JACOBS
UNIVERSITY

The DNA from a Communications and Information Theoretic Perspective

by

Dawit Andualem Nigatu

A thesis submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
in Electrical Engineering

Approved Dissertation Committee:

Prof. Dr.-Ing. Werner Henkel
Jacobs University Bremen

Prof. Dr. Marc-Thorsten Hütt
Jacobs University Bremen

Prof. Dr. Georgi Muskhelishvili
Agricultural University of Georgia

Prof. Dr. Hesham H. Ali
University of Nebraska at Omaha

Date of Defence: March 26, 2018

Computer Science and Electrical Engineering

Statutory Declaration

Family Name, Given/First Name:	Nigatu, Dawit Andualem
Matriculation number:	20330500
What kind of thesis are you submitting: Bachelor-, Master- or PhD-Thesis	PhD-Thesis

English: Declaration of Authorship

I hereby declare that the thesis submitted was created and written solely by myself without any external support. Any sources, direct or indirect, are marked as such. I am aware of the fact that the contents of the thesis in digital form may be revised with regard to usage of unauthorized aid as well as whether the whole or parts of it may be identified as plagiarism. I do agree my work to be entered into a database for it to be compared with existing sources, where it will remain in order to enable further comparisons with future theses. This does not grant any rights of reproduction and usage, however.

This document was neither presented to any other examination board nor has it been published.

German: Erklärung der Autorenschaft (Urheberschaft)

Ich erkläre hiermit, dass die vorliegende Arbeit ohne fremde Hilfe ausschließlich von mir erstellt und geschrieben worden ist. Jedwede verwendeten Quellen, direkter oder indirekter Art, sind als solche kenntlich gemacht worden. Mir ist die Tatsache bewusst, dass der Inhalt der Thesis in digitaler Form geprüft werden kann im Hinblick darauf, ob es sich ganz oder in Teilen um ein Plagiat handelt. Ich bin damit einverstanden, dass meine Arbeit in einer Datenbank eingegeben werden kann, um mit bereits bestehenden Quellen verglichen zu werden und dort auch verbleibt, um mit zukünftigen Arbeiten verglichen werden zu können. Dies berechtigt jedoch nicht zur Verwendung oder Vervielfältigung.

Diese Arbeit wurde noch keiner anderen Prüfungsbehörde vorgelegt noch wurde sie bisher veröffentlicht.

_____

Date, Signature

Dedication

To my late grandmother Biritu Bedada

Abstract

Beyond enabling the successful development of communication engineering, information theory has far-reaching applications in other disciplines, including molecular biology. Information theory has been effectively applied for analyzing and modeling biological systems and processes. Following the same framework, in this thesis, three related but distinct topics are studied. First, we modeled the transmission of genetic information assuming a codon-based mutation matrix as a communication channel and performed capacity computations. Furthermore, the severity of codon substitution errors was assessed by comparing mutation probabilities with chemical properties of amino acids using a dimension reduction technique. The second topic deals with the analysis of the relationship between the digital and analog information in bacterial genomes. The latter represents the three-dimensional information encoded by the physicochemical properties of the DNA. Here, the analog information is associated with thermodynamic stability. In addition, the spatial genomic sequence organization is studied in relation to selected functional classes of genes. Finally, a novel method of essential gene prediction based on machine-learning is proposed. Information-theoretic measures have been used as features and essentiality predictions were performed in both prokaryotes and eukaryotes. The obtained results show that gene essentiality annotations can be reliably transferred between both closely and distantly related species.

Acknowledgement

This dissertation is the result of my research in the transmission systems group at Jacobs University Bremen. Several people, whom I am very grateful to, have contributed to the successful conclusion of this work.

First and foremost, I would like to express my heartfelt thanks to my supervisor Prof. Dr.-Ing. Werner Henkel for his help, excellent supervision, and for the freedom I was granted throughout these years. I am very grateful to his constant encouragement, kind support, and valuable discussions. He has been a great mentor. Furthermore, I am grateful to Prof. Dr. Georgi Muskhelishvili, Prof. Dr. Marc-Thorsten Hütt, and Prof. Dr. Hesham H. Ali for accepting to join the dissertation committee and for taking the time to evaluate this dissertation.

I would also like to thank my current and former colleagues Nazia Islam, Oana Graur, Vladimir Burstein, Khodr Saaifan, and Attiya Mahmood for the friendship, fruitful discussions, and creating and maintaining a stimulating working environment. In addition, I have to express my gratitude to Prof. Dr. Georgi Muskhelishvili and Dr. Patrick Sobetzko for the successful interdisciplinary cooperation. I am also thankful to my friends, inside Bremen and all over the world, for making my life enjoyable and sharing all the good memories.

Special thanks goes to my late grandmother. I wouldn't have come this far if it was not for her efforts. I am also indebted to my relatives for the love, inspiration, and support.

I owe a great deal of gratitude to my dearest wife and best friend, Rahma. Without her love, encouragement, consistent prayers, and understanding, this work would not have become what it is. Sweetheart, I won't forget the sacrifices you made. I would also like to thank my little princess, Iman, for making the last year full of joy and happiness.

Finally, I would like to thank the German Research Foundation (DFG) for their financial support.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background and Motivation

In 1948, Claude Shannon founded information theory with his seminal paper *A Mathematical Theory of Communication* [Sha48]. Shannon provided a quantitative measure of information and a theoretical framework for communication systems. Since he defined information in relation to the probabilistic descriptions of the information source, it can be easily applied in any discipline. A year later, Henry Quastler started to develop a research field he called "information theory in biology" aiming to apply Shannon's information-theoretic concepts in molecular genetics [Kay00]. In 1953, the double-helical structure of the deoxyribonucleic acid (DNA) was discovered by Watson and Crick [WC+53]. They showed that genetic information is carried by the precise order of four nucleotide pairs. Since then, it has become apparent that information and communication theory can be used to study the transmission, storage, and processing of genetic information. Furthermore, the abstraction of the information contained in the DNA by the sequence of four letters made biology a computational (quantitative) science [Gam54; Yoc05]. Attracted by the Quastler's work, Yockey made a series of extensive studies on information-theoretic description of molecular biology concepts [Yoc74; Yoc92; Yoc05]. In 1953, Yockey and Quastler organized the first *Symposium on Information Theory in Biology* in which topics related to the measurement and storage of information as well as aging and radiation damages were discussed [Yoc+58].

Many researchers have followed this direction of research and used information and communication theory in biology. Interdisciplinary cooperation between biologists, information theorists, and communication engineers have enabled the modeling and analysis of biological sequences and systems. To name a few, Gérard Battail strongly argued about the existence of error correcting codes in the DNA [Bat97; Bat08]. Gatlin [Gat72], May et al. [May+04], Gong et al. [Gon+11], and Roman-Roldan et al. [RR+96] proposed genetic information transmission models of protein synthesis and evolution. Joachim Hagenauer's group employed mutual information computations to infer the links between genomic positions and diseases [Daw+06], has drawn parallels between binding site detection in the DNA and frame synchronization [WH07], and proposed compression methods for multiple genome alignments [Han+10]. Milenkovic and Vasić tried to show connections between gene regulatory

networks and the bipartite graph of an error-correction code, representing the DNA proofreading mechanism [MV04]. Grosse et al. [Gro+00] used mutual information profiles to distinguish between coding and non-coding DNA. Karlin and Mrázek [KM97] extracted phylogenetic signals from dinucleotide frequencies. Schneider developed the commonly used "sequence logos" which show patterns in genetic sequences using information-theoretic measures [SS90].

Due to the advancement in sequencing technologies and other system-based studies such as gene expression and interaction networks, there is an exponential growth in biological data collected in public databases. Hence, today, more than ever, there is a big demand for analysis and model development to extract knowledge. This thesis aims to use concepts from information theory and communication engineering to address selected problems in molecular biology.

1.2 Thesis Outline

The thesis covers three topics. In the first part, a channel model for the protein synthesis is proposed using a codon mutation matrix and the implications of substitutions in terms of chemical properties are analyzed. In the second part, the information content of the DNA is studied with respect to its digital and analog properties. The third topic we address is the prediction of essential and non-essential genes. We propose a novel machine learning based gene essentiality prediction method using information-theoretic feature extraction.

The thesis is structured as follows:

Chapter 2: Basic Concepts
This chapter introduces the fundamental concepts in information theory and molecular biology which will be required for understanding the subsequent chapters. The information-theory part includes Shannon's mathematical description of information and entropy, description of a communication channel, statistical inference, and Markov models and processes. The molecular biology part starts with basic definitions of nucleic acids. Then, basic biological process such as transcription, translation, and replication, are explained. Finally, the different types of permanent changes in genetic sequences, i.e., mutations, are described.

Chapter 3: A Channel Model from a Mutation Matrix
In this chapter, a channel model is proposed for the codon structure using a codon-based mutation matrix. The capabilities of the channel to preserve the genetic information encoded by the 64 codons and their mapping to the 20 amino acids are

analyzed. In addition, the relationship between mutational changes and the effects in the produced protein in terms of chemical properties are studied employing a dimension reduction and clustering technique.

Chapter 4: Digital Information and Thermodynamic Stability in Bacteria

In this chapter, we study the dual-coding nature of the DNA. We look into the relationship between the digital information encoded in the DNA sequences and the analog 3D information, representing thermodynamic stability. The Shannon and Gibbs entropy profiles are used to show the spatial distributions in both complete genomes and protein coding regions. Furthermore, distribution of selected functional classes of genes is associated with the entropies.

Chapter 5: Prediction of Essential Genes

This chapter deals with the prediction of essential genes in the three domains of life, Bacteria, Archaea (summeraized as prokaryotes), and Eukaryotes. We present a simple machine-learning based computational method using easily accessible information-theoretic features. Detailed results are shown using various performance evaluation method procedures. The transferability of gene essentiality annotations within and across multiple species, both closely and distantly related, are discussed.

Chapter 6: Conclusion

This chapter concludes the thesis. The main contributions are summeraized and possible directions for future work are outlined.

1.3 Publications

Parts of this work have been published in the following articles in scientific journals, conference proceedings, and a book chapter:

- D. Nigatu, A. Mahmood, and W. Henkel, "The empirical codon mutation matrix as a communication channel," *BMC Bioinformatics*, vol. 15, no. 80, 2014.

- D. Nigatu, W. Henkel, P. Sobetzko, G. Muskhelishvili, and A. Mahmood, "Relating digital information, thermodynamic stability, and classes of functional genes in E. coli," in *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Dec 2014, pp. 1338–1341.

- D. Nigatu, W. Henkel, P. Sobetzko, and G. Muskhelishvili. “Relationship between digital information and thermodynamic stability in bacterial genomes”. In: *EURASIP Journal on Bioinformatics and Systems Biology* vol. 1, no. 1, 2016.

- D. Nigatu and W. Henkel. “Prediction of essential genes based on machine learning and information theoretic features,” In: *Proceedings of BIOSTEC 2017 - BIOINFORMATICS*. 2017, pp. 81–92

- D. Nigatu, P. Sobetzko, M. Yousef, and W. Henkel. “Sequence-based information-theoretic features for gene essentiality prediction”. In: *BMC Bioinformatics* vol. 18, no. 473, 2017

- W. Henkel, G. Muskhelishvili, D. Nigatu, and P. Sobetzko. “The DNA from a coding perspective”. In: *ed. by Martin Bossert. Springer Lecture Notes in Bioengineering: Information- and Communication Theory in Molecular Biology*, June 20, 2017.

- D. Nigatu and W. Henkel. “Computational identification of essential genes in prokaryotes and eukaryotes”. In: *Springer Book of BIOINFORMATICS 2017*. Springer. 2017, *[Submitted]*

Other published works which are not discussed within this dissertation, but are still related include:

- A. Mahmood, N. Islam, D. Nigatu, and W. Henkel, "DNA inspired bi-directional Lempel-Ziv-like compression algorithms," 2014 8th International Symposium on Turbo Codes and Iterative Information Processing (ISTC), Bremen, 2014, pp. 162-166.

- M. Yousef, D. Nigatu, D. Levy, J. Allmer, and W. Henkel. “Categorization of species based on their microRNAs employing sequence motifs, information-theoretic sequence feature extraction, and *k*-mers”. In: *EURASIP Journal on Advances in Signal Processing* 2017.1 (2017), p. 70

2 Basic Concepts

Since this thesis is addressing an interdisciplinary topic, in this chapter, we present a brief introduction of selected concepts from both disciplines, biology and information theory. The primary aim is establishing a common language and understanding of the upcoming chapters. Section 2.1 provides the fundamentals in information theory and statistical methods. In Section 2.2, we present terminology and definitions of basic biological processes for a non-biologist reader.

2.1 Basic Concepts in Information Theory

Information theory was founded in 1948 by Shannon [Sha48]. He presented a mathematical framework for information transmission and established a foundation for modern communication systems. Shannon's tremendous achievement is the convenient definition of information based solely on the statistical properties of the transmitted message, irrespective of the meaning. The semantics of the message is "*irrelevant to the engineering problem*" [Sha48]. Shannon presented the essential components of a general communication system (Fig. 2.1).

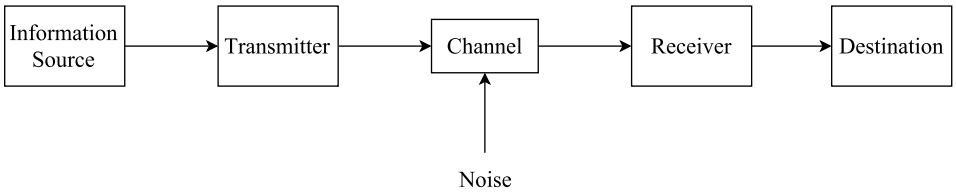

Fig. 2.1: Shannon's block diagram of a general communication system (reproduced from [Sha48])

It consists of

- an information source which produces the to be transmitted message,
- a transmitter which manipulates the message and produces a signal suitable for transmission over the channel,
- a channel which is the medium for transmitting or storing the signal,

- a receiver which reconstructs the transmitted message,

- a destination for whom the information is intended for.

2.1.1 Probability Theory

We briefly present basic concepts in probability theory. For detailed explanations we refer the reader to standard text books, such as [PP02; SW02], where the content of this section is also taken from.

Consider a random variable, denoted X, which takes values from a finite set $\mathcal{X}$. We present the following definitions and descriptions for a discrete random variable X. The continuous case can simply be obtained by replacing summations by integrals. Let $x \in \mathcal{X}$ be a realization of X. In a random experiment, the probability of x is denoted by $P(X = x)$. We abbreviate $P(X = x)$ by $P_X(x)$. $P(X = x)$ is known as the probability mass function (PMF) and it satisfies $\sum_{x\in\mathcal{X}} P_X(x) = 1$ and $P_X(x) \geq 0, \forall x \in \mathcal{X}$. The cumulative density function (CDF) is defined as

$$F_X(x) = P(X \leq x) = \sum_{x_i\in\mathcal{X}, x_i\leq x} P_X(x_i) \, . \tag{2.1}$$

For any given function $g(x)$, the expectation over the random variable X is given by

$$E\{g(x)\} = \sum_{x\in\mathcal{X}} g(x) \cdot P_X(x) \, . \tag{2.2}$$

The expected value of X, i.e., $\mu_X = E\{X\}$ is referred to as the mean. The variance is expressed as

$$\sigma^2 = E\{(X - \mu_X)^2\} \, , \tag{2.3}$$

σ is the standard deviation.

In combined experiments where two random variables X and Y are involved, the joint PMF of the pair (X, Y) over alphabets $\mathcal{X}$ and $\mathcal{Y}$ is denoted as $P_{XY}(x, y) = P(X = x, Y = y)$. $P_X(x) = \sum_{Y\in\mathcal{Y}} P_{XY}(x, y)$ and $P_Y(y) = \sum_{X\in\mathcal{X}} P_{X,Y}(x, y)$ are called the marginal PMFs. The random variables are said to be statistically independent, if

$$P_{XY}(x, y) = P_X(x) \cdot P_Y(y), \; \forall x, y \in \mathcal{X}, \mathcal{Y} \, . \tag{2.4}$$

The conditional PMF of Y given X is defined as

$$P_{Y|X}(y|x) = \frac{P_{XY}(x, y)}{P_X(x)} = \frac{P_{X|Y}(x|y)P_Y(y)}{P_X(x)} \, , \tag{2.5}$$

where $P_X(x) > 0$. The latter step follows from Bayes' theorem:

$$P_{XY}(x,y) = P_{X|Y}(x|y)P_Y(y) = P_{Y|X}(y|x)P_X(x) \ . \tag{2.6}$$

In the following sections, information theoretic concepts and quantities are introduced. For detailed descriptions we refer the reader to [CT91].

2.1.2 Information and Entropy

Hartley [Har28], in 1928, was the first to propose a quantitative measure of information. He defined information as $\log_b r$, where r is the number of possible outcomes. Shannon [Sha48] realized that Hartley's definition does not take into account the probabilities of the individual possible events. Hence, he defined information by associating it with the probabilities of events. Shannon's self-information of an event x is defined as

$$I(x) = \log_b \frac{1}{P_X(x)} \ . \tag{2.7}$$

The base b decides the unit of information and hence, does not have any influence on the information measure. If the logarithm is to the base 2, the unit is called *bit*. Unless stated otherwise, we will use $log = log_2$ throughout the thesis. The self-information measures the degree of surprise in observing a particular realization. Highly probable events provide less information whereas rare events result in a high degree of surprise and hence provide higher information.

The average information, called **entropy**, of a random variable X is defined as

$$\begin{aligned} H(X) &= \sum_{x \in \mathcal{X}} P_X(x) I(x) \\ &= \sum_{x \in \mathcal{X}} P_X(x) \log \frac{1}{P_X(x)} \ . \end{aligned} \tag{2.8}$$

Entropy can be intuitively expressed as the average uncertainty about the outcome of a random experiment. The lower the entropy the more certain we are about a random variable. Entropy is a non-negative quantity and can be shown that it is upper bounded by $\log |\mathcal{X}|$, where $|\mathcal{X}|$ is the cardinality (number of possible outcomes) of the set $\mathcal{X}$. The maximum entropy is achieved when X is uniformly distributed. For a deterministic process, i.e., $P_X(x_i) = 1 \wedge \forall j \neq i \ P_X(x_j) = 0$, $H(X) = 0$.

The **joint entropy** between two random variables $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ simply follows from Eq. (2.8).

$$H(X,Y) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P_{XY}(x,y) \log \frac{1}{P_{XY}(x,y)} \,. \tag{2.9}$$

Furthermore, the **conditional entropy** $H(Y|X)$ is the remaining uncertainty in Y given the random variable X and is defined as

$$\begin{aligned} H(Y|X) &= \sum_{x\in\mathcal{X}} P_X(x) H(Y|X=x) \,, & (2.10)\\ &= \sum_{x\in\mathcal{X}} P_X(x) \sum_{y\in\mathcal{Y}} P_{Y|X}(y|x) \log \frac{1}{P_{Y|X}(y|x)} \,, & (2.11)\\ &= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P_{XY}(x,y) \log \frac{1}{P_{Y|X}(y|x)} = E_{X,Y}\{\frac{1}{P_{Y|X}(y|x)}\} \,. & (2.12) \end{aligned}$$

The conditional entropy $H(Y|X)$ is upper bounded by $H(Y)$. It implies that conditioning can only reduce entropy. The maximum is achieved when X and Y are statistically independent. The so-called chain rule resulting from Bayes' theorem relates the joint and conditional entropies as follows:

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \tag{2.13}$$

2.1.3 Mutual Information and Kullback-Leibler Divergence

The **mutual information** is a very important information theoretic quantity which measures the information a random variable X contains about another random variable Y, and vise versa. The mutual information is mathematically defined as

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \,, & (2.14)\\ &= H(Y) - H(Y|X) \,, & (2.15)\\ &= \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} \,. & (2.16) \end{aligned}$$

From Eq. (2.14) one can understand mutual information as a reduction in uncertainty about X after Y is observed. Mutual information is symmetric, i.e., $I(X;Y) = I(Y;X)$, and is within the range $0 \leq I(X;Y) \leq \min(H(X), H(Y))$. The mutual information is zero when the two random variables are statistically independent.

Mutual information is a special case of the **Kullback-Leibler divergence** (D_{KL}), also known as relative entropy [KL51]. The Kullback-Leibler divergence is a measure

of the "distance" between any two distributions $P_X(x)$ and $Q_X(x)$. It is defined as follows:

$$D_{KL}(P_X(x)||Q_X(x)) = \sum_{x\in\mathcal{X}} P_X(x)\log\frac{P_X(x)}{Q_X(x)}\ . \tag{2.17}$$

However, it should be noted that D_{KL} is not a true distance (metric). Although $D_{KL}(P_X(x)||Q_X(x))$ is always non-negative and is zero if and only if $P_X(x) = Q_X(x)$, it is not symmetric and does not satisfy the triangle inequality.

As can be seen from Eq. (2.16), mutual information is a D_{KL} between the joint PMF, $P_{XY}(x,y)$, and the product of the marginal PMFs, $P_X(x)P_Y(y)$.

$$I(X;Y) = D_{KL}(P_{XY}(x,y)||P_X(x)P_Y(y))\ . \tag{2.18}$$

The conditional mutual information, i.e., the mutual information between two random variables X and Y conditioned on a third random variable Z, having a PMF $P_Z(z)$ is given by

$$\begin{aligned} I(X;Y|Z) &= \sum_{z\in\mathcal{Z}} P_Z(z) \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P_{XY|Z}(x,y|z)\log\frac{P_{XY|Z}(x,y|z)}{P_{X|Z}(x|z)P_{Y|Z}(y|z)} \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}} P_{XYZ}(x,y,z)\log\frac{P_Z(z)P_{XYZ}(x,y,z)}{P_{XZ}(x,z)P_{YZ}(y,z)} \end{aligned} \tag{2.19}$$

where $P_{XYZ}(x,y,z)$, $P_{XZ}(x,z)$, and $P_{YZ}(y,z)$ are the joint PMFs of the random variables shown in subscripts.

Now that we defined all the necessary information theoretic quantities, we can have a closer look at a communication system depicted in Fig. 2.1 and Shannon's channel coding theorem.

2.1.4 A Communication Channel

Information is transmitted between the transmitter and receiver over a noisy communication channel. If the channel is noiseless, the receiver will receive the exact copy of what was transmitted. The input to the channel is a message $\mathbf{x} = (x_1, x_2, \ldots, x_N)$, where $x_i \in \mathcal{X}$, generated by a source X described by $p_X(x)$. The output of the channel is a noisy version of the original message $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, where $y_i \in \mathcal{Y}$, modeled by a random variable Y. The channel is specified by a conditional probability density function $p_{Y|X}(\mathbf{y}|\mathbf{x})$ if it is continuous and a conditional PMF $P_{Y|X}(\mathbf{y}|\mathbf{x})$ if it is discrete.

A discrete memoryless channel (DMC) is defined by discrete input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$ and a PMF $P_{Y|X}(y|x)$, specifying the probability to observe y

given x was sent, denoted ($\mathcal{X}$, $P_{Y|X}(y|x)$, $\mathcal{Y}$). The channel is assumed memoryless. Hence, a vector output of the channel depends only on the current vector input and the channel transition probabilities can be factored as

$$P_{Y|X}(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{N} P_{Y|X}(y_i|x_i) \ . \tag{2.20}$$

The channel conditional probabilities are often conveniently described by a transition probability matrix $\mathbf{P}$. The channel is called symmetric, if the columns and rows of the transition matrix are permutations of each other.

Channel Capacity

The channel capacity is defined as the maximum rate at which a reliable (error-free) information transmission through the channel is possible. Capacity is determined by maximizing the mutual information between input (X) and output (Y) over all possible input probability distribution $P_X(x)$.

$$C = \sup_{P_X(x)} I(X;Y) \ . \tag{2.21}$$

Shannon's channel coding theorem states that if the information rate R is equal to or less than the channel capacity C, i.e., $R \leq C$, then there is a a coding technique which enables a communication through a noisy channel with arbitrarily small probability of error. However, the theorem does not provide a code construction.

Channel models

Among the various channel models available, we present here only the ones which are more useful in biological settings. The simplest but yet very useful channel model is a binary symmetric channel (BSC). The BSC is a channel with a binary input and output alphabet $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ and the transition probability matrix $\mathbf{P}_{BSC}$ with entries

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

The transition diagram of the BSC is shown in Fig. 2.2.

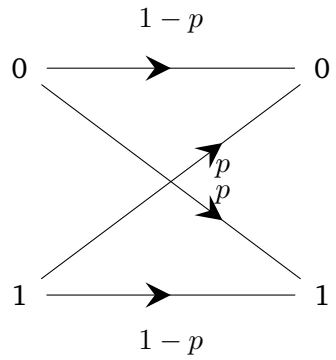

Fig. 2.2: Binary symmetric channel

The BSC can be extended to a q-ary symmetric discrete memoryless channel with alphabet $\mathcal{X} = \mathcal{Y} = \{1, 2, 3, \ldots, q\}$ of cardinality q. The transition matrix $\mathbf{P}_{QSC}$ will have entries

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

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

For modeling DNA sequence evolution, a quaternary symmetric channel is often employed because of the DNA alphabet, i.e., $\mathcal{X} = \mathcal{Y} = \{A, T, C, G\}$. The transition matrix is then given by

$$\mathbf{P}_{QSC} = \begin{bmatrix} P(A|A) & P(T|A) & P(C|A) & P(G|A) \\ P(A|T) & P(T|T) & P(C|T) & P(G|T) \\ P(A|C) & P(T|C) & P(C|C) & P(G|C) \\ P(A|G) & P(T|G) & P(C|G) & P(G|G) \end{bmatrix}.$$

2.1.5 Statistical Inference

Statistical inference deals with a statistical decision making about the parameters of the models based on observed data. Given a probabilistic model with one or more unknown parameters θ_i and observations $\mathbf{y} = [y_1, y_2, \ldots, y_n]$, statistical inference can be performed in two ways. The first approach is to estimate the values of the parameters from the observation $\mathbf{y}$ (parameter estimation). The second approach is to guess a value for θ_i and check the data if the value is correct (hypotheses testing). In this section, we will introduce basic concepts in parameter estimation and hypothesis testing. For further reading we recommend [PP02; CB02].

Parameter Estimation

Let $\mathbf{Y} = [Y_1, Y_2, \ldots, Y_n]$ be n random variables representing observations $\mathbf{y} = [y_1, y_2, \ldots, y_n]$. The joint PDF or PMF depends on the unknown parameter $\boldsymbol{\theta}$. The joint probability of a set of observations, conditioned on a choice for $\boldsymbol{\theta}$ is called the likelihood function

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{y}) \equiv P_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{y}|\boldsymbol{\theta}) \; . \tag{2.22}$$

The **maximum likelihood** (ML) estimation chooses the set of parameter values that most likely caused the observed data to occur. The ML estimate is formally defined as

$$\hat{\boldsymbol{\theta}}_{ML} = \arg\max_{\boldsymbol{\theta}} P_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{y}|\boldsymbol{\theta}) \; . \tag{2.23}$$

It is usually more convenient to maximize the log-likelihood function $\mathcal{LL}(\boldsymbol{\theta}; \mathbf{y}) = \ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{y})$. Since, $\ln$ is a monotonic function, the value of θ that maximizes $\ln \mathcal{L}(\boldsymbol{\theta}; \mathbf{y})$ will also maximize $\mathcal{L}(\boldsymbol{\theta}; \mathbf{y})$. Therefore, ML estimate can be obtained by solving

$$\hat{\boldsymbol{\theta}}_{ML} = \arg\max_{\boldsymbol{\theta}} \ln P_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{y}|\boldsymbol{\theta}) \; . \tag{2.24}$$

Example 2.1.1. Suppose $\mathbf{y} = [y_1, y_2, \ldots, y_N]$ is a DNA sequence with alphabet $\mathcal{S} = \{A, T, C, G\}$ in which base A occurs n_A times, T n_T times, C n_C times, and G n_G times. Assuming that Y_t are **independent** random variables, for $t = 1, 2, \ldots, n$. The independence assumption means that the probability of a base at a given location is the same regardless of what base precedes it. We like to estimate the base probabilities $\mathbf{p} = [p_A, p_T, p_C, p_G]$ from the observed sequence $\mathbf{y}$. The log-likelihood function is

$$\begin{aligned} \mathcal{LL}(\mathbf{p}; \mathbf{y}) &= \ln p_A^{n_A} p_T^{n_T} p_C^{n_C} p_G^{n_G} \\ &= \ln \prod_{i \in \mathcal{S}} p_i^{n_i} \; . \\ &= \sum_{i \in \mathcal{S}} n_i \ln p_i \end{aligned} \tag{2.25}$$

The ML estimate can be found by maximizing the log-likelihood as in Eq. (2.24), with the constraint that $\sum_{i \in \mathcal{S}} p_i = 1$. The resulting ML estimate is

$$\hat{p_i} = \frac{n_i}{N} \; , \tag{2.26}$$

which is the relative frequency (the fraction of times the corresponding symbol occurs) of the bases in the sequence.

In ML estimation the parameters are assumed to be deterministic. When unknown parameters $\boldsymbol{\theta}$ are realizations of discrete random variables $\boldsymbol{\Theta}$ distributed according

to $P_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$ (prior distribution), an optimal Bayesian estimator function is the so-called **maximum a-posteriori probability** (MAP). The MAP estimate is obtained as

$$\hat{\boldsymbol{\theta}}_{MAP} = \arg\max_{\boldsymbol{\theta}} P_{\boldsymbol{\Theta}|\mathbf{Y}}(\boldsymbol{\theta}|\mathbf{y}) = \arg\max_{\boldsymbol{\theta}} \frac{P_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{y}|\boldsymbol{\theta})P_{\boldsymbol{\Theta}}(\boldsymbol{\theta})}{P_{\mathbf{Y}}(\mathbf{y})} = \arg\max_{\boldsymbol{\theta}} P_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{y}|\boldsymbol{\theta})P_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \ , \tag{2.27}$$

where $P_{\boldsymbol{\Theta}|\mathbf{Y}}(\boldsymbol{\theta}|\mathbf{y})$ is called the posterior probability. In the last step, $P_{\mathbf{Y}}(\mathbf{y})$ is dropped because it doesn't depend on the maximization parameter $\boldsymbol{\theta}$. It can be seen that when the prior distribution $P_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$ is uniform, the ML and MAP estimators are identical.

Hypothesis Testing

We consider here a binary hypothesis testing where two mutually exclusive and opposing hypotheses are examined. For example, we might wish to test the assumption that the parameter $\theta = \theta_0$ against the assumption $\theta \neq \theta_0$. The two hypotheses are termed as the null hypothesis H_0 and the alternative hypothesis H_1. The steps involved in performing inference using hypothesis testing are presented below.

1. Identify the null and alternative hypotheses (i.e., state H_0 and H_1).

2. Specify the significance level α. Significance level is the probability of rejecting the null hypothesis when it is true. Typically, 0.01, 0.05, or 0.1 is selected.

3. Compute the so called test statistic. Test statistic is a function of the sample data that is used to make a decision about the rejection of the null hypothesis.

4. Determine the distribution of the test statistic under H_0.

5. Using the distribution of the test statistic, compute the so-called p-value. The p-value is the conditional probability of the tails and it determines the probability of getting a value that is at least as extreme as the one found from the sample data. Let T be the random variable representing the test statistics and t be the realization. The p-value is calculated as

   - Right-tailed test p-value $= P(T \geq t|H_0)$
   - Left-tailed test p-value $= P(T \leq t|H_0)$
   - Two-tailed test p-value $= P(T \geq |t| \, |H_0)$

6. Decide on rejecting or accepting H_0 based on the comparison of the p-value to the predefined significance level α.

$$\text{if } p\text{-value} \leq \alpha, \text{ reject } H_0.$$

Since we are making decisions based on the sample data only, two types of errors can occur. The first is called a type I error whereby H_0 is rejected given that H_0 is true. The other is called a type II error where we fail to reject H_1 given H_0 is false.

2.1.6 Markov Chains and Processes

Markov chains are often used to model statistical dependencies in biological sequences. The brief introduction to Markov chains and processes is presented in this section. Further explanations can be found in [EG06; CT91; Ser09; Dur+98; Ros14], where the content of this section is taken from.

Discrete-Time Markov Chains

A discrete-time stochastic process is a sequence of indexed random variables, denoted by $\{X_n : n \geq 0\}$. The collection of all possible values that X_n can assume is called the state space, denoted by $\mathcal{S}$. Each element $i \in \mathcal{S}$ is called a state. A stochastic process is described by the state space and the joint probability mass functions of the random variables. If there is some sort of dependence between the random variables, it is called a Markov process. The simplest dependency is a 1$^{\text{st}}$ order Markov process. In a 1$^{\text{st}}$ order Markov process, the future value of a random variable depends only on the current state and not on the past values. Formally, a discrete 1$^{\text{st}}$ order Markov process is a stochastic process $\{X_n : n \geq 0\}$ which satisfies the following property (also called Markov property):

$$P(X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) = P(X_{n+1} = x_{n+1} | X_n = x_n) \,. \tag{2.28}$$

In 1$^{\text{st}}$ order Markov process, the joint PMF of the random variables can be described as

$$\begin{aligned} p(x_1, x_2, \ldots, x_n) &= p(x_0)p(x_1|x_0)p(x_2|x_1)\ldots p(x_n|x_n - 1) \\ &= p(x_0)\prod_{i=1}^{n} p(x_i|x_{i-1}) \end{aligned} \tag{2.29}$$

In general, Eq. (2.28) can be generalized for any m as m^{th}-order Markov process in which the dependency is on the last m most recent past values. A Markov process is m^{th}-order, if

$$P(X_{n+1} = x_{n+1}|X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) = \\ P(X_{n+1} = x_{n+1}|X_n = x_n, \ldots, X_{n-m+1} = x_{n-m+1}) \; . \tag{2.30}$$

A Markov process is said to be a **Markov chain**, if the state space is discrete, i.e., finite or countable. If the conditional probability $p(x_{n+1}|x_n)$ does not depend on n and is constant, the Markov chain is termed as *time-invariant (time-homogeneous)*. A time-invariant Markov chain is characterized by its initial state and a probability transition matrix $\mathbf{\Pi} = [\Pi_{ij}],\ i, j \in \mathcal{S}$, where $\Pi_{ij} = P(X_{n+1} = j|X_n = i)$. The transition probabilities Π_{ij} denotes the probability that the chain being in state i moves after one-step into state j, and is referred to as a one-step transition probability. When leaving state i the chain must move to one of the states $j \in \mathcal{S}$. Hence, each row sums to one (every row of $\mathbf{\Pi}$ is a distribution), i.e.,

$$\sum_{j \in \mathcal{S}} \Pi_{ij} = 1 \; . \tag{2.31}$$

Let $\mathbf{p_s}(0) = [p_s^1(0), p_s^2(0), \ldots, p_s^{|\mathcal{S}|}(0)]$ be the initial distribution of the Markov chain over $\mathcal{S}$, where

$$p_s^i(0) = P(X_0 = i) \qquad i \in \mathcal{S} \; . \tag{2.32}$$

The state distribution at the k^{th} time step could be determined as

$$\mathbf{p_s}(k) = \mathbf{p_s}(k-1)\mathbf{\Pi} = \mathbf{p_s}(0)\mathbf{\Pi}^k \; . \tag{2.33}$$

If the Markov chain is regular, there exists a stationary (steady-state) distribution to which the system converges. A Markov chain is called regular if successive powers of the transition matrix $\mathbf{\Pi}$ contain only positive entries. A distribution $\mathbf{p_s}$ is said to be the stationary distribution of a Markov chain if

$$\mathbf{p_s} = \mathbf{p_s}\mathbf{\Pi} \tag{2.34}$$

Note that, $\mathbf{p_s}$ is the left eigenvector of $\mathbf{\Pi}$ corresponding to eigenvalue 1. To calculate the stationary distribution, we use the additional constraint $\sum_i p_s^i = 1$. In matrix form,

$$\mathbf{p_s}\mathbf{U} = \mathbf{1} \; , \tag{2.35}$$

where $\mathbf{U} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix}$ and $\mathbf{1} = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}$.

Adding Eq. (2.34) and Eq. (2.35),

$$\begin{aligned} \mathbf{p_s}(\mathbf{U}+\mathbf{\Pi}) &= \mathbf{p_s}+\mathbf{1} = \mathbf{p_s}\mathbf{I}+\mathbf{1}\ , \\ \mathbf{p_s}(\mathbf{U}+\mathbf{\Pi}-\mathbf{I}) &= \mathbf{1}\ , \\ \mathbf{p_s} &= \mathbf{1}(\mathbf{U}+\mathbf{\Pi}-\mathbf{I})^{-1}\ . \end{aligned} \tag{2.36}$$

$\mathbf{I}$ is the identity matrix.

Continuous-Time Markov Chains

A stochastic process $\{X(t) : t \geq 0\}$ with state space $\mathcal{S}$ and defined over a continuous-time $t \in [0, \infty)$ is said to be a continuous time Markov chain (CTMC), if it satisfies the Markovian property, i.e., the conditional pmf of the future $X(t+s)$ given the present $X(s)$ and the past $X(u)$, $u < s$, depends only on the present. For $i, j, x(u) \in \mathcal{S}$,

$$P(X(t+s) = j | X(s) = i, X(u) = x(u)) = P(X(t+s) = j | X(s) = i)\ . \tag{2.37}$$

If the conditional probability is independent of s, the CTMC is said to have a stationary or homogeneous transition probability. In a time-homogeneous CTMC, the conditional probabilities depend only on the time difference,

$$P(X(t+s) = j | X(s) = i) = P(X(t) = j | X(0) = i) = P_{ij}(t)\ . \tag{2.38}$$

The probabilities $P_{ij}(t)$ are called the transition probabilities and the $|\mathcal{S}| \times |\mathcal{S}|$ matrix

$$\mathbf{P}(t) = \begin{bmatrix} P_{00}(t) & P_{01}(t) & \cdots \\ P_{10}(t) & P_{11}(t) & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

is called the transition probability matrix. It is a stochastic matrix where each row sums to 1, $\sum_j P_{i,j}(t) = 1$ for all i.

Given two transition probability matrices $\mathbf{P}(t)$ and $\mathbf{P}(s)$, the Chapman-Kolmogorov theorem describes the state of the Markov process at time $t+s$. The process moves from state i to any state k after time t with probability $P_{ik}(t)$. Then, it moves to state j from state k with probability $P_{kj}(s)$. Thus, the probability

$$P_{ij}(t+s) = \sum_k P_{ik}(t) P_{kj}(s)\ , \tag{2.39}$$

in matrix form

$$\mathbf{P}(t+s) = \mathbf{P}(t)\mathbf{P}(s) \ . \tag{2.40}$$

It follows that a transition matrix $\mathbf{P}(t)$ at any integer time t can be calculated from a transition matrix $\mathbf{P}(1)$ of a unit time step as

$$\mathbf{P}(t) = \mathbf{P}(1)^t \ . \tag{2.41}$$

In time-homogeneous CTMC, $P_{ij}(t)$ is the probability of jumping from i to j during an interval time of duration t. Hence, we cannot speak about one-step transition matrices any more. Therefore, we describe the Markov process with the instantaneous transition rates as

$$\begin{aligned} \frac{d\mathbf{P}(t)}{dt} &= \lim_{\delta t \to 0} \frac{\mathbf{P}(t+\delta t) - \mathbf{P}(t)}{\delta t} \ , \\ &= \lim_{\delta t \to 0} \frac{\mathbf{P}(t)\mathbf{P}(\delta t) - \mathbf{P}(t)}{\delta t} \ , \\ &= \left[\lim_{\delta t \to 0} \frac{\mathbf{P}(\delta t) - \mathbf{I}}{\delta t} \right] \mathbf{P}(t) \ . \end{aligned} \tag{2.42}$$

The matrix of limits can be defined as a rate matrix, $\mathbf{Q}$, which describes the CTMC. Thus,

$$\frac{d\mathbf{P}(t)}{dt} = \mathbf{Q}\mathbf{P}(t) \ . \tag{2.43}$$

The rate matrix $\mathbf{Q} = \{Q_{ij}\}_{i,j \in \mathcal{S}}$ should satisfy the following properties:

- $0 \leq Q_{ij} < \infty \ \forall i \neq j$,
- $\sum_j Q_{ij} = 0 \ \forall i$.

