© 2024 Daniel A. Inafuku

INFORMATION THEORY AND MEREOLOGY WITH APPLICATIONS TO BIOLOGY

BY

DANIEL A. INAFUKU

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Physics
in the Graduate College of the
University of Illinois Urbana-Champaign, 2024

Urbana, Illinois

Doctoral Committee:

Professor Jun S. Song, Chair
Professor Kay L. Kirkpatrick, Director of Research
Professor Zoi Rapti
Professor S. Lance Cooper

Abstract

Molecular machines carry out many important biological processes in the cell, such as replication, transcription, and translation. In this thesis, we calculate information-theoretic quantities of several biomolecules using tools from classical Shannon information theory. In particular, we introduce models of the ribosome, DNA polymerase, and RNA polymerase as information-transmitting communication channels. We perform analytic calculations of bounds on the ribosome's channel capacity and an explicit formula for the capacity of DNA polymerase and RNA polymerase. We demonstrate that these enzymes safely operate at speeds below their capacity by comparing our calculations to experimentally observed rates. Finally, we conclude this thesis by presenting results on a model of mereological parthood and propose ways this model may be applied to knowledge representation in biomedical informatics.

To my parents, Shana, Lauren, and Sonny

Acknowledgments

My graduate journey has not been an easy one, but it was made easier through the contributions of many people. To be able to thank them here is an incredible privilege and something that I have looked forward to for a very long time.

I am incredibly grateful to my advisor and mentor Kay L. Kirkpatrick. Your patience and kindness gave me confidence and reassured me during many unstable times. I felt free to explore different avenues of research under your care, and for that I am very grateful. Your encouragement of and enthusiasm for my goals within and beyond academia gave me strength to see my Ph.D. through to the end. Thank you so much.

Thank you, Jun S. Song, for being generous with your time and mentorship as I conclude my program here. Your guidance has kept me accountable and helped fix my eyes towards the finish lines—both academic and professional—in a realistic manner.

I cannot overestimate my indebtedness towards S. Lance Cooper, who was always available to meet with me since my arrival to the University of Illinois right up through the present moment. You are a precious gem, and you make the department a much warmer place. In addition to your participation on my thesis committee, thank you for your insights and mentorship.

Thank you to Zoi Rapti for your membership on both my thesis and preliminary examination committees. I'm grateful for your expertise and participation.

I would also like to thank Seppe Kuehn, who participated as a member of my preliminary examination committee.

Additionally, thank you to Gina Lorenz for welcoming me into her lab. Although I ultimately found that I wasn't naturally adept at experiment, it was a pleasure to learn from you and begin my graduate journey under an excellent advisor.

Warm appreciation goes out to my research collaborators Onyema Osuagwu, Qier An, David A. Brewster, Mayisha Zeb Nakib, and Shubhang Goswami.

Thank you to the many other mentors who have positively influenced my academic path. Thank you to Francois Ramaroson for recognizing my mathematical talents. Thank you to Maggie Werner-Washburne for taking a genuine interest in my scientific career and being such a kind and generous guide.

For their sincere friendship and acquaintanceship, I extend my gratitude to ShayLyssa Alexander, Rita Garrido Menacho, Davneet Kaur, Albert Lam, Alexander R. Muñoz, Nicholas LaRacuente, Michael O'Boyle, Ki-Woo Park, Tanti Dorothea Sudibyo, Moonley Tran, and Courtnie Yokono.

I express warm gratitude to Smitha Vishveshwara and Latrelle Bright for allowing me to take part with them on a *Quantum Voyage* and share in their *Joy of Regathering*. It was a tremendous privilege to experience your creativity and brilliance firsthand.

To Siv Schwink, thank you for your guidance, mentorship, and collaborative spirit. It was and is a joy to work with you. Siv, you are such a wonderful person, and your belief in me truly gives me more confidence in my work. If I'm able to carry even an ounce of your joyous spirit into my future career, I'll know that I'm doing a good job.

A special acknowledgment goes out to my therapist, LG. Without you, I don't know if I would have been able to finish the Ph.D. program. You have been a source of incredible support and keen insights. I've learned so much from you. Thank you so much for your patience, empathy, and intelligence.

To my dear Sonny, I cannot express the joy that you've brought to my life. Besides the occasional glare during low moments, you are one of the few to never cast judgment on my character, instead loving me unconditionally. I am eternally grateful for your companionship. I hope we'll have the chance to share many more moments together, no matter how short.

Thank you all. I do hope to stay in touch with you in some way as I move into the future.

No set of acknowledgments would be complete without thanking my family. Thank you to my sisters, Shana and Lauren, for your support, encouragement, and smiles. I can come visit now. Thank you also to Roland and Jacob. If you can't make me laugh, I can always make you laugh, and that warms my heart.

I save the final personal acknowledgments for the two most important people in my life: my parents, Linda and Wendell. My life and everything I do would be impossible without your efforts, wisdom, and loving-kindness.

The author was partially supported by National Science Foundation Graduate Research Fellowship Award

No. DGE-1746047.

The work contained in this thesis was partially conducted on the territories of the Kickapoo-Mascouten, Miami, Massachusett, Canarsee, Potawatomi, and Native Hawaiian (Kānaka Maoli) peoples.

Table of contents

List of Abbreviations

aa-tRNA	aminoacyl-tRNA (charged tRNA)
BSC	binary symmetric channel
BA	Blahut-Arimoto (algorithm)
DNAP	DNA polymerase
RNAP	RNA polymerase
mRNA	messenger RNA
rRNA	ribosomal RNA
tRNA	transfer RNA

List of Symbols

$\mathbb{N}$	The natural numbers, i.e., $\mathbb{N} := \{1, 2, 3, \dots\}$
$\mathbb{Z}$	The integers
$\mathbb{R}$	The real numbers
$\mathbb{C}$	The complex numbers
$\lvert\mathcal{X}\rvert$	Cardinality of set $\mathcal{X}$
$\forall$	The "for all" quantifier
$\exists$	The "there exists" quantifier
$\subseteq$	The subset relation
log	The base-2 logarithm
p_X	Marginal probability distribution of random variable X
$H(X)$	The entropy of random variable X
$h(p)$	The binary entropy function: $h(p) = -p \log p - (1-p) \log(1-p)$
$H(Y\lvert X)$	The conditional entropy of Y given X
$I(X;Y)$	The mutual information of random variables X and Y
$\mathcal{C}$	Channel capacity or Shannon capacity
A	Nucleotide base adenine
C	Nucleotide base cytosine
G	Nucleotide base guanine
T	Nucleotide base thymine
U	Nucleotide base uracil
$\mathcal{G}$	The genetic code

Chapter 1

Introduction

1.1 Biological Information Theory

The idea that biology is in some sense fundamentally computational—often called the computational theory of mind [15]—is one that can be traced as far back as the 1940s, when McCulloch and Pitts devised the first artificial neuron in an attempt to make use of biologically inspired methods of calculation [48]. Their work was based on the analogy of the brain as a computing machine and anticipates the field of machine learning, a field devoted to solving sophisticated problems commonly associated with learning in living systems. The success of machine learning is reflected in the triumphs of the algorithms that it has produced, many of which support the concept that the computational theory of mind is a well-founded idea and that biology can indeed be understood from a computational viewpoint.

Yet, there are many arguments against the brain-as-computer analogy. In a famous article [72], the mathematician Alan Turing proposed a criterion for "intelligence." He argued that a purely information-processing machine might demonstrate intelligence by fooling a human interrogator into thinking that it itself is human by way of carefully selected responses. Many have countered such a criterion, their main arguments being that there exists a gap between pure syntax and semantics [8, 62]. Additionally, many machine learning algorithms fail in the face of seemingly simple tasks, such as image classification, tasks that are usually easy for humans [32]. It is clear that machine learning is still far away from its goal to mimic biology.

In a posthumously published report, Turing made a distinction between two types of machines: (a) "controlling" machines, or purely information processing ones, and (b) "active" machines, those producing a physical effect [71]. For example, Turing considered a telephone to be a controlling machine, whereas he

thought of a bulldozer as an example of an active machine. Interestingly, Turing placed the brain in the controlling category. However, we now know much more about biology than we did in Turing's time. The brain can certainly produce physical effects, such as external electromagnetic fields. Biology certainly produces physical effects at the microscopic scale also, where biomolecules are continually processing information and producing physical outputs. For example, the biomolecular complexes known as ribosomes convert the information found in DNA to physically active, functional proteins. This idea extends to other biomolecules that are critical for brain function, such as ion channels and neurotransmitters. Perhaps a new model that incorporates these agents' active natures can help us understand biology better. Conversely, it could also account for both the achievements and shortcomings of biologically inspired algorithms in artificial intelligence and lead to new—and ultimately better—algorithms.

1.1.1 A case for information theory

One starting point to incorporate action in biology is the field of information theory. At nearly the same time that Turing was exploring his computational ideas, the mathematician Claude Shannon was working on problems in telecommunications. Engineers at the time were interested in modeling message transmission through electrical cables in the presence of noise to improve accuracy and efficiency [51, 36]. Shannon laid out a series of fundamental ideas in a landmark 1948 paper, in which he rigorously defined information and characterized its abilities to be stored and transmitted [63].

The impact of information theory cannot be more apparent than it is today, with diverse applications in electrical engineering, statistical physics, machine learning, and quantum communication. As powerful as it is, it ignores the inherent meanings, or semantics, of the messages involved. This idea is stated clearly by Shannon himself in the opening line of his celebrated paper [63]:

> "The fundamental problem of communication...is that of reproducing at one point exactly or approximately a message...at another point... Frequently the messages have meaning... These semantic aspects of communication are irrelevant to the engineering problem."

We hope to make progress towards a notion of semantics in biology because, as mentioned above, biological systems not only process information passively, but also turn information into physical action. When a cell wants to create a protein, for example, it first copies the information contained in DNA to a molecule called messenger RNA (mRNA), which contains the information in triplets of nucleotide bases called codons. The mRNA is then processed by the ribosome, which chains together amino acids to form a polymer called a

polypeptide. One can consider the output information in the polypeptide as a string of amino acid symbols derived from a 20-letter amino acid alphabet. But amino acids are not merely symbols; they possess physical properties that dictate proper folding of the polypeptide critical for its function. In fact, improper folding of identical polypeptides can lead to non-functional or dysfunctional proteins. Incorporating this semantic content requires a new, active notion of information.

Information theory and biology have a rich history dating back to the 1950s, when Yockey attempted to apply Shannon's ideas to modern genetics [76, 77]. Indeed, protein synthesis, which describes the flow of genetic "instructions," is cast in information-theoretic language: we speak of the genetic "code," we say that information is "stored" in DNA, that it is "transcribed" to RNA, and "translated" to protein. Information theory has been fruitfully applied to biology in many ways, most notably in nucleotide sequence analysis and recognition of DNA binding sites [21].

Yet, most of these approaches do not feature an active component. Recently, Keller has provided a conceptual framework in which biomolecules possess information not only as symbols, but also within their physico-chemical structures and dynamics [43]. Translating this framework into a precise mathematical theory may enhance our understanding of biological systems as active information carriers. Because of its connections to statistical mechanics, information theory may also offer tractable methods to scale upwards from the microscopic to the macroscopic scales [50].

In this thesis (Chapters 4 and 5), we conduct an analysis of the information-processing abilities of several biomolecules, most notably the ribosome, using classical Shannon information theory.

1.2 Mereology as an application to biological and biomedical informatics

Representing information and ensuring the interoperability between different databases in biological and biomedical informatics is becoming increasingly important as the amount of data grows [13, 58]. One area where such information representation can have critical impacts is in patient care. Take, for example, the seemingly simple task of a laboratory test order by a physician (see Fig. 1.1). After a patient assessment, the physician initiates a request by making an entry into that patient's health electronic health record. To ensure that the laboratory staff can efficiently comply with the physician's request, the physician's entry is translated by some information system into a code understood by the laboratory staff. Once the request is

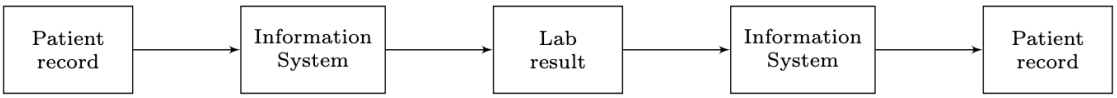

Figure 1.1: Basic scheme of a laboratory request by a physician for a patient

completed, the reverse process occurs, that is, the laboratory result is sent back through the information system and translated into language that medical staff can understand [5].

In spite of the relatively straightforward nature of this process, mistakes can accrue at each stage of information passage. If the physician's initial request does not correspond to the correct code read by the laboratory staff, the appropriate laboratory test cannot be carried out. A "common language"—that is, some degree of interoperability between the medical and laboratory staff's knowledge representations—must exist.

