the CDF of the model derived in Appendix B.2

$$F_L(\ell) = \mathrm{P}\left(L \leq \ell\right) = \sum_{f \in \mathcal{F}} p^{(f)} \left(1 - \left(1 - p_{stop}^{(f)}\right)^{\ell}\right)$$

and precede to get the quantile of order r by searching for the smallest integer such that

$$Q(r) = \inf\{\ell \in \{1, 2, 3 \ldots\} : F_L(\ell) \geq r\} \qquad 0 \leq r \leq 1.$$

For *EHEC* we determine the EDF introduced in Definition 2.3 given $\mathbf{x}$

$$F_{n_{ORF}}(x) = \frac{1}{n_{ORF}} \sum_{i=1}^{n_{ORF}} \mathbf{1}\{x_i \leq x\}, \qquad x \in \{1, 2, 3 \ldots\}$$

and find

$$Q(r) = \inf\{x \in \{1, 2, 3 \ldots\} : F_{n_{ORF}}(x) \geq r\} \qquad 0 \leq r \leq 1,$$

accordingly. The parameters are summarized in Table 3.6. A quantile separates the given data into subsets. For the interpretation of quantiles, we choose $Q(0.75)$, where 75% of all ORFs are predicted to be shorter than 34 codons in the model, which is the same value for *EHEC*. In summary, prediction and empirical data show a close correspondence comparing the investigated statistical parameters.

Parameter	*EHEC*	*Model*
n_{ORF}	187981	187505.62
$\bar{\ell}_{ORF}$	32.34	32.96
$Q(0.25)$	7.0	6.0
$Q(0.50)$	16.0	16.0
$Q(0.75)$	34.0	34.0
$Q(0.99)$	337.0	339.0

Table 3.6.: Statistical parameters of interest to compare the ORF distribution predicted by the model with the empirical distribution of *EHEC*. Unit of length values is codons.

Application to Overlapping Genes

If the hypothesis that overlapping genes in prokaryotes exist is true, there might be protein coding ORFs in the alternative reading frames, which have not been detected yet. But how

does our approach help? Due to the construction of the model, where the reading frame $\{+1\}$ is based on the protein coding aORFs of *EHEC*, the alternative reading frames contain so called non coding or not annotated ORFs (naORFs). In this paragraph we evaluate each reading frame individually, i.e., we investigate the geometric distribution and the survival probability for each reading frame separately. Therefore we compare the results of *EHEC* with the prediction of the IID model presented in Section 3.1.1 and determine following parameters:

- The number of ORFs per reading frame $n_{ORF}^{(f)}$ predicted by the model is given by frame wise evaluation of (3.5) with individual start and stop probabilities $p_{start}^{(f)}$ and $p_{stop}^{(f)}$ for each reading frame $f \in \{\pm 1, \pm 2, \pm 3\}$.
- The average ORF length per reading frame $\bar{\ell}_{ORF}^{(f)}$ is simply the expected value of the geometric distribution (3.2) changing the stop codon probability with each reading frame.
- And the survival bound for significance value $\alpha = 10^{-2}$ is also determined from the IID survival probability (Concept 3.3) with individual number of trials $n_{ORF}^{(f)}$ and stop probabilities $p_{stop}^{(f)}$.

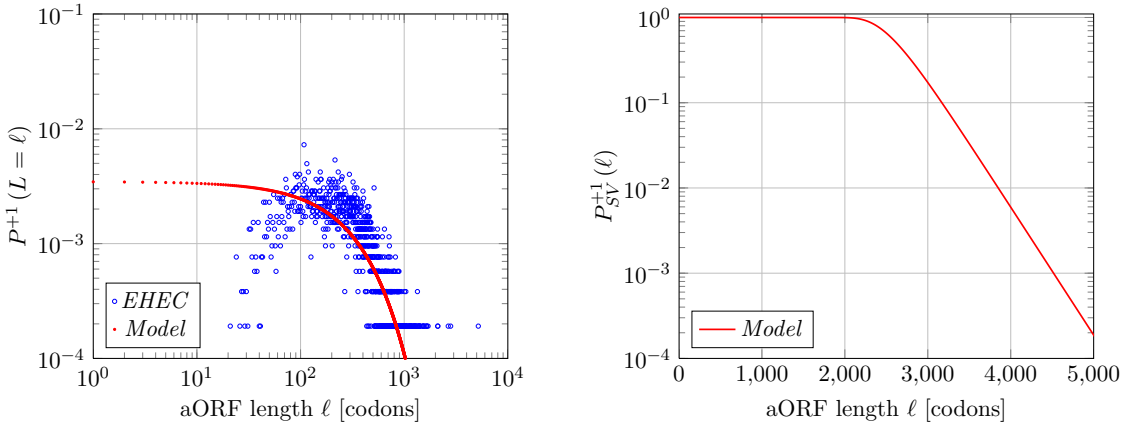
For separation of individual reading frames in *EHEC* we proceed as follows. An overlapping gene pair was defined as two protein coding sequences sharing the same DNA locus in different reading frames, i.e., all genes (aORFs) in *EHEC* that share some part of the genome with an naORF build potential overlapping gene pairs. As not all naORFs overlap with a gene, we call those that have an overlap shadow ORFs (sORFs). By definition all aORFs of *EHEC* are located in reading frame $\{+1\}$ and the corresponding frames of all sORFs are determined with respect to the protein coding frame. sORFs can be located in frames $\{-1, \pm 2, \pm 3\}$. Some statistical parameters are determined according to the previous paragraph, where we determined different parameters for all ORFs. The parameters are summarized in Table 3.7. The number of coding ORFs in reading frame $\{+1\}$ denoted

	Parameter	$+1$	$+2$	$+3$	-1	-2	-3
EHEC	n_{ORF}	5233	48643	19656	24454	35718	30142
Model	n_{ORF}	6045.55	55342.26	22729.93	28093.49	41114.85	34179.55
EHEC	$\bar{\ell}_{ORF}$	301.77	21.52	19.17	42.88	30.12	24.6
Model	$\bar{\ell}_{ORF}$	289.68	20.61	15.96	38.53	27.25	21.14
Model	B_{SV}	3850.0	314.0	228.0	566.0	409.0	312.0

Table 3.7.: Statistical parameters of interest for comparing the alternative reading frame distributions predicted by the model with the empirical distribution of *EHEC*. Unit of length values is codons.

with $n_{ORF}^{(+1)}$ is slightly over-estimated. The deviation can be explained from the fact that

the length distribution of the annotated genes in *EHEC* presented in Figure 3.14(a) does not follow a geometric distribution. The shortest gene in *EHEC* has a length of 21 codons,



(a) Probability distribution of aORF lengths.

(b) Survival probability for aORFs.

Figure 3.14.: (a) Probability distribution of aORF lengths of *EHEC* in comparison with the prediction of the model. (b) Survival probability to observe at least one aORF with length $L \geq \ell$.

whereas the geometric distribution takes all ORF lengths into account. Further, below a length of ~ 80 codons, less aORFs are annotated due to biological reasons [WAFS10], and additionally it is well known that current genome annotation programs rarely propose short genes [HPMR$^+$10]. We can get closer to the observed number, by adding the constraint that we are interested in ORF lengths $\geq t = 21$ codons only. Therefore the number of ORFs is multiplied with the probability to draw an ORF of length $L \geq t$ given in (3.7)

$$n_{ORF}^{(+1)} \mid_t = n_{ORF}^{(+1)} \cdot \left(1 - p_{stop}^{(+1)}\right)^{t-1},$$

which results in $n_{ORF}^{(+1)} \mid_{t=21} = 5641.57$. From Figure 3.14(b) and $B_{SV}^{(+1)}$ in Table 3.7 we see that the genes of *EHEC* completely determine the survival probability of all ORFs, as the largest ORFs observed in the organism are in fact genes.
Unfortunately the model can only predict the number and length distribution of not annotated ORFs in alternative reading frames and for *EHEC* we can only determine shadow ORFs per reading frame, therefore the model prediction of $n_{ORF}^{(f)}$ for $f \in \{-1, \pm 2, \pm 3\}$ is always larger than the observation of *EHEC*, as not all naORFs are sORFs.
If overlapping genes exits, we expect that the parameters of the model deviate from parameters of natural genomes. Random mutational events that lead to a minimal initial stage of beneficial protein expression, will be enforced by selection to prevent a decay back to the original state. The average ORF length $\bar{\ell}_{ORF}^{(f)}$ in each reading frame of the model is exactly the mean recurrence time t_M already shown in Table 3.5 on page 55. *EHEC* on the other

hand has larger ORF lengths than expected. In Figure 3.15(a) to Figure 3.15(e) on page 61 the curves of ORF lengths per reading frame are presented, the naORFs for the model deviate significantly from the sORFs of *EHEC* where the distribution has a longer tail.
Due to the different probabilities to obtain stop codons in the alternative frames the survival probabilities differ for each frame Figure 3.15(f). The survival bound $B_{SV}^{(f)}$ in Table 3.7 gives specific limits for each reading frame. For instance, a long ORF of 600 codons observed in reading frame $\{-1\}$ is more probable than one of the same size in reading frame $\{+2\}$. Especially for reading frame $\{-1\}$ the observation of long ORFs overlapping protein-coding sequences is not new and was already described, e.g., by Merino *et al.* [MBPB94]. The sorting of the survival bounds matches the results of the waiting time given in Table 3.5 on page 55. The question remains if evolutionary forces have shaped those long sORFs.
According to [LK07, CW07, JC04] there are significant differences between the frequency and length distribution of overlapping genes where the sORFs are located in reading frames $\{+2\}$ and $\{+3\}$. Short overlaps ($1-5$ base pairs) are rare in $\{+2\}$ and plenty in $\{+3\}$, whereas for long overlaps (≥ 7 base pairs) the reverse relation holds. In contrast to [CW07, Kra00] this extent of long overlaps is not reasoned by a selective constraint, but rather due to the start and stop codon frequencies in the coding regions, e.g., [FHP14, SGL08]. If we compare the length distributions in Figure 3.15(a) and Figure 3.15(b) we also observe a large extent of short overlaps in reading frames $\{+3\}$ compared with $\{+2\}$. On the other hand, the survival probability Figure 3.15(f) reveals that frame $\{+3\}$ has in general a smaller survival bound than $\{+2\}$, which leads accordingly to shorter naORFs and therefore shorter overlaps. In his paper Sabath [SGL08] examined same-strand overlapping genes in bacteria, he showed that differences between the frequencies can be explained by compositional factors as codon usage without invoking selection. Kingsford *et al.* [KDS07] investigated the opposite strand and explained the length distributions by reverse complementary stop codon frequencies. We follow their line of argumentation applying the survival probabilities that are based on start- and stop codon probabilities.

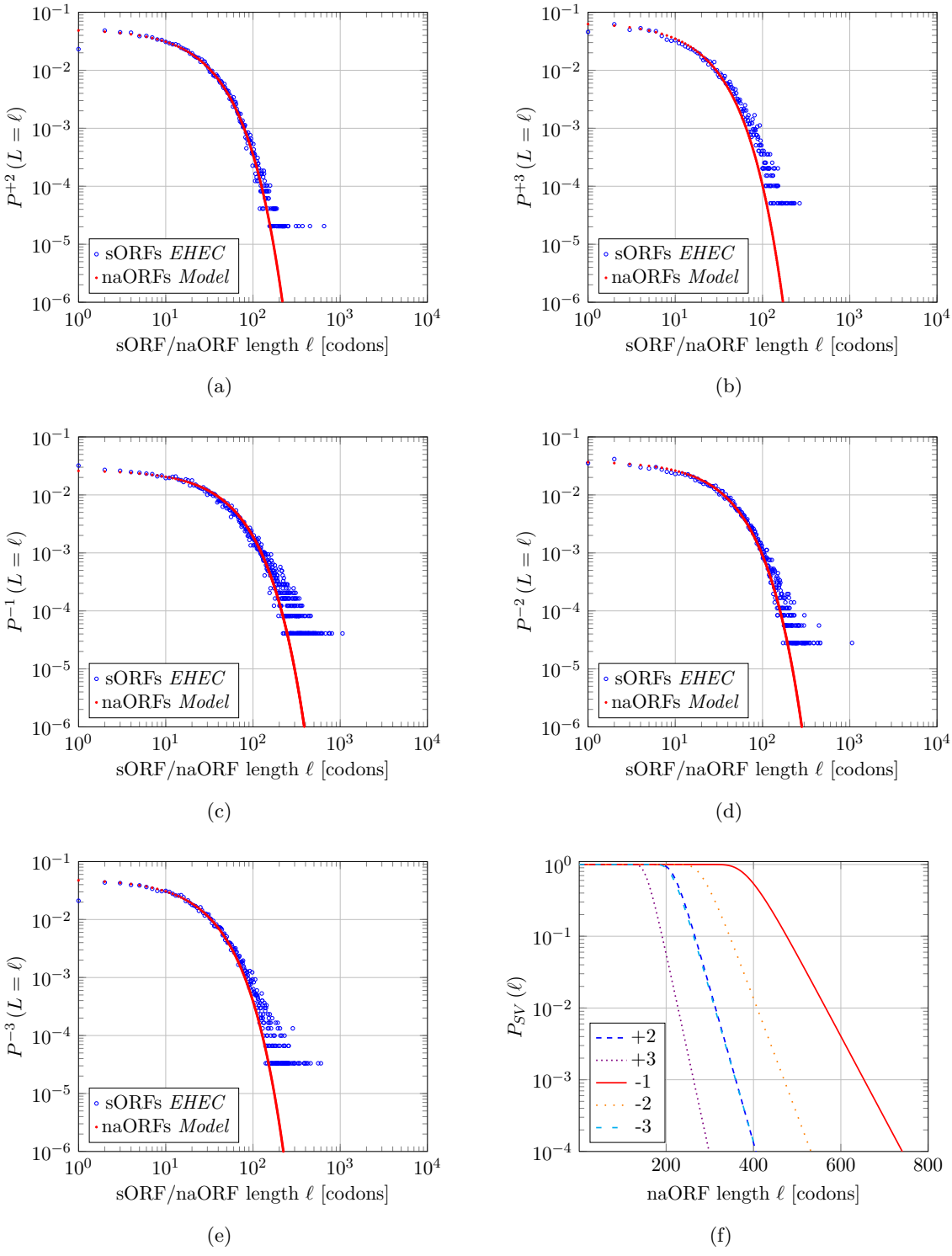
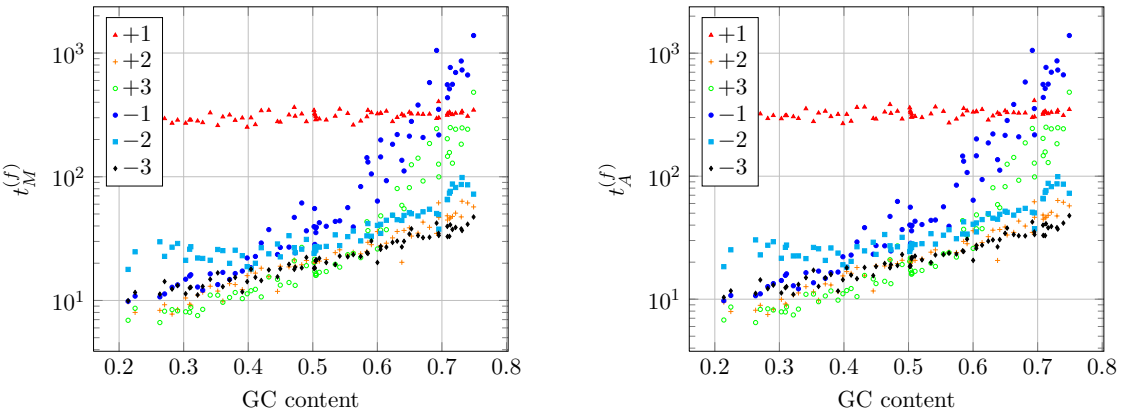


Figure 3.15.: (a)-(e) Probability distribution of sORF lengths of *EHEC* in comparison with the prediction of the naORF length of the model. (f) Survival probability for each alterantive reading frame.

3.3.2. Application to Multiple Organisms

We further evaluate the usability of the model by application to a total of 71 prokaryotes with GC contents ranging from 21.4% to 74.9%. The complete list of investigated species can be found in Appendix B.4. All parameters we investigate in the following paragraphs were already introduced in the previous section at the example of *EHEC* and are now extended to different organisms[2].

Waiting Time We compare the average waiting time from Concept 3.6 in Figure 3.16(a) with the arithmetic mean of the absorption time in Figure 3.16(b). As already mentioned in Section 3.2.2, all transient states have the same absorption time, which leads to similar distributions for both values.



(a) Mean recurrence time

(b) Mean absorption time

Figure 3.16.: The mean recurrence time (a) is the average waiting time to return to a stop codon. The arithmetic mean of absorption time over all transient states (b) is the time until an absorbing state is reached.

Based on the result, we again replace $\mathcal{MC}\left(\boldsymbol{\pi}^{(f)}, \boldsymbol{P}^{(f)}\right)$ by the Finite Mixture Model $\mathcal{MC}_0\left(\boldsymbol{\pi}^{(f)}\right)$.

ORF length distribution in All Reading Frames For the individual models we assess several characteristics that are important for the global description of the genomes [Li99, MAG09, LSBGO98, OM96]. For instance, we compare the average ORF lengths over different GC contents, ORF frequencies and the 75% ORF length quantiles of the model predictions with bacterial genomes. The probability distributions are compared using a Quantile-Quantile-Plot (QQ-Plot, [WG68]). The 75% quantile of the ORF lengths predicted by the model is compared with the natural genomes in Figure 3.17(a). The dashed straight

[2] All organisms where downloaded between 04/09/14 and 04/19/14 [oM14b].

line with slope one indicates not only correlation, but virtually identical distributions. When comparing natural genomes with the model, most genomes show an excellent correspondence, only genomes with a large GC content deviate to some extent. The particular genomes are labelled in Figure 3.17(a). As excellent correlation is found in Figure 3.17(b), when the number of ORFs predicted by the model is compared with the number of ORFs found in the bacterial genomes.

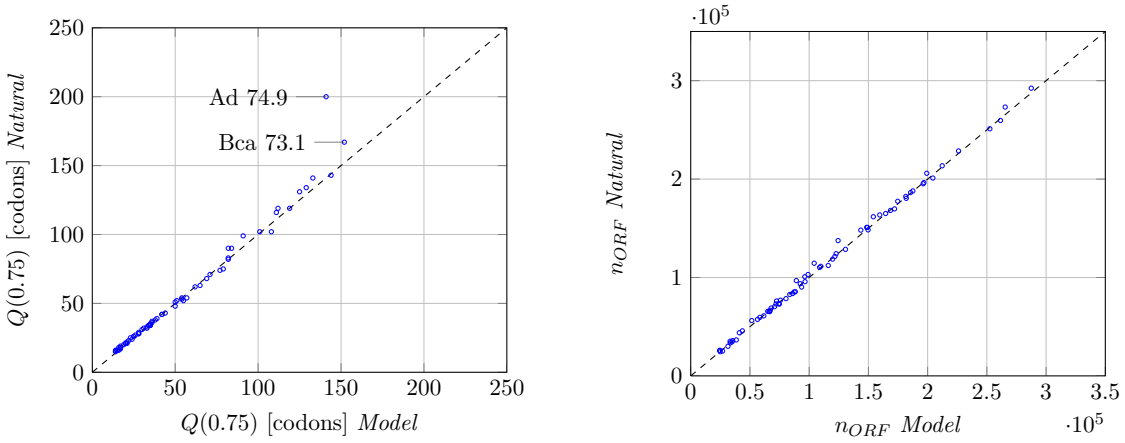
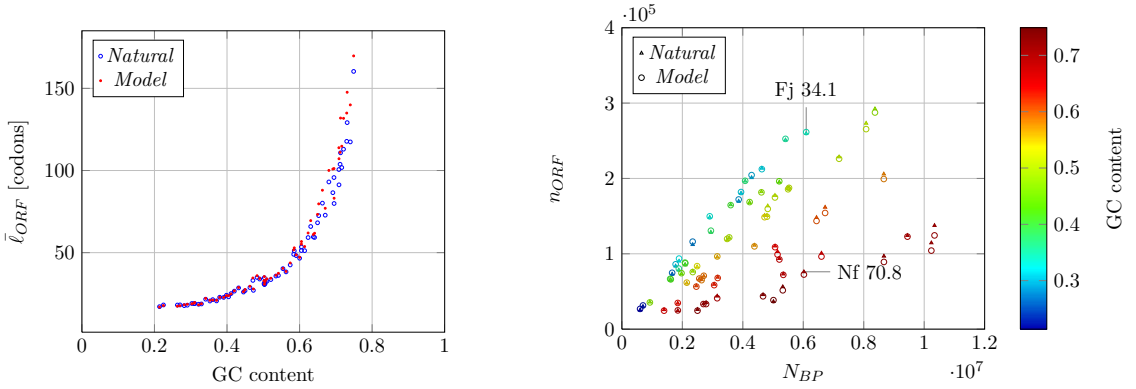


Figure 3.17.: (a) Quantile-Quantile Plot for comparison of 75% quantile of ORF lengths observed with prediction from the model. Individual data points are labeled with an abbreviated species name and GC content according to Appendix B.4. (b) Comparison of ORF numbers.

We examine the influence of the GC content on the predictions of the model in more detail, which is according to [OM96, Sil97] an important factor for the expected number of ORFs as well as the average ORF length. It is obvious that GC rich sequences have less stop codons, since stop codons $\{TAA, TAG, TGA\}$ are AT-rich and therefore increasingly rare [OM96]. This is also discussed in more detail later in Figure 3.23 on page 68. Decreasing probabilities for stop codons cause longer ORFs, which is observable in Figure 3.18(a) on page 64. The model reproduces this correlation, when calculating the expected ORF length following (3.19), which depends only on the start and stop codon probabilities and is independent of the sequence length, for sufficiently large sequences. From (3.17) together with (3.16) we know the number of ORFs expected for the model is influenced by the total length of a genome. Trivially, a larger genome can harbour more ORFs than a short one. In Figure 3.18(b), we obtain the correlation of ORF numbers increasing with the sequence length. When adding information about the GC content, the colour pattern emerges. This is due to the fact, that the number of ORFs further depends on the start- and stop codon probabilities, that are directly connected with the GC content, more precisely the codon usage considered. For example, if two organisms with similar genome lengths $N_{BP} \sim 6 \cdot 10^6$ base pairs are

compared, the number of ORFs expected for the organism with small GC content 0.34 is $n_{ORF} \sim 2.6 \cdot 10^5$, a magnitude higher than for an organism with GC 0.71 that has only $n_{ORF} \sim 7.6 \cdot 10^4$ ORFs. This effect was already shown in Figure 3.5.

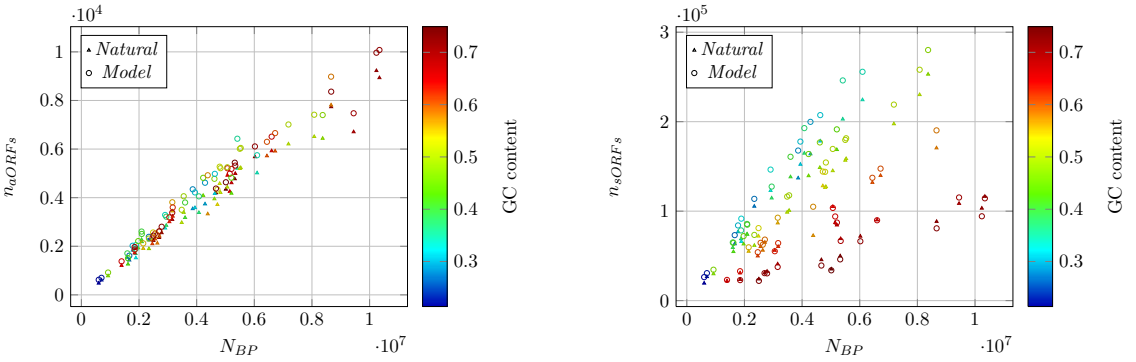


(a) Average ORF length.

(b) Number of ORFs.

Figure 3.18.: (a) Average ORF lengths over GC content predicted by the model compared to natural genomes. (b) Influence of GC content and sequence length on the number of ORFs. Individual data points are labeled with an abbreviated species name and GC content according to Appendix B.4.

It has been previously observed that the total gene number is also correlated with the genome size [MOM01, FNT03, LK07]. This behaviour is also visible in Figure 3.19(a). According to [FNT03, JC04] there is a robust relationship between the number of overlapping genes and the genome size, but for the sORFs/naORFs Figure 3.19(b) there is additionally the influence of the GC content observable as in Figure 3.18(b).

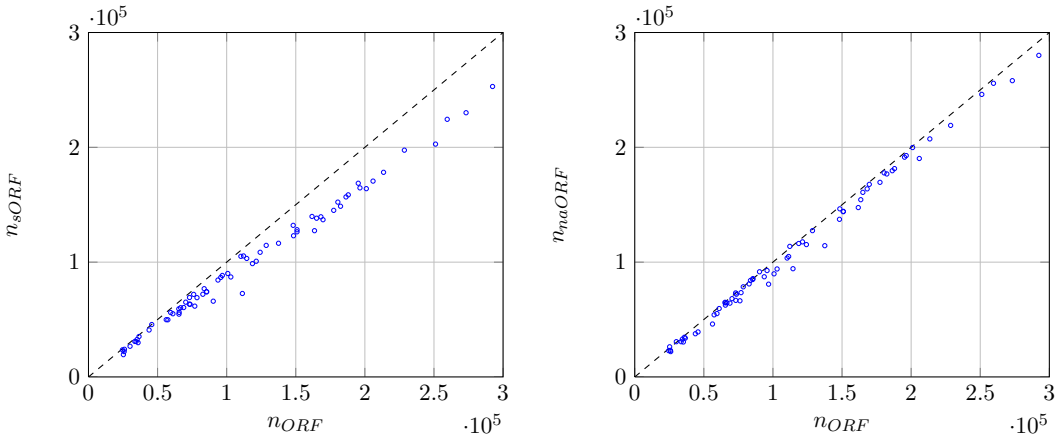


(a) Number of aORFs.

(b) Number of sORFs.

Figure 3.19.: Influence of GC content and sequence length on the number of aORFs (a) and sORFs of natural genomes compared with naORFs of the model prediction (b).

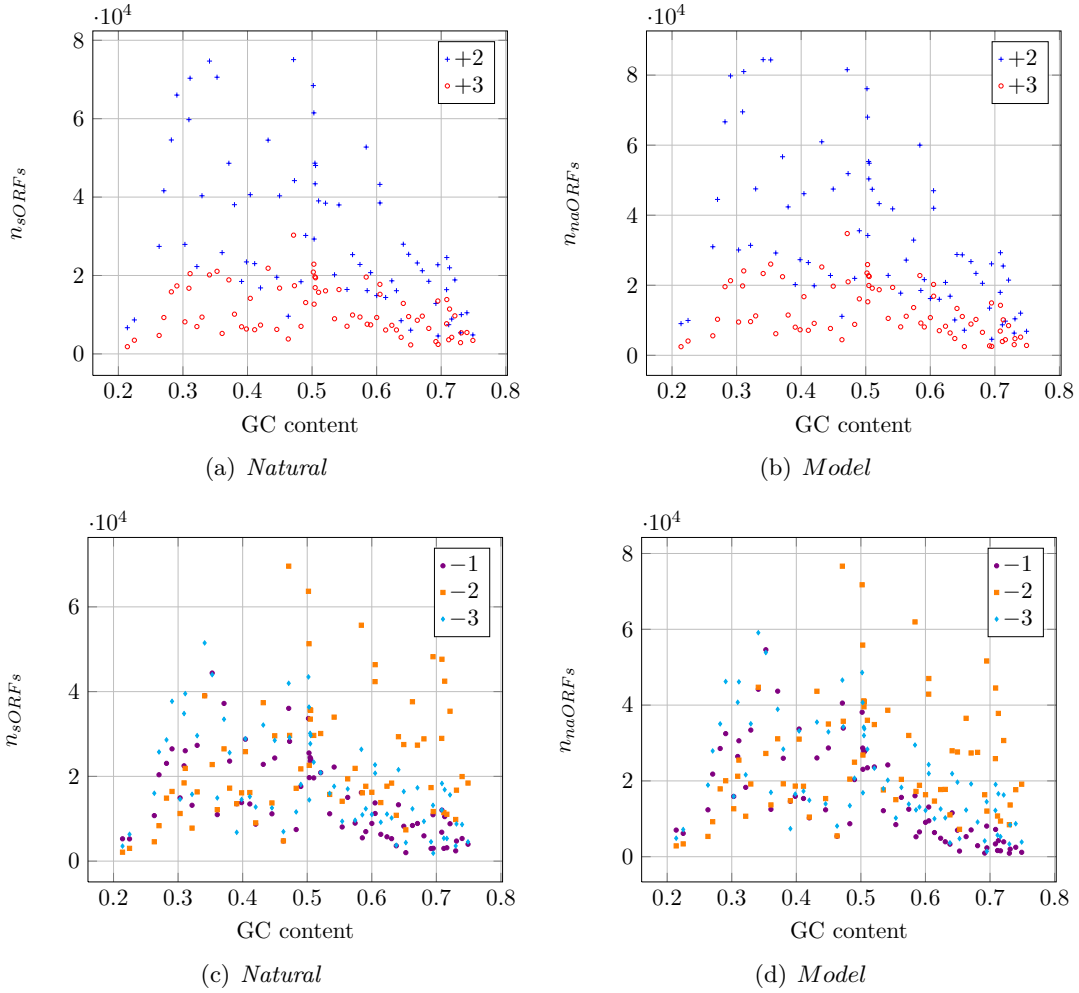
Further, there is a correlation between the total number of ORFs and the number of sORFs observed in natural genomes or naORFs predicted by the model (see Figure 3.20). A similar observation can be found in [FNT03, JC04], where [JC04] also showed that overlapping genes are not a result of misannotation.



(a) *Natural.* (b) *Model.*

Figure 3.20.: Correlation between the total number of ORFs and the number of sORFs (a) of the natural genomes and naORFs (b) predicted by the model.

Sabath *et al.* [SGL08] revealed that long overlaps in reading frame {+2} are negatively correlated with the genomic GC content, whereas long {+3} overlaps show no significant correlation. We repeated this analysis and compared the results with the predictions of the model in Figure 3.21 on page 66. Our predictions reproduce the overlap frequencies of Sabath, who analyzed 167 bacterial genomes. As we are not limited to an analysis of the forward strand, we can also determine the reverse strand. However, the distribution of the sORF and naORF numbers over the GC content is arbitrary and does not reveal further insights. A closer look at the results of Sabath reveals a negative correlation coefficient of $r = -0.28$ that is significant with $p < 0.001$. Unfortunately the significant value is not specified. We introduced those paramters in Section 2.1.2. For the interpretation, we calculate the coefficient of determination $r^2 = 0.0784$ that uncovers that only 7.8% of the total variation the overlap frequency can be explained by variation of the GC content. Further, application of such a correlation test is sensitive to the data distribution and is only applicable if the data are approximately normally distributed. Therefore, we resign to make a declaration regarding correlation for comparison with Sabath.



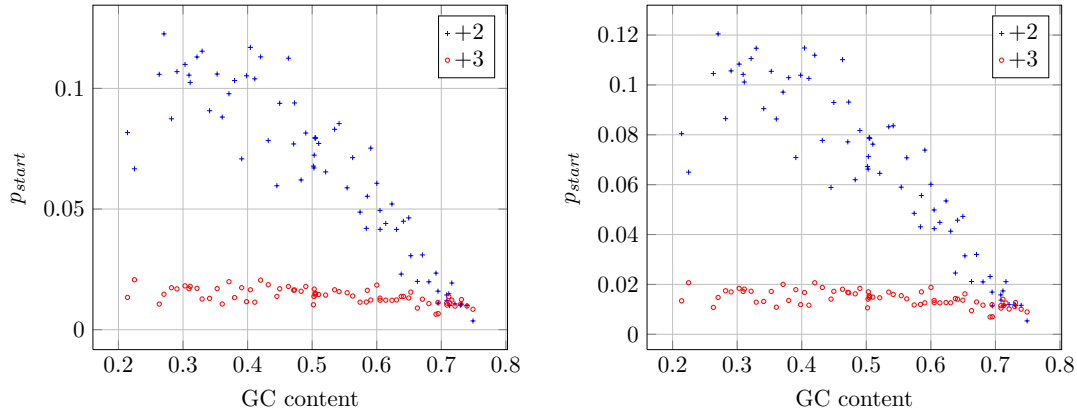
(a) *Natural* (b) *Model*

(c) *Natural* (d) *Model*

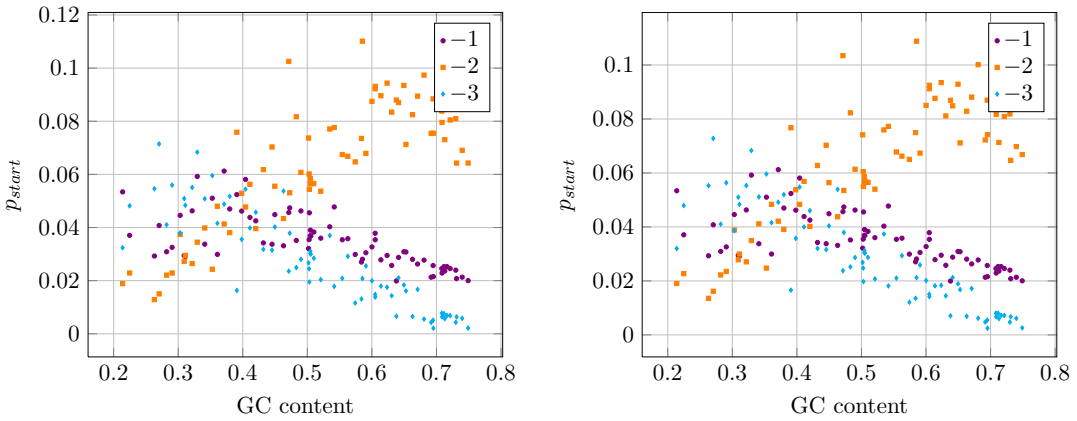
Figure 3.21.: Forward strand frequency of sORFs (a) and naORFs (b) against GC content of natural organisms. Reverse strand of sORFs (c) and naORFs (d).

Furthermore, Sabath investigated the start and stop codon frequencies on the forward strand. It turned out that the start codon probability in reading frame {+2} is higher than in frame {+3} over all GC content values. As the stop codon probabilities in both frames are similar to each other, he suggests the difference between the amount of long overlaps in both frames is primarily influenced by the frequencies of the start codons in both reading frames. This matches our findings of the survival probabilities. We repeated again the analysis of the start and stop codon probabilities for the forward (Figure 3.22) and the reverse strand (Figure 3.23 on page 68). Cock *et al.* [CW10] support [SGL08] investigating the evolution of same strand overlapping genes. In their study, they revealed that alternative start codons are more probable to reproduce the overlap distributions than omission of a stop codon by mutation, insertion or deletion of a nucleotide. Some studies contradict these results

[FWT99, FNT03, SSVC05, FHP14] suggesting that elongation of the stop codon is more probable. We do not mind which theory of evolution is more probable, as our model depends on both, the start- and stop codon frequencies. From our point of view, the fact that our model reproduces the distributions observed in the natural genome as well as results from literature is of interest.