The diagonal entries are calculated as

$$Q_{ii} = -\sum_{j \neq i} Q_{ij} \ . \tag{2.44}$$

The diagonal entries Q_{ii} are always negative and correspond to the rate with which the Markov chain leaves the state i.

Solving the differential equation in Eq. (2.43) together with the initial condition $\mathbf{P}(0) = \mathbf{I}$, $\mathbf{I}$ is the identity matrix, we get

$$\mathbf{P}(t) = e^{\mathbf{Q}t} \tag{2.45}$$

If we denote the distribution of the states at time 0 by $\boldsymbol{\pi}^{(0)}$, the distribution after time t is computed as

$$\boldsymbol{\pi}^{(t)} = \boldsymbol{\pi}^{(0)}\mathbf{P}(t) \ . \tag{2.46}$$

The stationary distribution $\boldsymbol{\pi}$, where the initial and target distribution are equal, is calculated as

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}(t) \;, \tag{2.47}$$

or equivalently using the rate matrix

$$\boldsymbol{\pi}\mathbf{Q} = 0 \;. \tag{2.48}$$

For modeling molecular evolution, the CTMC is assumed to be *time-reversible* [Yan06]. Reversibility implies that the expected amount of changes from state i to state j in steady state is equal to the amount of change from state j to i.

$$\pi_i q_{i,j} = \pi_j q_{j,i} \qquad \forall i, j \in \mathcal{S} \;, \tag{2.49}$$

with $\boldsymbol{\Pi}^* = \text{diag}\{\boldsymbol{\pi}\}$ it can be written in matrix form as

$$\boldsymbol{\Pi}^*\mathbf{Q} = \boldsymbol{\Pi}^*(\mathbf{S}\mathbf{Q}) = (\mathbf{S}\boldsymbol{\Pi}^*)^\intercal\mathbf{Q} = \mathbf{Q}^\intercal\boldsymbol{\Pi}^* \;, \tag{2.50}$$

where the rate matrix, $\mathbf{Q} = \mathbf{S}\boldsymbol{\Pi}^*$, is decomposed into a product of a symmetric matrix $\mathbf{S}$ and $\boldsymbol{\Pi}^*$.

2.2 Basic Concepts in Molecular Biology

This section provides an overview of the basic processes in molecular biology and definition of terms used in this thesis. The aim here is not to extensively introduce the biological concepts but rather to familiarize non-biologists with the terminology and basic definitions, so that the subsequent chapters can be understood. The material presented here is available in standard text books of molecular biology and genetics [All07; HA09; Har05].

Organisms can be classified in to three domains called the Archaea, the Bacteria, and the Eukarya, based on their phylogenetic relationships [Woe+90]. The domains, being the highest taxonomic rank of organisms, show shared characteristics and evolutionary differences. Archaea and bacteria are grouped together as prokaryotes. The central difference between eukaryotes and prokaryotes is that eukaryotic cells contain membrane-bound organelles, including the nucleus and prokaryotic cells do not. Eukaryotes include all higher multicellular and complex organisms, e.g., plants, animals, fungi, or humans.

2.2.1 Nucleic Acids

There are two types of nucleic acids, deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). DNA is a double stranded structure found in all cells, containing the genetic information of the living organism. It consists of building blocks called nucleotides. The nucleotides are made of a sugar phosphate moiety and one of the four nitrogenous bases attached to the sugars. These bases are called Adenine, Thymine, Cytosine, and Guanine (A, T, C, G). The hydrogen-bonded bases of opposite strands are stacked into chains connected by alternating phosphate and sugar groups of the nucleotides. Each strand has so-called $3'$ (three prime) and $5'$ (five prime) ends. The two strands are anti-parallel with the so-called leading strand oriented in $3'$ to $5'$ direction, whereas the lagging strand runs from the $5'$ end to the $3'$ end [HH09]. A figure showing the structure of the DNA is presented in Fig. 2.3.

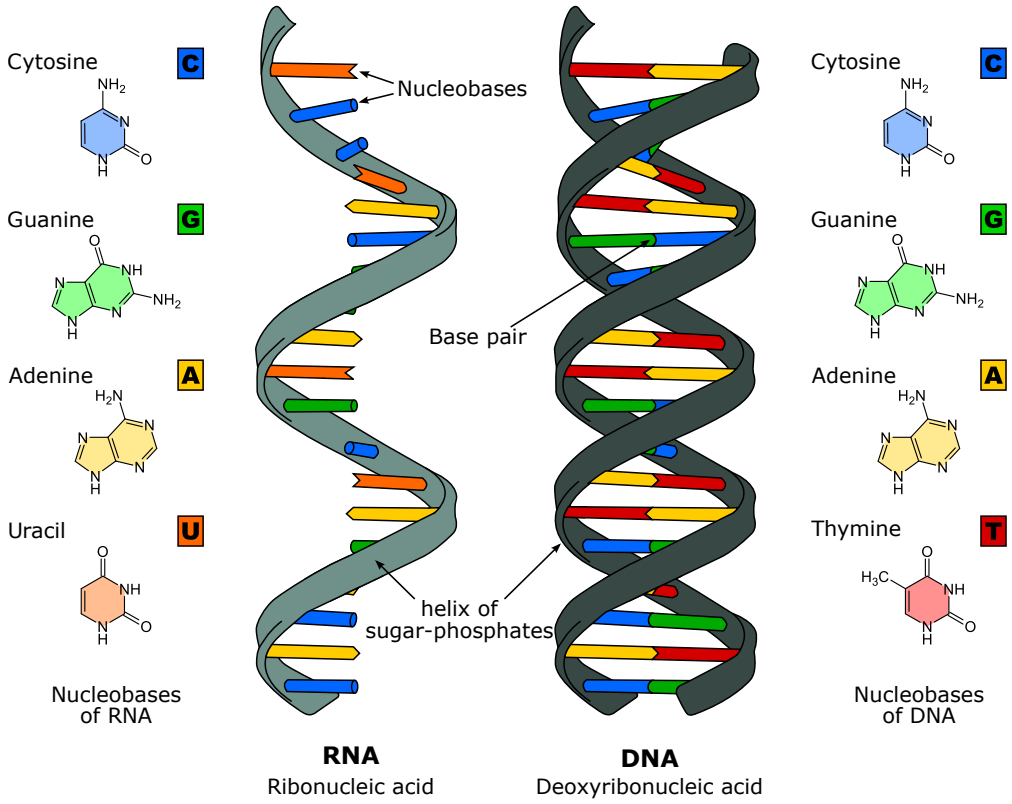

Fig. 2.3: The structures of DNA and RNA [Com10].

The two strands are complementary to each other. According to the Watson-Crick pairing rule, A is always paired with T and G is always paired with C [WC+53]. This means, if we know the sequence of nucleotides on one strand, the sequence in the complementary strand is known right away. The bases are linked by hydrogen bonds. G-C pairs have three hydrogen bonds whereas A-T pairs have two hydrogen bonds. The additional hydrogen bond makes the G-C pairs slightly more stable

than A-T pairs. Therefore, the GC-rich regions of the DNA are more stable. The bases are further classified into two groups - purines and pyrimidines. As can be seen in Fig. 2.3, purines have two carbon rings and four nitrogen atoms, A and G are purines, while T and C are pyrimidines that contain a single carbon ring and two nitrogen atoms. Purines are denoted as $R = \{A, G\}$ whereas pyrimidines are denoted by $Y = \{C, T\}$.

RNA is a similar molecule, except that the RNA is single-stranded (some viruses, e.g., retrovirus, have double stranded RNA) and the base Uracil (U) is present instead of Thymine (T). The major functions of the RNA molecule are transfer of information to different parts of the cell and providing a template to protein synthesis.

The complete set of genetic information that the organism carries in its DNA is called the genome. A genome might contain one or more chromosomes. Most bacteria, including the ones studied here, contain a single chromosome arranged in a circular fashion. Generally, the genome size of prokaryotes is smaller than that of eukaryotes. To put this into perspective, the genome size of the most studied bacteria *Escherichia coli (E. coli)* is around 4.6×10^6 base pairs (bp) whereas a human genome contains about 3.2×10^9 bp.

2.2.2 The Flow of Biological Information

Francis Crick [Cri58] states that the flow of biologic information is from the DNA towards proteins and called the process the central dogma of molecular biology (Fig. 2.4). The sequences of bases aligned in a segment of a DNA, called a gene, carry the directions for building proteins. Proteins carry out nearly every aspect of cellular function, including transporting other molecules, offer structural support, and directing chemical reactions. The process of synthesizing proteins or in some cases a functional RNA from the information encoded in the DNA is referred to as gene expression [HH09]. Gene expression consists of two steps, transcription and translation.

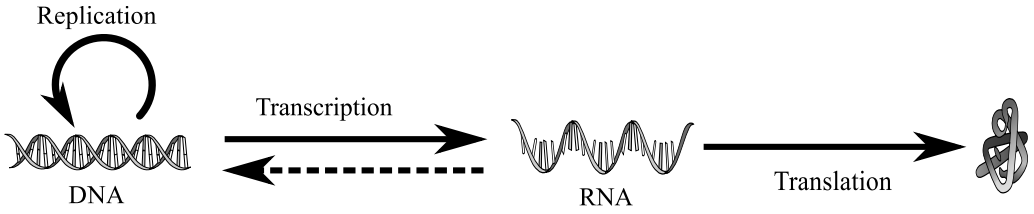

Fig. 2.4: The central dogma of molecular biology: The dashed arrow shows reverse transcription, a special cases [Com12]

Transcription

The RNA polymerase enzyme unwinds the DNA molecule and the transcription process begins. In transcription, the gene sequence is copied into messenger RNA (mRNA) using the template strand of the DNA. The enzyme involved in this reaction is known as RNA polymerase. RNA polymerase reads the DNA sequence of the template strand in the $3'$ to $5'$ direction, thus the new RNA strand is synthesized from the $5'$ to the $3'$ end. The synthesized mRNA has obviously the same sequence as the coding strand except that the base uracil (U) is used instead of thymine (T). An example of a transcription process is shown in Fig. 2.5.

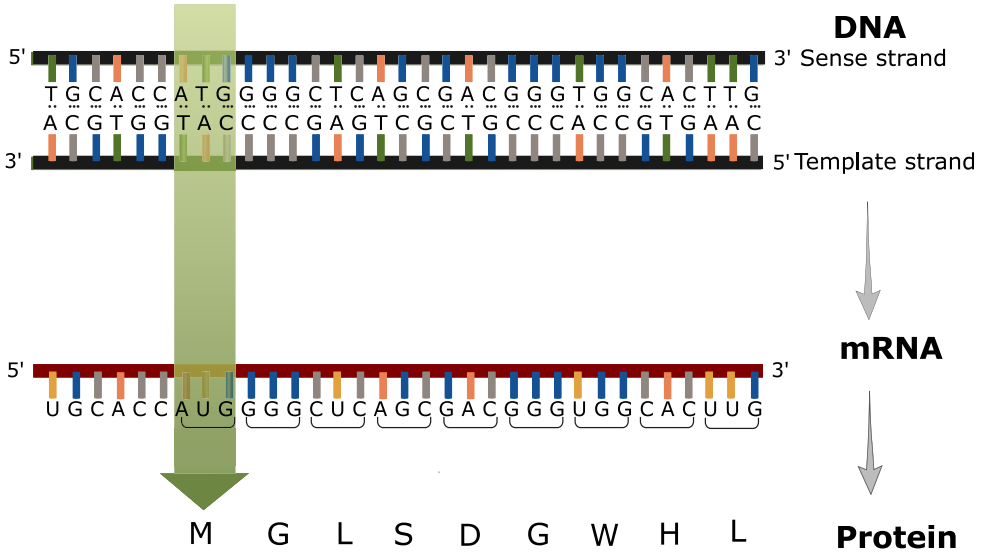

Fig. 2.5: Transcription and translation processes. The standard genetic code is used to synthesize proteins from mRNA. Modified from [All07] [1]

Translation

In the translation phase, the ribosome translates the sequence of mRNA molecule to amino acids, reading the sequence in groups of three bases (codons). There are 20 naturally occurring amino acids [2]. The chart in Fig. 2.6 is refereed to as the genetic code which shows the codon to amino acid mappings. The genetic code offers an encoding redundancy by allowing multiple codons (up to six codons) to encode the same amino acids. Codons encoding the same amino acid are called synonymous codons. The process starts when the smaller ribosomal subunit is attached to the translation initiation site, usually AUG. Then, the transfer RNA (tRNA) binds to the mRNA. The tRNA contains an anticodon complementary to the mRNA to which it binds and the corresponding amino acid is attached to it. Next, the large ribosomal

[2]There are two more amino acids, selenocysteine and pyrrolysine, which are sometimes synthetically incorporated into proteins. They are encoded by the stop codons UGA and UAG.

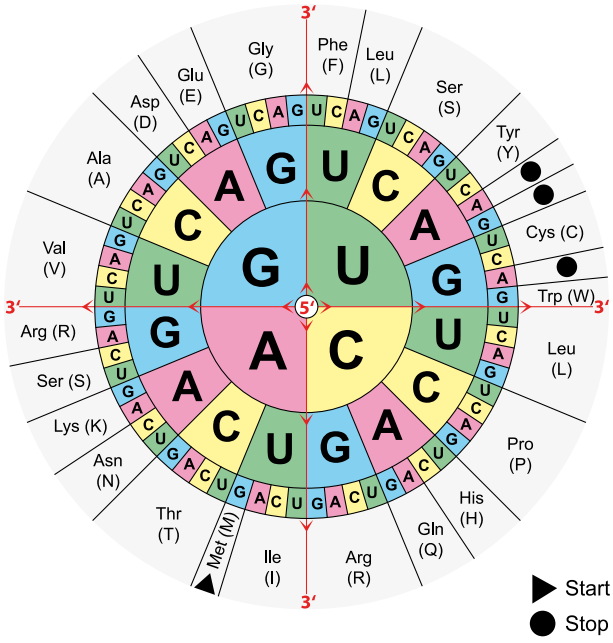

Fig. 2.6: Codon-amino acid encoding chart [Com09].

subunit binds to create the P-site (peptidyl) and A-site (aminoacyl). The first tRNA occupies the P-site and the second tRNA enters to the A-site. After that, the tRNA at the P-site transfers the amino acid it carries to the second tRNA at the A-site and exits. The ribosome then moves along the mRNA and the next tRNA enters. This process will continue until a stop codon (UAG, UAA, or UGA) signals the end of the mRNA molecule. Finally, the amino acids are connected by a peptide bond and folded in a certain way to create proteins of specific functions. The whole process is shown in Fig. 2.5.

Replication

The other process involving the transfer of information is DNA replication, which is the copying of the double stranded DNA. The process of replication is depicted in Fig. 2.7. Replications starts at a specific sequence of nucleotides called the origin of replication (oriC), when the enzymes called DNA helicases recognize and bind to the site. The DNA helicase unwinds the DNA by breaking the hydrogen bonds. After the two strands are separated, each strand will be used as a template for synthesizing the complimentary strand, producing two identical copies. The enzyme called DNA polymerase reads the strands in the $3'$ to $5'$ direction placing the corresponding nucleotides along the way. On the leading strand template, addition of complementary nucleotides is continuous. However, in the lagging strand, the DNA ploymerase has to move in opposite direction (for DNA polymerase, reading

only makes sense in the $3'$ to $5'$ direction) and hence, DNA synthesis occurs in short and separated fragments (Okazaki fragments)[OO80].

Since the chromosome in bacteria is usually circular, the replication process is bidirectional. As the DNA helicase unwinds the two strands, a moving replication fork is created in both directions, away from the oriC. The two replication forks meet at the opposite end of the chromosome called replication terminus (Ter).

The replication process has to be of a very high fidelity in order to preserve the genetic information over many generations [Pra08]. It is reported that the enzyme DNA polymerase makes about 10^{-6} errors per base pair per cell division. However, most of the mistakes are corrected by DNA repair mechanisms known as proofreading and mismatch repair, which brings the errors down to between 10^{-8} and 10^{-10} per base pair per cell division [Pra08][Alb+13, Chapter 6]. In the following section, the different types of genetic errors and mutations will be discussed.

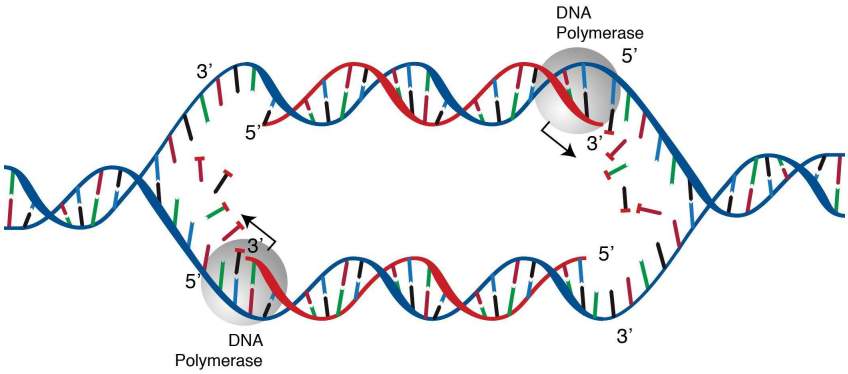

Fig. 2.7: DNA replication [HNHGRI].

2.2.3 Mutations

DNA is susceptible to changes and modifications, i.e., it is mutable. Changes in the nucleotide sequence are called mutations. Mutations can happen due to mistakes during DNA replication or are caused by physical and chemical agents in the environments, such as ultraviolet radiation. Most of the errors are detected and repaired by proofreading and mismatch repair mechanisms of the cell. The common repairing technique is called nucleotide excision repair. The DNA around and including the wrong base is removed and replaced with the correct bases using the intact complementary strand as a template. However, it should not be forgotten that mutations are the driving force behind evolution by allowing organisms to adapt to the environment.

Most commonly, mutations are substitutions (point mutations), in which a nucleotide is replaced by one of the other three nucleotides. Depending on the chemical class

of the original and the replaced nucleotides, substitution mutations are of two types. The first type is referred to as transitions in which a purine is changed to a purine or a pyrimidine to a pyrimidine. The other type is called transversion in which a purine is changed to a pyrimidine or vice versa. Transitions and transversion errors are depicted in Fig. 2.8. In general, transitions are more common than transversions. A base substitution in the coding region can be silent, meaning that the encoded amino acid is still the same. For example, if ACG is changed to ACA or ACC in the mRNA, the produced amino acid will not be changed (threonine in this case). Changes in the third position of a codon often causes a silent substitution, due to the degeneracy of the genetic code (see Fig. 2.6). A substitution can also result in an encoding of a different amino acid, and hence different protein. It is called a missense mutation. For instance, sickle-cell anemia is caused by a substitution of GAG in the mRNA, which specifies glutamate, by GUG resulting in a valine residue in the beta-globin protein. The last type of substitution mutation is called nonsense mutation. This is a mutation to a stop (termination) codon and it will result in the premature termination of the amino acid chain.

The other type of mutations result from an insertion or deletion of one or more bases. When the inserted or deleted chunk is not a multiple of 3, because of the resulting frame shift in the coding sequence, a lot of amino acid changes will be introduced. In addition, it often leads to a premature stop codon. These insertions and deletions (InDels) are also known as frameshift mutations.

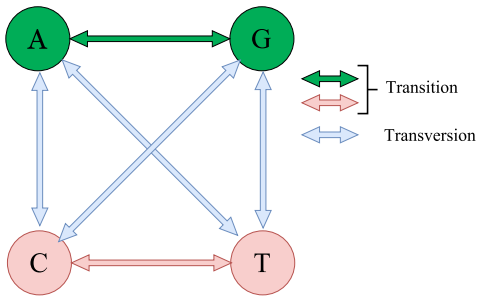

Fig. 2.8: Transition and transversion mutations.

3 A Channel Model from a Mutation Matrix

This chapter deals with the analysis of a mutation probability matrix called the empirical codon mutation matrix. In the first part, a hypothetical assumption of the matrix as a communication channel is made and computations of mutual information, capacity, and optimal codon distribution are performed. In the second part, we employ a dimension reduction and a clustering method to compare mutation and chemical distances, with the aim of checking whether highly probable mutations are between chemically similar codons or amino acids. We start, in Section 3.1, by introducing existing nucleotide, codon, and protein-based models of evolution. Then in Section 3.2, biological models of communication are described. In Section 3.3, after our codon-based-model is presented, an exponent for the mutation matrix to allow for an error-free transmission of the genetic message is computed. In Section 3.4, a description of the dimension reduction method used in this work is presented and the chemical and mutation matrices are compared. Finally, we summarize the findings and point to a future work in this direction in Section 3.5.

3.1 Models of Molecular Evolution

Ever since the time of Charles Darwin, understanding the evolutionary relationships between all organisms has been a central theme in biology [NK00]. Darwin in 1859 put forth a theory of evolution based on the idea of a universal common descent, stating that all life on earth descended from the last universal common ancestor. Furthermore, he pointed out that the primary driving force of evolution is natural selection. The classical way of reconstructing evolutionary relationships (phylogenetic tree reconstructions) between species relied on comparative study of anatomical and physiological features. However, the results of the classical methods are relatively subjective and are not satisfactory because of the complexity of morphological and physiological characters [NK00]. After the advancements in molecular biology, evolution is now studied by comparisons of the DNA of the organisms. The advantage of DNA-based methods over the classical approach is the possibility of comparing any group of organisms, including bacteria, plants, and animals, which otherwise was impossible to do. In addition, compared to

the morphological features, the genome of the organisms provide a much larger information on evolution.

As mentioned in the previous chapter, mutations play an important role in shaping evolution, allowing the organism to adapt to certain environmental conditions [Van03, Chapter 1]. Mutations are caused by events such as base substitutions, insertions and deletions, and recombination and become eventually fixed in populations of species. However, the most predominant mutations are point mutations where single DNA bases are substituted by another base. A substitution in the DNA sequence can be a transition or a transversion. There are four possible transition errors and eight possible transversion errors (see Fig. 2.8). Thus, if mutations would occur randomly, a transversion would be two times more likely than a transition. However, in most cases, transitions occur more frequently than transversions [Fit67; VK77]. Considering that substitutions are the most frequent evolutionary events, many researchers have studied and proposed various models of substitution errors.

Currently, thanks to the revolution in sequencing techniques in the last decades, evolutionary information is easily inferred from comparisons of DNA or amino acid sequences. If two sequences are similar, it is assumed that they have evolved from a common ancestor (homology). For this reason, mathematical models of molecular evolution have been developed which describe temporal changes in the biological sequences. Markov models are often used to model evolutionary substitutions in biological sequences. The Markov models can be applied on both DNA and protein sequences. In DNA sequences, the states of the Markov chains can be either individual nucleotides (i.e., 4 states) or codons (i.e., 61 or 64 states), whereas in protein sequences, the 20 amino acids are considered as states. Since the evolution is happening over a continuous time period, continuous time Markov chains will be used. We will next present the substitution models at nucleotide, codon, and amino acid levels in the literature.

3.1.1 Nucleotide-Based Evolutionary Models

In nucleotide models of evolution, each site of the DNA sequence, independent of other sites, is assumed to evolve according to a Markov chain with state space $\{A, G, C, T\}$. The first and simplest model is proposed by Jukes and Cantor in 1969 [JC69], hence the model is referred to as JC69 model. The JC69 model assumes that each nucleotide has the same mutation rate, i.e., each nucleotide mutates to any nucleotide at a rate of μ per site. Thus, the rate matrix for the ordered set $\{A, G, C, T\}$ is

$$\mathbf{Q_{JC69}}(t) = \frac{1}{4}\begin{bmatrix} -3\mu & \mu & \mu & \mu \\ \mu & -3\mu & \mu & \mu \\ \mu & \mu & -3\mu & \mu \\ \mu & \mu & \mu & -3\mu \end{bmatrix}.$$

The corresponding transition probability matrix $\mathbf{P_{JC69}}(t) = [P_{ij}(t)]\ \forall i, j \in \{A, T, C, G\}$ calculated from the matrix exponential in Eq. (2.45) is

$$P_{ij}(t) = \begin{cases} \frac{1}{4} + \frac{3}{4}\exp^{-\mu t} & \text{if } i = j \\ \frac{1}{4} - \frac{1}{4}\exp^{-\mu t} & \text{if } i \neq j \end{cases}.$$

The entries $P_{ij}(t)$ should be interpreted as follows. If we let every site (position) i of a long DNA sequence to evolve for a duration t, the proportion of nucleotide j in the sequence will be $P_{ij}(t)$. In the limiting case where $t \to \infty$, irrespective of the starting nucleotide frequencies, $P_{ij}(t) = 1/4, \forall i, j$. This implies that so many substitutions have occurred at every position and that led to a random steady state distribution $\boldsymbol{\pi} = [\pi_A, \pi_G, \pi_C, \pi_T]$, with probability 1/4 for every nucleotide. Note that, the JC69 model has only a single parameter.

The other model of DNA evolution is called K80. The K80 model, proposed by Kimura in 1980 [Kim80], takes care of the differences in transition and transversion substitution rates. The estimate of the transition/transversion ratio for the DNA of various organisms ranges from 0.89 to 18.67 [PB97]. This indicates that transitions are more frequent than transversions. Denoting the transition error rate by α and the transversion error rate by β, the rate matrix for the ordered set $\{A, G, C, T\}$ is given by

$$\mathbf{Q_{K80}}(t) = \frac{1}{4}\begin{bmatrix} -(\alpha+2\beta) & \alpha & \beta & \beta \\ \alpha & -(\alpha+2\beta) & \beta & \beta \\ \beta & \beta & -(\alpha+2\beta) & \alpha \\ \beta & \beta & \alpha & -(\alpha+2\beta) \end{bmatrix}.$$

The total substitution rate for any base is therefore $\alpha + 2\beta$. The entries of the transition probability matrix $\mathbf{P_{K80}}(t)$ are

$$P_{ij}(t) = \begin{cases} \frac{1}{4} + \frac{1}{4}\exp^{-\beta t} + \frac{1}{2}\exp^{-2(\alpha+\beta)t} & \text{if } i = j \text{ no mutation} \\ \frac{1}{4} + \frac{1}{4}\exp^{-\beta t} - \frac{1}{2}\exp^{-2(\alpha+\beta)t} & \text{if } i \to j \text{ transition} \\ \frac{1}{4} - \frac{1}{4}\exp^{-4\beta t} & \text{if } i \to j \text{ transversion} \end{cases}.$$

The JC69 and K80 models have symmetric substitution rate matrices and have uniform stationary distributions. This is not true in almost all real data sets. Hence, a number of other methods which will allow unequal base frequencies have been proposed. To mention a few, FEL81 [Fel81], HKY85 [Has+85], and TN93 [TN93]. The most general model, however, is called general time-reversible (GTR) model [Tav86], proposed by Travaré in 1986. It has a stationary distribution $\boldsymbol{\pi} = [\pi_A, \pi_G, \pi_C, \pi_T]$ and a rate matrix

$$\mathbf{Q_{GTR}}(t) = \begin{bmatrix} * & \eta\pi_G & \delta\pi_C & \beta\pi_T \\ \eta\pi_A & * & \varepsilon\pi_C & \gamma\pi_T \\ \delta\pi_A & \varepsilon\pi_G & * & \alpha\pi_T \\ \beta\pi_A & \gamma\pi_G & \alpha\pi_C & * \end{bmatrix},$$

where α, β, γ, δ, ε, and η denote the rates of substitutions between $T \rightleftharpoons C$, $T \rightleftharpoons A$, $T \rightleftharpoons G$, $C \rightleftharpoons A$, $C \rightleftharpoons G$, and $A \rightleftharpoons G$, respectively. The diagonal elements are calculated in such a way that the row sums equal zero. The transition diagram is shown in Fig. 3.1. The GTR model has 9 free parameters: 6 for the rates and 3 for the nucleotide frequencies.

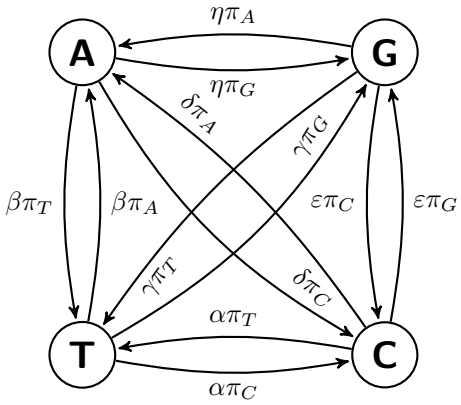

Fig. 3.1: Graphical representation of the GTR model

The other models could be derived from the GTR method by imposing various constraints. For instance, considering only differences in transition, i.e., $\eta = \alpha$, and transversion, i.e.,$\gamma = \delta = \varepsilon = \beta$, rates, the GTR model reduces to what is referred to as the HKY model. The HKY model has 5 parameters and is known to work well for most cases. If a further assumption of $\pi_A = \pi_G = \pi_C = \pi_T = \frac{1}{4}$ is imposed, the HYK model reduces to the K80 model.

3.1.2 Protein-Based Evolutionary Models

Unlike the evolutionary models at the nucleotide levels, most of the models of protein evolution are empirical. The amino acid substitution rates are estimated from a large set of protein sequences. The amino acid based Markov chains are thus

20×20. Dayhoff et al. [Day+78] estimated the first such model in 1972, resulting in the widely used point accepted mutations (PAM) matrix. An accepted mutation is the replacement of a single amino acid by another amino acid which is "accepted" by natural selection. The PAM matrix is derived from an ungapped multiple alignments of closely related protein sequences that are at least 85% identical. To construct the matrix, they used alignments of 71 families of proteins containing 1572 mutations. The alignments are used to produce the most parsimonious phylogenetic trees for inferring evolutionary relationships. A parsimonious tree is the one which requires the fewest evolutionary changes. After the reconstruction of the evolutionary tree, the number of substitutions of the amino acid i by j, A_{ij}, is counted. Then, the relative mutability, which is the probability that the amino acid will change in a given small evolutionary time. The relative mutability m_j is computed as the ratio of the number of times each amino acid has changed to the number of times it occurred in the sequence. Finally, the PAM mutation matrix is determined. The off-diagonal entries of the matrix M_{ij}, which represent the probability that each amino acid is replaced by another amino acid, are computed as

$$M_{ij} = \lambda m_j \frac{A_{ij}}{\sum_{i=1}^{20} A_{ij}} \,. \tag{3.1}$$

The diagonal entries are determined as

$$M_{ii} = 1 - \lambda m_i \,. \tag{3.2}$$

The constant λ is used to calibrate the matrix such that it represents changes in terms of evolutionary time scale. PAM1 represents a time period over which 1 % of the amino acids are expected to undergo accepted point mutations. Since PAM1 represents the one-time step transition, the PAM matrices for higher evolutionary distances are obtained by exponentiation, i.e., PAMn = PAM1^n. PAM250, i.e., 2.5 point accepted mutations per amino acid, is the most commonly used matrix.

Mostly, the PAM matrices are used for scoring sequence alignments, so that the optimal alignment is selected from a set of possible alignments. Therefore, the PAM mutation matrices are further converted into scoring matrices by using log-likelihood ratios of the observed amino acid exchanges to the probabilities in random substitutions.

The other amino acid based parametric substitution model was proposed in 1992 by Henikoff and Henikoff and is referred to as block substitution matrix (BLOSUM) [HH92]. The BLOSUM matrices are not based on evolutionary distances rather they are based on multiple alignment, without gaps, of highly conserved region in proteins (using the Blocks database). They used 504 groups and 2205 blocks of

proteins and clustered the sequences together if they are more than 62 % identical, hence, the matrix is named as BLOSUM62. In general, unlike the PAM matrices, the matrix can be constructed for any percent identity. Most commonly, BLOSUM90 (at 90 % percent identity) is well suited for comparing closely related sequences, while BLOSUM30 is employed for distantly related sequences.

Later on, many other amino acid empirical substitution matrices were proposed. To mention a few, Vogt et al.'s substitution matrix is based on chemical properties of the amino acid side chains [Vog+95], Risler et al. proposed another one using 3D structural alignments [Ris+88], and Gonnet et al. produced a 400×400 dipeptide substitution matrix [Gon+94]. Whelan and Goldman (WAG) proposed a novel approach to estimate amino acid replacement matrices from a large database of aligned protein sequences in 2001 [WG01]. It combines the estimation of transition and scoring matrices by a maximum-likelihood approach. The WAG matrix also assumes that all positions in the sequence evolve independently according to a stationary, time homogeneous, and reversible Markov process. They used a counting method proposed in [Jon+92], similar to the one used in PAM, which gives a near-optimal tree (maximum parsimonious). Then, the transition probabilities are optimized by a maximum-likelihood approach, under the assumption that the optimal tree topologies are known.

3.1.3 Codon-Based Evolutionary Models

When applied to protein coding regions, the nucleotide and amino acid based models ignore important evolutionary information. The nucleotide models assume each site evolves independently, neglecting the codon structure imposed by the genetic code. However, it is known that there are evolutionary differences among the three codon positions [BG07]. The amino acid based models on the other hand cannot take into account synonymous changes which result in the preservation of amino acids. This will omit important evolutionary information associated with selective pressure. Thus, codon-based evolutionary models have been proposed to alleviate this problems and offer accurate description of evolutionary processes. For an elegant summary and review of issues involving codon models, we refer the reader to [CS12].

The first codon models proposed in 1994 were parametric [MG94; GY94]. The parameters used to describe the models are the ratio of non-synonymous to synonymous substitution rates ($w = \frac{dN}{dS}$) which describe selection, codon equilibrium frequencies $\pi_i,\ i \in \mathcal{C}$ ($\mathcal{C}$ is the set of all codons), and transition/transversion ratio κ. If $w > 1$, the change of the amino acid is considered to result in a positive selection, i.e., the amino acid substitution increases the protein fitness. Whereas $w < 1$ and

$w = 1$ imply negative and neutral selection, respectively. The instantaneous rate of change from codon i to codon j of the most popular the Goldman and Yang model is

$$Q_{ij} = \begin{cases} 0 & \text{if more than one change,} \\ \pi_j & \text{if synonymous transversion,} \\ \kappa\pi_j & \text{if synonymous transition,} \\ w\pi_j & \text{if non-synonymous transversion,} \\ w\kappa\pi_j & \text{if non-synonymous transition.} \end{cases} \tag{3.3}$$

Schneider et al. [Sch+05] in 2005 proposed the first empirical codon model. The complete substitution matrix is estimated from alignments of orthologous sequences following a similar approach used in constructing PAM matrices. The genome sequence of five vertebrates namely human (*Homo sapiens*), mouse (*Mus musculus*), chicken (*Gallus gallus*), frog (*Xenopus tropicalis*), and zebrafish (*Brachydanio rerio*) was investigated. This model was shown to perform better than empirical amino acid based models for sequence alignment, especially for closely related species where amino acid replacements are relatively rare.

Other semi-empirical codon-based models combine empirical rates of substitution with parameters that provide flexibility. Doron-Faigenboim and Pupko [DFP07] used an empirical amino acid substitution matrix and incorporated the parameters w, κ, and π_i to produce a 61×61 matrix. Kosiol et al. [Kos+07] proposed a semi-empirical model which is constructed by using an empirical codon model and introducing parameters which allow different transition/transversion and non-synonymous/synonymous rate ratios. In 2011, Zoller and Schneider [ZS11] proposed another semi-empirical codon model based on principal component analysis (PCA) of sequence alignment data. The PCA is done to identify the most relevant parameters for codon substitution models. They constructed their substitution matrix using a linear combination of the most important principal components with the determined parameters as coefficients.

3.2 Biological Communication Models

Several researchers have proposed channel models for biological information transfer. The models are mostly based on analogies between communication systems of data storage and transmission depicted in Fig. 2.1 and the biological flow of information from DNA to mRNA to proteins (the central dogma of molecular biology shown in Fig. 2.4) [Gat72; Yoc92; RR+96; May+04; Gon+11].

Gatlin's communication model: Gatlin's work in 1972 [Gat72] was the first, to our knowledge, to explore information theoretic aspects of biological information processing systems. Gatlin's model assumes the DNA base sequence as an encoded message generated by a source, the steps to protein production, i.e., transcription and translation, as a channel, and the amino acid sequence as a received message. The model is presented in Fig. 3.2. Although this model takes into account the transcription and translation processes, the role of DNA replication is not clearly incorporated.

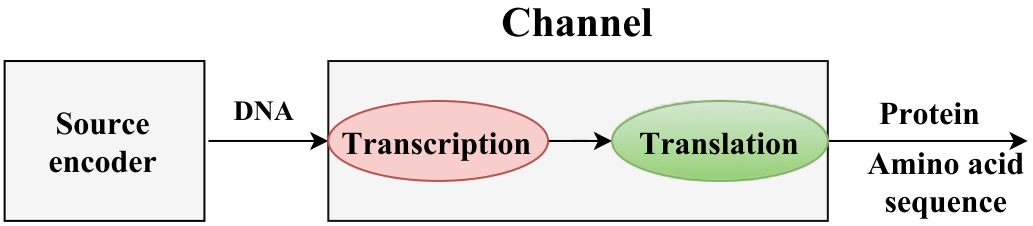

Fig. 3.2: Gatlin's communication model.

Yockey's communication model: Yockey described the central dogma and other biological information theoretic aspects [Yoc92; Yoc74]. He viewed the flow of information from DNA to RNA to proteins as a communication system and employed entropy, rate, and capacity calculations with a transition matrix he developed by considering base changes of equal probability. A detailed analysis of the application of information theory to molecular biology can be found in his book [Yoc92]. Yockey's model is based on a data storage model whereby the genetic information system is paralleled to that of a Turing machine. As shown in Fig. 3.3, the genetic message is generated by a stationary Markov process and recorded in the DNA sequence, similar to tape-recording [Yoc05]. Transcription is assumed as an encoding procedure to produce an mRNA code. The mRNA is considered as a channel which transmits the genetic message to the ribosomes, which act as a decoder. The decoding procedure from the mRNA code to the 20 letters of the protein sequence is called translation.

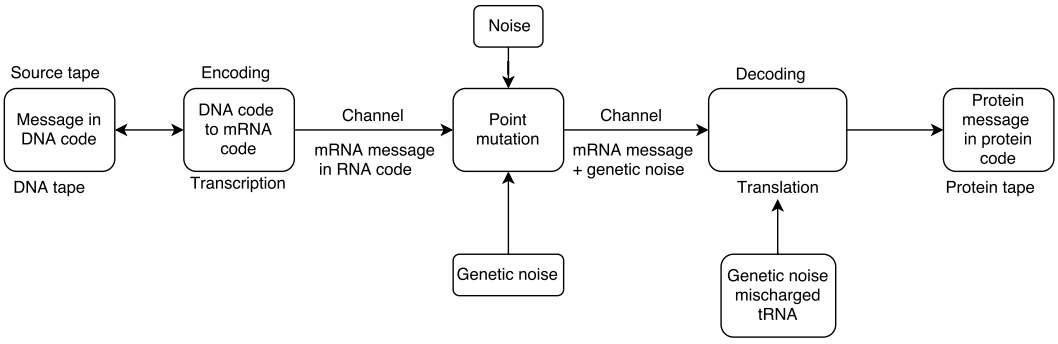

Fig. 3.3: Yockey's communication model [Yoc92].

Roman-Roldanv et al.'s communication model: Roman-Roldanv et al. [RR+96] also proposed a communication channel model whereby the genetic code is viewed

as a channel in which DNA sequences are transmitted and proteins are received. Hence, it is a DNA-protein communication channel. They defined the genetic information source as an ergodic and stationary source that generates messages from a finite alphabet and the transmission channel is assumed to be stationary and memoryless. This model is similar to Gatlin's in the way that DNA rather than mRNA is at the input to the channel. The problem with this approach is that it does not explain the existence of non-coding DNA. The genetic channel transition probability for the noiseless (mutation free) case is represented as

$$p(A_i|B_1B_2B_3) = \begin{cases} 1 & \text{if } p(A_i|B_1B_2B_3) \text{ is part of the genetic code,} \\ 0 & \text{otherwise} \end{cases}, \tag{3.4}$$

where A_i is the ith amino acid and $B_1B_2B_3$ specifies a codon.