Interestingly, ideas from philosophy—in particular, ontology, the study of being and reality—can help us tackle this interoperability problem. The goal of classical ontology is to develop conceptualizations of the world and how these conceptualizations can be classified. In this laboratory test order example, we see that each stage (e.g., the medical staff, the information system, the lab result) has its own "conceptualization of reality" or way of representing the same information. Ontologies in computer are key components to organizing data in biology, including those that classify phenotypic data [12], chemical compounds [20], anatomy, [35], and genes and gene functions [16].

Mereology, the study of parthood and a subdiscipline of ontology, may help to bridge the gap between different ontologies and ensure interoperability. In particular, it would useful to formulate a system that could quantify keywords and index terms within an ontology. These ranking and parthood perspectives could potentially describe biological systems having hierarchical, subset-like structures, such as phylogenies and taxonomies.

In this thesis (Chapter 6), we build on results for a mereological model of parthood first introduced by Schumm *et al.* in Ref. [61].

1.3 Thesis outline

This thesis encompasses two major topics: biological information theory and mereology.

In Chapters 2-5, we introduce the subject of information theory and study the information-processing abilities of the ribosome (Chapter 4), as well as of DNA polymerase and RNA polymerase (Chapter 5).

Chapters 2 and 3 are dedicated to preliminary surveys covering the principal rudiments of the biological

systems we will study, as well as the required information-theoretic tools we will use.

Chapter 2: Information Theory Preliminaries

Chapter 2 establishes relevant definitions and theorems from information theory that we will use to quantify the information-processing abilities of certain biomolecules. This chapter culminates in the statement of Shannon's noisy channel coding theorem.

Chapter 3: Information-Carrying Biomolecules, Replication, Transcription, & Translation

Chapter 3 presents a brief review of three important information-carrying biomolecules—DNA, RNA, and protein—as well as the processes that convert the information in these biomolecules from one form to another as described by the central dogma of molecular biology—namely, DNA replication, transcription, and translation. These concepts will be relevant to Chapters 4 and 5.

Chapter 4: The Channel Capacity of the Ribosome

Chapter 4 presents original work with collaborators and is based on the journal article "The channel capacity of the ribosome," by **Inafuku, D.A.**, Kirkpatrick, K.L., Osuagwu, O., An, Q., Brewster, D.A., Zeb Nakib, M., which is published in the journal *Physical Review E* [39]. We present an information-theoretic model of the ribosome and derive some theoretical limits on its function.

Chapter 5: The Channel Capacities of DNA Polymerase and RNA Polymerase

Chapter 5 includes previously unpublished work on the information transmission abilities of DNA polymerase and RNA polymerase, the primary enzymes responsible for DNA replication and transcription, respectively. We devise models for these enzymes that are similar to the model introduced in Chapter 4.

Chapter 6: Mereological Measures on a Finite Space

We introduce this thesis' second major topic, mereology, the philosophy study of parthood, and present some original results on a mereological model. We also propose applications of our model to biological ontologies in biomedical informatics, such as the gene ontology. The work in this chapter extends work outlined in Ref. [61].

Chapter 7: Summary & Conclusion

We conclude this thesis in this chapter by providing a concise summary of this thesis' primary research contributions.

Chapter 2

Information Theory Preliminaries

Information theory is fundamentally a statistical theory and is often considered to be a branch of applied probability theory. It is chiefly concerned with two types of processes: (1) data compression and (2) data transmission. They are complementary: the core idea of compression is to remove redundancy in data to store it efficiently, whereas the core idea of transmission is to add redundancy to data to transmit it accurately.

In a landmark paper in 1948, Claude Shannon showed that there are mathematical limits on both compression and transmission [63]. He summarized these limits in two main theorems—the Source Coding Theorem for data compression, and the Noisy Channel Coding theorem for data transmission. We will focus on the Noisy Channel Coding theorem for the transmission of information by biomolecules. Before stating the theorem and to prime the reader for the methods and results in Chapters 4 and 5, we present a short survey of the relevant quantities and concepts in information theory. We will start with some basic definitions. For a comprehensive review on information theory, we refer the reader to Refs. [18, 59].

2.1 Basic information-theoretic quantities

Definition 2.1.1 (Entropy)**.** *Let X be a discrete random variable taking values in $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ and having probability mass function $p(x)$. The entropy $H(X)$ of the random variable X is defined by*[1]

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_b p(x), \tag{2.1}$$

[1] Here, we define $0 \log_2 0 = 0$ since $\lim_{p \to 0^+} p \log_2 p = 0$.

The units of entropy are determined by the base of the logarithm b. If $b = 2$, the unit is called the bit (short for binary digit). Other common units are nats ($b = e$), bans ($b = 10$), and trits ($b = 3$). For the remainder of this thesis, we will work with bits and implicitly assume that $b = 2$ unless otherwise stated.

Entropy can be thought of as a measure of how "surprised" one would be upon observing a random variable X. If X takes on a highly probable value, one is less surprised than if X takes on a less probable value. Since X may take on multiple values, the entropy gives the average value of this surprise. For this reason, entropy is also traditionally thought of as the *uncertainty* of X: the more uncertain we are of an outcome, the more surprised we are upon observing it. It is a standard exercise to show that Eq. (2.1) satisfies intuitive properties of uncertainty, namely that it is positive, is continuous in probability, and is additive in the probabilities of independent events.

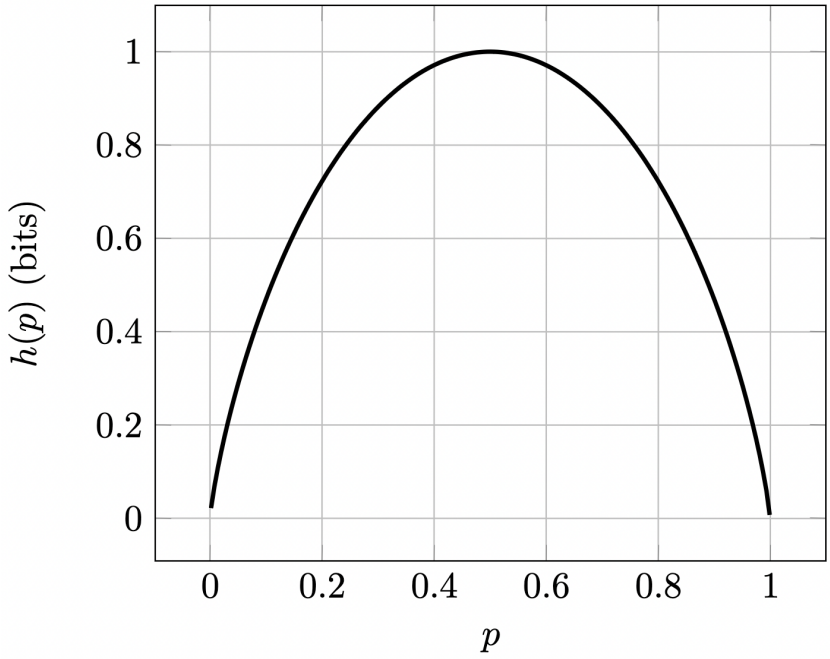

Figure 2.1: Binary entropy function $h(p)$ vs. Probability of obtaining heads p

Example 2.1.2. Suppose one has a coin with two possible states, "heads" (H) and "tails" (T), with associated probabilities $\mathbb{P}(X = \mathrm{H}) = p$ and $\mathbb{P}(X = \mathrm{T}) = 1 - p$, respectively. Calculating the entropy of X, we find that

$$H(X) = h(p) := -p \log p - (1 - p) \log(1 - p). \tag{2.2}$$

Eq. (2.2) is called the *binary entropy function* and is often denoted by $h(p)$. As seen in Fig. (2.1), $h(p)$ is maximized when $p = 1/2$. This observation agrees with our intuitive understanding of entropy as uncertainty,

since the more uncertain we are of an outcome, the greater our surprise should be. As $p \to 0$, we are more certain to obtain tails. Similarly, as $p \to 1$, we are more certain to obtain heads. In these limiting cases, $h(p) = 0$, meaning that one would gain 0 bits (i.e., no information) upon observation because one is always guaranteed to get a particular result, there being no uncertainty. In the case of a fair coin ($p = 1/2$), we have $H(X) = \log_2(2) = 1$, and we say that one would gain 1 bit of information upon observing that a coin is either "H" or "T." This idea generalizes to systems of N possible states, in which case the uniform distribution maximizes the entropy.

Entropy can also be associated with more the one random variable. For example, the conditional entropy below quantifies the amount of information in a second random variable given that one knows the value of the first.

Definition 2.1.3 (Conditional entropy)**.** *Let X and Y be discrete random variables taking values in finite sets $\mathcal{X}$ and $\mathcal{Y}$, respectively, and having joint distribution $p(x, y)$. The conditional entropy of Y given X is*

$$H(Y|X) := -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \left(\frac{p(x, y)}{p_X(x)} \right), \tag{2.3}$$

where p_X is the marginal distribution of X.

Definition 2.1.4 (Mutual information)**.** *Let X and Y be discrete random variables taking values in finite sets $\mathcal{X}$ and $\mathcal{Y}$, respectively, and having joint distribution $p(x, y)$. The mutual information $I(X; Y)$ is defined by*

$$I(X; Y) := \sum_{x \in \mathcal{X},\, y \in \mathcal{Y}} p(x, y) \log \left(\frac{p(x, y)}{p_X(x) p_Y(y)} \right), \tag{2.4}$$

where p_X, p_Y are the respective marginal distributions of X, Y.

Intuitively, mutual information describes the degree of dependence of X and Y. If $I(X; Y) = 0$, then X and Y are independent, so that $p(x, y) = p_X(x) p_Y(y)$. In general, mutual information quantifies the "difference" between the marginal distributions p_X, p_Y and their joint distribution $p(x, y)$. It is easily shown that

$$I(X; Y) = H(Y) - H(Y|X). \tag{2.5}$$

Given random variables X and Y, their respective marginal distributions p_X and p_Y, and a conditional probability distribution $p_{Y|X}$, it is important to recognize that, in general,

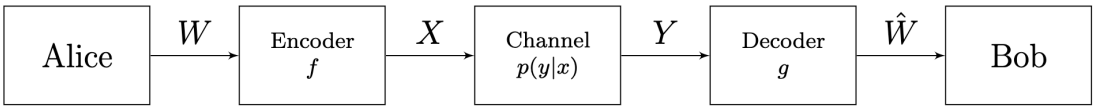

Figure 2.2: Basic communication scheme: Alice sends a string/message W to Bob, who receives $\hat{W}$. The goal is for $W = \hat{W}$.

1. $I(X;Y)$ can be viewed as a function of $p_{Y|X}$ and the marginal distribution of either X or Y, i.e., $I(X;Y) = I(p_{Y|X}, p_X)$ or $I(X;Y) = I(p_{Y|X}, p_Y)$;

2. $I(X;Y)$ is a continuous function of p_X; and

3. for a fixed conditional probability distribution $p_{Y|X}$, $I(X;Y)$ is a concave function of p_X.

These properties will be important for us when we calculate the channel capacities of different biomolecules in Chapters 4 and 5.

We can use these definitions to build a general communication scheme (see Fig. 2.2). Suppose we have two people, one of whom is a message sender, Alice, and the other a message receiver, Bob. Alice would like to send Bob a message $W \in [M] := \{1, 2, \ldots, M\}$. We can think of $[M]$ as the set of all possible messages. Physically, Alice must send the message across some medium, or channel, such as an electrical cable or air. To send the source message in a form that the channel will recognize, Alice must convert their symbols (taken from a finite set $\mathcal{X}$) into an appropriate form using an encoding function $f : [M] \to \mathcal{X}^n$ that sends W to an element in a set of input channel symbols $\mathcal{X}^n$ for some $n \in \mathbb{N}$.

The channel outputs symbols from some finite set $\mathcal{Y}$. Unfortunately, any chosen medium will introduce noise into the encoded message, so that it could become "garbled." Therefore, we identify the channel with some conditional probability distribution $p_{Y|X} : \mathcal{Y} \times \mathcal{X} \to [0, 1]$. (For simplicity, we will often write this distribution as $p(y|x)$.)

The sequence of output symbols is then converted back to source symbols using a decoding function $g : \mathcal{Y}^n \to [M]$, producing a (potentially) corrupted message $\hat{W}$, which is received by Bob.

To make these ideas precise, we make the following definitions.

Definition 2.1.5 (Discrete channel)**.** *A discrete channel is a triple $(\mathcal{X}, p(y|x), \mathcal{Y})$, where $\mathcal{X}$ and $\mathcal{Y}$ are finite sets and $p_{Y|X} : \mathcal{Y} \times \mathcal{X} \to [0, 1]$ is a conditional probability distribution.*

We interpret $\mathcal{X}$ as a set of input symbols and $\mathcal{Y}$ as a set of output symbols. A discrete *memoryless* channel refers to a channel whose input symbols do not depend on previous output symbols (for the sake of brevity,

we omit a formal definition, but we refer the reader to Ref. [18]).