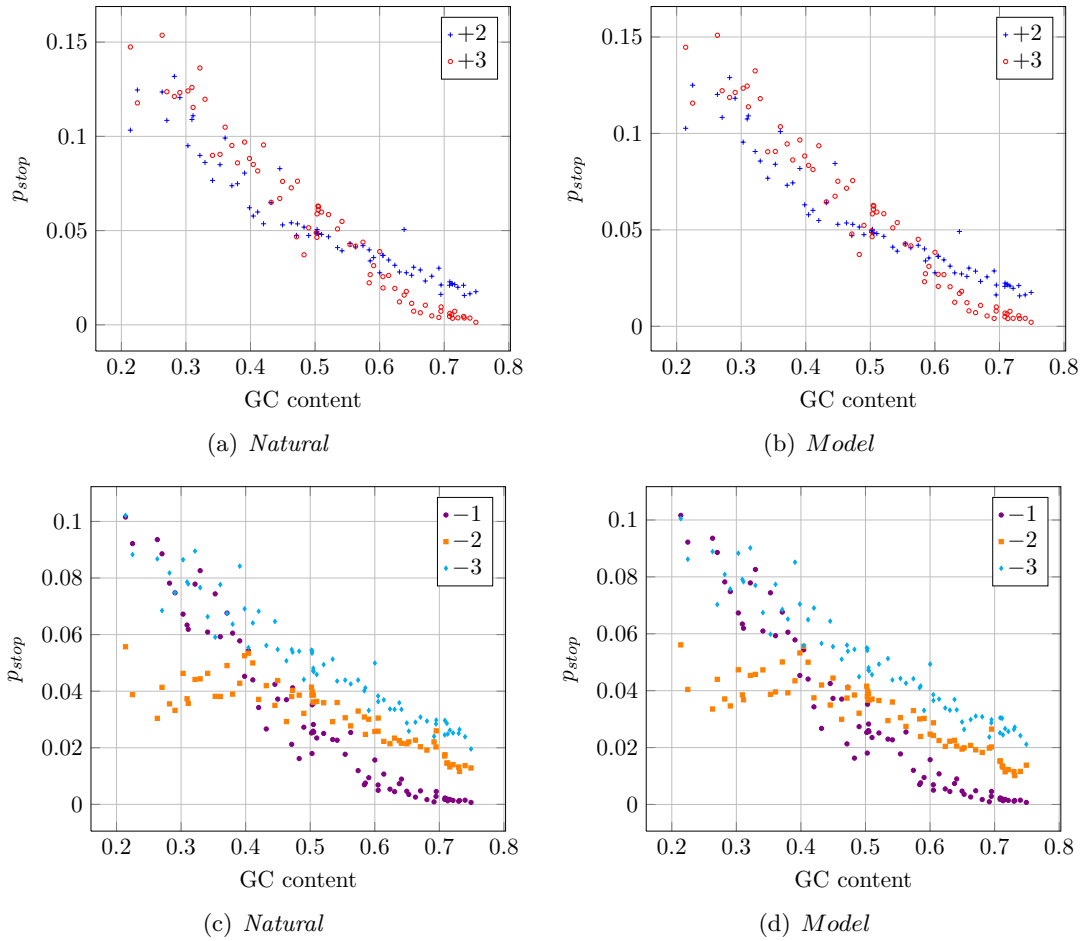
(a) *Natural* (b) *Model*



(c) *Natural* (d) *Model*

Figure 3.22.: Forward strand probability for start codon (a) and (b) against GC content of natural organisms. Reverse strand probability for start codon (c) and (d).

Robustness of predictions In this paragraph we discuss some aspects on the robustness of the models investigated. We show that the model is even applicable for *Mycoplasma mycoides* (NC_005364) an organism of 1211703 base pairs with a GC-content of 24% that belongs to a very peculiar group of bacteria with a parasitic life style [Raz06]. In contrast to most organisms investigated yet, the standard genetic code introduced in Section 2.4.2



(a) *Natural* (b) *Model*

(c) *Natural* (d) *Model*

Figure 3.23.: Forward strand probability for stop codon (a) and (b) against GC content of natural organisms. Reverse strand probability for stop codon (c) and (d).

is replaced by a modified translation table (table 4) that has only *TAA* and *TAG* as stop codons. The remaining stop codon *TGA* is read as amino acid Tryptophan instead. Adaption of the model to this species, results in following global parameters Table 3.8 and Table 3.9. Further information on results can be found in Appendix B.5.

Pallejà *et al.* [PHB08] analyzed genes in prokaryotic genomes with overlaps larger or equal to 60 base pairs. It turned out that most long overlaps investigated, e.g., in [FNT03, FWT99, JC04, SSVC05, CW07] where caused by artifical extensions of genes due to misannotations. However, they also identified some correct overlaps. They classified same strand overlapping genes into five categories of misannotations. Therefore 715 overlapping genes, where compared with their orthologs. Orthologs are homologue sequences found in different species. In the following enumeration, we recite the different categories [PHB08] and show how our model is

Parameter	*Mycoplasma*	*Model*
n_{ORF}	55570	59059.24
$\bar{\ell}_{ORF}$	19.52	18.51
$Q(0.25)$	4.0	3.0
$Q(0.50)$	9.0	7.0
$Q(0.75)$	17.0	16.0
$Q(0.99)$	263.898	212

Table 3.8.: Statistical parameters of interest to compare the ORF distribution predicted by the model with the empirical distribution of *Mycoplasma mycoides*. Unit of length values is codons.

	Parameter	+1	+2	+3	−1	−2	−3
Mycoplasma	n_{ORF}	1017.0	15890	4654	10235	3801	11773
Model	n_{ORF}	1174.42	19639.73	5773.16	12435.47	4890.11	15146.35
Mycoplasma	$\bar{\ell}_{ORF}$	324.14	11.1	10.81	12.79	36.93	11.86
Model	$\bar{\ell}_{ORF}$	302.47	11.12	9.15	11.82	34.06	10.13
Model	B_{SV}	3525.0	155.0	116.0	160.0	441.0	138.0

Table 3.9.: Statistical parameters of interest to compare the alternative reading frame distributions predicted by the model with the empirical distribution of *Mycoplasma mycoides*. Unit of length values is codons.

able to handle those errors.

A 57% of the downstream genes have wrong start codons compared to their orthologs. Their argumentation is mainly based on the fact that most genes tend to start with *ATG* which results in an overrepresentation of gene candidates, if weak start codons *GTG* and *TTG* are used for the analysis.
In our model, we use a weighing factor for each start codon, therefore we add the expected number of occurrences (3.21) during preparation of the transition matrix. For our analysis of the natural genomes this could have an influence, but in fact the results only change slightly, if we repeat some investigation for *EHEC* with single start codon *ATG* in Appendix B.6. Therefore, our model can also be applied to start codon *ATG* only and still gives observations close to natural genomes.

B 23% of the pairs could be mapped to a single ortholog gene. Frameshift mutation or sequencing error caused a fragmentation into the overlapping pair.
In our opinion this is not necessarily a contradiction to the existence of an overlapping gene pair. If the fragmentation happens by insertion or deletion of a nucleotide that leads to a frameshift, an overlapping gene pair evolved by a random mutation. For our argumentation line, a directed mutation is not inevitably needed. In fact, we show in

Chapter 4 that a random mutation can lead to two protein-coding sequences that are both selected during evolution.

C 9.5% of the upstream genes are extended to a wrong stop codon compared to the orthologs. This is either caused by a frameshift or point mutations at the real stop codon.
For this category we repeat the argument of category B: We do not distinguish if an overlapping gene pair evolved by a random mutation or a directed one.

D 0.5% of pairs are within the same reading frame having different start codons.
In our investigation of natural genomes we choose the longest run from start codon to stop codon, therefore we do not get an overlap of ORFs that both end at the same stop codon.

E 10% of pairs are a combination of category A & C. For our investigations the corresponding arguments still hold.

3.4. Optimality Investigations of Alternative Genetic Codes

The standard genetic code has a highly nonrandom structure, which is important for the robustness of the code. For an overview see, e.g., [KN09]. Results and discussions of the following parts were already published in Mir *et al.* [2]. Haig and Hurst [HH91] and Freeland [FH98] obtained estimates showing that the standard genetic code outperforms most random codes regarding mistranslation errors, this means that most codons that differ by only one letter are assigned to the correct amino acid or to chemically related ones. Itzkovitz and Alon [IA07] presented an empirical investigation on alternative genetic codes, studied their translational frame-shift errors and found that 99.3% of alternative codes aborted translation after such an error later than the standard code. We will follow this line of argumentation to compare different codes regarding the existence of long overlapping genes. Instead of an empirical investigation, we apply our model.

Alternative genetic codes We consider alternative genetic codes following Itzkovitz and Alon [IA07]. Therefore each codon position can be permuted individually in order to generate alternative codes keeping the structure of the standard genetic code implied by Figure 3.24. In detail, the number of codons per amino acid is preserved and the effect of misread errors on the translated protein is minimized. This is ensured as codons that differ by only one letter are mapped to the same or to chemically related amino acids. Additionally amino acids with a simple chemical structure have more codons assigned to them. The number of permutations is $(4!)^3$ which leads to 13824 possible codes, but as a permutation of T and C in the third codon position always results in the same code, the number of alternative codes is only 6912.

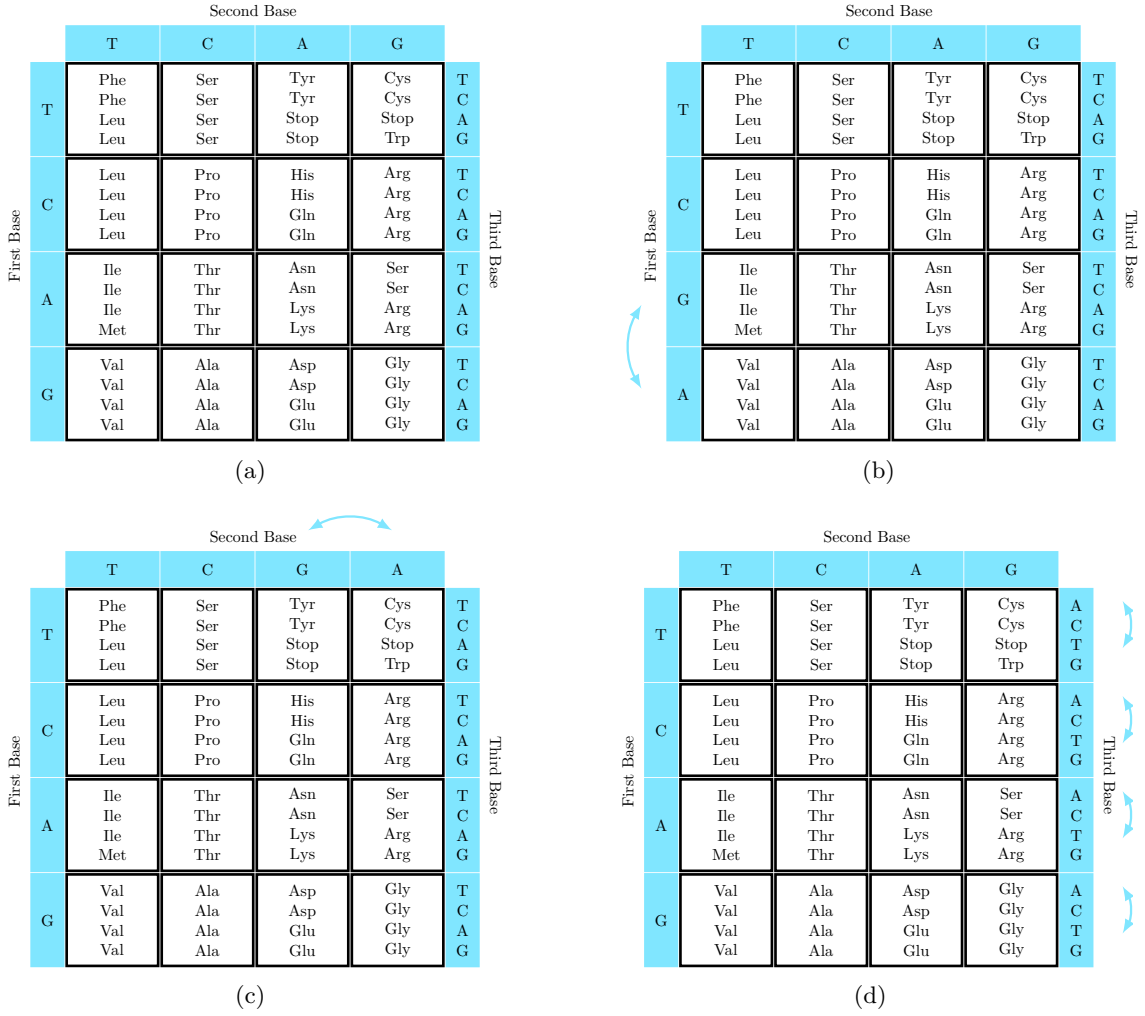


Figure 3.24.: Pattern of alternative genetic codes. (a) Standard genetic code. (b) Alternative genetic code obtained by permutation of A and G in first position. (c) Alternative genetic code obtained by permutation of A and G in second position. (d) Alternative genetic code obtained by permutation of A and T in third position. (modified version of [IA07]).

Input transition matrix In order to generate a transition matrix, that can be applied to different genetic codes, we determine the empirical amino acid distribution $\boldsymbol{\pi}_{\mathcal{A}_{21}}^{(+1)} = \{\pi_a^{(+1)}\}_{a\in\mathcal{A}_{21}}$ of *EHEC*. Each amino acid $a \in \mathcal{A}_{21}$, has an individual set of codons $\{c_1, c_2, \ldots, c_{|a|}\}$ based on some genetic code $\mathcal{G}$. We assume, that the codon frequency $f_C(c)$ within each amino acid is uniformly distributed

$$f_C(c) = \left(\frac{\pi_a^{(+1)}}{|a|}\right) \qquad \forall c \mapsto a \in \mathcal{A}_{21} = \mathcal{G}\left(\mathcal{C}_{64}\right).$$

Note, as in *EHEC* all nucleotides have approximately the same frequency of occurrence this assumption is sufficient. The transition matrix in reading frame $\{+1\}$ is determined from independent and identically distributed (IID) codons, where $\boldsymbol{P}^{IID} = \{P_{C_{j-1},C_j}\} = f_{C_j}(c_j)$. The generated transition matrix is then used as input to the model described in Section 3.2.1, which is used to determine the individual codon distributions per reading frame $\boldsymbol{\pi}^{(f)}$. We further assume that only one start codon exists, which is always the one assigned to amino acid Met independently of the the genetic code used as well as three stop codons.

Investigation of translational frame-shift errors Before we apply our model to investigate if the genetic code is optimal regarding the evolution of long sORFs, we repeat the investigation of Itzkovitz and Alon [IA07] regarding frame-shift translation errors. We show that our model reproduces their findings investigating a possible selection advantage of the standard genetic code over alternative genetic codes. First, we have to explain briefly the biological background leading to such an error. During the translation process, a new protein is synthesized while the ribosome slides codon wise along the mRNA sequence. In case a frame-shift error occurs, the ribosome changes the reading frame ignoring one or two nucleotides which leads to a nonsense protein sequence. A genetic code that aborts translation quickly after such a frame-shift error could have a selective advantage [SP04]. Clearly, a fast termination depends on the stop codon probability in the new reading frame that was entered during the frame-shift. In [IA07] they proceeded as follows: They determined all 61×61 di-codon combinations based on the amino acid frequency in proteins of 134 sequenced genomes [PFM+04]. Afterwards they obtained the stop codon probabilities after a one- or two base pair shift by summing all di-codons containing a stop codon at position $2-4$ and $3-5$ respectively. The average over both cases, results in their waiting time.
For comparison we use the codon usage $\boldsymbol{\pi}^{(f)}$. After a translational frame-shift error, we most likely end up in a transient state. The average waiting time to end up in a stop codon starting in a transient state, is the mean absorption time $t_A^{(f)}$ of (3.23). In Figure 3.25(a) on page 74 we determined for each of the 6912 alternative codes the average waiting time over both reading frames $\{+2, +3\}$,

$$\bar{t}_A = \frac{1}{2}\left(t_A^{(+2)} + t_A^{(+3)}\right).$$

Note, that a frame-shift error can only occur on the same strand. In Figure 3.25(a) as histogram of the average waiting time is shown. The standard genetic code (STD) has

$\bar{t}_A^{STD} = 16.25$ codons. Only 214 of 6912 alternative codes have an average absorption time shorter or equal to the standard genetic code, therefore the standard genetic code aborts translation faster than 96.9% of the alternative genetic codes considered. It is near-optimal regarding translational frame-shift errors. Our resulting figure is very close to [IA07, Supplementary Material Figure S1].

Investigation of long sORFs Itzkovitz and Alon [IA07] additionally found out that the standard code is near optimal for encoding additional information within a protein sequence. However, their work did not consider how the code behaves during simultaneously encoding of two proteins. We are interested in the question if the standard genetic code is optimized to support the existence of long sORFs. In reality, an ORF starts in a start codon and ends in a stop codon, but here we investigate somehow an upper bound on the ORF length by considering the distance between two stop codons. This is obviously the mean recurrence time $t_M^{(f)}$ introduced in Concept 3.6 based on $\boldsymbol{\pi}^{(f)}$, which is the average waiting time to reach a stop codon for the first time starting in one of the three stop codons. In Figure 3.25 on page 74 we show the mean recurrence time for each reading frame. The longer the recurrence time, the longer are potential sORFs in the investigated frame. Reading frames $\{+2, +3\}$ in Figure 3.25(b) and Figure 3.25(c) match the results of the translational frame-shift errors. For the negative strand, we obtain from Figure 3.25(d) to Figure 3.25(f), the largest mean recurrence time in the order $t_M^{(-1)} > t_M^{(-2)} > t_M^{(-3)}$, which matches the survival probabilities of Figure 3.15(f).
More detailed, in frame $\{-1\}$ comparing the standard genetic code with the alternative codes, 31.4% of all codes allow larger sORFs. There are 20.8% of alternative codes larger in frame $\{-2\}$ and 65.2% in reading frame $\{-3\}$. We reveal that the standard genetic code is not optimized regarding long sORF, as it is not located at the margins of any reading frame. If we compare both criteria of optimality, we determine that 384 codes are located at $t_M^{(-1)} \geq 80$ and 16 of them have $\bar{t}_A \leq \bar{t}_A^{STD}$ aborting translation earlier after a frame-shift error Figure 3.25(d). From Figure 3.25(e) we set $t_M^{(-2)} \geq 39$ as bound finding that this holds for 384 codes, where 24 of them have better absorption times than the standard code. Finally, from Figure 3.25(f) we use the gap at $t_M^{(-3)} \geq 33$ that holds for 192 alternative codes, where 8 abort translation earlier than the standard genetic code after a frame-shift error on the forward strand. We summarize, that if the length of sORFs in alternative frames would be a criterion for optimization, there are better codes than the standard genetic code, even under the condition that optimality of fast termination after a translational frame-shift error has to be conserved.

Investigating the Entropy of Alternative Genetic Codes In the context of evolution it is a well known fact, that a certain amino acid is conserved with highest probability if the second codon position is fixed, followed by codon position one and then three, e.g., Xia [Xia83]. This is reflected in their substitution rates. A closer look at the alternative genetic codes reveals the same connection. Assume we fix position $i \in \{1, 2, 3\}$ of a codon to one of four nucleotides $n_i \in \mathcal{B}$. What entropy is left about the chosen amino acid X? We determine for

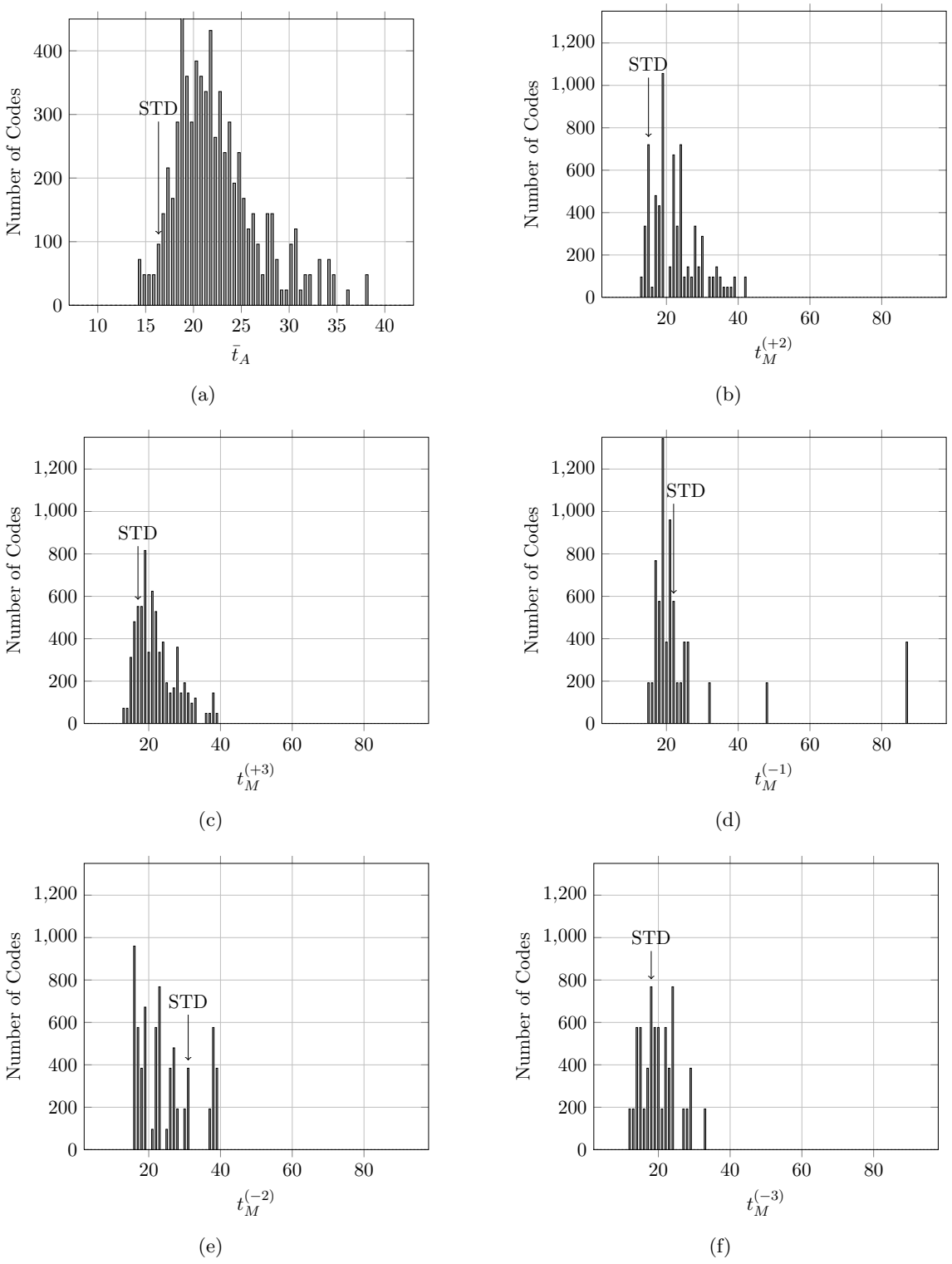


Figure 3.25.: (a) Impact of translational frame-shift errors on standard (STD) and alternative genetic codes. (b) to (f) Impact of different genetic codes on sORF length.

each amino acid of the alphabet $\mathcal{A}_{21}$ the conditional entropy Definition 2.13 introduced in Section 2.2.1

$$H(X|n_i) = -\sum_{a\in\mathcal{A}_{21}} P(X=a|n_i)\log_2 P(X=a|n_i).$$

For each codon position $i \in \{1,2,3\}$, we take the average over all four nucleotides

$$\bar{H}(X|i) = \frac{1}{4}\sum_{n_i=\mathcal{B}} H(X|n_i),$$

in order to get the conditional entropy per position and end up with

$$\bar{H}(X|2) = 2.23 < \bar{H}(X|1) = 2.44 < \bar{H}(X|3) = 3.71.$$

Interestingly all alternative genetic codes result in the same entropies, which is due to the fact that all of them share the block-wise structure of the standard genetic code.

3.5. Summary

In this chapter, we introduced a basic model, assuming that ORF lengths of IID nucleotide sequences follow a geometric distribution. However, for comparison with natural genomes this model is inappropriate. Therefore, we presented as more realistic model, a discrete-time Markov chain based on transition probabilities between codons of annotated genes. Further, we showed that the Markov chain model reveals virtually identical results as a model based solely on start and stop codon composition of annotated genes and the sequence length of a genome. Therefore, we could apply again geometric distributions, this time weighted according to six different reading frames. Finally, we presented encompassing simulation results. Therefore, we started with a detailed analysis of a single prokaryotic organism called *EHEC*, proceeded with a large scale analysis over a wide range of different bacteria and concluded with a more abstract analysis of alternative genetic codes applying our model predictions.
With our models, we are able to reproduce statistical properties of bacterial genomes without empirical curve fitting and showed excellent agreement of model predictions and empirical data of natural genomes. Statistical properties of ORFs are important in the context of overlapping genes. Can long ORFs overlapping annotated genes in prokaryotic genomes be explained from a random process? Using the developed model, we showed that the length distribution of sORFs deviates significantly from the observation in bacterial genomes. We observed a larger number of long sORFs than expected from statistical analysis. Those signals might have been eliminated by mutations if no selection pressure were stabilizing those long sORFs. However, those deviations call for functional explanations. We have shown that shorter sORFs can also be significant, depending on the reading frame where they appear. As already mentioned, most sORFs escaped discovery, as they were dismissed as false positives by genome annotation programs. Any observation indicating that genomes deviate from our model are a good starting point for further analysis of overlapping gene

candidates since these differences may have been shaped during evolution. Unfortunately, verification of protein-coding sequences is still a challenging and time consuming task. It turned out that the standard genetic code is not optimized to produce long sORFs in any alternative reading frame.

CHAPTER 4

Evolution Channels of Overlapping Genes

Molecular Evolution studies the impact of nucleotide substitutions on the DNA, RNA or protein level over time. Over the years, while the interest in the investigation of overlapping genes increased, their evolution moved into the focus of research. Several mechanisms explain the origins of overlaps [Kra00]: One reason, could be the reduction of the genome size conserving the number of encoded proteins [SSVC05, MY78]. Another, the generation of new proteins conserving the sequence length [MY78, SG00], also [KG92, Kra02, SS79] consider that overlaps have an important influence on the evolution of new genes. Further, functionally related genes could be jointly expressed or regulate the expression [NBE$^+$83, CSD$^+$98, ZSF$^+$02]. Evolution of overlapping coding regions is complicated, since a single mutation can affect two genes simultaneously [FHP14, Kra00].
Independent of the reason of the development of overlapping genes, it is of interest how they evolve. Mainly, the loss of a start or a stop codon within a protein coding sequence due to deletion, point mutation or a frame-shift leads to an elongation of the gene resulting in an overlap. For [CW07, CW10, SGL08] a mutation in the start codon is the main reason for an overlap in contrast to [FWT99, RSS$^+$02, SSVC05] who propagate a mutation in the stop codon. Our focus of interest regarding evolution is slightly different.
A fundamental task in molecular evolution, is to measure and analyze the intensity of selection acting on protein coding genes. Selection pressure is a force that influences the evolution of population. Investigating selection intensities in an overlapping gene pair is challenging, since both regions are interconnected [MY78, SGL08, HS95, PDIGP97] to a certain degree that depends on the type of overlap [Kra00].
This chapter is concerned with the question, how selective pressure within the protein coding reading frame $\{+1\}$ influences the alternative reading frames on both strands $\{-1, \pm 2, \pm 3\}$. Further, we are interested if it is possible to protect two protein coding sequences in different frames simultaneously? Therefore, we need to know, if selective pressure acting on an overlapping gene pair can be quantified? We explore those questions, applying a stochastic model within the protein coding reading frame, that is used to predict the consequent behaviour in each alternative reading frame.
Sequence evolution of protein coding genes can be described on nucleotide level, e.g., [JC69,

Kim80, Fel81], amino acid level, e.g., [DSO78, HH92], or on codon level, e.g., [GY94, MG94]. Within this chapter, we chose the latter approach assuming a continuous-time Markov process as suggested by Yang and Nielsen [YN00] which is based on [GY94]. An extended model was already used by Sabath *et al.* [SGL08], who developed a framework for analysis of overlapping genes to study the evolution of a random protein-coding sequence. In contrast to our approach, Sabath investigated the selection intensities of overlapping genes assuming that each gene of the overlapping pair faces selection independently, which leads to a model with many parameters. Similarly impractical is the model for the substitution process of sequences presented by Pedersen *et al.* [PJ01]. Several studies analyzed selection intensities in virus genomes within overlapping gene regions investigating how nonsynonymous and synonymous mutations influence two reading frames simultaneously showing that a high rate of nonsynonymous mutations in one reading frame falls onto synonymous substitutions in an alternative frame at the same time, e.g. [GD02, HWdS$^+$01, HH05, Pav06]. An empirical study of selection pressure in overlapping genes of prokaryotic organisms was published, e.g., by [RSS$^+$02].
We hypothesize that selection pressure in the protein-coding reading frame $\{+1\}$ is correlated with the reverse reading frame $\{-2\}$, where in fact many examples of overlapping genes found so far are located, e.g., [TBCM09, FBW$^+$14]. Precisely, we observe a strong coupling of substitution rates in these frames. If the evolution process is regarded as a communication process of protein synthesis over time, where anything that changes the expected result is summarized as noise, information theoretic measures can be applied to quantify the amount of information that can be transmitted over the evolution channel [Yoc92].

First, in Section 4.1 we introduce the evolution channel of codon substitutions in the protein coding reading frames and derive the corresponding channels that arise within the alternative reading frames. Afterwards, the selection pressure is investigated in Section 4.2 presenting estimation methods from literature and their application within our framework. Further the consequences within the alternative reading frames are discussed. In Section 4.3 information theoretic measures that quantify the information that can be transmitted over an evolutionary channel are presented and discussed. Each section contains simulation results of the model organism *EHEC*. Investigation of the genetic code in Section 4.4 reveals an explanation of some findings presented in the previous sections. Finally, Section 4.5 concludes the chapter. Some concepts applied in this chapter are well-known from literature, for a detailed introduction to the topic *Modelling of Molecular Evolution* see Part I from Yang [Yan06]. Results obtained in this chapter are in parts already published in [3].

4.1. Information Transmission over Time

Sequence evolution is modelled by a continuous-time Markov process, where each site evolves independently of each other without memory, such that the future of evolution depends only on the actual state of the chain. To gain knowledge about molecular evolution, we have to assume homogeneity, stationarity and reversibility of the considered process, although these

assumptions are not completely appropriate [LG98]. However, this procedure is state-of-the art. A review of rate and transition matrices used to describe sequence evolution can be found in, e.g., [LG98, Yan06, MA09]. Commonly used Markov models based on nucleotide substitutions with different constraints on substitution rates are, e.g., JC69 [JC69], K80 [Kim80], F81 [Fel81], HKY85 [HKY85, HYK84], TN93 [TN93], GTR [Tav86, Yan94, Zha94], UNREST [Yan94] and further [Bla85, TK81, GIN82, Fel95]. More biological reality is introduced in the codon based models of [GY94, MG94, YNH98, PWC98], where the influence of nucleotide substitutions on the encoded amino acid sequence is taken into account using the structure of the genetic code. Further Markov models of amino acid replacements are studied, but in contrast to the nucleotide and codon approaches above, amino acid models are empirical only, e.g., PAM [DSO78], BLOSUM [HH92] and further [JTT92, CAJ$^+$94, AH96, WG01]. Codon based models are our method of choice, as they combine point mutations on the nucleotide level with selective pressure operating on the protein level [DSS09].
In Section 4.1.1 a continuous-time Markov process of codon substitutions is introduced, that is used to model an evolution channel. Section 4.1.2 shows an extension and adaption of the model to alternative reading frames. Finally, in Section 4.1.3 we present some simulation results at the example of the *EHEC* genome to analyze our model.

4.1.1. Evolution Channel of Codon Substitutions

We use a continuous-time Markov process $\{C_t\}_{t\geq 0}$ introduced in Section 2.3.2 to describe codon evolution within protein-coding DNA sequences. Two well-known models of codon evolution were proposed by Goldman & Yang [GY94] and Muse & Gaut [MG94]. The rate matrix of the model considered here is a simplified version of the Goldman & Yang model, introduced by [YN00], where the following definitions can be found. The concepts and definitions presented here, are published in [Yan06]. The nucleotide alphabet is denoted with $\mathcal{B} = \{A, C, G, T\}$. The state space of the considered Markov chain is the codon alphabet without stop codons $\mathcal{C}_{61} = \mathcal{B}^3 \setminus \{TAG, TGA, TAA\}$, as transitions to stop codons are not allowed inside functional proteins and therefore not considered in the chain. The Markov chain is characterized by the substitution rate matrix $\boldsymbol{Q} = \{Q_{xy}\}_{x,y\in\mathcal{C}_{61}}$, where Q_{xy} is the instantaneous rate of change from codon x to codon y

$$\mathrm{P}\left(X_{t+\Delta t} = y | X_t = x\right) = Q_{xy}\Delta t \quad \forall\, x, y \in \mathcal{C}_{61}, \quad x \neq y.$$

We assume a time-homogeneous process (Definition 2.34). The rates are specified by

$$Q_{xy} = \begin{cases} 0 & \text{if } x \text{ and } y \text{ differ at more than one position,} \\ \pi_y & \text{if } x \text{ and } y \text{ differ by a synonymous transversion,} \\ \kappa\pi_y & \text{if } x \text{ and } y \text{ differ by a synonymous transition,} \\ \omega\pi_y & \text{if } x \text{ and } y \text{ differ by a non-synonymous transversion,} \\ \kappa\omega\pi_y & \text{if } x \text{ and } y \text{ differ by a non-synonymous transition,} \end{cases} \tag{4.1}$$

where κ is the transition/transversion rate ratio, ω is the nonsynonymous/synonymous rate ratio and π_y is the equilibrium frequency of codon $y \in \mathcal{C}_{61}$. We assume independent

occurrence of mutations at each codon position such that simultaneous changes at more than one position are neglectable. Interchanges of nucleotides $A \leftrightarrow G$ or $C \leftrightarrow T$ are called *transitions* and all remaining nucleotide substitutions are *transversions*. For a detailed explanation of transitions and transversions see Section 2.4.2 especially Figure 2.12 on page 31. Within the chosen model synonymous substitutions are distinguished from nonsyonymous substitutions. A nucleotide or codon replacement is called *synonymous* if the corresponding amino acid is not affected, whereas a *nonsynonymous* substitution changes the encoded amino acid [MY80]. The advantage of evolution models working on the protein level is that synonymous and nonsynonymous substitutions rates are selected differently such that their comparison provides understanding of natural selection. A detailed analysis of these parameters follows in Section 4.2.
Note, that the substitution rate matrix has all properties of Definition 2.35, e.g., we know the row sums of the rate matrix $\boldsymbol{Q}$ have to be zero, which determines the main diagonal of the matrix to

$$Q_{xx} = -\sum_{y:y\neq x} Q_{xy},$$

which is the substitution rate at which the Markov chain leaves codon x. The expected number of nucleotide substitutions per codon is given by

$$\lambda = -\sum_{x} \pi_x Q_{xx}$$

and therefore, the distance between two sequences after time t is $d = \lambda \cdot t$. The rate matrix is commonly multiplied by a scaling factor in order to normalize the average substitution rate λ to one, having the advantage that now time and distance are the same.
We also know from Section 2.3.2 that a transition matrix over any time $t > 0$ can be associated with every continuous-time Markov process. The evolution transition matrix $\boldsymbol{P}(t)$ has all properties of Definition 2.33 and is determined for state space $\mathcal{C}_{61}$ according to (2.14) by

$$\boldsymbol{P}(t) = \{P_{xy}(t)\}_{x,y\in\mathcal{C}_{61}} = e^{\boldsymbol{Q}t} = \boldsymbol{U}\mathrm{diag}\{e^{\lambda_1 t}, e^{\lambda_2 t}, \ldots, e^{\lambda_{|\mathcal{C}_{61}|} t}\}\boldsymbol{U}^{-1}, \tag{4.2}$$

where λ_i are the eigenvalues of $\boldsymbol{Q}$ and columns of $\boldsymbol{U}$ are the right eigenvectors of $\boldsymbol{Q}$. This connection holds, if the rate matrix can be decomposed. The CMTC is time-reversible as given in Concept 2.3. $P_{xy}(t)$ is the probability that input codon x has become y after time t. Starting from the initial condition at time $t = 0$ with $\boldsymbol{P}(0) = \boldsymbol{I}$, the diagonal elements $P_{yy}(t)$ decrease from one to π_y as $t \to \infty$, while $P_{xy}(t)$ increases from zero to π_y irrespectively of starting codon x.
Note, that the stationary distribution $\boldsymbol{\pi} = \{\pi_i\}_{i\in\mathcal{C}_{61}}$ can be extracted according to Concept 2.2 from the evolution rate matrix or the evolution transition matrix solving

$$\boldsymbol{\pi} = \boldsymbol{\pi}\boldsymbol{P}(t) \tag{4.3}$$

or

$$\boldsymbol{\pi}\boldsymbol{Q} = 0. \tag{4.4}$$

If the chain is stationary, the total amount of flow into a state equals the total amount out of the state.