May et al.'s communication model: May et al. [May+04] argues that the Gatlin's, Yockey's, Roman-Roldanv et al.'s models ignore the important role of the replication process. In addition, Yockey's model presents transcription as an encoding procedure. However, the transcription process may introduce errors and therefore it is inconsistent with the notion of encoders in communication systems. Thus, May et al. proposed a communication model depicted in Fig. 3.4. The DNA is viewed as an encoded message, which is transmitted through an error-introducing genetic channel of the replication process. The translation and transcription processes are considered as parts of the decoding process producing the received protein message.

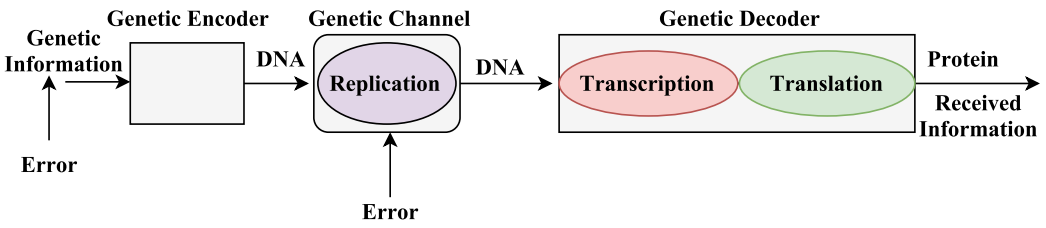

Fig. 3.4: May et al.'s communication model (modified from [May+04]).

Gong et al.'s communication model: Gong et al. [Gon+11] pointed out that in normal communication systems, apart from degradation from channel effects, the transmitted and received information are the same. This is not the case in the previous DNA-to-protein models. From a communication perspective, the DNA-to-protein system is a decoding procedure. Therefore, they introduced an abstract and biologically non-existent source-channel encoder that produces an encoded DNA message from the protein source information. Other than the depiction of proteins as the source information and the lumped genetic noise arising from transcription, translation, replication, and point mutations, their model is similar to May et al.'s (Fig. 3.4). Using the PAM matrix and a channel matrix they produced they performed capacity calculations.

3.3 A Mutation Matrix as a Communication Channel

Among the models of evolution and communication described in the previous sections, we selected Schneider et al.'s Empirical Codon Mutation (ECM) matrix for performing channel capacity computation. We also assumed a communication model similar to May et al.'s and Gong et al.'s, in which the codon-based channel is followed by a decoder performing transcriptional and translational processes using the standard genetic code. However, the input and outputs of the channel are codons. The system model is shown in Fig. 3.5.

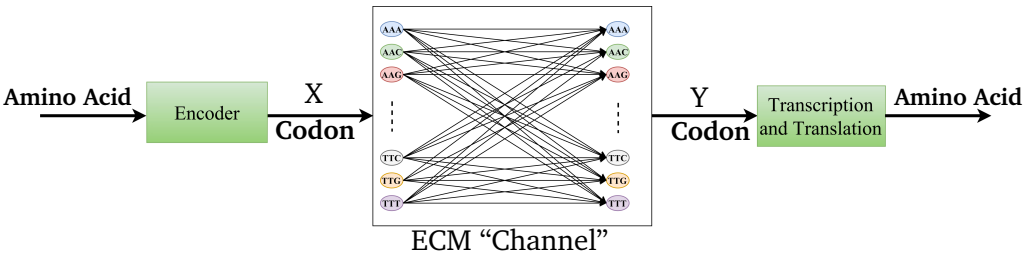

Fig. 3.5: A codon-based communication model with the ECM matrix as a "channel". The communication model is adopted from [Gon+11]).

We choose a codon-based model because of two reasons. The first reason is, compared to protein-based models, the codon level models demonstrate mutational changes among the codons and this gives us more information by highlighting the tendency of mutations between codons encoding the same amino acid (synonymous changes) as well as the mutational effects between codons that code for different amino acids (non-synonymous changes). The second reason is that the codon-based models capture the evolutionary differences in the three codon positions. The ECM matrix was constructed by summarizing biological mutations for about 300 Million years. Thus, it should provide a fair estimate of mutation rates. Furthermore, as described in Section 3.1.3, differences among the three codon positions are inherently included in the model.

3.3.1 Capacity of the Codon Mutation Matrix

In order to compute the mutation probability in the ECM matrix, 17502 alignments of sequences from five vertebrate genomes yielded 8.3 million aligned codons from which the number of substitutions between codons were counted [Sch+05]. This matrix has 64×64 entries stating the mutation probability of each codon to every other codon. Basically, the substitution from sense codons to stop codons is not included in the ECM matrix, which makes the matrix block diagonal with a 61×61

matrix for coding codons and a 3×3 entries for substitutions between stop codons. Therefore, we will consider only substitutions between coding codons and regard the ECM matrix as 61×61. From the communication perspective, this mutation matrix describes channel transition probabilities $\mathbf{P}(y|x)$.

There is also another matrix in [Sch+05], which gives the actual count of substitutions observed. From this substitution count matrix **C**, we obtained the biological probability distribution of the codons as

$$\mathbf{p}_x = \frac{\sum\limits_j C_{ij}}{\sum\limits_i \sum\limits_j C_{ij}} \ . \tag{3.5}$$

Thereafter, we combined the codons which encode for the same amino acid and computed the probability distribution of amino acids, denoted $\mathbf{p}_a$. Using this distribution, the to be preserved information content of the 64 codons representing the 20 amino acids can be computed as

$$R_{20} = -\sum_{i=1}^{20} \mathbf{p}_a(i) \log_2(\mathbf{p}_a(i)) = 4.1875 \text{ bit} \ , \tag{3.6}$$

which is less than the maximum value of $\log_2(20) = 4.3219$ bit. Likewise, the required rate obtained by using the amino acid probability distribution provided by King & Jukes in [KJ69], derived from 5492 residues of 53 vertebrate polypeptides is 4.2033 bit. Thus, it is reasonable to look for a capacity that is at least greater than 4.1875.

According to Shannon's channel coding theorem (see Section 2.1.4), a communication through a noisy channel of capacity C at an information rate of R is possible with an arbitrarily small probability of error, if $R < C$ [Sha48]. Hence, the channel capacity has to, at least, exceed the value of R_{20}.

In communication systems, the channel capacity is determined by maximizing the mutual information $I(X;Y)$ between input (X) and output (Y) over the input probability distribution $\mathbf{p}_x$.

$$C = \sup_{\mathbf{p}_x} I(X;Y) \ . \tag{3.7}$$

For solving the optimization problem, the Arimoto-Blahut algorithm was employed [Ari72], [Bla72]. The Arimoto-Blahut algorithm is briefly described below.

The Arimoto-Blahut Algorithm

The Arimoto-Blahut algorithm is an iterative numerical algorithm that monotonically converges to the capacity value. To compute the capacity, it is starting from any arbitrary input probability distribution $\mathbf{p}_x$ (usually uniform) and performs the following two steps until the algorithm converges.

1. Compute a quantity related to the mutual information per input symbol

$$c(x_j) \coloneqq \exp \sum_k p(y_k|x_j) \log \frac{p(y_k|x_j)}{\sum_j p(x_j)p(y_k|x_j)} \ , \tag{3.8}$$

   This results from a Lagrange multiplier step in [Bla72].

2. Update the input probability distribution according to

$$p(x_j) \coloneqq \frac{p(x_j)c(x_j)}{\sum_k p(x_k)c(x_k)} \ . \tag{3.9}$$

The termination criteria is based on the lower and upper bounds of channel capacity,

$$\log \left(\sum_j p(x_j)c(x_j) \right) \leq C \leq \log \left(\max_{x_j} c(x_j) \right) \ . \tag{3.10}$$

The iterations are terminated when the upper and lower bounds are equal up to a certain accuracy.

The mutual information which measures the mutual dependence between input and output codon distributions is calculated using Eq. (2.15) as the difference between the entropy of the codon distribution at the output of the ECM "channel" $H(Y)$ and the conditional entropy $H(Y|X)$, referred to as prevarication or irrelevance. $H(Y)$ is computed as

$$H(Y) = -\sum_{i=1}^{61} p_{y_i} \log_2(p_{y_i}), \tag{3.11}$$

where p_{y_i} is the output probability distribution of the i^{th} codon. The conditional entropy $H(Y|X)$ between input and output distribution of codons is computed as

$$H(Y|X) = -\sum_{i=1}^{61} p(x_i) \sum_{j=1}^{61} p(y_j|x_i) \log_2 p(y_j|x_i). \tag{3.12}$$

$p(y_j|x_i)$ is the conditional probability between codons, which is given by the empirical codon mutation (ECM) matrix.

However, in the system we are considering, the input distribution (i.e. probability distribution of codons) is not something to adjust. It is defined by nature. Therefore, we determine the mutual information corresponding to the mutation "channel" matrix for a biological codon frequency obtained by Eq. (3.5). Moreover, we would like to find the optimal input probability distribution of the 61 codons to maximize the mutual information by solving Eq. (3.7) and compare it with the biological distribution.

The mutual information between the input and output of the ECM channel using the biological codon distribution and the optimal codon distribution (i.e. the capacity) is 2.39 and 2.66, respectively. This is below the required rate of 4.1875. Hence, we apply an exponent F to the ECM matrix and compute an exponential factor for the ECM matrix that would still allow for preserving the genetic information given the redundancy that is present in the codon-to-amino acid mapping. This gives an insight on how such a mutation matrix relates to the preservation of a species in an information-theoretic sense. In other words, we want to compute the exponent of the ECM matrix that is needed to match the required rate of 4.1875, i,e., we stepwise reduce it until it satisfies the rate requirement. Hereto, we use the singular value decomposition (SVD) yielding

$$[\mathbf{P}(y|x)]^F = \mathbf{U}(\mathbf{\Sigma})^F \mathbf{V}^\mathsf{T}, \tag{3.13}$$

where $\mathbf{U}, \mathbf{V}$ are unitary matrices, $\mathbf{\Sigma}$ is a diagonal matrix with nonnegative real numbers in the diagonal, and F is an exponent to be fine-tuned. The value of the exponent is changed in steps from zero to one. A value of 1 means the original ECM matrix is used.

The capacity obtained by optimizing the codon distribution, the mutual information based on the observed biological codon distribution, and the required rate are shown together in Fig. 3.6. When the exponent of the ECM matrix is reduced, the output codon distribution changes and the prevarication $H(Y|X)$ will be smaller. As a result, the capacity increases. The maximal exponent which satisfies the rate requirement of 4.1854 bit for an error-free "transmission" using the biological codon frequency is found to be ≈ 0.26. At the same exponent, the optimized "channel" capacity is 4.2586 bit. It can also be seen that the capacity curve is very close to the one found by using the biological codon distribution. This indicates that the biological probability distribution is almost optimally "chosen" to achieve the capacity of the "channel".

It is not surprising that the exponent is not one, since the matrix was obtained comparing five different vertebrate DNAs, the times corresponding to time spans between 40 M – 350 M years. However, the exponent is not extremely small, which indicates that the matrix is at least roughly in agreement with information-theoretic

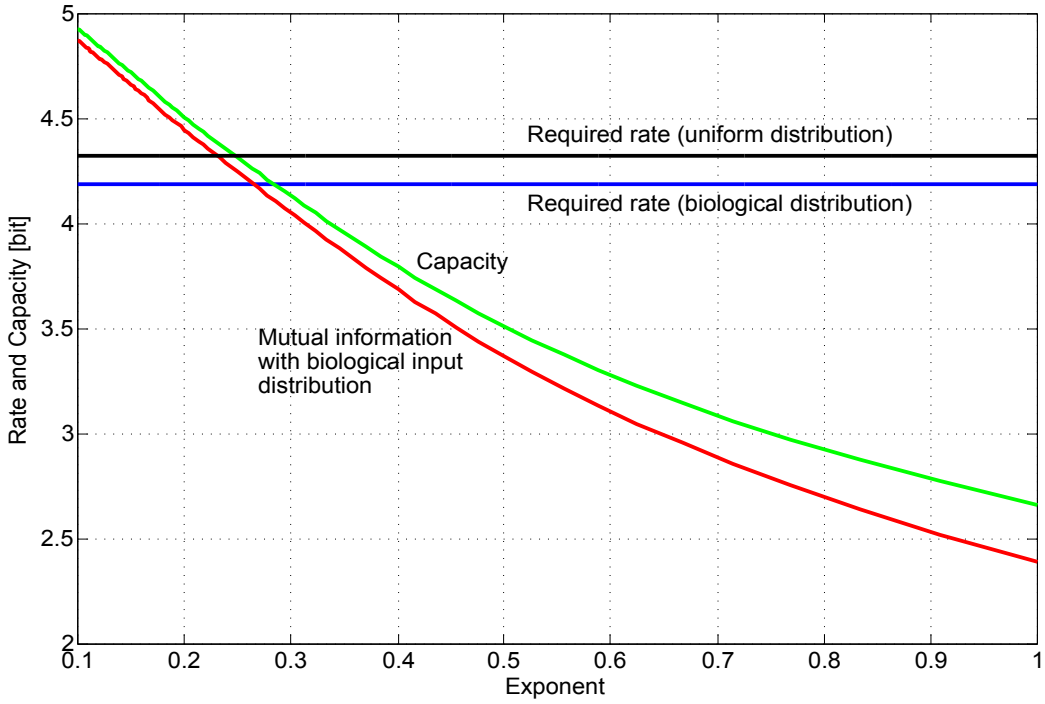

Fig. 3.6: Capacity and mutual information using the biological codon distribution as a function of an exponential factor. The rate requirements with a uniform and the biological distributions are also presented.

calculations. One may also see this as an argument to recompute the matrix using the obtained exponent.

The optimal capacity-achieving codon distribution and the observed biological codon distribution are both shown in Fig. 3.7. The corresponding values are also tabulated in Table 3.1 and Table 3.2.

Once the optimized codon distribution is obtained using the Arimoto-Blahut algorithm, to note the similarity to the biological distribution, we applied the so called Kullback–Leibler divergence (D_{KL}) [CT91]. D_{KL}, defined in Eq.(2.17), is a quantitative measure of how similar a probability distribution P is to a model distribution Q. D_{KL} is non-negative and gives a zero result when the distributions are perfectly matched. Technically speaking, D_{KL} measures the average number of extra bits required (coding penalty) for using a code based on Q instead of P.

The D_{KL} between the optimal and biological distributions is 0.0926 bit, which is not a very small difference (comparable with the D_{KL} of two Gaussians of equal mean and a variance differing by a factor of two) but still, similarities are obvious. Both of the probability distributions satisfy the rate requirement of 4.1875 bit. In addition, the distribution among synonymous codons is very similar. To mention one example,

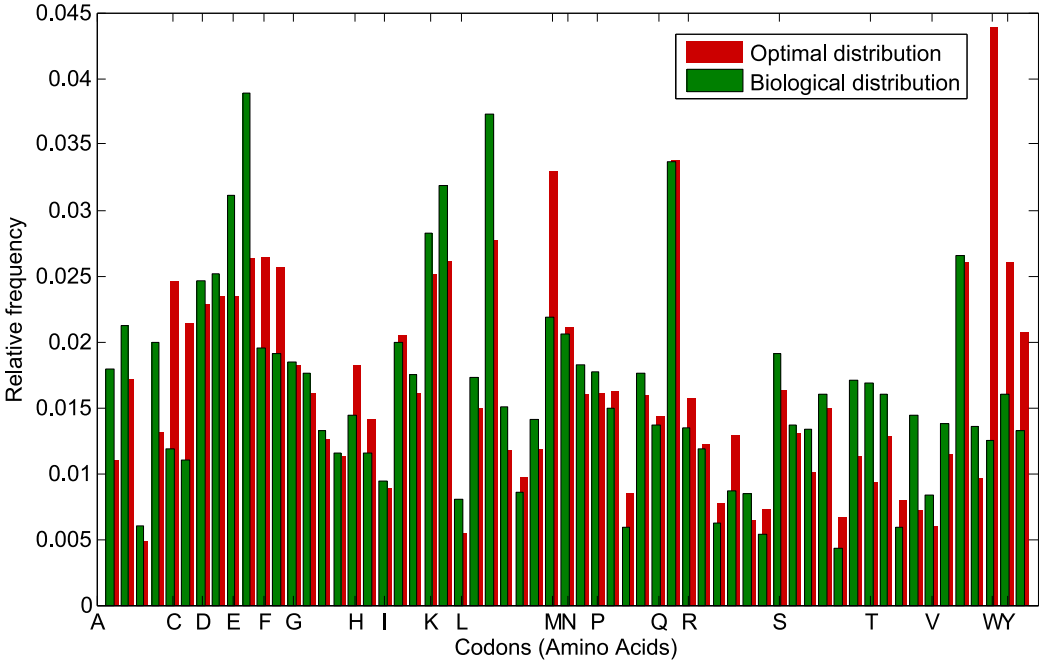

Fig. 3.7: The biological and optimal probability distribution of codons, the codons belong to the same encoded amino acid (one letter symbol) are represented by the consecutive bins. The synonymous codons are alphabetically arranged.

Tab. 3.1: Biological codon relative frequency. The codon relative frequency of the five vertebrate genomes (human, mouse, chicken, frog, and zebrafish) from the data presented by Schneider A., Cannarozzi G., and Gonnet G. [Sch+05].

	Codon	Freq.	Codon	Freq.	Codon	Freq.	Codon	Freq.	
T	TTT	0.0191	TCT	0.0171	TAT	0.0132	TGT	0.0110	T
	TTC	0.0196	TCC	0.0160	TAC	0.0160	TGC	0.0119	C
	TTA	0.0085	TCA	0.0133	TAA	0.0003	TGA	0.0003	A
	TTG	0.0141	TCG	0.0043	TAG	0.0001	TGG	0.0125	G
C	CTT	0.0150	CCT	0.0176	CAT	0.0116	CGT	0.0054	T
	CTC	0.0173	CCC	0.0150	CAC	0.0144	CGC	0.0087	C
	CTA	0.0080	CCA	0.0178	CAA	0.0137	CGA	0.0062	A
	CTG	0.0373	CCG	0.0059	CAG	0.0337	CGG	0.0085	G
A	ATT	0.0175	ACT	0.0144	AAT	0.0182	AGT	0.0136	T
	ATC	0.0200	ACC	0.0160	AAC	0.0206	AGC	0.0191	C
	ATA	0.0094	ACA	0.0169	AAA	0.0282	AGA	0.0135	A
	ATG	0.0219	ACG	0.0059	AAG	0.0319	AGG	0.0118	G
G	GTT	0.0136	GCT	0.0200	GAT	0.0252	GGT	0.0115	T
	GTC	0.0138	GCC	0.0213	GAC	0.0246	GGC	0.0176	C
	GTA	0.0084	GCA	0.0179	GAA	0.0311	GGA	0.0184	A
	GTG	0.0265	GCG	0.0060	GAG	0.0389	GGG	0.0133	G
	T		C		A		G		

codons encoding Alanine (A) in decreasing order of abundance, is GCC, GCT, GCA, and GCG, for both the biological and the capacity-achieving distributions.

Tab. 3.2: Calculated codon relative frequency. The codon relative frequency that maximizes the mutual information between input and output and yielding a capacity close to what is required for preserving the information content of amino acids. An exponential factor of 0.26 is applied to the ECM matrix.

	Codon	Freq.	Codon	Freq.	Codon	Freq.	Codon	Freq.	
T	TTT	0.0257	TCT	0.0113	TAT	0.0207	TGT	0.0215	T
	TTC	0.0264	TCC	0.0150	TAC	0.0260	TGC	0.0247	C
	TTA	0.0097	TCA	0.0100	TAA	*	TGA	*	A
	TTG	0.0119	TCG	0.0066	TAG	*	TGG	0.0439	G
C	CTT	0.0118	CCT	0.0159	CAT	0.0141	CGT	0.0073	T
	CTC	0.0150	CCC	0.0162	CAC	0.0183	CGC	0.0129	C
	CTA	0.0054	CCA	0.0161	CAA	0.0144	CGA	0.0077	A
	CTG	0.0277	CCG	0.0085	CAG	0.0337	CGG	0.0065	G
A	ATT	0.0162	ACT	0.0071	AAT	0.0160	AGT	0.0130	T
	ATC	0.0205	ACC	0.0128	AAC	0.0212	AGC	0.0163	C
	ATA	0.0088	ACA	0.0093	AAA	0.0251	AGA	0.0157	A
	ATG	0.0330	ACG	0.0079	AAG	0.0261	AGG	0.0122	G
G	GTT	0.0096	GCT	0.0132	GAT	0.0234	GGT	0.0114	T
	GTC	0.0114	GCC	0.0172	GAC	0.0228	GGC	0.0162	C
	GTA	0.0060	GCA	0.0110	GAA	0.0235	GGA	0.0183	A
	GTG	0.0260	GCG	0.0048	GAG	0.0263	GGG	0.0126	G
	T		C		A		G		

The exponent applied to the ECM matrix can be understood as changing the mutation rate. The mutation rate is calculated by summing the off-diagonal entries of the matrix, i.e.,

$$\text{Mutation rate} = 1 - \sum_{i=j} p_i P_{ij} \ . \tag{3.14}$$

The ECM matrix has a mutation rate of 65 %. Exponentiation of the matrix with an exponent between 0 and 1, will result in smaller off-diagonal entries and hence the mutation rate is reduced. For an exponent bigger than 1, the off-diagonal elements will be larger and the mutation rate is increased. The optimal exponent 0.26 corresponds to a mutation rate of 29 %.

3.4 Comparison of Mutation and Chemical distances

The ECM matrix shows the substitution probabilities of a codon to every other codon. Our aim here is to see how these changes relate to chemical differences among the amino acids which the codons encode according to the genetic code. To do that, we have selected an amino acid based chemical distance matrix proposed by Grantham [Gra74] which estimates the chemical difference between amino acids that combines three chemical properties: composition, polarity, and molecular volume. The three

chemical properties are selected because of their high correlation with amino acid substitution frequencies. In Euclidean space with these three chemical properties at the axes, this matrix gives the distance ($D_{ij}^{(c)}$) between the ith and jth amino acid.

To compare the mutation and chemical distances, the Euclidean distance between codons has to be determined. To model the Euclidean distance between codons, we have used the pairwise error probability expression by assuming a Gaussian i.i.d "channel" with a constant standard deviation (σ) in relating mutation probabilities to distances. The pairwise error probability (PEP) then results in

$$P_{ij} = \frac{1}{2}\text{erfc}\frac{D_{ij}^{(m)}}{\sqrt{2}\sigma} \ , \tag{3.15}$$

where $D_{ij}^{(m)}$ is the Euclidean distance between ith and jth codons, σ corresponds to a standard deviation, and P_{ij} is the mutation probability between ith and jth codons.

With the PEP expression, the distance matrix between codons of size 61×61 was acquired, i.e.,

$$D_{ij}^{(m)} = \sqrt{2}\sigma\text{erfc}^{-1}(2P_{ij}) \ . \tag{3.16}$$

The high dimensionality of the matrices makes it very difficult for making comparisons. Hence, we will first reduce the number of dimensions to two or three. We used a technique called classical multidimensional scaling (CMDS) [BG05] to reduce the dimensions of the matrices. In the following sections, we present the mathematics behind the CMDS and the results of the dimension reduction.

3.4.1 Classical Multidimensional Scaling

In this section, the mathematics behind CMDS technique will be described. The reference used for this section is [Che10].

Assume that we have observed $n \times n$ Euclidean distance matrix $\mathbf{D} = [d_{ij}]$ derived from a raw $n \times p$ data matrix $\mathbf{X}$. With CMDS, the aim is to recover the original data matrix of n points in p dimensions from the distance matrix. However, since distances are invariant to change in location, rotation, and reflections, those are arbitrary. Define an $n \times n$ matrix $\mathbf{B}$ such that

$$\mathbf{B} = \mathbf{X}\mathbf{X}^T \ . \tag{3.17}$$

The elements of $\mathbf{B}$ are given by

$$b_{ij} = \sum_{k=1}^{p} x_{ik} x_{jk} \ . \tag{3.18}$$

Similarly, since D is a distance matrix, the squared Euclidean distances can be written as

$$\begin{aligned} d_{ij}^2 &= \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \ , \\ &= \sum_{k=1}^{p} x_{ik}^2 + \sum_{k=1}^{p} x_{jk}^2 - 2 \sum_{k=1}^{p} x_{ik} x_{jk} \ , \\ &= b_{ii} + b_{jj} - 2 b_{ij} \ . \end{aligned} \tag{3.19}$$

If we can rewrite the b_{ij}s in terms of the d_{ij}'s, $\mathbf{X}$ can be derived from $\mathbf{B}$. However, unless a location constraint is introduced, a unique solution cannot be found to determine $\mathbf{B}$ from $\mathbf{D}$. Commonly, the center of the columns of $\mathbf{X}$ are set to the origin, i.e.,

$$\sum_{i=1}^{n} x_{ik} = 0 \ , \forall k. \tag{3.20}$$

The added constraint will also mean that the sum of the terms in any row of $\mathbf{B}$ is zero.
Let T be the trace of $\mathbf{B}$ and observe that

$$\sum_{i=1}^{n} d_{ij}^2 = T + n b_{jj} \ , \tag{3.21}$$

$$\sum_{j=1}^{n} d_{ij}^2 = n b_{ii} + T \ , \tag{3.22}$$

$$\sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}^2 = 2nT \ . \tag{3.23}$$

Solving for b_{ij},

$$b_{ij} = -\frac{1}{2} \left[d_{ij}^2 - \frac{1}{n} \sum_{j=1}^{n} d_{ij}^2 - \frac{1}{n} \sum_{i=1}^{n} d_{ij}^2 + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}^2 \right] \tag{3.24}$$

Applying singular value decomposition (SVD) on $\mathbf{B}$,

$$\mathbf{B} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}' = \mathbf{V} \mathbf{\Lambda}_\mathbf{1}^{\frac{1}{2}} \mathbf{\Lambda}_\mathbf{1}^{\frac{1}{2}} \mathbf{V}' \ . \tag{3.25}$$

Using only the 2 (or 3) biggest eigenvalues, λ_1, λ_2 (λ_3) and the corresponding eigenvectors u_1 and u_2 (u_3) we obtain

$$\mathbf{X} = \mathbf{V_1} \mathbf{\Lambda}_\mathbf{1}^{\frac{1}{2}} \ , \tag{3.26}$$

where $\mathbf{\Lambda_1} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$ and $\mathbf{V_1} = [\mathbf{u_1}\mathbf{u_2}]$ for 2 dimensions.

3.4.2 Dimension Reduced Matrices and Observations

The two-dimensional (2-D) plots of the mutation and chemical distance matrices are shown in figures 3.8 and 3.9, respectively. The 3-D plot of the ECM mutation matrix shown in Fig. 3.12 is a better representation but is hard to visualize on paper. The codons encoding the same amino acid are bundled together. From the matrix alone, one can already observe interesting properties. For instance, with a middle T, mutations seem to be more probable at both sides of the codon. From the 2-D representations, we were able to see similar clusterings, which means that highly probable mutations are between codons of similar chemical properties, at least in terms of polarity, chemical composition, and molecular volume. Also, the clusterings of amino acids are mostly consistent with the common Taylor classification shown in Fig. 3.10, which classifies amino acids based on their physicochemical properties [Tay86]. Using these observations we can deduce that most of the mutational changes will not lead to a significant change of chemical properties. However, there are also some inconsistencies where lower mutation distances come together with higher chemical distances and vice versa. The explanation for the inconstancies can be obtained by studying the chemical properties of the amino acids in details. For example, “cysteine”, which allows for disulfide bridges between cysteine residues within a polypeptide and has, hence, a special outlier position in the chemical distance matrix. Large chemical distances together with small mutation distances will, of course, require some protection mechanisms that one has to localize. The inconsistencies are listed below.

Large chemical distance but small mutation distance:

- **C** with **“all others”**
- **G** with **E**
- **S** with **{P,T,A}**
- **{D,N}** with **E**
- **{D,N}** with **G**
- **{Q,H}** with **{W,Y}**
- **K** with **N**

Small chemical distance but large mutation distance:

- **{W,Y}** with **{F,L,M,I,V}**

- **{P,T,A}** with **{Q,H,R}**

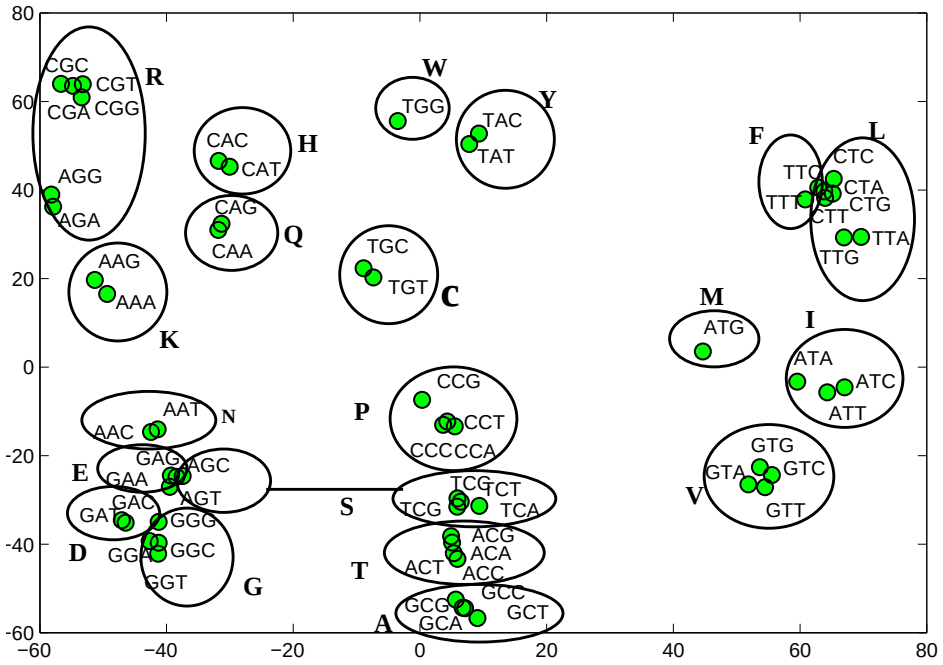

Fig. 3.8: 2-D plot of the mutation distance matrix.

To have a detailed view into the groupings, we performed a simple hierarchical clustering of the amino acids using the mutation and chemical distances. The Unweighted Pair Group Method with Arithmetic Mean (UPGMA), proposed by Sokal and Michener [Sok58], was employed to produce the rooted trees in Fig. 3.11. In the UPGMA method, at each iteration, the pairs with the smallest distance are clustered together and the distance matrix is updated considering the already clustered points as a single object. The process continues until two clusters are left.

The CMDS method works best if the eigenvalues used for reconstruction are very large compared to the unused eigenvalues. However, in our case the eigenvalues are not decaying very quickly, and hence the error in 2-D representation is significant, with a root mean squared error of around half the mean distance. Including more dimensions for representation would improve the accuracy. However, it will be difficult to visualize. The 3-D plot of the ECM mutation matrix is shown in Fig. 3.12. Another option would be to apply possibly better suited dimension reduction and clustering methods.

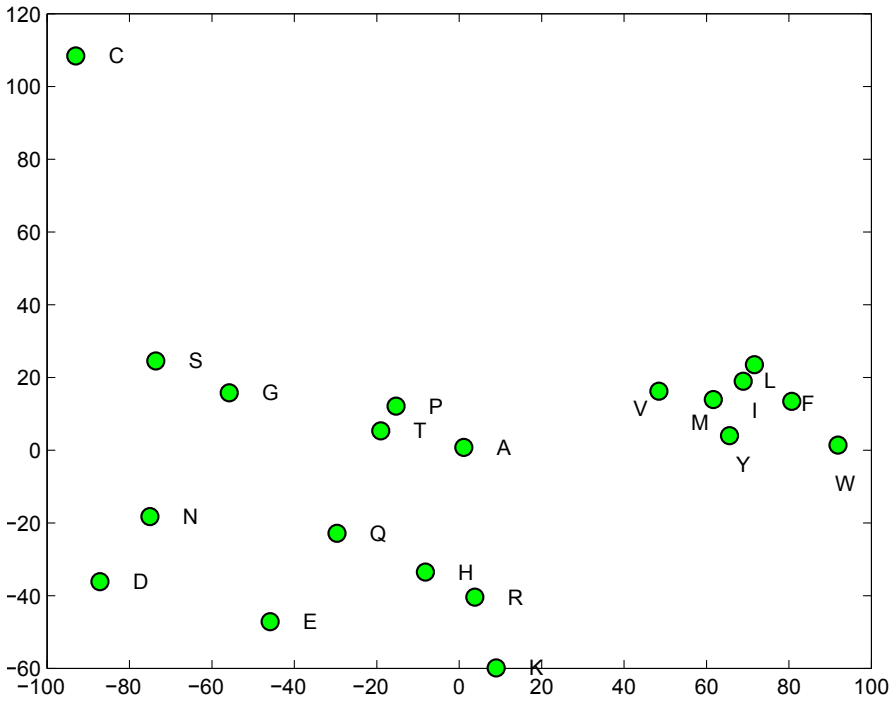

Fig. 3.9: 2-D plot of the chemical distance matrix.

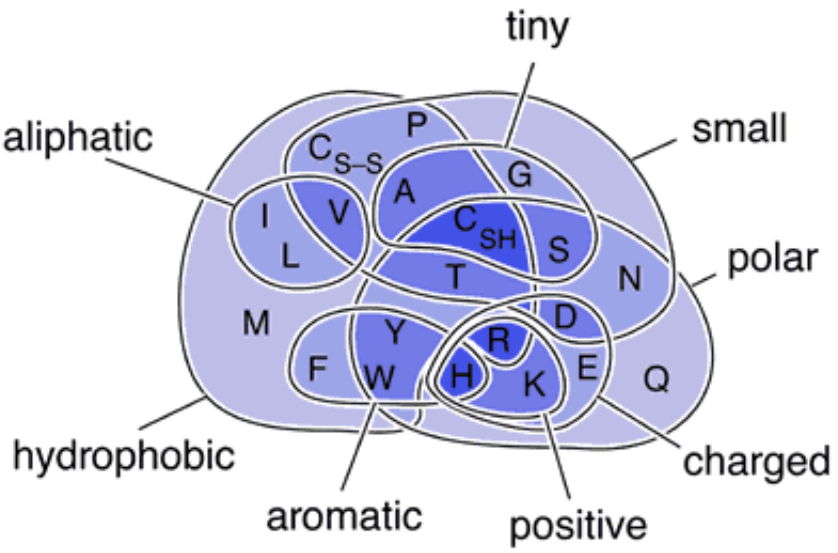

Fig. 3.10: Taylor classification of amino acids [Com07].

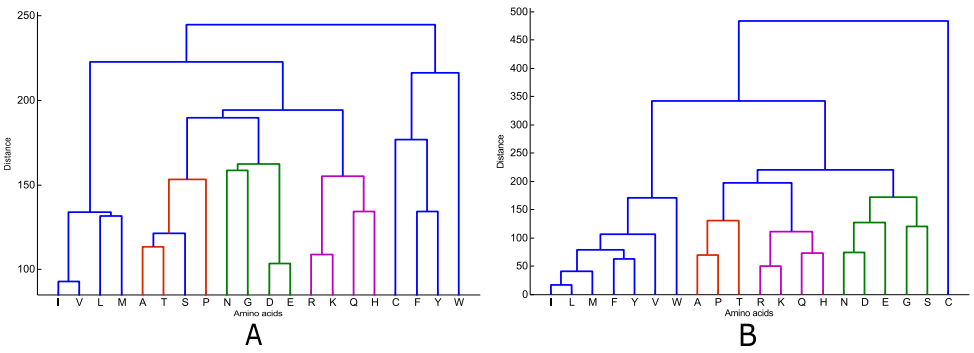

Fig. 3.11: Clustering of amino acids using A) mutation B) chemical distance

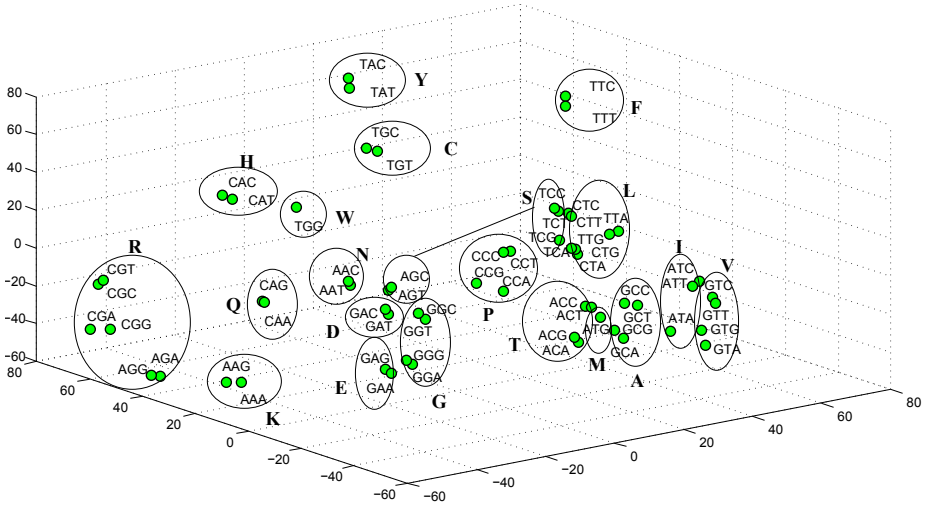

Fig. 3.12: 3-D plot of the mutation distance matrix.

3.5 Summary

In this chapter, we regarded a codon mutation probability matrix, empirical codon substitution matrix (ECM), as a communication channel and performed capacity computations. In addition, the relationship between mutation and chemical distances applying dimension reduction on the corresponding mutation and chemical distance matrices was investigated. From the codon usage in five vertebrates, we found the rate required to preserve the genetic information represented by the 20 amino acids to be 4.1875 bit. This is less than the 4.323219 bit, assuming a uniform amino acid distribution. The entries of the ECM matrix are considered as the transition probabilities of the genetic channel and an exponent was introduced to fine-tune such that the rate requirements are satisfied. It was found that an exponent of 0.26, corresponding to a mutation rate of 29 %, leads to a capacity of 4.1875 bit. The optimal codon distribution which results for the desired channel capacity was determined and compared to the biological distribution. Although the two distributions are not too similar, the relative abundances of the synonymous codons are identical. In addition, in terms of the mutual information, the biological distribution is not too far from the optimal capacity-achieving distribution. This shows that that the biological distribution is well “chosen”.

A comparison between chemical properties of amino acids and mutation probabilities of codons was carried out using the classical multidimensional scaling method. The results showed that most of the highly probable mutations will not lead to a dramatic change of chemical properties. However, some inconsistencies were also observed. Thus, further studies of the severeness of the mutations and possible protection mechanism to counteract the effects is required. In addition, the error introduced in representing 64-dimensional data with two dimensions is significant. This is due

to the slow decay of the eigenvalues of the data. Therefore, another dimension reduction and clustering method with a better performance may be applied in the future.

4 Digital Information and Thermodynamic Stability in Bacteria

In this chapter, we present an analysis on the digital information content of genomes and the couplings to an analog type of information represented by thermodynamic stability. In Section 4.1, we first introduce the motivations and concepts of analog and digital information. Then, in Section 4.2, we present the entropy measures used to quantify the information. In Section 4.5.1, different functional classes of genes are introduced. In Section 4.3, the entropy measures are applied to complete genomes of four selected bacterial genomes. In Section 4.4, the entropy profiles are analyzed considering only the coding sequences, removing the non-coding parts. In Section 4.5, the relation of the digital and analog information to the functional classes of genes is investigated. In Section 4.6, we show an application of the Gibbs entropy for the identification of coding and non-coding regions. Finally, we conclude and point out future works in Section 4.7.