Definition 2.1.6 ((M, n) code)**.** *Let $M, n \in \mathbb{N}$ and let $(\mathcal{X}, p(y|x), \mathcal{Y})$ be a discrete memoryless channel. An (M, n) code is a triple $([M], f, g)$ consisting of:*

1. *An index set $[M] := \{1, 2, \ldots, M\}$*
2. *A function $f : [M] \to \mathcal{X}^n$.*
3. *A function $g : \mathcal{Y}^n \to [M]$.*

We call f an encoding function/encoder, g a decoding function/decoder, and the elements of $\mathcal{X}^n$ codewords. n is called the (block) length of the code.

When the sets $\mathcal{X}$ and $\mathcal{Y}$ are understood, we will often refer to $p(y|x)$ as the channel rather than $(\mathcal{X}, p(y|x), \mathcal{Y})$. When $|\mathcal{X}| = n$ and $|\mathcal{Y}| = m$, it is often convenient to write $p(y|x)$ as a $n \times m$ matrix, which we call the *channel matrix*:

$$p(y|x) = \begin{pmatrix} p(y_1|x_1) & p(y_2|x_1) & \cdots & p(y_m|x_1) \\ p(y_1|x_2) & p(y_2|x_2) & \cdots & p(y_m|x_2) \\ \vdots & \vdots & \ddots & \vdots \\ p(y_1|x_n) & p(y_2|x_n) & \cdots & p(y_m|x_n) \end{pmatrix}, \tag{2.6}$$

where, of course, each row of $p(y|x)$ sums to one so that it is stochastic[2]:

$$\sum_{j=1}^{n} p(y_j|x_i) = 1. \tag{2.7}$$

Example 2.1.7 (Binary symmetric channel, Part 1)**.** The binary symmetric channel (BSC) is the simplest model of a channel with noise. It is defined by the finite alphabet sets $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ and the conditional probability distribution

$$p(y|x) = \begin{cases} 1 - p\,, & y = x \\ p \quad , & y \neq x. \end{cases} \tag{2.8}$$

Bits are sent through the channel (see Fig. 2.3) with the goal of recovering them on the receiving side. An error occurs when 0 gets flipped to a 1 and vice versa with some probability $p \in [0, 1]$.

[2]Strictly speaking, a stochastic matrix is a square matrix. Here, we refer to any matrix whose rows sum to 1 as stochastic.

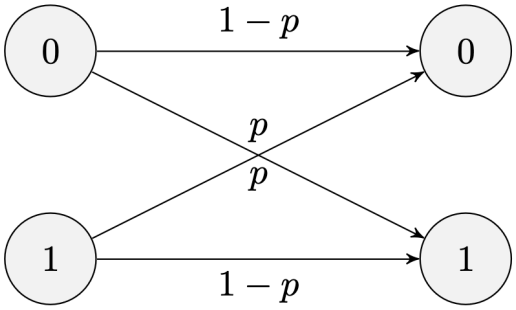

Figure 2.3: Transition diagram for the binary symmetric channel with error probability p. Bits are sent (left) through the channel to be received (right).

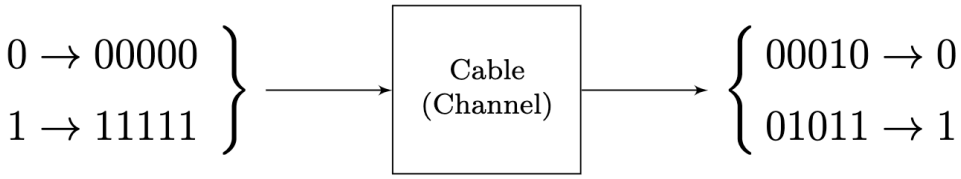

Figure 2.4: Sending a binary message using a repetition code.

Because the channel is probabilistic in any communication scheme, it is very much possible that Bob could receive an incorrect message from Alice. The challenge is for Alice to send Bob the message W as efficiently and accurately as possible. Ideally, the goal is for $W = \hat{W}$. How can Alice do this? Consider the following example.

Example 2.1.8 (Repetition code)**.** Suppose Alice is sending a binary message through a cable (see Fig. 2.4). If Alice sends a 0 and Bob receives a 1, Bob has no way of knowing if Alice sent a 0 due to noise. That is, the channel could "flip" the bit from 0 to 1 (or vice versa). To avoid sending an incorrect bit, Alice could *encode* 0 and 1 as a longer string of 0s and 1s, respectively. In other words, instead of sending individual 0s and 1s directly through the channel, Alice could send $f(0) = 00000$ and $f(1) = 11111$. This type of encoding f is called a *repetition code*. Assuming that the bit flip probability is low and that each flip is independent of the others, Bob could take a majority rule to decode the received bit string. By choosing this type of encoding—in particular, one having a large block length ($n = 5$ here)—Alice can send their message accurately. In fact, Alice could repeat the bits in the encoding as many times as they would like to send the message with as high an accuracy as they want.

Example 2.1.8 illustrates a conspicuous fact: Although Alice can achieve an arbitrarily high accuracy, the more bits they use in their encoding, the longer it will take for them to send the message! Thus, there is a

trade-off between accuracy and efficiency. Shannon proved a remarkable theorem—the Noisy Channel Coding theorem—that places a limit on the degree to which one can transmit both accurately and efficiently. Before stating the theorem, we introduce the following definitions. In what follows, $([M], f, g)$ is a given (M, n) code for a given discrete memoryless channel $(\mathcal{X}, p(y|x), \mathcal{Y})$.

Definition 2.1.9 (Conditional probability of error)**.** *The conditional probability of error λ_i given that the message $i \in [M]$ was sent is*

$$\lambda_i := \Pr(g(Y^n) \neq i | X^n = x^n(i)) := \sum_{y^n} p(y^n|x^n(i)) \mathbf{1}(g(y^n) \neq i), \tag{2.9}$$

where "Pr" denotes probability and $\mathbf{1}$ *denotes the indicator function.*

Definition 2.1.10 (Maximal probability of error)**.** *Given an (M, n) code, the maximal probability of error is*

$$\lambda^{(n)} := \max_{i \in [M]} \lambda_i, \tag{2.10}$$

where λ_i is the conditional probability of error given that the message $i \in [M]$ was sent.

Definition 2.1.11 (Rate of a code)**.** *The rate of an (M, n) code is*

$$R := \frac{\log M}{n}. \tag{2.11}$$

When sending a message through a channel, we want to add redundancy to the message to increase the probability that our message can be accurately decoded by the receiver (cf. Example 2.1.8). Since M is the number of possible messages, the idea in channel coding is that we encode the messages using n symbols, where $n > M$. The ratio R quantifies the degree to which n exceeds M. The logarithm in the numerator converts M into units of information.

Definition 2.1.12 (Rate achievability)**.** *A rate $R \geq 0$ is said to be achievable if there exists a sequence $\{(\lceil 2^{nR} \rceil, n)\}_{n=1}^{\infty}$ such that $\lambda^{(n)} \to 0$ as $n \to \infty$, where $\lceil \cdot \rceil$ denotes the ceiling function.*

Definition 2.1.13 (Channel capacity)**.** *Consider the discrete memoryless channel $(\mathcal{X}, p(y|x), \mathcal{Y})$ and let X and Y be random variables in $\mathcal{X}$ and $\mathcal{Y}$. The channel capacity $\mathcal{C}$ is given by*

$$\mathcal{C} = \sup_{p_X} I(X; Y), \tag{2.12}$$

where the supremum is taken over all possible input distributions $p_X(x)$.

The units of channel capacity are often expressed in bits per use.

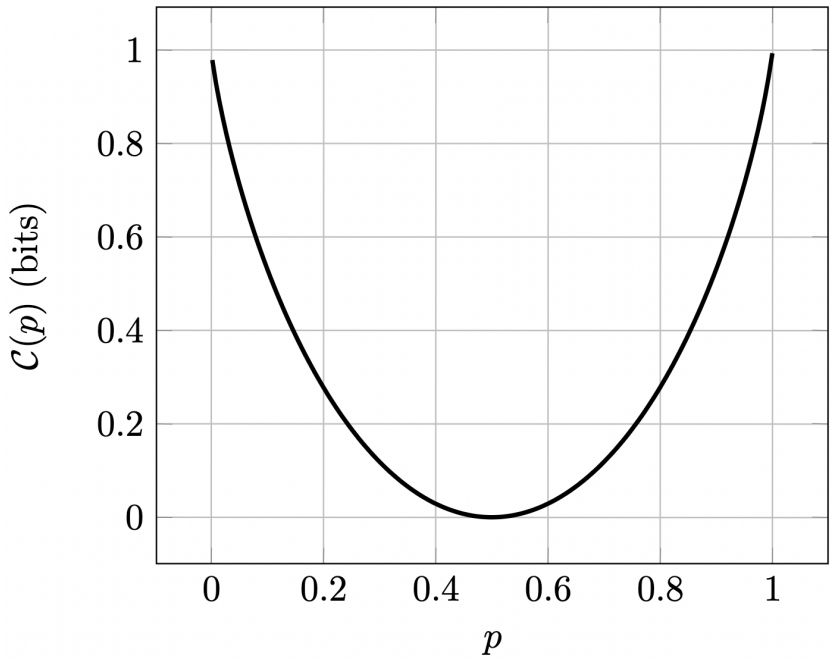

Figure 2.5: BSC's channel capacity as a function of error probability p

Example 2.1.14 (Binary Symmetric Channel, Part 2)**.** It is straightforward to show that $\mathcal{C} = 1 - h(p)$ for the BSC. The channel capacity represents the upper bound on the rate at which one can transmit reliably (i.e., with arbitrary error probability). If the channel is noiseless ($p = 0$), then $\mathcal{C} = 1$ bit per transmission (see Fig. 2.5). If bit flip is guaranteed ($p = 1$), we obtain the same capacity; note that this is still a noiseless channel. If, however, $p = 1/2$, then $\mathcal{C} = 0$ and no information is gained by the receiver. This observation makes sense, since the receiver can never be confident of what bit was sent. In fact, we confirm our intuitive notion of information, since we have not reduced any uncertainty:

$$\text{Information} = I(X;Y) = H(Y) - H(Y|X) = 0. \tag{2.13}$$

Shannon's great insight was to show that the channel capacity $\mathcal{C}$ is the supremum of the set of all achievable rates. Using our definitions above, we are now ready to state Shannon's theorem.

Theorem 2.1.15 (Shannon's Noisy Channel Coding theorem)**.** *For a discrete memoryless channel, all rates less than the capacity are achievable. In particular, for all* $R < \mathcal{C}$*, there exists a sequence of codes*

$\{(\lceil 2^{nR} \rceil, n)\}_{n=1}^{\infty}$ *such that* $\lambda^{(n)} \to 0$. *Conversely, any sequence of* $\lceil (2^{nR}, n) \rceil$ *codes with* $\lambda^{(n)} \to 0$ *must satisfy* $R < \mathcal{C}$.

Essentially, this theorem states that, given a channel, for all $\varepsilon > 0$, if $R < \mathcal{C}$, there exists a way of encoding and decoding such that the probability of error of the channel is less than ε. That is to say, if we transmit at a rate less than the capacity, we can make the error probability as small as we wish, i.e., arbitrarily small.

It is important to note that Theorem 2.1.15 is nonconstructive: just because we satisfy the hypotheses of the theorem does not mean that it gives us a code yielding our chosen error probability. The theorem only guarantees the existence of such a code.

Theorem 2.1.15 will form the basis of our conclusions about information-theoretic results we obtain in Chapters 4 and 5 for some biological information channels.

2.2 Summary

In this chapter, we have introduced some basic information-theoretic quantities, including

- the entropy $H(X)$ of a random variable X, which captures the information contained in a probability distribution;
- the mutual information $I(X;Y)$ of two random variables X and Y; and
- the channel capacity $\mathcal{C} := \sup_{p_X} I(X;Y)$.

Furthermore, we introduced a model of information transmission, the discrete memoryless channel.

This chapter culminated in the statement of Shannon's celebrated Noisy Channel Coding theorem, which states (informally) that if we transmit information over a channel at a rate below a certain number (the channel capacity), then the channel is capable of transmitting information with an arbitrarily small probability of error.

Chapter 3

Information-Carrying Biomolecules, Replication, Transcription, & Translation

In chapter 2, we introduced a mathematical definition of information, the entropy, and used this quantity to construct a model for information transmission, the discrete memoryless channel. We described this model using additional quantities such as the mutual information and the channel capacity. This chapter culminated in the statement of Shannon's Noisy Channel Coding theorem, which states that the channel capacity is an upper bound below which reliable transmission is achievable.

In the current chapter, we introduce the biological systems to which we will apply the information-theoretic tools laid out in chapter 2. In particular, we describe the principal biological molecules often considered to carry information, such as DNA, RNA, and protein, as well as the processes that are considered to transmit the information contained in these molecules, such as replication, transcription, and translation.