Up to now, we repeated well-known concepts. Now, we consider an evolutionary framework. In the protein-coding reading frame $\{+1\}$ choose arbitrary values for the parameters $\kappa^{(+1)}$, $\omega^{(+1)}$ and $\boldsymbol{\pi}^{(+1)}$ to define the rate matrix $\boldsymbol{Q}^{(+1)}$ applying (4.1). We assume that codons are transmitted independently over the so called evolution channel, referring to the communication theoretic term [Sha48]. The evolution channel for a fixed and arbitrarily chosen parameter $t > 0$ is defined by the evolution transition matrix $\boldsymbol{P}^{(+1)}(t)$ solving the matrix exponential presented above or in (2.14).

Concept 4.1 (Evolution Channel). We call a DMC $\left(\mathcal{X}, \boldsymbol{P}^{(f)}(t), \mathcal{Y}\right)$ an *Evolution Channel* in reading frame $f \in \{\pm 1, \pm 2, \pm 3\}$, consisting of input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$, where $\boldsymbol{P}^{(f)}(t)$ specifies the conditional probabilities $p_{Y|X}(y|x)$ at time t to observe $y \in \mathcal{Y}$ given that $x \in \mathcal{X}$ was sent. □

Details on a DMC were given in Definition 2.16. Note that the matrix $\boldsymbol{Q}^{(+1)}$ is also a channel matrix. Each codon in reading frame $\{+1\}$ consists of three nucleotides $C_n = \left(X_n^{(1)}, X_n^{(2)}, X_n^{(3)}\right)$ and evolves to codon $\tilde{C}_n = \left(Y_n^{(1)}, Y_n^{(2)}, Y_n^{(3)}\right)$ resulting in the scheme presented in Figure 4.1 for two consecutive codons C_{j-1}^{+1} and C_j^{+1}, that are transmitted independently over evolution channel $\left(\mathcal{C}_{61}, \boldsymbol{P}^{(+1)}(t), \mathcal{C}_{61}\right)$.

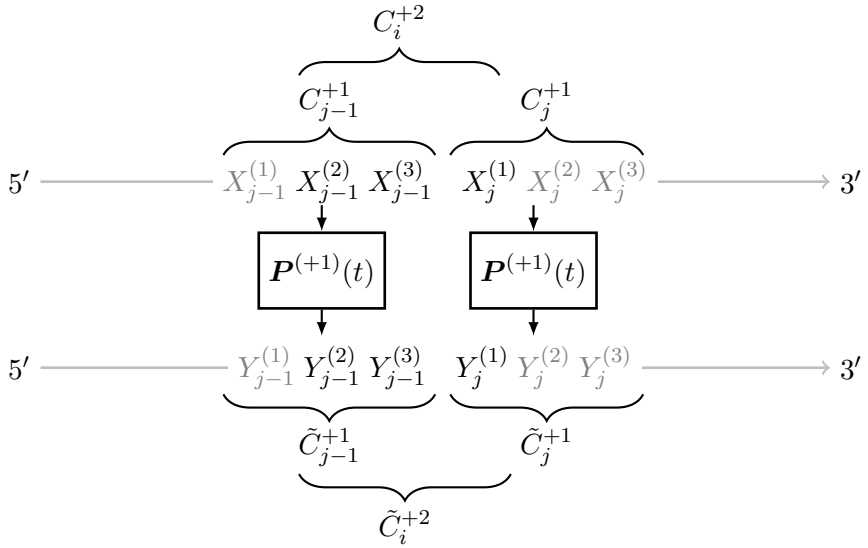


Figure 4.1.: Scheme of transitions of two consecutive codons in sequence direction on forward strand and in time direction.

4.1.2. Evolution Channel in Reading Frames

Based on the Evolution Channel of protein-coding reading frame $\{+1\}$, we investigate the corresponding evolution channel within the alternative reading frames $\mathcal{F}_A = \{-1, \pm 2, \pm 3\}$, i.e., the shifted and reverse complement reading frames. Given rate matrix $\boldsymbol{Q}^{(+1)}$ based on input parameters $\kappa^{(+1)}, \omega^{(+1)}, \boldsymbol{\pi}^{(+1)}$, we determine the corresponding rate matrices per reading frame denoted by $\boldsymbol{Q}^{(f)}$, $f \in \mathcal{F}_A$ knowing $\boldsymbol{\pi}^{(+1)}$. This task can be solved either via the transition matrices $\boldsymbol{P}^{(f)}(t)$, $f \in \mathcal{F} = \{\pm 1, \pm 2, \pm 3\}$ or directly using the rate matrices. The advantage, if we relinquish transition matrices is, that we are independent of the evolutionary time t.
In order to calculate the rate matrices within alternative reading frames, we first have to determine the joint rate matrix of two independent Markov processes as intermediate result, therefore we proceed as follows: We already know, that the transition matrix at time t can be obtained from the rate matrix via

$$\boldsymbol{P}(t) = e^{\boldsymbol{Q}t}$$

and accordingly the reverse operation

$$\boldsymbol{Q} = \frac{1}{t} \ln\left(\boldsymbol{P}(t)\right). \tag{4.5}$$

The joint probability matrix of two independent Markov processes 1 and 2 can be obtained from

$$\boldsymbol{P}_{12}(t) = e^{\boldsymbol{Q}_1 t} \otimes e^{\boldsymbol{Q}_2 t}, \tag{4.6}$$

where $\otimes$ denotes the so called Kronecker product [ZD13, LT85], which is defined by

Definition 4.1 (Kronecker product). For matrices $\boldsymbol{A} = \{a_{ij}\} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{B} = \{b_{ij}\} \in \mathbb{R}^{p \times q}$, their Kronecker product is defined as

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{pmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \dots & a_{1n}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \dots & a_{2n}\boldsymbol{B} \\ \vdots & \vdots & & \vdots \\ a_{m1}\boldsymbol{B} & a_{m2}\boldsymbol{B} & \dots & a_{mn}\boldsymbol{B} \end{pmatrix} \in \mathbb{R}^{mp \times nq}.$$

□

The Kronecker product of matrix exponentials can be replaced by the so called Kronecker sum [SW81, ZD13, LT85] to simplify the matrix exponential of the Kronecker product [SW81, ZD13].

Definition 4.2 (Kronecker sum). For square matrices $\boldsymbol{A} = \{a_{ij}\} \in \mathbb{R}^{a \times a}$ and $\boldsymbol{B} = \{b_{ij}\} \in \mathbb{R}^{b \times b}$, the Kronecker sum is defined as

$$\boldsymbol{A} \oplus \boldsymbol{B} = \boldsymbol{A} \otimes \boldsymbol{I}_b + \boldsymbol{I}_a \otimes \boldsymbol{B},$$

where $\boldsymbol{I}_a$ is the $a \times a$ identity matrix and $\boldsymbol{I}_b$ accordingly. □

Concept 4.2 (Kronecker product of matrix exponential).

$$e^{\boldsymbol{A}} \otimes e^{\boldsymbol{B}} = e^{\boldsymbol{A} \oplus \boldsymbol{B}}$$

□

We obtain the corresponding combined rate matrix either via the combined transition matrix of the process or directly from the rate matrices by performing some trivial transformations

$$\begin{aligned}
\boldsymbol{Q}_{12} &\overset{(a)}{=} \frac{1}{t} \ln \boldsymbol{P}_{12}(t) && \text{(a) applying (4.5)} \\
&\overset{(b)}{=} \frac{1}{t} \ln \left(e^{\boldsymbol{Q}_1 t} \otimes e^{\boldsymbol{Q}_2 t} \right) && \text{(b) applying (4.6)} \\
&\overset{(c)}{=} \frac{1}{t} \ln \left(e^{\boldsymbol{Q}_1 t \oplus \boldsymbol{Q}_2 t} \right) && \text{(c) applying Concept 4.2} \\
&= \frac{1}{t} \left(\boldsymbol{Q}_1 t \oplus \boldsymbol{Q}_2 t \right) \\
&\overset{(d)}{=} \frac{1}{t} \left(\boldsymbol{Q}_1 t \otimes \boldsymbol{I}_{q_2} + \boldsymbol{I}_{q_1} \otimes \boldsymbol{Q}_2 t \right) && \text{(d) applying Definition 4.2,}
\end{aligned}$$

where $\boldsymbol{Q}_1$ has the type $(q_1 \times q_1)$ and $\boldsymbol{Q}_2$ type $(q_2 \times q_2)$. Note, the combined rate matrix $\boldsymbol{Q}_{12}$ has dimension $q_1 q_2 \times q_1 q_2$ and the parameter of time cancels out.
Applied to our setup, we determine the corresponding di-codon matrix of two similar rate matrices $\boldsymbol{Q}^{(+1)}$ by

$$\boldsymbol{Q}^{(+1)}_{\text{di-codon}} = \left(\boldsymbol{Q}^{(+1)} \otimes \boldsymbol{I}_Q + \boldsymbol{I}_Q \otimes \boldsymbol{Q}^{(+1)} \right), \tag{4.7}$$

where $\boldsymbol{I}_Q$ is the identity matrix with the dimension 61×61, leading to a 3721×3721 di-codon matrix. At the example of reading frame $\{+2\}$, we now show, how those di-codons have to be merged in order to get the codon transition in the corresponding reading frame. In Figure 4.1 on page 81 the connection between frame $\{+1\}$ and frame $\{+2\}$ is indicated, where the nucleotides that belong to the codons of interest are black and the remaining nucleotides are gray. Assume, we are interested in the rate from codon C_i^{+2} to codon $\tilde{C}_i^{+2}$ with realizations $c_i^{+2} = \left(x_{j-1}^{(2)} x_{j-1}^{(3)} x_j^{(1)} \right)$ and $\tilde{c}_i^{+2} = \left(y_{j-1}^{(2)} y_{j-1}^{(3)} y_j^{(1)} \right)$. We obtain

$$\begin{aligned}
&\boldsymbol{Q}^{(+2)} \left(y_{j-1}^{(2)} y_{j-1}^{(3)} y_j^{(1)} | x_{j-1}^{(2)} x_{j-1}^{(3)} x_j^{(1)} \right) = \boldsymbol{Q}^{(+2)} \left(\tilde{c}_i^{+2} | c_i^{+2} \right) \\
&= \frac{\sum_{x_{j-1}^{(1)}, x_j^{(2)}, x_j^{(3)}, y_{j-1}^{(1)}, y_j^{(2)}, y_j^{(3)} \in \mathcal{B}} \pi^{(+1)}_{\text{di-codon}}(x_{j-1}^{(1)} c_i^{+2} x_j^{(2)} x_j^{(3)}) \cdot \boldsymbol{Q}^{(+1)}_{\text{di-codon}}(y_{j-1}^{(1)} \tilde{c}_i^{+2} y_j^{(2)} y_j^{(3)} | x_{j-1}^{(1)} c_i^{+2} x_j^{(2)} x_j^{(3)})}{\sum_{x_{j-1}^{(1)}, x_j^{(2)}, x_j^{(3)} \in \mathcal{B}} \pi^{(+1)}_{\text{di-codon}}(x_{j-1}^{(1)} c_i^{+2} x_j^{(2)} x_j^{(3)})},
\end{aligned}$$

where the probability for a di-codon in $\{+1\}$ that consists of two consecutive codons c_{j-1} and c_j is determined under the IID assumption using the stationary distribution $\boldsymbol{\pi}^{(+1)}$ by $\pi^{(+1)}_{\text{di-codon}}(c_{j-1} c_j) = \pi^{(+1)}_{c_{j-1}} \cdot \pi^{(+1)}_{c_j} \quad \forall\, c_{j-1}, c_j \in \mathcal{C}_{61}$. Summarized, the rate matrix in a certain reading frame is based on rate matrix $\boldsymbol{Q}^{(+1)}$ and the stationary distribution $\boldsymbol{\pi}^{(+1)}$.
The rate matrices in the remaining reading frames denoted with $\boldsymbol{Q}^{(f)}$, $f \in \{-1, -2, \pm 3\}$ can be determined accordingly, the corresponding equations can be found in Appendix C. Note

that $\boldsymbol{Q}^{(f)}$ is a 64×64 matrix for $f = \{\pm 2, \pm 3\}$ and a 61×61 matrix for $f = \{\pm 1\}$. Since every frame has its own codon alphabet, we denote the corresponding state space in each reading frame $f \in \mathcal{F}$ with $\mathcal{C}^{(f)}$.

Given the rate matrix $\boldsymbol{Q}^{(f)}$ of a time-continuous Markov chain the corresponding stationary distribution $\boldsymbol{\pi}^{(f)}$ in each reading frame is given by (4.4) and the corresponding transition matrix $\boldsymbol{P}^{(f)}(t)$ for evolutionary time t is determine by (4.2).

4.1.3. Simulation Results

Throughout this chapter, we chose *Escherichia coli* O157:H7 EDL933 (*EHEC*, Accession number NC_002655) as a model organism. The bacterium has a GC content of 50.4% and a genome of length 5528445 base pairs. We take the empirical codon usage $\hat{\boldsymbol{\pi}}^{(+1)}$ of *EHEC* as input stationary distribution $\boldsymbol{\pi}^{(+1)}$ to rate matrix $\boldsymbol{Q}^{(+1)}$. The genomic data are publicly available, further information on *EHEC* as well as detailed information on the steps needed to extract the empirical codon usage $\hat{\boldsymbol{\pi}}^{(+1)}$ is given in Section 3.3.1 especially paragraph *Data preparation of prokaryotic genome data*, where the elements $\{\hat{\pi}_i^{(+1)}\}_{i \in \mathcal{C}_{64}}$ are determined in (3.22). Since the codon usage runs over alphabet $\mathcal{C}_{64}$, we have to remove the stop codon probabilities to get the input distribution $\boldsymbol{\pi}^{(+1)}$ over $\mathcal{C}_{61}$.
The simulation framework is as follows: Given input parameters $\kappa, \omega^{(+1)}, \boldsymbol{\pi}^{(+1)}$, we determine the rate matrix $\boldsymbol{Q}^{(+1)}$ and together with input parameter t the corresponding transition matrix $\boldsymbol{P}^{(+1)}(t)$. Note, that input parameter $\boldsymbol{\pi}^{(+1)}$ is fixed over all simulations, while parameters κ, t run over different values. Since the nonsynonymous/synonymous rate ratio $\omega^{(f)}$ is of special interest for our investigations, we dedicate Section 4.2 to its interpretation. We set the following parameters: The transition/transversion rate ratio $\kappa \in \{1.0, 3.75, 5.0\}$ and the nonsynonymous to synonymous rate ratio $\omega^{(+1)} \in \{0.3, 3.0\}$, whereas the time t takes values in $[0.1, 0.2, \ldots, 10]$. The parameters are chosen as suggested in, e.g., [YN00, ZY06, ZLY06]. In the following, we will have a closer look at the assumptions that have to be fulfilled as well as the different parameters of the model.

Assumptions for the model The continuous-time Markov chain is time-reversible, this property implies that the process is stationary over time. This is the case, if the detailed balance equation is fulfilled, i.e., for each parameter set, we control if for each frame $f \in \mathcal{F}$

$$\pi_x^{(f)} Q_{xy}^{(f)} = \pi_y^{(f)} Q_{yx}^{(f)} \qquad \forall\, x, y \in \mathcal{C}^{(f)} \text{ with } x \neq y$$

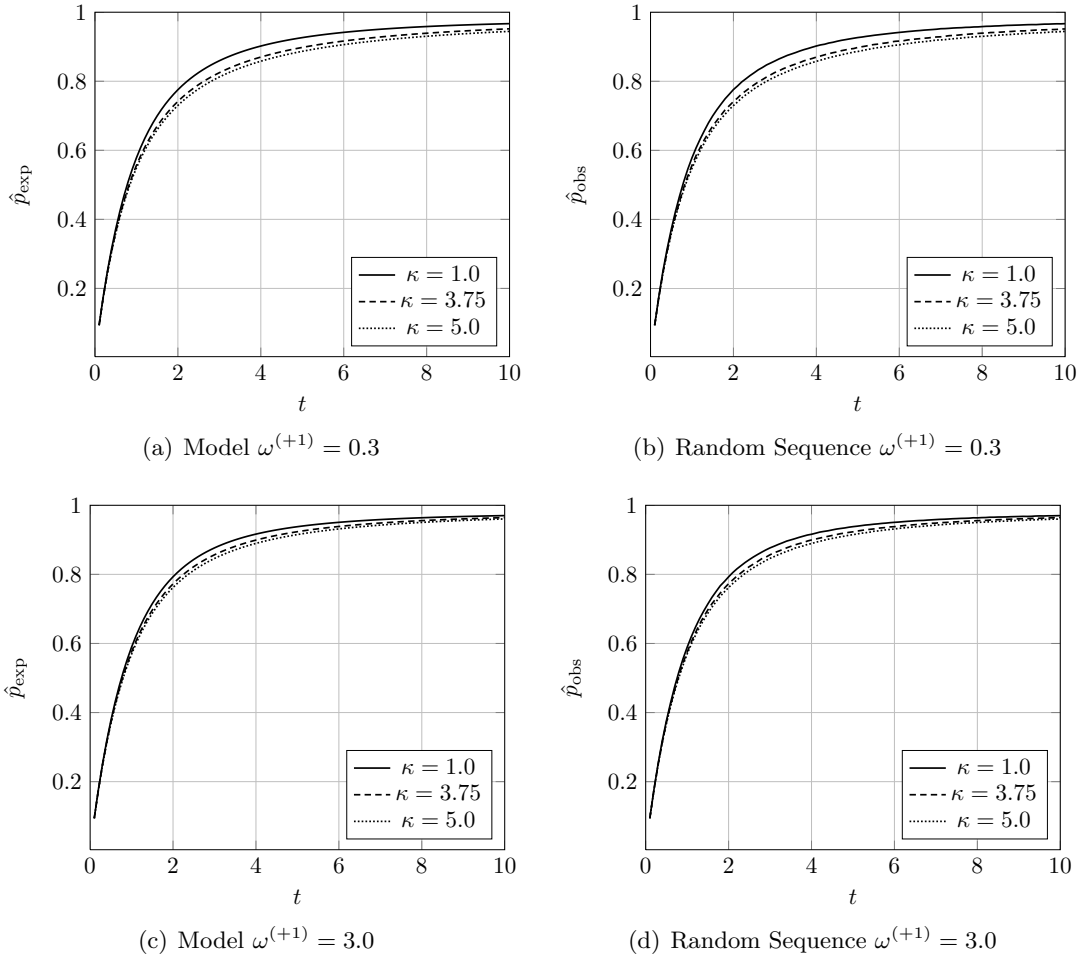
holds. This means that the process looks the same if time runs into future direction or back into the past [Yan06]. The considered state space $\mathcal{C}^{(f)}$ depends on the reading frame, since for frames $\{\pm 1\}$ only 61 codons are possible and for the remaining frame 64 codons.
The matrix exponential can be determined by (4.2) if the continuous-time Markov chain is decomposable. This is ensured, if for each parameter set the eigenvalues of $\boldsymbol{Q}^{(f)}$ in each reading frame have algebraic multiplicity one, i.e., the $n \times n$ rate matrix, has n distinct eigenvalues which implies linearly independent eigenvectors. The number n depends again on the reading frame.

We determined the stationary distribution $\boldsymbol{\pi}^{(f)}$ as well as the rate matrix per alternative reading frame $\boldsymbol{Q}^{(f)}, f \in \mathcal{F}_A$ as presented in Section 4.1.2. In fact all rate matrices on forward and backward strand considered in this simulation have distinct eigenvalues and the detailed balance equation is fulfilled. According to [Kei79] all eigenvalues of a rate matrix, that belong to a reversible Markov chain, are real.

Evolutionary time and transition/transversion rate ratio Here, we investigate how progress of evolutionary time t influences the number of substituted codons comparing a sequence $\boldsymbol{x}$ before the evolution channel with sequence $\boldsymbol{y}$ afterwards. Therefore, we determine the transition matrix per reading frame $\boldsymbol{P}^{(f)}(t), f \in \mathcal{F}_A$ as presented in Section 4.1.2.
We already know that the rate matrix is multiplied by a scaling factor, normalizing the average substitution rate per codon to $\lambda = 1$ such that the distance between two sequences $\boldsymbol{x}$ and $\boldsymbol{y}$ equals evolutionary time t, since $d = \lambda \cdot t$ holds. Time t or distance d can be estimated using the proportion of different sites between two sequences expected by the model following the principle of [Yan06]:

$$\hat{p} = \underbrace{\sum_{x \neq y} f_{xy}(t)}_{\hat{p}_{\text{obs}}} = \sum_{x \neq y} \pi_x^{(f)} P_{xy}^{(f)}(t) = \underbrace{\sum_{x} \pi_x^{(f)} (1 - P_{xx}^{(f)}(t))}_{\hat{p}_{\text{exp}}},$$

where $f_{xy}(t)$ is the relative frequency of observing a site with codon x in $\boldsymbol{x}$ and codon y after time t in $\boldsymbol{y}$. Unfortunately, this simple distance measure underestimates the true number of substitutions with growing time t. The reason is that there are different types of codon replacements, e.g., the measure can not distinguish if only one or multiple substitutions occurred at a given site. Further, it may also happen that a substitution changes back to the original codon, e.g., [Hur02].
We compare the model prediction $\hat{p}_{\text{exp}}$ based on $\boldsymbol{P}^{(f)}(t)$ and $\boldsymbol{\pi}^{(f)}$ in reading frames $\{+1\}$ with $\hat{p}_{\text{obs}}$ of a simulated random sequence. The procedure to determine the random sequence is as follows: Given input parameters $\kappa, \omega^{(+1)}, \boldsymbol{\pi}^{(+1)}, t$, we determine the rate matrix $\boldsymbol{Q}^{(+1)}$ and the corresponding transition matrix $\boldsymbol{P}^{(+1)}(t)$. We generate in reading frame $\{+1\}$ a random sequence $\boldsymbol{x}_{\text{in}} = (c_0 c_1 \dots c_{N_C-1})$ of length $N_C = 10^6$ codons, where each codon $c_n \in \mathcal{C}_{61}$ is drawn IID from given input distribution $\boldsymbol{\pi}^{(+1)}$. Each codon is then transmitted independently over evolution channel $\left(\mathcal{C}_{61}, \boldsymbol{P}^{(+1)}(t), \mathcal{C}_{61}\right)$. Input sequence $\boldsymbol{x}_{\text{in}}$ evolves to output sequence $\boldsymbol{y}_{\text{out}}$, such that a comparison of both results in $\hat{p}_{\text{obs}}$. The results are presented in Figure 4.2 on page 86. Both parameters κ and $\omega^{(+1)}$ have an influence on the distance, but if we compare the model predictions with the simulation results of the random sequence, $\hat{p}_{\text{exp}}$ matches $\hat{p}_{\text{obs}}$.
We are especially interested how the parameter set of the evolution channel in the protein-coding reading frame $\{+1\}$ influences the corresponding channels and therefore the so called selection pressure in the different reading frames. The evolutionary time that passes in each reading frame has to be the same in order to make the corresponding parameter $\omega^{(f)}$ comparable. This separated consideration is possible since in each reading frame $\boldsymbol{P}^{(f)}(t)$ is determined for the same input time t from the time invariant rate matrix $\boldsymbol{Q}^{(f)}$. As remaining



(a) Model $\omega^{(+1)} = 0.3$

(b) Random Sequence $\omega^{(+1)} = 0.3$

(c) Model $\omega^{(+1)} = 3.0$

(d) Random Sequence $\omega^{(+1)} = 3.0$

Figure 4.2.: Expected proportion of different sites $\hat{p}_{\text{exp}}$ between two sequences separated by time t under the given model for different parameter sets ((a),(c)). Observed proportion of different sites $\hat{p}_{\text{obs}}$ based on a random sequence of IID codons for different parameter sets ((b),(d)). Only reading frame $\{+1\}$ is presented.

parameter, the transition/transversion rate ratio κ has to be discussed. Since this is a value that operates on the nucleotide level, κ does not change while considering alternative reading frames and has therefore the same value as given in the protein-coding frame. For this reason, only reading frame $\{+1\}$ was considered in Figure 4.2 and the next section deals with the influence of $\omega^{(+1)}$ on $\omega^{(f)}$ in detail.

4.2. Selection Pressure During Evolution

Understanding the mechanism behind sequence evolution of DNA is a challenge addressed in research over the last decades, e.g., in [Kim68, ZP65]. To measure selective strength or pressure of molecular evolution on a protein, the nonsynonymous and synonymous substitution rates among protein-coding sequences are estimated [ZY06]. The development of the basic concepts date back to 1980, when Miyata and Yasunags [MY80] and Perler *et al.* [PEL$^+$80] published their methods to estimate evolutionary rates and sequence divergence. Nowadays the selection pressure on the protein level is described by the ratio $\omega = \frac{d_N}{d_S}$, the number of nonsynonymous substitutions per synonymous site d_N to the number of synonymous substitutions per synonymous site d_S. The corresponding definitions can be found in many publications, e.g., [Hur02, YB00, Yan06]. Differences between sequences on the nucleotide level that lead to different amino acids are called nonsynonymous substitutions, whereas differences that leave the encoded amino acid unchanged are synonymous or silent substitutions. To measure the variation between sequences, the synonymous and nonsynonymous changes are counted. Afterwards, the values are normalized to the number of possible synonymous and nonsynonymous changes to correct for the redundancy of the code [Hur02]. In fact this ensures that neutral evolution leads to $\omega = 1$. Further we distinguish between two scenarios: If deleterious mutations are eliminated, such that the protein sequence is protected, we get $\omega < 1$, called negative or purifying selection. On the other hand if advantageous mutations changing the protein sequence occur, we get $\omega > 1$ which is denoted as positive or adaptive selection. Inference whether synonymous changes appeared randomly or were driven by selection is not possible [YB00]. Over the years, several methods have been developed to determine ω, either by maximum likelihood estimation [GY94] or by heuristic counting methods introduced in [MY80, PEL$^+$80]. For the counting method, there have been numerous improvements published that incorporate different levels of biases like transition to transversion bias or codon frequency bias,e.g., [NG86, LWL85, Li93, PB93, Com95, Ina95, YN00, ZLY06]. Since those counting methods contain steps and assumptions that cannot be justified precisely, they are also called approximate methods [YN00]. The main advantage of the maximum likelihood method is, that it does not involve approximate steps. Unfortunately most methods produce different estimation results, as they are not very robust towards the model assumptions. More recently Bayesian estimation methods are investigated, e.g., [ARY14]. Nevertheless the maximum likelihood method and because of their simplicity the counting methods are widely applied to real data, e.g., [NBC$^+$05, GCFW08, BRBR12, GDA$^+$13] and, e.g., [GGMS03, SWTW10, GRS$^+$13]. Detailed comparisons and reviews of different methods, their assumptions, biases and problems, can be found in, e.g., [NG86, YB00, MA09, Ina95, ZY06].

In Section 4.2.1 we present two methods to determine substitution rates, that are state of the art in literature. Those methods are then applied to the alternative reading frames in Section 4.2.2 and finally we present simulation results and interpretations regarding selection pressure in overlapping genes in Section 4.2.3.

4.2.1. Estimation of Substitution Rates

The number of methods to estimate the selection pressure is enormous, therefore we concentrate on two different concepts. As example for the counting method, we present the well-known method of Nei and Gojobory [NG86] denoted as NG, that is a simplified version of the concept presented by Miyata and Yasunaga [MY80]. For the maximum likelihood (ML) estimation of the substitution rates we follow Goldman and Yang [GY94]. According to [YN00] NG is the most commonly used counting method and ML the general method of choice to estimate ω. For both approaches the selection pressure is estimated based on a comparison of two protein-coding sequences.

Counting method NG The counting method to estimate the number of synonymous and nonsynonymous substitutions per site consists of three major steps [Yan06]:

1. Count synonymous (S) and nonsynonymous (N) sites (*counting sites*).
2. Count synonymous (S_d) and nonsynonymous (N_d) differences (*counting differences*).
3. Correct for multiple hits.

Here we deviate slightly from the typical presentation of NG in order to apply the method to our purposes easier and to emphasize how to implement the algorithm.
In the first step, called *counting sites*, we start with the building of a basic dictionary for the synonymous sites B_S and the nonsynonymous sites B_N that depends on the genetic code only. Therefore each of the three nucleotide sites within a codon is mapped to one of the two categories (synonymous mutation, nonsynonymous mutation). In detail we proceed as follows: For each codon $c \in \mathcal{C}_{64}$ that maps to an amino acid $a_c \in \mathcal{A}_{21}$, determine for each of the three positions within the codon, how many of the three possible nucleotide substitutions change the original amino acid a_c. For example

- $c = TTG \mapsto a_c = Leu$
  For position 1 following three nucleotide replacements are possible: $ATG \mapsto Met$, $GTG \mapsto Val$, $CTG \mapsto Leu$, therefore $\frac{1}{3}$ of all nucleotide substitutions are synonymous and do not change the original amino acid $a_c = Leu$. For position 2 we determine $TAG \mapsto Stop$, $TGG \mapsto Trp$, $TCG \mapsto Ser$, therefore 0 nucleotide substitutions are synonymous. And for position 3 we get $TTA \mapsto Leu$, $TTC \mapsto Phe$, $TTT \mapsto Phe$, such that $\frac{1}{3}$ of all nucleotide substitutions are synonymous and do not change amino acid $a_c = Leu$.
  Overall we have for TTG the proportion per position $\left(p^{(1)}_{TTG}, p^{(2)}_{TTG}, p^{(3)}_{TTG}\right) = \left(\frac{1}{3}, 0, \frac{1}{3}\right)$.
- $c = TTC \mapsto a_c = Phe$
  For position 1 $ATC \mapsto Ile$, $GTC \mapsto Val$, $CTC \mapsto Leu$, results in 0 synonymous substitutions. Also for position 2 we determine $TAC \mapsto Tyr$, $TGC \mapsto Cys$, $TCC \mapsto Ser$, that are 0 synonymous substitutions. Only for position 3 we get $TTA \mapsto Leu$, $TTG \mapsto Leu$, $TTT \mapsto Phe$, such that $\frac{1}{3}$ of all nucleotide substitutions are synonymous and do

not change amino acid $a_c = Phe$.
Overall we have for TTC $\left(p^{(1)}_{TTC}, p^{(2)}_{TTC}, p^{(3)}_{TTC}\right) = \left(0, 0, \frac{1}{3}\right)$.

For each codon combination, the sum over the average per codon position yields the entry of the basic dictionary for the synonymous sites B_S

$$B_S(x,y) = \sum_{i=1}^{3} \left(\frac{p_x^{(i)} + p_y^{(i)}}{2} \right) \qquad \forall\, x, y \in \mathcal{C}_{64}, \tag{4.8}$$

for the example we get $B_S(TTG, TTC) = \sum_{i=1}^{3} \left(\frac{1}{6}, 0, \frac{1}{3}\right) = 0.5$, which corresponds to a mutation rate from codon TTG to TTC or the other way round.
The basic dictionary of the nonsynoymous sites B_N can be determined directly from the average per position of B_S, since every position has to sum to one, i.e.,

$$B_N(x,y) = \sum_{i=1}^{3} \left(1 - \frac{p_x^{(i)} + p_y^{(i)}}{2} \right) = 3 - B_S(x,y) \qquad \forall\, x, y \in \mathcal{C}_{64}, \tag{4.9}$$

such that for the example $B_N(TTG, TTC) = \sum \left(\frac{5}{6}, 1, \frac{2}{3}\right) = 2.5$ holds, since $B_S + B_N = 3$ is always fulfilled.
Given two DNA sequences of length n denoted with $\boldsymbol{x} = (x_0 x_1 \cdots x_{n-1})$ and $\boldsymbol{y} = (y_0 y_1 \cdots y_{n-1})$, where $x_i, y_i \in \mathcal{C}_{64}$, the synonymous and nonynonymous sites can be determined by a codon wise comparison, i.e.,

$$S = \sum_{i=0}^{n-1} B_S(x_i, y_i) \tag{4.10}$$

$$N = \sum_{i=0}^{n-1} B_N(x_i, y_i) = 3n - S \tag{4.11}$$

Since B_S and B_N depend only on the genetic code, the dictionaries have to be determined only once and can then be applied to all sequences that have to be considered.

In the second step, *counting differences*, we build a basic dictionary for the synonymous differences B_{S_d} and the nonsynonymous differences B_{N_d} that depends again only on the genetic code. Each difference between two codons has to be categorized. For each codon pair $x, y \in \mathcal{C}_{64}$ we determine the Hamming distance $\mathrm{d_H}[x, y]$.

- If $\mathrm{d_H}[x, y] = 0$, the codons as well as their corresponding amino acids are identical, we set $B_{S_d}(x, y) = B_{N_d}(x, y) = 0$.
- If $\mathrm{d_H}[x, y] = 1$, the corresponding amino acids of the codons $x \mapsto a_x \in \mathcal{A}_{21}$ and $y \mapsto a_y \in \mathcal{A}_{21}$ either differ, i.e., $a_x \neq a_y$ which results in $B_{S_d}(x, y) = 0$, $B_{N_d}(x, y) = 1$ or are equal, i.e., $a_x = a_y$ such that $B_{S_d}(x, y) = 1$, $B_{N_d}(x, y) = 0$.
  Example: For $x = TTG \mapsto a_x = Leu$ and $y = TTC \mapsto a_y = Phe$ we have one point mutation that leads to a nonsynonymous amino acid, which results in $B_{S_d}(TTG, TTC) = 0$ and $B_{N_d}(TTG, TTC) = 1$.