4.1 Introduction

The double-helical DNA polymer is the carrier of the genetic information required for the reproduction of any organism. This information is inscribed by the sequence of four bases. The unique succession of the base pairs (letters) in a gene dictating the production of RNA molecules and proteins provides for digital type of information. The digital nature of the information can also be seen in gene expression where the information indication whether a gene is expressed or not corresponds to the "on-or-off" type digital logic [MT13]. However, there is another type of information, termed "analog code", that coexists with the digital code and is related to physicochemical properties of the DNA [MT13; Tra+12]. This three-dimensional information emerges as a result of dynamic structural and topological variations of the chromosomal DNA and is involved in facilitating and regulating the gene expression, chromosome compaction, and replication [Sob+13; TM13; Son+11]. The analog nature of this information is obvious because it is the additive interactions of successive base steps rather than individual base pairs which determine the physicochemical properties of the polymer. These properties, including DNA thermodynamic stability and

supercoiling, are by definition continuous properties that play a central role in determining the strength of gene expression [Sob+13].

The two types of information are intrinsically coupled by the primary DNA sequence. The physicochemical properties characterizing the analog information are largely sequence dependent. Preferred direction for bending (anisotropy), stiffness, thermodynamic stability, and supercoiling are among the properties that are essentially dependent on the DNA sequence organization [Koo+86; Zhu83; Sob+13]. Previous studies provided compelling arguments concerning the peculiar relationship or interdependence between the two types of DNA information [TM15; Tra+12; Sob+13; MT13; TM13; Mus15].

The average information content of the genome can be measured using Shannon entropy [Sha48]. This information is related to the digital code in the DNA. So far, researchers have extensively applied this information-theoretic measure for studying a wide variety of topics in molecular biology and bioinformatics, including DNA pattern recognition, gene prediction, sequence alignment, and comparative genomics [Akh+13; Cha+05; Sch10; SS; Sch+86; RR+96; Cap+04; Hag+04]. However, due to the existence of an equally important analog code, solely looking at the base or codon composition in DNA sequences only tells a part of the story and the complete description of the underlying coding structure will not be achieved. For this, it is vital to look jointly into both the digital and analog information types encoded in the nucleotide sequence.

It is asserted that the relative stability of the DNA duplex structure relies on its base sequence [Bre+86][SJH04]. Stacking between adjacent base pairs and pairing between complimentary bases determine the thermodynamic stability of the DNA [Pro+04][Yak+06]. Since the stability of the DNA appears as a decisive factor in most of the biological processes, and due to the availability of thermodynamic parameters to describe DNA stability, such as Santalucia's unified nearest-neighbor (NN) thermodynamic stability parameters (free energies) of Watson-Crick base pairs in 1 M NaCl [San98], in this work, analog information will be measured in relation to thermodynamic stability.

We base our analyses and observations on four selected bacterial genomes, namely *Escherichia coli K12 MG1655* (accession NC_000913), *Bacillus subtilis subsp. subtilis str. 168* (accession NC_000964), *Salmonella enterica subsp. enterica serovar Typhimurium DT104* (accession NC_022569), and *Streptomyces coelicolor A3(2)* (accession NC_003888). The general goal is to understand the interrelationship between the sequence organization and thermodynamic property of the genomic sequence in the genomes of the four selected bacteria. Sequence data and the corresponding annotations were taken from GenBank genomes (ftp://ftp.ncbi.nih.gov/genomes

/Bacteria/). Shannon's block entropy is used here to measure the digital information, whereas Gibbs' entropy is employed to measure the analog information. Boltzmann probability distribution is used to convert the DNA stacking energies into probabilities for Gibbs entropy computations. To further relate the two forms of information to gene function, we also incorporated in our analyses spatial distributions of the anabolic, catabolic. aerobic, and anaerobic genes. By doing so, we hoped to reveal the connections between analog and digital information types, as well as its possible functional meaning.

4.2 Shannon, Boltzmann, and Gibbs Entropies

First, in our study, the genome sequence is rearranged to start at the origin (OriC) of replication. Then, the entropy of chunks of the DNA sequence is computed by scanning the complete genome with a sliding window. To examine the effect of the window size, results are shown for window sizes of 100 kb, 250 kb, and 500 kb. Within a window, all possible words of the given block size (N) are counted. To account for all adjacent base interactions, neighboring base pairs are considered. That is, if the nucleotide sequence is "AGCTAG" and the block size is 3 base pairs (bp), AGC, GCT, CTA, and TAG are counted. In this section, the methodology is presented for a block size of three ($N = 3$), other block sizes are handled likewise. The Shannon entropy quantifies the average information content of the genomic sequence from the distribution of symbols (words) of the source [CT91]. It is mathematically given as

$$H_N = -\sum_i P_s^{(N)}(i) \log_2 P_s^{(N)}(i) \,, \tag{4.1}$$

where $P_s^{(N)}(i)$ is the probability (relative frequency) to observe the i^{th} word of block size N inside the window and the summation is over all possible nucleotide words of length N. Essentially, if we take a block size of 3 bp (i.e. codons), the sum will range up to 64. We count the frequency of every triplet in the window and normalize it to the total number of codons. As described in Chapter 2, the Shannon entropy is maximum when all words occur with equal probabilities, and it is zero when one of the symbols occurs with probability one.

Ledwig Boltzmann was the first to give a statistical explanation of the physical (thermodynamic) entropy by relating it to the number of possible arrangements of molecules (microstates) belonging to a macrostate [BS99]. The celebrated formula reads

$$S_B = k_B \ln \Omega \,. \tag{4.2}$$

k_B is the Boltzmann constant which gives this entropy a thermodynamic unit of measure, $k_B = 1.38 \times 10^{-23}$J/K, and Ω is the number of accessible microstates. Boltzmann's entropy is defined for a system based on a microcanonical ensemble in which the macrostate is of a fixed number of particles, volume, and energy. All states are accessed equally likely with the same energy [Rei85]. Gibbs devised a generic entropy definition over a more general probability distribution of the possible states (canonical ensemble). The Gibbs entropy is defined as

$$S_G = -k_B \sum_i P_G(i) \ln P_G(i) \ , \tag{4.3}$$

where the sum is over all microstates and $P_G(i)$ is the probability that the molecule is in the i^{th} state. It can easily be seen that for a uniform distribution of states, the Gibbs entropy reduces to the Boltzmann entropy.

Gibbs' entropy has a similar form as Shannon's entropy except for the Boltzmann constant. Nevertheless, unlike the Shannon case where the probability $P_s^{(N)}$ is defined according to the frequency of occurrence, we associated the probability distribution with thermodynamic stability quantified by the nearest-neighbor free energy parameters. We used Sanatluca's unified free energy parameters for di-nucleotide steps at $37°C$ as in [San98], presented here in Table 4.1. For block sizes greater than two, the energies are computed by adding the involved di-nucleotides. For instance, if the block size is three and the sequence is ATTGC, the energies of AT, TT, TG, and GC will be added. This way, we have a list of codons with their corresponding energies, providing 64 energy states denoted by $E(i)$[1]. Assuming a random process behind the construction of the DNA, with a certain probability, one would obtain molecules with certain energies. If there are n_i codons in the i^{th} energy state, we assumed that the probability for having a certain energy state follows the Boltzmann distribution given by

$$P_G(i) = \frac{n_i e^{-\frac{E(i)}{k_B T}}}{\sum_j n_j e^{-\frac{E(j)}{k_B T}}} \ . \tag{4.4}$$

T is the temperature in Kelvin. Although we are aware that the Boltzmann distribution gives the most probable distribution of energy (the one pertaining to the equilibrium state) for states having a random distribution of energies (e.g. ideal gas), which is not the case here, we just used it to have a representation of stability (energy) in an entropy-like expression.

To see how the Gibbs entropy captures the stability, we generated a random nucleotide sequence of length 100 kb with a specific GC content. By changing the GC

[1]It should be noted that a state here is loosely defined, just for relating the energies to probabilities.

Tab. 4.1: The thermodynamic stability parameters of Watson-Crick base pairs in 1 M NaCl at $37°C$ [San98]

Sequence	kcal/mol
AA/TT	-1.00
AT/TA	-0.88
TA/AT	-0.58
CA/GT	-1.45
GT/CA	-1.44
CT/GA	-1.28
GA/CT	-1.30
CG/GC	-2.17
GC/CG	-2.24
GG/CC	-1.84

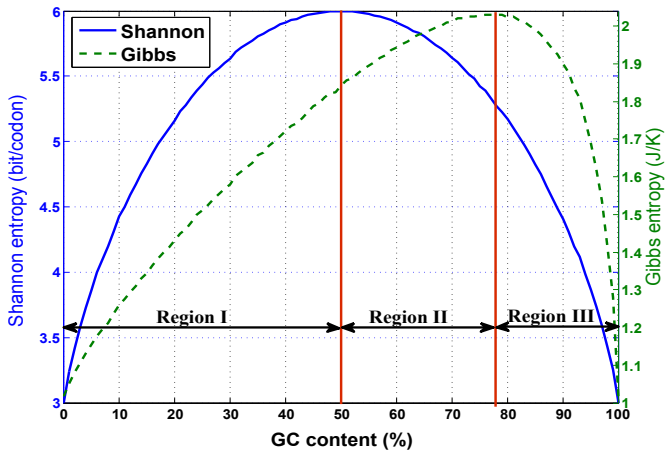

Fig. 4.1: Shannon and Gibbs entropies as a function of GC content.

content from 0 % to 100 %, the Shannon and Gibbs entropies are calculated from the frequency distribution of the codons in the generated sequence. The result is shown in Fig. 4.1. The Shannon entropy function is symmetric with the maximum at 50 %. It tells us how random the sequence is. By comparing it with the maximum value, we can tell how diverse the sequence is, but it does not distinguish between AT and GC. However, except for larger GC content values (in Region III), the Gibbs entropy curve is uniformly related to the GC content. If we are operating in regions I and II (the GC content of organisms typically cannot be greater than 80 %), the higher the Gibbs entropy, the higher the GC content and hence it measures stability. One has to be careful about the maximum point of the Gibbs entropy. The indicated maximum point in Fig. 4.1 is only valid for this randomly generated sample. For other realistic genome sequences the maximum might move elsewhere.

4.3 Shannon vs. Gibbs Entropy Applied on Complete Genomes

The aim is to compare the analog information, quantifying relative stability and measured with the Gibbs entropy (applying Boltzmann statistics to convert the stacking or melting energies to probabilities), with the digital Shannon information. To do so, the block size was set to 3 bp and a sliding window was shifted 4 kb at a time along the complete genome starting from the origin of replication (OriC) as the center of the first window. The Shannon entropy is calculated using overlapping codons (i.e. with a shift of 1 bp).

To support our qualitative statements of comparisons, localized Pearson correlation coefficients are incorporated in the figures. The local cross-correlation coefficients are calculated by taking 100 points to the left and right of the corresponding position. Pearson's correlation coefficient between two vectors x and y is calculated as

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}\,, \tag{4.5}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of x and y, respectively. The correlation coefficient is between $+1$ and -1. If $r_{xy} = 0$, x and y are uncorrelated. The sign of the correlation coefficient indicated the nature of the correlation, positive or negative.

The Shannon and Gibbs entropies in the *E. coli* genome are plotted for window sizes of 100 kb, 250 kb, and 500 kb in figures 4.2-4.4. Since the nucleotide sequence is rearranged to start at the origin, the terminus region (Ter) will be exactly in the middle. This is also evidently visible from the shape of Gibbs entropy curve in which the lowest point is around the terminus, attributed to the AT-richness. Smaller windows lead to high fluctuations and are not easy to compare. Likewise, a very large window will hinder the visibility of the differences as a result of the smoothing effect it creates. Additional results for block sizes 2 and 5 base pairs and a window size of 250 kb are shown in figures 4.5 and 4.6, respectively. For 2 base pairs, the entropies are anti-correlated in all regions. The change to 3 base pairs (Fig. 4.3) has caused three regions to have a positive correlation and these regions remain correlated in this way for higher block sizes. In addition, the vicinity of the terminus region has shown an extremely high anti-correlation. Similarly, there is no significant change when moving to 4 or 5 base pair blocks. It is very significant that the overall

shape of the curves as well as the positions of the troughs and crests remained unaffected by changes in both the block and window sizes.

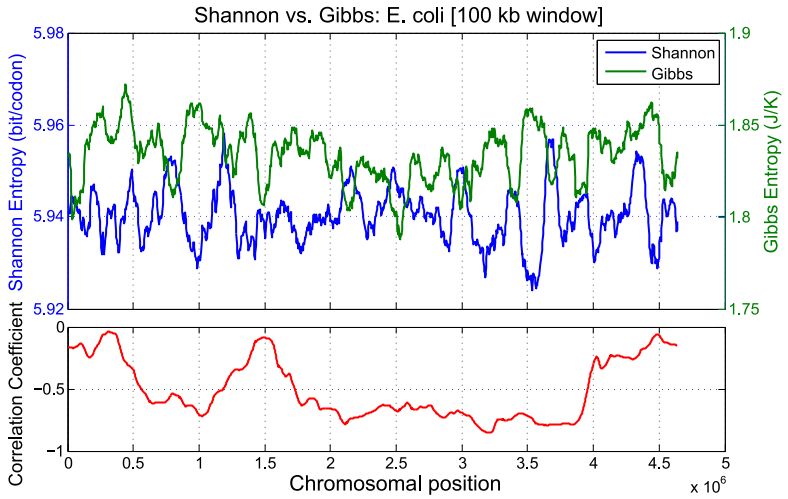

Fig. 4.2: Shannon and Gibbs entropy profiles of *E. coli* for 100 kb sliding windows. The start and end positions are the origin of replication (oriC) whereas the terminus (Ter) is in the middle.

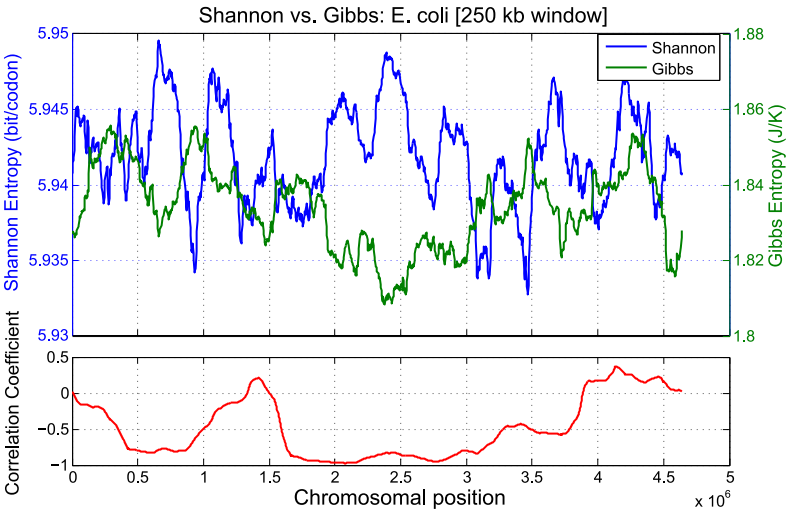

Fig. 4.3: Shannon and Gibbs entropy profiles of *E. coli* for 250 kb sliding windows.

The changes in entropy along the genome might seem very small. For example, in Fig. 4.3, the Shannon entropy in the *E. coli* genome (250 kb window) ranges from 5.9327 and 5.9494, which is a change of only 0.0167. To assess how significant the observed changes ($\Delta SE_{\text{observed}}$) are compared with the changes in entropy (ΔSE) in random sequences, we have calculated the Z-score. However, the random model has to be selected in such a way that it preserves the biological sequence complexity as much as possible. Otherwise, any order present in the real genome will be lost and the resulting Shannon entropies will just be the maximum. We have fragmented the genome into genes and intergenic regions and produced 1000 random "genomes" by shuffling the positions of the fragments. For each random genome, the ΔSE is calculated and the distribution of shuffled ΔSEs is obtained. Finally, the Z-score of

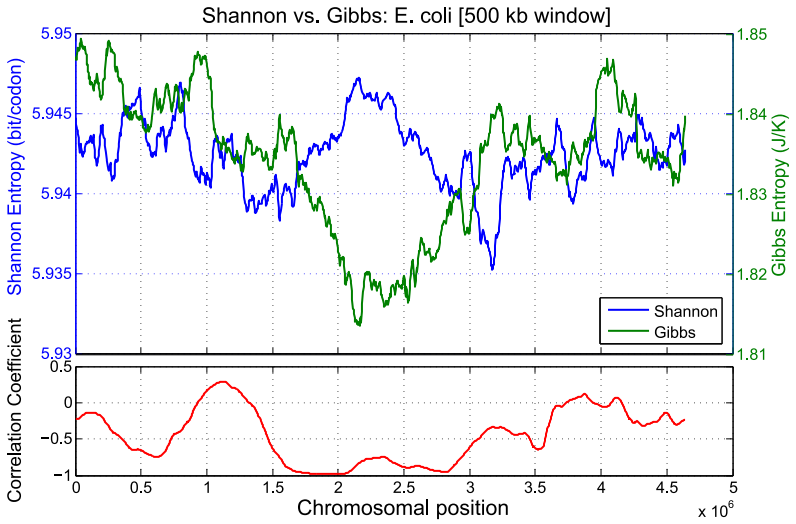

Fig. 4.4: Shannon and Gibbs entropy profiles of *E. coli* for 500 kb sliding windows.

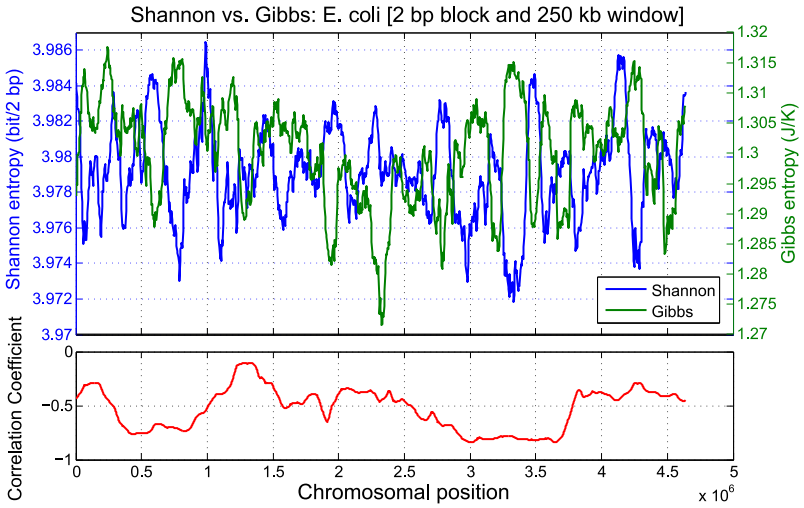

Fig. 4.5: Shannon and Gibbs entropy profiles of *E. coli* for 2 bp block size.

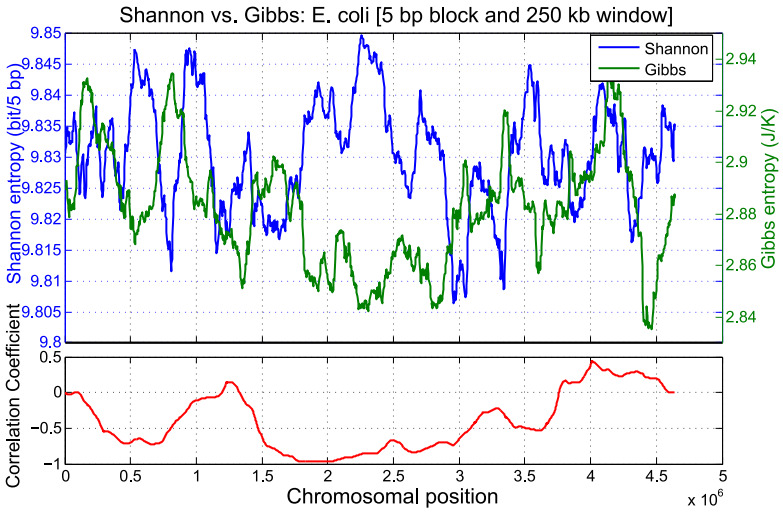

Fig. 4.6: Shannon and Gibbs entropy profiles of *E. coli* for 5 bp block size.

$\Delta SE_{\text{observed}}$ is obtained as $(\Delta SE_{\text{observed}} - \text{Mean}(\Delta SE)/\text{Std}(\Delta SE))$. For $\Delta SE_{\text{observed}}$ = 0.0167 (Fig. 4.3) the Z-score is 3.74 and none of the randomized genomes have exceeded the $\Delta SE_{observed}$. This shows that the observed changes in entropies, even though very small, are highly significant and can safely be used to show differences in certain parts of the genome.

The two entropies are mostly anti-correlated in *E. coli*, with a stronger magnitude around the terminus. The terminus region is characterized by high Shannon entropy and low Gibbs entropy, that is, the sequence is more random and less stable. This means that the codon composition of the sequence has become slightly more balanced, which is due to an increase in AT-rich codons. Similarly, there are also positions where the Shannon entropy is relatively low and the Gibbs entropy is higher (e.g. around position 0.8 Mbp) which means a codon bias towards being more GC-rich. In general, our interpretation for a block size of 3 bp is that whenever both entropies increase, this means that both the GC content and the randomness have increased, and the sequence is more stable due to the usage of more GC-rich codons. However, if there is a decrease in the Gibbs entropy while the Shannon entropy is higher, the sequence has become less stable (AT-rich) and more random as a result of an increase in usage of AT-rich codons.

The Shannon and Gibbs entropy profiles for *B. subtilis* and *S. typhimurium* for a window size of 500 kb are shown in figures 4.7 and 4.8, respectively. Since *S. typhimurium* and *E. coli* are close relatives in phylogeny, the Gibbs and Shannon entropy profiles in *S. typhimurium* show a behavior very similar to *E. coli* being mostly anti-correlated. In contrast, in the evolutionarily more distant gram-positive bacterium *B. subtilis* the two entropies are highly positively correlated.

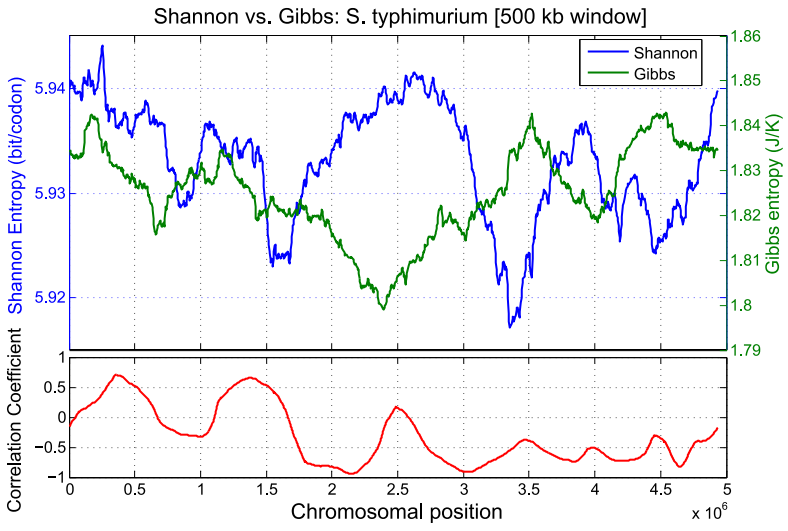

Fig. 4.7: Shannon and Gibbs entropy profiles of *S. typhimurium*.

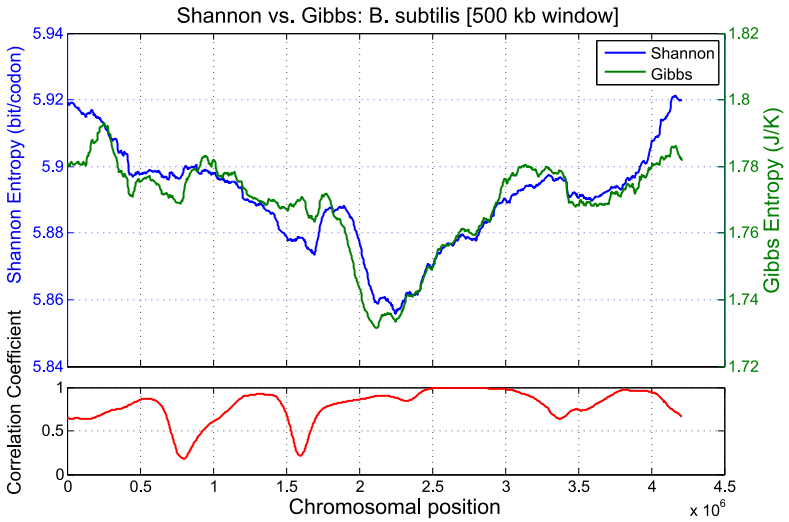

Fig. 4.8: Shannon and Gibbs entropy profiles of *B. subtilis*.

The relationship between the two entropies mostly depends on the GC content of the organism. This can be seen from Fig. 4.1. If the GC content is less than 50 %, there will be a direct relationship between the two. An increase in the number of AT-rich codons will reduce the Gibbs entropy (stability) and at the same time the Shannon entropy will decrease because of the skewed codon distribution. If Shannon entropy increases as a result of having more GC-rich codons, the Gibbs entropy will also increase. For organisms having a slightly more than 50 % average GC content, the entropies will have opposite behaviors. Most of the sequence will be slightly GC-rich and a further increase in GC content would mean an increase of Gibbs entropy. At the same time the Shannon entropy will decrease as a result of the decrease in the variability of the sequence. *E. coli* and *S. typhimurium* have an average GC content of 51 % and 53 %, respectively. Hence, for most regions anti-correlation is observed. However, since the GC contents are in the vicinity of 50 % and locally it can be less than 50 %, the entropies may become positively correlated in some regions. For *B. subtilis* however, the average GC content is 43.5 % and as a result the two entropies are entirely correlated with a global correlation coefficient of 0.9. In the region from the maximum point of Gibbs entropy to 100 % GC (Region III in Fig. 4.1), as the stability (GC content) increases the Gibbs entropy decreases. Therefore, the Gibbs entropy will not be in the same direction as the thermodynamic stability. The plot in Fig. 4.1 is done considering the codon distribution of a randomly generated sample sequence. However, for sequences containing mixtures of AT and GC, the maximum can be anywhere on the right hand side. Therefore, when applying the Gibbs entropy measure on highly GC rich genomes, one can end up in the last operating region where the Shannon and Gibbs entropies follow the same directions. This effect can be seen in Fig. 4.9 where the Shannon entropy profile of *S. coelicolor*, a highly GC-rich linear genome (average GC content is 72.12 %), is plotted with both the Gibbs entropy and the local GC profiles.

The origin is located in the middle of the linear *S. coelicolor* but to be consistent with the plots of the other bacteria, the data is rearranged to have an orientation of OriC-Ter-OriC, although the actual genome is not arranged as a ring. The increase in the GC content makes the sequence more stable. Accordingly, both the Shannon and Gibbs entropies will decrease. Hence, one should mirror the Gibbs entropy to use it as a stability measure. From Fig. 4.9B, it can be seen that the Shannon entropy is perfectly anti-correlated with the GC content, and hence the stability.

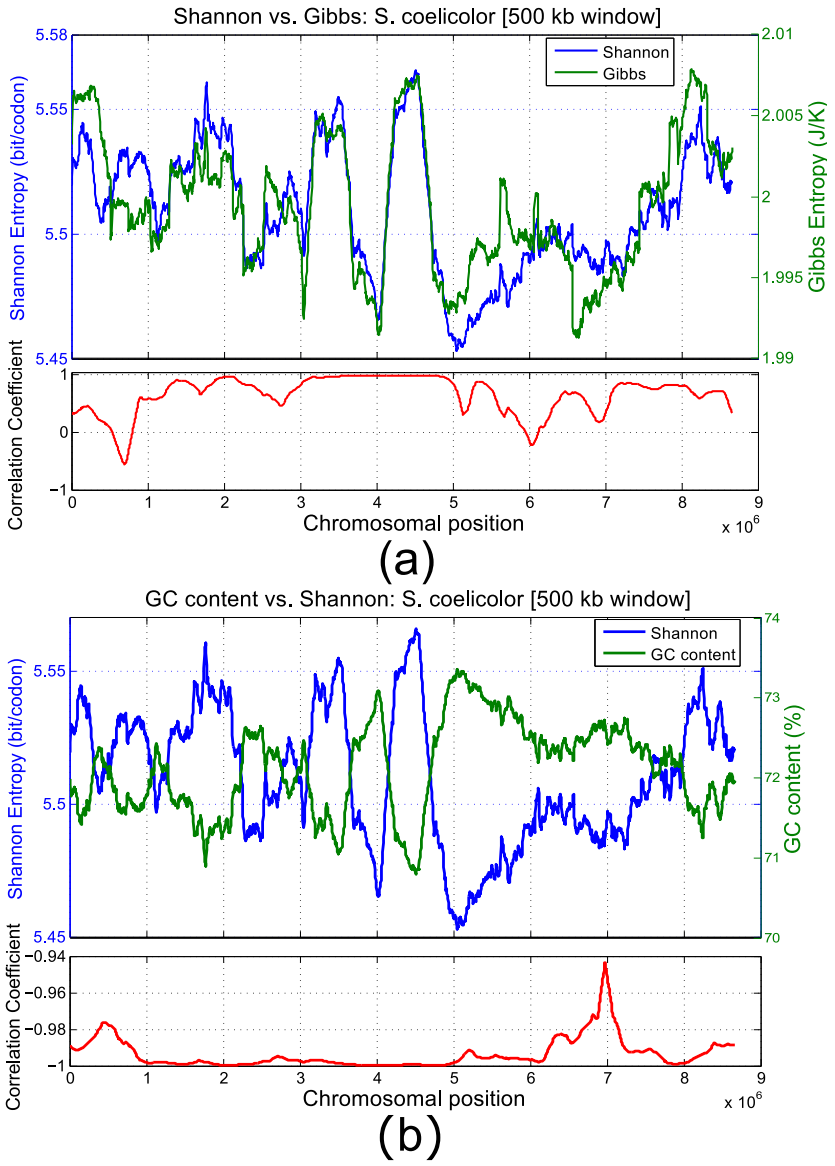

Fig. 4.9: Shannon entropy, Gibbs entropy, and GC profiles of *S. coelicolor*. The linear genome is rearranged in the oriC-Ter-oriC orientation.

The spatial DNA sequence organization can also be observed from the Gibbs entropy profiles. The decreasing gradient from the origin of replication to the terminus shows that most stable DNA is encoded near the origin and less stable at the terminus. This pattern is consistent with the gradients of gene order, DNA melting energy, and distribution of DNA binding sites for DNA gyrase, an enzyme introducing

negative supercoils into the DNA. A highly conserved pattern was observed in α- and γ-Proteobacterial genomes is the gene order along the OriC-Ter axis [Sob+12; Sob+13; TM13]. The anabolic genes that are highly expressed during exponential growth are located in the vicinity of the origin of replication, whereas catabolic genes are predominantly located close to the terminus. In γ-Protebacterial genomes [Sob+13] [TM15], regions close to the origin have a high average melting energy while the regions around the terminus have a low average melting energy. There is also a high concentration of DNA gyrase binding sites in the vicinity of the origin of replication which creates a gradient of average negative superhelicity decreasing towards the terminus, in both replichores [Wan02; Jeo+04; Sob+12].

4.4 Shannon Entropy in the Protein Coding Sequences

So far, the Shannon entropy is computed considering overlapping triplets in the complete genomes. We now only take the protein coding sequences (CDS) of the four genomes and compute the Shannon entropy using both the distribution of non-overlapping triplets (codons) and the corresponding translated amino acid distribution. In a given window, the protein coding genes in both strands are collected and the frequencies of the codons are counted. The base sequences of genes in the complementary strand are complimented and reversed before the counting so that the computed Shannon entropies reflect the actual codon and amino acid composition encoded in the region.

The codon to amino acid translation is carried out using the standard genetic code. The results are shown in Fig. 4.10. Almost for all bacteria, the entropy profiles per codon and per amino acid positively correlate. However, there are regions where the two are negatively correlated (e.g. Fig. 4.10b & d around positions 2.4 Mb and 6 Mb, respectively). The positive correlation can trivially be explained as a direct linear mapping between codons and amino acids. There is a certain level of expected positive correlation between the two profiles. However, since the number of codons encoding a similar amino acid (synonymous codons) varies (ranging from 1 to 6), a change in the frequency distribution of codons may not necessarily affect the amino acid distribution. In *E. coli* and *S. typhimurium*, a high Shannon entropy in the Ter-proximal region reflects the relatively more random nature of the codon and amino acid composition. Except for *S. coelicolor*, the terminus region has the highest amino acid entropy which means that the amino acid distribution in the Ter region is more balanced.

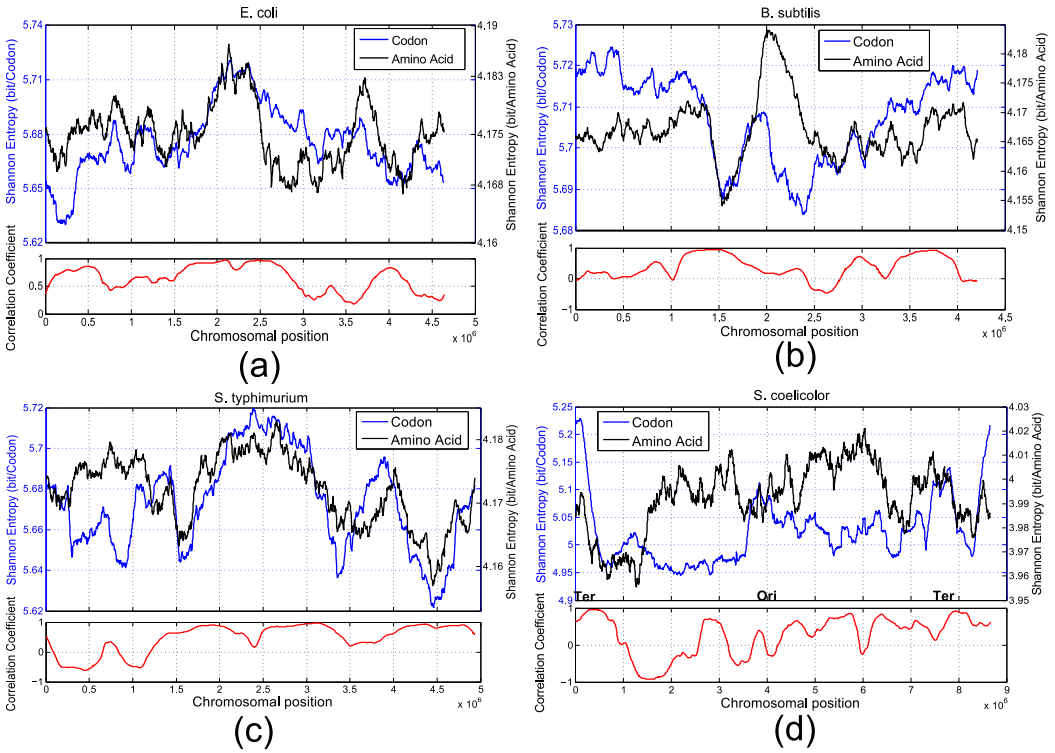

Fig. 4.10: Shannon entropy profiles per codon and amino acid in the coding sequences of the four bacteria. *E. coli* (a), *B. subtilis* (b), *S. typhimurium* (c), and *S. coelicolor* (d). Window size is 500 kb.

The regulatory sequence organization requirement of having an AT-rich terminus region and GC-rich origin is achieved by the selective usage of either synonymous codons or amino acids [Mus15]. For example, the amino acid serine is encoded by AGT, TCA, TCT, AGC, TCC, and TCG. The first three codons are AT-rich whereas the last three are GC-rich. Similarly, the amino acids could also be classified as AT and GC-rich. Amino acids such as proline, encoded by CCT, CCC, CCA, and CCG, can be regarded as a GC-rich amino acid. Likewise, lysine which is encoded by AAA and AAG could be regarded as an AT-rich amino acid. A less stable sequence around the terminus can be attained by using more AT-rich amino acids, which will in turn affect the distribution of amino acids (it will be biased towards the AT-rich ones) or the AT-rich codons among the synonymous ones without affecting the amino acid composition. In *E. coli* and *S. typhimurium*, the high Shannon entropy of codons and amino acids at the terminus (Fig. 4.10) indicates the more uniform codon as well as amino acid distributions. Thus, it appears that the less stable nature of the DNA in this region can be tolerated by allowing the synonymous codon usage. To reveal this selective codon usage, we counted the frequencies of the synonymous codons within two 500 kb windows, one located at the origin and another at the terminus. Here, only non-overlapping triplets (codons) in the coding sequence were considered. Figure 4.11 (a and b) shows the synonymous codon usage in *E. coli* for amino acids serine and leucine. Note that in the Ter region the frequency of the

AT-rich codons have increased whereas the GT-rich ones have decreased. Although leucine is most often encoded by CTG, since it is a GC-rich triplet, the number of CTG codons has decreased considerably. This observation is pertinent also to the other amino acids. The terminus region of *B. subtilis* is also less stable and has the highest Shannon entropy of amino acids. Although the Shannon entropy of codons in the Ter region is not higher than around the origin, the selective usage of codons still occurs. As shown in Fig. 4.11, compared to the origin of replication, the frequency of AT-rich codons has increased in the terminus region. It is noteworthy that the low GC content of the organism by itself favors the use of AT-rich codons. For encoding serine and leucine, *B. subtilis* uses almost twice as many AT-rich codons as GC-rich ones (see Fig. 4.11c & d). This explains the observed low Shannon entropy of codons at the terminus region shown in Fig. 4.10b.

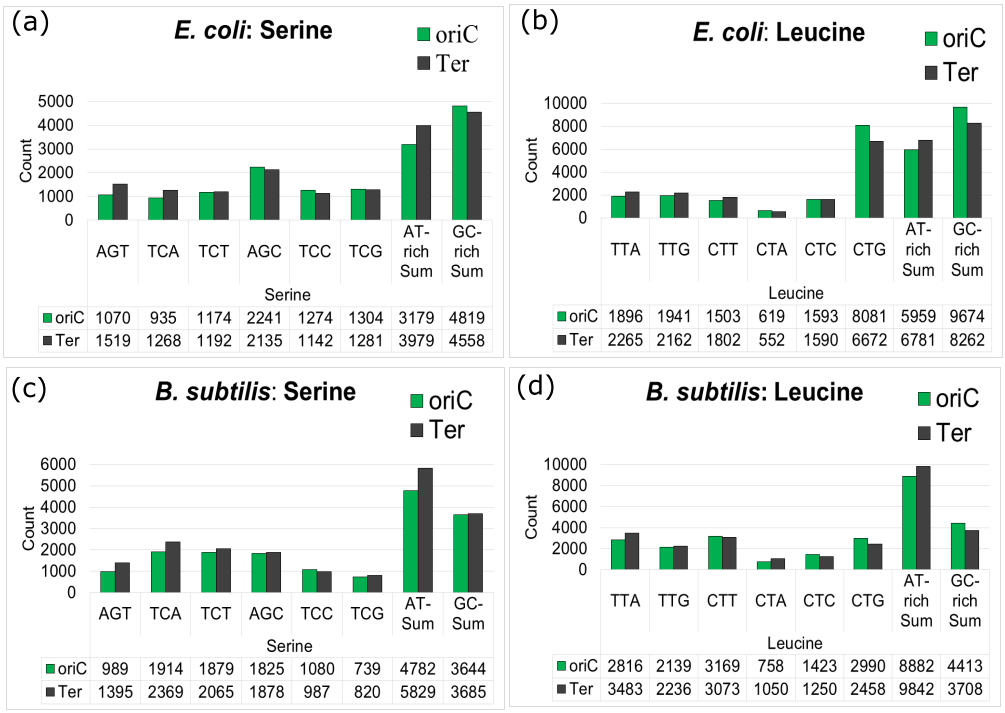

Fig. 4.11: Synonymous codon usage in *E. coli* (A and B) and *B. subtilis* (C and D) at origin and terminus regions. AT-rich sum and GC-rich sum are the total number of AT and GC-rich codons, respectively.