It is important to note that organisms are often categorized into two taxonomic "superkingdoms"—eukaryotes and prokaryotes. Eukaryotes and prokaryotes differ in many ways, most notably in the presence of a membrane-bound nucleus in the eukaryotic cell, whereas prokaryotic cells lack such a nucleus. Because of their divergent evolutionary histories, eukaryotes and prokaryotes often differ in the biochemical reactions and processes that they carry out. However, many of these processes are fundamentally the same, differing only in

certain details. For the processes considered in this chapter, we will treat processes shared by eukaryotes and prokaryotes—such as replication, transcription, and translation—as fundamentally identical, unless otherwise stated. This perspective is justified by the fact that these processes involve many of the same types of biomolecules, such as DNA and RNA, as well as the fact that products of these processes (e.g., protein) are the essentially the same.

The contents of chapters 2 and 3 provide the settings of chapters 4 and 5, where we model the molecules involved in replication, transcription, and translation as discrete memoryless channels. For a more detailed treatment of the basically biology covered in this chapter, we refer the reader to Refs. [2, 14].

3.1 DNA, RNA, and Protein

As mentioned in chapter 1, in biology one often says that information is "contained" in DNA, is "transcribed" to the intermediate molecule RNA, and finally "translated" to protein. In this section, we will describe both the structure and function of these information-carrying molecules.

3.1.1 Deoxyribonucleic acid (DNA)

Often described as the "blueprint" of an organism and the basis of heredity, deoxyribonucleic acid (DNA) is found in nearly every cell of every living organism. DNA's main role is to store all of the symbolic information that an organism needs to function in a sequence of "letters." These sequences are written using chemical letters usually denoted by the symbols A, C, G, and T. Just as information in computer science and information theory is written in strings of 0s and 1s (bits), it is in this sense that we say that DNA stores information. The only difference is that DNA is written in a 4-letter chemical alphabet, whereas traditional information uses the 2-letter alphabet of bits.

DNA consists of two strands of linear polymers. Each strand is made up of a chain of chemical units called nucleotides, which consists of a phosphate group, a five-carbon sugar called 2-deoxyribose (from which DNA derives the first part of its name), and one of four nitrogen-containing bases: adenine (A), cytosine (C), guanine (G), or thymine (T) (see Fig. 3.1).

The carbon atoms of deoxyribose can be labelled clockwise from $1'$ to $5'$ (see Fig. 3.1). In a nucleotide, the nitrogenous base is covalently bound to the sugar's $1'$ carbon, and the phosphate group is covalently bound to the $5'$ carbon. The strand is extended in either direction when one nucleotide's phosphate group

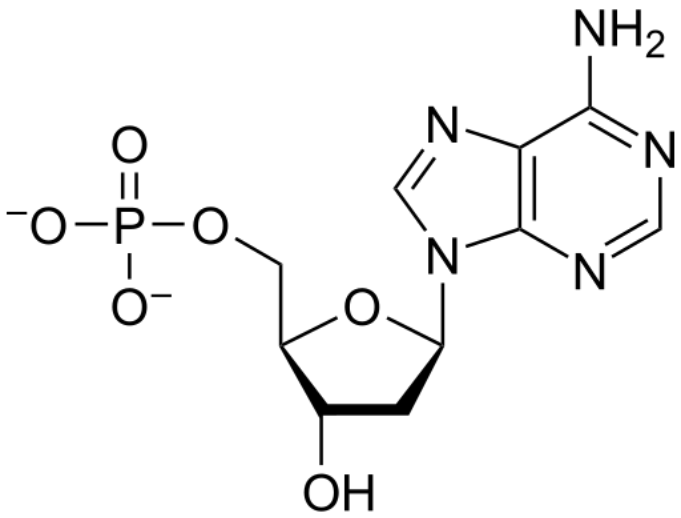

Figure 3.1: Chemical structure of a single nucleotide. The nitrogenous base (in this case, adenine) is bound to the 1′ carbon, and the phosphate group is bound to the 5′ carbon. A polynucleotide chain is formed when a phosphate group from a neighboring nucleotide covalently bonds to the 3′ carbon, replacing the hydroxyl (-OH) group.

binds to the 3′ carbon of a neighboring nucleotide (see Fig. 3.2). In this way, the alternating phosphate groups and sugars are said to form a "backbone."

The two strands of DNA are held together when one strand's bases interact with another strand's bases. In particular, a base on one strand will be held to the other when these bases hydrogen bond to each other, so that the phosphate-sugar backbone lies external to the internally paired bases. By their structure, A preferentially bonds to T (and vice versa), whereas C preferentially bonds to G (and vice versa). We call these preferential bonding rules complementary (or standard) base-pairing. Because of complementary base-pairing, knowing one sequence of bases on one strand automatically tells you the sequence of bases on the other. Thus, it is often said that one strand acts as a "template" for the other.

In addition, each strand possesses an orientation arising from the chemical structure of the 2-deoxyribose sugars: the deoxyribose sugar ring is not symmetric. Therefore, we naturally denote the strand end on the side facing the the 3′ carbon the 3′ end. Likewise, we call the strand end on the side facing the 5′ carbon the 5′ end. This directionality, much like a computer might read strings of bits (0s and 1s)—or indeed, just how like one might read English text from left to right—allows DNA to be read in a linear fashion. Moreover, when two strands bases are bound, they are bound in such a way that their backbones run antiparallel to one another. That is, if one reads one strand in, say, the 5′-to-3′ direction, the complementary strand will be oriented in the 3′-to-5′ direction relative to the first strand.

DNA does not exist as a flattened structure (see Fig. 3.3). In three dimensions, the DNA double strand

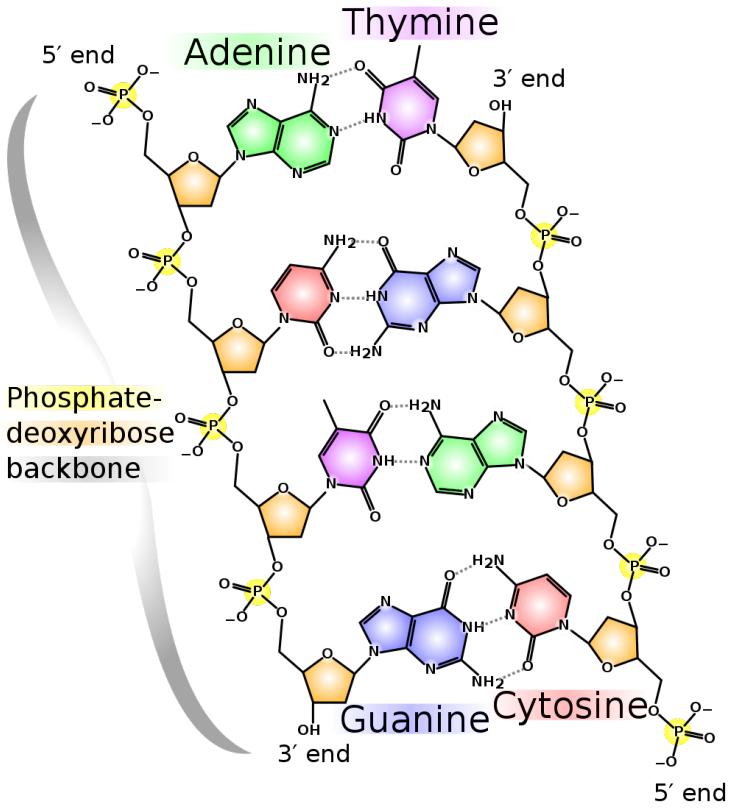

Figure 3.2: Double-stranded DNA. The "side rails" of the DNA "ladder" are made up of a backbone of phosphate groups (yellow) and deoxyribose sugars (orange). The nucleotide bases adenine (A, green), cytosine (C, red), guanine (G, blue), and thymine (T, purple) make up the "rungs." The bases are held together by hydrogen bonds (black dashes). The order of the As, Cs, Gs, and Ts encode biological information. (Courtesy of Mariana Ruiz under Creative Commons CC0 License)

adopts a characteristic helical shape because of the hydrogen bonding between its nucleotides. To minimize the binding energy (or equivalently, to maximize its packing efficiency), the two antiparallel strands wrap around each other in a right-handed helix, completing one full turn in approximately every ten base pairs. This configuration allows the helix to maintain a constant distance between the strands along the entire length of the backbone. One could imagine the DNA helix as a twisted ladder whose side rails are made by the phosphate-sugar backbone and whose rungs are formed by the complementary base pairs.

Taken together, segments of As, Cs, Gs, and Ts in the DNA sequence make up genes, units of biological information that ultimately guide the function of cells. For example, as we will see, a gene may code for the production of a certain protein, which ultimately carries out a particular function. We refer to all the information contained in an organism's DNA—and oftentimes, the DNA itself— as the organism's genome.

Next, we turn to another information-carrying molecule, ribonucleic acid, which is often seen as an

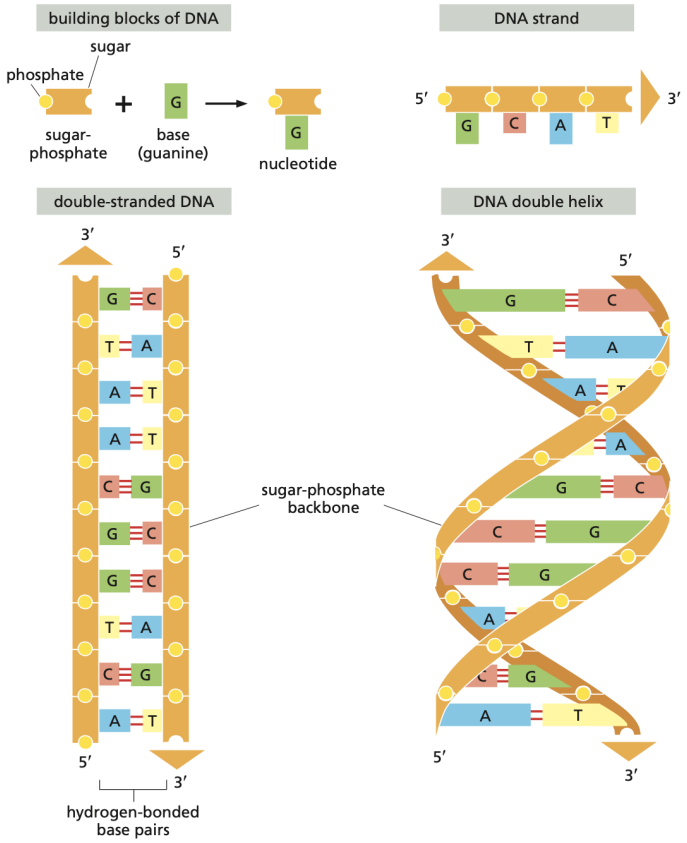

Figure 3.3: Diagram of DNA visualized in a "flattened," two-dimensional representation (left) and its three-dimensional helical structure (right). Hydrogen bonding causes the double strand to twist into its characteristic helical shape. (Figure adapted from Ref. [2].)

intermediary between DNA and protein in the process of gene expression.

3.1.2 Ribonucleic acid (RNA)

Ribonucleic acid (RNA) is another important information-containing molecule. RNA is similar to DNA in that it is also a polynucleotide polymer, having a backbone of alternating phosphate groups and sugars, with nitrogenous bases linked to the sugars. However, whereas DNA exists as a double-stranded polymer, RNA is a single-stranded polymer. Furthermore, the sugars in RNA are not deoxyribose sugars; rather, they are a kind of sugar called ribose. Like deoxyribose, ribose is also a five-carbon sugar and is differentiated from deoxyribose by the presence of hydroxyl group (-OH) on the $2'$ carbon (see Fig. 3.4). To distinguish between a nucleotide having a deoxyribose sugar and one having a ribose sugar, sometimes the terms deoxyribonucleotide and ribonucleotide are used, respectively.

Like DNA, RNA also uses a 4-letter chemical alphabet, including adenine, cytosine, and guanine. Unlike DNA, however, RNA uses the nucleotide uracil (U) rather than thymine. Uracil plays a similar role to thymine in that uracil also hydrogen bonds complementarily to adenine (see Fig. 3.6). In fact, uracil and thymine possess nearly identical chemical structures, except that thymine has a methyl ($-CH_3$) group on its carbon ring.

Because RNA is single-stranded and not bound to another strand like DNA, RNA's phosphate-sugar backbone can bend flexibly. Thus, it can adopt different complex three-dimensional shapes through complementary (and in fact, non-complementary) base pairing within itself (see Fig. 3.5). This ability allows RNA to have a much wider range of possible functions than DNA. Besides information storage and transmission, some of these functions include biochemical catalysis, forming structural elements, and gene regulation.

RNA is also inherently more chemically unstable than DNA because of several factors. First, since RNA is single-stranded, its bases are more exposed than DNA. Second, the presence of the hydroxyl group on the ribonucleotide's $2'$ carbon (a functional group that DNA lacks) renders it more reactive and vulnerable to attack by hydrolysis, wherein a water molecule reacts with the nucleotide. Third, RNA's uracil is more reactive than DNA's thymine because lack of a methyl ($-CH_3$) group. Thymine's methyl group confers some resilience, making it less susceptible to mutations. These properties makes DNA more suitable for long-term information storage.