- If $\mathrm{d_H}[x, y] = 2$, take the average over the two possible pathways built from one-point mutations.
  Example: Consider both pathways to go from $x = TTT \mapsto a_x = Phe$ to $y = GTA \mapsto a_y = Val$. The first pathway is $x = TTT \mapsto a_x = Phe$ to $GTT \mapsto Val$ to $y = GTA \mapsto a_y = Val$, i.e., one nonsynonymous and one synonymous replacement. The second path $x = TTT \mapsto a_x = Phe$ to $TTA \mapsto Leu$ to $y = GTA \mapsto a_y = Val$ has two nonsynonymous replacements. Take the average over both pathways $B_{S_d}(TTT, GTA) = \frac{1+0}{2} = 0.5$, $B_{N_d}(TTT, GTA) = \frac{1+2}{2} = 1.5$
- If $\mathrm{d_H}[x, y] = 3$, take the average over the six possible pathways built from one-point mutations.

Go codon wise through the DNA sequences $\boldsymbol{x}$ and $\boldsymbol{y}$, where $x_i, y_i \in \mathcal{C}_{64}$ and sum the corresponding differences up

$$S_d = \sum_{i=0}^{n-1} B_{S_d}(x_i, y_i), \tag{4.12}$$

$$N_d = \sum_{i=0}^{n-1} B_{N_d}(x_i, y_i). \tag{4.13}$$

Since B_{S_d} and B_{N_d} depend only on the genetic code, the dictionaries have to be determined only once and then can be applied to all sequences that have to be considered.

In the last step, called *correcting for multiple hits*, we first determine the fraction of synonymous (D_S) and nonsynonymous (D_N) sites that differ by

$$D_S = \frac{S_d}{S} \qquad \text{and} \qquad D_N = \frac{N_d}{N}. \tag{4.14}$$

This correction factor is necessary, since a comparison of two sequences underestimates the true number of mutations that occurred over time as multiple substitutions can not be distinguished from single mutations [Hur02, ZY06]. To correct for those multiple hits, the number of synonymous substitutions per synonymous site (d_S) and nonsynonymous substitutions per nonsynonymous site (d_N) are estimated using the Jukes-Cantor formula [JC69]

$$d_S = \begin{cases} \infty & \text{if } D_S > \frac{3}{4} \\ -\frac{3}{4} \ln\left(1 - \frac{4}{3} D_S\right) & \text{else} \end{cases} \tag{4.15}$$

$$d_N = \begin{cases} \infty & \text{if } D_N > \frac{3}{4} \\ -\frac{3}{4} \ln\left(1 - \frac{4}{3} D_N\right) & \text{else .} \end{cases} \tag{4.16}$$

Note, the Jukes-Cantor method was originally invented for a substitution model of nucleotides and its application in this scenario is highly controversial. However, the effect of this correction is minor and therefore the bias introduced can be neglected [YN00].

The estimated nonsynonymous/synonymous rate ratio based on the NG method is given by

$$\omega_{\mathrm{NG}} = \frac{d_N}{d_S}. \tag{4.17}$$

ML Estimation A detailed description of the procedure presented here can be found in [YN00, Yan06]. The maximum likelihood (ML) method [GY94] uses two DNA sequences to estimate the parameters $t, \kappa, \omega, \boldsymbol{\pi}$.
Given two DNA sequences of length n denoted with $\boldsymbol{x} = (x_0 x_1 \cdots x_{n-1})$ and $\boldsymbol{y} = (y_0 y_1 \cdots y_{n-1})$, where $x_i, y_i \in \mathcal{C}_{61}$. The codon frequencies are usually estimated from the observed data, where different methods based on the nucleotide or codon level exist, e.g., a simple assumption is the uniform distribution over all codons. We denote the estimate as $\hat{\boldsymbol{\pi}} = \{\hat{\pi}_x\}$. For further methods to measure the codon frequencies see, e.g., [Yan06, PDMS10]. The remaining parameters are estimated numerically by maximizing the log likelihood function

$$\ell(t, \kappa, \omega) = \sum_{x,y \in \mathcal{C}_{61}} n_{xy} \log\left(f_{xy}(t)\right) = \sum_{x,y \in \mathcal{C}_{61}} n_{xy} \log\left(\hat{\pi}_x \cdot P_{xy}(t)\right), \tag{4.18}$$

where n_{xy} is the number of sites occupied by codon x in $\boldsymbol{x}$ and codon y in $\boldsymbol{y}$ and f_{xy} the corresponding joint probability. This maximization is usually implemented iteratively where in each maximization step, the transition matrix $P_{xy}(t)$ is determined from (4.1) and (4.2) with the current values of t, κ, ω and $\hat{\boldsymbol{\pi}}$. The resulting estimates, denoted with $\hat{t}, \hat{\kappa}, \hat{\omega}$, are then used to calculate ω_{ML}. Therefore, the proportion of synonymous substitutions, that is the sum over all codon pairs x and y with $x \neq y$ that code for the same amino acid is calculated by

$$\rho_S = \sum_{x \neq y, a_x = a_y} \hat{\pi}_x Q_{xy}\left(\hat{\pi}_y, \hat{\kappa}, \hat{\omega}\right), \tag{4.19}$$

where a_x is the amino acid encoded by codon x and the rate matrix is calculated from (4.1) using $\hat{\kappa}, \hat{\omega}$ and $\hat{\boldsymbol{\pi}}$ which is indicted by the parenthesis in the equation. The proportion of nonsynonymous substitutions is accordingly

$$\rho_N = \sum_{x \neq y, a_x \neq a_y} \hat{\pi}_x Q_{xy}\left(\hat{\pi}_y, \hat{\kappa}, \hat{\omega}\right), \tag{4.20}$$

which corresponds to $\rho_N = 1 - \rho_S$ since the rate matrix is normalized. Note, the numbers of synonymous and nonsynonymous substitutions per codon after time $\hat{t}$ are $S_d = \hat{t}\rho_S$ and $N_d = \hat{t}\rho_N$, where $S_d + N_d = \hat{t}$ holds. The proportion of synonymous sites is defined as the proportion of synonymous mutations on the nucleotide level when no selective constraints exist at the amino acid level. This can be calculate from (4.19) fixing $\omega = 1.0$

$$\rho_S^1 = \sum_{x \neq y, a_x = a_y} \hat{\pi}_x Q_{xy}\left(\hat{\pi}_y, \hat{\kappa}, 1.0\right), \tag{4.21}$$

and the nonsynonymous sites accordingly

$$\rho_N^1 = \sum_{x \neq y, a_x \neq a_y} \hat{\pi}_x Q_{xy}\left(\hat{\pi}_y, \hat{\kappa}, 1.0\right). \tag{4.22}$$

The numbers of synonymous and nonsynonymous nucleotides sites per codon are therefore given by $S = 3\rho_S^1$ and $N = 3\rho_N^1$. The number of synonymous substitutions per synonymous site is

$$d_S = \frac{S_d}{S} = \frac{\hat{t}\rho_S}{3\rho_S^1} \tag{4.23}$$

and the number of nonysnonymous substitutions per nonsynonymous site

$$d_N = \frac{N_d}{N} = \frac{\hat{t}\rho_N}{3\rho_N^1}. \tag{4.24}$$

The ratio compares the proportions of synonymous and nonsynonymous substitutions that occured during natural selection with the expected ratio if there had been no selection on the protein

$$\omega_{\mathrm{ML}} = \frac{d_N}{d_S} = \frac{\frac{\rho_N}{\rho_S}}{\frac{\rho_N^1}{\rho_S^1}} = \frac{\rho_N \rho_S^1}{\rho_N^1 \rho_S}. \tag{4.25}$$

Note that the ratio is independent of the time parameter.

4.2.2. Estimation of Substitution Rates in Reading Frames

In the previous Section 4.2.1 we introduced two state of the art methods to estimate nonsynonymous/synonymous rate ratios ω_{NG} and ω_{ML}. Since we are interested in the estimation of selection pressure in alternative reading frames, we describe now, how both methods can be adapted in our scenario to determine $\omega_{\mathrm{NG}}^{(f)}$ and $\omega_{\mathrm{ML}}^{(f)}$ for frames $f \in \mathcal{F}$. Only the differences to the former description are presented. Additionally, we present a third method that results directly from the ML method, denoted with $\omega_{\mathrm{Q}}^{(f)}$.

Counting method NG Given the stationary distribution $\boldsymbol{\pi}^{(f)}$ and the corresponding transition matrix $\boldsymbol{P}^{(f)}(t)$ for evolutionary time t in each reading frame, the joint PMF that codons x evolves to codon y and are therefore observed together is

$$p_{XY}(x, y) = \pi_x^{(f)} P_{xy}^{(f)}(t), \tag{4.26}$$

where the chosen codon alphabet $\mathcal{C}^{(f)}$ depends on the reading frame. For each codon, we build the weighted sum of the dictionaries according to the joint PMF to determine synonymous (S) and nonsynonymous (N) *sites*

$$S = \sum_{x,y \in \mathcal{C}^{(f)}} p_{XY}(x, y) \cdot B_S(x, y)$$

and

$$N = \sum_{x,y \in \mathcal{C}^{(f)}} p_{XY}(x, y) \cdot B_N(x, y),$$

where B_S and B_N were previously calculated in (4.8) and (4.9). Accordingly for the synonymous (S_d) and nonsynonymous (N_d) *differences*

$$S_d = \sum_{x,y \in \mathcal{C}^{(f)}} p_{XY}(x,y) \cdot B_{S_d}(x,y)$$

$$N_d = \sum_{x,y \in \mathcal{C}^{(f)}} p_{XY}(x,y) \cdot B_{N_d}(x,y)$$

where B_{S_d} and B_{N_d} were previously calculated. Now, we proceed with (4.14), (4.15) and (4.16) to determine finally

$$\omega_{\mathrm{NG}}^{(f)} = \frac{d_N}{d_S}. \tag{4.27}$$

ML Estimation Given the stationary distribution $\boldsymbol{\pi}^{(f)}$ and the corresponding transition matrix $\boldsymbol{P}^{(f)}(t)$ for evolutionary time t in each reading frame. Further we know the codon alphabet $\mathcal{C}^{(f)}$ that depends on the reading frame. We proceed exactly as presented in the original version of the ML estimation. The only difference is, that parameters κ and t are both assumed to be known input values of the model not changing within alternative frames and the codon frequency per reading frame $\boldsymbol{\pi}^{(f)}$ is also given. Only $\hat{\omega}$ is estimated maximizing (4.18)

$$\ell(\omega) = \sum_{x,y \in \mathcal{C}^{(f)}} p_{XY}(x,y) \log\left(\pi_x^{(f)} \cdot P_{xy}(t)\right),$$

where $p_{XY}(x,y)$ is the joint PMF determined by (4.26) based on the $\boldsymbol{P}^{(f)}(t)$ and $P_{xy}(t)$ are determined from (4.1) and (4.2) with the current value of ω. Note: Instead of maximizing $\ell(\omega)$, we can also minimize $-\ell(\omega)$, e.g., by applying Nelder-Mead simplex algorithm [NM65]. We determine the rate matrix in (4.19) and (4.20) with $\boldsymbol{\pi}^{(f)}, \kappa$ given and $\hat{\omega}$ estimated by

$$\rho_S^{(f)} = \sum_{\substack{x,y \in \mathcal{C}^{(f)} \\ x \neq y, a_x = a_y}} \pi_x^{(f)} Q_{xy}\left(\pi_y^{(f)}, \kappa, \hat{\omega}\right) \tag{4.28}$$

and

$$\rho_N^{(f)} = \sum_{\substack{x,y \in \mathcal{C}^{(f)} \\ x \neq y, a_x \neq a_y}} \pi_x^{(f)} Q_{xy}\left(\pi_y^{(f)}, \kappa, \hat{\omega}\right). \tag{4.29}$$

Accordingly applying (4.21) and (4.22) to

$$\rho_S^{1,(f)} = \sum_{\substack{x,y \in \mathcal{C}^{(f)} \\ x \neq y, a_x = a_y}} \pi_x^{(f)} Q_{xy}\left(\pi_y^{(f)}, \kappa, 1.0\right) \tag{4.30}$$

and

$$\rho_N^{1,(f)} = \sum_{\substack{x,y \in \mathcal{C}^{(f)} \\ x \neq y, a_x \neq a_y}} \pi_x^{(f)} Q_{xy}\left(\pi_y^{(f)}, \kappa, 1.0\right). \tag{4.31}$$

The ratio per reading frames is given following (4.25) by

$$\omega_{\mathrm{ML}}^{(f)} = \frac{\rho_N^{(f)} \rho_S^{1,(f)}}{\rho_N^{1,(f)} \rho_S^{(f)}}. \tag{4.32}$$

Rate matrix Given the stationary distribution $\boldsymbol{\pi}^{(f)}$ and the corresponding rate matrix $\boldsymbol{Q}^{(f)}$ in each reading frame. Further, we know the codon alphabet $\mathcal{C}^{(f)}$ that depends on the reading frame. We follow closely the procedure presented in the ML estimation, but skip the maximization of the likelihood function. Instead we access directly (4.19) to determine the proportion of synonymous substitutions, that is the sum over all codon pairs x and y with $x \neq y$ that code for the same amino acid

$$\rho_S^{(f)} = \sum_{\substack{x,y \in \mathcal{C}^{(f)} \\ x \neq y, a_x = a_y}} \pi_x^{(f)} Q_{xy}^{(f)}, \tag{4.33}$$

where a_x is the amino acid encoded by codon x. Accordingly from (4.20)

$$\rho_N^{(f)} = \sum_{\substack{x,y \in \mathcal{C}^{(f)} \\ x \neq y, a_x \neq a_y}} \pi_x^{(f)} Q_{xy}^{(f)}. \tag{4.34}$$

We assume that κ is given as input parameter to the protein-coding reading frame $\{+1\}$ and therefore also known in the alternative reading frames. Applying (4.21), we determine the proportion of synonymous sites without selective constraints on the amino acid level by

$$\rho_S^{1,(f)} = \sum_{\substack{x,y \in \mathcal{C}^{(f)} \\ x \neq y, a_x = a_y}} \pi_x^{(f)} Q_{xy} \left(\pi_y^{(f)}, \kappa, 1.0\right), \tag{4.35}$$

where the rate matrix is calculated from (4.1) using κ which is known and $\omega = 1.0$ which is fixed. Accordingly

$$\rho_N^{1,(f)} = \sum_{\substack{x,y \in \mathcal{C}^{(f)} \\ x \neq y, a_x \neq a_y}} \pi_x^{(f)} Q_{xy} \left(\pi_y^{(f)}, \kappa, 1.0\right). \tag{4.36}$$

The ratio per reading frames is given following (4.32) by

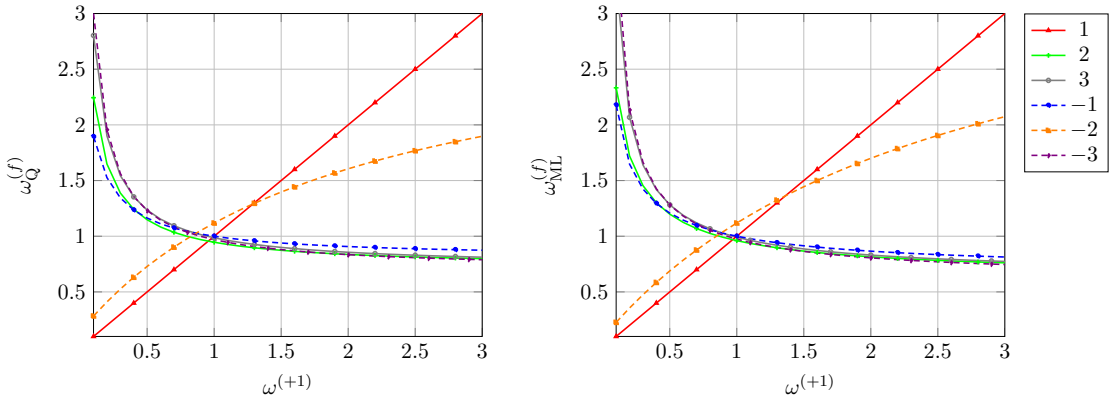
$$\omega_{\mathrm{Q}}^{(f)} = \frac{\rho_N^{(f)} \rho_S^{1,(f)}}{\rho_N^{1,(f)} \rho_S^{(f)}}. \tag{4.37}$$

4.2.3. Simulation Results

To investigate the influence of selection pressure during evolution on each reading frame, we compare the different methods presented above. As mentioned in Section 4.1.3, we refer to the codon usage of EHEC as input stationary distribution $\boldsymbol{\pi}^{(+1)}$ to rate matrix

$\boldsymbol{Q}^{(+1)}$. Further, we fix the transition/transversion rate $\kappa = 1.0$ and the time $t = 1.0$. The nonsynonymous/synonymous rate ratio $\omega^{(+1)}$ takes values in $[0.1, 0.2, \ldots, 3.0]$.
For each input parameter set, we obtain the rate matrix $\boldsymbol{Q}^{(f)}$, the transition matrix $\boldsymbol{P}^{(f)}(t)$ and the stationary distribution $\boldsymbol{\pi}^{(f)}$ following Section 4.1.2.

Estimation of selection pressure based on maximum likelihood method In Figure 4.3, we compare the calculation of the nonsynonymous/synonymous rate ratio $\omega_{\mathrm{Q}}^{(f)}$ of (4.37) for $f \in \mathcal{F}$ based on the rate matrix with the maximum likelihood method $\omega_{\mathrm{ML}}^{(f)}$ of (4.32) that is based on the transition matrix. The comparison reveals, that both methods lead to similar results. Further both methods result in a straight line with slope one for reading frame $\{+1\}$, where the input value of $\omega^{(+1)}$ is directly reproduced.

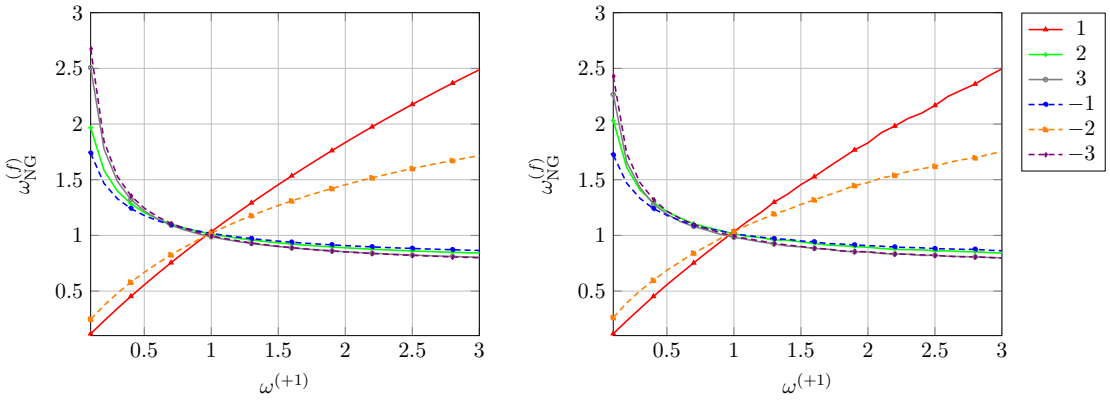


(a) Model (b) Model

Figure 4.3.: Estimation of nonsynonymous/synonymous rate ratio per reading frame $\omega^{(f)}$, based on the rate matrix (a) and the maximum likelihood approach (b). The input parameters are $\kappa = 1.0$ and $t = 1.0$. Protection of protein coding frame $\{+1\}$ is directly coupled with reading frame $\{-2\}$ sustaining sequences in both frames.

Selection against nonsynonymous substitutions on the DNA level with $\omega^{(+1)} < 1$, called purifying selection, sustains the protein sequence. Both methods show, that a protection of the protein coding reading frame, protects simultaneously the sequence in frame $\{-2\}$. In contrast to that, the remaining alternative reading frames face adaptive selection $\omega^{(f)} > 1$, $f \in \{-1, +2, \pm 3\} = \mathcal{F} \setminus \{+1, -2\}$ such that their sequence evolves freely. The opposite scenario is observed for adaptive selection in the protein coding reading frame. Values $\omega^{(+1)} > 1$ lead to $\omega^{(-2)} > 1$ implying that new information can be induced in frames $\{+1, -2\}$ simultaneously, whereas the other frames start slightly below the neutral mutation line and are better protected as $\omega^{(+1)}$ approaches 3.0.

Estimation of selection pressure based on counting method In Figure 4.4(a) we apply the counting method of Nei and Gojobori (NG) $\omega_{\mathrm{NG}}^{(f)}$ following (4.27) based on the transition matrix and the stationary distribution. Of particular interest is, that for increasing values $\omega^{(+1)} \to 3.0$, the NG method is not able to reproduce the input parameter in reading frame $\{+1\}$, such that $\omega^{(+1)} \neq \omega_{\mathrm{NG}}^{(+1)}$ holds. This deviation is due to a bias in the NG estimation method, that is well known in literature. According to [ZY06, YN00] an overestimation of d_S and an underestimation of d_N result for NG in an underestimated ω, especially for adaptive selection. When time increases this effect is even stronger. According to Ina [Ina95], who presented a detailed comparison of different estimation methods, the major cause for this bias, is that the number of synonymous sites in underestimated and the number of nonsynonymous sites overestimated. Finally [YN00] reveals that the substitutions model underlying NG acts on nucleotide level, which ignores that the probability that transitions and transversions lead to synonymous replacements depends on the codon position. Despite this bias, the behaviour of $\omega_{\mathrm{NG}}^{(f)}$ is consistent with the observations in Figure 4.4(a).



(a) Model (b) Random Sequence

Figure 4.4.: Estimation of nonsynonymous/synonymous rate ratio per reading frame $\omega^{(f)}$ using the NG counting method, based on the model (a) and a random sequence of IID codons (b). The input parameters are $\kappa = 1.0$ and $t = 1.0$. Bias is caused by chosen estimation method.

Further, in Figure 4.4, we compare the NG predictions of the model with the results obtained from the random sequence $\boldsymbol{x}_{\mathrm{in}}$, that is transmitted over $\left(\mathcal{C}_{61}, \boldsymbol{P}^{(+1)}(t), \mathcal{C}_{61}\right)$ resulting in $\boldsymbol{y}_{\mathrm{out}}$. Details on the random sequence can be found in Section 4.1.3. The NG estimate of reading frame $\{+1\}$ is based on $\boldsymbol{x}_{\mathrm{in}}, \boldsymbol{y}_{\mathrm{out}}$ following the procedure that resulted in (4.17), where we go codon wise through the sequences. Frame $\{+2, +3\}$, are obtained accordingly, shifting $\boldsymbol{x}_{\mathrm{in}}, \boldsymbol{y}_{\mathrm{out}}$ by one respectively two nucleotides. For the reverse frames, we use the reverse complement sequences as input to the procedure to obtain the NG estimates. Although NG reveals biased estimates, the predictions of the model in Figure 4.4(a) for the selection pressure per reading frame obtains almost identical estimates as the observations in the

random sequence pair in Figure 4.4(b).

The estimation of parameter ω should be perceived with great care. According to [YB00], a comparison of different estimation methods for ω reveals, that assumptions matter more than methods and approximate methods can lead to extreme biases for distantly related sequences. It also happens that different methods lead to varying estimation results, even if the sequences show high similarity [YB00]. Note: The NG method assumes equal codon frequencies, that have a major influence on the estimation of parameter $\omega_{\mathrm{NG}}^{(f)}$ according to [ZY06]. The ML method is based on a numerical maximization step to obtain $\omega_{\mathrm{ML}}^{(f)}$. Therefore, we prefer the rate matrix method $\omega_{\mathrm{Q}}^{(f)}$.

4.3. Quantifying Noise During Evolution

In the central dogma of molecular biology, the process of protein synthesis is expressed as a communication process (see Figure 2.3 on page 23). The combination of information theory and molecular biology dates back to 1974, when Yockey showed how the term of "information" in this context can be described mathematically by the application of Shannon's information theory [Yoc74]. Later, in 1992, he expanded his considerations [Yoc92] and revisited them again in 2005 [Yoc05]. The basic ideas and arguments of this part are based on [Yoc92, Yoc05], where we assume that the concepts presented in Chapter 2 are known. Noise in terms of information and communication theory disturbs useful information during transmission over a noisy channel, whereas genetic noise in molecular biology comprises anything that changes the expected result of protein synthesis. Typical sources of genetic noise are errors during the transcription or translation process or point mutations in the evolution channel. Here, the evolution of a sequence is considered as a communication process over time, transmitting codons independently over a channel with transition matrix $\boldsymbol{P}^{(f)}(t)$, where only point mutations of nucleotides occur as genetic noise and change the output sequence. This communication process should be evaluated, dealing with the following kind of questions:
A protein is transmitted over an evolutionary channel, how long is this information conserved over time in each reading frame? If no error correction processes exist, after what time can the output sequence no longer be distinguished from a random sequence?
Besides the channel coding theorem, Shannon [Sha48] shows how to measure the amount of information that can be transmitted over a noisy channel. Based on Shannon, Yockey [Yoc92] proposed to use the conditional entropy to measure the effect of genetic noise. Further, the mutual information is suggested to measure the relationship between any two sequences. More precisely, the similarity between an input sequence and an output sequence is compared using the mutual information.
In Section 4.3.1 we present how the effects of genetic noise perceive in alternative reading frames. The sequence similarity as well as the channel capacity is investigated in Section 4.3.2 and finally in Section 4.3.3 the corresponding simulation results and interpretations are presented.

4.3.1. Information Loss over Time

In order to examine the evolutionary channel, we start from the transition matrix $\boldsymbol{P}^{(f)}(t)$ in each reading frame $f \in \mathcal{F}$. Since we are interested in transmission of protein sequences and we know that different codons not necessarily lead to different amino acids, we consider amino acids instead of codons.
Given the codon evolution matrix $\boldsymbol{P}^{(f)}(t)$, the probabilities can be summarized to determine the amino acid evolution matrix $\boldsymbol{P}^{(f,\mathcal{A}_{20})}(t) = \{P_{xy}^{(f,\mathcal{A}_{20})}(t)\}_{x,y\in\mathcal{A}_{20}}$, based on the stationary distribution $\boldsymbol{\pi}^{(f)}$. Note, the stop codon probabilities are removed in all frames, leading to 20×20 matrices, since $|\mathcal{A}_{20}| = 20$.
Each pair of amino acids $x, y \in \mathcal{A}_{20}$ has an individual set of codons $\mathcal{C}_x$ and $\mathcal{C}_y$.

$$P_{xy}^{(f,\mathcal{A}_{20})}(t) = \frac{\sum_{i\in\mathcal{C}_x}\sum_{j\in\mathcal{C}_y}\pi_i^{(f)}P_{ij}^{(f)}(t)}{\sum_{i\in\mathcal{C}_x}\pi_i^{(f)}}. \tag{4.38}$$

After summing up the probabilities that belong to the same amino acid and normalization of the vector to remove the stop codon probabilities, we determine the corresponding stationary amino acid distribution, which we denote as $\boldsymbol{\pi}^{(f,\mathcal{A}_{20})}$.
Now the amino acid transition matrices over $\mathcal{A}_{20}$ can be directly applied to measure the genetic noise of the channel. The conditional entropy between two random variables X, Y was introduced in Definition 2.13. Based on those equation, we here determine the conditional entropy between input amino acid X and output amino acid Y in frame f. Assume, we know that a certain amino acid $a \in \mathcal{A}_{20}$, was transmitted, then the entropy conditioned on amino acid a is given by

$$H_t^{(f)}(Y|X=a) = \sum_{y\in\mathcal{A}_{20}} P_{ay}^{(f,\mathcal{A}_{20})}(t)\log_2 P_{ay}^{(f,\mathcal{A}_{20})}(t).$$

Knowing that a specific amino acid was transmitted, how much of this knowledge is lost with increasing evolutionary time t? Since the comparison of 20 values over time is inconvenient, we apply uniform weighting denoted with $p_X(x) = \frac{1}{|20|}, \forall\, x \in \mathcal{A}_{20}$, which results in

$$H_t^{(f)}(Y|X) = \sum_{x\in\mathcal{A}_{20}} p_X(x)H_t^{(f)}(Y|X=x). \tag{4.39}$$

The corresponding output distribution per reading frame is given by

$$p_Y^{(f)}(y) = \sum_{x\in\mathcal{A}_{20}} p_X(x)P_{xy}^{(f,\mathcal{A}_{20})}(t). \tag{4.40}$$

Similarly, we could also use $\boldsymbol{\pi}^{(f,\mathcal{A}_{20})}$ as weighting distribution which results in the same output distribution, but since we do not want to compare biases in amino acid distributions in the individual frames, we choose uniform weighting. Note that (4.39) is bounded as follows

$$H_t^{(f)}(Y|X) \leq H_t^{(f)}(Y) \leq \log_2(20) = 4.32\,[\text{bit}], \tag{4.41}$$

with $H_t^{(f)}(Y)$ given in Definition 2.12 to

$$H_t^{(f)}(Y) = -\sum_{y\in\mathcal{A}_{20}} p_Y^{(f,\mathcal{A}_{20})}(y)\log_2 p_Y^{(f,\mathcal{A}_{20})}(y), \tag{4.42}$$

reaching its maximal value if the amino acids Y are uniformly distributed.

4.3.2. Sequence Similarity over Time

We use the amino acid evolution matrix $\boldsymbol{P}^{(f,\mathcal{A}_{20})}(t)$ determined above from $\boldsymbol{P}^{(f)}(t)$ and $\boldsymbol{\pi}^{(f)}$ applying (4.38) to measure the similarity between amino acid sequences. As suggested by Yockey [Yoc92], we use the mutual information defined between two random variables X, Y in Definition 2.15 calculating

$$I_t^{(f)}(X;Y) = H_t^{(f)}(Y) - H_t^{(f)}(Y|X), \tag{4.43}$$

from (4.39), (4.40) and (4.42). This corresponds to (2.9)

$$I_t^{(f)}(X;Y) = \sum_{y\in\mathcal{A}_{20}}\sum_{x\in\mathcal{A}_{20}} p_X(x)P_{xy}^{(f,\mathcal{A}_{20})}(t)\log_2\frac{P_{xy}^{(f,\mathcal{A}_{20})}(t)}{\sum_{k\in\mathcal{A}_{20}} p_X(k)P_{ky}^{(f,\mathcal{A}_{20})}(t)}, \tag{4.44}$$

where p_X is again the uniform distribution.
We know from Definition 2.17 that the channel capacity measures the amount of information that can be transmitted over a channel. Since the capacity of a discrete memoryless channel is the maximal mutual information over all input distributions, it determines in the scenario of evolution the largest reachable sequence similarity.
An iterative algorithm to determine the capacity numerically by finding the input distribution $\boldsymbol{p}^{(f)} = \{p_x^{(f)}\}_{x\in\mathcal{A}_{20}}$ that maximizes the mutual information for a given channel matrix $\boldsymbol{P}^{(f,\mathcal{A}_{20})}(t)$ is the Blahut-Arimoto Algorithm [Ari06]. The algorithm converges monotonically to capacity. A tutorial that describes the algorithm and its derivations in a readable manner is the basic for the following part [Mis09].
First, define the $|20| \times |20|$ a matrix $\boldsymbol{\phi}$

$$\boldsymbol{\phi} = [\phi_{yx}]_{\forall x,y\in\mathcal{A}_{20}} = \left[\frac{p_x^{(f)}P_{xy}^{(f,\mathcal{A}_{20})}(t)}{\sum_{k\in\mathcal{A}_{20}} p_k^{(f)}P_{ky}^{(f,\mathcal{A}_{20})}(t)}\right]_{\forall x,y\in\mathcal{A}_{20}}$$

and the generalized equivocation

$$J\left(\boldsymbol{P}^{(f,\mathcal{A}_{20})}(t), \boldsymbol{p}^{(f)}, \boldsymbol{\phi}\right) = -\sum_{x\in\mathcal{A}_{20}} p_x^{(f)}\sum_{y\in\mathcal{A}_{20}} P_{xy}^{(f,\mathcal{A}_{20})}(t)\log_2(\phi_{yx}).$$

This is used for another definition of capacity given by

$$\mathcal{C}(\boldsymbol{P}^{(f,\mathcal{A}_{20})}(t)) = \max_{p_x^{(f)}}\max_{\boldsymbol{\phi}\in\boldsymbol{\Phi}}\left[H\left(\boldsymbol{p}^{(f)}\right) - J\left(\boldsymbol{P}^{(f,\mathcal{A}_{20})}(t), \boldsymbol{p}^{(f)}, \boldsymbol{\phi}\right)\right],$$

where $\mathbf{\Phi}$ is the set of all possible $\boldsymbol{\phi}$ and the entropy is given according to Definition 2.12 as

$$H\left(\boldsymbol{p}^{(f)}\right) = -\sum_{x\in\mathcal{A}_{20}} p_x^{(f)} \log_2 p_x^{(f)}.$$

Then for each reading frame, we call procedure BLAHUT-ARIMOTO($\boldsymbol{P}^{(f,\mathcal{A}_{20})}(t)$) in Algorithm 1 to determine the corresponding channel capacity $\mathcal{C}^{(f)}$ per reading frame. When the algorithm starts with a uniform distribution, the estimation error $e(t_I)$ after iteration t_I is bounded by $e(t) < \frac{\log_2|\mathcal{A}_{20}| - H\left(\boldsymbol{p}^{(f)}\right)}{t_I}$, where $\boldsymbol{p}^{(f)}$ is the unknown probability distribution that achieves capacity.