4.5 Sequence Organization in Relation to Gene Function

We have shown that the sequence organization is mainly dependent on the physicochemical property requirements to serve certain functions. For example, the less stable and hence AT-rich terminus region is assumed to absorb the positive superhelicity generated by the convergence of the two replisomes during replication

[TM13]. We have also analyzed the spatial sequence organization in relation to other functional requirements. We chose two functional classes of genes - anabolic and catabolic genes - connected to energy and resource supply of the cell.

4.5.1 Functional Classes of Genes

To further associate the digital and analog information with protein function. We considered functional classes of genes which are related to the nature of the chemical reaction in the living cell and oxygen usage (cellular respiration). If the chemical reaction requires energy, it is called anabolic and the genes involved in this type of reaction are called anabolic genes. Accordingly, if the reaction releases energy, it is called catabolic and the corresponding genes are named catabolic genes. Catabolic reactions produce energy by breaking down complex compounds to smaller molecules in stages of energy and resource shortages [Tor+04]. Conversely, the anabolic reactions use energy to maintain life by building complex compounds from simpler ones. The other functional classes of genes are related to the mode of cellular respiration. Aerobic respiration requires oxygen whereas anaerobic respiration does not use oxygen [BP10]. Hence, the genes expressed in the presence and absences of oxygen are termed as aerobic and anaerobic genes, respectively.

The functional gene groups were taken from the Gene Ontology (GO) tree provided by the RegulonDB database. Anabolic genes: biosynthesis of macromolecules (GID000000120); catabolic genes: degradation of macromolecules (GID000000057); aerobic genes: aerobic respiration (GID000000068); anaerobic genes: anaerobic respiration (GID000000069). To have a possibility of comparison between the bacteria, the orthologues of anabolic and catabolic genes were considered. The corresponding functional groups where counted in 500 kb sliding windows using a 4 kb shift. The window size was chosen so as to have a significant number of genes and obtain a smooth curve.

The distribution of anabolic and catabolic genes of *E. coli* are plotted along with Gibbs entropy (thermodynamic stability) in Fig. 4.12. We used a 500 kb window and counted the number of genes of the corresponding functional group and normalized it to the total number of genes in the window. The gene frequencies are further normalized to the range 0 and 1 to plot them on a similar scale. Interestingly, the distribution of anabolic and catabolic genes are strongly related to the Gibbs entropy. Anabolic genes and Gibbs entropy are highly correlated (note the similarity in the profiles and also the magnitude of the correlation coefficient in Fig. 4.12). It seems that catabolic genes are encoded by sequences of low thermodynamic stability, while anabolic genes are favorably encoded by DNA sequences of high thermodynamic stability. Highly stable DNA sequences require an extra input energy, for instance,

to open up the DNA strands for transcription, and hence, the anabolic genes are activated during the fast growth in rich medium. In such a way, energy consuming functions and energy availability are coupled [Sob+13]. In addition, anabolic and catabolic genes show an opposite chromosomal distribution pattern reflecting their mutually exclusive roles in bacterial metabolism. There are two symmetric regions flanking the origin of replication (0.5 Mb and 4.2 Mb) that show a deviation from the general pattern of a decreasing trend of anabolic genes towards the terminus. These regions are known to harbor highly transcribed stable RNA (rRNA) genes. The transcription dynamics of stable RNA operons form large DNA structures called transcription foci [Ber+10]. Most likely, the optimal thermodynamic coding of anabolic genes is affected by the presence of these structures. The regions are relatively enriched with catabolic genes, preserving the opposite genomic distribution of anabolic and catabolic genes.

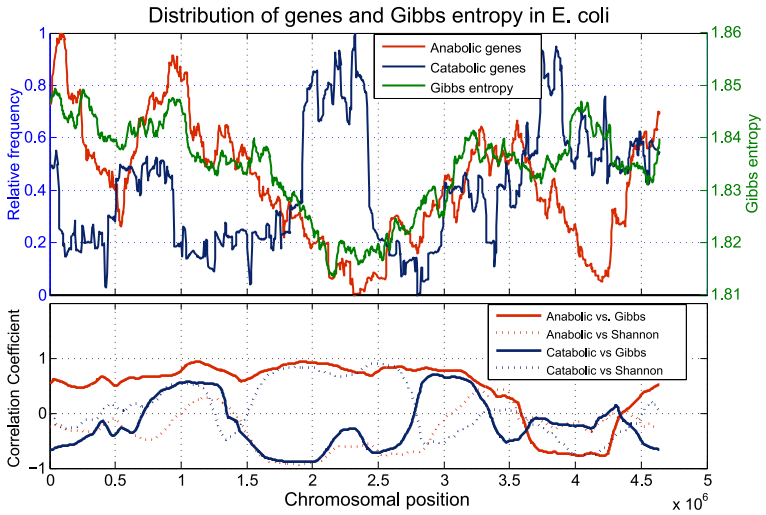

Fig. 4.12: Distribution of anabolic and catabolic genes along with Gibbs entropy in *E. coli*. The correlations with the entropies are also shown. The number of the genes are normalized to the total number of genes within the 500 kb window.

The distribution of the aerobic genes (Fig. 4.13) has also similar increasing and decreasing patterns as the Gibbs entropy, except for the quantization effects resulting from a very low number of genes. The dependency of the anaerobic genes with the entropies, however, is not so uniform and obvious to see as for anabolic and aerobic genes. However, the relationships in distinct regions of the chromosome further show how function directs spatial organization.

The genomic distribution of the orthologues of anabolic and catabolic genes in *B. subtilis* and *S. typhimurium* are presented in figures 4.14 and 4.15, respectively. In *B. subtilis*, at the terminus region, both anabolic and catabolic genes anti-correlate with both the Gibbs and the Shannon entropies. The right replichore shows a very high correlation between the entropies and the functional classes of genes. At the terminus, although the sequence is less stable, a high number of anabolic and

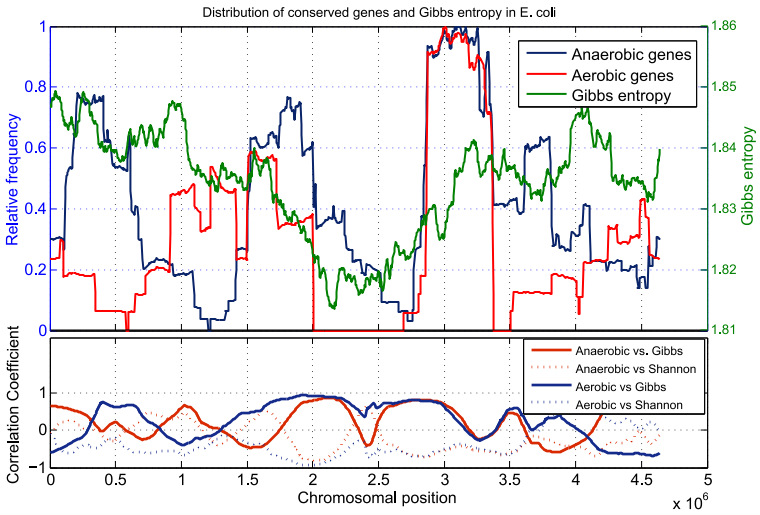

Fig. 4.13: Distribution of aerobic and anaerobic genes along with Gibbs entropy in *E. coli*. The correlations with the entropies are also shown. The number of the genes are normalized to the total number of genes within the 500 kb window.

catabolic genes were observed, which is not consistent with the results obtained in *E. coli*. However, since *B. subtilis* and *E. coli* have different life stiles (e.g. occurrence of the process of septation in the former) and diverged about one billion years ago, substantial differences in genome organization are to be expected. The high correlation of Gibbs entropy and anabolic genes in *E. coli* supports the view that the genomic sequence organization is largely determined by the process of replication [TM13]. However, *B. subtilis* is known for its property of sporulation, which imposes constraints on the organization of the genome and chromosome segregation [Wan+13]. Also, it uses different replication factories and possesses different and much more numerous sigma factors [Kum+15]. Thus, we assume that the observed anti-correlation (Fig. 4.14) is due, at least in part, to these differences. The profiles of anabolic and catabolic genes of *S. typhimurium*, shown in Fig. 4.15, are also mostly anti-correlated with the Gibbs entropy. However, around the terminus region, catabolic genes are anti-correlated with the Gibbs entropy in all of the analyzed bacteria. Although there is no ubiquitous relationship that explains how the functional groups are spatially organized, the obtained plots yield qualitative relations between digital and analog properties of the DNA sequence at specific sites in the chromosomes.

4.6 Gibbs Entropy for Identification of Coding Regions

The Gibbs entropy profiles can further be used as a tool for detecting coding and non-coding regions. Generally, because of the AT-richness of the promoters as well as the

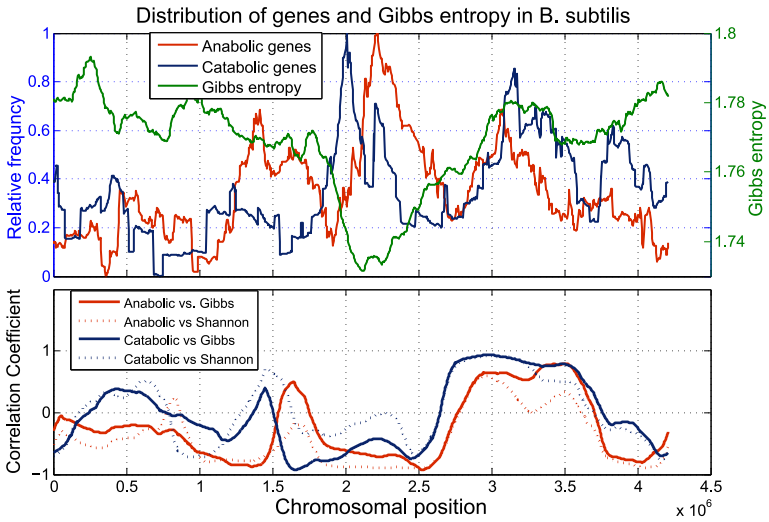

Fig. 4.14: Distribution of anabolic and catabolic genes in *B. subtilis*. The correlations with the entropies are also shown. The number of the genes are normalized to the total number of genes within the 500 kb window.

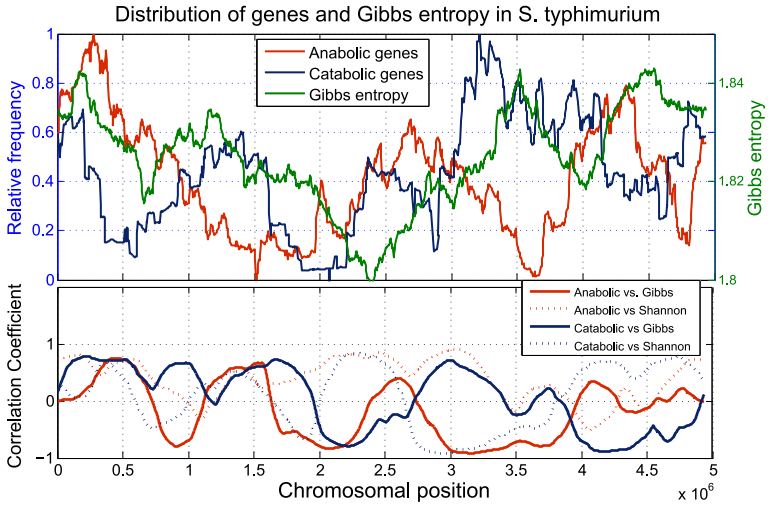

Fig. 4.15: Distribution of anabolic and catabolic genes in *S. typhimurium*. The correlations with the entropies are also shown. The number of the genes are normalized to the total number of genes within the 500 kb window.

$5'$ and $3'$ gene flanking regions, the coding sequences are GC-rich compared to the corresponding non-coding sequences [RB10][TM15]. Since the Boltzman probability distribution gives more weight to AT-rich sequences (see Eq. 4.4), the Gibbs entropy will have smaller values at the non-coding regions. To demonstrate this, we have used a smaller sliding window (400 bp) with a 50 bp shift on a segment of *E. coli* genome containing 12 genes. The results are presented in Fig. 4.16. The minimum points with low thermodynamic stability match with the non-coding regions of the genome (the gaps between genes in the annotation at the top). Stability and melting temperature profiles have been previously used for identification of various genomic regions (e.g. see [Kha+14] and [KB09]). However, our method produces a

significant variation in Gibbs entropy more clearly pointing out the differences in coding and non-coding regions of the genome.

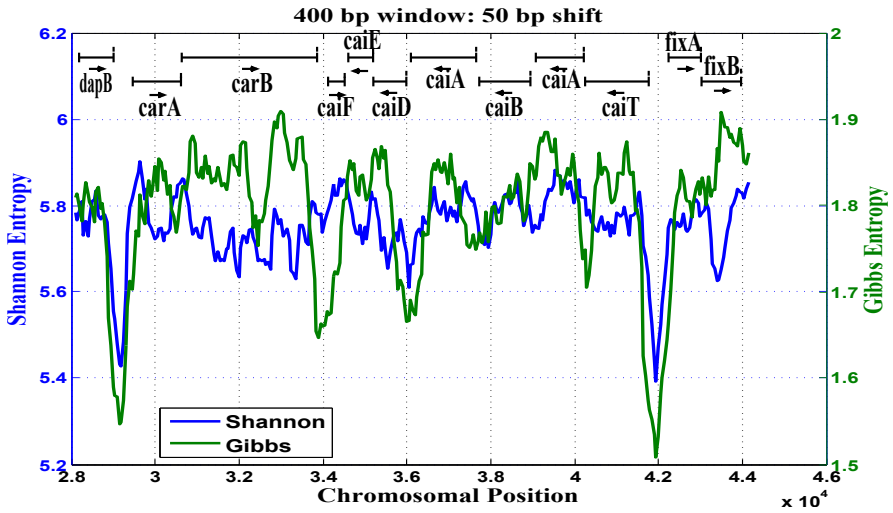

Fig. 4.16: Shannon and Gibbs entropy profiles in a segment of *E. coli* genome. The 12 genes located in the segment are annotated. Note that the troughs of the Gibbs entropy are exactly at the non-coding positions.

4.7 Summary

In addition to the digital type of linear genetic code encoding the proteins, DNA contains a continuous or analog type of information resulting from the physico-chemical properties of the DNA polymer. The analog information depends on the additive interactions of consecutive base steps rather than the individual bases. Hence, integrated analysis of the analog and digital DNA information types not only provides an additional angle to interpreting and understanding the genome sequence organization but also provides a way to integrate and consolidate the structural and functional data. In this study, we analyzed the spatial organization and relationships between the digital and analog properties of the DNA sequence along with the functional classes (anabolic, catabolic, aerobic, anaerobic) of genes in four bacterial species.

In *E. coli*, Shannon and Gibbs entropies are mostly anti-correlated. Especially, the two entropies are almost exactly opposite around the terminus. The results show that the global patterns of the entropies are more or less preserved independent of changing the window and block sizes. The observed gradient of Gibbs entropy from the origin to the terminus in both replichores is partly due to the GC content based selective usage of synonymous codons. The gradient of thermodynamic stability has been previously related to the process of replication and the demand to utilize the anabolic and catabolic genes at different stages of the growth cycle, facilitated by

their location on the opposite chromosomal ends [MT13; Sob+13; TM13; Mus15]. Another core finding is the relation between the genomic distribution of anabolic and catabolic genes and the Gibbs entropy. In *E. coli*, anabolic genes are highly correlated with the Gibbs entropy whereas around the terminus region, catabolic genes are anti-correlated with Gibbs entropy in all analyzed bacteria. The observed patterns are very similar, implying a clear connection between functional gene types and DNA thermodynamic stability and, due to the correlation between entropies, also to statistical properties, i.e., the information content. We have also demonstrated the application of Gibbs entropy for the distinction of coding and non-coding regions based on the differences in DNA thermodynamic stability. While we propose this here, we think that verification of this proposal merits a separate study. The gram-negative enterobacterium *S. typhimurium* is closely related to *E. coli* and therefore, it shows profiles very similar to *E. coli*. However, the AT-rich genome of the gram-positive soil bacterium *B. subtilis* exhibits different properties of organization. In *B. subtilis*, the Shannon and Gibbs entropy profiles are highly correlated. The distributions of the orthologues of anabolic and catabolic genes are also anti-correlated with the Gibbs entropy. *S. coelicolor* is a gram-positive bacterium with a life-stile resembling fungi and containing two large plasmids in addition to the linear genome. The peculiarity of *S. coelicolor* is that the distribution of different types of genes reveals a central core comprising half of the chromosome and containing all the essential genes, whereas genes encoding apparently non-essential functions lie in the arms [Ben+02]. Notably, this biphasic structure of the chromosome does not align with the position of oriC. These peculiarities may affect the relationship between the analog and digital DNA information in organizing the genetic function in the highly GC-rich genome of *S. coelicolor*. Nevertheless, we observed that also in *S. coelicolor* the Shannon entropy is perfectly anti-correlated with the GC content (Fig. 4.9B). Taken together, our data strongly support the notion that the organization of the genetic code in the genome is dictated by thermodynamic and information-theoretic properties of the genomic sequence. Digital and analog DNA information types are tightly intertwined, which on evolutionary timescale can adopt different relationships depending on the type and life stile of a bacterium.

5 Prediction of Essential Genes

The subset of genes which are necessary for the viability and reproduction of an organism are called essential genes (EGs). Detection of these genes is very crucial for understanding the minimal requirements for maintaining life [Koo00; Ita95]. Since the disruption or deletion of EGs results in the death of the organism, EGs of pathogens can be used as potential drug targets [CL02; Lam+03]. Furthermore, studies on EGs are very important in synthetic biology for re-engineering microorganisms and creating cells with a minimal genome [Hut+16].

In this chapter, we present a machine learning based EG predictor using novel information-theoretic features derived exclusively from the DNA sequences of the genes. In Section 5.1, a literature review of existing essential gene prediction methods and the motivation for the study are presented. In Section 5.2, the machine learning algorithms used in this study are briefly described. The sources for the essential/non-essential annotated gene data and the genome sequences of the species are outlined in Section 5.3. In sections 5.4 and 5.5, the features used for prediction are specified. After the classification procedures and performance evaluation methods are specified in Section 5.6, prediction results for essential genes in Bacteria, Archaea, and Eukarya are described in sections 5.7, 5.8, and 5.9, respectively. Finally, the chapter is summarized in Section 5.10.

5.1 Background

Genome-wide systematic or random experimental laboratory procedures such as transposon mutagenesis [Sal+04], single gene knockout [Che+15; Gia+02], and RNA interference [CA05] are used to identify the EGs. Recently, the CRISPR (clustered regularly interspaced short palindromic repeats) gene-editing technology has also been used [Blo+15; Har+15; Wan+15]. Although the experimental methods are fairly accurate, they are often time-consuming and expensive. Moreover, gene essentiality results of the experimental methods may depend on growth conditions [D'E+09]. To bypass these constraints, various computational prediction methods have been proposed. The earliest computational methods were based on comparative genomics in which gene essentiality annotations are transferred among species through homology mappings [MK96; Zha+16]. However, homology mappings have

the limitation that mappings are only between conserved orthologs and often these conserved genes are non-essential. For instance, *E. coli* and *A. baylyi* have 1198 orthologs, which is 36 % of the gene set in *A. baylyi*, and only 195 genes are common EGs (Fig. 5.1). More recently, when lists of genes for model organisms became available in public databases (such as DEG [Luo+14], CEG [Ye+13], and OGEE [Che+12]), researchers have studied the characteristics and features of EGs and deployed machine-learning based prediction methods.

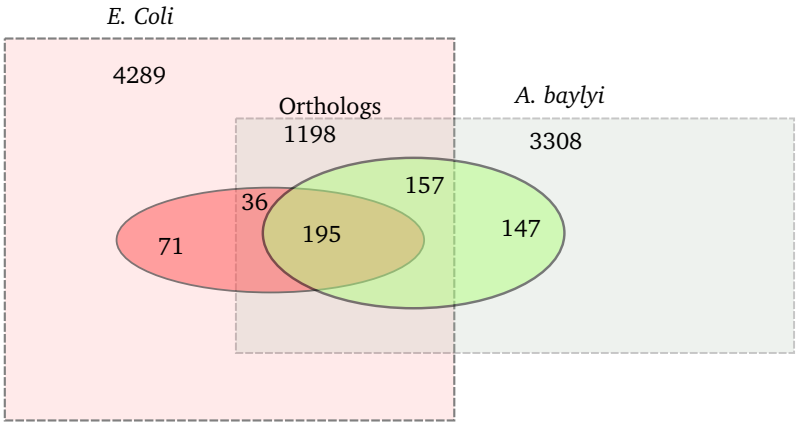

Fig. 5.1: Comparison of conserved orthologs and shared essential genes between *E. coli* and *A. baylyi*. Modified from [Den+11].

A wide range of features has been associated with gene essentiality. The features can be broadly categorized into sequence information (e.g. GC content, protein length, and codon composition) [Nin+14; Son+14; Yu+17], network topology (e.g. degree centrality and clustering coefficient) [Pla+10; AL09; Lu+14; Che+14], homology (e.g. number of paralogs)[Wei+13; Che+13; Son+14], gene expression (e.g. mRNA expression level and fluctuations in gene-expression) [Den+11; Che+14], cellular localization (e.g. cytoplasmic score and outer membrane score)[Che+14; PM11; Liu+17], functional domain (e.g. domain enrichment)[Den+11], and physicochemical property (e.g. molecular weight and number of moles of amino acids) [PM11; Liu+17].

Except for the sequence-based and sequence-derived features, which can be obtained directly from the DNA or protein sequences, the others require pre-computed experimental data. Network topology based features require the availability or construction of protein-protein interaction, gene regulatory networks, or metabolic networks. Similarly, the gene expression and functional domain features demand the expression data and a search in protein domain databases such as PROSITE and PFAM. Although experimental and genetic network information is available for the well-studied species, they are not available for all organisms, especially not for the newly sequenced and under-studied. Hence, predictors relying only on sequence information are of special importance.

A number of researchers have proposed sequence-based essential gene predictors [Nin+14; Son+14; Li+17; Liu+17; PM11; Yu+17; Wei+13; Guo+17]. Ning et al. [Nin+14] used nucleotide, di-nucleotide, codon, and amino acid frequencies along with what is known as CodonW features. The CodonW features, which are sequence-derived, are obtained from a codon usage analysis software (http://codonw.sourceforge.net). However, some of the CodonW features are not purely obtainable from the DNA or protein sequence. For instance, the Codon Adaptation Index (CAI) is a measure of the relative adaptability of the codon usage of a gene compared to the codon usage of highly expressed genes [SL87]. That means, one needs to first distinguish the highly expressed genes in the organism. Due to its effectiveness, the CAI feature is used by all sequence-based predictors. Ning et al. performed cross-validation experiments considering 16 bacteria species. The other very effective essential gene predictor based solely on sequence and sequence-derived properties is Song et al.'s ZUPLS [Son+14]. ZUPLS uses features from the so-called Z-curve, sequence-based (e.g. size, CAI, and strand), homology mapping, and domain enrichment scores. Cross-organism results were shown using models trained on *E. coli* and *B. subtilis*. Among the sequence-based methods, ZUPLS seems to be the best method. Palaniappan and Mukherjee [PM11], in 2011, proposed a machine learning based EG predictor using sequence and physico-chemical properties, plus cellular localization information. In addition to predictions of EGs between organisms, they showed results at higher taxonomic levels. In 2017, Liu et al. [Liu+17], using a feature which measures long-range correlation (the Hurst exponent) and similar features to [PM11] made an extensive study on 31 bacteria and presented detailed results. Yu et al. [Yu+17] and Li et al. [Li+17] used a different set of features based on fractals and inter-nucleotide distance sequences. In 2013, a method called Geptop (gene essentiality prediction tool based on orthology and phylogeny) [Wei+13] was proposed and due to the high accuracy and the availability of a Web server, it is the most used computational tool. Geptop identifies orthologs by the reciprocal best hit method and computes evolutionary distance between genomes using the Composition Vector (CV) method [XH09]. Then, an essentiality score is defined and a threshold-based classification is performed.

Other computational methods which use sequence information together with network topology and gene expression include the works of Deng et al. [Den+11] and Cheng et al. [Che+14; Che+13]. Deng et al. [Den+11] used thirteen features. Along with the sequence dependent features such as protein length and number of codons, they used features related to network topology, gene-expression, homology, phylogenetics, and protein domain knowledge. A combination of four machine-learning algorithms (Naive Bayes, logistic regression, C4.5 decision tree, and CN2 rule) were applied. They showed the effective transferability of essentiality annotations among *E. coli, B. subtilis, A. baylyi,* and *P. aeruginosa*. Cheng et al. [Che+14] proposed a novel computational method which is based on Naive Bayes classifier, logistic

regression, and a genetic algorithm. They have used a combination of network topology, gene expression, and sequence-related features and reciprocally predicted EGs among 21 species and obtained excellent results.

Although most essential gene predictions were applied to prokaryotes, there were also computational predictors applied to eukaryotic species. Chen and Xu [CX05], in 2005, proposed a protein dispensability prediction based on rates of evolution, protein-protein interaction connectivity, gene-expression, and gene duplication. They used Neural Networks and Support Vector Machines (SVMs) for classifying genes of *Saccharomyces cerevisiae*. Serginghaus et al. [Ser+06] investigated the predictability of the EGs of the yeast *S. cerevisiae* using 14 features, which are accessible from the genomic sequence data. They used a combination of seven learning algorithms. In addition, EGs of the closely related yeast *S. mikataea* were predicted using a model trained on *S. cerevisiae*. In 2012, Yuan et al. [Yua+12] used 491 features derived from the sequence, gene expression, and protein interaction networks and performed lethality phenotype predictions in mice. Their models produced a very good prediction accuracy. Lloyd et al. [Llo+15] analyzed relationships between lethality and various gene properties including network connectivity, gene copy number, and gene expression levels in *Arabidopsis thaliana*. Using these features and machine learning models, the EGs of *A. thaliana* were predicted and also cross-organism predictions to transfer essentiality annotations to *Oryza sativa* and *S. cerevisiae* were performed. Recently, Guo et al. [Guo+17] showed that using only sequence information, human EGs can be accurately predicted. They employed an SVM algorithm and used the so-called λ- interval Z-curve features [Guo+03], reflecting nucleotide composition and associations.

In the present work, we propose machine learning based prediction using information-theoretic features which rely solely on sequence information. The information-theoretic features are entropy (Shannon and Gibbs), mutual information (MI), conditional mutual information (CMI), Kullback-Leibler divergence (KLD), and Markov model (M) based. These quantities measure the compositional, structural, and organizational properties in the DNA sequences. The entropy computations will highlight the degree of randomness and thermodynamic stability of the genes. In Chapter 4, we have analyzed the application and implication of Shannon and Gibbs entropies in bacterial genomes. MI has been extensively used in various computational biology and bioinformatics applications. For instance, MI profiles were used as genomic signatures to reveal phylogenetic relationships between genomic sequences [Bau+08] and for identification of coding and non-coding DNA [Gro+00], as a metric of phylogenetic profile similarity [DM03], and for identification of single nucleotide polymorphisms (SNPs) [Hag+04]. Hence, MI and CMI features make use of sequence organization and dependencies and capture the differences between essential and non-essential genes. The Markov features are selected for measuring

statistical dependencies. We performed EG predictions in the three domains of life: Bacteria, Archaea, and Eukarya. 15 bacteria, 1 archeaon, and 4 eukaryotes were analyzed. Moreover, with the hope of increasing the prediction performance, other non-information-theoretic features, which can be easily obtained from the genetic sequences and known to have a correlation with gene essentiality, were included. The added features are related to optimized stop codon usage, gene length, and GC content. The predictive power of the five feature sets, both individually and collectively, was assessed.

5.2 Machine Learning Algorithms

Two of the most commonly used and powerful machine learning algorithms were used alternatively for classification, Support Vector Machines (SVMs) and Random Forest. We start by giving an overview of the two algorithms.

5.2.1 Support Vector Machines

Boser et al. introduced SVMs in 1992 [Bos+92]. Due to the high accuracy in handwritten digit recognition (1.1 % test error), SVM became very popular [Bot+94]. Since then, SVMs have been successful in various applications including bioinformatics. In this section, we present SVM for binary classification briefly. A more comprehensive description can be found in [Bur98; Bis06].

Linearly separable training data

Suppose given a training data set, such as the one shown in Fig. 5.2, the SVM classifier finds a hyperplane, among the many possibilities, that has the largest margin possible between the classes. For this reason, SVMs are regarded as a maximum margin classifier. The position of the decision boundary is fully specified by a small subset of the data. This subset of points are referred to as support vectors.

Let $\mathbf{x_i} = \{x_1, x_2, \ldots, x_n\}$, $\mathbf{x_i} \in \mathbb{R}^n$ be the data set and $\mathbf{y_i} \in \{-1, +1\}$ be the class labels of $\mathbf{x_i}$. The task is to find a decision boundary (i.e. a hyperplane) which linearly separates the two classes. The hyperplane is described by the equation

$$\mathbf{w}^T\mathbf{x} + b = 0 \,, \tag{5.1}$$

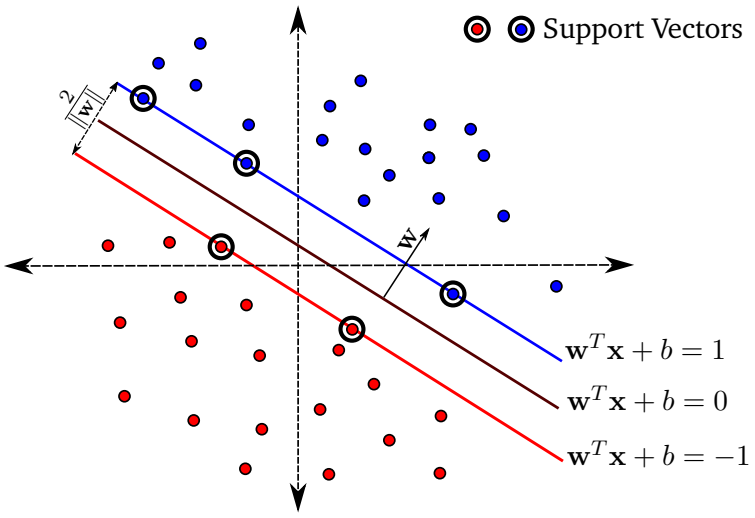

Fig. 5.2: SVM classifier for a linearly separable data

where $\mathbf{w}$ is normal to the hyperplane and $b/\|\mathbf{w}\|$ is the perpendicular distance between the origin and the hyperplane. $\mathbf{w}$ is also known in the machine learning literature as the weight vector. Any point above the hyperplane, i.e ., $\mathbf{w}^T\mathbf{x} + b > 0$ belongs to $y_i = +1$, whereas any point below the hyperplane, i.e., $\mathbf{w}^T\mathbf{x} + b < 0$ corresponds to $y_i = -1$. In a linearly separable training data, two hyperplanes that separate the classes can be selected. Then, the distance between the two parallel hyperplanes can be maximized. Suppose the following constraints are satisfied for every training data set:

$$\begin{aligned} \mathbf{w}^T\mathbf{x} + b \geq +1 \text{ for } y_i = +1 \\ \mathbf{w}^T\mathbf{x} + b \leq -1 \text{ for } y_i = -1 \end{aligned} \tag{5.2}$$

The two inequalities in Eq. (5.2) can be combined as

$$y_i(\mathbf{w}^T\mathbf{x_i} + b) \geq 1 \; \forall i \; . \tag{5.3}$$

The distance between the hyperplanes (the margin) can be shown to be $\frac{2}{\|\mathbf{w}\|}$. Let $\mathbf{x_1}$ be an arbitrary point which lies on $\mathbf{w}^T\mathbf{x} + b = -1$. The closest point to $\mathbf{x_1}$ on the line $\mathbf{w}^T\mathbf{x} + b = +1$ is $\mathbf{x_2} = \mathbf{x_1} + \lambda\mathbf{w}$. $\lambda\mathbf{w}$ is the line segment connecting $\mathbf{x_1}$ and $\mathbf{x_2}$. Thus, $\lambda\|\mathbf{w}\|$ is the distance between the hyperplanes. Solving for λ,

$$\begin{aligned} &\mathbf{w}^T\mathbf{x_2} + b = +1 \; , \\ &\mathbf{w}^T(\mathbf{x_1} + \lambda\mathbf{w}) + b = +1 \; , \\ &\mathbf{w}^T\mathbf{x_1} + b + \lambda\mathbf{w}^T\mathbf{w} = +1 \; , \\ &-1 + \lambda\mathbf{w}^T\mathbf{w} = +1 \; , \\ &\lambda\mathbf{w}^T\mathbf{w} = 2 \; , \\ &\lambda = \frac{2}{\mathbf{w}^T\mathbf{w}} = \frac{2}{\|\mathbf{w}\|^2} \; . \end{aligned} \tag{5.4}$$

The distance between the two hyperplanes, $\lambda\|\mathbf{w}\|$, is, thus, $\frac{2}{\|\mathbf{w}\|^2} \times \|\mathbf{w}\| = \frac{2}{\|\mathbf{w}\|}$.

To obtain the largest margin, we should maximize $\frac{2}{\|\mathbf{w}\|}$. Thus, the decision boundary can be obtained by solving the constrained optimization problem

$$\begin{aligned} \underset{\mathbf{w},b}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{w}\| \\ \text{subject to} \quad & y_i(\mathbf{w}^T\mathbf{x_i} + b) \geq 1 \end{aligned} \tag{5.5}$$

or since square root is a monotonic function, it is equivalent to solving

$$\begin{aligned} \underset{\mathbf{w},b}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{subject to} \quad & y_i(\mathbf{w}^T\mathbf{x_i} + b) \geq 1 \end{aligned} . \tag{5.6}$$

Non-separable training data

If the training data is not perfectly linearly separable, there will not be a feasible solution to the quadratic optimization problem. Thus, to allow misclassification of noisy or difficult points, we need to relax the constraints in Eq. (5.3). To do so, a slack variable $\xi_i \geq 0$ is introduced and the quadratic optimization problem becomes

$$\begin{aligned} \underset{\mathbf{w},b}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i \\ \text{subject to} \quad & y_i(\mathbf{w}^T\mathbf{x_i} + b) \geq 1 - \xi_i,\ \xi_i \geq 0 \end{aligned} . \tag{5.7}$$

The parameter C controls the tradeoff between error and margin. A smaller C corresponds to a lower penalty to errors.

Non-linear classification with kernels

To perform non-linear classification, the data vectors, $\mathbf{x_i}$, are transformed into a high dimensional feature space where the training data is linearly separable. The quadratic optimization problem remains unchanged but all $\mathbf{x_i}$ are replaced by $\phi(\mathbf{x_i})$,

$$\begin{aligned} \underset{\mathbf{w},b}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i \\ \text{subject to} \quad & y_i(\mathbf{w}^T\phi(\mathbf{x_i}) + b) \geq 1 - \xi_i,\ \xi_i \geq 0 \end{aligned} , \tag{5.8}$$

where ϕ is the higher-dimensional mapping function. In solving the optimization problem (in the Lagrangian) the data vectors appear in an inner product. After

the mapping functions are applied, the inner product will be between the vectors $\phi(\mathbf{x_i}) and \phi(\mathbf{x_j})$. Since computations in the high-dimensional space can be computationally very intense, what is referred to as a "kernel trick" is applied [Bos+92]. As long as there is a way to get the inner product, the explicit mapping is not necessary. The kernels are special functions which operate on the data vectors $\mathbf{x_i}$ and $\mathbf{x_j}$ to produce a value equivalent to the inner products of the high-dimensional vectors. A kernel function $K(\mathbf{x_i}, \mathbf{x_j})$ is thus defined by

$$K(\mathbf{x_i}, \mathbf{x_j}) = \phi(\mathbf{x_i})^T \phi(\mathbf{x_j}) \ .$$

The two most commonly used kernel functions used are the radial basis function (RBF) and polynomial kernels.

- RBF kernel with parameter $\gamma = \frac{1}{2\sigma^2}$ and $\sigma > 0$

$$K(\mathbf{x_i}, \mathbf{x_j}) = \exp(-\gamma \|\mathbf{x_i} - \mathbf{x_j}\|^2) \ . \tag{5.9}$$

- A polynomial kernel with degree d

$$K(\mathbf{x_i}, \mathbf{x_j}) = (\mathbf{x_i}^T \mathbf{x_j} + 1)^d \ . \tag{5.10}$$

When applying the SVM algorithm, the proper parameters have to be selected. Typically, a grid-search is performed to find optimal parameters, the right kernel function and hyper-parameters C, γ, and d. As shown earlier, a small C provides a soft margin where errors are allowed to be made while fitting the model to the training set. A large C corresponds to a hard margin where training errors are not allowed. The γ parameter of the RBF kernel defines the influence of the training samples on the decision boundary.

5.2.2 Decision Tree

Decision Trees are very simple and intuitive supervised classification algorithms. Classification is achieved by asking a series of questions which split the sample into more homogeneous sub-samples. Every node poses a question and the possible answers lead to a child. Technically, an exponential number of trees can be constructed and finding the optimal tree is computationally impossible as it is an NP-complete problem. However, the problem is solved by an efficient greedy search algorithms such as CART [Bre+84], ID3 [Qui86], and C4.5 [Qui93]. The decision tree is grown by making successive locally optimum choices on which feature to use

for partitioning [Tan+06]. Starting from the root of the tree, a decision on which attribute best partitions the data, i.e., finding the most informative feature that leads to a pure (almost pure) subgroup. Most commonly, the feature which reduces the uncertainty (entropy) the most is selected first. Hence, mutual information, also called information gain in machine learning and data mining literature, is used to measure the expected entropy reduction as a result of splitting with a given feature. If we denote the whole training set by S. The information gain (IG), i.e., the expected drop in entropy, after splitting with an attribute A is

$$IG(S, A) = H(S) - H(S|A) \ . \tag{5.11}$$

Once the decision tree is constructed, assigning a class label for new data is straight forward. The appropriate branches based on the test conditions are followed until a leaf of the tree is encountered.

For a better understanding lets see the following example.

Example 5.2.1 (A toy example for constructing a decision tree [1])**.** Suppose we have a training dataset of 14 genes shown in Table 5.1. The task is to construct a decision tree to determine whether genes are essential (EG) or not (NEG). Each gene is characterized by three attributes (features). The attributes are the length of the protein, the DNA strand the gene is located on, and the GC content . A possible

Tab. 5.1: A sample training dataset for gene essentiality prediction

Length	GC content	Strand	Essentiality
Medium	0.58	+	NEG
Long	0.53	-	EG
Medium	0.57	+	NEG
Medium	0.53	+	EG
Short	0.51	+	EG
Short	0.51	-	NEG
Medium	0.57	-	NEG
Long	0.59	+	EG
Short	0.56	+	EG
Short	0.54	+	EG
Medium	0.54	-	EG
Long	0.59	-	EG
Long	0.54	+	EG
Short	0.56	-	NEG

decision tree for the data presented in Table 5.1 is shown in Fig. 5.3. The training set S contains 9 EGs and 5 NEGs. Hence, the entropy, $H(S) = -\frac{9}{15}\log_2\frac{9}{15} - \frac{5}{15}\log_2\frac{5}{15} = 0.940$. At the root of the tree, the information gain of the three attributes are determined.