To summarize, RNA differs from DNA in the following ways:

1. RNA is single-stranded, whereas DNA is double-stranded.

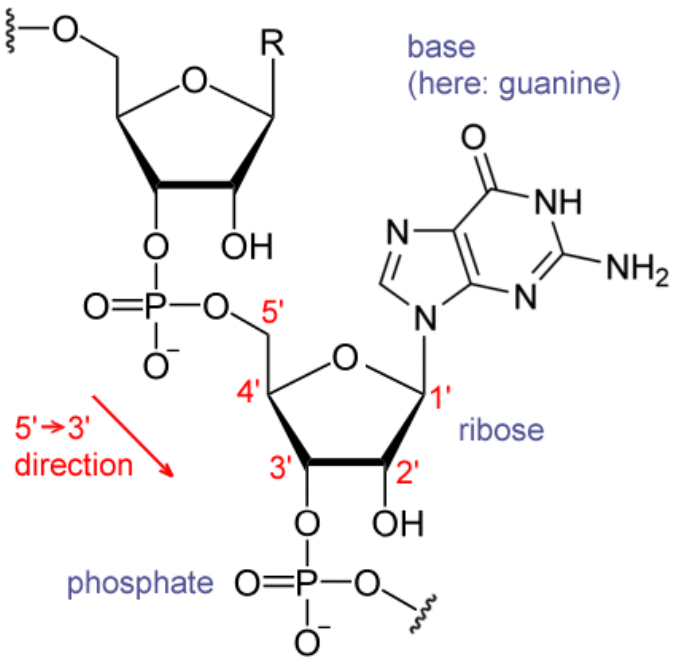

Figure 3.4: Section of an RNA strand showing two ribose sugars. Ribose is distinguished from deoxyribose by a hydroxyl group (-OH) on the 2′ carbon. Like DNA, phosphate groups are bound to the 5′ carbon and bases are bound to the 1′ carbon. Wavy lines denote continued repeating units of the strand. Courtesy of Narayanese under a CC BY-SA 3.0 License.

2. RNA contains the sugar ribose rather than deoxyribose.

3. The nucleotide uracil replaces thymine in RNA, so that RNA is written using A, C, G, and U.

4. RNA is more chemically unstable than DNA.

In terms of function, there are several different types of RNAs, three of which are relevant to this thesis. The first type is messenger RNA (mRNA), which, after being formed during transcription from DNA, is decoded by a biomolecular machine called the ribosome during translation to form protein. The second type is transfer RNA (tRNA), which brings in amino acids to the ribosome during translation. The third type is ribosomal RNA, which, as its name suggests, forms major structural components of the ribosome. There are more types of RNAs, but they will not appear in this thesis, so we will not mention them here.

Having reviewed two types of information-carrying polynucleotides, DNA and RNA, we will next discuss another important type of polymer of a different nature: protein.

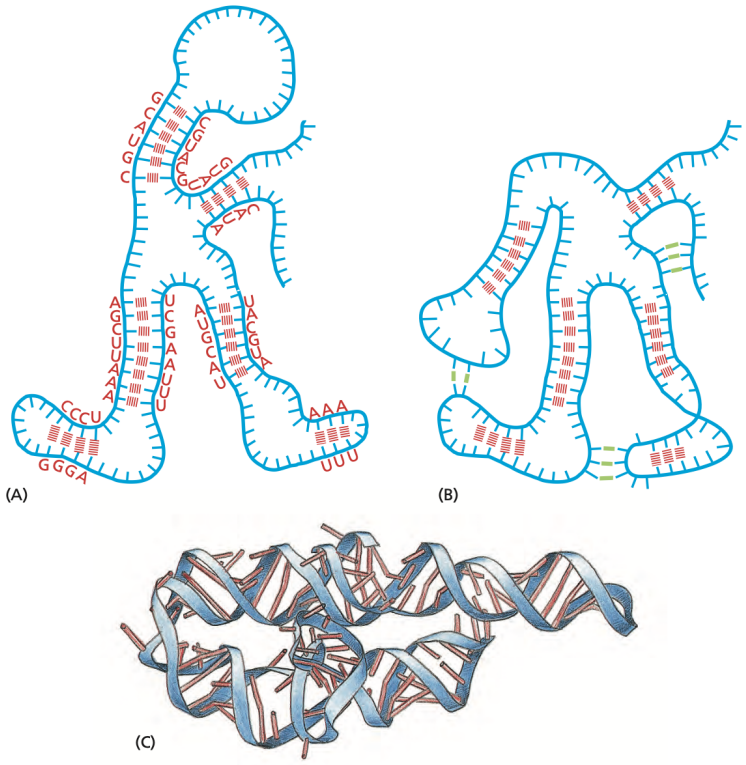

Figure 3.5: Diagram of RNA folding. Because RNA is single-stranded, its backbone is free to bend and can fold when its bases interact with each other. (A) RNA can fold when its bases form complementary base pairing (red lines). (B) RNA can also fold using non-complementary base pairing (green lines). (C) Example of RNA folding in three dimensions. (Figure adapted from Ref. [2].)

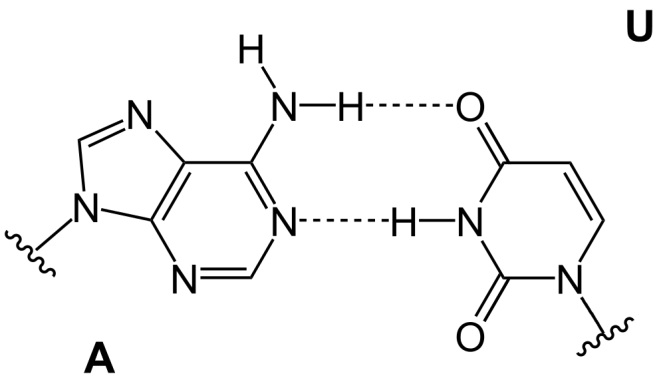

Figure 3.6: Uracil (U, right) hydrogen bonding with adenine (A, left). Uracil is a nucleotide found in RNA, where it plays the role that thymine does in DNA. The wavy lines indicate where the bases bond covalently to the sugar portion of the RNA backbone.

3.1.3 Protein

Proteins are among the most functionally diverse chemicals in biology. Like DNA and RNA, proteins are also complex polymers. We've seen that RNA, because of its single-stranded nature, has a flexible backbone that allows it to have a wide variety of functions such as enzymatic and structural roles in addition to its information storage ability. Like RNA, proteins also carry out such roles. However, proteins use a 20-letter chemical alphabet rather than 4-letter chemical alphabet, making them much more diverse in both structure and function. Proteins catalyze biochemical reactions (e.g., enzymes); act as structural elements (e.g., actin in muscle); produce force and motion (e.g., dynein and kinesin); send chemical messages (e.g., hormones and neurotransmitters); and act as essential components of the immune system (e.g., antibodies), among other roles. Proteins make up the bulk of cell's dry mass and carry out most of an organism's functions.

At the most basic level, proteins consist of chains of chemical building blocks called amino acids. In principle, there are infinitely many kinds of amino acids, but in practice, only 21 different types are used in biological systems.

Each amino acid consists of a central carbon atom (called the α carbon) to which are attached three important functional groups: an acidic carboxyl group (-COOH), a basic amino group (-NH_3), and a unique side chain (-R) that determines that amino acid's identity (see Fig. 3.7). The side chain of an amino acid confers a unique character: some amino acids are hydrophobic, whereas others are hydrophilic. Some are positively charged, whereas others are negatively charged (see Fig. 3.8). These properties determine how amino acids interact with each other and their environment.

Biologists and biochemists use two conventions to refer to amino acids: a three-letter abbreviation and a

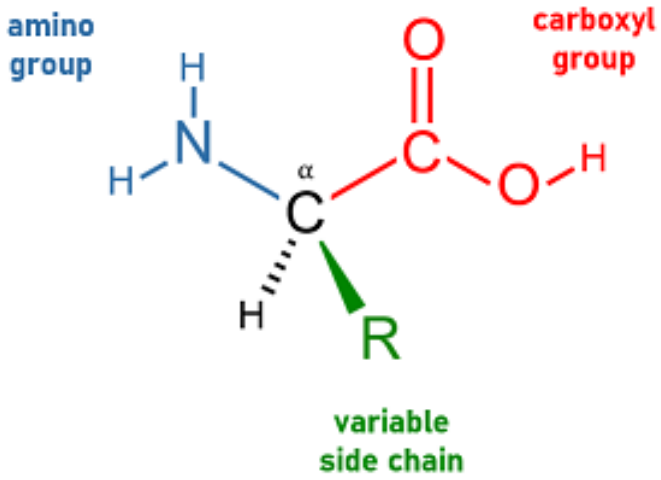

Figure 3.7: Structure of a generic amino acid. There are three important functional groups: an acidic carboxyl group (right, red), a basic amino group (left, blue), and a variable side chain group R (bottom, green) that determines the precise amino acid. Courtesy of Yassine Mrabet under a CC BY-SA 4.0 License.

single-letter abbreviation. For example, the amino acid methionine is abbreviated both as "Met" and "M." Another amino acid, tyrosine, is abbreviated "Tyr" and "Y." In this thesis, we use the three-letter convention.

Amino acids covalently bond to each other to form chains called peptides. This bonding occurs when the acidic carboxyl group of one amino acid reacts with the basic amino group of another amino acid (see Fig. 3.9). The covalent bond between two amino acids is called peptide bond. For convenience, a chain of two amino acids is referred to as a dipeptide, a chain of three amino acids is a tripeptide, and anything longer is a polypeptide. In principle, there is no limit to how long an amino acid can be. (The longest known polypeptide, a protein called titin, can reach over 38,000 amino acids long.)

Before a polypeptide can perform a function—that is, become a protein—it must fold into a particular shape. This shape is largely dictated by the interactions of the amino acids within the polypeptide chain, that is to say, the specific sequence of amino acids. A single substitution of one amino acid for another can have drastic consequences. For example, many diseases can be traced to single amino acid substitutions.

Much like how DNA and RNA have orientations—each has a $5'$ end and a $3'$ end—peptides are also oriented: one end has an amino ($-NH_2$) group, which we call the N-terminus, and the other end has a carboxyl ($-COOH$) group, which we call the C-terminus.

(It is often said that the amino acid sequence of a polypeptide determines the folding of a protein—a principle commonly referred to as Anfinsen's Dogma. However, this statement is not necessarily true.)

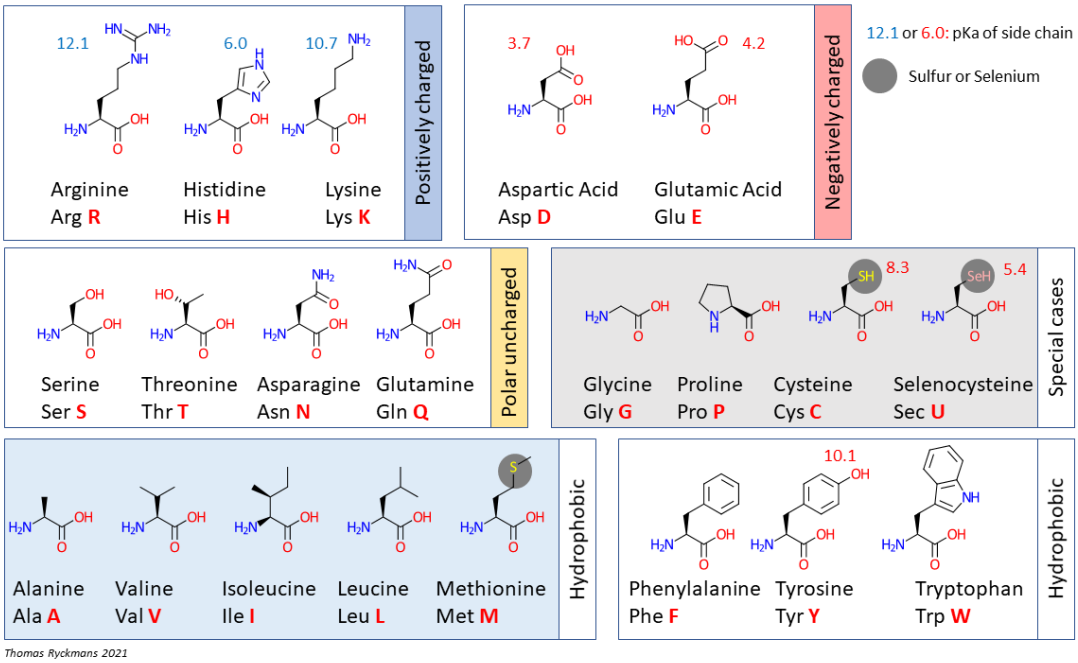

Figure 3.8: Table of the 21 proteinaceous amino acids. The amino acids are organized by their chemical properties, such as their charge or hydrophobic nature. Both the three- and single-letter abbreviations for each amino acid are given. Courtesy of Thomas Ryckmans under a CC BY-SA 4.0 License.