Algorithm 1 Compute channel capacity per reading frame

procedure BLAHUT-ARIMOTO($P_{xy}(t)$)
  $t_I = 1$
  Initialize $p_x^{t_I} = \frac{1}{|\mathcal{A}_{20}|}\ \forall x \in \mathcal{A}_{20}$, $\mathcal{C}^{t_I-1} = 0$
  while true **do**
    Maximize $H\left(\boldsymbol{p}^{t_I}\right) - J\left(P_{xy}(t), \boldsymbol{p}^{t_I}, \boldsymbol{\phi}\right)$ with respect to $\boldsymbol{\phi} \in \mathbf{\Phi}$ by

$$\boldsymbol{\phi}^{t_I} \leftarrow \left[\frac{P_{xy}(t)\cdot p_x^{t_I}}{\sum_{k\in\mathcal{A}_{20}} P_{ky}(t) p_k^{t_I}}\right]_{\forall x,y\in\mathcal{A}_{20}}$$

    and compute capacity

$$\mathcal{C}^{t_I} \leftarrow H\left(\boldsymbol{p}^{t_I}\right) - J\left(P_{xy}(t), \boldsymbol{p}^{t_I}, \boldsymbol{\phi}^{t_I}\right)$$

    if $|\mathcal{C}^{t_I} - \mathcal{C}^{t_I-1}| < 10^{-8}$ **then**
      return $\mathcal{C}^{t_I}$
    end if
    Maximize $H\left(\boldsymbol{p}\right) - J\left(P_{xy}(t), \boldsymbol{p}, \boldsymbol{\phi}^{t_I}\right)$ with respect to $\boldsymbol{p}$ by

$$p_x^{t_I+1} \leftarrow \frac{r_x^{t_I}}{\sum_{k\in\mathcal{A}_{20}} r_k^{t_I}}, \quad \forall x \in \mathcal{A}_{20},$$

    where

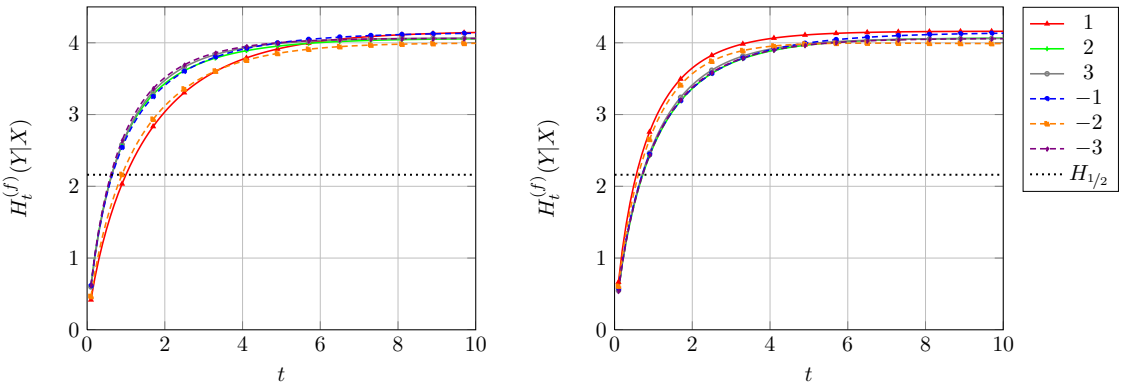
$$r_x^{t_I} = 2^{\sum_{y\in\mathcal{A}_{20}} P_{xy}(t)\cdot\log_2(\phi_{yx}^{t_I})}$$

    $t_I = t_I + 1$
    if $t_I > 10^3$ **then**
      return -1
    end if
  end while
end procedure

4.3.3. Simulation Results

We investigate the influence of genetic noise using a similar simulation scenario as in Section 4.1.3 or Section 4.2.3. The codon usage of EHEC is used as input stationary distribution $\boldsymbol{\pi}^{(+1)}$ to rate matrix $\boldsymbol{Q}^{(+1)}$. The transition/transversion rate is fixed to $\kappa = 1.0$ and the time t takes values in $[0.1, 0.2, \ldots, 10]$. Purifying selection is modelled by $\omega^{(+1)} = 0.3$ and adaptive selection by $\omega^{(+1)} = 3.0$.

Investigation of information loss During evolution, a loss of information with increasing time is expected. This corresponds to an increase of uncertainty, since the genetic noise distorts the information which amino acid was originally transmitted. A comparison, how long an amino acid is conserved in each reading frame, is obtained applying the conditional entropy (4.39) which leads to Figure 4.5.



(a) $\kappa = 1.0, \omega^{(+1)} = 0.3$ (b) $\kappa = 1.0, \omega^{(+1)} = 3.0$

Figure 4.5.: Information loss over time per reading frame results in an increasing uncertainty. The conditional entropy measures the genetic noise, that distorts the transmitted amino acid. In (a) reading frames $\{+1, -2\}$ are protected, as $\omega^{(+1)} = 0.3$ results in $\omega^{(-2)} < 1$. For (b) the opposite holds. Line $H_{1/2}$ marks the point, where half of the information is lost.

We quantify this information loss by determining the time that is needed to loose half of the transmitted information. According to (4.41) $H_t^{(f)}(Y|X)$ is bounded by $\log_2(20)$, such that we determine for each reading frame time $t_{1/2}^{(f)}$, when

$$H_t^{(f)}(Y|X) \geq \frac{\log_2(20)}{2} = H_{1/2}$$

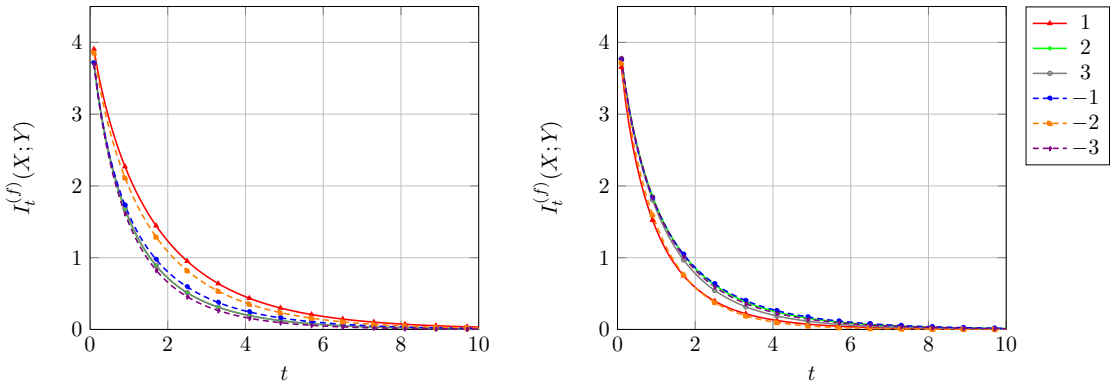
is reached. The results of this approximative information loss are summarized in Table 4.1.

$\omega^{(+1)}$	$t_{1/2}^{(f)}$					
	$+1$	$+2$	$+3$	-1	-2	-3
0.3	1.0	0.7	0.7	0.7	0.9	0.7
3.0	0.6	0.8	0.8	0.8	0.7	0.8

Table 4.1.: Time for each reading frame $t_{1/2}^{(f)}$, where half of the information is lost, such that $H_{1/2} = {}^{\log_2(20)}/_2$ is reached.

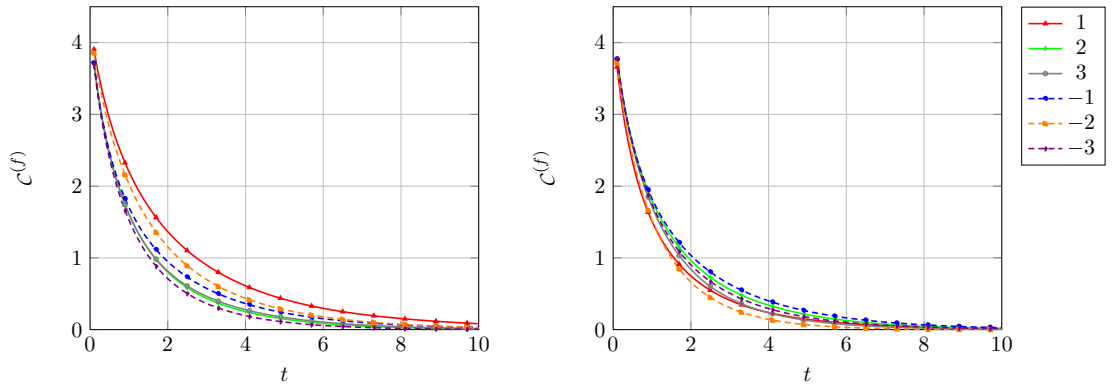
In view of the simulation results of the selection pressure $\omega^{(f)}$, especially Figure 4.3 on page 95, we interpret Figure 4.5 and Table 4.1 as follows: Reading frames, that are protected, meaning that $\omega^{(f)} < 1$ show a slower increasing uncertainty, i.e., they store information longer than unprotected frames, where $\omega^{(f)} > 1$. Detailed, in Figure 4.5(a), we set $\omega^{(+1)} = 0.3$ which leads to $\omega^{(-2)} < 1$ and remaining frames $\omega^{(f_r)} > 1$, $f_r \in \mathcal{F} \setminus \{+1, -2\}$. This corresponds to the order of time $t_{1/2}^{(f_r)} < t_{1/2}^{(-2)} < t_{1/2}^{(+1)}$. A protection of the protein coding reading frame corresponds to a protection of reading frame $\{-2\}$, resulting in slower loss of information than observed within the remaining reading frames. The opposite holds for Figure 4.5(b), where $\omega^{(+1)} = 3.0$ results in $\omega^{(-2)} > 1$ and $\omega^{(f_r)} < 1$. The time in Table 4.1 is ordered accordingly $t_{1/2}^{(+1)} < t_{1/2}^{(-2)} < t_{1/2}^{(f_r)}$.

Investigation of sequence similarity The amount of information, that can be transmitted over an evolution channel, is directly connected to the mutual information. We measure the similarity between the amino acid sequence before evolution and the resulting sequence after evolution by estimation of the mutual information $I_t^{(f)}(X;Y)$ per reading frame $f \in \mathcal{F}$ between input variable X and output variable Y applying (4.43). The results are presented in Figure 4.6.
The results obtained have to be interpreted, again, in relation to the selection pressure $\omega^{(f)}$. In Figure 4.6(a), we set $\omega^{(+1)} = 0.3$ resulting in $\omega^{(-2)} < 1$ and $\omega^{(f_r)} > 1$, $f_r \in \mathcal{F} \setminus \{+1, -2\}$. Although the sequence similarity decreases as evolutionary time passes in all reading frames, most information is transmitted in protein coding frame $\{+1\}$ followed by $\{-2\}$. For the remaining frames, the mutual information decreases faster. In Figure 4.6(b), where $\omega^{(+1)} = 3.0$ is set, we obtain the fastest decreasing sequence similarity in frames $\{+1, -2\}$, while the remaining frames are protected with $\omega^{(f_r)} < 1$ which corresponds to a slower decreasing mutual information.
But what is the maximal amount of information that can be transmitted over the evolution channel per reading frame? We deal with this question in Figure 4.7, applying the Blahut-Arimoto Algorithm 1 to determine the channel capacity $\mathcal{C}^{(f)}$ per reading frame.
In Table 4.2 on page 104 we see the comparison of the ratio $\omega^{(f)}$, the mutual information $I_t^{(f)}(X;Y)$ and the capacity $\mathcal{C}^{(f)}$, for input parameter set $\kappa = t = 1.0$ and $\omega^{(+1)} = 0.3$ respectively $\omega^{(+1)} = 3.0$. For each reading frame the absolute difference between the mutual information and the capacity is denoted as Δ in Table 4.2. The channel capacity shows the largest amount of information, that can be transmitted over the channel. We observe that



(a) $\kappa = 1.0, \omega^{(+1)} = 0.3$ (b) $\kappa = 1.0, \omega^{(+1)} = 3.0$

Figure 4.6.: Sequence similarity per reading frame decreases as time passes. The mutual information for a uniform input distribution over amino acids measures the amount of information that can be transmitted over the evolution channel. In (a) reading frames $\{+1, -2\}$ are protected, as $\omega^{(+1)} = 0.3$ results in $\omega^{(-2)} < 1$. For (b) the opposite holds.



(a) $\kappa = 1.0, \omega^{(+1)} = 0.3$ (b) $\kappa = 1.0, \omega^{(+1)} = 3.0$

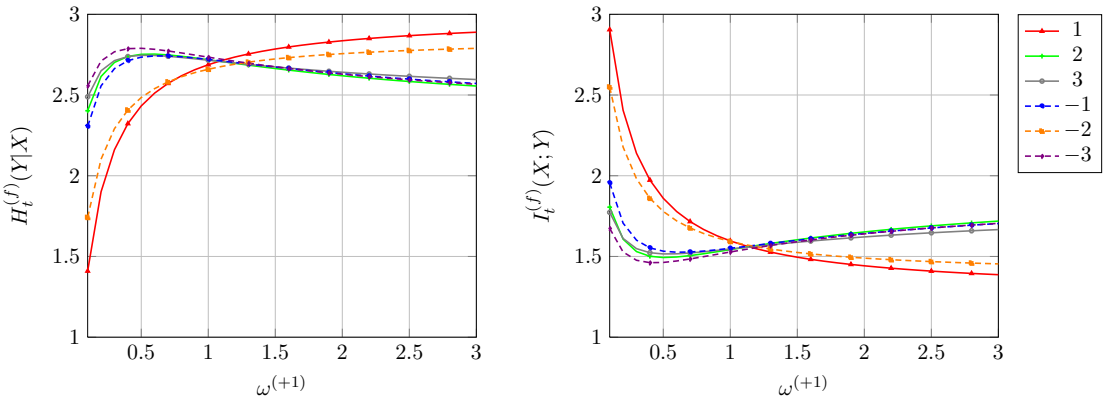
Figure 4.7.: Channel capacity over time per reading frame decreases. The maximal amount of information, that can be transmitted over the evolution channel is measured for protected frames $\{+1, -2\}$ in (a) and for protected frames $\{-1, +2, \pm 3\}$ in (b).

for $\omega^{(+1)} = 0.3$, the largest mutual information is maintained in reading frame $\{+1\}$, where also the channel capacity is largest. Reading frame $\{-2\}$ maximizes its potential, whereas $\{-1\}$, has largest distance to the channel optimum. For $\omega^{(+1)} = 3.0$ reading frames $\{+1, -2\}$ prefer nonsynonymous codon substitutions, which results in the smallest channel capacity and the fastest loss of information.

	+1	+2	+3	−1	−2	−3
$\omega^{(f)}$	0.3000	1.3880	1.5446	1.3441	0.5256	1.5633
$I_t^{(f)}(X;Y)$	2.1391	1.5303	1.5483	1.6009	1.9811	1.4764
$\mathcal{C}^{(f)}$	2.2046	1.6175	1.6085	1.7054	2.0304	1.5285
Δ	0.0655	0.0872	0.0602	0.1045	0.0493	0.0521
$\omega^{(f)}$	3.0000	0.7988	0.8117	0.8744	1.8977	0.7892
$I_t^{(f)}(X;Y)$	1.3871	1.7190	1.6664	1.7040	1.4540	1.7050
$\mathcal{C}^{(f)}$	1.5111	1.7971	1.7122	1.8268	1.5278	1.7475
Δ	0.1240	0.0781	0.0458	0.1228	0.0738	0.0425

Table 4.2.: Comparison of selection pressure, mutual information and capacity with respect to different scenarios, where $\omega^{+1} = 0.3$ results in $\omega^{-2} < 1$ and $\omega^{+1} = 3.0$ results in $\omega^{-2} > 1$. Parameters $\kappa = 1.0$ and $t = 1.0$ are fixed.

Comparison of conditional entropy and mutual information In order to investigate, if the results obtained above hold for different values of the selection pressure, we set $t = 1.0$ and $\kappa = 1.0$, while the nonsynonymous/synonymous rate ratio $\omega^{(+1)}$ ranges from $[0.1, 3.0]$. The conditional entropy in Figure 4.8(a) and the mutual information Figure 4.8(b) reflects the influence of the selection pressure within the different reading frames, investigated in Figure 4.3.



(a) $\kappa = 1.0, t = 1.0$ (b) $\kappa = 1.0, t = 1.0$

Figure 4.8.: Conditional entropy (a) and mutual information (b) over different values of nonsynonymous/synonymous rate ratio $\omega^{(+1)}$. Approximately, $\omega^{(+1)} < 1$ results in $\omega^{(-2)} < 1$ and $\omega^{(f_r)} > 1$, for $f_r \in \mathcal{F} \setminus \{+1, -2\}$. In contrast $\omega^{(+1)} > 1$ reveals $\omega^{(-2)} > 1$ and $\omega^{(f_r)} < 1$.

4.4. Influence of Genetic Code

The protection of reading frame $\{+1\}$ together with $\{-2\}$ can be attributed to a simple property of the genetic code. In fact, the first and second codon positions have the largest influence on the encoded amino acid and they fall onto the second respectively first codon position in the $\{-2\}$ reading frame. This indicates that a conservation of a sequence in reading frame $\{-2\}$ might be solely an artefact, providing not necessarily evidence for functionality.

Influence of synonymous codon replacements For a detailed analysis of this property of the genetic code, we considered all tuples of $\mathcal{C}_{64}^3$ possible codon combinations of three adjacent codons $C_{j-1}^{+1}C_j^{+1}C_{j+1}^{+1}$ in reading frame $\{+1\}$. Each codon corresponds to an unique amino acid, such that in each reading frame a specific amino acid sequence is obtained. In frames $\{\pm 2, \pm 3\}$ two amino acids are given, and in frames $\{\pm 1\}$ three amino acids.
We replace the codon that is located in the middle C_j^{+1} by all synonymous codons, such that the amino acid in $\{+1\}$ is conserved. Codons C_{j-1}^{+1} and C_{j+1}^{+1} are fixed. Now, we determine in each alternative reading frame $f \in \mathcal{F}_A$, if the original amino acid sequence that was induced from $\{+1\}$ changes, answering the question, if a synonymous replacement in $\{+1\}$ leads to a synonymous replacement within a certain frame. The number of point mutations between two synonymous codons lies between $[1, 3]$, such that a replacement of C_j^{+1} influences one or two codons in each alternative reading frame, depending on the number of point mutations. For example if CTC is replaced by TTG, the first and last nucleotide changes, such that at most two amino acids can be changed within each reading frame except $\{-1\}$, whereas replacement of CTC by CTT has influence on only one amino acid in each alternative reading frame. For each tuple and each alternative reading frame $f \in \mathcal{F}_A$, we determine the number of synonymous replacements in $\{+1\}$ that did not change the amino acid in f as well as the number of amino acids in f that could have changed. Afterwards we summarize all tuples, that have the same codon replacement in the middle position C_j^{+1}.
Now we have: The original codon in reading frame $\{+1\}$ denoted with $c^{ORG} \in \mathcal{C}_{64}$ that is a realization of C_j^{+1} and the corresponding synonymous codons $\boldsymbol{c}^{SYN} = \left(c_1^{SYN}, c_2^{SYN}, \ldots\right)$. We know for each replacement in each frame $f \in \mathcal{F}_A$, the number of unchanged amino acids $N_{SYN}^{(f)}(c^{ORG}, c_i^{SYN})$, $i \in \{1, 2, \ldots\}$ as well as the number of amino acid that could have been changed $N^{(f)}(c^{ORG}, c_i^{SYN})$, $i \in \{1, 2, \ldots\}$. SYN means here, that the synonymous codon replacement in $\{+1\}$ leads to a synonymous codon replacement in f and ORG refers to the original codon. We summarize the relative frequency of synonymous replacements over all codon pairs by

$$h_{SYN}^{(f)} = \frac{\sum_{c^{ORG} \in \mathcal{C}_{64}} \sum_{c_i^{SYN} \in \boldsymbol{c}^{SYN}} N_{SYN}^{(f)}(c^{ORG}, c_i^{SYN})}{\sum_{c^{ORG} \in \mathcal{C}_{64}} \sum_{c_i^{SYN} \in \boldsymbol{c}^{SYN}} N^{(f)}(c^{ORG}, c_i^{SYN}}, \tag{4.45}$$

where $c^{ORG} \cup \boldsymbol{c}^{SYN}$ covers all codons that belong to a certain amino acid. The results are summarized in Table 4.3 on page 106. for each alternative reading frame. We observe that

the probability to preserve the amino acid in a certain frame under synonymous codon replacements in reading frame $\{+1\}$ is by far the largest in $\{-2\}$ with 54% followed by $\{+2\}$ with only 17%.

	$+2$	$+3$	-1	-2	-3
$h_{SYN}^{(f)}$	0.17	0.05	0.09	0.54	0.02

Table 4.3.: Influence on amino acid within alternative reading frame under synonymous codon replacements in reading frame $\{+1\}$.

Each amino acid $a \in \mathcal{A}_{21}$ can be individually interpreted, denoted as $h_{SYN}^{(f)}(a)$. For example *Tyr* has only two codons *TAT* and *TAC*, if we replace them the amino acid in frames $\{+2, -1, -3\}$ is always changed, while the probability that the amino acid is preserved in $\{+3\}$ is 12.5% and in $\{-2\}$ 87.5%. Comparable results are obtained for all amino acids, that contain only two codons, i.e., *Phe*, *Cys*, *Asn*, *Asp*, *Glu*, *His*, *Lys*, where the probability in $\{-2\}$ is always $> 80\%$, except *Gln* where $\{-1, -2\}$ are both 100%. Representative for amino acids with more than two codons, we show the distributions of *Pro* that has 4 codons in Figure 4.9 and of *Ser* with 6 codons in Figure 4.10 on page 108. A clear deviation between reading frame $\{-2\}$ and $\{-1, +2, \pm 3\}$ is observable, which also holds for many other amino acids.

Degree of freedom regarding codon choice The property of the genetic code to preserve automatically the amino acid in reading frame $\{-2\}$ entails a severe drawback. It is challenging to embed information in the overlapping reading frame $\{-2\}$, when the protein-coding sequence in $\{+1\}$ consists of fixed amino acids. In order to quantify this effect, we proceed as follows. Assume two adjacent amino acids $a_1, a_2 \in \mathcal{A}_{21}$ should be encoded in $\{+1\}$. Obviously each amino acid corresponds to an individual number of codons that can be freely chosen without changing a_1 and a_2. By harnessing this effect, it is possible to change the encoded amino acid that is induced in an alternative reading frame. We count for each amino acid pair $a_1, a_2 \in \mathcal{A}_{21}$ how many different amino acids can be freely chosen within each alternative reading frame $\mathcal{F}_A$ and take the average over all 21^2 pairs to determine the degree of freedom $h_{DOF}^{(f)}$ presented in Table 4.4.

	$+2$	$+3$	-1	-2	-3
$h_{DOF}^{(f)}$	2.94	2.93	2.67	1.59	3.12

Table 4.4.: Degree of freedom to choose an amino acid within alternative reading frames under fixed amino acids in reading frame $\{+1\}$.

It turns out that the number of amino acids that can be freely chosen is smallest in $\{-2\}$ with an average of 1.59 amino acids and largest in reading frame $\{-3\}$ with an average of 3.12 amino acids. It is worth noting that in general it is possible to embed information even

in protein-coding sequences, e.g., [HB13]. The investigations above leads to the conjecture that it is difficult to encode information in reading frame $\{-2\}$, but once information is embedded it is most easily protected.

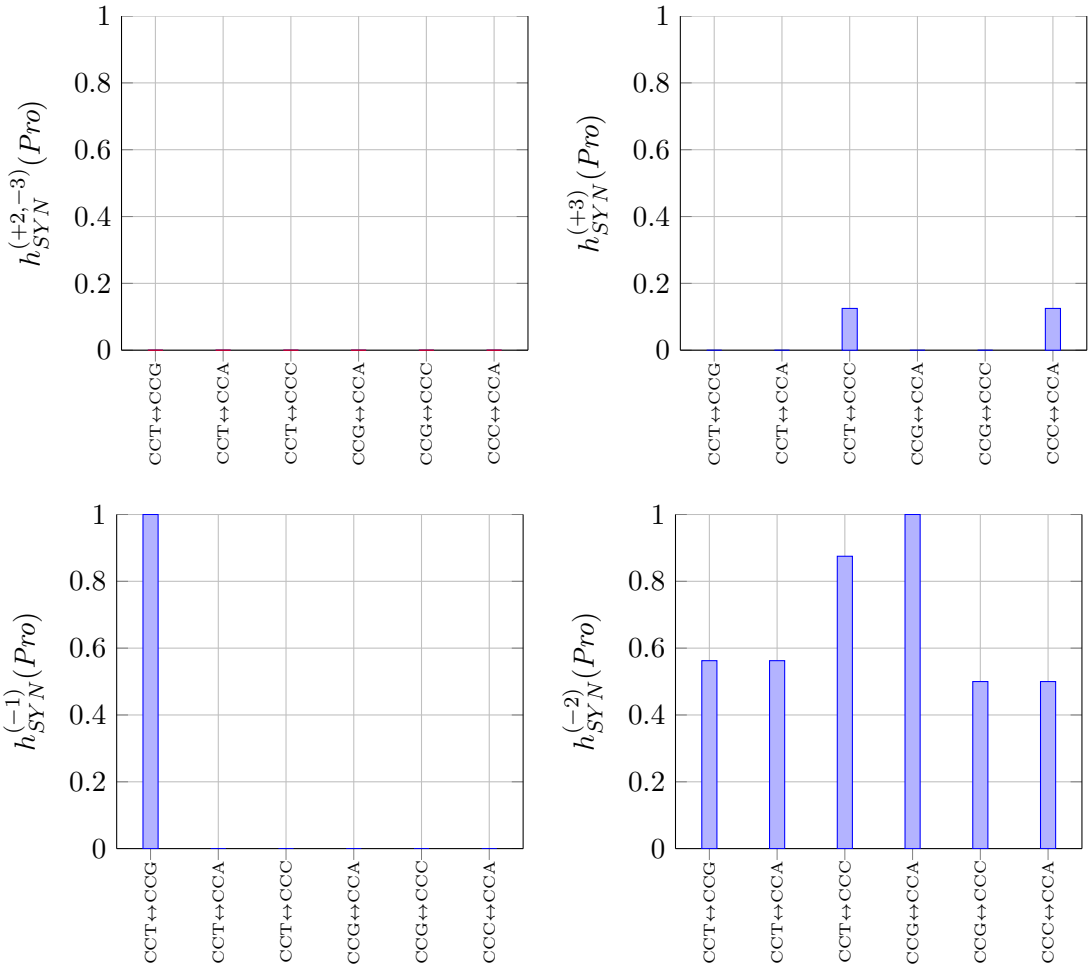


Figure 4.9.: Probability to perverse amino acids in alternative reading frames, if codons of amino acid *Pro* are replaced in reading frame $\{+1\}$. Note frame $\{+2, -3\}$ are consolidated, since both frames do not have synonymous replacements for *Pro*.

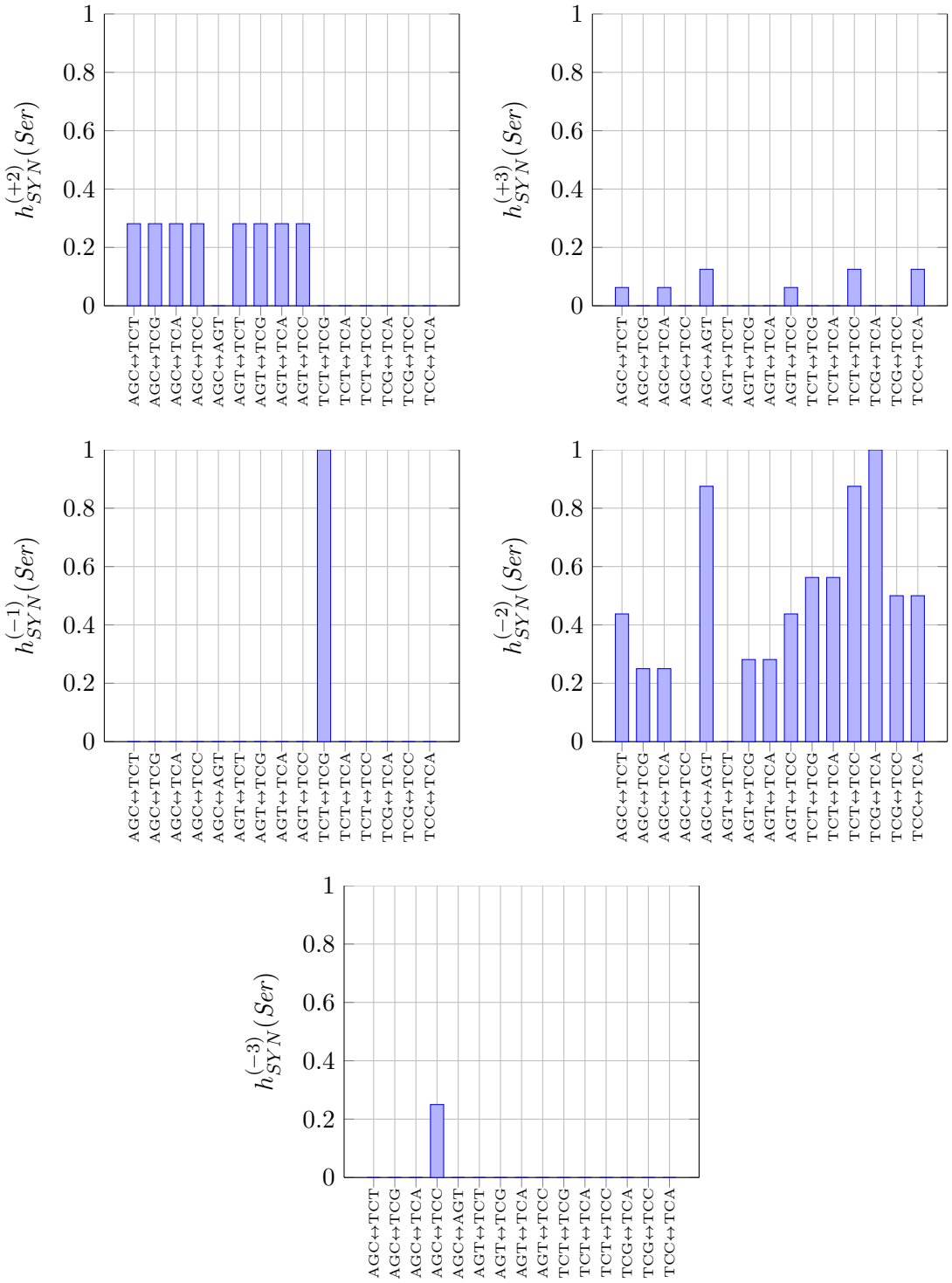


Figure 4.10.: Probability to perverse amino acids in alternative reading frames, if codons of amino acid *Ser* are replaced in reading frame {+1}.

4.5. Summary

In this chapter, we investigated information transmission over evolutionary time. We introduced a model to determine the codon evolution in different reading frames based on the protein-coding reading frame $\{+1\}$. Understanding the evolution of overlapping protein-coding regions is extremely challenging, since one DNA segment codes for two proteins which are translated in different reading frames simultaneously. Therefore, a mutation at any position affects both proteins [Kra00, MY80]. Evolutionary adaption of proteins is studied under the assumption that adaption requires more nonsynonymous replacements, i.e., nucleotide mutations at positions where the amino acid is changed, than synonymous replacements, i.e., mutations that preserve the amino acid [KP08]. This effect is quantified by the the nonsynonymous/synonymous substitution rate, which is an indicator for selective pressure within protein coding sequences. Our point of interest was slightly different: Assuming we are given a gene that evolves over time, we are interested in the evolutionary constraints implied within alternative reading frames. We applied our model to predicted the selection pressure in each reading frame, gaining interesting insights how the reading frames are connected and how they are mutually influenced. Our investigation reveals that a purifying selection in reading frame $\{+1\}$, which is a protection of the protein coding reading frame, is combined with a preserved sequence in reading frame $\{-2\}$. Meanwhile, in the remaining frames nonsynonymouus substitutions are preferred. On the other hand, if $\{+1\}$ is exposed to adaptive selection, the sequence in reading frame $\{-2\}$ is also free to evolve, while in the remaining frames the sequences are protected. The obtained results turned out to be robust applying different estimation methods for selection pressure.
We confirmed our findings by investigations of information theoretic quantities, that are used to analyze the evolution channels in each reading frame. We obtain that in case of purifying selection of the protein coding reading frame, the genetic noise is minimal in frames $\{+1, -2\}$ and the sequence similarity before and after evolution is largest. Further, those frames transmit the largest amount of information. Converse results are obtained for adaptive selection in $\{+1, -2\}$. In fact, over time the divergence of a sequence increases for all parameters and in all reading frames, but we showed that this change happens slower in case of purifying selection. It should be mentioned, that a protection of the reading frames $\{+1, -2\}$, does not imply that both sequences are translated into proteins, since functionality could also evolve later. Independent of the question, how or if an overlapping gene pair evolves, our observations indicate the special role of the $\{-2\}$ reading frame in combination with the protein coding frame. Interestingly, two recently experimentally verified examples of overlapping gene pairs in prokaryotes, *yaaW/htga* by [FBW+14] and *dmdR1/adm* by [TBCM09], are located in frame $\{-2\}$. On the one hand, we showed that it is theoretically possible to protect this pair by simply controlling the selection pressure within the protein-coding reading frame. But on the other hand, our investigations regarding the properties of the genetic code indicates, that the evolution of those two overlapping genes might be an artefact. Nevertheless, from our point of view these findings do not doubt the existence or importance of overlapping genes in prokaryotes, but raises the question if - in some cases - their evolution and protection can be attributed to a coincidence.

CHAPTER 5

Concluding Remarks

WITHIN THIS THESIS, we studied aspects of overlapping genes and their evolution. Since the existence of overlapping genes in prokaryotes was doubted for a long time, and the number of confirmed candidates increases slowly, our group in cooperation with molecular biologists, decided to encourage the process of finding new overlapping genes in prokaryotes. Since the number of overlapping gene candidates, that can be found simply by searching for all overlapping Open Reading Frames (ORFs), is enormous, it is necessary to get a deeper understanding of the theory of overlapping genes. The main goal of our cooperation is to predict candidates that justify the expensive and time-consuming effort of experimental verification. The functionality should be found in the laboratory, by knock-out experiments. Here, both potential overlapping genes are inactivated independently by introduction of a stop codon, a procedure that sounds simple, but isn't. Afterwards the biologist has to find out, what influence the inactivation of the gene had and under which environmental conditions a change is observed, in order to determine its function. This step is comparable to search for a needle in a haystack, which makes good candidates mandatory.

In Chapter 3 the statistical properties of bacterial genomes were investigated. We derived an analytical model based on the codon composition of genes and the genome length. The model was applied to the genome of *EHEC* presenting a detailed analysis of statistics like length distribution, number and average length of ORFs. Further, we compared the bacterial genomes of 71 organisms, showing that the model can be applied over a wide range of GC contents. Surprisingly, the predictions obtained from the model, were in large parts in excellent agreement with the natural genomes, although the model is based on a random process. We applied the model to overlapping genes, where an annotated ORF (aORF) and a shadow ORF (sORF) are located in the same DNA region, investigating the behaviour of sORFs in different reading frames. Deviations of the sORF length distributions from the values expected from the model indicate that long sORFs can not be explained from a random process, pointing towards functionality. Those sORFs are overlapping gene candidates, that should be further analyzed. It turned out, that depending on the reading frames, different sORF lengths are significant. It should be mentioned, that starting from the pure genome

database and the positions of the annotated genes, we did not depend on existing software, such that each step of data analysis was carefully considered. This is necessary, since from an engineering point of view in the biological environment many assumptions are vague or implicitly specified, such that we invested a lot of time in building a reliable simulation framework.

Chapter 4 deals with information transmission over evolutionary time. We analyzed a model of codon evolution to predict the codon substitution channels that are implied in each reading frame. The amount of information that can be transmitted over time was quantified by application of information theoretic measures. The obtained channel matrices can be used, to determine the selection pressure based on the protein-coding reading frame. The results indicate, that a simultaneous protection of an overlapping gene pair is theoretically possible, but not inevitably based on a "conscious decision" that in both genes a functionality should be sustained.

The results presented in this thesis, give interesting insights into statistical aspects of overlapping genes. However, it is a first step in the direction of understanding the theory behind overlapping genes. Our findings have been already applied to select some good candidates and now it's up to the biologists to evaluate their quality.