[1] Modified from the famous prediction of playing tennis example of [Mit+97]

Information gain calculations for the attribute "Length":
5 short genes, of which 3 are EGs and 2 are NEGs.
5 medium genes, of which 2 are EGs and 3 are NEGs.
4 long genes, of which all of them are EGs.

$$
\begin{aligned}
&P(\text{Length} = \text{Short}) = \frac{5}{14} \,, \\
&P(\text{Length} = \text{Medium}) = \frac{5}{14} \,, \\
&P(\text{Length} = \text{Long}) = \frac{4}{14} = \frac{2}{7} \,, \\
&H(S|\text{Length} = \text{Short}) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.971 \,, \\
&H(S|\text{Length} = \text{Medium}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971 \,, \\
&H(S|\text{Length} = \text{Long}) = 0 \,, \\
&H(S|\text{Length}) = \sum_{\forall x} P(\text{Length} = x) H(S|\text{Length} = x) = 0.693 \,, \\
&IG(S, \text{Length}) = H(S) - H(S|\text{Length}) \,, \\
&IG(S, \text{Length}) = 0.940 - 0.693 = 0.247 \,.
\end{aligned}
$$

Information gain calculations for the attribute "Strand":
8 genes on the + strand, of which 6 are EGs and 2 are NEGs.
6 genes on the - strand, of which 3 are EGs and 3 are NEGs.

$$
\begin{aligned}
&P(\text{Strand} = +) = \frac{8}{14} = \frac{4}{7} \,, \\
&P(\text{Strand} = -) = \frac{6}{14} = \frac{3}{7} \,, \\
&H(S|\text{Strand} = +) = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4} = 0.811 \,, \\
&H(S|\text{Strand} = -) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1 \,, \\
&H(S|\text{Strand}) = \sum_{\forall x} P(\text{Strand} = x) H(S|\text{Strand} = x) = 0.892 \,, \\
&IG(S, \text{Strand}) = H(S) - H(S|\text{Strand}) \,, \\
&IG(S, \text{Length}) = 0.940 - 0.892 = 0.048 \,.
\end{aligned}
$$

Information gain calculations for the attribute "GC content":
Since GC content is continuous attribute, it should be discretized. For this example, we look at values less than 0.55 and greater than 0.55.
8 genes on the + strand, of which 6 are EGs and 2 are NEGs.
6 genes on the - strand, of which 3 are EGs and 3 are NEGs.

$$
\begin{aligned}
&P(\text{GC content} > 0.55) = P(\text{GC content} < 0.55) = \frac{1}{2}\,, \\
&H(S|\text{GC content} > 0.55) = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} = 0.985\,, \\
&H(S|\text{GC content} < 0.55) = -\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7} = 0.592\,, \\
&H(S|\text{Strand}) = \sum_{\forall x} P(\text{Strand} = x)H(S|\text{Strand} = x) = 0.788\,, \\
&IG(S, \text{Strand}) = H(S) - H(S|\text{Strand})\,, \\
&IG(S, \text{Length}) = 0.940 - 0.788 = 0.152\,.
\end{aligned}
$$

The attribute with the largest information gain is Length. Hence, we split based on the length at the root node. After that, the algorithm runs recursively on the non-leaf branches, until everything is fully classified.

Although decision trees are powerful machine learning algorithms which are easy to understand and visualize, they tend to overfit the training data and hence cannot generalize well. One way to deal with the overfitting problem is to stop growing the tree before it completely classifies the training data. The other way is to prune the tree after it is fully grown, i.e., by collapsing nodes which do not provide a lot of information. The other problem associated with decision tree learning is instability.

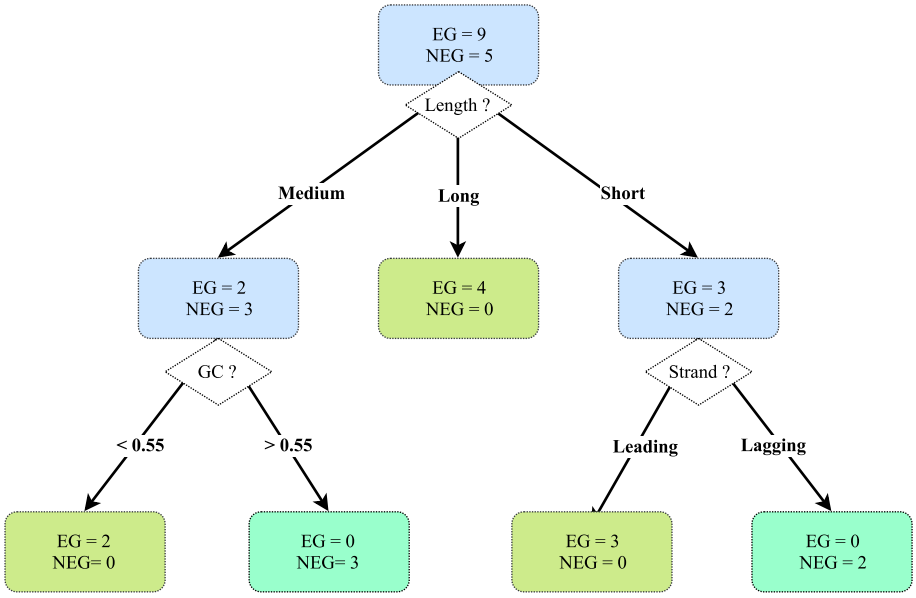

Fig. 5.3: A learned decision tree for predicting EGs.

Small changes in the input data can cause large changes in the tree (due to the changes in the information gain values) [LB02]. Since we want our models to be robust to noise, stability is a desirable property.

The overfitting (high variance) and stability problems can also be mitigated through what is known as an ensemble learning approach. Ensemble learning methods build multiple models (base classifiers) and final predictions are obtained by aggregating the classification results. The two well known ensemble techniques are bagging (bootstrap aggregating) [Bre96] and boosting [Sch+98]. Both bagging and boosting involve the generation of models using subsamples randomly selected (with replacement) from the training data. The difference between the two is in the way each classifier is built. In bagging each model is constructed independently where as in boosting, models are constructed sequentially. Subsequent base classifiers are trained on samples selected based on the prediction outcomes of the previous classifier (misclassified samples have a high chance of being selected).

5.2.3 Random Forest

Random Forests, introduced by Breiman in 2001 [Bre01], are ensemble learning algorithms which realize classification using bagging. The base classifiers are as the name suggests decision trees. The key idea in the random forest algorithm is the introduction of an additional randomization to the bagging procedure to reduce the correlation between the trees in the forest. In addition to the randomization through the sampling with replacement, the features considered for finding the best split are also randomly chosen. In regular decision trees, the best among all features are used for partitioning. In random forests, nodes are split using the best among a randomly selected subset of features.

The steps for a random forest algorithm are:

1. Sub-sampling the training set: Draw *numTrees* samples from the training data with replacement.

2. For each sub-sampled trainin g data, grow a decision tree. At each node, the best split is calculated based on a randomly chosen subset of features (*numFeatures*).

3. Predict new examples by majority voting or averaging the prediction results of the *numTrees* trees.

5.3 Data Sources

The data for EGs and non EGs (NEGs) for the 15 bacteria and 1 archaeon were downloaded from the database of EGs (DEG 13.5). DEG collects a comprehensive list of essential and non-essential genes identified by various researchers through experimental gene knockout and silencing methods [Luo+14]. In DEG, although the EGs dataset for eukaryotes are available, the list of NEGs are not included. One way to deal with this is, as done in most of the gene essentiality prediction studies, to regard all other genes as NEGs. Since some studies consider and test a small number of genes (small-scale screenings), taking the untested genes as non-essential could be misleading. Hence, for eukaryotic EG predictions, we used the dataset presented by the database of Online GEne Essentiality (OGEE) [Che+11]. The species used in this study along with the number of EGs and NEGs are listed in Table 5.2. The genome sequences were obtained from the NCBI database (ftp://ftp.ncbi.nih.gov/genomes/). We selected the bacterial species studied by Ning at al. [Nin+14] to allow for easy performance comparisons.

Tab. 5.2: Names and abbreviations of the species used in this study. The accession numbers along with the number of essential and non-essential genes are listed.

No.	Organism	Abbr.	Accession No.	EGs	NEGs
1	*Acinetobacter baylyi ADP1*	AB	NC_005966	499	2594
2	*Bacillus subtilis 168*	BS	NC_000964	271	3904
3	*Escherichia coli MG1655*	EC	NC_000913	296	4077
4	*Francisella novicida U112*	FN	NC_008601	392	1329
5	*Haemophilus influenzae Rd KW20*	HI	NC_000907	642	512
6	*Helicobacter pylori 26695*	HP	NC_000915	323	1135
7	*Mycobacterium tuberculosis H37Rv*	MT	NC_000962	614	2552
8	*Mycoplasma genitalium G37*	MG	NC_000908	381	94
9	*Mycoplasma pulmonis UAB CTIP*	MP	NC_002771	310	322
10	*Pseudomonas aeruginosa UCBPP-PA14*	PA	NC_008463	335	960
11	*Salmonella enterica serovar Typhi*	SE	NC_004631	353	4005
12	*Salmonella typhimurium LT2*	ST	NC_003197	230	4228
13	*Staphylococcus aureus N315*	SA	NC_002745	302	2281
14	*Staphylococcus aureus NCTC 8325*	SA2	NC_007795	351	2541
15	*Vibrio cholerae N16961*	VC	NC_002505	779	2943
16	*Methanococcus maripaludis S2*	MM	NC_005791	519	1077
17	*Caenorhabditis elegans*	CEL	NC_003280	742	10704
18	*Drosophila melanogaster*	DRO	NT_033779	267	13514
19	*Homo sapiens*	HSA	NC_000015	1632	19897
20	*Mus musculus*	MUS	NC_000081	4289	4592

5.4 Information-Theoretic Features

In computational biology and bioinformatics, information-theoretic quantities have been widely used to model, analyze, and/or measure both structural and organizational properties in biological sequences. In this thesis, we used IT quantities to produce features which will enable the classification of essential and non-essential genes. Four feature sets were used. 7 entropy, 17 mutual information (MI), 65 conditional mutual information (CMI), and 2 Markov model (M) related. Here, we present a brief description of the information-theoretic quantities used in this work. A detailed description can be found in standard information theory text books [CT91].

5.4.1 Mutual Information

The mutual information measures the information shared by two random variables (see Section 2.1.3). It is the amount of information provided by one random variable about the other. Here, mutual information was used to measure the information between consecutive bases X and Y and is mathematically defined as

$$I(X,Y) = \sum_{x\in\Omega}\sum_{y\in\Omega} P(x,y)\log_2\frac{P(x,y)}{P(x)P(y)}\,, \tag{5.12}$$

where Ω is the set of nucleotides $\{A,T,C,G\}$, $P(x,y)$ is the joint probability, and $P(x)$ and $P(y)$ are the marginal probabilities. The probabilities are estimated from their relative frequencies in the corresponding gene sequences. Along with the total mutual information computed according to Eq. (5.12), for each base pair (x,y), the quantity $P(x,y)\log_2\frac{P(x,y)}{P(x)P(y)}$ is calculated and used as a feature. Therefore, a total of 17 MI-related features were calculated.

5.4.2 Conditional Mutual Information

The mutual information between two random variables X and Y conditioned on a third random variable Z is given by Eq. (2.19). The three positions in a DNA triplet are regarded as the random variables X, Z, and Y, respectively. The mutual information between the bases at the first and the third position conditioned on the base in the middle is calculated according to Eq. (2.19) and used as a feature. In addition, for each possible triplet, the quantity $P(x,y,z)\log_2\frac{P(z)P(x,y,z)}{P(x,z)P(y,z)}$ was calculated. This resulted in a total of 65 CMI-based features.

5.4.3 Entropy and Relative Entropy

Shannon and Gibbs Entropies

In Chapter 4, Shannon and Gibbs entropies were applied on the complete genomes to measure the average informtion content and thermodynamic stability. Here, we used them as features to measure the information content and stability of the genes. Shannon and Gibbs entropies of the genes, for block sizes of 2 (base pairs) and 3 (triplates), were calculated according to Eq. (4.1) and Eq. (4.3), respectively.

Kullback-Leibler Divergence

As presented in Chapter 2, the Kullback-Leibler Divergence (KLD) [KL51] measures the similarity of a probability distribution to a model distribution (2.17). The frequencies of the nucleotides, di-nucleotides, and tri-nucleotides in a given gene sequence were compared against the corresponding frequencies in the genome of the organism used for training the model (background distributions).

In total, 7 entropy-related features were calculated.

5.4.4 Markov Model

Assuming that the gene sequences in the essential and non-essential classes are generated by two separate Markov sources, we construct a Markov chain and use the scores of the genes as Markov features. The training set is subdivided into a subset containing the essential and non-essential samples. Thereafter, each subset is used to generate a Markov chain of a preselected or estimated order m ($MC_+(m)$ and $MC_-(m)$ for essential and non-essential genes, respectively).

The first step is to estimate the correct orders of the Markov chains.

Markov order estimation

Numerous Markov chain order estimators have been put forth in the literature. We have assessed the performances of selected estimators [Ton75; Kat81; PS05; DD05a; Men+11] on DNA sequence data and the estimator proposed by Papapetrou and

Kugiumtzis [PK13] was chosen. The order estimation is based on the CMI given in Eq. (2.19). A Markov chain of order m_E has the following property.

$$P(x_n|x_{n-1},\dots,x_{n-m_E},x_{n-m_E-1},\dots) = P(x_n|x_{n-1},\dots x_{n-m_E}) \,. \tag{5.13}$$

For any $m \leq L$, since the n^{th} and $(n-m)^{\text{th}}$ nucleotides are dependent, the CMI between the two conditioned on the $m-1$ bases in the middle will be greater than zero. Conversely, for any $m > L$, two nucleotides are independent and the CMI will be zero. Using this observation, Papapetrou and Kugiumtzis have proposed both parametric and non-parametric significance testing procedures [PK13; PK16]. Compared to other approximations, the results of the gamma distribution were best suited [PK16]. Hence, we used the gamma distribution based parametric approach for estimating the orders. In a symbol sequence of length N, $\hat{I}(X;Y|Z)$, the estimate of the CMI, is approximated by the gamma distribution as

$$\hat{I}(X;Y|Z) \approx \Gamma(\frac{|Z|}{2}(|X|-1)(|Y|-1), \frac{1}{N\ln 2}) \,. \tag{5.14}$$

The gamma distribution is used as the distribution of the null hypothesis, $H_0 : CMI(m) = 0$. Since $CMI \geq 0$ always holds, one-sided parameter testing is performed. Thus, the p-value (see Chapter 2.1.5) is computed from the complementary cumulative distribution of the gamma distribution in Eq. (5.14). H_0 is rejected if the p-value is less than the nominal significance level ($\alpha = 0.05$). Starting from order zero, the null hypothesis is checked and if it is rejected, the next order is checked and the process continues until the null hypothesis is accepted.

After the orders are estimated, the transition probabilities of the two Markov chains are empirically estimated using the so-called Lidstone estimator [Lid20; DD05b]. Let $N_x(\mathbf{v})$ denote the number of times a word $\mathbf{v}$ of length m appears in a training sequence $\mathbf{x}$. The probability that the next nucleotide is a, where $a \in \Omega = \{A, C, G, T\}$, conditioned on the context $\mathbf{v} \in \Omega^m$ is

$$P_{v,a} = \frac{N_x(\mathbf{v}a) + \delta}{N_x(\mathbf{v}) + 4\delta} \,. \tag{5.15}$$

The parameter δ assigns a pseudo count to unseen symbols to avoid zero probabilities. We experimentally checked and found that small values of δ are better suited and consequently set $\delta = 0.001$. After the two Markov chains have been constructed, the Markov features are computed by scoring every gene using the generated Markov chains. If we represent a sequence of length L as $b_1, b_2, b_3, ..., b_L$, the score is calculated as

$$Score = \sum_{i=1}^{L-m} P(b_i b_{i+1} \dots b_{i+m}) \log_2(\frac{P(b_{i+m}|b_i b_{i+1} \dots b_{i+m-1})}{P(b_{i+m})}) \,. \tag{5.16}$$

The score measures how likely the sequence is generated by the given m_E^{th} and m_N^{th} order Markov chains. The scores of the gene sequence on the Markov chains $MC_+(m_E)$ and $MC_-(m_N)$ were used as features. For intra-organism predictions, the Markov orders were estimated from the training sets whereas for cross-organism gene essentiality predictions, order estimation increased the computational complexity without improving the result. Hence, we decided to use a fixed order Markov chain. After experimenting with orders 1 up to 6, order 1 (i.e., $m_E = m_N = 1$) was selected.

5.5 Other Simple Sequence-Based Features

To further increase the prediction performances, among the frequently used and easily accessible features, the GC content, length of the protein, and GC3 (GC content in the 3rd position of the codons) were computed. In addition, features related to stop codon usage were included. As in [Ser+06; Son+14; Yua+12], we calculated the number of "close-to-stop" codons, which are codons a single third-nucleotide substitution away from one of the three stop codons (TAA, TAG, TGA). The five codons differing by a single base to the stop codons are TAC, TAT, TGT, TGC, and TGG. Hence, the total frequency of these codons is used as a feature. The idea is to measure how likely it is for the protein to be terminated when a substitution error occurs. In a similar direction, to include the case where a single insertion or deletion occurs causing a frame-shift, we added a new set of features. We computed the number of stop codons and the position of the first stop codon in the other two reading frames. In total, 8 features are included. For brevity, we call this set of features Stop + Len + GC or non-IT features.

5.6 Classification Approach and Performance Evaluation

In this section, we present the procedures used for training and testing along with the employed performance evaluation methods. The flowchart in Fig. 5.4 presents the classification procedures. To avoid the scaling differences between features, all features were standardized to zero mean and unit variance (feature scaling) prior to the training and testing of the classifier. If a Random Forest classifier is used, feature scaling is not needed.

Typically, the number of NEGs is significantly larger than that of the EGs. To balance the two classes, various schemes of under- and over-sampling approaches could be

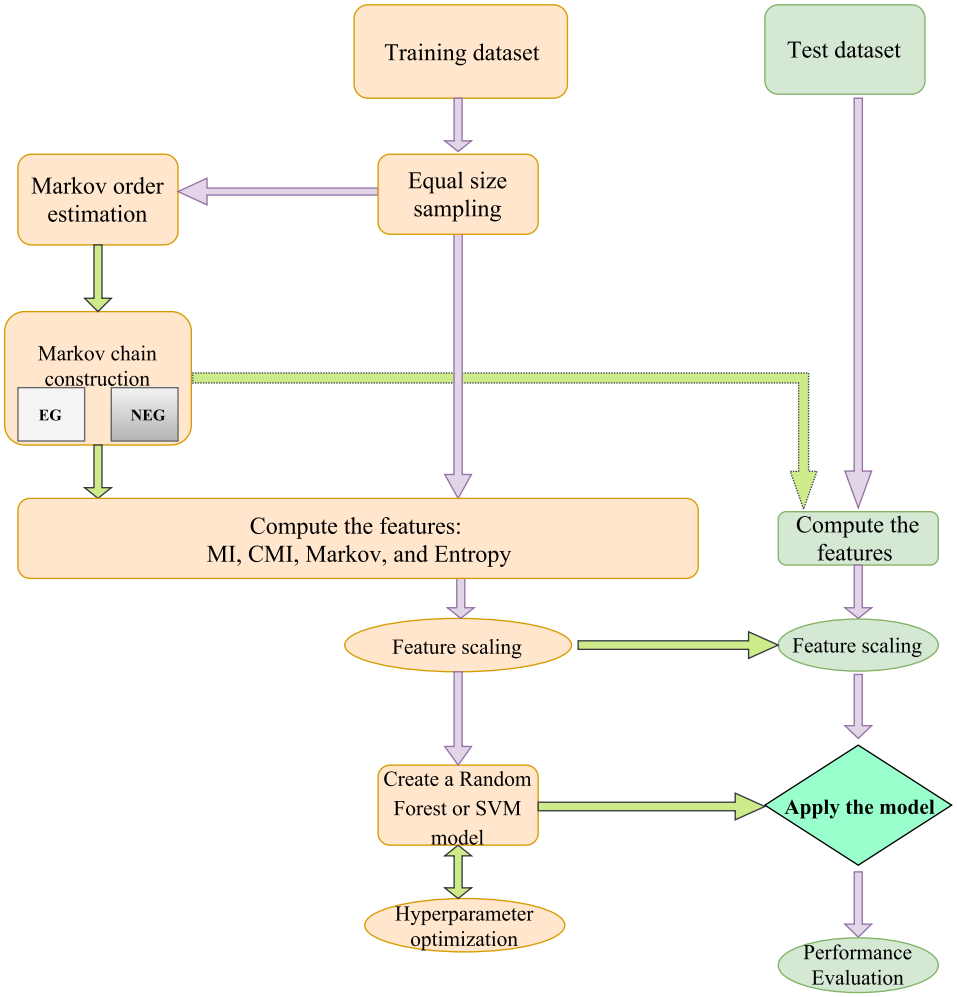

Fig. 5.4: A flow chart of the classification procedure [NH17].

taken. Since it was shown in [Yu+17] that the choice of a balancing approach does not influence the performance of essential gene predictions, we followed a random under-sampling approach. The minimum number of iterations needed to cover all genes in the majority class (mostly NEGs) can be determined using what is known as the Clarke and Carbon formula [CC76], i.e.,

$$N = \frac{\log(1-p)}{\log(1-\#EG/\#NEG)}\ , \tag{5.17}$$

where p is the probability that a given gene is represented in the samples. $\#EG$ and $\#NEG$ are the number of essential and non-essential genes, respectively. This formula is widely used to calculate the coverage statistics in genome sequencing.

The training and testing data sets can emanate either from the same (intra-organism) or from different species (cross-organism). In cross-organism predictions, classifiers were trained on one (or more) organism (s) and tested on another, whereas in intra-organism predictions 80 % of the data is used for training the models and 20

% is used for testing. The random selections were repeated 100 times, i.e., 100-fold Monte Carlo cross-validations were performed for model establishment.

The Area Under the Curve (AUC) of the Receiver Operating characteristic Curve (ROC) was used to evaluate the performance of our classifier. The ROC plots the true positive rate versus false positive rate. It shows the trade-off between sensitivity Sn and specificity Sp for all possible thresholds. Other performance evaluation criteria such as F-measure and Accuracy were also calculated. However, these parameters depend on the selected threshold value. Therefore, we mainly used the AUC score for analyzing the performance of the classifier. For a given threshold, the true positive (TP), false negative (FN), false positive (FP), and true negative (TN) predictions are determined and Sn, Sp, F-measure, Accuracy (Acc), and Positive Predictive Value (PPV) are calculated as follows:

$$\begin{aligned} Sn &= \frac{TP}{TP+FN}\,, \\ Sp &= \frac{TN}{TN+FP}\,, \\ Acc &= \frac{TP+TN}{TP+FN+TN+FP} \\ \text{F-measure} &= \frac{2TP}{2TP+FN+FP} \\ PPV &= \frac{TP}{TP+FP}\,. \end{aligned} \tag{5.18}$$

Feature preparation and computations were performed using Python 3.5.2, scikit-learn module [Ped+11]. We implemented a Random Forest classifier using the data analytics platform Konstanz Information Miner (KNIME 3.3.1) [Ber+07]. Information gain is used as a splitting criterion.

5.7 Essential Gene Prediction in Bacteria

5.7.1 Intra-Organism Predictions

The bacteria EC, BS, and MP were selected to assess the intra-organism prediction performance in which models are trained and tested on data obtained from the same species. This setup is especially useful when a portion of the essentiality annotations are performed using an experimental method and machine learning procedures are performed to complete the analysis.

EC has 296 EGs and 4077 NEGs. The Random Forest classifier was trained on 80 % of the data and the remaining 20 % were used to test the model. All five

feature groups (MI, CMI, Entropy, Markov, Stop + Len + GC), both individually and collectively, were used and 100 iterations were performed. The average ROC curves using the Random Forest model are shown in Fig. 5.5. The combination of all features produced a very good AUC score of 0.86. Equally good results were achieved by using CMI features alone. The results of MI and Markov features were also satisfactory while the entropy and the non-IT features yielded a relatively small prediction accuracy.

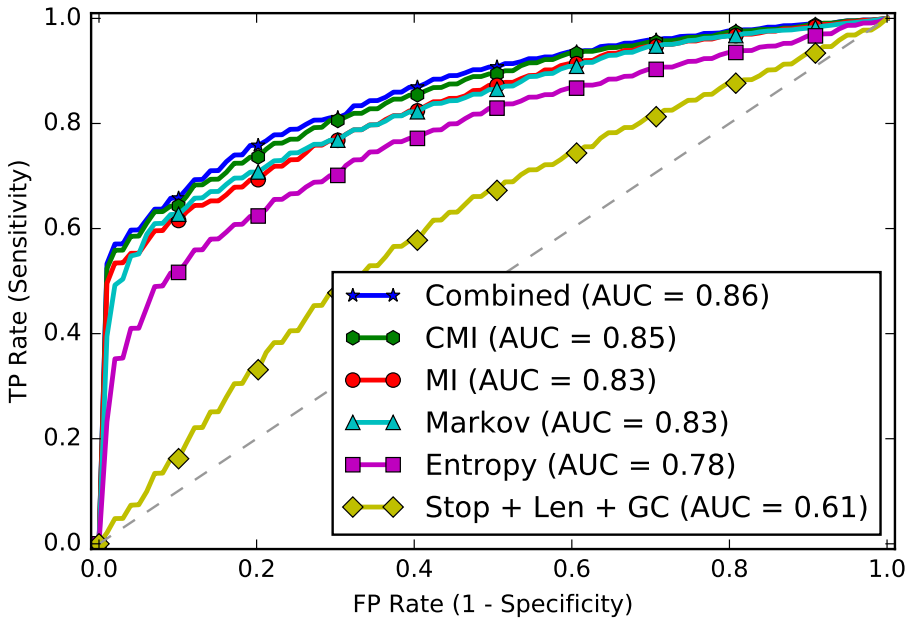

Fig. 5.5: EG predictions in *E. coli*. The estimated Markov orders were 5 for both EGs and NEGs.

The SVM algorithm resulted in an almost similar performance. Using the entropy, mutual information, and Markov features and a linear kernel with parameter $C = 0.85$, an average AUC score of 0.85 was obtained.

Similarly, the prediction results of our proposed method applied to the bacterium MP are presented in Fig. 5.6. Using the complete feature set, an AUC score of 0.82 was obtained. Taken separately, all feature groups provided a score greater than 0.72. The result shows that the non-IT features (Stop + Len + GC) also have the ability to distinguish between essential and non-essential genes with a decent accuracy (0.72). A much higher AUC score of 0.90 was achieved for the bacteria BS (Fig. 5.7). Both MI and CMI features achieved a score of 0.87, while the entropy and Stop + Len + GC features yielded 0.79 and 0.76, respectively.

To check if the estimated Markov order is indeed better than the other orders, we have used features from the scores of fixed Markov orders ranging from 1 to 7 and tested the performance of our classifier. A linear SVM ($C = 0.5$) is applied on a

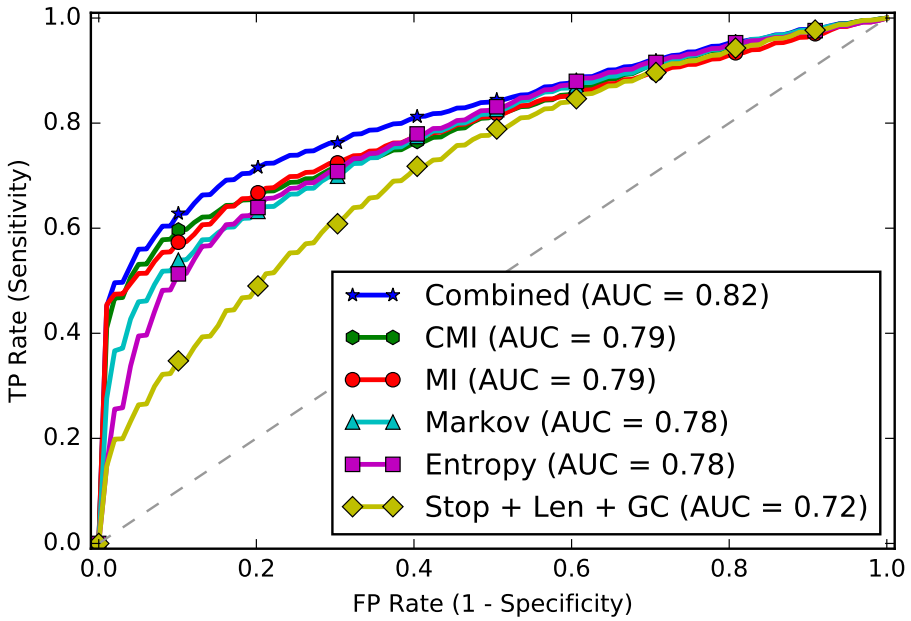

Fig. 5.6: EG predictions in *M. pulmonis*. The estimated Markov orders were 6 for both EGs and NEGs.

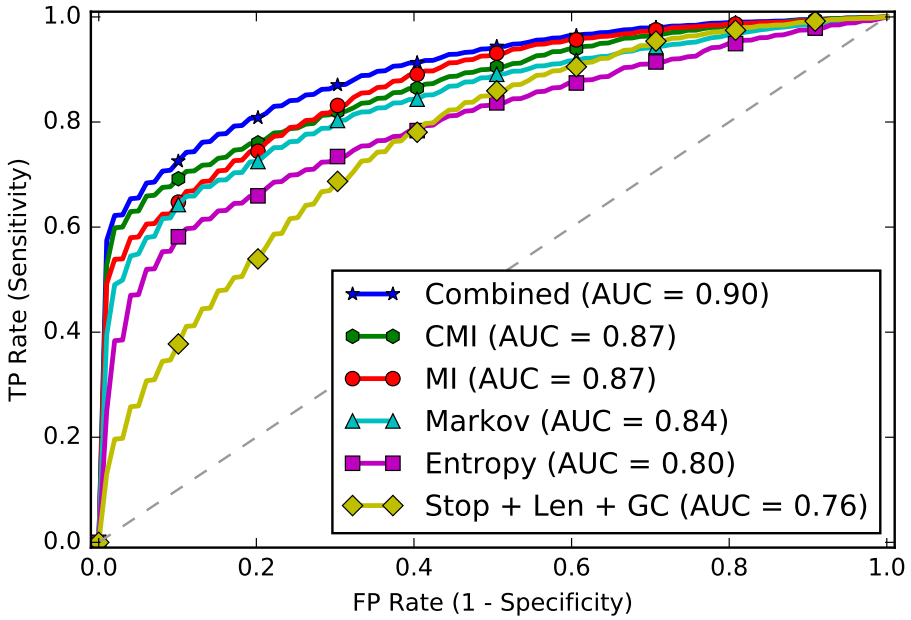

Fig. 5.7: The average ROC curves of *B. subtilis* EG prediction. The estimated Markov orders were 5 for EGs and 4 for NEGs.

balanced dataset using a single sample. The AUC scores for the prediction of EGs in *B. subtilis*, *E. coli*, and *M. pulmonis* are shown in Fig. 5.8. In *E. coli*, the estimated order is 5 and the maximum AUC score was obtained by using orders 4 and 5 (AUC = 0.83). In *B. subtilis*, the estimated order is order 4 for NEGs and order 5 for EGs. The maximum AUC score of 0.83 is observed for order 6. However, the performance using the estimated orders, on the same sample, is also 0.83. This is a good example

to show how the order estimation finds the optimal performance, rather than a random choice of an order. *M. pulmonis* has an estimated order of 6 and it is with order 6 Markov chain the maximum AUC is obtained (0.8). Therefore, this shows that the estimated order achieves the best possible AUC score.

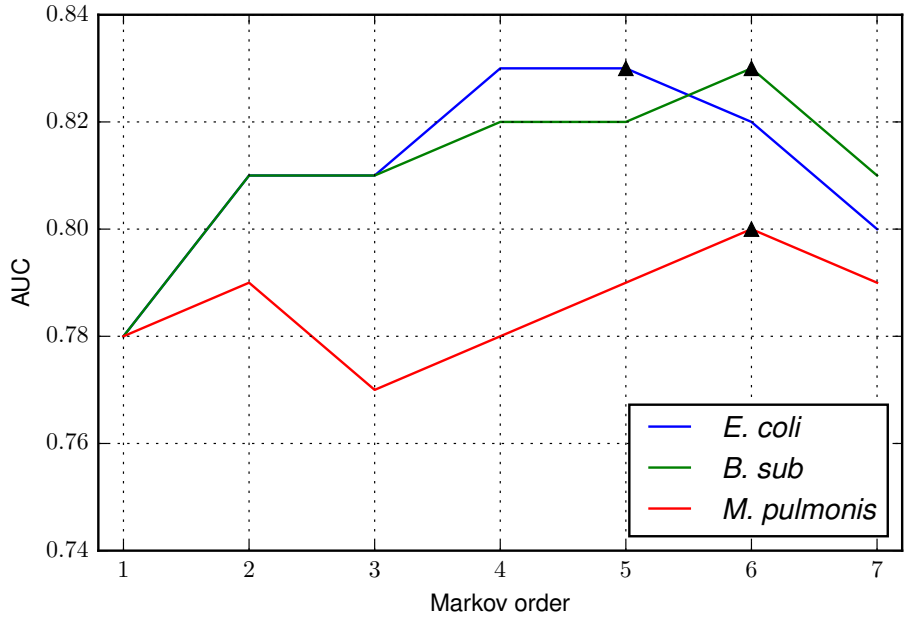

Fig. 5.8: The AUC scores of different Markov orders. The black triangles indicate the maximum AUC score.

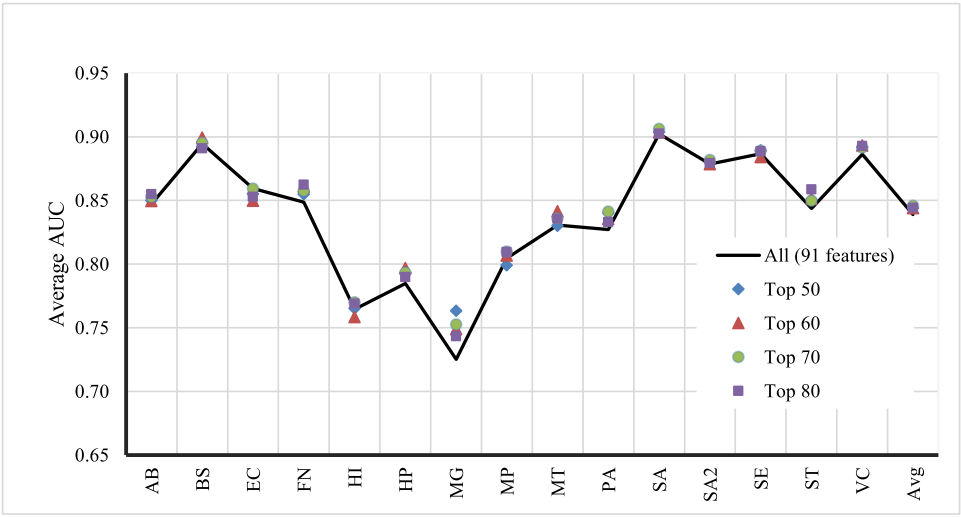

Fig. 5.9: Average AUC scores of intra-organism essential gene predictions in 15 bacteria species. The prediction performance of the top 50, 60, 70, and 80 features based on information gain is also shown.

The average AUC scores of a 100-fold Monte Carlo cross-validation experiment on the 15 bacteria using only the information-theoretic features are presented in Fig. 5.9. The values range between 0.73 and 0.90, 0.84 on average. Except for three bacteria, namely HI, HP, and MG, the AUC scores are more than 0.80. We also performed

a feature selection experiment using the information gain rankings, selecting the top 50, 60, 70, and 80 features (Fig. 5.9). Using the top 70 features provided the better accuracy on average. For MG taking only the top 50 features yielded a 4 % gain. The result demonstrates that fewer features can be used to improve the computational complexity without affecting the accuracy of the predictions. Overall, the improvement gained by feature selection is not significant. Therefore, in the remaining parts of this work, feature selection is not considered. To assess the contributions of the different feature types to the classification task, the information gain rankings for all species were collected and a global feature ranking was obtained. The top 20 features consists of 8 MI, 8 CMI, 3 entropy, and 1 Markov features. This shows that all feature classes contribute to the high prediction performances.

Compared to Ning et al.'s [Nin+14] essentiality predictor which uses only sequence based and sequence-derived features, our method yielded better AUC scores. The AUC scores for EC and MP were improved from 0.82 to 0.86 and from 0.74 to 0.80, respectively. Similarly, in comparison with the inter-nucleotide distance sequences based essential gene predictor proposed by Li et al. [Li+17], our method provided an improvement of up to 9 %. For EC, the AUC score is improved from 0.80 to 0.86, for BS from 0.81 to 0.89, for SE from 0.80 to 0.89, and for SA from 0.88 to 0.90. In addition, our average AUC score (0.84) was also much better than Yu et al.'s [Yu+17] fractal features based predictor (0.77 on 27 selected bacteria).

5.7.2 Cross-Organism Predictions

So far, both the training and test sets were taken from a single genome. In this section, models trained on a given organism (or groups) are used to predict the essential and non-essential genes of another distinct organism. Cross-organism predictions are more realistic and useful in *ab initio* identification of EGs. Two approaches were taken. The first approach is a pairwise cross-organism prediction in which models trained on one species are used to predict the essential and non-essential genes of every other species, separately. The other approach is a leave-one-species-out procedure whereby genes of the 14 bacteria are collectively used for establishing a model and EGs of the remaining bacterium are predicted.

Pairwise Predictions

Pairwise cross-organism predictions among the 15 bacteria were performed to see how well essentiality annotations can be transferred between both closely and distantly related species. The 15×15 average AUC matrix is presented in Fig. 5.10. The bacteria are also grouped together according to the phylogenetic tree constructed

using the PhyloT tree generator (http://phylot.biobyte.de/index.html). The overall prediction performances were very good (AUC scores of up to 0.92 were obtained). However, cross-predictions among MT and MG, MP, FN, and HP are very bad, even sometimes worse than a random guess. As described in [Che+14; Zha+16], larger evolutionary distance, differences in growth conditions, phenotypes, and lifestyles, and poor quality of the training data may have led to poor performances.

			Proteobacteria							Terrabacteria group						
			Gammaproteobacteria								Bacillales		Mycoplasmatales			
	Pseudomonadales		Enterobacteriaceae								Staphylococcus					
				Salmonella												
	AB	**PA**	**EC**	**SE**	**ST**	**FN**	**HI**	**VC**	**HP**	**BS**	**SA**	**SA2**	**MG**	**MP**	**MT**	**Average**
AB		0.80	0.86	0.87	0.82	0.82	0.76	0.71	0.77	0.87	0.86	0.84	0.69	0.74	0.77	**0.80**
PA	0.78		0.82	0.83	0.80	0.78	0.75	0.70	0.67	0.78	0.80	0.79	0.65	0.63	0.78	**0.75**
EC	0.84	0.81		0.90	0.84	0.80	0.75	0.73	0.73	0.86	0.85	0.84	0.68	0.62	0.79	**0.79**
SE	0.84	0.81	0.89		0.85	0.78	0.74	0.73	0.74	0.86	0.87	0.85	0.70	0.63	0.79	**0.79**
ST	0.79	0.79	0.85	0.85		0.76	0.76	0.68	0.70	0.81	0.82	0.81	0.67	0.63	0.79	**0.77**
FN	0.78	0.70	0.79	0.77	0.76		0.75	0.77	0.79	0.75	0.84	0.82	0.73	0.75	0.45	**0.75**
HI	0.73	0.76	0.74	0.74	0.75	0.78		0.66	0.75	0.76	0.79	0.80	0.70	0.74	0.62	**0.74**
VC	0.76	0.73	0.81	0.80	0.75	0.77	0.73		0.77	0.73	0.70	0.69	0.65	0.70	0.63	**0.73**
HP	0.73	0.62	0.70	0.68	0.68	0.81	0.76	0.81		0.71	0.74	0.74	0.71	0.76	0.33	**0.70**
BS	0.84	0.78	0.86	0.87	0.83	0.79	0.75	0.71	0.75		0.88	0.86	0.70	0.70	0.79	**0.79**
SA	0.83	0.75	0.85	0.85	0.80	0.83	0.76	0.71	0.76	0.86		0.90	0.73	0.79	0.70	**0.79**
SA2	0.83	0.76	0.85	0.85	0.80	0.83	0.75	0.73	0.78	0.85	0.92		0.73	0.80	0.72	**0.80**
MG	0.69	0.58	0.66	0.62	0.63	0.79	0.74	0.66	0.72	0.66	0.76	0.74		0.72	0.41	**0.67**
MP	0.71	0.65	0.67	0.67	0.65	0.80	0.75	0.73	0.77	0.71	0.77	0.78	0.70		0.44	**0.70**
MT	0.78	0.81	0.82	0.83	0.82	0.64	0.73	0.70	0.55	0.78	0.79	0.79	0.45	0.45		**0.71**
Gram	-	-	-	-	-	-	-	-	-	+	+	+	+	+	N	

Fig. 5.10: Pairwise cross-organism predictions results. 15×15 average AUC scores are presented. Rows indicate organisms used for training while columns are test organisms. The phylogenetic relationship and the taxonomic classification of the bacteria are also shown.