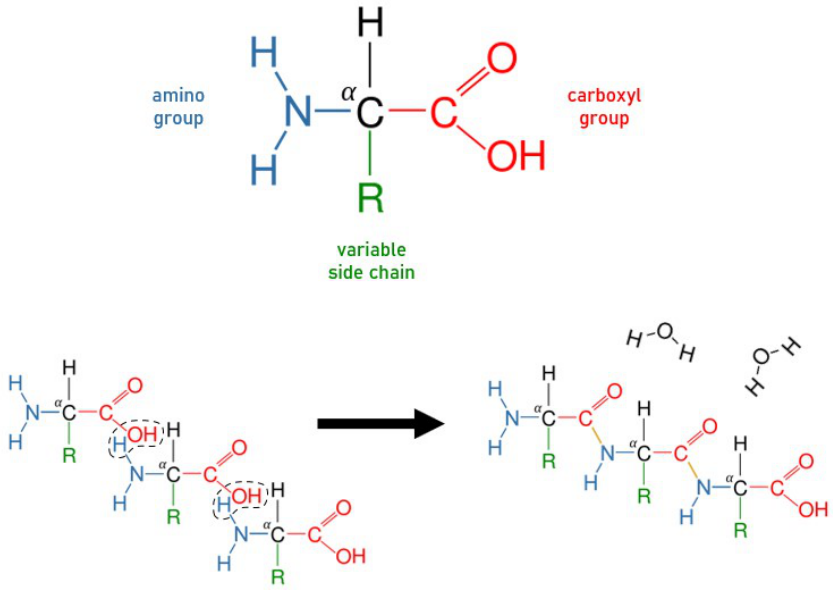

Figure 3.9: Peptide bond synthesis. The acidic carboxyl group of one amino acid reacts with the basic amino group of another amino acid. This reaction results in a chained dipeptide and the release of one molecule of water. Here, two such reactions are shown, producing a tripeptide.

3.1.4 Summary of DNA, RNA, and Protein

We have surveyed three major information-carrying molecules—DNA, RNA, and protein.

DNA is a polynucleotide that stores information in a linear sequence composed of four letters, A, C, G, and T. It exists as a double strand of antiparallel strands that forms a helix in three dimensions.

Like DNA, RNA is a polynucleotide. Unlike DNA, however, RNA is single-stranded and uses uracil in place of thymine and ribose in place of deoxyribose. In addition to storing information, RNA can perform additional functions because of a flexible backbone that allows it to fold into complex three-dimensional shapes.

Protein is a polymer composed of units called amino acids, of which there are ∼20 types. Like RNA, proteins fold into different shapes. But because proteins have access to a more diverse 20-letter amino acid alphabet rather than RNA's 4-letter nucleotide alphabet, proteins can adopt many more shapes, allowing them to perform a wider array of functions in the cell.

3.2 The Central Dogma of Molecular Biology

In Section 3.1, we introduced three important classes of biomolecules: DNA, RNA, and protein. In this section, we describe the dynamic processes that convert the information contained in these biomolecules from one form to another.

These processes are summarized concisely by the so-called central dogma of molecular biology, a paradigm describing the flow information in biological systems. In the central dogma (see Fig. 3.10), information stored in DNA can copied back into DNA by an enzyme called DNA polymerase in a process called **DNA replication**. In addition to being replicated, the information stored in DNA can be transferred to RNA by an enzymatic "transcriber" called RNA polymerase in a process aptly named **transcription**. **Translation**, the "third" step in the central dogma, is the process by which information in RNA is converted to a physical polypeptide. This process is carried about by a complex biomolecular machine composed of both RNA and protein called the ribosome.

It is important to acknowledge that there are several exceptions to the central dogma. For example, certain viruses can convert the information in RNA to DNA in a process called reverse transcription, and there are also some viruses having RNA-based genomes that perform RNA replication. However, these processes are relatively rare, and we will consider only the three main processes of the central dogma, namely,

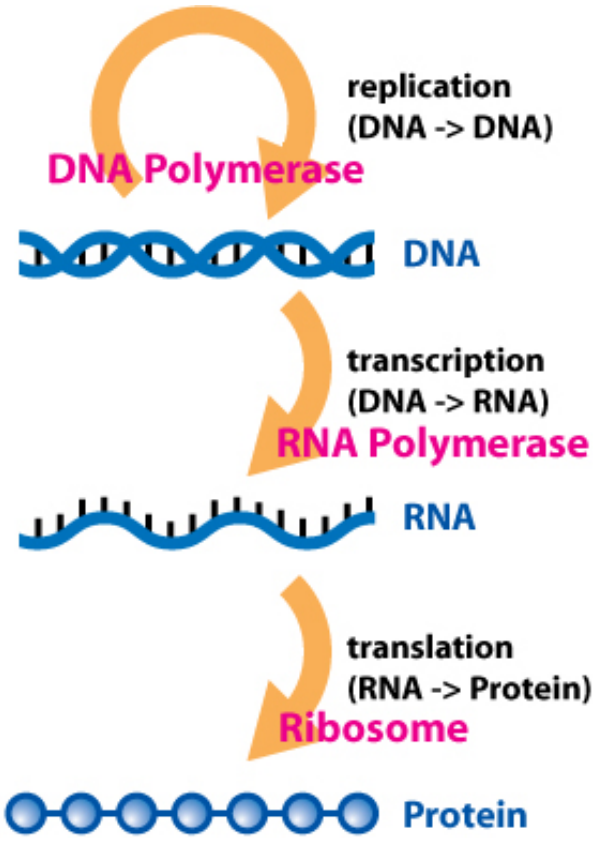

Figure 3.10: The central dogma of molecular biology. Information in DNA can be replicated into DNA, transcribed from DNA into RNA, and translated from RNA into protein. DNA replication is mediated by DNA polymerase. Transcription is mediated by RNA polymerase. And translation is mediated by the ribosome. Courtesy of Daniel Horspool under a CC BY-SA 3.0 License.

DNA replication, transcription, and translation.

3.2.1 DNA replication

Cells reproduce in several different ways. For example, in eukaryotes, tissues grow when their cells proliferate during mitosis. And in prokaryotes (bacteria and archaea), new individuals arise through binary fission. In each of these cases, a cell must copy its entire genome before it divides so that each resulting cell contains all the information it needs to eventually conduct its own reproduction and protein synthesis.

(For an excellent overview of DNA replication, please see the video at the following link: https://www.youtube.com/watch?v=TNKWgcFPHqw.)

DNA replication is carried out by a class of specialized enzymes called DNA polymerases (DNAPs). When a cell is ready to copy its DNA, the DNA must be "unzipped," which is carried out by an enzyme known as helicase. Helicase, aided by another enzyme called topoisomerase, moves along DNA, unzipping the double-stranded DNA, exposing its nucleotide bases and forming a so-called "replication fork" (see Fig. 3.11). The fork produces two single-stranded sections, one of which is called the leading strand and the other the lagging strand.

DNAP can only synthesize DNA in the 5'-to-3' direction. The leading strand is the strand for which DNAP can synthesize new DNA continuously towards the replication fork.

For the leading strand, an enzyme known as primase binds to DNA behind the replication fork and lays down a sequence of RNA called a primer, which signals the starting location where the addition of new bases will be added. Next, DNAP continuously moves along the strand towards the fork, sequentially adding new nucleotide bases.

For the lagging strand, which has an opposition orientation to the leading strand, DNAP cannot synthesize new DNA continuously. Instead, primase lays down a primer behind the replication fork. Subsequently, DNAP, signalled by this primer, synthesizes new DNA in the $5'$-to-$3'$ direction. Primase lays down another primer nearer to the fork than the previous primer, and DNAP synthesizes from the new primer to the previous primer. The series of newly synthesized, discrete chunks of DNA on the lagging strand are called Okazaki fragments. This process repeats itself.

Finally, enzymes called exonucleases degrade the RNA primers, and DNAP comes in to fill in the empty regions once occupied by the primers with new DNA.

This process on both strands continues until the entire DNA molecule is copied. At the end of replication,

a new enzyme, ligase, covalently links the Okazaki fragments together. Having started with a single DNA molecule, we are now left with two copies of DNA, each containing the same information.

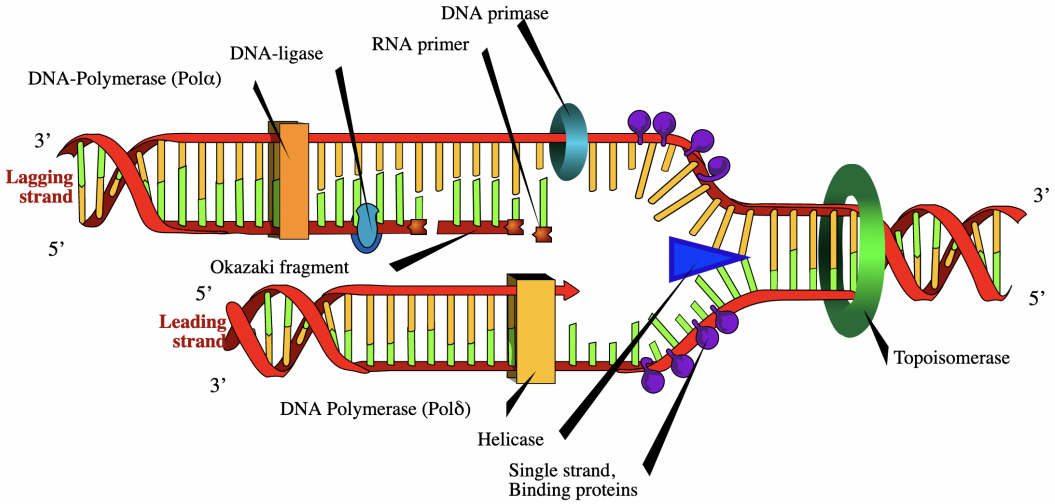

Figure 3.11: DNA replication. DNAP (yellow cuboid) synthesizes DNA in the 5'-to-3' (relative to the newly synthesized strand) direction. DNA is synthesized continually along the leading strand (bottom), whereas DNA is synthesized in discrete chunks called Okazaki fragments along the lagging strand (top). (Courtesy of Mariana Ruiz under Creative Commons CC0 License)

3.2.2 Transcription

The instructions to produce a protein product are contained in DNA. However, DNA does not directly produce proteins. Before a protein can be made, the information in the original protein-encoding gene in DNA is first copied to RNA in a process called transcription, which is mediated by an enzyme called RNA polymerase (RNAP). In this way, RNA acts as an intermediary between DNA and protein.

Both DNA and RNA are written in nucleotide letters, although RNA's uracil takes the role of DNA's thymine. Because of the conversion from one language into one that is nearly identical, information is often said to be "transcribed" from DNA to RNA. Transcription, along with translation (described in the next section), is one way that a cell expresses the genes contained in DNA. Depending on the cell or tissue type being examined, as well as the environmental conditions they are found in, different genes can be expressed at different rates, allowing the cell to respond to different stimuli according to its needs.

(For an excellent visualization of transcription, we suggest the video found at this link: https://www.youtube.com/watch?v=_Zyb8bpGMR0&t=60s.)

Transcription is very similar to DNA replication in that nucleotides are added to a growing polynucleotide chain by RNAP through complementary base-pairing to a template DNA sequence. However, transcription

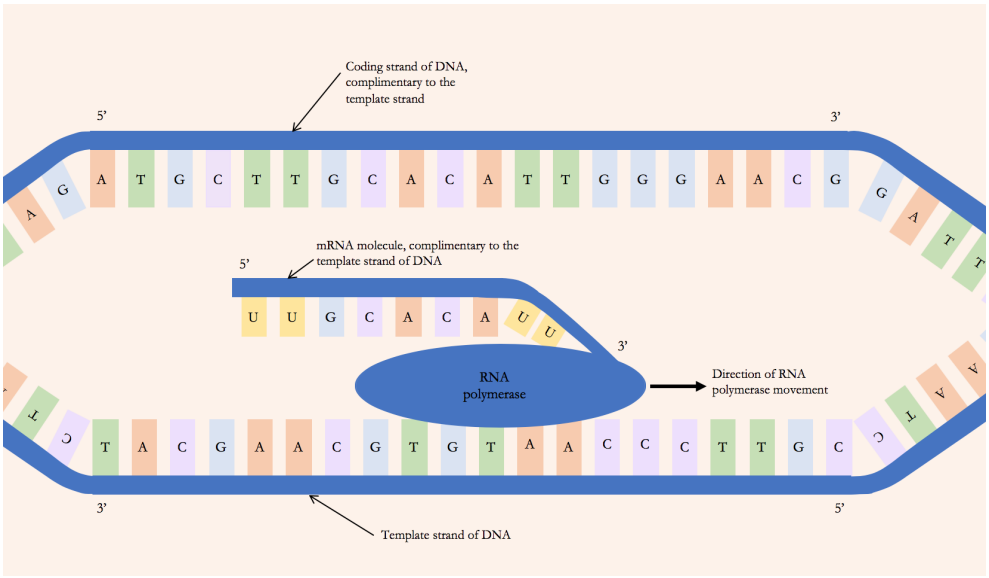

Figure 3.12: Diagram of transcription showing the transcription bubble. RNA polymerase moves along DNA's antisense strand (bottom) in the 3′-to-5′ (relative to the antisense strand) direction to produce an mRNA transcript. Licensed by Kep17 under a CC BY-SA 4.0 License.

is limited to only a specific gene region relevant to the protein product instead of the entire genome as in replication.