APPENDIX A

Details regarding Markov Chains and Processes

A.1. Proof of Theorem 2.1

We sketch the proof based on [Fel68], page 391. Let i and j be two arbitrary states of an irreducible Markov chain, i.e., there exists a $k \geq 0$ such that $P_{ij}^{(k)} > 0$ for any $i, j \in \mathcal{S}$. Suppose there exist integers k, l such that $\alpha = P_{ij}^{(k)} > 0$ and $\beta = P_{ji}^{(l)} > 0$. Therefore it holds that $P_{ii}^{(k+l)} > \alpha\beta > 0$, which means that $k + l$ is a multiple of the period d_i of state i. Suppose that m is an arbitrary value that satisfies $P_{jj}^{(m)} > 0$. Then the probability to go from i to j return to j in m steps and return to i is given by $P_{ii}^{(k+m+l)} > 0$ which leads to $k + l + m$ being a multiple of d_i and also m being a multiple of d_i. As $m \geq d_j$, the period of state j is automatically $d_j \geq d_i$. Interchanging the roles of i and j leads to $d_i \geq d_j$, which concludes the proof that $d_i = d_j$ holds.

A.2. Proof of Theorem 2.2

For the proof we need additional terms and theorems: First, we have to introduce the additional term of a spectral radius (pages 283, 490, 497 [Mey01]).
For a square matrix $\boldsymbol{A}$, the number $\rho(\boldsymbol{A}) = \max_{\lambda \in \sigma(\boldsymbol{A})} |\lambda|$ is called the spectral radius, where $\sigma(\boldsymbol{A})$ is the set of distinct eigenvalues of $\boldsymbol{A}$. The spectral radius is bounded by the matrix norm $\rho(\boldsymbol{A}) \leq ||\boldsymbol{A}||$, where $||\boldsymbol{A}||$ is the largest absolute row sum.
According to the Perron-Frobenius Theorem for non-negative irreducible matrices $\boldsymbol{A}$ with dimension $n \times n$, the following statements are true (page 673 in [Mey01]):

- There exists a positive eigenvalue $\lambda_{\max} > 0$, such that all other eigenvalues of matrix $\boldsymbol{A}$ are $|\lambda_i| \leq \lambda_{max}, \forall \lambda_i \neq \lambda_{\max}$, i.e., $\lambda_{\max} = \rho(\boldsymbol{A}) \in \sigma(\boldsymbol{A})$.
- The eigenvalue $\lambda_{\max}$ has algebraic multiplicity one.
- An eigenvector exists $\boldsymbol{x}^T > 0$ such that $\boldsymbol{A}\boldsymbol{x}^T = \lambda_{\max}\boldsymbol{x}^T$ is fulfilled.

- The unique vector defined by $\boldsymbol{A}\boldsymbol{p}^T = \lambda_{\max}\boldsymbol{p}^T$ with $\boldsymbol{p}^T > 0$, where $||\boldsymbol{p}^T||_1 = \sum_{i=1}^{n} |p_i| = 1$ holds, is called the Perron vector. There are no non-negative eigenvectors for $\boldsymbol{A}$ except for positive multiples of $\boldsymbol{p}^T$, regardless of the eigenvalue.

If there is only one eigenvalue $\lambda_{\max}$ on the spectral circle with radius $\rho(\boldsymbol{A})$, $\boldsymbol{A}$ is called a primitive matrix (page 674 in [Mey01]). A non-negative irreducible matrix $\boldsymbol{A}$ with $\lambda_{\max} = \rho(\boldsymbol{A})$ is primitive if and only if $\lim_{k\to\infty} (\boldsymbol{A}/\lambda_{\max})^k$ exists, in which case

$$\lim_{k\to\infty} \left(\frac{\boldsymbol{A}}{\lambda_{\max}}\right)^k = \frac{\boldsymbol{p}^T\boldsymbol{q}}{\boldsymbol{q}\boldsymbol{p}^T} > \boldsymbol{0}, \tag{A.1}$$

where $\boldsymbol{p}^T$ is the Perron vector of $\boldsymbol{A}$ and $\boldsymbol{q}^T$ the Perron vector of $\boldsymbol{A}^T$. Further, note that an irreducible Markov chain, for which the corresponding transition matrix is primitive, is called an aperiodic chain (page 694 in [Mey01]).

Those findings can now be applied to the irreducible aperiodic non-negative transition matrix $\boldsymbol{P}$. From Definition 2.21, we know that the transition matrix $\boldsymbol{P}$ is a stochastic matrix that fulfills $\sum_j P_{ij} = 1 \; \forall i, j \in \mathcal{S}$, which is equivalent to $\boldsymbol{P}\boldsymbol{e}^T = \boldsymbol{e}^T$, where $\boldsymbol{e}^T$ is the column vector of all ones that corresponds to an eigenvalue of one. Since here $\rho(\boldsymbol{P}) \geq 1$ and as the spectral radius is bounded by the largest absolute row sum $\rho(\boldsymbol{P}) \leq ||\boldsymbol{P}|| = 1$ it follows that $\lambda_{\max} = 1$ is the largest eigenvalue of the stochastic matrix. Further, since $\boldsymbol{P}$ is primitive, there is only one eigenvalue on the spectral circle $\rho(\boldsymbol{P})$.
The corresponding Perron vector is the uniform distribution $\boldsymbol{e}^T/n$, that fulfills $||\boldsymbol{e}^T/n||_1 = 1$. We denote the Perron vector of $\boldsymbol{P}^T$ with $\boldsymbol{\pi}^T$ and apply (A.1)

$$\lim_{k\to\infty} \boldsymbol{P}^k = \frac{\left(\boldsymbol{e}^T/n\right)\boldsymbol{\pi}}{\boldsymbol{\pi}\left(\boldsymbol{e}^T/n\right)} = \frac{\boldsymbol{e}^T\boldsymbol{\pi}}{\boldsymbol{\pi}\boldsymbol{e}^T} = \boldsymbol{e}^T\boldsymbol{\pi} = \begin{pmatrix} \boldsymbol{\pi} \\ \boldsymbol{\pi} \\ \vdots \\ \boldsymbol{\pi} \end{pmatrix} > \boldsymbol{0}.$$

From (2.11)

$$\begin{aligned} \boldsymbol{p}^{(k)} &= \boldsymbol{p}^{(0)}\boldsymbol{P}^k \\ \lim_{k\to\infty} \boldsymbol{p}^{(k)} &= \boldsymbol{p}^{(0)} \lim_{k\to\infty} \boldsymbol{P}^k = \underbrace{\boldsymbol{p}^{(0)}\boldsymbol{e}^T}_{\sum_i p_i^{(0)}=1}\boldsymbol{\pi} = \boldsymbol{\pi}, \end{aligned}$$

which concludes the proof. The derivation follows [Mey01] pages 689 and 691.

A.3. Solve Matrix Exponential

This derivation is well-known. There are numerous lecture courses and presentations and papers dealing with this problem, e.g., [MVL03]. Given a homogeneous differential equation with scalar a

$$x'(t) = ax(t)$$

and initial condition $x(0) = 1$, by integration the solution is $x(t) = e^{at}$ with Taylor series expansion

$$e^{at} = \sum_{k=0}^{\infty} \frac{at^k}{k!}$$

In case of continuous-time Markov chains, we replace the scalar by the square rate matrix $\boldsymbol{Q}$

$$e^{\boldsymbol{Q}t} = \sum_{k=0}^{\infty} \frac{t^k}{k!}\boldsymbol{Q}^k = I + t\boldsymbol{Q} + \frac{t^2}{2}\boldsymbol{Q}^2 + \frac{t^3}{6}\boldsymbol{Q}^3 + \cdots$$

The series converges for all square matrices $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ with finite state space for any t.

We consider the eigenvalue decomposition of the rate matrix, that exists if $\boldsymbol{Q}$ has distinct eigenvalues, according to: Theorem 7.2.5 on page 512 in [Mey01] shows that a square matrix is diagonalizable if and only if for each eigenvalue of the matrix the geometric multiplicity equals the algebraic multiplicity. Further from Theorem 7.2.2 on page 511 in [Mey01] it follows, that a simple eigenvalue, i.e., where algebraic multiplicity is one, has geometric multiplicity of one. If a $n \times n$ matrix has n distinct eigenvalues, then each eigenvalue is simple, leading to Corollary 7.2.6 on page 514 [Mey01] that states, that a matrix is diagonizable if no eigenvalue of the matrix is repeated.

Under the assumption of distinct eigenvalues, we write $\boldsymbol{Q} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{-1}$, where $\boldsymbol{U} = \left(\boldsymbol{u}_1^T, \ldots, \boldsymbol{u}_{|\mathcal{S}|}^T\right)$ denotes the matrix of right eigenvectors and $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_{|\mathcal{S}|})$ the diagonal matrix with the distinct eigenvalues of $\boldsymbol{Q}$.

$$e^{\boldsymbol{Q}t} = \sum_{k=0}^{\infty} \frac{t^k}{k!}\boldsymbol{Q}^k = \sum_{k=0}^{\infty} \frac{t^k}{k!}(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{-1})^k$$

With property of decomposition

$$(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{-1})^k = (\boldsymbol{U}\boldsymbol{\Lambda}\underbrace{\boldsymbol{U}^{-1})(\boldsymbol{U}}_{\boldsymbol{I}}\boldsymbol{\Lambda}\boldsymbol{U}^{-1})\ldots(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{-1}) = \boldsymbol{U}\boldsymbol{\Lambda}^k\boldsymbol{U}^{-1}$$

we get

$$e^{\boldsymbol{Q}t} = \boldsymbol{U}\left(\sum_{k=0}^{\infty} \frac{t^k}{k!}\boldsymbol{\Lambda}^k\right)\boldsymbol{U}^{-1} = \boldsymbol{U}e^{\boldsymbol{\Lambda}t}\boldsymbol{U}^{-1}$$

with

$$e^{\boldsymbol{\Lambda}t} = \begin{pmatrix} e^{\lambda_1 t} & & & \\ & e^{\lambda_2 t} & & \\ & & \ddots & \\ & & & e^{\lambda_{|\mathcal{S}|} t} \end{pmatrix}$$

hence

$$\boldsymbol{P}(t) = e^{\boldsymbol{Q}t} = \boldsymbol{U}e^{\boldsymbol{\Lambda}t}\boldsymbol{U}^{-1}$$

concludes the derivation.

APPENDIX B

Details regarding Statistical ORF Properties

B.1. Statistical Parameters of Geometric Distribution

The infinite geometric series

$$\sum_{k=0}^{\infty} q^k = \frac{1}{1-q} \qquad |q| < 1, \tag{B.1}$$

as well as the partial sum

$$\sum_{k=0}^{n} q^k = \frac{1-q^{n+1}}{1-q} \qquad |q| < 1. \tag{B.2}$$

are common knowledge.

Geometric Series (B.1) applied in (3.7) Starting from

$$P(L \geq \ell') = p_{stop} \sum_{\ell=\ell'}^{\infty} (1-p_{stop})^{\ell-1} = p_{stop} (1-p_{stop})^{\ell'-1} \underbrace{\sum_{\ell=0}^{\infty} (1-p_{stop})^{\ell}}_{\frac{1}{1-(1-p_{stop})}=\frac{1}{p_{stop}}} = (1-p_{stop})^{\ell'-1}.$$

Expected value of geometric distribution (3.2) From the definition of expectation Definition 2.5 and the ORF length distribution in (3.1) we determine

$$\begin{aligned}\mathbf{E}[L] &= \sum_{\ell=1}^{\infty} \ell \cdot P(L=\ell) = \sum_{\ell=1}^{\infty} \ell \cdot (1-p_{stop})^{\ell-1} p_{stop} = p_{stop} \sum_{\ell=1}^{\infty} \ell \cdot (1-p_{stop})^{\ell-1} \\ &= p_{stop} \sum_{\ell=0}^{\infty} \ell \cdot (1-p_{stop})^{\ell-1} = p_{stop} \sum_{\ell=0}^{\infty} \left(\frac{\mathrm{d}}{\mathrm{d}(1-p_{stop})} (1-p_{stop})^{\ell} \right) \\ &= p_{stop} \frac{\mathrm{d}}{\mathrm{d}(1-p_{stop})} \underbrace{\sum_{\ell=0}^{\infty} (1-p_{stop})^{\ell}}_{(B.1)} = p_{stop} \underbrace{\frac{\mathrm{d}}{\mathrm{d}(1-p_{stop})} \left(\frac{1}{1-(1-p_{stop})} \right)}_{\frac{1}{(p_{stop})^2}} = \frac{1}{p_{stop}}.\end{aligned}$$

Cumulative Distribution Function of Geometric Distribution From the definition of the cumulative distribution function (CDF) in Definition 2.4 and the ORF length distribution in (3.1) we determine

$$\begin{aligned}\mathrm{P}(L \leq \ell) &= \sum_{k=1}^{\ell} (1-p_{stop})^{k-1} p_{stop} = p_{stop} \sum_{k=1}^{\ell} (1-p_{stop})^{k-1} \\ &= p_{stop} \underbrace{\sum_{k=0}^{\ell-1} (1-p_{stop})^{k}}_{\frac{1-(1-p_{stop})^{\ell}}{1-(1-p_{stop})}} = 1-(1-p_{stop})^{\ell}.\end{aligned}$$

B.2. Statistical Parameters of Weighted Geometric Distribution

Expected value of weighted geometric distribution (3.19) Extension of expected value to the weighted case applying the results of Appendix B.1 together with

$$p^{(f)} = \frac{n_{ORF}^{(f)}}{\sum_{j\in\mathcal{F}} n_{ORF}^{(j)}}$$

and

$$n_{ORF}^{(f)} = N_C \cdot p_{stop}^{(f)} \cdot \frac{p_{start}^{(f)}}{p_{start}^{(f)} + p_{stop}^{(f)}}$$

we determine

$$\begin{aligned}
\mathbf{E}[L] &= \sum_{\ell=1}^{\infty} \ell \sum_{f\in\mathcal{F}} p^{(f)} \left(1 - p_{stop}^{(f)}\right)^{\ell-1} \cdot p_{stop}^{(f)} \\
&= \sum_{f\in\mathcal{F}} p^{(f)} \sum_{\ell=1}^{\infty} \ell \left(1 - p_{stop}^{(f)}\right)^{\ell-1} \cdot p_{stop}^{(f)} \\
&= \sum_{f\in\mathcal{F}} p^{(f)} \frac{1}{p_{stop}^{(f)}} \\
&= \sum_{f\in\mathcal{F}} \frac{n_{ORF}^{(f)}}{\sum_{j\in\mathcal{F}} n_{ORF}^{(j)}} \cdot \frac{1}{p_{stop}^{(f)}} \\
&= \sum_{f\in\mathcal{F}} \frac{N_C \cdot p_{stop}^{(f)} \cdot \frac{p_{start}^{(f)}}{p_{start}^{(f)}+p_{stop}^{(f)}}}{\sum_{j\in\mathcal{F}} N_C \cdot p_{stop}^{(j)} \cdot \frac{p_{start}^{(j)}}{p_{start}^{(j)}+p_{stop}^{(j)}}} \cdot \frac{1}{p_{stop}^{(f)}} \\
&= \sum_{f\in\mathcal{F}} \frac{\frac{p_{start}^{(f)}}{p_{start}^{(f)}+p_{stop}^{(f)}}}{\sum_{j\in\mathcal{F}} p_{stop}^{(j)} \cdot \frac{p_{start}^{(j)}}{p_{start}^{(j)}+p_{stop}^{(j)}}} \\
&= \sum_{f\in\mathcal{F}} \frac{p_{start}^{(f)}}{p_{start}^{(f)} + p_{stop}^{(f)}} \left(\sum_{j\in\mathcal{F}} p_{stop}^{(j)} \cdot \frac{p_{start}^{(j)}}{p_{start}^{(j)} + p_{stop}^{(j)}} \right)^{-1}.
\end{aligned}$$

Adaption of (3.7) to weighted case The probability to draw an ORF of length $L \geq \ell'$ in any reading frame is given by

$$\begin{aligned} P\left(L \geq \ell'\right) &= \sum_{\ell=\ell'}^{\infty} P\left(L = \ell\right) = \sum_{\ell=\ell'}^{\infty} \sum_{f \in \mathcal{F}} p^{(f)} \left(1 - p_{stop}^{(f)}\right)^{\ell-1} \cdot p_{stop}^{(f)} \\ &= \sum_{f \in \mathcal{F}} p^{(f)} p_{stop}^{(f)} \sum_{\ell=\ell'}^{\infty} \left(1 - p_{stop}^{(f)}\right)^{\ell-1} \\ &= \sum_{f \in \mathcal{F}} p^{(f)} \left(1 - p_{stop}^{(f)}\right)^{\ell'-1}, \end{aligned} \tag{B.3}$$

over all six frames $\mathcal{F} \in \{\pm 1, \pm 2 \pm 3\}$.

Cumulative Distribution Function of weighted case Extension of CDF to the weighted case applying the results of Appendix B.1 and the ORF length distribution in Concept 3.7 we determine

$$\begin{aligned} \mathrm{P}\left(L \leq \ell\right) &= \sum_{k=1}^{\ell} \sum_{f \in \mathcal{F}} p^{(f)} \left(1 - p_{stop}^{(f)}\right)^{k-1} p_{stop}^{(f)} \\ &= \sum_{f \in \mathcal{F}} p^{(f)} \sum_{k=1}^{\ell} \left(1 - p_{stop}^{(f)}\right)^{k-1} p_{stop}^{(f)} \\ &= \sum_{f \in \mathcal{F}} p^{(f)} \left(1 - \left(1 - p_{stop}^{(f)}\right)^{\ell}\right). \end{aligned}$$

B.3. Transition matrices in Section 3.2.1

A detailed derivation of (3.11) and the remaining forward reading frame {+3} (B.4).

Transition matrix $\boldsymbol{P}^{(+2)}$

$$
\begin{aligned}
&P_{C_i^{+2}|C_{i-1}^{+2}}\left(n_j^{(2)}n_j^{(3)}n_{j+1}^{(1)}|n_{j-1}^{(2)}n_{j-1}^{(3)}n_j^{(1)}\right)\\
&=\frac{\sum_{n_{j-1}^{(1)}\in\mathcal{B}}\sum_{n_{j+1}^{(2)},n_{j+1}^{(3)}\in\mathcal{B}}P_{C_{j-1}^{+1}}\left(n_{j-1}^{(1)}n_{j-1}^{(2)}n_{j-1}^{(3)}\right)P_{C_j^{+1}|C_{j-1}^{+1}}\left(n_j^{(1)}n_j^{(2)}n_j^{(3)}|n_{j-1}^{(1)}n_{j-1}^{(2)}n_{j-1}^{(3)}\right)P_{C_{j+1}^{+1}|C_j^{+1}}\left(n_{j+1}^{(1)}n_{j+1}^{(2)}n_{j+1}^{(3)}|n_j^{(1)}n_j^{(2)}n_j^{(3)}\right)}{\sum_{n_{j-1}^{(1)}\in\mathcal{B}}\sum_{n_j^{(2)},n_j^{(3)}\in\mathcal{B}}P_{C_{j-1}^{+1}}\left(n_{j-1}^{(1)}n_{j-1}^{(2)}n_{j-1}^{(3)}\right)P_{C_j^{+1}|C_{j-1}^{+1}}\left(n_j^{(1)}n_j^{(2)}n_j^{(3)}|n_{j-1}^{(1)}n_{j-1}^{(2)}n_{j-1}^{(3)}\right)}\\
&=\frac{\sum_{n_{j-1}^{(1)}\in\mathcal{B}}P_{C_{j-1}^{+1}}\left(n_{j-1}^{(1)}n_{j-1}^{(2)}n_{j-1}^{(3)}\right)\sum_{n_{j+1}^{(2)},n_{j+1}^{(3)}\in\mathcal{B}}P_{C_j^{+1}|C_{j-1}^{+1}}\left(n_j^{(1)}n_j^{(2)}n_j^{(3)}|n_{j-1}^{(1)}n_{j-1}^{(2)}n_{j-1}^{(3)}\right)P_{C_{j+1}^{+1}|C_j^{+1}}\left(n_{j+1}^{(1)}n_{j+1}^{(2)}n_{j+1}^{(3)}|n_j^{(1)}n_j^{(2)}n_j^{(3)}\right)}{\sum_{n_{j-1}^{(1)}\in\mathcal{B}}P_{C_{j-1}^{+1}}\left(n_{j-1}^{(1)}n_{j-1}^{(2)}n_{j-1}^{(3)}\right)\sum_{n_j^{(2)},n_j^{(3)}\in\mathcal{B}}P_{C_j^{+1}|C_{j-1}^{+1}}\left(n_j^{(1)}n_j^{(2)}n_j^{(3)}|n_{j-1}^{(1)}n_{j-1}^{(2)}n_{j-1}^{(3)}\right)}\\
&=\frac{\sum_{n_{j-1}^{(1)},n_{j+1}^{(2)},n_{j+1}^{(3)}\in\mathcal{B}}P_{C_j^{+1}|C_{j-1}^{+1}}\left(n_j^{(1)}n_j^{(2)}n_j^{(3)}|n_{j-1}^{(1)}n_{j-1}^{(2)}n_{j-1}^{(3)}\right)P_{C_{j+1}^{+1}|C_j^{+1}}\left(n_{j+1}^{(1)}n_{j+1}^{(2)}n_{j+1}^{(3)}|n_j^{(1)}n_j^{(2)}n_j^{(3)}\right)}{\sum_{n_{j-1}^{(1)},n_j^{(2)},n_j^{(3)}\in\mathcal{B}}P_{C_j^{+1}|C_{j-1}^{+1}}\left(n_j^{(1)}n_j^{(2)}n_j^{(3)}|n_{j-1}^{(1)}n_{j-1}^{(2)}n_{j-1}^{(3)}\right)},
\end{aligned}
$$

and transition matrix $\boldsymbol{P}^{(+3)}$

$$
\begin{aligned}
&P_{C_i^{+3}|C_{i-1}^{+3}}\left(n_j^{(3)}n_{j+1}^{(1)}n_{j+1}^{(2)}|n_{j-1}^{(3)}n_j^{(1)}n_j^{(2)}\right)\\
&=\frac{\sum_{n_{j-1}^{(1)},n_{j-1}^{(2)}\in\mathcal{B}}\sum_{n_{j+1}^{(3)}\in\mathcal{B}}P_{C_{j-1}^{+1}}\left(n_{j-1}^{(1)}n_{j-1}^{(2)}n_{j-1}^{(3)}\right)P_{C_j^{+1}|C_{j-1}^{+1}}\left(n_j^{(1)}n_j^{(2)}n_j^{(3)}|n_{j-1}^{(1)}n_{j-1}^{(2)}n_{j-1}^{(3)}\right)P_{C_{j+1}^{+1}|C_j^{+1}}\left(n_{j+1}^{(1)}n_{j+1}^{(2)}n_{j+1}^{(3)}|n_j^{(1)}n_j^{(2)}n_j^{(3)}\right)}{\sum_{n_{j-1}^{(1)},n_{j-1}^{(2)}\in\mathcal{B}}\sum_{n_j^{(3)}\in\mathcal{B}}P_{C_{j-1}^{+1}}\left(n_{j-1}^{(1)}n_{j-1}^{(2)}n_{j-1}^{(3)}\right)P_{C_j^{+1}|C_{j-1}^{+1}}\left(n_j^{(1)}n_j^{(2)}n_j^{(3)}|n_{j-1}^{(1)}n_{j-1}^{(2)}n_{j-1}^{(3)}\right)}\\
&=\frac{\sum_{n_{j-1}^{(1)},n_{j-1}^{(2)},n_{j+1}^{(3)}\in\mathcal{B}}P_{C_j^{+1}|C_{j-1}^{+1}}\left(n_j^{(1)}n_j^{(2)}n_j^{(3)}|n_{j-1}^{(1)}n_{j-1}^{(2)}n_{j-1}^{(3)}\right)P_{C_{j+1}^{+1}|C_j^{+1}}\left(n_{j+1}^{(1)}n_{j+1}^{(2)}n_{j+1}^{(3)}|n_j^{(1)}n_j^{(2)}n_j^{(3)}\right)}{\sum_{n_{j-1}^{(1)},n_{j-1}^{(2)},n_j^{(3)}\in\mathcal{B}}P_{C_j^{+1}|C_{j-1}^{+1}}\left(n_j^{(1)}n_j^{(2)}n_j^{(3)}|n_{j-1}^{(1)}n_{j-1}^{(2)}n_{j-1}^{(3)}\right)}.
\end{aligned}
\tag{B.4}
$$

For the reverse strand we turn the direction of the strand, such that we also read from left to right

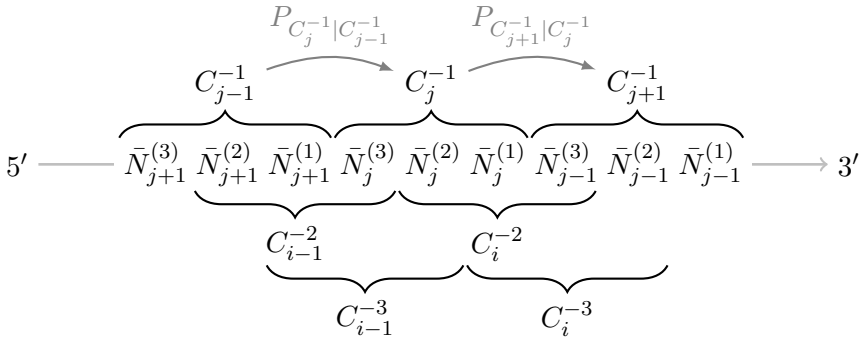


Figure B.1.: Reverse strand turned.

We get transition matrix $\boldsymbol{P}^{(-2)}$

$$P_{C_i^{-2}|C_{i-1}^{-2}}\left(\bar{n}_j^{(2)}\bar{n}_j^{(1)}\bar{n}_{j-1}^{(3)}|\bar{n}_{j+1}^{(2)}\bar{n}_{j+1}^{(1)}\bar{n}_j^{(3)}\right)$$
$$= \frac{\sum_{\bar{n}_{j+1}^{(3)},\bar{n}_{j-1}^{(2)},\bar{n}_{j-1}^{(1)}\in\mathcal{B}} P_{C_j^{-1}|C_{j-1}^{-1}}\left(\bar{n}_j^{(3)}\bar{n}_j^{(2)}\bar{n}_j^{(1)}|\bar{n}_{j+1}^{(3)}\bar{n}_{j+1}^{(2)}\bar{n}_{j+1}^{(1)}\right) P_{C_{j+1}^{-1}|C_j^{-1}}\left(\bar{n}_{j-1}^{(3)}\bar{n}_{j-1}^{(2)}\bar{n}_{j-1}^{(1)}|\bar{n}_j^{(3)}\bar{n}_j^{(2)}\bar{n}_j^{(1)}\right)}{\sum_{\bar{n}_{j+1}^{(3)},\bar{n}_j^{(2)},\bar{n}_j^{(1)}\in\mathcal{B}} P_{C_j^{-1}|C_{j-1}^{-1}}\left(\bar{n}_j^{(3)}\bar{n}_j^{(2)}\bar{n}_j^{(1)}|\bar{n}_{j+1}^{(3)}\bar{n}_{j+1}^{(2)}\bar{n}_{j+1}^{(1)}\right)}, \tag{B.5}$$

and transition matrix $\boldsymbol{P}^{(-3)}$

$$P_{C_i^{-3}|C_{i-1}^{-3}}\left(\bar{n}_j^{(1)}\bar{n}_{j-1}^{(3)}\bar{n}_{j-1}^{(2)}|\bar{n}_{j+1}^{(1)}\bar{n}_j^{(3)}\bar{n}_j^{(2)}\right)$$
$$= \frac{\sum_{\bar{n}_{j+1}^{(3)},\bar{n}_{j+1}^{(2)},\bar{n}_{j-1}^{(1)}\in\mathcal{B}} P_{C_j^{-1}|C_{j-1}^{-1}}\left(\bar{n}_j^{(3)}\bar{n}_j^{(2)}\bar{n}_j^{(1)}|\bar{n}_{j+1}^{(3)}\bar{n}_{j+1}^{(2)}\bar{n}_{j+1}^{(1)}\right) P_{C_{j+1}^{-1}|C_j^{-1}}\left(\bar{n}_{j-1}^{(3)}\bar{n}_{j-1}^{(2)}\bar{n}_{j-1}^{(1)}|\bar{n}_j^{(3)}\bar{n}_j^{(2)}\bar{n}_j^{(1)}\right)}{\sum_{\bar{n}_{j+1}^{(3)},\bar{n}_{j+1}^{(2)},\bar{n}_j^{(1)}\in\mathcal{B}} P_{C_j^{-1}|C_{j-1}^{-1}}\left(\bar{n}_j^{(3)}\bar{n}_j^{(2)}\bar{n}_j^{(1)}|\bar{n}_{j+1}^{(3)}\bar{n}_{j+1}^{(2)}\bar{n}_{j+1}^{(1)}\right)}. \tag{B.6}$$

B.4. List of Species

This table of species was already published in [1, Supplementary Table S1]. One new organism was added.

Table B.1.: Bacterial species investigated in this study. Species names, accession numbers, GC-content and length of the analyzed organism.

Species	Abbr.	GC [%]	Length [base pairs]	Accession
Candidatus *Phytoplasma mali*	CPm	21.4	601943	NC_011047
Wigglesworthia brevipalpis	Wb	22.5	697724	NC_004344
Streptobacillus moniliformis DSM 12112	Strm	26.3	1662578	NC_013515
Arcobacter butzleri RM4018	Ab	27	2341251	NC_009850
Clostridium botulinum A str. ATCC 19397	Cb	28.2	3863450	NC_009697
Clostridium difficile 630	Cdif	29.1	4290252	NC_009089
Campylobacter jejuni RM1221	Cj	30.3	1777831	NC_003912
Clostridium acetobutylicum ATCC824	Ca	30.9	3940880	NC_003030
Clostridium ljungdahlii DSM 13528	Clj	31.1	4630065	NC_014328
Francisella tularensis subsp. *holarctica* FTNF002-00	Fth	32.2	1890909	NC_009749
Staphylococcus aureus subsp. *aureus* JH1	Staa	33	2906507	NC_009632
Flavobacterium johnsoniae UW101	Fj	34.1	6096872	NC_009441
Bacillus cereus ATCC 14579	Bce	35.3	5411809	NC_004722
Anaerococcus prevotii DSM 20548	Ap	36.1	1883067	NC_013171
Legionella longbeachae NSW150	Llo	37.1	4077332	NC_013861
Listeria monocytogenes EGDe	Lm	38	2944528	NC_003210
Helicobacter pylori HPAG1	Hp	39.1	1596366	NC_008086
Streptococcus pneumoniae TCH8431/19A	Strp	39.8	2088772	NC_014251
Acinetobacter sp. ADP1	Asp	40.4	3598621	NC_005966
Streptococcus suis 05ZYH33	Strs	41.1	2096309	NC_009442
Gardnerella vaginalis 409-05	Gv	42	1617545	NC_013721
Bacteroides fragilis NCTC 9343	Bf	43.2	5205140	NC_003228
Campylobacter curvus 525.92	Cc	44.5	1971264	NC_009715

Continued on next page

Table B.1 – continued from previous page

Species	Abbr.	GC [%]	Length [base pairs]	Accession
Xenorhabdus bovienii SS-2004	Xb	45	4225498	NC_013892
Tropheryma whipplei TW08/27	Tw	46.3	925938	NC_004551
Haliscomenobacter hydrossis DSM 1100	Hh	47.1	8371686	NC_015510
Yersinia enterocolitica subsp. *enterocolitica* 8081	Yee	47.3	4615899	NC_008800
Porphyromonas gingivalis W83	Pg	48.3	2343476	NC_002950
Geobacillus thermodenitrificans NG80-2	Gth	49	3550319	NC_009328
Spirosoma linguale DSM 74	Sl	50.2	8078757	NC_013730
Nitrosococcus oceani ATCC 19707	No	50.3	3481691	NC_007484
Paenibacillus sp. JDR-2	Ps2	50.3	7184930	NC_012914
Escherichia coli O157:H7 EDL933	EcE	50.4	5528445	NC_002655
Escherichia coli O157:H7 Sakai	EcS	50.5	5498450	NC_002695
Pectobacterium wasabiae WPP163	Pw	50.5	5063892	NC_013421
Shigella sonnei Ss046	Shs	51	4825265	NC_007384
Salmonella enterica subsp. *e.* serovar Agona str. SL483	See	52.1	4798660	NC_011149
Corynebacterium diphtheriae NCTC 13129	Cdiph	53.5	2488635	NC_002935
Magnetococcus sp. MC-1	Msp	54.2	4719581	NC_008576
Mobiluncus curtisii ATCC 43063	Mc	55.4	2146480	NC_014246
Renibacterium salmoninarum ATCC 33209	Rsa	56.3	3155250	NC_010168
Cronobacter turicensis z3032	Ct	57.4	4384526	NC_013282
Paenibacillus mucilaginosus KNP414	Pm	58.4	8663821	NC_015690
Bifidobacterium dentium Bd1	Bd	58.5	2636367	NC_013714
Granulibacter bethesdensis CGDNIH1	Gbe	59.1	2708355	NC_008343
Propionibacterium acnes KPA171202	Pac	60	2560265	NC_006085
Pseudomonas fluorescens PfO-1	PfO-1	60.5	6438405	NC_007492
Pseudomonas fluorescens SBW25	Pfl25	60.5	6722539	NC_012660
Corynebacterium jeikeium K411	CjK	61.4	2462499	NC_007164
Laribacter hongkongensis HLHK9	Lh	62.4	3169329	NC_012559
Erythrobacter litoralis HTCC2594	Eli	63.1	3052398	NC_007722

Continued on next page

Table B.1 – continued from previous page

Species	Abbr.	GC [%]	Length [base pairs]	Accession
Thermanaerovibrio acidaminovorans DSM 6589	Ta	63.8	1848474	NC_013522
Mycobacterium abscessus ATCC 19977T	Ma	64.1	5067172	NC_010397
Xanthomonas campestris pv. *campestris*	Xcc	65	5079002	NC_007086
Burkholderia sp. 383 3	Bsp	65.3	1395069	NC_007509
Pseudomonas aeruginosa LESB58	Pae	66.3	6601757	NC_011770
Gordonia bronchialis DSM 43247	Gbr	67.1	5208602	NC_013441
Bordetella bronchiseptica RB50	Bbr	68.1	5339179	NC_002927
Rhodobacter sphaeroides KD131	Rs	69.2	3152792	NC_011963
Thermus thermophilus HB8	Tth	69.5	1849742	NC_006461
Haliangium ochraceum DSM 14365	Ho	69.5	9446314	NC_013440
Nocardia farcinica IFM10152	Nf	70.8	6021225	NC_006361
Streptosporangium roseum DSM 43021	Sr	70.9	10341314	NC_013595
Kocuria rhizophila DC2201	Kr	71.2	2697540	NC_010617
Amycolatopsis mediterranei U32	Am	71.3	10236715	NC_014318
Kytococcus sedentarius DSM 20547	Ks	71.6	2785024	NC_013169
Streptomyces coelicolor A3(2)	Scoe	72.1	8667507	NC_003888
Micrococcus luteus NCTC 2665	Mlu	73	2501097	NC_012803
Beutenbergia cavernae DSM 12333	Bca	73.1	4669183	NC_012669
Geodermatophilus obscurus DSM 43160	Go	74	5322497	NC_013757
Anaeromyxobacter dehalogenans 2CP-C	Ad	74.9	5013479	NC_007760

B.5. Application to *Mycoplasma mycoides*

A special organism we studied is *Mycoplasma mycoides* (Accession number NC_005364) [Raz06] introduced in Section 3.3.2. For the presentation of the results we follow Section 3.3.1.