Although close evolutionary distance and similar lifestyles provide common essential gene characteristics, the results for the distantly related species were also good. For instance, BS and EC diverged over a billion years ago [CP02], before the divergence of plants and animals, and yet highly accurate predictions were possible (AUC score of 0.86). In addition, models trained based on the taxonomic orders Bacillales (BS, SA, SA2) and Enterobacterales (EC, SE, ST) produced better overall performances. Hence, future blind essentiality predictions of a new species can be done using one of these bacteria.

The performance of our predictor is as good as the other existing state-of-the art gene essentiality predictors which use homology, gene-expression, and network topology based features in addition to sequence-derived information. Note that sequence similarity searching is computationally expensive. The comparison to Deng et al.'s [Den+11] and Song et al.'s [Son+14] ZUPLS classifiers among AB, BS, EC, and PA is shown in Table 5.3. On average, our method is slightly better than Deng et al's (2 %) in AUC score. ZUPLS is the best method among the sequence-based predictors and

on average it is only 3 % better than our method. However, since a database search for homology and domain information are not required, our method could be more advantageous in case of limited computational power. In addition, our approach is much better in terms of PPV scores.

Tab. 5.3: Comparison of the prediction performance (average AUC score) among AB, BS, EC and PA.

Train	Test	Deng et al. [Den+11]		Song et al. [Son+14]		Our method	
		AUC	PPV	AUC	PPV	AUC	PPV
AB	EC	0.89	0.75	0.91	0.64	0.86	0.99
BS	AB	-	-	0.86	0.74	0.84	0.77
BS	EC	0.86	0.73	0.91	0.64	0.86	0.87
BS	PA	-	-	0.81	0.44	0.78	0.59
EC	AB	0.8	0.85	0.86	0.79	0.84	0.65
EC	BS	0.8	0.93	0.93	0.74	0.86	0.65
EC	PA	-	-	0.81	0.47	0.81	0.77
PA	EC	0.82	0.47	-	-	0.82	0.90
Average		**0.83**	**0.75**	**0.87**	**0.64**	**0.84**	**0.77**

Cheng et al. [Che+13] and Liu et al. [Liu+17] made pairwise predictions on 21 and 31 species, respectively, providing 21×21 and 31×31 AUC matrices. We filtered out the common bacterial species and compared the results. Here, it should be noted that, in all the three methods, the classifiers for each species are trained independently and tested on every other species. Hence, taking the sub-group (15×15) and comparing the results is fair. Looking at the distribution of the AUC scores and the corresponding mean AUC values, our predictor (0.75) was 14 % better than Liu et al.'s (0.61) while Cheng et al.'s predictor (0.79), being the best essentiality predictor, was 4 % better than ours. Considering that Cheng et al. used network, gene expression, and homology information, the AUC scores of our method are very good. To compare the performances of the three methods, the distribution of the pairwise AUC scores of the 15 common species is plotted in Fig. 5.11.

Leave-one-species-out predictions

In the leave-one-species-out approach, we predicted the essential/non-essential genes of one species using a model trained on the remaining 14 bacterial annotated genes. This approach is also very practical for blind essentiality annotations of new organisms. We performed this analysis using both SVM and Random Forest classifiers.

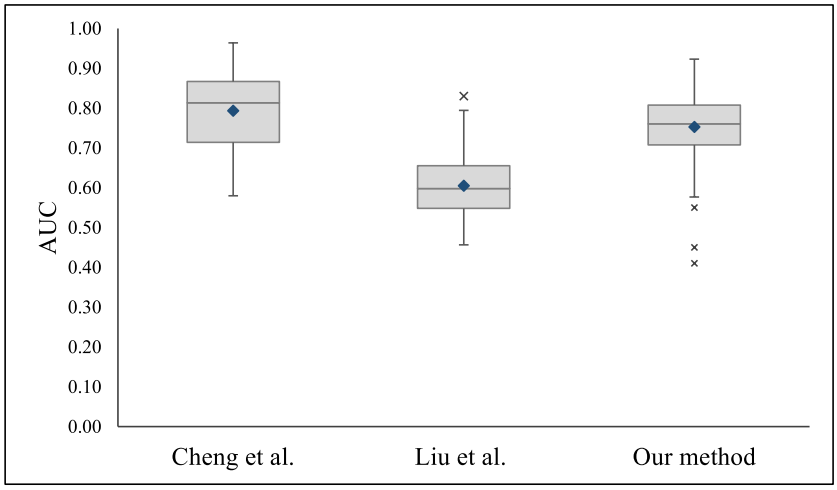

Fig. 5.11: A comparison between pairwise prediction results of our method and two existing methods, proposed by Cheng et al. [Che+13] and Liu et al. [Liu+17]. The diamond markers show the mean values.

The prediction performance of our method is shown in Table. 5.4. Apart from MG whose AUC score is 0.68, very good results (AUC $\geq$ 0.75) were obtained for all other species. Both machine learning algorithms yielded a similar 0.8 average AUC score and comparable results on individual species. This shows that the high prediction accuracy of our method is due to the ability of the information-theoretic features to capture gene essentiality/non-essentiality attributes.

Three studies have used leave-one-species-out approach to assess the performance of their models. Palaniappan and Mukherjee [PM11] in 2011, Geptop [Wei+13] in 2013, and Liu et al. [Liu+17] in 2017. The average AUC score has a 10 % and 19 % improvement over Liu et al.'s and Palaniappan and Mukherjee's, respectively. Our method is also comparable to Geptop. However, for well-studied organisms like EC and BS, Geptop is significantly better. Along with the homology- and phylogeny-based predictor, in [Wei+13], the results of another method, called integrative compositional information predictor, were reported. Codon and amino acid compositions and CodonW features (158 features) were used. Compared to this method which used sequence composition features, our method is slightly better.

Cross-Validation on All Bacteria

The other most common method to asses the prediction accuracy of machine learning models is a 5-fold cross-validation. After the total data consisting of 6078 EGs and 33477 NEGs is divided into 5 separate subsamples, each subsample is tested on a model trained on the combination of the other 4 subsamples. An average AUC score

Tab. 5.4: Leave-one-species-out results using SVM (rbf kernel) and Random Forest classifiers. The average AUC scores of four existing methods are also presented for comparison. Geptop* is a sequence composition based predictor presented along with Geptop [Wei+13].

	Our method		Liu et al.	Palaniappan and Mukherjee	Geptop (homology)	Geptop* (Composition)
Training on (No. of species)	14		30	14	18	18
	Random Forest	SVM	SVM	SVM	Score based	Score based
AB	0.81	0.83	0.75	0.74	0.85	0.79
BS	0.84	0.84	0.77	0.58	0.95	0.81
EC	0.87	0.88	0.83	0.65	0.95	0.84
FN	0.83	0.83	0.67	0.66	0.84	0.74
HI	0.75	0.77	0.54	0.46	0.57	0.59
HP	0.75	0.74	0.52	0.59	0.60	0.64
MG	0.68	0.66	0.60	0.64	0.72	0.56
MP	0.75	0.74	0.64	0.61	0.87	0.76
MT	0.80	0.77	0.70	0.49	0.73	0.77
PA	0.80	0.80	0.65	0.66	0.80	0.79
SA	0.88	0.90	0.81	0.66	0.84	0.86
SA2	0.86	0.85	0.80	-	0.88	0.83
SE	0.86	0.86	0.69	-	0.95	0.86
ST	0.81	0.79	0.84	0.53	0.71	0.69
VC	0.75	0.72	0.69	-	0.61	0.72
Avg	**0.80**	**0.80**	**0.70**	**0.61**	**0.79**	**0.75**

of 0.88 was obtained. Again, in comparison with Ning et al. [Nin+14] (0.82 AUC) and Palaniappan and Mukherjee [PM11] (0.8 AUC), our method is superior.

5.7.3 Cross-Taxonomic Predictions

Palaniappan and Mukherjee [PM11] tested the generalization ability of their classifiers across taxonomic boundaries. We made a similar assessment on our classifier at higher taxonomic level. Species belonging to a similar taxonomic order are grouped together (see Fig. 5.10) and cross-taxon and leave-one-taxon-out tests were made. The four taxonomic orders are Bacillales (BS, SA, and SA2), Enterobacterales (EC, SE, and ST), Mycoplasmatales (MG and MP), and Pseudomonadales (AB and PA). Species without a taxonomic pair were left out of this taxonomic analysis. The cross-taxonomic results are depicted in Fig. 5.12. The cross-taxonomic results are as good as the cross-organism counterparts. For example, the prediction of EC using BS yielded 0.86 AUC score and predicting Enterobacterales using Bacillales also yielded 0.85. In the leave-one-taxon-out setting, very accurate results were obtained. For Bacillales and Enterobacterales the average AUC scores were 0.85 whereas Mycoplasmatales and Pseudomonadales had 0.78 and 0.80, respectively. In comparison to Palaniappan and Mukherjee, our classifier produced an outstanding performance (Fig. 5.13).

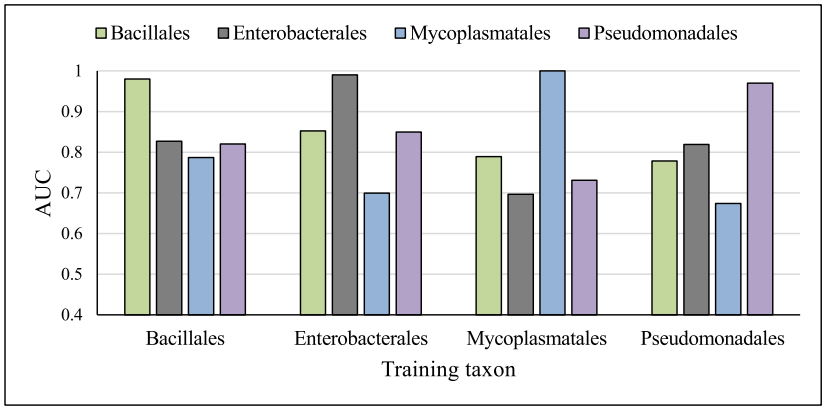

Fig. 5.12: Cross-taxon prediction results. A model trained on each taxon is used to predict the essential genes of the four taxon groups (including self-test).

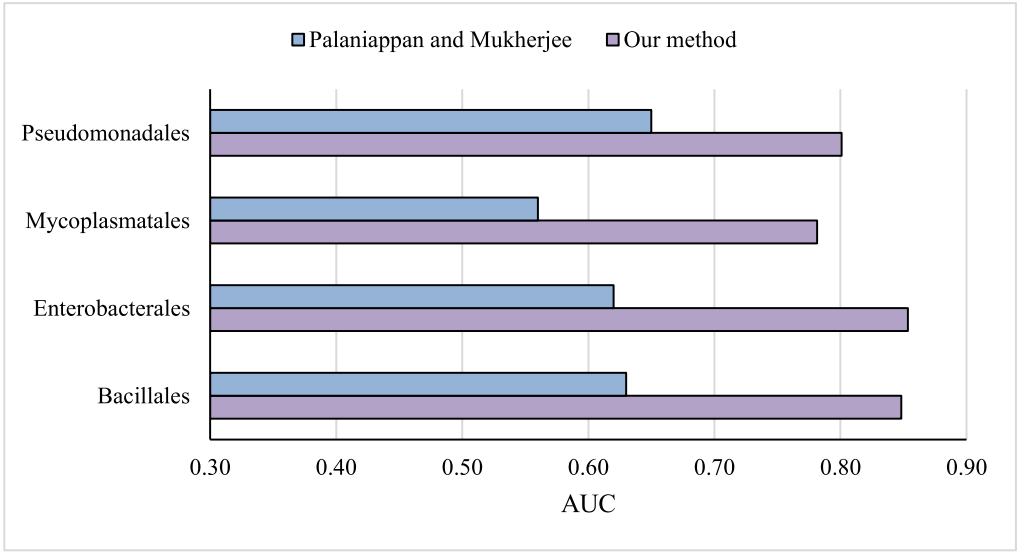

Fig. 5.13: Leave-one-taxon out predictions of our method and an existing method [PM11].

5.8 Essential Gene Prediction in Archaea

Methanococcus maripaludis (MM) is the only archaeon whose EGs and NEGs are available in DEG, generated experimentally by Sarimento et al. [Sar+13]. We trained the RF classifier using 80 % of the genes, which are randomly selected, and predicted the remaining 20 %. The ROC curves of the five feature groups along with their combination are shown in Fig. 5.14. Using all features, an average AUC score of 0.73 was obtained, which is good but smaller than the values obtained for most of the bacteria. This can be due to the reduced quality of the data. Sarimento et al. [Sar+13] could not confidently specify weather 419 genes are essential or non-essential. Hence, most of these genes are regarded as NEGs.

We further predicted the EGs of MM using the known essential and non-essential genes of the bacteria EC and BS. The achieved AUC scores were not satisfactory, 0.59

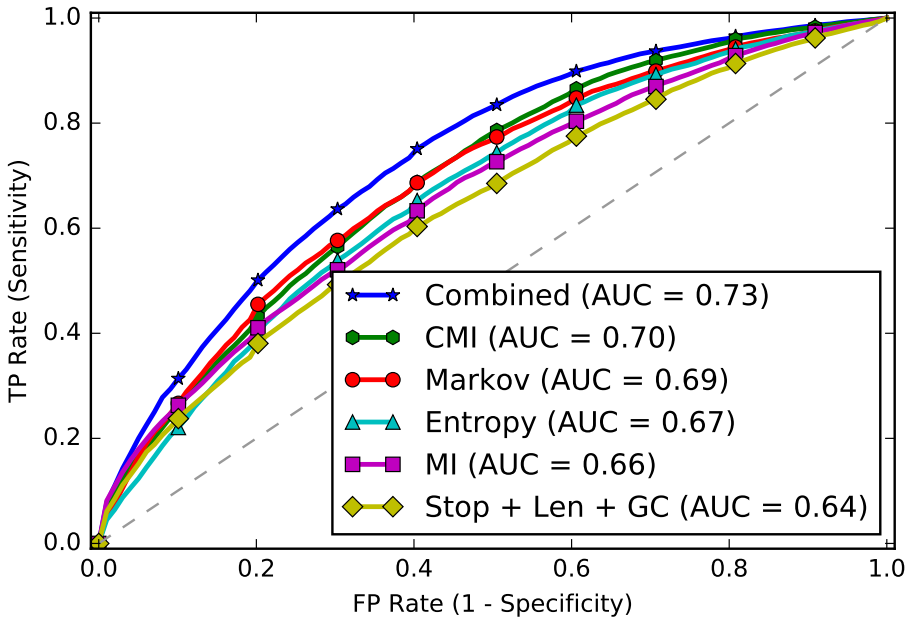

Fig. 5.14: The average ROC curves of EG prediction in *Methanococcus maripaludis*. The estimated Markov order is 6.

using EC and 0.64 using BS. The decline in performance is expected because of the inherent differences in the genetic makeup between bacteria and archaea.

5.9 Essential Gene Prediction in Eukaryotes

Genome-scale gene deletion experiments and systematic screens using RNA interference (RNAi) were applied to determine the EGs of relatively simpler eukaryotes, such as yeast and *C. elegans*. However, the RNAi screening technique has not been successful in mammals [Fra15]. Besides, EGs in higher species can only be identified in connection to the indispensability in specific cell types, typically tumor-specific EGs. The introduction of the CRISPR-Cas9 genome editing method has enabled the identification of human EGs [Har+15; Blo+15; Wan+15]. Using the annotated gene sequences provided by the experimental methods, computational predictions of EGs in yeast, human, fruit fly, worm, and mouse were performed.

Prediction of EGs in *Schizosaccharomyces pombe*

The fission yeast *Schizosaccharomyces pombe* is regarded as a very important model organism for the study of eukaryotic molecular and cellular biology [ZL95]. According to DEG, It has 1260 essential and 3573 non-essential genes. The Random Forest classifier was trained using 80 % of the data and is tested on the remaining 20 %,

performing 50-fold Monte Carlo cross-validation steps. The average ROC curve is shown in Fig. 5.15. A very good average AUC score of 0.86 was obtained.

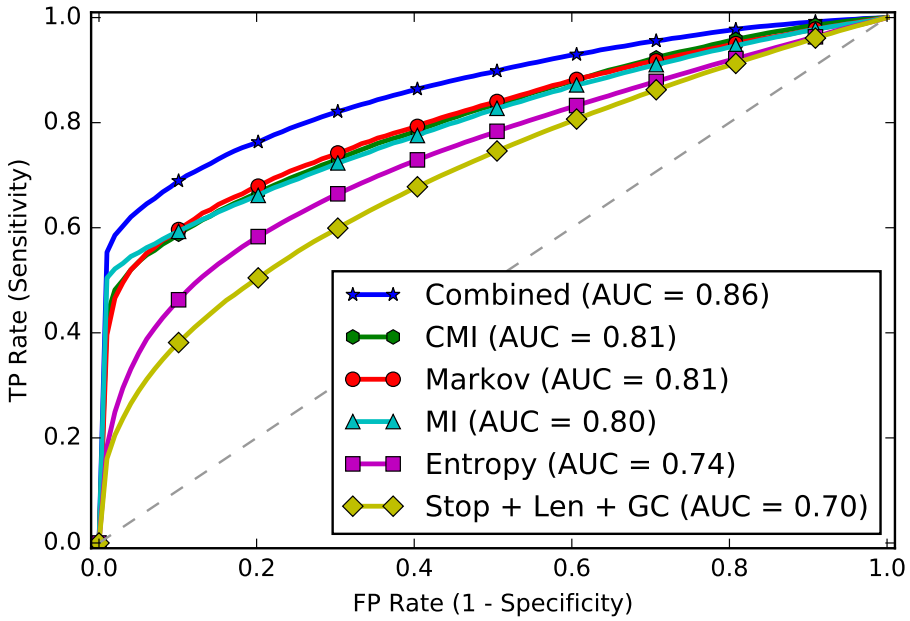

Fig. 5.15: The average ROC curves of EG prediction in *Schizosaccharomyces pombe* (fission yeast).

Prediction of EGs in *Homo sapiens*

Experimental studies in human gene essentiality are performed with a purpose of identifying the EGs in different cell lines. Mostly, gene essentiality is determined in relation to the proliferation and viability in various human cancer cell lines, such as ovarian, colon, and chronic myeloid leukemia (CML). Hence, the characterized set of EGs is cell-specific and does not indicate essentiality in all cell types [BA15]. In the OGEE database, 18 data sets from 7 separate studies are provided. Some of the data sets are small scale and cover only a limited portion of the genes while some of the studies are genome-wide, investigating around 20,000 genes. Since a gene can be designated essential in one data set and non-essential in the other, OGEE adopts a third category named conditionally EGs (CEG). A specific gene was covered by up to 11 data sets and if all the studies do not agree on the essentiality, it is labeled as CEG.

Out of 21,529 genes, 182 are EGs, 6,985 are CEGs, and 14,362 are NEGs. To categorize the CEGs as essential or non-essential, we adopted a simple majority voting scheme. That is, a CEG is regarded as EG if it is essential in a majority of cell lines. This resulted in 1,632 EGs. We trained RF classifiers for each of the five feature sets and their combinations. The data was split into 80 % for training and

20 % for testing and 100 trials with different sets were performed. The ROC curves are presented in Fig. 5.16. Using all the available features, a decent AUC of 0.76 was obtained. Similar to the prokaryotic EG prediction, the mutual information features provided the largest contribution. However, the non-IT features were better than the entropy and Markov features. Guo et al. [Guo+17] predicted human EGs utilizing only intrinsic sequence information and obtained an AUC score of 0.89. However, the approach used by Guo et al. is somewhat different. They prepared the positive data set based on a majority decision on essentiality in 11 cell lines. A gene is considered essential if it essential in more than 6 cell lines. Otherwise, the gene is totally discarded rather than taking it as a negative sample. Afterwards, they obtained 1,516 EGs and 10,499 NEGs. Considering that even the CRISPR-based experimental method proposed by Wang et al. [Wan+15] yielded a 0.78 AUC score when validated using known EGs of a yeast genome, our results are good.

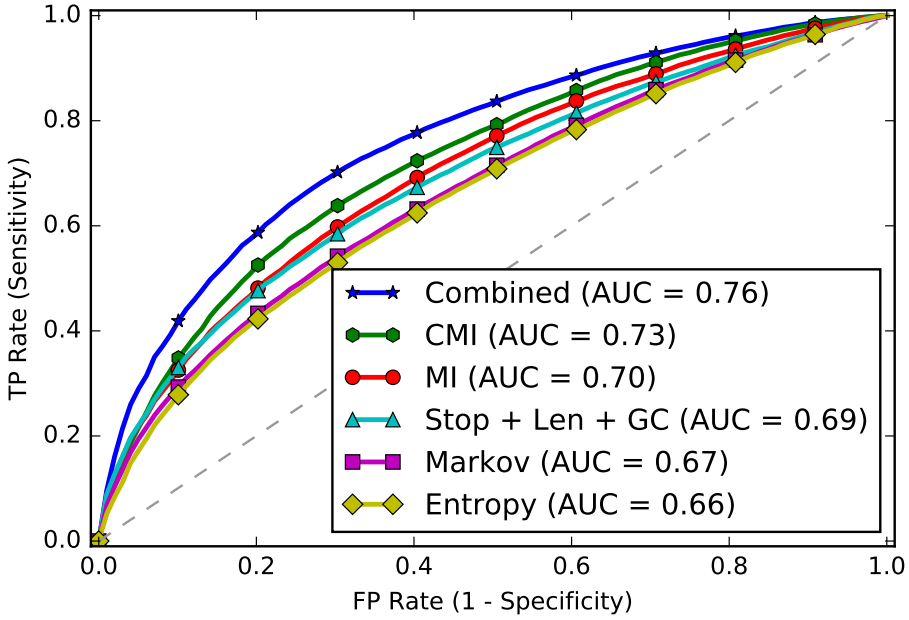

Fig. 5.16: The average ROC curves of *H. sapiens* EG prediction.

Prediction of EGs in *Drosophila melanogaster*

The other commonly used model organism in developmental biology studies is *Drosophila melanogaster* (DRO). Although there are two data sets providing essentiality annotations for DRO, one of them tested only 437 genes. Hence, we only took the large-scale results obtained using double stranded RNAi on embryonic hemocyte (blood cell) lines. Among the 13781 tested genes, only 267 were found to be EGs. This is the smallest reported percentage of EGs among eukaryotic species. The prediction results are presented in Fig. 5.17. The combined features yielded a

very high AUC score (0.87) and the performances of the individual feature groups are also satisfactory.

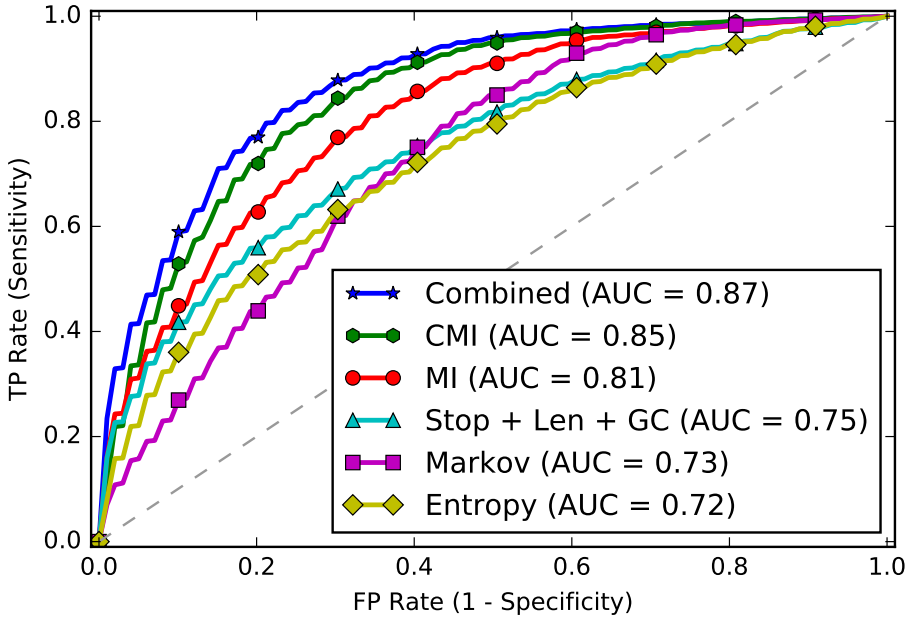

Fig. 5.17: The average ROC curves of *Drosophila melanogaster* EG prediction.

Prediction of EGs in *Caenorhabditis elegans*

Next, we tested the ability of our method to classify EGs of *Caenorhabditis elegans* (CEL). The CEL dataset in the OGEE database contains 742 EGs and 10704 NEGs. Intra-organism predictions were performed and the ROC curves of 100 iterations are presented in Fig. 5.18. The prediction scores by all of the feature sets were very high (AUC ≥ 0.74). Taking all the features, an AUC score of 0.85 was obtained.

Prediction of EGs in *Mus musculus*

The *Mus musculus* (MUS) dataset has 4289 EGs and 4592 NEGs. Through a similar validation procedure, we predicted the EGs. The ROC curves are shown in Fig. 5.19. The combined AUC score was relatively low (0.66). This could be either due to the uncharacteristically almost equal number of EGs and NEGs or the poor predictive power of the used features. We have roughly checked the quality of the annotations by comparing it to another dataset provided by Dickinson et al. [Dic+16]. In this dataset, through a developmental gene knockout study performed on 1751 genes, 410 were found to be essential genes and 1143 were non-essential genes. Roughly

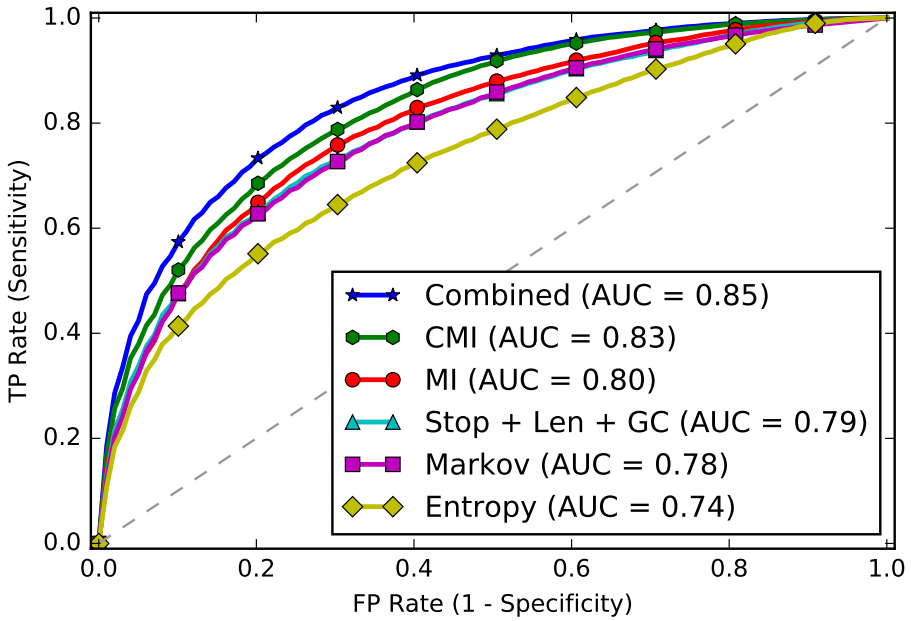

Fig. 5.18: The average ROC curves of EG predictions in *C. elegans* (worm).

124 genes which were regarded as essential in the OGEE dataset are non-essential. This shows that the list of EGs does not reflect the absolute minimal set. Moreover, the mouse genome contains around 23,000 protein-coding genes but only essentiality annotations for 8881 were provided.

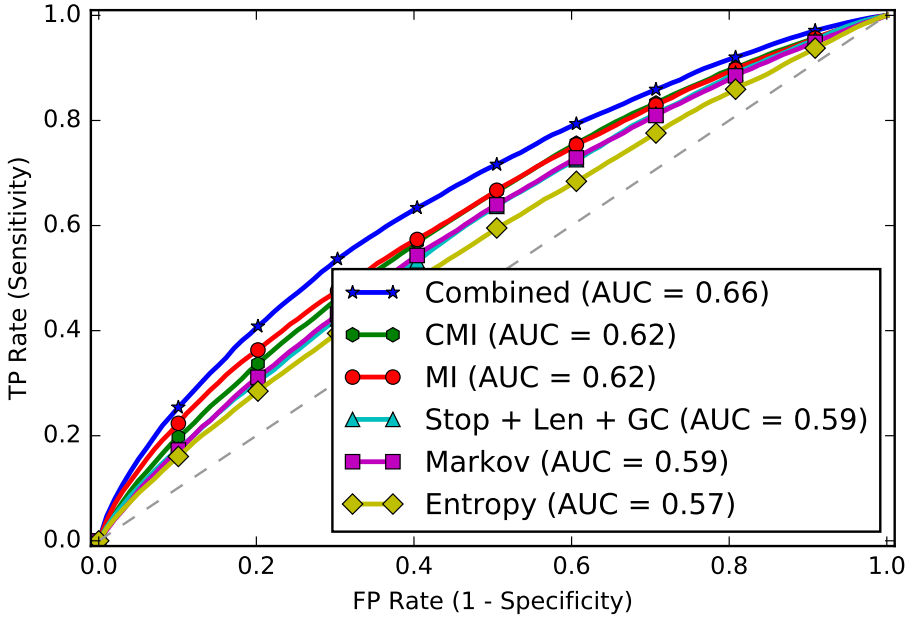

Fig. 5.19: The average ROC curves of EG prediction in *Mus musculus*.

Cross-organism prediction between Eukaryotes

To test the transferability of EG annotations among the eukaryotic species, cross-organism predictions were performed by training classifiers using DRO, CEL, and HSA data sets. Prediction of CEL using models trained on DRO and HSA yielded an average AUC score of 0.72 and 0.73, respectively. However, our method failed to predict human EGs using both DRO and CEL models. Prediction of DRO EGs using a classifier trained on HSA was also not possible (AUC = 0.47), while an AUC score of 0.68 was obtained using CEL. All in all, cross-organismic EG predictions in eukaryotes were not as successful as they were in bacteria.

5.10 Summary

Computational prediction of EGs in the three domains of life, Bacteria, Archaea, and Eukarya, was performed. We used novel information-theoretic features derived exclusively from the DNA sequences of the genes. We analyzed the prediction performances of four feature sets. The feature sets are based on entropy, mutual information, conditional mutual information, and Markov models. Other commonly used, sequence-based, non-information-theoretic features related to stop codon usage, GC content, and protein length were also additionally utilized. Two powerful machine learning algorithms (SVM and Random Forest) were used for the classification task. Performance evaluations were carried out considering both intra- and cross-organism predictions. Other computational methods depend on the availability of experimental data, such as gene-expression and protein interaction networks. Since these data are not available for newly sequenced and under-/un-studied organisms, their application is limited to well-studied model organisms. EG predictors based on homology and functional domain information, despite being sequence-based, require a computationally expensive sequence alignment and data base search. Hence, relying only on the sequence information, our method provides a very simple and effective prediction of EGs applicable to any organism.

Applied to bacteria, our proposed method yielded a very good prediction performance. The prediction accuracy is better than most existing predictors which rely only on sequence information and it is as good as the methods using network topology, homology, and gene-expression in addition to sequence-based features. Extensive performance evaluation using different setups were performed on selected 15 bacterial species. In intra-organism predictions, very high AUC scores ranging from 0.73 to 0.9 were obtained. In cross-organism pairwise predictions, the vast majority of the results are very good. Scores as high as 0.92 and mean AUC of 0.75 were achieved. However, due to factors such as high evolutionary distance,

different lifestyles, growth conditions, and phenotypes there were very few poor results [Den+11]. Based on the results, for future blind predictions, we suggest using one of the well-studied bacteria, such as *B. subtilis* and *E. coli* (the essentiality annotations are of high quality). In addition, 5-fold cross-validation and leave-one-species-out experiments have yielded average AUC scores of 0.88 and 0.80, respectively. Furthermore, our model performed very well at higher taxonomic ranks (order). An average score of 0.82 in cross-taxon and 0.78 in leave-one-taxon-out predictions, which is significantly superior to the previously published result having average AUC of 0.62.

In the archaeon *Methanococcus maripaludis*, an average AUC score of 0.77 was obtained. In eukaryotes, we studied the possible prediction of EGs in yeast (*Schizosaccharomyces pombe*), humans (*Homo sapiens*), fruit fly (*Drosophila melanogaster*), mouse (*Mus musculus*), and worm (*Caenorhabditis elegans*). EG predictions in *Schizosaccharomyces pombe* yelded a very high AUC score of 0.86. In *H. sapiens*, the gene essentiality data is based on cancer cell lines. Hence, there are some genes whose essentiality is conditional, i.e., essential in one cell line and non-essential in another. Our model classified the human genes with an AUC score of 0.76, which is a decent performance but not as good as the method proposed by Guo et al. [Guo+17]. EGs of *D. melanogaster* and *C. elegans* were predicted with a very good accuracy, AUC scores of 0.87 and 0.85, respectively. However, in *M. musculus*, the prediction result was worse, AUC 0.66. We suspect that the reason for this is the reduced quality of the annotations in the data. Almost 50 % of the 8881 investigated genes were EGs, which is inconsistent to the percentage of essential genes in other species. Cross-organism predictions among eukaryotes were also not as successful as they were between bacteria. Only *C. elegans* was predicted with a good accuracy using the other species.

To conclude, we demonstrated that information-theoretic features, which can be easily derived from the genetic sequences, allow the classification of EGs and NEGs in both prokaryotes and eukaryotes.

6 Conclusion

In this thesis, we addressed three problems in computational biology using concepts from information theory and communication engineering. The first one is a channel model for the transfer of genetic information from DNA to RNA to proteins. The second topic we studied is the digital information content of the bacterial genomes together with the thermodynamic stability and spatial organization of functional groups of genes. The third topic deals with the prediction of essential and non-essential genes. We proposed a simple and reliable computational method for the identification of essential genes.

We started by studying two aspects of the empirical codon mutation (ECM) matrix, which shows substitution rates between codons. In the first part, the matrix was assumed to model a communication channel. The required rate defined by the amino acid distribution in a set of species is 4.1875 bit. The mutual information between the input and output of the ECM "channel" is well below what is required. Hence, we introduced an exponent to the matrix and asked the question, at what exponent the required rate would still be satisfied? The exponent was found to be 0.26. The exponent corresponds to a mutation rate of 29 %. The channel capacity at the optimal exponent was 4.2586 bit, which is very close to the one found using the biological codon distribution. This shows that the biological codon distribution is optimally "chosen" by nature. The obtained result further implies that a reliable communication through the genetic "channel" is only possible if the mutation rate of the channel is less than 29 %. In the second part, we compared the mutation probabilities between codons to the chemical properties of the resulting amino acids. A dimension reduction technique called classical multidimensional scaling was used to reduce the 64×64 ECM matrix and the 20×20 chemical distance matrix. A comparison between the two-dimensional representations of the matrices revealed that, as expected, most mutations are to synonymous codons or to chemically similar amino acids. However, we also found some inconsistencies showing highly probable mutations leading to a chemically different amino acids and vice versa. Although some of the inconsistencies can be explained by studying the chemical properties in details, it could also be that a further protection mechanism is involved to minimize mutations between codons having larger chemical distance. In a future work, better dimension reduction and clustering methods, which may be better suited for our data, can be applied.

After that, the digital information is measured employing Shannon entropy and is analyzed in connection to the analog type of information characterizing the physicochemical properties of the DNA polymer. Additionally, the spatial organization of selected functional classes of genes, anabolic, catabolic, aerobic, anaerobic, were studied. In this thesis, analog information was associated with thermodynamic stability and measured using Gibbs entropy. This integrative analysis reveals the genomic sequence organization and DNA encodings with respect to certain functional constraints. The relationships between Shannon and Gibbs entropies were investigated in four bacterial genomes, *E. coli*, *B. subtilis*, *S. typhimurium*, and *S. coelicolor*. In *E. coli*, Shannon and Gibbs entropies are mostly anti-correlated, especially around the terminus. The terminus region is thermodynamically less stable, i.e., it is relatively AT-rich. This is achieved by a selective usage of synonymous codons and/or amino acids. If a certain amino acid is encoded around the terminus region, among the codons encoding the amino acid, the AT-rich ones are preferred. In addition, from the significantly high correlation between the distribution of anabolic genes and Gibbs entropy, it seems that anabolic genes are preferentially encoded by sequences of high thermodynamic stability whereas catabolic genes prefer DNA locations with low thermodynamic stability. The distribution of aerobic genes is also highly correlated to the Gibbs entropy profiles. Since *S. typhimurium* is closely related to *E. coli*, it shows similar properties. In the evolutionarily more distant *B. subtilis*, however, the relationships between the entropies are different. This leads us to the conclusion that the digital and analog information of the DNA are coupled and the relationships depend on the lifestyle and type of the bacterium. Furthermore, we demonstrated the possibility of using Gibbs entropy to the detection of coding and non-coding regions. However, this was only a preliminary study and should be investigated more in a future work.

Finally, we presented a novel gene essentiality prediction method. Information-theoretic quantities such as entropy, mutual information, Markov models, and Kullback-Leibler divergence have been used as features to show structural and compositional properties which can highlight differences between essential and non-essential genes. Two supervised machine learning algorithms (Support Vector Machines and Random Forests) were employed to perform the classification task. We showed the applicability of our method in both prokaryotes and eukaryotes. Predictions were performed among and between 15 bacteria, 1 archaeon, and 4 eukaryotic species. The performances of the predictors were analyzed in both intra- and cross-organism/taxon settings. We demonstrate that gene essentiality annotations can be transferred between both closely and distantly related organisms with a reasonable accuracy. The obtained results were better than previously published sequence-based predictors. Our method is also as good as those essential gene prediction methods using additional features related to network topology, gene-expression, and homology. Considering the complexity of the concept of gene essentiality (it is

context dependent and involves the interaction of multiple proteins and pathways), the obtained results using only sequence information are excellent. Although the experimental data such as gene/protein interaction networks and gene-expression can provide very important information in identifying essential genes, the data is mostly unavailable for understudied or unstudied organisms. Furthermore, most bacteria among the total diversity are uncultured, i.e., cannot be grown in synthetic media. Hence, as our prediction method relies solely on the genomic sequence, it can be easily applied to any species. In addition, we believe that the proposed usage of information-theoretic quantities as features can be applied in other classification problems. In [You+17], in a collaboration with a group in Israel, we used our information-theoretic method successfully for pre-miRNA detection. Similarly, in a future work, our approach can be used in other applications.

Bibliography

[Akh+13] S. Akhter, B. A. Bailey, P. Salamon, R. K. Aziz, and R. A. Edwards. “Applying Shannon’s information theory to bacterial and phage genomes and metagenomes”. In: *Scientific reports* 3 (2013) (cit. on p. 50).

[AL09] M. L. Acencio and N. Lemke. “Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information”. In: *BMC bioinformatics* 10.1 (2009), p. 1 (cit. on p. 70).

[Alb+13] B. Alberts, D. Bray, K. Hopkin, et al. *Essential cell biology*. Garland Science, 2013 (cit. on p. 23).

[All07] L. Allison. *Fundamental Molecular Biology*. Wiley, 2007 (cit. on pp. 18, 21).

[Ari72] S. Arimoto. “An algorithm for computing the capacity of arbitrary discrete memoryless channels”. In: *Information Theory, IEEE Transactions on* 18.1 (1972), pp. 14–20 (cit. on p. 35).