Transcription can be broken into three steps: initiation, elongation, and termination. A simplified diagram of transcription is given in Fig. 3.12.

The start of transcription is marked by initiation, wherein the enzyme RNAP binds to a special section of DNA called a promoter. This binding occurs through interactions along the external edges of the DNA's double helix's bases. RNAP then unzips a region of DNA called the transcription bubble by breaking the hydrogen bonds between DNA's two strands. One strand, called the antisense strand, will act as a template for adding nucleotides through complementary base pairing. The other strand, called the sense strand, will not act as a template.

The next step is elongation, during which a messenger RNA transcript is synthesized. After forming the transcription bubble, the RNAP enzyme moves along the DNA antisense strand in the 3′-to-5′ direction one nucleotide at a time, "dragging" the transcription bubble along with it. Diffusing nucleotides bind to bases within the transcription bubble through complementary base pairing, and RNAP covalently links nucleotides together sequentially.

Elongation continues until RNAP encounters a nucleotide sequence called a terminator, at which point RNAP and the mRNA transcript dissociate from DNA.

3.2.3 Translation

In the previous section, we showed how the information stored in DNA is converted to an mRNA transcript. Once a transcript is synthesized, it can be acted upon by molecular machines called ribosomes to produce proteins during the next process, translation. In going from RNA to protein, information is taken from a nucleotide language and converted to an amino acid language—hence the term "translation."

Because of a lack of a nucleus, translation in prokaryotes can occur right after—and even during—transcription. However, in eukaryotes, because transcription occurs within the nucleus, the mRNA must migrate out of the nucleus before translation can occur. Moreover, eukaryotic mRNA transcripts often undergo varying degrees of post-transcriptional processing before this migration, whereas prokaryotic mRNA transcripts do not. Although these post-transcriptional processes are important and can ultimately lead to different protein products, they will not be relevant to our work, and thus we will omit them.

In a nutshell, an mRNA transcript is translated by a biomolecular complex called the ribosome, which "reads" the transcript by recognizing triplets of nucleotides called codons, of which there are $4^3 = 64$. Each codon binds to a complementary nucleotide triplet called an anticodon. Anticodons are part of diffusing molecules called translational RNAs (tRNAs), to which are attached corresponding amino acids.

Unlike DNA replication and transcription, which involve a one-to-one correspondence given by complementary base-pairing, the correspondence in translation is dictated by a kind of "dictionary" known as the genetic code (Table 3.1), which maps each codon to one amino acid. In this way, a codon sequence is converted into an amino acid sequence. There is one special genetic code used by nearly all present-day organisms. We call this code the standard genetic code. There are, however, some alternative genetic codes, such as those in mitochondria.

(For an excellent visual overview of translation, we suggest the video found using the following link: https://www.youtube.com/watch?v=5bLEDd-PSTQ.)

Before we describe translation in more detail, we first discuss some of the "machinery" that carries out this process.

Ribosomes and tRNA

The primary mediator of translation is a biomolecular complex called the ribosome, which is composed of RNA, called ribosomal RNA (rRNA), as well as special ribosomal proteins. There are subtle differences between eukaryotic and prokaryotic ribosomes, but they are largely similar in both structure and function.

1st Base	2nd Base								3rd Base
	U		C		A		G		
U	UUU	Phe/F	UCU	Ser/S	UAU	Tyr/Y	UGU	Cys/C	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu/L	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG		UGG	Trp/W	G
C	CUU		CCU	Pro/P	CAU	His/H	CGU	Arg/R	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln/Q	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile/I	ACU	Thr/T	AAU	Asn/N	AGU	Ser/S	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys/K	AGA	Arg/R	A
	AUG	Met/M	ACG		AAG		AGG		G
G	GUU	Val/V	GCU	Ala/A	GAU	Asp/D	GGU	Gly/G	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu/E	GGA		A
	GUG		GCG		GAG		GGG		G

Table 3.1: The standard genetic code, which maps RNA codons to amino acids. Codons are written such that bases closer to the 5′ end of an mRNA transcript come first.

One difference, for example, is that eukaryotic ribosomes are slightly more massive than their prokaryotic counterparts. Both types are composed of two subunits, a large one and a small one (see Fig. 3.13). The small subunit acts as a frame that facilitates pairing of transcript codons with tRNA anticodons, whereas the large subunit catalyzes the reaction that joins consecutive amino acids. The subunits exist as separate entities when not translating but must come together for translation to occur.

The complete ribosome has four binding sites: an mRNA binding site and three tRNA binding sites called the A-site, the P-site, and the E-site ("A" for "aminoacyl," "P" for "peptidyl," and "E" for "exit"), whose roles we describe in more detail in the next section. In short, these sites are where the ribosome carries out its main role of catalyzing the reaction between consecutive amino acids.

While the ribosome joins amino acids together, tRNAs are responsible for carrying these amino acids to the ribosome (see Fig. 3.14). These tRNA molecules are transcripts roughly 80 nucleotides long. Through intramolecular complementary base pairing, tRNAs fold into "L" shapes in three dimensions, which are often described as a "cloverleaf" structures when drawn in two dimensions. There are two important regions of unpaired bases in a tRNA molecule: the anticodon loop and the acceptor stem. The anticodon loop contains the anticodon, which, as we have mentioned, complementarily binds to a codon on mRNA. The acceptor stem is the tRNA's 3′ end, to which is attached an amino acid, whose identity is dictated by the genetic code. Thus, a codon associates with its corresponding amino acid indirectly, with a tRNA molecule acting as a kind of intermediary.

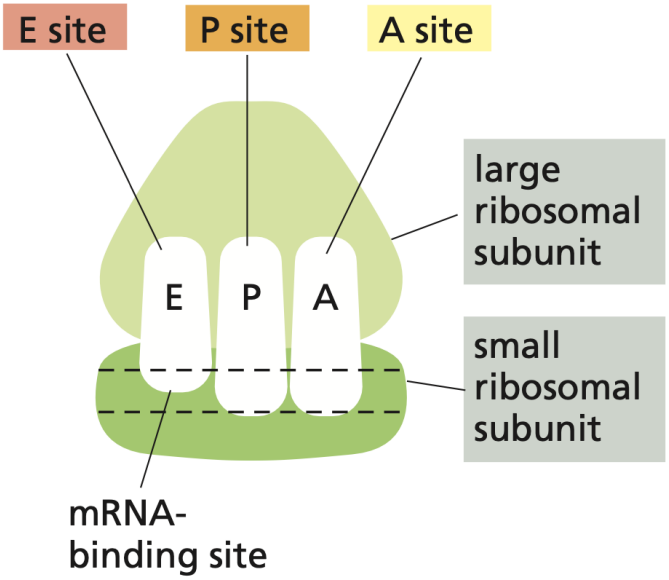

Figure 3.13: Cartoon of a complete ribosome showing the two subunits and three binding tRNA binding sites and the mRNA binding site. (Figure adapted from Ref. [2])

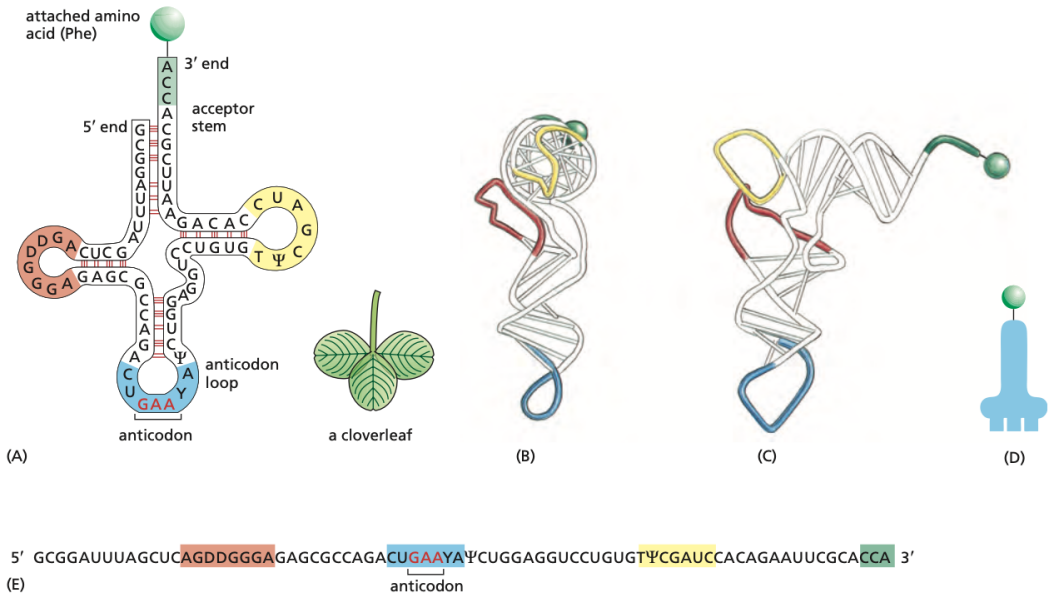

Figure 3.14: Diagram of a charged aminoacyl-tRNA (aa-tRNA). (A) The two-dimensional, "flattened" structure of an aa-tRNA molecule. Base pairings are indicated by red lines. The anticodon, which will ultimately bind to mRNA, is contained within a region near the center of the aa-tRNA's nucleotide sequence; whereas the amino acid is attached to the aa-tRNA's 3′ end. The aa-tRNA resembles a "cloverleaf." (B,C) The aa-tRNA molecule folds into a three-dimensional "L" shape. (D) Cartoon depiction of an aa-tRNA molecule. Its amino acid is represented by the green ball. (E) The linear, one-dimensional sequence of the tRNA, with color-coded regions corresponding to the color-coded regions found in (A-D). (Figure adapted from Ref. [2])

Amino acids are attached to their corresponding tRNAs by enzymes called aminoacyl-synthetases. When joined together with their amino acid partners, biologists call these tRNAs aminoacyl-tRNAs (aa-tRNAs) and say that these tRNAs are "charged." The bond between the tRNA and its amino acid is important, because the bond will later provide the energy needed to bind the amino acid to another amino acid during protein synthesis.

Translation details

Translation can be broken down into three steps: initiation, elongation, and termination. In initiation, the ribosome assembles around the mRNA transcript to begin translation. In elongation, the ribosome moves along the transcript, linking together amino acids and elongating this chain to form a polypeptide. In termination, the ribosomes reads a stop codon on the transcript and dissociates. For simplicity, we will focus largely on prokaryotic translation. Otherwise, the fundamental process of the translation in both eukaryotes and prokaryotes are similar.

As mentioned above, the ribosome exists as two separate subunits when not translating. These subunits must assemble on an mRNA transcript for translation to start. Translation is initiated at the start codon AUG, whose presence is usually signalled by special nucleotide sequence—the Kozak sequence in eukaryotes and the Shine-Dalgarno sequence in prokaryotes. These nucleotides, along with special molecules called initiation factors, help to assemble the full ribosome around the start codon. Initiation also requires a special initiator tRNA, which, by the genetic code, carries the amino acid methionine.

Elongation is the second phase of translation, during which the majority of a protein is synthesized. During this phase, a new amino acid is added to the growing, elongating chain for every repetition of the following three major steps (see Fig. 3.15):

- Step 1 (tRNA binding): An aa-tRNA molecule arrives to the ribosome's location via diffusion and binds to the mRNA's codon in the vacant A-site. At this stage, a growing amino acid-tRNA chain, called peptidyl-tRNA, is located in the P-site and the next aa-tRNA is located in the A-site.

- Step 2 (peptide bond formation): The bond between the most recently added amino acid at the growing peptide's C-terminus and the tRNA of the P-site's peptidyl-tRNA is broken. A peptide bond between the C-terminal amino acid of the growing chain and the new amino acid is formed. This peptide bond formation is catalyzed by the ribosome's large subunit. The peptide chain is now bonded to the tRNA at the A-site, and the preceding tRNA is in the P-site.