Prediction of Codon Usage and Amino Acid Distribution:

Frame	$\|\|Model - Mycoplasma\|\|$	$D(Mycoplasma\|\|Model) + D(Model\|\|Mycoplasma)$
$+1$	$4.48 \cdot 10^{-4}$	$3.23 \cdot 10^{-5}$
$+2$	$7.94 \cdot 10^{-3}$	$1.95 \cdot 10^{-3}$
$+3$	$3.19 \cdot 10^{-3}$	$8.51 \cdot 10^{-4}$
-1	$5.27 \cdot 10^{-4}$	$1.81 \cdot 10^{-5}$
-2	$2.58 \cdot 10^{-2}$	$1.17 \cdot 10^{-2}$
-3	$1.12 \cdot 10^{-2}$	$2.34 \cdot 10^{-3}$
Frame	$\|\|RND - Mycoplasma\|\|$	$D(Mycoplasma\|\|RND) + D(RND\|\|Mycoplasma)$
$+1$	$1.17 \cdot 10^{-1}$	$6.42 \cdot 10^{-1}$

Table B.2.: Kullback-Leibler divergence and Euclidean distance of the amino acid distributions of *Mycoplasma mycoides* and the predictions of the model. Random amino acid distribution based on the GC content of the organism.

Start- and Stop codon probabilities:

	p_{start}			p_{stop}		
Frame	$Mycoplasma$	$Model$	Δ_{start}	$Mycoplasma$	$Model$	Δ_{stop}
$+1$	0.0246	0.0241	$4.4 \cdot 10^{-4}$	0.0031	0.0033	$2.3 \cdot 10^{-4}$
$+2$	0.1067	0.1058	$8.5 \cdot 10^{-4}$	0.0896	0.0900	$3.7 \cdot 10^{-4}$
$+3$	0.0164	0.0164	$9.4 \cdot 10^{-5}$	0.1097	0.1093	$4.6 \cdot 10^{-4}$
-1	0.0484	0.0484	$3.5 \cdot 10^{-6}$	0.0846	0.0846	$8.2 \cdot 10^{-6}$
-2	0.0194	0.0206	$1.2 \cdot 10^{-3}$	0.0267	0.0294	$2.6 \cdot 10^{-3}$
-3	0.0592	0.0605	$1.3 \cdot 10^{-3}$	0.0950	0.0987	$3.7 \cdot 10^{-3}$

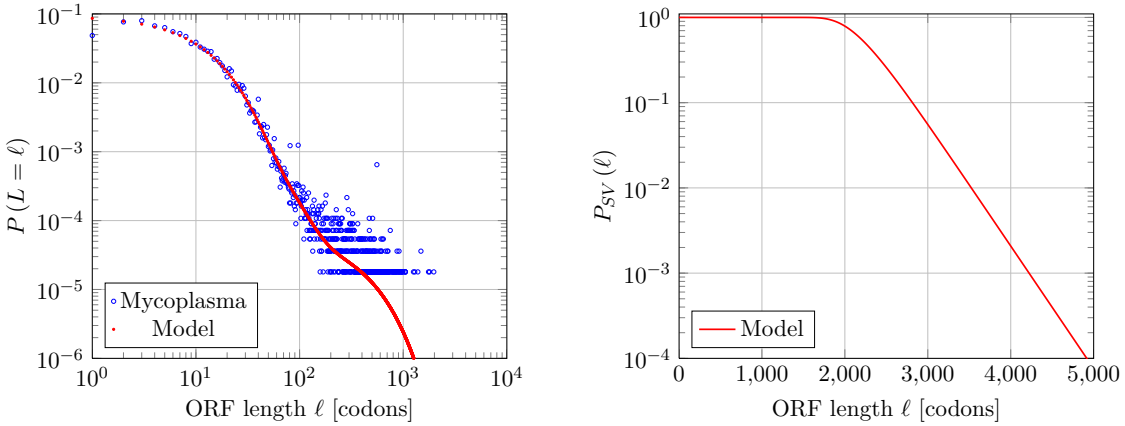
Table B.3.: Comparison of start and stop codon probabilities of *Mycoplasma mycoides* (empirical) with the model predictions. Additionally, the absolute deviation is presented.

Waiting Time:

Frame	*Mycoplasma*	*Model*		
	$\hat{t}_M$	t_M	t_A	s_A
$+1$	325.14	302.47	324.91	1.14
$+2$	11.16	11.12	10.44	$8.4 \cdot 10^{-1}$
$+3$	9.11	9.15	9.12	$3.5 \cdot 10^{-1}$
-1	11.82	11.82	11.86	$3.5 \cdot 10^{-1}$
-2	37.43	34.06	34.42	$7.9 \cdot 10^{-1}$
-3	10.53	10.13	10.18	$4.0 \cdot 10^{-1}$

Table B.4.: Mean recurrence time t_M compared with mean absorbing time t_A over all transient states together with sample standard deviation s_A.

ORF length distribution and Survival Probability:



(a) Probability distribution of ORF lengths.

(b) Survival probability.

Figure B.2.: (a) Probability distribution of ORF lengths of *Mycoplasma mycoides* in comparison with the prediction of the model. (b) Survival probability to observe at least one ORF with length $L \geq \ell$.

B.6. Application to single start codon *ATG*

We repeat the analysis of *Escherichia coli* O157:H7 EDL933 (*EHEC*, Accession number NC_002655) allowing *ATG* as start codon. For the presentation of the results we follow Section 3.3.1.

Prediction of Codon Usage and Amino Acid Distribution:

Frame	$\|\|Model - EHEC\|\|$	$D(EHEC\|\|Model) + D(Model\|\|EHEC)$
+1	$3.77 \cdot 10^{-4}$	$1.43 \cdot 10^{-5}$
+2	$4.26 \cdot 10^{-3}$	$1.32 \cdot 10^{-3}$
+3	$5.22 \cdot 10^{-3}$	$9.87 \cdot 10^{-4}$
−1	$3.99 \cdot 10^{-4}$	$4.99 \cdot 10^{-6}$
−2	$4.72 \cdot 10^{-3}$	$5.80 \cdot 10^{-4}$
−3	$4.20 \cdot 10^{-3}$	$5.71 \cdot 10^{-4}$
Frame	$\|\|RND - EHEC\|\|$	$D(EHEC\|\|RND) + D(RND\|\|EHEC)$
+1	$9.33 \cdot 10^{-2}$	$3.59 \cdot 10^{-1}$

Table B.5.: Kullback-Leibler divergence and Euclidean distance of the amino acid distributions of *EHEC* and the predictions of the model. Random amino acid distribution based on the GC content of the organism.

Start- and Stop codon probabilities:

	p_{start}			p_{stop}		
Frame	$EHEC$	$Model$	Δ_{start}	$EHEC$	$Model$	Δ_{stop}
+1	0.0662	0.0660	$2.3 \cdot 10^{-4}$	0.0033	0.0034	$1.5 \cdot 10^{-4}$
+2	0.0797	0.0788	$9.1 \cdot 10^{-4}$	0.0484	0.0485	$8.8 \cdot 10^{-5}$
+3	0.0150	0.0153	$2.8 \cdot 10^{-4}$	0.0630	0.0627	$3.5 \cdot 10^{-4}$
−1	0.0371	0.0371	$4.0 \cdot 10^{-5}$	0.0256	0.0256	$7.6 \cdot 10^{-5}$
−2	0.0565	0.0571	$5.8 \cdot 10^{-4}$	0.0363	0.0368	$5.2 \cdot 10^{-4}$
−3	0.0299	0.0302	$3.2 \cdot 10^{-4}$	0.0473	0.0475	$1.8 \cdot 10^{-4}$

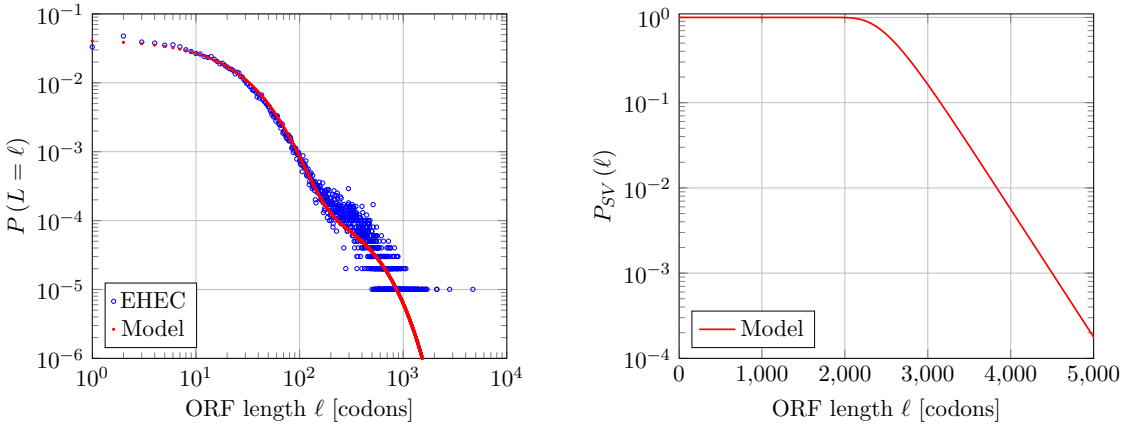
Table B.6.: Comparison of start and stop codon probabilities of *EHEC* (empirical) with the model predictions. Additionally, the absolute deviation is presented.

Waiting Time:

Frame	*EHEC*	*Model*		
	$\hat{t}_M$	t_M	t_A	s_A
$+1$	302.93	289.86	302.84	$7.9 \cdot 10^{-1}$
$+2$	20.66	20.62	20.5	$5.2 \cdot 10^{-1}$
$+3$	15.86	15.95	15.97	$3.8 \cdot 10^{-1}$
-1	39.12	39.0	39.78	$4.7 \cdot 10^{-1}$
-2	27.54	27.16	27.98	$4.6 \cdot 10^{-1}$
-3	21.15	21.07	21.6	$5.5 \cdot 10^{-1}$

Table B.7.: Mean recurrence time t_M compared with mean absorbing time t_A over all transient states together with sample standard deviation s_A.

ORF length distribution and Survival Probability:



(a) Probability distribution of ORF lengths.

(b) Survival probability.

Figure B.3.: (a) Probability distribution of ORF lengths of *EHEC* in comparison with the prediction of the model. (b) Survival probability to observe at least one ORF with length $L \geq \ell$.

Parameter	$EHEC$	$Model$
n_{ORF}	99578	98784.51
$\bar{\ell}_{ORF}$	39.69	40.2
$Q(0.25)$	7.0	7.0
$Q(0.50)$	17.0	17.0
$Q(0.75)$	37.0	37.0
$Q(0.99)$	444.818	504

Table B.8.: Statistical parameters of interest to compare the ORF distribution predicted by the model with the empirical distribution of *EHEC*. Unit of length values is codons.

	Parameter	+1	+2	+3	−1	−2	−3
$EHEC$	n_{ORF}	4380.0	23780	6203	11570	17696	9526
$Model$	n_{ORF}	5650.59	32625.87	8370.39	15783.7	23869.81	12484.15
$EHECg$	$\bar{\ell}_{ORF}$	301.93	22.28	22.31	44.65	30.39	25.62
$Model$	$\bar{\ell}_{ORF}$	289.86	20.62	15.95	39.0	27.16	21.07
$Model$	B_{SV}	3832.0	303.0	212.0	551.0	393.0	290.0

Table B.9.: Statistical parameters of interest to compare the alternative reading frame distributions predicted by the model with the empirical distribution of *EHEC*. Unit of length values is codons.

APPENDIX C

Details regarding Evolution

Rate matrices in Section 4.1.2

Assume, we are interested in the rate from codon C_i^{+3} to codon $\tilde{C}_i^{+3}$ with realizations $c_i^{+3} = \left(x_{j-1}^{(3)} x_j^{(1)} x_j^{(2)}\right)$ and $\tilde{c}_i^{+3} = \left(y_{j-1}^{(3)} y_j^{(1)} y_j^{(2)}\right)$.

$$\boldsymbol{Q}^{(+3)}\left(y_{j-1}^{(3)} y_j^{(1)} y_j^{(2)} | x_{j-1}^{(3)} x_j^{(1)} x_j^{(2)}\right) = \boldsymbol{Q}^{(+3)}\left(\tilde{c}_i^{+3} | c_i^{+3}\right)$$
$$= \frac{\sum_{x_{j-1}^{(1)}, x_{j-1}^{(2)}, x_j^{(3)}, y_{j-1}^{(1)}, y_{j-1}^{(2)}, y_j^{(3)} \in \mathcal{B}} \pi^{(+1)}_{\text{di-codon}}(x_{j-1}^{(1)} x_{j-1}^{(2)} c_i^{+3} x_j^{(3)}) \cdot \boldsymbol{Q}^{(+1)}_{\text{di-codon}}(y_{j-1}^{(1)} y_{j-1}^{(2)} \tilde{c}_i^{+3} y_j^{(3)} | x_{j-1}^{(1)} x_{j-1}^{(2)} c_i^{+3} x_j^{(3)})}{\sum_{x_{j-1}^{(1)}, x_{j-1}^{(2)}, x_j^{(3)} \in \mathcal{B}} \pi^{(+1)}_{\text{di-codon}}(x_{j-1}^{(1)} x_{j-1}^{(2)} c_i^{+3} x_j^{(3)})},$$

where the probability for a di-codon in $\{+1\}$ that consists of two consecutive codons c_{j-1} and c_j is determined under the IID assumption using the stationary distribution $\boldsymbol{\pi}^{(+1)}$ by $\pi^{(+1)}_{\text{di-codon}}(c_{j-1}c_j) = \pi^{(+1)}_{c_{j-1}} \cdot \pi^{(+1)}_{c_j} \quad \forall\, c_{j-1}, c_j \in \mathcal{C}_{61}$.

For the reverse reading frames, we first determine the reverse rate matrix $\boldsymbol{Q}^{(-1)}$ directly from $\boldsymbol{Q}^{(+1)}$. Given the rate from $C_n^{+1} = \left(X_n^{(1)}, X_n^{(2)}, X_n^{(3)}\right)$ to $\tilde{C_n}^{+1} = \left(Y_n^{(1)}, Y_n^{(2)}, Y_n^{(3)}\right)$ in reading frame $\{+1\}$, this corresponds in reverse direction to a transition from codon $C_n^{-1} = \left(\bar{X}_n^{(3)}, \bar{X}_n^{(2)}, \bar{X}_n^{(1)}\right)$ to $\tilde{C_n}^{+1} = \left(\bar{Y}_n^{(3)}, \bar{Y}_n^{(2)}, \bar{Y}_n^{(1)}\right)$, where $\bar{X}, \bar{Y}$ denote complement versions of nucleotides X, Y. For all $\left(x_j^{(1)} x_j^{(2)} x_j^{(3)}\right), \left(y_j^{(1)} y_j^{(2)} y_j^{(3)}\right) \in \mathcal{C}_{61}$ determine

$$\boldsymbol{Q}^{(-1)}\left(\bar{y}_j^{(3)} \bar{y}_j^{(2)} \bar{y}_j^{(1)} | \bar{x}_j^{(3)} \bar{x}_j^{(2)} \bar{x}_j^{(1)}\right) = \boldsymbol{Q}^{(+1)}\left(y_j^{(1)} y_j^{(2)} y_j^{(3)} | x_j^{(1)} x_j^{(2)} x_j^{(3)}\right).$$

Accordingly the reverse stationary distribution $\boldsymbol{\pi}^{(-1)}$ is

$$\pi^{(-1)}_{\left(\bar{x}_j^{(3)} \bar{x}_j^{(2)} \bar{x}_j^{(1)}\right)} = \pi^{(+1)}_{\left(x_j^{(1)} x_j^{(2)} x_j^{(3)}\right)} \quad \forall \left(x_j^{(1)} x_j^{(2)} x_j^{(3)}\right) \in \mathcal{C}_{61}.$$

The rate matrices for $\{-2, -3\}$ are calculated similar to those of $\{+2, +3\}$, replacing $\boldsymbol{\pi}^{(+1)}$ and $\boldsymbol{Q}^{(+1)}$ by $\boldsymbol{\pi}^{(-1)}$ and $\boldsymbol{Q}^{(-1)}$:

$$\boldsymbol{Q}^{(-2)}\left(\bar{y}_j^{(2)}\bar{y}_j^{(1)}\bar{y}_{j-1}^{(3)}|\bar{x}_j^{(2)}\bar{x}_j^{(1)}\bar{x}_{j-1}^{(3)}\right) = \boldsymbol{Q}^{(-2)}\left(\tilde{c}_i^{-2}|c_i^{-2}\right)$$
$$= \frac{\sum_{\bar{x}_j^{(3)},\bar{x}_{j-1}^{(2)},\bar{x}_{j-1}^{(1)},\bar{y}_j^{(3)},\bar{y}_{j-1}^{(2)},\bar{y}_{j-1}^{(1)}\in\mathcal{B}} \pi^{(-1)}_{\text{di-codon}}(\bar{x}_j^{(3)}c_i^{-2}\bar{x}_{j-1}^{(2)}\bar{x}_{j-1}^{(1)}) \cdot \boldsymbol{Q}^{(-1)}_{\text{di-codon}}(\bar{y}_j^{(3)}\tilde{c}_i^{-2}\bar{y}_{j-1}^{(2)}\bar{y}_{j-1}^{(1)}|\bar{x}_j^{(3)}c_i^{-2}\bar{x}_{j-1}^{(2)}\bar{x}_{j-1}^{(1)})}{\sum_{\bar{x}_j^{(3)},\bar{x}_{j-1}^{(2)},\bar{x}_{j-1}^{(1)}\in\mathcal{B}} \pi^{(-1)}_{\text{di-codon}}(\bar{x}_j^{(3)}c_i^{-2}\bar{x}_{j-1}^{(2)}\bar{x}_{j-1}^{(1)})},$$
$$\boldsymbol{Q}^{(-3)}\left(\bar{y}_j^{(1)}\bar{y}_{j-1}^{(3)}\bar{y}_{j-1}^{(2)}|\bar{x}_j^{(1)}\bar{x}_{j-1}^{(3)}\bar{x}_{j-1}^{(2)}\right) = \boldsymbol{Q}^{(-3)}\left(\tilde{c}_i^{-3}|c_i^{-3}\right)$$
$$= \frac{\sum_{\bar{x}_j^{(3)},\bar{x}_j^{(2)},\bar{x}_{j-1}^{(1)},\bar{y}_j^{(3)},\bar{y}_j^{(2)},\bar{y}_{j-1}^{(1)}\in\mathcal{B}} \pi^{(-1)}_{\text{di-codon}}(\bar{x}_j^{(3)}\bar{x}_j^{(2)}c_i^{-3}\bar{x}_{j-1}^{(1)}) \cdot \boldsymbol{Q}^{(-1)}_{\text{di-codon}}(\bar{y}_j^{(3)}\bar{y}_j^{(2)}\tilde{c}_i^{-3}\bar{y}_{j-1}^{(1)}|\bar{x}_j^{(3)}\bar{x}_j^{(2)}c_i^{-3}\bar{x}_{j-1}^{(1)})}{\sum_{\bar{x}_j^{(3)},\bar{x}_j^{(2)},\bar{x}_{j-1}^{(1)}\in\mathcal{B}} \pi^{(-1)}_{\text{di-codon}}(\bar{x}_j^{(3)}\bar{x}_j^{(2)}c_i^{-3}\bar{x}_{j-1}^{(1)})},$$

with

$$\boldsymbol{Q}^{(-1)}_{\text{di-codon}} = \left(\boldsymbol{Q}^{(-1)} \otimes \boldsymbol{I}_Q + \boldsymbol{I}_Q \otimes \boldsymbol{Q}^{(-1)}\right)$$

and $\pi^{(-1)}_{\text{di-codon}}(c_{j-1}c_j) = \pi^{(-1)}_{c_{j-1}} \cdot \pi^{(-1)}_{c_j}$.

Bibliography

References

[AH96] Jun Adachi and Masami Hasegawa. Model of amino acid substitution in proteins encoded by mitochondrial DNA. *Journal of Molecular Evolution*, 42:459–468, 1996.

[Ari06] Suguru Arimoto. An Algorithm for Computing the Capacity of Arbitrary Discrete Memoryless Channels. *IEEE Transactions on Information Theory*, 18(1):14–20, September 2006.

[art] Free Clip art. Genetic code. Available online at `http://www.clker.com/cliparts/M/p/0/K/I/B/genetic-code-rna-black-and-white.svg`; visited on September 30th 2014.

[ARY14] Konstantinos Angelis, Mario D. Reis, and Ziheng Yang. Bayesian estimation of nonsynonymous/synonymous rate ratios for pairwise sequence comparisons. *Molecular Biology and Evolution*, pages msu142+, April 2014.

[BAJ+11] Alberts Brude, Johnson Alexander, Lewis Julian, Raff Martin, Roberts Keith, and Walter Peter. *Molekularbiologie der Zelle*. Wiley - VCH, fifth edition, 2011.

[Bal13] Wikimedia Commons Madeleine Price Ball. DNA chemical structure, 2013. Available online at `http://commons.wikimedia.org/wiki/File:DNA_chemical_structure.svg#mediaviewer/File:DNA_chemical_structure.svg`; visited on September 25th 2014.

[Bar76] Bart G. Barrell. Overlapping genes in bacteriophage $\Phi X174$. *Nature*, 264:34–41, 1976.

[Bat97] Gérard Battail. Does information theory explain biological evolution? *Europhysics Letters*, 40(3):343, 1997.

[BCLWS08] Nicholas D. Bonawitz, Marc Chatenay-Lapointe, Christopher M. Wearn, and Gerald S. Shadel. Expression of the *rDNA*-encoded mitochondrial protein *Tar1p* is stringently controlled and responds differentially to mitochondrial respiratory demand and dysfunction. *Current Genetics*, 54(2):83–94, 2008.

[BIB+97] Frederick R. Blattner, Guy Plunkett III, Craig A. Bloch, Nicole T. Perna, Valerie Burland, Monica Riley, Julio Collado-Vides, Jeremy D. Glasner, Christopher K. Rode, George F. Mayhew, Jason Gregor, Nelson Wayne Davis, Heather A. Kirkpatrick, Michael A. Goeden, Debra J. Rose, Bob Mau, and Ying Shao. The Complete Genome Sequence of *Escherichia coli* K-12. *Science*, 277:1453–1462, 1997.

[Bla85] B. Edwin Blaisdell. A method for estimating from two aligned present day DNA sequences their ancestral composition and subsequent rates of composition and subsequent rates of substitution, possibly different in the two lineages, corrected for multiple and parallel substitutions at the same site. *Journal of Molecular Evolution*, 22:69–81, 1985.

[BM13] Gordon M. Bennett and Nancy A. Moran. Small, smaller, smallest: the origins and evolution of ancient dual symbioses in a Phloem-feeding insect. *Genome Biology and Evolution*, 5(9):1675–1688, 2013.

[Bos99] Martin Bossert. *Channel Coding for Telecommunications*. John Wiley & Sons, New York, first edition, 1999.

[Bos12] Martin Bossert. *Einführung in die Nachrichtentechnik*. Oldenbourg Verlag, first edition, 2012.

[BRBR12] Emmanuel Buschiazzo, Carol Ritland, Jorg Bohlmann, and Kermit Ritland. Slow but not low: genomic comparisons reveal slower evolutionary rate and higher dN/dS in conifers compared to angiosperms. *BMC Evolutionary Biology*, 12(1):8, 2012.

[BSMM01] Ilja N. Bronstein, Konstantin A. Semendjajew, Gerhard Musiol, and Heiner Mühlig. *Taschenbuch der Mathematik*. Verlag Harri Deutsch, Frankfurt am Main, 2001.

[BSN02] Martin Behrens, Jalaluddin Sheikh, and James P. Nataro. Regulation of the overlapping *pic/set* locus in *Shigella flexneri* and enteroaggregative *Escherichia coli*. *Infection and Immunity*, 70:2915–2925, 2002.

[CAJ+94] Ying Cao, Jun Adachi, Axel Janke, Pääbo Svante, and Hasegawa Masami. Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene. *Journal of Molecular Evolution*, 39:519–527, 1994.

[CB85] Athel Cornish-Bowden. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Research*, 13(9):3021–3030, May 1985.

[CCA+02] Paulo S.R. Coelho, Bryan Anthony C., Kumar Anuj, Shadel Gerald S., and Snyder Michael. A novel mitochondrial protein, *Tar1p*, is encoded on the antisense strand of the nuclear *25S rDNA*. *Genes & Development*, 16(21):755–60, November 2002.

[CHI10] Han-Yu Chuang, Matan Hofree, and Trey Ideker. A decade of systems biology. *Annual Review of Cell and Developmental Biology*, 26:721–44, January 2010.

[Cin75] Erhan Cinlar. *Introduction to Stochastic Processes*. Prentice-Hall, Inc., 1975.

[Com95] Joseph M. Comeron. A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. *Journal of Molecular Evolution*, 41(6):1152–9, 1995.

[Cri58] Francis Crick. On protein synthesis. *Symposia of the Society for Experimental Biology*, 12:138–163, 1958.

[CSD+98] Paul R. Cooper, Nancy J. Smilinich, Colleen D. Day, Norma J. Nowak, Laura H. Reid, R. Scott Pearsall, Mark Reece, Dirk Prawitt, John Landers, David E. Housman, Andreas Winterpacht, Bernhard U. Zabel, Jerry Pelletier, Bernard E. Weissman, Thomas B. Shows, and Michael J. Higgins. Divergently Transcribed Overlapping Genes Expressed in Liver and Kidney and Located in the 11p15.5 Imprinted Domain. *Genomics*, 49(1):38 – 51, 1998.

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

[CVB10] Nicola Chirico, Alberto Vianelli, and Robert Belshaw. Why genes overlap in viruses. *Proceedings of the Royal Society B: Biological Sciences*, 277(1701):3809–3817, December 2010.

[CW07] Peter J.A. Cock and David E. Whitworth. Evolution of Gene Overlaps: Relative Reading Frame Bias in Prokaryotic Two-Component System Genes. *Journal of Molecular Evolution*, 64(4):457–462, 2007.

[CW10] Peter J. A. Cock and David E. Whitworth. Evolution of relative reading frame bias in unidirectional prokaryotic gene overlaps. *Molecular Biology and Evolution*, 27(4):753–6, 2010.

[DBPS07] Arthur L. Delcher, Kirsten A. Bratke, Edwin C. Powers, and Steven L. Salzberg. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, 23(6):673–679, 2007.

[DGH+06] Zaher Dawy, Bernhard Göbel, Joachim Hagenauer, Cristophe Andreoli, Thomas Meitinger, and Jakob C. Müller. Gene mapping and marker clustering using Shannon's mutual information. *IEEE Transactions on Computational Biology and Bioinformatics*, 3(1):47–56, January-March 2006.

[DGHM05] Zaher Dawy, Faruck Gonzàles, Joachim Hagenauer, and Jakob C. Müller. Modeling and Analysis of Gene Expression Mechanisms: A Communication Theory Approach. In *IEEE International Conference on Communications (ICC 2005)*, Seoul, South Corea, May 2005.

[DSO78] Margaret O. Dayhoff, R. M. Schwartz, and B. C. Orcutt. Chapter 22: A model of evolutionary change in proteins. In *in Atlas of Protein Sequence and Structure*, volume 5, page 345–352, Washington DC, 1978.

[DSS09] Wayne Delport, Konrad Scheffler, and Cathal Seoighe. Models of coding sequence evolution. *Briefings in Bioinformatics*, 10(1):97–109, 2009.

[DTA08] John Dougherty, Ioan Tabus, and Jaakko Astola. Inference of Gene Regulatory Networks Based on a Universal Minimum Description Length. *EURASIP Journal on Bioinformatics and Systems Biology*, 2008(1):482090, 2008.

[FBW+14] Lea Fellner, Niklas Bechtel, Michael A. Witting, Svenja Simon, Philippe Schmitt-Kopplin, Daniel A. Keim, Siegfried Scherer, and Klaus Neuhaus. Phenotype of htgA (mbiA), a recently evolved orphan gene of *Escherichia coli* and *Shigella*, completely overlapping in antisense to yaaW. *FEMS Microbiology Letters*, 350(1):57–64, 2014.

[Fel68] William Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley, third edition, 1968.

[Fel81] Joseph Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981.

[Fel95] Joseph Felsenstein. PHYLIP (Phylogenetic inference package) ver. 3.57, 1995.

[FH98] Stephen J. Freeland and Laurence D. Hurst. The genetic code is one in a million. *Journal of Molecular Evolution*, 47(3):238–248, 1998.

[FHP14] Miguel Fonseca, D. James Harris, and David Posada. Origin and Length Distribution of Unidirectional Prokaryotic Overlapping Genes. *G3*, 4:19–27, January 2014.

[FNT03] Yoko Fukuda, Yoichi Nakayama, and Masaru Tomita. On dynamics of overlapping genes in bacterial genomes. *Gene*, 323:181–187, December 2003.

[FWT99] Yoko Fukuda, Takanori Washio, and Masaru Tomita. Comparative study of overlapping genes in the genomes of Mycoplasma genitalium and Mycoplasma pneumoniae. *Nucleic Acids Research*, 27:1847–1853+, 1999.

[Gal68] Robert G. Gallager. *Information Theory and Reliable Communication*. John Wiley & Sons, New York, 1968.

[GCFW08] Guangtao Ge, Lenore Cowen, Xiaochuan Feng, and Giovanni Widmer. Protein Coding Gene Nucleotide Substitution Pattern in the Apicomplexan Protozoa *Cryptosporidium parvum* and *Cryptosporidium hominis*. *Comparative and Functional Genomics*, 2008.

[GD02] Sébastien Guyader and Danièle G. Ducray. Sequence analysis of Potato leafroll virus isolates reveals genetic stability, major evolutionary events and differential selection pressure between overlapping reading frame products. *Journal of General Virology*, 83:1799–807, 2002.

[GDA+13] Pierrea Gladieux, Benjamin Devier, Gabriela Aguileta, Corinne Cruaud, and Tatiana Giraud. Purifying selection after episodes of recurrent adaptive diversification in fungal pathogens. *Infection, Genetics and Evolution*, 17:123–131, 2013.

[GF95] Roderic Guigo and James W. Fickett. Distinctive sequence features in protein coding genic non-coding, and intergenic human DNA. *Journal of Molecular Biology*, 253(1):51–60, 1995.

[GGMS03] M. Rosario Garcia-Gil, Merja Mikkonen, and Outi Savolainen. Nucleotide diversity at two phytochrome loci along a latitudinal cline in *Pinus sylvestris*. *Molecular Ecology*, 12(5):1195–206, 2003.

[Gil94] Nicholas W. Gillham. *Organelle Genes and Genomes*. Oxford University Press, USA, September 1994.

[GIN82] Takashi Gojobori, Kazushige Ishii, and Masatoshi Nei. Estimation of Average Number of Nucleotide Substitutions when the Rate of Substitution varies with Nucleotide. *Journal of Molecular Evolution*, 18:414–423, 1982.

[GRS+13] Christopher J. Graves, Vera I. D. Ros, Brian Stevenson, Paul D. Sniegowski, and Dustin Brisson. Natural Selection Promotes Antigenic Evolvability. *PLoS Pathogens*, 9(11):e1003766, November 2013.

[GS97] Charles M. Grinstead and J. Laurie Snell. *Introduction to Probability*. American Mathematical Society, Providence, 1997.

[GS01] Geoffrey Grimmett and David Stirzaker. *Probability and Random Processes*. Oxford University Press, New York, third edition, 2001.

[GY94] Nick Goldman and Ziheng Yang. A Codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution*, 11:725–736, 1994.

[HB13] David Haughton and Félix Balado. BioCode: Two biologically compatible Algorithms for embedding data in non-coding and coding regions of DNA. *BMC Bioinformatics*, 14:121+, 2013.

[HDG+04] Joachim Hagenauer, Zaher Dawy, Bernhard Göbel, Pavol Hanus, and Jakob C. Müller. Genomic Analysis using Methods from Information Theory. In *IEEE Information Theory Workshop (ITW 2004)*, pages 55–59, San Antonio, USA, October 2004.

[HGD+07] Pavol Hanus, Bernhard Göbel, Janis Dingel, Johanna Weindl, Juergen Zech, Zaher Dawy, Joachim Hagenauer, and Jakob C. Müller. Information and communication theory in molecular biology. *Electrical Engineering*, 90(2):161–173, 2007.

[HH91] David Haig and Laurence D. Hurst. A quantitative measure of error minimization in the genetic code. *Journal of molecular evolution*, 33(5):412–417, November 1991.

[HH92] Steven Henikoff and Jorja G. Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89:10915–10919, 1992.

[HH05] Austin L. Hughes and Mary Ann K. Hughes. Patterns of nucleotide difference in overlapping and non-overlapping reading frames of papillomavirus genomes. *Virus Res*, 113:81–88, 2005.

[HKY85] Masami Hasegawa, Hirohisa Kishino, and Taka-aki Yano. Dating of the Human-Ape Splitting by a Molecular Clock of Mitochondrial-DNA. *Journal of Molecular Evolution*, 22:160–174, 1985.

[HPMR+10] Matthew R. Hemm, Brian J. Paul, Juan Miranda-Rios, Aixia Zhang, Nima Soltanzad, and Gisela Storz. Small stress response proteins in *Escherichia coli*: proteins missed by classical proteomic studies. *Journal of Bacteriology*, 192:46–58, 2010.

[HS95] Jotun Hein and Jens Stovlbaek. A maximum-likelihood approach to analyzing nonoverlapping and overlapping reading frames. *Journal of Molecular Evolution*, 40(2):181–9, 1995.

[Hur02] Laurence D. Hurst. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends in Genetics*, 18(9), September 2002.

[HWdS+01] Austin L. Hughes, Kristi Westover, Jack da Silva, David H. O'Connor, and David I. Watkins. Simultaneous positive and purifying selection on overlapping reading frames of the tat and vpr genes of simian immunodeficiency virus. *Journal of Virology*, 75:7966–72, 2001.

[HYK84] Masami Hasegawa, Taka-aki Yano, and Hirohisa Kishino. A new molecular clock of mitochondrial DNA and the evolution of Hominoids. *Proceedings of the Japan Academy, Series B*, 60:95–98, 1984.

[IA07] Shalev Itzkovitz and Uri Alon. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. *Genome Research*, February 2007.

[IHS+00] Yoshio Inokuchi, Akikazu Hirashima, Yasuhiko Sekine, Laszlo Janosi, and Akira Kaji. Role of ribosome recycling factor (RRF) in translational coupling. *The EMBO Journal*, 19(14):3788–98, 2000.

[Ina95] Yasuo Ina. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. *Journal of Molecular Evolution*, 40(2):190–226, February 1995.

[JC69] Thomas H. Jukes and Charles R. Cantor. *Evolution of Protein Molecules*. Academy Press, 1969.

[JC04] Zackary I. Johnson and Sallie W. Chisholm. Properties of overlapping genes are conserved across microbial genomes. *Genome Research*, 14(11):2268–2272, November 2004.