[BA15] C. Boone and B. J. Andrews. “The indispensable genome”. In: *Science* 350.6264 (2015), pp. 1028–1029 (cit. on p. 98).

[Bat08] G. Battail. “Genomic error-correcting codes in the living world”. In: *Biosemiotics* 1.2 (2008), pp. 221–238 (cit. on p. 1).

[Bat97] G. Battail. “Does information theory explain biological evolution?” In: *EPL (Europhysics Letters)* 40.3 (1997), p. 343 (cit. on p. 1).

[Bau+08] M. Bauer, S. M. Schuster, and K. Sayood. “The average mutual information profile as a genomic signature”. In: *BMC bioinformatics* 9.1 (2008), p. 1 (cit. on p. 72).

[Ben+02] S. D. Bentley, K. F. Chater, A.-M. Cerdeno-Tarraga, et al. “Complete genome sequence of the model actinomycete Streptomyces coelicolor A3 (2)”. In: *Nature* 417.6885 (2002), pp. 141–147 (cit. on p. 68).

[Ber+07] M. R. Berthold, N. Cebron, F. Dill, et al. “KNIME: The Konstanz Information Miner”. In: *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. Springer, 2007 (cit. on p. 87).

[Ber+10] M. Berger, A. Farcas, M. Geertz, et al. “Coordination of genomic structure and transcription by the main bacterial nucleoid-associated protein HU”. In: *EMBO Rep.* 11.1 (2010), pp. 59–64 (cit. on p. 64).

[BG05] I. Borg and P. J. Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005 (cit. on p. 41).

[BG07] L. Bofkin and N. Goldman. “Variation in evolutionary processes at different codon positions”. In: *Molecular biology and evolution* 24.2 (2007), pp. 513–521 (cit. on p. 30).

[Bis06] C. M. Bishop. “Pattern Recognition and Machine Learning (Information Science and Statistics) Springer-Verlag New York”. In: *Inc. Secaucus, NJ, USA* (2006) (cit. on p. 73).

[Bla72] R. Blahut. “Computation of channel capacity and rate-distortion functions”. In: *Information Theory, IEEE Transactions on* 18.4 (1972), pp. 460–473 (cit. on pp. 35, 36).

[Blo+15] V. A. Blomen, P. Májek, L. T. Jae, et al. “Gene essentiality and synthetic lethality in haploid human cells”. In: *Science* 350.6264 (2015), pp. 1092–1096 (cit. on pp. 69, 97).

[Bos+92] B. E. Boser, I. M. Guyon, and V. N. Vapnik. “A training algorithm for optimal margin classifiers”. In: *Proceedings of the fifth annual workshop on Computational learning theory*. ACM. 1992, pp. 144–152 (cit. on pp. 73, 76).

[Bot+94] L. Bottou, C. Cortes, J. S. Denker, et al. “Comparison of classifier methods: a case study in handwritten digit recognition”. In: *Pattern Recognition, 1994. Vol. 2-Conference B: Computer Vision & Image Processing., Proceedings of the 12th IAPR International. Conference on*. Vol. 2. IEEE. 1994, pp. 77–82 (cit. on p. 73).

[BP10] J. E. Burgess and B. I. Pletschke. “ANAEROBIC AND AEROBIC RESPIRATION”. In: *MEDICAL AND HEALTH SCIENCES-Volume XV* (2010), p. 78 (cit. on p. 63).

[Bre+84] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. “Classification and regression trees”. In: (1984) (cit. on p. 76).

[Bre+86] K. J. Breslauer, R. Frank, H. Blöcker, and L. A. Marky. “Predicting DNA duplex stability from the base sequence”. In: *Proceedings of the National Academy of Sciences* 83.11 (1986), pp. 3746–3750 (cit. on p. 50).

[Bre01] L Breiman. “Random Forests Machine Learning. 45: 5–32”. In: *View Article PubMed/NCBI Google Scholar* (2001) (cit. on p. 80).

[Bre96] L. Breiman. “Bagging predictors”. In: *Machine learning* 24.2 (1996), pp. 123–140 (cit. on p. 80).

[BS99] R. Bowley and M. Sánchez. *Introductory statistical mechanics*. Oxford: Clarendon Press, 1999 (cit. on p. 51).

[Bur98] C. J. Burges. “A tutorial on support vector machines for pattern recognition”. In: *Data mining and knowledge discovery* 2.2 (1998), pp. 121–167 (cit. on p. 73).

[CA05] L. M. Cullen and G. M. Arndt. “Genome-wide screening for gene function using RNAi in mammalian cells”. In: *Immunology and cell biology* 83.3 (2005), pp. 217–223 (cit. on p. 69).

[Cap+04] E. Capriotti, P. Fariselli, I. Rossi, and R. Casadio. “A Shannon entropy-based filter detects high-quality profile–profile alignments in searches for remote homologues”. In: *PROTEINS: Structure, Function, and Bioinformatics* 54.2 (2004), pp. 351–360 (cit. on p. 50).

[CB02] G. Casella and R. L. Berger. *Statistical inference*. Vol. 2. Duxbury Pacific Grove, CA, 2002 (cit. on p. 11).

[CC76] L. Clarke and J. Carbon. “A colony bank containing synthetic CoI EI hybrid plasmids representative of the entire E. coli genome”. In: *Cell* 9.1 (1976), pp. 91–99 (cit. on p. 86).

[Cha+05] C.-H. Chang, L.-C. Hsieh, T.-Y. Chen, et al. “Shannon information in complete genomes”. In: *Journal of bioinformatics and computational biology* 3.03 (2005), pp. 587–608 (cit. on p. 50).

[Che+11] W.-H. Chen, P. Minguez, M. J. Lercher, and P. Bork. “OGEE: an online gene essentiality database”. In: *Nucleic acids research* 40.D1 (2011), pp. D901–D906 (cit. on p. 81).

[Che+12] W.-H. Chen, P. Minguez, M. J. Lercher, and P. Bork. “OGEE: an online gene essentiality database”. In: *Nucleic acids research* 40.D1 (2012), pp. D901–D906 (cit. on p. 70).

[Che+13] J. Cheng, W. Wu, Y. Zhang, et al. “A new computational strategy for predicting essential genes”. In: *BMC genomics* 14.1 (2013), p. 910 (cit. on pp. 70, 71, 93, 94).

[Che+14] J. Cheng, Z. Xu, W. Wu, et al. “Training set selection for the prediction of essential genes”. In: *PloS one* 9.1 (2014), e86805 (cit. on pp. 70, 71, 92).

[Che+15] L. Chen, X. Ge, and P. Xu. “Identifying essential Streptococcus sanguinis genes using genome-wide deletion mutation”. In: *Gene Essentiality: Methods and Protocols* (2015), pp. 15–23 (cit. on p. 69).

[CL02] A. F. Chalker and R. D. Lunsford. “Rational identification of new antibacterial drug targets that are essential for viability using a genomics-based approach”. In: *Pharmacology & therapeutics* 95.1 (2002), pp. 1–20 (cit. on p. 69).

[CP02] C. Condon and H. Putzer. “The phylogenetic distribution of bacterial ribonucleases”. In: *Nucleic Acids Research* 30.24 (2002), pp. 5339–5346 (cit. on p. 92).

[Cri58] F. H. Crick. “On protein synthesis.” In: *Symposia of the Society for Experimental Biology*. Vol. 12. 1958, p. 138 (cit. on p. 20).

[CS12] G. M. Cannarozzi and A. Schneider. *Codon evolution: mechanisms and models*. Oxford University Press, 2012 (cit. on p. 30).

[CT91] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. New York, NY, USA: Wiley-Interscience, 1991 (cit. on pp. 7, 14, 38, 51, 82).

[CX05] Y. Chen and D. Xu. “Understanding protein dispensability through machine-learning analysis of high-throughput data”. In: *Bioinformatics* 21.5 (2005), pp. 575–581 (cit. on p. 72).

[D’E+09] M. A. D’Elia, M. P. Pereira, and E. D. Brown. “Are essential genes really essential?” In: *Trends in microbiology* 17.10 (2009), pp. 433–438 (cit. on p. 69).

[Daw+06] Z. Dawy, B. Goebel, J. Hagenauer, et al. “Gene mapping and marker clustering using Shannon’s mutual information”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 3.1 (2006), p. 47 (cit. on p. 1).

[Day+78] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt. “A model of evolutionary change in proteins”. In: *Natl Biomedical Research Foundation, Washington, D.C.* 5 (1978), pp. 345–352 (cit. on p. 29).

[DD05a] D. Dalevi and D. Dubhashi. “The Peres-Shields order estimator for fixed and variable length markov models with applications to DNA sequence similarity”. In: *International Workshop on Algorithms in Bioinformatics*. Springer. 2005, pp. 291–302 (cit. on p. 83).

[DD05b] D. Dalevi and D. Dubhashi. “The Peres-Shields order estimator for fixed and variable length markov models with applications to DNA sequence similarity”. In: *International Workshop on Algorithms in Bioinformatics*. Springer. 2005, pp. 291–302 (cit. on p. 84).

[Den+11] J. Deng, L. Deng, S. Su, et al. “Investigating the predictability of essential genes across distantly related organisms using an integrative approach”. In: *Nucleic acids research* 39.3 (2011), pp. 795–807 (cit. on pp. 70, 71, 92, 93, 103).

[DFP07] A. Doron-Faigenboim and T. Pupko. “A combined empirical and mechanistic codon model”. In: *Molecular biology and evolution* 24.2 (2007), pp. 388–397 (cit. on p. 31).

[Dic+16] M. E. Dickinson, A. M. Flenniken, X. Ji, et al. “High-throughput discovery of novel developmental phenotypes”. In: *Nature* 537.7621 (2016), p. 508 (cit. on p. 100).

[DM03] S. V. Date and E. M. Marcotte. “Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages”. In: *Nature biotechnology* 21.9 (2003), pp. 1055–1062 (cit. on p. 72).

[Dur+98] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998 (cit. on p. 14).

[EG06] W. J. Ewens and G. R. Grant. “Statistical Methods in Bioinformatics”. In: *Journal of Applied Statistics* 33.8 (2006) (cit. on p. 14).

[Fel81] J. Felsenstein. “Evolutionary trees from DNA sequences: a maximum likelihood approach”. In: *Journal of molecular evolution* 17.6 (1981), pp. 368–376 (cit. on p. 28).

[Fit67] W. M. Fitch. “Evidence suggesting a non-random character to nucleotide replacements in naturally occurring mutations”. In: *Journal of molecular biology* 26.3 (1967), pp. 499–507 (cit. on p. 26).

[Fra15] A. Fraser. “Essential human genes”. In: *Cell systems* 1.6 (2015), pp. 381–382 (cit. on p. 97).

[Gam54] G. Gamow. “Possible relation between deoxyribonucleic acid and protein structures”. In: *Nature* 173.4398 (1954), pp. 318–318 (cit. on p. 1).

[Gat72] L. L. Gatlin. “Information theory and the living system”. In: (1972) (cit. on pp. 1, 31, 32).

[Gia+02] G. Giaever, A. M. Chu, L. Ni, et al. “Functional profiling of the Saccharomyces cerevisiae genome”. In: *nature* 418.6896 (2002), pp. 387–391 (cit. on p. 69).

[Gon+11] L. Gong, N. Bouaynaya, and D. Schonfeld. “Information-theoretic model of evolution over protein communication channel”. In: *Computational Biology and Bioinformatics, IEEE/ACM Transactions on* 8.1 (2011), pp. 143–151 (cit. on pp. 1, 31, 33, 34).

[Gon+94] G. H. Gonnet, M. A. Cohen, and S. A. Benner. “Analysis of amino acid substitution during divergent evolution: the 400 by 400 dipeptide substitution matrix”. In: *Biochemical and biophysical research communications* 199.2 (1994), pp. 489–496 (cit. on p. 30).

[Gra74] R Grantham. “Amino acid difference formula to help explain protein evolution”. In: *Science* 185.4154 (1974), pp. 862–864 (cit. on p. 40).

[Gro+00] I. Grosse, H. Herzel, S. V. Buldyrev, and H. E. Stanley. “Species independence of mutual information in coding and noncoding DNA”. In: *Physical Review E* 61.5 (2000), p. 5624 (cit. on pp. 2, 72).

[Guo+03] F.-B. Guo, H.-Y. Ou, and C.-T. Zhang. “ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes”. In: *Nucleic acids research* 31.6 (2003), pp. 1780–1789 (cit. on p. 72).

[Guo+17] F.-B. Guo, C. Dong, H.-L. Hua, et al. “Accurate prediction of human essential genes using only nucleotide composition and association information”. In: *Bioinformatics* 33.12 (2017), pp. 1758–1764 (cit. on pp. 71, 72, 99, 103).

[GY94] N. Goldman and Z. Yang. “A codon-based model of nucleotide substitution for protein-coding DNA sequences.” In: *Molecular biology and evolution* 11.5 (1994), pp. 725–736 (cit. on p. 30).

[HA09] P. G. Higgs and T. K. Attwood. *Bioinformatics and molecular evolution*. John Wiley & Sons, 2009 (cit. on p. 18).

[Hag+04] J Hagenauer, Z Dawy, B Gobel, P Hanus, and J Mueller. “Genomic analysis using methods from information theory”. In: *Information Theory Workshop, 2004. IEEE*. IEEE. 2004, pp. 55–59 (cit. on pp. 50, 72).

[Han+10] P. Hanus, J. Dingel, G. Chalkidis, and J. Hagenauer. “Compression of whole genome alignments”. In: *IEEE Transactions on Information Theory* 56.2 (2010), pp. 696–705 (cit. on p. 1).

[Har+15] T. Hart, M. Chandrashekhar, M. Aregger, et al. “High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities”. In: *Cell* 163.6 (2015), pp. 1515–1526 (cit. on pp. 69, 97).

[Har05] R. C. Hardison. *Working with Molecular Genetics*. 2005 (cit. on p. 18).

[Har28] R. V. Hartley. “Transmission of information”. In: *Bell System technical journal* 7.3 (1928), pp. 535–563 (cit. on p. 7).

[Has+85] M. Hasegawa, H. Kishino, and T.-a. Yano. “Dating of the human-ape splitting by a molecular clock of mitochondrial DNA”. In: *Journal of molecular evolution* 22.2 (1985), pp. 160–174 (cit. on p. 28).

[HH09] M. Hoefnagels and M. Hoefnagels. *Biology: concepts and investigations*. McGraw-Hill Higher Education, 2009 (cit. on pp. 19, 20).

[HH92] S. Henikoff and J. Henikoff. “Amino acid substitution matrices from protein blocks”. In: *Proc. Natl. Acad. Sci. USA* 89 (1992), pp. 10915–10919 (cit. on p. 29).

[Hut+16] C. A. Hutchison, R.-Y. Chuang, V. N. Noskov, et al. “Design and synthesis of a minimal bacterial genome”. In: *Science* 351.6280 (2016), aad6253 (cit. on p. 69).

[Ita95] M. Itaya. “An estimation of minimal genome size required for life”. In: *FEBS letters* 362.3 (1995), pp. 257–260 (cit. on p. 69).

[JC69] T. H. Jukes and C. R. Cantor. “Evolution of protein molecules”. In: *Mammalian protein metabolism* 3.21 (1969), p. 132 (cit. on p. 26).

[Jeo+04] K. S. Jeong, J. Ahn, and A. B. Khodursky. “Spatial patterns of transcriptional activity in the chromosome of Escherichia coli”. In: *Genome Biol* 5.11 (2004), R86 (cit. on p. 60).

[Jon+92] D. T. Jones, W. R. Taylor, and J. M. Thornton. “The rapid generation of mutation data matrices from protein sequences”. In: *Computer applications in the biosciences: CABIOS* 8.3 (1992), pp. 275–282 (cit. on p. 30).

[Kat81] R. W. Katz. “On some criteria for estimating the order of a Markov chain”. In: *Technometrics* 23.3 (1981), pp. 243–249 (cit. on p. 83).

[Kay00] L. E. Kay. *Who wrote the book of life?: A history of the genetic code*. Stanford University Press, 2000 (cit. on p. 1).

[KB09] G Khandelwal and J Bhyravabhotla. “A phenomenological model for predicting melting temperatures of DNA sequences.” In: *PloS one* 5.8 (2009), e12433–e12433 (cit. on p. 66).

[Kha+14] G. Khandelwal, R. A. Lee, B Jayaram, and D. L. Beveridge. “A Statistical Thermodynamic Model for Investigating the Stability of DNA Sequences from Oligonucleotides to Genomes”. In: *Biophysical journal* 106.11 (2014), pp. 2465–2473 (cit. on p. 66).

[Kim80] M. Kimura. “A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences”. In: *Journal of molecular evolution* 16.2 (1980), pp. 111–120 (cit. on p. 27).

[KJ69] J. King and T. Jukes. “Non-Darwinian evolution.” In: *Science (New York, NY)* 164.3881 (1969), p. 788 (cit. on p. 35).

[KL51] S. Kullback and R. A. Leibler. “On information and sufficiency”. In: *The annals of mathematical statistics* 22.1 (1951), pp. 79–86 (cit. on pp. 8, 83).

[KM97] S. Karlin and J. Mrázek. “Compositional differences within and between eukaryotic genomes”. In: *Proceedings of the National Academy of Sciences* 94.19 (1997), pp. 10227–10232 (cit. on p. 2).

[Koo+86] H.-S. Koo, H.-M. Wu, and D. M. Crothers. “DNA bending at adenine· thymine tracts”. In: *Nature* 320.6062 (1986), pp. 501–506 (cit. on p. 50).

[Koo00] E. V. Koonin. “How many genes can make a cell: The minimal-gene-Set concept 1”. In: *Annual review of genomics and human genetics* 1.1 (2000), pp. 99–116 (cit. on p. 69).

[Kos+07] C. Kosiol, I. Holmes, and N. Goldman. “An empirical codon model for protein sequence evolution”. In: *Molecular biology and evolution* 24.7 (2007), pp. 1464–1479 (cit. on p. 31).

[Kum+15] S. Kumar, M. Vendruscolo, A. Singh, D. Kumar, and A. Samal. “Analysis of the hierarchical structure of the B. subtilis transcriptional regulatory network”. In: *Molecular BioSystems* 11.3 (2015), pp. 930–941 (cit. on p. 65).

[Lam+03] G. Lamichhane, M. Zignol, N. J. Blades, et al. “A postgenomic method for predicting essential genes at subsaturation levels of mutagenesis: application to Mycobacterium tuberculosis”. In: *Proceedings of the National Academy of Sciences* 100.12 (2003), pp. 7213–7218 (cit. on p. 69).

[LB02] R.-H. Li and G. G. Belford. “Instability of decision tree classification algorithms”. In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2002, pp. 570–575 (cit. on p. 80).

[Li+17] Y. Li, Y. Lv, X. Li, W. Xiao, and C. Li. “Sequence comparison and essential gene identification with new inter-nucleotide distance sequences”. In: *Journal of Theoretical Biology* 418 (2017), pp. 84–93 (cit. on pp. 71, 91).

[Lid20] G. J. Lidstone. “Note on the general case of the Bayes-Laplace formula for inductive or a posteriori probabilities”. In: *Transactions of the Faculty of Actuaries* 8.182-192 (1920), p. 13 (cit. on p. 84).

[Liu+17] X. Liu, B.-J. Wang, L. Xu, H.-L. Tang, and G.-Q. Xu. “Selection of key sequence-based features for prediction of essential genes in 31 diverse bacterial species”. In: *PloS one* 12.3 (2017), e0174638 (cit. on pp. 70, 71, 93, 94).

[Llo+15] J. P. Lloyd, A. E. Seddon, G. D. Moghe, M. C. Simenc, and S.-H. Shiu. “Characteristics of plant essential genes allow for within-and between-species prediction of lethal mutant phenotypes”. In: *The Plant Cell* 27.8 (2015), pp. 2133–2147 (cit. on p. 72).

[Lu+14] Y. Lu, J. Deng, J. C. Rhodes, H. Lu, and L. J. Lu. “Predicting essential genes for identifying potential drug targets in Aspergillus fumigatus”. In: *Computational biology and chemistry* 50 (2014), pp. 29–40 (cit. on p. 70).

[Luo+14] H. Luo, Y. Lin, F. Gao, C.-T. Zhang, and R. Zhang. “DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements”. In: *Nucleic acids research* 42.D1 (2014), pp. D574–D580 (cit. on pp. 70, 81).

[May+04] E. E. May, M. A. Vouk, D. L. Bitzer, and D. I. Rosnick. “An error-correcting code framework for genetic sequence analysis”. In: *Journal of the Franklin Institute* 341.1 (2004), pp. 89–109 (cit. on pp. 1, 31, 33).

[Men+11] M. Menéndez, L Pardo, M. Pardo, and K. Zografos. “Testing the order of Markov dependence in DNA sequences”. In: *Methodology and computing in applied probability* 13.1 (2011), pp. 59–74 (cit. on p. 83).

[MG94] S. V. Muse and B. S. Gaut. “A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome.” In: *Molecular Biology and Evolution* 11.5 (1994), pp. 715–724 (cit. on p. 30).

[Mit+97] T. M. Mitchell et al. “Machine learning. 1997”. In: *McGraw Hill* 45.37 (1997), pp. 870–877 (cit. on p. 77).

[MK96] A. R. Mushegian and E. V. Koonin. “A minimal gene set for cellular life derived by comparison of complete bacterial genomes”. In: *Proceedings of the National Academy of Sciences* 93.19 (1996), pp. 10268–10273 (cit. on p. 69).

[MT13] G. Muskhelishvili and A. Travers. “Integration of syntactic and semantic properties of the DNA code reveals chromosomes as thermodynamic machines converting energy into information”. In: *Cellular and Molecular Life Sciences* 70.23 (2013), pp. 4555–4567 (cit. on pp. 49, 50, 68).

[Mus15] G. Muskhelishvili. *DNA Information: Laws of Perception*. Heidelberg, Germany: Springer, 2015 (cit. on pp. 50, 61, 68).

[MV04] O. Milenkovic and B. Vasic. “Information theory and coding problems in genetics”. In: *Information Theory Workshop* (2004) (cit. on p. 2).

[NH17] D. Nigatu and W. Henkel. “Prediction of Essential Genes based on Machine Learning and Information Theoretic Features”. In: *Proceedings of BIOSTEC 2017 - BIOINFORMATICS*. 2017, pp. 81–92 (cit. on p. 86).

[Nin+14] L. W. Ning, H. Lin, H. Ding, et al. “Predicting bacterial essential genes using only sequence composition information”. In: *Genet. Mol. Res* 13 (2014), pp. 4564–4572 (cit. on pp. 70, 71, 81, 91, 95).

[NK00] M. Nei and S. Kumar. *Molecular evolution and phylogenetics*. Oxford university press, 2000 (cit. on p. 25).

[OO80] T. Ogawa and T. Okazaki. “Discontinuous DNA replication”. In: *Annual review of biochemistry* 49.1 (1980), pp. 421–457 (cit. on p. 23).

[PB97] A. Purvis and L. Bromham. “Estimating the transition/transversion ratio from independent pairwise comparisons with an assumed phylogeny”. In: *Journal of molecular evolution* 44.1 (1997), pp. 112–119 (cit. on p. 27).

[Ped+11] F. Pedregosa, G. Varoquaux, A. Gramfort, et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. on p. 87).

[PK13] M. Papapetrou and D. Kugiumtzis. “Markov chain order estimation with conditional mutual information”. In: *Physica A: Statistical Mechanics and its Applications* 392.7 (2013), pp. 1593–1601. arXiv: 1301.0148 (cit. on p. 84).

[PK16] M. Papapetrou and D. Kugiumtzis. “Markov chain order estimation with parametric significance tests of conditional mutual information”. In: *Simulation Modelling Practice and Theory* 61 (2016), pp. 1–13 (cit. on p. 84).

[Pla+10] K. Plaimas, R. Eils, and R. König. “Identifying essential genes in bacterial metabolic networks with machine learning methods”. In: *BMC systems biology* 4.1 (2010), p. 1 (cit. on p. 70).

[PM11] K. Palaniappan and S. Mukherjee. “Predicting" Essential" Genes across Microbial Genomes: A Machine Learning Approach”. In: *Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on*. Vol. 2. IEEE. 2011, pp. 189–194 (cit. on pp. 70, 71, 94–96).

[PP02] A. Papoulis and S. U. Pillai. *Probability, random variables, and stochastic processes*. Tata McGraw-Hill Education, 2002 (cit. on pp. 6, 11).

[Pra08] L. Pray. “DNA replication and causes of mutation”. In: *Nature education* 1.1 (2008), p. 214 (cit. on p. 23).

[Pro+04] E. Protozanova, P. Yakovchuk, and M. D. Frank-Kamenetskii. “Stacked–unstacked equilibrium at the nick site of DNA”. In: *Journal of molecular biology* 342.3 (2004), pp. 775–785 (cit. on p. 50).

[PS05] Y. Peres and P. Shields. “Two new Markov order estimators”. In: *ArXiv Mathematics e-prints* (June 2005). eprint: `math/0506080` (cit. on p. 83).

[Qui86] J. R. Quinlan. “Induction of decision trees”. In: *Machine learning* 1.1 (1986), pp. 81–106 (cit. on p. 76).

[Qui93] J. R. Quinlan. *C4. 5: Programs for Machine Learning*. Morgan Kaufmann, 1993 (cit. on p. 76).

[RB10] V. Rangannan and M. Bansal. “High-quality annotation of promoter regions for 913 bacterial genomes”. In: *Bioinformatics* 26.24 (2010), pp. 3043–3050 (cit. on p. 66).

[Rei85] F. Reif. *Fundamentals of Statistical and Thermal Physics*. International student edition. McGraw-Hill Book, 1985 (cit. on p. 52).

[Ris+88] J. L. Risler, M. O. Delorme, H. Delacroix, and A. Henaut. “Amino acid substitutions in structurally related proteins a pattern recognition approach: Determination of a new and efficient scoring matrix”. In: *Journal of molecular biology* 204.4 (1988), pp. 1019–1029 (cit. on p. 30).

[Ros14] S. M. Ross. *Introduction to probability models*. Academic press, 2014 (cit. on p. 14).

[RR+96] R. Roman-Roldan, P. Bernaola-Galvan, and J. Oliver. “Application of information theory to DNA sequence analysis: a review”. In: *Pattern recognition* 29.7 (1996), pp. 1187–1194 (cit. on pp. 1, 31, 32, 50).

[Sal+04] N. R. Salama, B. Shepherd, and S. Falkow. “Global transposon mutagenesis and essential gene analysis of Helicobacter pylori”. In: *Journal of bacteriology* 186.23 (2004), pp. 7926–7935 (cit. on p. 69).

[San98] J. SantaLucia. “A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics”. In: *Proc. Natl. Acad. Sci.* 95.4 (1998), pp. 1460–1465 (cit. on pp. 50, 52, 53).

[Sar+13] F. Sarmiento, J. Mrázek, and W. B. Whitman. “Genome-scale analysis of gene function in the hydrogenotrophic methanogenic archaeon Methanococcus maripaludis”. In: *Proceedings of the National Academy of Sciences* 110.12 (2013), pp. 4726–4731 (cit. on p. 96).

[Sch+05] A. Schneider, G. M. Cannarozzi, and G. H. Gonnet. “Empirical codon substitution matrix”. In: *BMC bioinformatics* 6.1 (2005), p. 1 (cit. on pp. 31, 34, 35, 39).

[Sch+86] T. D. Schneider, G. D. Stormo, L. Gold, and A. Ehrenfeucht. “Information content of binding sites on nucleotide sequences”. In: *J. Mol. Biol.* 188 (1986), pp. 415–431 (cit. on p. 50).

[Sch+98] R. E. Schapire, Y. Freund, P. Bartlett, W. S. Lee, et al. “Boosting the margin: A new explanation for the effectiveness of voting methods”. In: *The annals of statistics* 26.5 (1998), pp. 1651–1686 (cit. on p. 80).

[Sch10] T. D. Schneider. “A brief review of molecular information theory”. In: *Nano communication networks* 1.3 (2010), pp. 173–180 (cit. on p. 50).

[Ser+06] M. Seringhaus, A. Paccanaro, A. Borneman, M. Snyder, and M. Gerstein. “Predicting essential genes in fungal genomes”. In: *Genome research* 16.9 (2006), pp. 1126–1135 (cit. on pp. 72, 85).

[Ser09] R. Serfozo. *Basics of applied stochastic processes*. Springer Science & Business Media, 2009 (cit. on p. 14).

[Sha48] C. E. Shannon. “A mathematical theory of communication”. In: *The Bell System Technical Journal* 27.July 1928 (1948), pp. 379–423. arXiv: 9411012 [chao-dyn] (cit. on pp. 1, 5, 7, 35, 50).

[SJH04] J. SantaLucia Jr and D. Hicks. “The thermodynamics of DNA structural motifs”. In: *Annu. Rev. Biophys. Biomol. Struct.* 33 (2004), pp. 415–440 (cit. on p. 50).

[SL87] P. M. Sharp and W.-H. Li. “The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications”. In: *Nucleic acids research* 15.3 (1987), pp. 1281–1295 (cit. on p. 71).

[Sob+12] P. Sobetzko, A. Travers, and G. Muskhelishvili. “Gene order and chromosome dynamics coordinate spatiotemporal gene expression during the bacterial growth cycle”. In: *Proceedings of the National Academy of Sciences* 109.2 (2012), E42–E50 (cit. on p. 60).

[Sob+13] P. Sobetzko, M. Glinkowska, A. Travers, and G. Muskhelishvili. “DNA thermodynamic stability and supercoil dynamics determine the gene expression program during the bacterial growth cycle”. In: *Mol BioSyst* 9.7 (2013), pp. 1643–1651 (cit. on pp. 49, 50, 60, 64, 68).

[Sok58] R. R. Sokal. “A statistical method for evaluating systematic relationship”. In: *University of Kansas science bulletin* 28 (1958), pp. 1409–1438 (cit. on p. 44).

[Son+11] N. Sonnenschein, M. Geertz, G. Muskhelishvili, and M.-T. Hütt. “Analog regulation of metabolic demand”. In: *BMC systems biology* 5.1 (2011), p. 40 (cit. on p. 49).

[Son+14] K. Song, T. Tong, and F. Wu. “Predicting essential genes in prokaryotic genomes using a linear method: ZUPLS”. In: *Integrative Biology* 6.4 (2014), pp. 460–469 (cit. on pp. 70, 71, 85, 92, 93).

[SS] T. D. Schneider and J. Spouge. “Information content of individual genetic sequences”. In: *J. Theor. Biol.* 189 (), pp. 427–441 (cit. on p. 50).

[SS90] T. D. Schneider and R. M. Stephens. “Sequence logos: a new way to display consensus sequences”. In: *Nucleic acids research* 18.20 (1990), pp. 6097–6100 (cit. on p. 2).

[SW02] H. Stark and J. Woods. *Probability and Random Processes with Applications to Signal Processing*. Prentice Hall, 2002 (cit. on p. 6).

[Tan+06] P.-N. Tan et al. *Introduction to data mining*. Pearson Education India, 2006 (cit. on p. 77).

[Tav86] S. Tavaré. “Some probabilistic and statistical problems in the analysis of DNA sequences”. In: *Lectures on mathematics in the life sciences* 17 (1986), pp. 57–86 (cit. on p. 28).

[Tay86] W. R. Taylor. “The classification of amino acid conservation”. In: *Journal of theoretical Biology* 119.2 (1986), pp. 205–218 (cit. on p. 43).

[TM13] A. Travers and G. Muskhelishvili. “DNA thermodynamics shape chromosome organisation and topology”. In: *Biochem Soc Trans* 41 (2013), pp. 548–553 (cit. on pp. 49, 50, 60, 63, 65, 68).

[TM15] A. Travers and G. Muskhelishvili. “DNA structure and function”. In: *FEBS Journal* 282.12 (2015), pp. 2279–2295 (cit. on pp. 50, 60, 66).

[TN93] K. Tamura and M. Nei. “Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees.” In: *Molecular biology and evolution* 10.3 (1993), pp. 512–526 (cit. on p. 28).

[Ton75] H. Tong. “Determination of the order of a Markov chain by Akaike’s information criterion”. In: *Journal of Applied Probability* (1975), pp. 488–497 (cit. on p. 83).

[Tor+04] G. J. Tortora, B. R. Funke, C. L. Case, and T. R. Johnson. *Microbiology: an introduction*. Vol. 9. Benjamin Cummings San Francisco, CA, 2004 (cit. on p. 63).

[Tra+12] A. Travers, G. Muskhelishvili, and J. Thompson. “DNA information: from digital code to analogue structure”. In: *Philos Transact A Math Phys Eng Sci* 370.1969 (2012), pp. 2960–86 (cit. on pp. 49, 50).

[Van03] A.-M. Vandamme. “Basic concepts of molecular evolution”. In: *The Phylogenic Handbook-A practical approach to DNA and protein phylogeny* (2003), pp. 1–23 (cit. on p. 26).

[VK77] F. Vogel and M. Kopun. “Higher frequencies of transitions among point mutations”. In: *Journal of molecular evolution* 9.2 (1977), pp. 159–180 (cit. on p. 26).

[Vog+95] G. Vogt, T. Etzold, and P. Argos. “An assessment of amino acid exchange matrices in aligning protein sequences: the twilight zone revisited”. In: *Journal of molecular biology* 249.4 (1995), pp. 816–831 (cit. on p. 30).

[Wan+13] X. Wang, P. M. Llopis, and D. Z. Rudner. “Organization and segregation of bacterial chromosomes”. In: *Nature Reviews Genetics* 14.3 (2013), pp. 191–203 (cit. on p. 65).

[Wan+15] T. Wang, K. Birsoy, N. W. Hughes, et al. “Identification and characterization of essential genes in the human genome”. In: *Science* 350.6264 (2015), pp. 1096–1101 (cit. on pp. 69, 97, 99).

[Wan02] J. C. Wang. “Cellular roles of DNA topoisomerases: a molecular perspective”. In: *Nature Reviews Molecular Cell Biology* 3.6 (2002), pp. 430–440 (cit. on p. 60).

[WC+53] J. D. Watson, F. H. Crick, et al. “Molecular structure of nucleic acids”. In: *Nature* 171.4356 (1953), pp. 737–738 (cit. on pp. 1, 19).

[Wei+13] W. Wei, L.-W. Ning, Y.-N. Ye, and F.-B. Guo. “Geptop: a gene essentiality prediction tool for sequenced bacterial genomes based on orthology and phylogeny”. In: *PloS one* 8.8 (2013), e72343 (cit. on pp. 70, 71, 94, 95).

[WG01] S. Whelan and N. Goldman. “General empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach”. In: *Molecular Biology and Evolution* 18 (2001), pp. 691–699 (cit. on p. 30).

[WH07] J. Weindl and J. Hagenauer. “Applying techniques from frame synchronization for biological sequence analysis”. In: *Communications, 2007. ICC'07. IEEE International Conference on*. IEEE. 2007, pp. 833–838 (cit. on p. 1).

[Woe+90] C. R. Woese, O. Kandler, and M. L. Wheelis. “Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya.” In: *Proceedings of the National Academy of Sciences* 87.12 (1990), pp. 4576–4579 (cit. on p. 18).

[XH09] Z. Xu and B. Hao. “CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes”. In: *Nucleic acids research* 37.suppl_2 (2009), W174–W178 (cit. on p. 71).

[Yak+06] P. Yakovchuk, E. Protozanova, and M. D. Frank-Kamenetskii. “Base-stacking and base-pairing contributions into thermal stability of the DNA double helix”. In: *Nucleic acids research* 34.2 (2006), pp. 564–574 (cit. on p. 50).

[Yan06] Z. Yang. *Computational molecular evolution*. Oxford University Press, 2006 (cit. on p. 18).

[Ye+13] Y.-N. Ye, Z.-G. Hua, J. Huang, N. Rao, and F.-B. Guo. “CEG: a database of essential gene clusters”. In: *BMC genomics* 14.1 (2013), p. 1 (cit. on p. 70).

[Yoc+58] H. P. Yockey et al. “Symposium on information theory in biology, Gatlinburg, Tennessee, October 29-31, 1956”. In: *Symposium on Information Theory in Biology (1956: Gatlinburg, Tenn.)* 574.01. Pergamon Press, Symposium Publications Division, 1958 (cit. on p. 1).

[Yoc05] H. P. Yockey. *Information theory, evolution, and the origin of life*. Cambridge University Press, 2005 (cit. on pp. 1, 32).

[Yoc74] H. P. Yockey. “An application of information theory to the central dogma and the sequence hypothesis”. In: *Journal of Theoretical Biology* 46.2 (1974), pp. 369–406 (cit. on pp. 1, 32).

[Yoc92] H. P. Yockey. *Information Theory and Molecular Biology*. Cambridge University Press, 1992 (cit. on pp. 1, 31, 32).

[You+17] M. Yousef, D. Nigatu, D. Levy, J. Allmer, and W. Henkel. “Categorization of species based on their microRNAs employing sequence motifs, information-theoretic sequence feature extraction, and k-mers”. In: *EURASIP Journal on Advances in Signal Processing* 2017.1 (2017), p. 70 (cit. on p. 107).

[Yu+17] Y. Yu, L. Yang, Z. Liu, and C. Zhu. “Gene essentiality prediction based on fractal features and machine learning”. In: *Molecular BioSystems* 13.3 (2017), pp. 577–584 (cit. on pp. 70, 71, 86, 91).

[Yua+12] Y. Yuan, Y. Xu, J. Xu, R. L. Ball, and H. Liang. “Predicting the lethal phenotype of the knockout mouse by integrating comprehensive genomic data”. In: *Bioinformatics* 28.9 (2012), pp. 1246–1252 (cit. on pp. 72, 85).

[Zha+16] X. Zhang, M. L. Acencio, and N. Lemke. “Predicting Essential Genes and Proteins Based on Machine Learning and Network Topological Features: A Comprehensive Review”. In: *Frontiers in Physiology* 7 (2016), p. 75 (cit. on pp. 69, 92).

[Zhu83] V. B. Zhurkin. “Specific alignment of nucleosomes on DNA correlates with periodic distribution of purine—pyrimidine and pyrimidine—purine dimers”. In: *Febs Letters* 158.2 (1983), pp. 293–297 (cit. on p. 50).

[ZL95] Y. Zhao and H. B. Lieberman. “Schizosaccharomyces pombe: a model for molecular studies of eukaryotic genes”. In: *DNA and cell biology* 14.5 (1995), pp. 359–371 (cit. on p. 97).

[ZS11] S. Zoller and A. Schneider. “A new semi-empirical codon substitution model based on principal component analysis of Mammalian sequences”. In: *Molecular biology and evolution* (2011), msr198 (cit. on p. 31).

Websites

[Che10] S. W. Cheng. *Multidimensional Scaling (MDS)*. 2010 (cit. on p. 41).

[Com07] W. Commons. *Venn Diagram of Amino Acids*. File:Amino Acids Venn Diagram.png. 2007. URL: https://commons.wikimedia.org/wiki/File:AminoAcidsVennDiagram.png (cit. on p. 45).

[Com09] W. Commons. *Codons sun*. File: Aminoacids table.svg. 2009. URL: https://commons.wikimedia.org/wiki/File:Aminoacids_table.svg (cit. on p. 22).

[Com10] W. Commons. *Comparison of a single-stranded RNA and a double-stranded DNA with their corresponding nucleobases*. File: Difference DNA RNA-EN.svg. 2010. URL: https://commons.wikimedia.org/wiki/File:Difference_DNA_RNA-EN.svg (cit. on p. 19).

[Com12] W. Commons. *Comparison of a single-stranded RNA and a double-stranded DNA with their corresponding nucleobases*. 2012. URL: https://commons.wikimedia.org/wiki/File:Central_dogma_of_molecular_biology.svg (cit. on p. 20).

[HNHGRI] N. I. of Health. National Human Genome Research Institute. *Talking Glossary of Genetic Terms*. URL: https://www.genome.gov/glossary/ (cit. on p. 23).