[JPF+06] Klaus T. Jensen, Lise Petersen, Søren Falk, Pernille Iversen, Peter Andersen, Michael Theisen, and Anders Krogh. Novel overlapping coding sequences in Chlamydia trachomatis. *FEMS Microbiology Letters*, 265(1):106–117, 2006.

[JTT92] David T. Jones, William R. Taylor, and Janet M. Thornton. The rapid generation of mutation data matrices from protein sequences. *Computer applications in the biosciences: CABIOS*, 8(3):275–282, June 1992.

[Kau69] Stuart A. Kauffman. Metabolic Stability and Epigenesis in Randomly Constructed Nets. *Journal of Theoretical Biology*, 22:437–467, 1969.

[Kau74] Stuart A. Kauffman. The Large Scale Structure and Dynamics of Genetic Control Circuits: an Ensemble Approach. *Journal of Theoretical Biology*, 44:167–190, 1974.

[KCGC02] Samuel Karlin, Chingfer Chen, Andrew J. Gentles, and Michael Cleary. Associations between human disease genes and overlapping gene groups and multiple amino acid runs. *Proceedings of the National Academy of Sciences of the United States of America*, 99(26):17008–17013, 2002.

[KCH+09] Dae-Soo Kim, Chi-Young Cho, Jae-Won Huh, Heui-Soo Kim, and Hwan-Gue Cho. EVOG: a database for evolutionary analysis of overlapping genes. *Nucleic Acids Research*, 37:698–702, 2009.

[KDS07] Carl Kingsford, Arthur L. Delcher, and Steven L. Salzberg. A unified model explaining the offsets of overlapping and near-overlapping prokaryotic genes. *Molecular Biology and Evolution*, 24(9):2091–8, 2007.

[Kei79] Julian Keilson. *Markov Chain Models – Rarity and Exponentiality*. Springer-Verlag, New York, 1979.

[Kel12] Wikimedia Commons Kelvinsong. Diagram of RNA translation, 2012. Available online at `http://commons.wikimedia.org/wiki/File:Protein_synthesis.svg`; visited on September 30th 2014.

[KG92] Paul K. Keese and Adrian Gibbs. Origins of genes: "big bang" or continuous creation? *Proceedings of the National Academy of Sciences of the United States of America*, 89(20):9489–9493, October 1992.

[Kim68] Motoo Kimura. Evolutionary Rate at the Molecular Level. *Nature*, 217(5129):624–626, 1968.

[Kim80] Motoo Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16:111–120, 1980.

[KKBS14] Johannes Georg Klotz, David Kracht, Martin Bossert, and Steffen Schober. Canalizing Boolean Functions Maximize Mutual Information. *IEEE Transactions on Information Theory*, 60(4):2139–2147, 2014.

[KN09] Eugene V. Koonin and Artem S. Novozhilov. Origin and evolution of the genetic code: the universal enigma. *International Union of Biochemistry and Molecular Biology Life*, 61(2):99–111, February 2009.

[KP08] Sergey Kryazhimskiy and Joshua B. Plotkin. The population genetics of dN/dS. *PLoS Genetics*, 4(12):e1000304+, December 2008.

[Kra00] David C. Krakauer. Stability and Evolution of Overlapping Genes. *Evolution; International Journal of Organic Evolution*, 54(3):731–739, 2000.

[Kra02] David C. Krakauer. Evolutionary principles of genomic compression. *Comments Theoretical Biology*, 7:215–236, 2002.

[KSP+09] Wook Kim, Mark W. Silby, Sam O. Purvine, Julie S. Nicoll, Kim K. Hixson, Matt Monroe, Carrie D. Nicora, Mary S. Lipton, and Stuart B. Levy. Proteomic detection of non-annotated protein-coding genes in *Pseudomonas fluorescens* Pf0-1. *PloS ONE*, 4(12):e8455+, December 2009.

[KTR+04] Janakiraman Krishnamurthy, Chad Torrice, Matthew R. Ramsey, Grigoriy I. Kovalev, Khalid Al-Regaiey, Lishan Su, and Norman E. Sharpless. Ink4a/Arf expression is a biomarker of aging. *The Journal of Clinical Investigation*, 114(9):1299–1307, November 2004.

[LG98] Pietro Liò and Nick Goldman. Models of molecular evolution and phylogeny. *Genome Research*, 8(12):1233—1244, December 1998.

[Li93] Wen-Hsiung Li. Unbiased estimation of the rates of synonimous and nonsynonimous substitution. *Journal of Molecular Evolution*, 36:96–99, 1993.

[Li99] Wentian Li. Statistical properties of open reading frames in complete genome sequences. *Computers and Chemistry*, 23(3-4):283–301, June 1999.

[LK07] Fabrizio Lillo and David C. Krakauer. A statistical analysis of the threefold evolution of genomic compression through frame overlaps in prokaryotes. *Biology Direct*, 2:22+, September 2007.

[LLB+01] Eric S. Lander, Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jeniffer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, and William et al. FitzHugh. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001.

[LS03] Scott W. Lowe and Charles J. Sherr. Tumor suppression by Ink4a–Arf: progress and puzzles. *Current Opinion in Genetics & Development*, 13(1):77–83, 2003.

[LSBGO98] Wentian Li, Gustavo Stolovitzky, Pedro Bernaola-Galvan, and Jose L. Oliver. Compositional heterogeneity within, and uniformity between, DNA sequences of yeast chromosomes. *Genome Research*, 9:916–928, 1998.

[LT85] Peter Lancaster and Miron Tismenetsky. *The theory of matrices: with applications*. Computer science and applied mathematics. Academic Press, Orlando, 1985.

[LWL85] Wen-Hsiung Li, Chung-I Wu, and Chi-Cheng Luo. A new method for estimating synonymous and non-synonymous rates of nucleotide substitutions considering the relative likelihood of nucleotide and codon changes. *Molecular Biology and Evolution*, 2:150–174, 1985.

[MA09] Carolin Kosiol Maria Anisimova. Investigating Protein-Coding Sequence Evolution with Probabilistic Codon Substitution Models. *Molecular Biology and Evolution*, (2):255–271, 2009.

[MAG09] Michael W. McCoy, Andrew P. Allen, and James F. Gillooly. The random nature of genome architecture: Predicting Open Reading Frame distributions. *PLoS ONE*, 4(7):e6456, July 2009.

[MB14] Nancy A. Moran and Gordon M. Bennett. The Tiniest Tiny Genomes. *Annual Review of Microbiology*, 68:195–215, September 2014.

[MBPB94] Enrique Merino, Paulina Balbás, José L. Puente, and Francisco Bolivar. Antisense overlapping open reading frames in genes from bacteria to humans. *Nucleic Acids Research*, 22(10):1903–1908, 1994.

[Med14] Nobelprize.org. Nobel Media. The Nobel Prize in Physiology or Medicine 1962, 2014. Available online at `http://www.nobelprize.org/nobel_prizes/medicine/laureates/1962/`; visited on October 17th 2014.

[Mey01] Carl D. Meyer. *Matrix Analysis and Applied Linear Algebra*. SIAM, 2001.

[MFS+00] Anette McVeigh, Alessio Fasano, Daniel A. Scott, Sandra Jelacic, Steve L. Moseley, Donald C. Robertson, and Stephen J. Savarino. IS1414, an *Escherichia coli* insertion sequence with a heat-stable enterotoxin gene embedded in a transposase-like gene. *Infection and Immunity*, 68(10):5710–5715, 2000.

[MG94] Spencer V. Muse and Brandon S. Gaut. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution*, 11(5):715–724, 1994.

[MGA83] Julio Montoya, George L. Gaines, and Guiseppe Attardi. The pattern of transcription of the human mitochondrial rRNA genes reveals two overlapping transcription units. *Cell*, 34(1):151 – 159, 1983.

[MGR93] Dominique Missiakas, Costa Georgopoulos, and Satish Ratina. The *Escherichia coli* Heat Shock Gene *htpY*: Mutational Analysis, Cloning, Sequencing, and Transcriptional Regulation. *Journal of Bacteriology*, 175(9):2613–2624, 1993.

[Mis09] Manoj Mishra. Tutorial on "An Algorithm for Computing the Capacity of Arbitrary Discrete Memoryless Channels" by Suguru Arimoto. Website, 2009. Available online at `https://www3.nd.edu/~jnl/ee80653/tutorials/manoj.pdf`; visited on September 18th 2014.

[MLM05] Izabela Makalowska, Chiao F. Lin, and Wojciech Makalowski. Overlapping genes in vertebrate genomes. *Computational Biology and Chemistry*, 29:1–2, February 2005.

[MOM01] Alex Mira, Howard Ochman, and Nancy A. Moran. Deletional bias and the evolution of bacterial genomes. *Trends in Genetics*, 17(10):589–596, October 2001.

[MS77] Florence J. MacWilliams and Neil J.A. Sloane. *The Theory of Error–Correcting Codes*. North-Holland, Amsterdam, Netherlands, 1977.

[MV04] Olgica Milenkovic and Bane Vasic. Information theory and coding problems in genetics. In *Information Theory Workshop, 2004. IEEE*, pages 60–65, October 2004.

[MVL03] Cleve Moler and Charles Van Loan. Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later. *SIAM Review*, 45(1):3–49, January 2003.

[MY78] Takashi Miyata and Teruo Yasunaga. Evolution of overlapping genes. *Nature*, 272:532–535, 1978.

[MY80] Takashi Miyata and Teruo Yasunaga. Molecular Evolution of mRNA: A Method for estimating Evolutionary Rates of Synonymous and Amino Acid Substitutions from Homologous Nucleotide Sequences and its Application. *Genetics*, 16:641–657, 1980.

[NBC+05] Rasmus Nielsen, Carlos Bustamante, Andrew G. Clark, Stephen Glanowski, Timothy B. Sackton, Melissa J. Hubisz, Adi Fledel-Alon, David M. Tanenbaum, Daniel Civello, Thomas J. White, John J. Sninsky, Mark D. Adams, and Michele Cargill. A Scan for Positively Selected Genes in the Genomes of Humans and Chimpanzees. *PLoS Biology*, 3(6):e170, 2005.

[NBE+83] Staffan Normark, Sven Bergström, Thomas Edlund, Thomas Grundström, Bengtake Jaurin, Frederik P. Lindberg, and Olof Olsson. Overlapping genes. *Annual Review of Genetics*, 17:499–525, 1983.

[NG86] Masatoshi Nei and Takashi Gojobori. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, 3(5):418–426, September 1986.

[NM65] John A. Nelder and Roger Mead. A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313, 1965.

[Nor98] James R. Norris. *Markov Chains (Cambridge Series in Statistical and Probabilistic Mathematics)*. Cambridge University Press, 1998.

[OM96] Jose L. Oliver and Antonio Marn. A relationship between GC content and coding-sequence length. *Journal of Molecular Evolution*, 43:216–223, 1996.

[oM13] United States National Library of Medicine. National Center for Biotechnology Information, 2013. Available online at `http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi#SG11`; visited on October 1st 2014.

[oM14a] United States National Library of Medicine. GenBank, 2014. Available online at `http://www.ncbi.nlm.nih.gov/genbank`; visited on March 5th 2014.

[oM14b] United States National Library of Medicine. National Center for Biotechnology Information, 2014. Available online at `http://http://www.ncbi.nlm.nih.gov`; visited on March 5th 2014.

[Par04] Emanuel Parzen. Quantile Probability and Statistical Data Modeling. *Statistical Science*, 19(4):652–662, November 2004.

[Pat99] László Patthy. Genome evolution and the evolution of exon-shuffling–a review. *Gene*, 238(1):103–114, September 1999.

[Pav06] Angelo Pavesi. Origin and evolution of overlapping genes in the family *Microviridae*. *Journal of General Virology*, 87(4):1013–1017, 2006.

[PB93] Pekka Pamilo and Nestor O. Bianchi. Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. *Molecular Biology and Evolution*, 10(2):271–81, 1993.

[PDIGP97] Angelo Pavesi, Bettina De Iaco, Maria Ilde Granero, and Alfredo Porati. On the Informational Content of Overlapping Genes in Prokaryotic and Eukaryotic Viruses. *Journal of Molecular Evolution*, 44(6):625–631, 1997.

[PDMS10] Sergei Kosakovsky Pond, Wayne Delport, Spencer V. Muse, and Konrad Scheffler. Correcting the Bias of Empirical Frequency Parameter Estimators in Codon Models. *PLoS ONE*, 7(5):e11230, 2010.

[PEL+80] Francine Perler, Argiris Efstratiadis, Peter Lomedico, Walter Gilbert, Richard Kolodner, and Jerry Dodgson. The evolution of genes: the chicken preproinsulin gene. *Cell*, 20(2):555 – 566, 1980.

[PFM+04] Itsik Pe'er, Clifford E. Felder, Orna Man, Israel Silman, Joel L. Sussman, and Jacques S. Beckmann. Proteomic signatures: Amino acid and oligopeptide compositions differentiate among phyla. *Proteins: Structure, Function, and Bioinformatics*, 54(1):20–40, 2004.

[PHB08] Albert Palleja, Eoghan Harrington, and Peer Bork. Large gene overlaps in prokaryotic genomes: result of functional constraints or mispredictions? *BMC Genomics*, 9(1):335+, July 2008.

[PJ01] Anne-Mette Pedersen and Jens L. Jensen. A dependent-rates model and an MCMC-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. *Molecular Biology and Evolution*, 18(5):763–76, 2001.

[PPB+01] Nicole T. Perna, Guy Plunkett, Valerie Burland, Bob Mau, Jeremy D. Glasner, Debra J. Rose, George F. Mayhew, Peter S. Evans, Jason Gregor, Heather A. Kirkpatrick, Gyorgy Posfai, Jeremiah Hackett, Sara Klink, Adam Boutin, Ying Shao, Leslie Miller, Erik J. Grotbeck, N. Wayne Davis, Alex Lim, Eileen T. Dimalanta, Konstantinos D. Potamousis, Jennifer Apodaca, Thomas S. Anantharaman, Jieyi Lin, Galex Yen, David C. Schwartz, Rodney A. Welch, and Frederick R. Blattner. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature*, 409:529–533, 2001.

[Pri14] Gravetter J. Privitera. *Statistics for the Behavioral Sciences*. SAGE Publications, 2014.

[PWC98] Annet-Mette K. Pedersen, Carsten Wiuf, and Freddy B. Christiansen. A codon-based model designed to describe lentiviral evolution. *Molecular Biology and Evolution*, 15:1069–1081, 1998.

[Raz06] Shmuel Razin. The genus *Mycoplasma* and related genera (class mollicutes). In M. Dworkin, S. Falkow, E. Rosenberg, K.-H. Schleifer, and E. Stackebrandt, editors, *The Prokaryotes: A Handbook on the Biology of Bacteria*, number 3rd edition, pages 896–904. New York: Springer, 2006.

[REHA11] N. Rodríguez-Ezpeleta, M. Hackenberg, and A.M. Aransay. *Bioinformatics for High Throughput Sequencing*. SpringerLink: Bücher. Springer, 2011.

[Ros10] Sheldon M. Ross. *Introduction to Probability Models, Tenth Edition*. Academic Press, 2010.

[Rot06] Ron Roth. *Introduction to Coding Theory*. Cambridge University Press, Cambridge, MA, USA, 2006.

[RSS+02] Igor B. Rogozin, Alexey N. Spiridonov, Alexander V. Sorokin, Yuri I. Wolf, I. King Jordan, Roman L. Tatusov, and Eugene V. Koonin. Purifying and directional selection in overlapping prokaryotic genes. *Trends in Genetics*, 18(5):228 – 232, 2002.

[RT08] Oleg Reva and Burkhard Tümmler. Think big – giant genes in bacteria. *Environmental Microbiology*, 10:768–777, 2008.

[SAB+77] Frederick Sanger, G. M. Air, Bart G. Barrell, N. L. Brown, A. R. Coulson, C. A. Fiddes, C. A. Hutchison, P. M. Slocombe, and M. Smith. Nucleotide sequence of bacteriophage $\Phi X174$ DNA. *Nature*, 265(5596):687–695, 1977.

[Sch06] Steffen Schober. Stability of Attractor Cycles in Random Boolean Networks. In *European Conference on Complex Systems (ECCS 06)*, September 2006.

[SDHM05] Michel Sarkis, Zaher Dawy, Joachim Hagenauer, and Jakob C. Müller. Gene Clustering using Independent Component Analysis. In *IEEE International Workshop on Genomic Signal Processing and Statistics*, Newport, USA, May 2005.

[SFD+82] Christine Sacerdot, Guy Fayat, Philippe Dessen, Mathias Springer, Jacqueline A. Plumbridge, Marianne Gninberg-Manago, , and Sylvain Blanquetl. Sequence of a 1.26-kb DNA fragment containing the structural gene for *E.coli* initiation factor IF3: presence of an AUU initiator codon. *The EMBO Journal*, 1(3):311–315, 1982.

[SG00] Dmitry V. Scherbakov and Maria B. Garber. Overlapping genes in bacterial and phage genomes. *Molecular Biology*, 34(4):485–495, 2000.

[SGH86] Charlotte A. Spencer, R. Daniel Gietz, and Ross B. Hodgetts. Overlapping transcription units in the dopa decarboxylase region of *Drosophila*. *Nature*, 322:279–281, 1986.

[SGL08] Niv Sabath, Dan Graur, and Giddy Landan. Same-strand overlapping genes in bacteria: compositional determinants of phase bias. *Biology Direct*, 3:36+, August 2008.

[Sha40] Claude E. Shannon. *An algebra for theoretical genetics*. PhD thesis, Massachusetts Institute of Technology, Dept. of Mathematics, 1940.

[Sha48] Claude E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27:379–423, 623–656, October 1948.

[She01] Charles J. Sherr. The INK4a/ARF network in tumour suppression. *Nature Reviews*, 2:731–737, 2001.

[She06] Charles J. Sherr. Divorcing ARF and p53: an unsettled case. *Nature Reviews*, 6:663–673, 2006.

[Sil97] John Silke. The majority of long non-stop reading frames on the antisense strand can be explained by biased codon usage. *Gene*, 194(1):143–155, 1997.

[SL08] Mark W. Silby and Stuart B. Levy. Overlapping protein-encoding genes in *Pseudomonas fluorescens* Pf0-1. *PLoS Genetics*, 4:e1000094, 2008.

[SP80] Robert A. Smith and John S. Parkinson. Overlapping genes at the cheA locus of *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America*, 77(9):5370—5374, September 1980.

[SP04] Hervé Seligmann and David D. Pollock. The ambush hypothesis: hidden stop codons prevent off-frame gene reading. *DNA and cell biology*, 23(10):701–705, October 2004.

[Spo10] Wikimedia Commons Sponk. Comparison of a single-stranded RNA and a double-stranded DNA with their corresponding nucleobases, 2010. Available online at `http://commons.wikimedia.org/wiki/File:Difference_DNA_RNA-EN.svg`; visited on September 25th 2014.

[SRL04] Mark W. Silby, Paul B. Rainey, and Stuart B. Levy. IVET experiments in *Pseudomonas fluorescens* reveal cryptic promoters at loci associated with recognizable overlapping genes. *Microbiology*, 150:518–520, 2004.

[SRS03] Wojciech Szpankowski, Wenhui Ren, and Lukasz Szpankowski. An optimal DNA segmentation based on the MDL principle. In *Bioinformatics Conference, 2003. CSB 2003. Proceedings of the 2003 IEEE*, pages 541–546, August 2003.

[SS79] Christian Sander and Georg E. Schulz. Degeneracy of the information contained in amino acid sequences: Evidence from overlaid genes. *Journal of Molecular Evolution*, 13(3):245–252, 1979.

[SS90] Thomas D. Schneider and R. Michael Stephens. Sequence Logos: A New Way to Display Consensus Sequences. *Nucleic Acids Research*, 18:6097–6100, 1990.

[SSVC05] Kishore R. Sakharkar, Meena K. Sakharkar, Chandra Verma, and Vincent T. K. Chow. Comparative study of overlapping genes in bacteria, with special reference to *Rickettsia prowazekii* and *Rickettsia conorii*. *International Journal of Systematic and Evolutionary Microbiology*, 55(Pt 3):1205–9, 2005.

[Sul09] Wikimedia Commons Sulai. DNA, transcription, 2009. Available online at `http://commons.wikimedia.org/wiki/File:DNAtranscription.svg`; visited on September 29th 2014.

[SW81] Willi-Hans Steeb and F. Wilhelm. Exponential functions of Kronecker products and trace calculation. *Linear and Multilinear Algebra*, 9(4):345–346, 1981.

[SWTW10] Tamara Schenekar, Kathrin A Winkler, Jennifer L Troyer, and Steven Weiss. Isolation and Characterization of the CYP2D6 Gene in Felidae with Comparison to Other Mammals. *Journal of Molecular Evolution*, 72:222–231, 2010.

[Tav86] Simon Tavaré. Some probabilistic and statistical problems on the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, 17:57–86, 1986.

[Tay90] Richard Taylor. Interpretation of the Correlation Coefficient: A Basic Review. *Journal of Defense Modeling and Simulation*, 1:35–39, 1990.

[TBCM09] Sedef Tunca, Carlos Barreiro, Juan-Jose R. Coque, and Juan F. Martin. Two overlapping antiparallel genes encoding the iron regulator DmdR1 and the Adm proteins control siderophore and antibiotic biosynthesis in *Streptomyces coelicolor* A3(2). *FEBS Journal*, 276(17):4814–4827, 2009.

[TK81] Naoyuki Takahata and Motoo Kimura. A model of evolutionary based substitutions and its application with special reference to rapid change of pseudogenes. *Genetics*, 98(3):641–657, 1981.

[TN93] Koichiro Tamura and Masatoshi Nei. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, 10:512–526, 1993.

[VAM+01] J. Craig Venter, Mark D. Adams, Eugene W. Myers, Peter W. Li, Richard J. Mural, Granger G. Sutton, Hamilton O. Smith, Mark Yandell, Cheryl A. Evans, and Robert A. et al. Holt. The Sequence of the Human Genome. *Science*, 291(5507):1304–1351, 2001.

[VMG+04] Vamsi Veeramachaneni, Wojckiech Makalowski, Michael Galdzicki, Raman Sood, and Izabela Makalowska. Mammalian overlapping genes: the comparative perspective. *Genome Research*, 14:280–286, 2004.

[WAFS10] Andrew Warren, Jeremy Archuleta, Wu C. Feng, and Joao Setubal. Missing genes in the annotation of prokaryotic genomes. *BMC Bioinformatics*, 11(1):131, March 2010.

[WC53] James D. Watson and Francis H. Crick. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171:737–738, 1953.

[Wei14] Eric W. Weisstein. Hypothesis Testing, 2014.

[WF86] Trevor Williams and Mike Fried. A mouse locus at which transcription from both DNA strands produces mRNAs complementary at their 3′ ends. *Nature*, 322:275–279, 1986.

[WG68] Maike B. Wilk and Ram Gnanadesikan. Probability plotting methods for the analysis of data. *Biometrika*, 55(1):1–17, March 1968.

[WG01] Simon Whelan and Nick Goldman. A general empirical model of protein evolution derived from multiple protein families using a maximum likelihood approach. *Molecular Biology and Evolution*, 18:691–699, 2001.

[WHD+07] Johanna Weindl, Pavol Hanus, Zaher Dawy, Jürgen Zech, Joachim Hagenauer, and Jakob C. Müller. Modeling DNA-binding of Escherichia coli sigma70 exhibits a characteristic energy landscape around strong promoters. *Nucleic Acid Research*, October 2007.

[Xia83] Xuhua Xia. The Rate Heterogeneity of Nonsynonymous Substitutions in Mammalian Mitochondrial Genes. *Molecular Biology*, pages 336–344, 1983.

[Yan94] Ziheng Yang. Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution*, 10:105–111, 1994.

[Yan06] Ziheng Yang. *Computational molecular evolution*. Oxford University Press, Oxford, 2006.

[YB00] Ziheng Yang and Joseph P. Bielawski. Statistical methods for detecting molecular adaptation. *Trends In Ecology and Evolution*, 15(12):496–503, December 2000.

[YN00] Ziheng Yang and Rasmus Nielsen. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution*, 17(1):32–43, January 2000.

[YNH98] Ziheng Yang, Rasmus Nielsen, and Masami Hasegawa. Models of amino acid substitution and applications to mitochondrial protein evolution. *Molecular Biology and Evolution*, 15(12):1600–1611, 1998.

[Yoc74] Hubert P. Yockey. An application of information theory to the central dogma and the sequence hypothesis. *Journal of Theoretical Biology*, 46(2):369 – 406, 1974.

[Yoc92] Hubert P. Yockey. *Information Theory and Molecular Biology.* Cambridge University Press, 1992.

[Yoc05] Hubert P. Yockey. *Information Theory, Evolution, and the Origin of Life.* Cambridge University Press, New York, 2005.

[YSR+07] Shibu Yooseph, Granger Sutton, Douglas B. Rusch, Aaron L. Halpern, Shannon J. Williamson, Karin Remington, Jonathan A. Eisen, Karla B. Heidelberg, Gerard Manning, Weizhong Li, Lukasz Jaroszewski, Piotr Cieplak, Christopher S. Miller, Huiying Li, Susan T. Mashiyama, Marcin P. Joachimiak, Christopher van Belle, John-Marc M. Chandonia, David A. Soergel, Yufeng Zhai, Kannan Natarajan, Shaun Lee, Benjamin J. Raphael, Vineet Bafna, Robert Friedman, Steven E. Brenner, Adam Godzik, David Eisenberg, Jack E. Dixon, Susan S. Taylor, Robert L. Strausberg, Marvin Frazier, and J. Craig Venter. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biology*, 5(3):e16+, March 2007.

[Zac] Wikispaces AP Biology Zack. What is a transcription unit? Available online at `http://bioap.wikispaces.com/Ch+17+Collaboration`; visited on September 29th 2014.

[ZD13] Huamin Zhang and Feng Ding. On the Kronecker Products and Their Applications. *Journal of Applied Mathematics*, 2013, 2013.

[Zha94] Andrey Zharkikh. Estimation of evolutionary distance between nucleotide sequences. *Journal of Molecular Evolution*, 39:315–329, 1994.

[Zha00] Jianzhi Zhang. Protein-length distributions for the three domains of life. *Trends in Genetics*, 16(3):107–109, March 2000.

[ZLY06] Zhang Zhang, Jun Li, and Jun Yu. Computing Ka and Ks with a consideration of unequal transitional substitutions. *BMC Evolutionary Biology*, 6(1):44+, June 2006.

[ZP65] Emile Zuckerkandl and Linus Pauling. Evolutionary divergence and convergence in proteins. *Evolving Genes and Proteins*, 97:97–166, 1965.

[ZSF+02] Yu Zheng, Joseph D. Szustakowski, Lance Fortnow, Richard J. Roberts, and Simon Kasif. Computational Identification of Operons in Microbial Genomes. *Genome Research*, 12:1221–1230, 2002.

[ZY06] Zhang Zhang and Jun Yu. Evaluation of Six Methods for Estimating Synonymous and Nonsynonymous Substitution Rates. *Genomics, Proteomics & Bioinformatics*, 4(3):173–181, 2006.

Nomenclature

List of Acronyms

aORF	Annotated ORF
bp	Base pairs
CDF	Cumulative distribution function
CTMC	Continuous-time Markov chain
DMC	Discrete memoryless channel
DNA	Deoxyribonucleic acid
DTMC	Discrete-time Markov chain
EDF	Empirical distribution function
EHEC	Escherichia coli bacterium
IID	Independent and identically distributed (random variable)
IUD	Independent and uniformly distributed (random variable)
mRNA	Messenger RNA
naORF	Not annotated ORF
NCBI	National Center for Biotechnology Information
NLM	United States National Library of Medicine
oORF	Overlapping ORF
ORF	Open reading frame
PMF	Probability mass function
RNA	Ribonucleic acid
RND	Random amino acid distribution
sORF	Shadow ORF
STD	Standard genetic code
tRNA	Transfer RNA

Notation

X	Capital character denotes random variable
x	Small character denotes realization of random variable X
$(0,1)$	Real number between 0 and 1
$D(\cdot\|\|\cdot)$	Kullback-Leibler divergence
$F_X(\cdot)$	Cumulative distribution function
$F_n(\cdot)$	Empirical distribution function
$H(\cdot)$	Entropy
$H(\cdot\|\cdot)$	Conditional entropy
$I(\cdot;\cdot)$	Mutual information
$P\left(X=x\right)$	Probability of event $X=x$
$Q(r)$	Quantile function of r
$[a,b)$	Set of integers $\{a, a+1, \ldots b-1\}$
X_n	Random variable with time index n
x_n	Realization of random variables with time index n
$\{X_n\}_{n\geq 0}$	Discrete-time stochastic process time index n
$\{X_t\}_{t\geq 0}$	Continuous-time stochastic process time index t
$\mathcal{X}$	Calligraphic character denotes finite set or alphabet
σ	Standard deviation
σ^2	Variance
σ_{XY}	Covariance
ϱ_{XY}	Correlation coefficient
$p_X(x)$	Probability mass function
r_{xy}	Empirical correlation coefficient
$\|\cdot\|$	Cardinality of a finite set or alphabet
$\|\|\cdot-\cdot\|\|$	Euclidean distance or ℓ^2 norm
$\mathbf{E}\left[\cdot\right]$	Expected value
$\mathrm{d_H}\left[\cdot,\cdot\right]$	Hamming distance
$\boldsymbol{X}$	Matrix $\boldsymbol{A}$ with entries X_{ij}
$\boldsymbol{x}$	(Row) vector $\boldsymbol{x}=(x_0x_1\ldots x_{n-1})$ of length n

List of Symbols

A	Adenine nucleotide
B_{SV}^{GC}, B_{SV}	Survival bound
C	Cytosine nucleotide
G	Guanine nucleotide
GC	Number of bases G and C related to N_{BP}
$P\left(L\geq\ell'\right)$	Probability of ORF with length $\geq\ell'$

$P^f(L=\ell)$	Probability of ORF with length ℓ per reading frame f
$P^f_{SV}(\ell)$	Survival probability per reading frame f
T	Thymine nucleotide
U	Uracile nucleotide
$\mathcal{F}_A = \{-1, \pm 2, \pm 3\}$	Alternative reading frames without $+1$
$\mathcal{MC}_{CT}(\boldsymbol{\pi}, \boldsymbol{Q})$	Continuous-time Markov chain determined by $\boldsymbol{\pi}$ and $\boldsymbol{Q}$
$\mathbf{E}[L] = \bar{\ell}_{ORF}$	Average ORF length
$\mathcal{MC}(\boldsymbol{\pi}, \boldsymbol{P})$	Discrete-time Markov chain determined by $\boldsymbol{\pi}$ and $\boldsymbol{P}$
$\mathcal{MC}_0(\boldsymbol{\pi})$	Discrete-time Markov chain without memory
$\mathbb{R}_0^+$	Set of non-negative real numbers
$\mathcal{F} = \{\pm 1, \pm 2, \pm 3\}$	Reading frames
$\boldsymbol{P}^{(f)}(t)$	Transition matrix of CTMC in reading frame f at time t
$\boldsymbol{Q}^{(f)}$	Rate matrix of CTMC in reading frame f
$\boldsymbol{P}^{(f)}_{abs}$	Absorbing Markov chain per reading frame f
$\mathbb{Z}^+$	Set of positive integers $\{1, 2, \ldots\}$
$\mathbb{Z}_0^+$	Set of positive integers with zero $\{0, 1, 2, \ldots\}$
$\mathcal{A}_{21}$	Amino acid alphabet with stop label
$\mathcal{A}_{20}$	Amino acid alphabet without stop label
$\mathcal{C}_{64}$	Codon alphabet
$\mathcal{C}_{61}$	Codon alphabet without stop codons
$\mathcal{C}^{(f)}$	Channel capacity per reading frame f
N_{BP}	Length of sequence in base pairs
N_C	Length of sequence in codons
$\mathcal{G}$	Standard genetic code
$\boldsymbol{p}^{(k)}$	State distribution of Markov chain after k steps
$\boldsymbol{\pi}^{(f)}$	Codon usage per reading frame f
$\boldsymbol{\pi}^{(f)}_{\mathcal{A}_{21}}$	Amino acid distribution per reading frame f
$i \in \mathcal{S}$	State of Markov chain
$n^{(f)}_{ORF}$	Expected number of ORFs per reading frame f
n^{rel}_{ORF}	Relative number of ORFs expected
$p^{(f)}_{start}$	Start codon probability per reading frame f
$p^{(f)}_{stop}$	Stop codon probability per reading frame f
$t^{(f)}_A$	Mean absorbing time per reading frame f
$t^{(f)}_M$	Mean recurrence time per reading frame f
$\mathbb{R}$	Set of real numbers
$\mathcal{S}$	State space of Markov chain
$\boldsymbol{P}$	Transition matrix of Markov chain with element $\{P_{ij}\}_{i,j \in \mathcal{S}}$
$\boldsymbol{Q}$	Rate matrix of Markov chain with element $\{Q_{ij}\}_{i,j \in \mathcal{S}}$
$\boldsymbol{P}^{(f)}$	Transition matrix of Markov chain in reading frame f
$\mathbb{Z}$	Set of integers $\{\ldots, -2, -1, 0, 1, 2, \ldots\}$
$\mathcal{B}$	Nucleotide alphabet $\{A, C, G, T\}$

$\boldsymbol{\pi}$	Stationary distribution of a Markov chain

Katharina Schilling

Der Lebenslauf wurde aus Gründen des Datenschutzes entfernt.

✉ *katharina.mir@uni-ulm.de*

List of Publications

Journal Publications

Katharina Mir, Klaus Neuhaus, Martin Bossert, and Steffen Schober. Short Barcodes for Next Generation Sequencing. *PLoS ONE*, 8(12):e82933, December 2013.

Katharina Mir, Klaus Neuhaus, Siegfried Scherer, Martin Bossert, and Steffen Schober. Predicting Statistical Properties of Open Reading Frames in Bacterial Genomes. *PLoS ONE*, 7(9):e45103, September 2012.

Katharina Mir and Steffen Schober. Selection Pressure in Alternative Reading Frames. *PLoS ONE*, 9(10):e108768, October 2014.

Conference Contributions

Katharina Mir and Steffen Schober. Investigation of genetic code optimality for overlapping protein coding sequences. In *Proc. 8th International Symposium on Turbo Codes and Iterative Information Processing (ISTC)*, Bremen, Germany, August 2014.

Steffen Schober and Katharina Mir. Inference of Boolean networks using spectral techniques. In *Proc. Sixth International Workshop on Computational Systems Biology (TICSP Series)*, pages 147–151, Aarhus, Denmark, June 2009.

Steffen Schober, Katharina Mir, and Martin Bossert. Reconstruction of Boolean genetic regulatory networks consisting of canalyzing or low sensitivity functions. In *Proc. International ITG Conference on Source and Channel Coding*, Siegen, Germany, January 2010.

Steffen Schober, Katharina Mir, Klaus Neuhaus, and Martin Bossert. Design of short barcodes for next generation sequencing of DNA and RNA. In *Genomic Signal Processing and Statistics (GENSIPS)*, pages 31–34, Washington DC, December 2012.

Place, Date

Katharina Schilling

✉ *katharina.mir@uni-ulm.de*