- Step 3 (large subunit translocation): Once a codon has been "read" by the ribosome and the appropriate amino acid attached to the growing peptide, the ribosome must move, or translocate, to the next codon in the mRNA codon sequence. Translocation proceeds in two substeps. First, the large subunit moves in the 5′-to-3′ direction by three nucleotides so that the new peptidyl-tRNA is now located in the large subunit's P-site and the preceding tRNA is located in the large subunit's E-site.
- Step 4 (small subunit translocation): Second, the small subunit moves in the 5′-to-3′ direction by three nucleotides, "catching up" to the large subunit. This process is accompanied by the release of the tRNA from the E-site. The ribosome's empty A-site is now ready to accept a new aa-tRNA.

These four steps repeat over and over, and new amino acids are added to the peptide as dictated by the mRNA codon sequence and the genetic code. It is important to note that the specificity of the codon-anticodon pairing is a kind of proofreading step, as incorrect pairings do not trigger the necessary conformational changes in the ribosome needed for elongation to proceed. Thus, incoming aa-tRNAs can dissociate from the A-site after Step 1 (i.e., be rejected from the A-site) and before Step 2 if an incorrect match occurs. In addition, when an aa-tRNA arrives in Step 1, a time delay is introduced as the new amino acid is properly oriented. This time delay is longer for incorrect pairings, increasing the likelihood that the aa-tRNA will be rejected from the A-site.

As the polypeptide is synthesized, it exits through a tunnel in the large subunit and begins to fold into its proper three-dimensional shape dictated by the interactions of the amino acids in its sequence.

Finally, translation ends with termination, which occurs when the ribosome encounters a stop codon: UAA, UAG, or UGA. When a stop codon enters the A-site, special enzymes called release factors bind to the ribosome, triggering the addition of water to the peptide chain's C-terminus instead of a new amino acid. The newly made protein is released and the ribosome dissociates from the mRNA transcript.

3.2.4 Summary

In this chapter, we have described in detail the principal information-carrying molecules often encountered in biology—DNA, RNA, and protein. Moreover, we have described the processes that convert information from one kind of biomolecule to another—DNA to DNA (replication), DNA to RNA (transcription), and RNA to protein (translation).

The material contained in this chapter prepares the reader for novel work presented in Chapter 4, in which we introduce a model of the ribosome as a discrete memoryless channel, as well as in Chapter 5, in

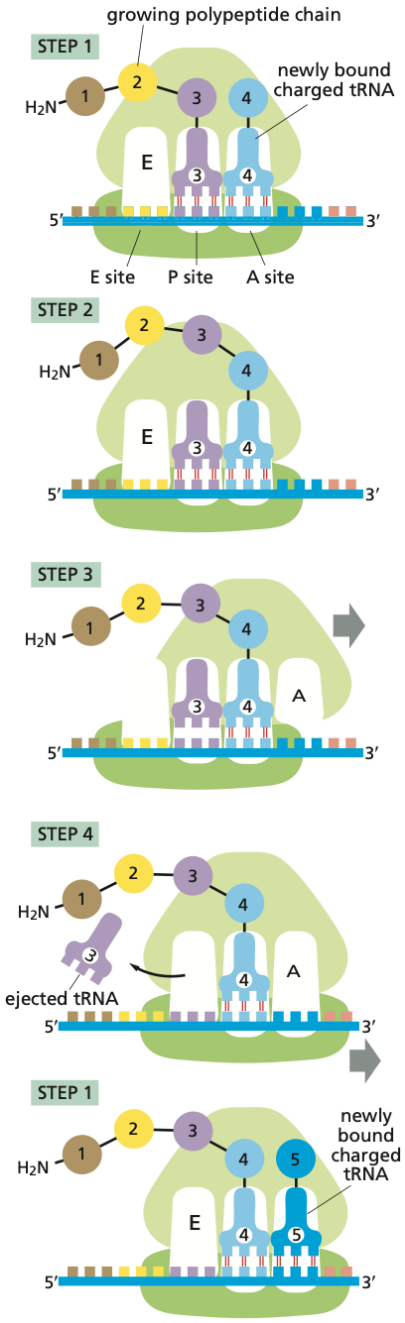

Figure 3.15: Cartoon diagram of the ribosome (large subunit: light green, small subunit: dark green) translating during elongation. Step 1: An aa-tRNA (blue) diffuses to the ribosome's A-site and binds to mRNA through a complementary codon-anticodon pairing. A growing polypeptide attached to a tRNA (peptidyl-tRNA, purple) is located at the P-site. Step 2: The ribosome's large subunit (light green) catalyzes the breakage of the peptide-tRNA bond of the P-site's peptidyl-tRNA and the formation of a new peptide bond between the previous amino acid (purple dot labelled "3") and the new one (blue dot labelled "4") brought by the aa-tRNA in the A-site. Step 3: The large subunit (light green) translocates in the 5′-to-3′ direction (to the right), moving the first tRNA (purple) from the P-site to the E-site and the new peptidyl-tRNA (blue) from the A-site to the P-site. Step 4: The small subunit (dark green) "catches up" to the large subunit and the first tRNA is ejected from the E-site. The process now returns to Step 1 and the cycle repeats until termination is signalled by a stop codon. (Figure is adapted from Ref. [2])

which we introduce a similar model for DNAP and RNAP.

Chapter 4

The Channel Capacity of the Ribosome

In this chapter, we introduce a model of the ribosome as an information channel and show that it operates within information-theoretic limits, which allows the ribosome to translate both accurately and quickly. In particular, we derive explicit bounds on the ribosome's channel capacity. We verify these analytic bounds numerically and compute a closer numerical approximation. Finally, we compare our analytical and numerical results to experimentally observed translation rates, showing that these rates lie below our calculated values. These results explain the ribosome's ability to translate quickly without sacrificing accuracy. To our knowledge, this work is the first to compare the ribosome's channel capacity to experimental translation rates. For the remainder of this chapter, "log" denotes the base-2 logarithm so that the units of information are bits, unless stated otherwise.

The work in Sections 4.1-4.5 is primarily based on the journal article "Channel capacity of the ribosome," by **Inafuku, D.A.**, Kirkpatrick, K.L., Osuagwu, O., An, Q., Brewster, D.A., Zeb Nakib, M., *Physical Review E* 108(4), 044404 (2023) [39], which was published on 09 October 2023. Kirkpatrick, Osuagwu, and Inafuku conceived of and developed the project. Inafuku, Kirkpatrick, and Osuagwu performed the analytical calculations and wrote the manuscript. An, Brewster, and Zeb Nakib performed the numerical calculations.

4.1 Motivation & Background

The ribosome is a Brownian nano-machine that assembles proteins from codon sequences in messenger RNA (mRNA, and codons are nucleotide triplets), matching each codon to an anticodon and through that to an amino acid by a kind of look-up table (i.e., the genetic code) in the physical form of transfer RNA (tRNA) [19, 57]. After joining the codon with its anticodon tRNA, the ribosome catalyzes the peptide bond formation, producing an amino acid string that folds into a protein.

A ribosome is a one-way, almost deterministic, finite transducer (in the terminology of Aho [1]): almost deterministic in the sense that rare errors occur approximately once in 1,000 to 10,000 codons [27, 25, 55, 52, 44, 2]. Ribosomes process between about 99.9% to 99.99% of codons accurately, thanks to proof-reading mechanisms, and errors often result in premature abandonment of translation. Ribosomes usually halt correctly at stop codons but occasionally get stalled if a stop codon is missing, damaged, or misread. Such stalling can be deadly for a cell, but there are mechanisms in eukaryotes for rescue [40, 41, 78].

In addition, the ribosome is a memoryless finite-state machine having 64 codon symbols and 20 amino acid states: memoryless because the ribosome's current state determines its next action, and that next action is energetically favorable [60]. Because it links the amino acids in an ordered chain, there are combinatorially many possible output proteins. In theory, a ribosome could make more than 20^{50} outputs (50-2,000 amino acids being the length of a typical protein [2], and some can potentially reach 38,000 amino acids long [4, 45]); although in practice, the ribosome is limited by the information it is fed by the mRNA sequences.

Ribosomes operate quickly, translating a codon in about 50 milliseconds and producing a typical-length polypeptide on the order of minutes [57, 9, 2]. The polypeptide then folds into its functional protein form, with the fastest folding times being on the order of microseconds [46].

The natural interpretation of protein synthesis as a process of information transmission is widespread and may contribute to our understanding of the ribosome's simultaneously high accuracy and speed. Applications of information theory are numerous: efforts have been made at the neuron, network [42], and system levels in a variety of ways with names such as information bottleneck [68], information distortion [22], effective information [38], consistent information [17], teleosemantic information [11], and positional information [69, 26, 56, 67]. But there is no consensus yet on which are the most useful interpretations, and they are all problematic [43].

Calculations of information-theoretic quantities focusing generally on gene expression and protein synthesis have been conducted previously using specially constructed channel matrices [76, 77, 23]. We build on these

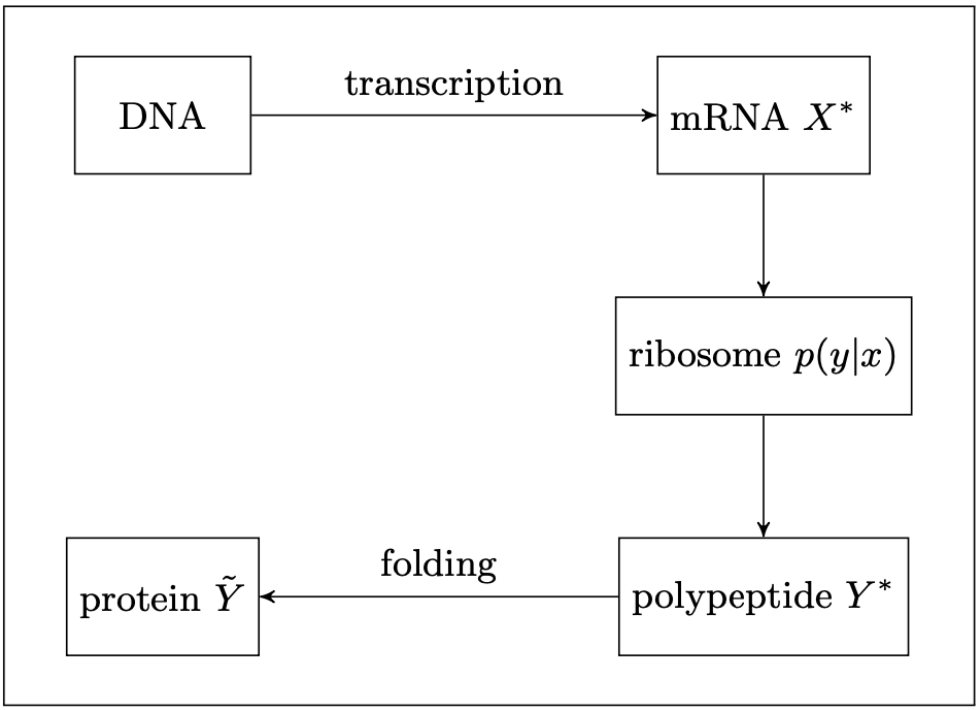

Figure 4.1: The ribosome as an information-theoretic channel: input string X^* is translated into an output string Y^*, which folds into protein $\tilde{Y}$.

results by introducing a novel channel matrix for the ribosome and show that it operates at rates below its channel capacity and satisfies the hypotheses of Shannon's Noisy Channel Coding theorem, allowing the ribosome to transmit information with an arbitrary degree of error. We do so by modeling the ribosome as an information channel, calculating bounds on the ribosome's channel capacity, and comparing the capacity with experimentally determined translation rates. These results provide explanations for the ribosome's high accuracy despite its high translation rate.

4.2 The ribosome as an information channel

In order to model the ribosome as an information channel, we must first choose the input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$. During translation, the ribosome takes information from codons to amino acids. Thus, a natural choice for the input alphabet is $\mathcal{X} := \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}^3$, and a natural choice for the output alphabet is $\mathcal{Y} := \{\text{Met, Leu, } \ldots, \text{ Ser, Stop}\}$, where "Met," "Leu," $\ldots$, "Ser" are the 20 proteinogenic amino acids and "Stop" is the stop symbol. Notice that each codon is specified by $\log 64 = 6$ bits and each amino acid is specified by $\log 21 = 4.3923\ldots$ bits.

Given $\mathcal{X}$ and $\mathcal{Y}$, we model the ribosome as a discrete memoryless channel by specifying its conditional

probability distribution $p(y|x)$ (see Fig. 4.1):

$$p(y|x) = \begin{cases} 1 - r \ , & y = \mathcal{G}(x) \\ \frac{r}{20} \ , & y \neq \mathcal{G}(x), \end{cases} \tag{4.1}$$

where $\mathcal{G} : \mathcal{X} \to \mathcal{Y}$ is the standard genetic code, and $r \in [0, 1]$ is the probability of error. $\mathcal{G}$ is constructed from a table of the standard genetic code [14]. For example, $\mathcal{G}(\text{AUG}) = \text{Met}$, where Met is the amino acid methionine. The transmission diagram of the ribosomal channel is given by Fig. 4.2.

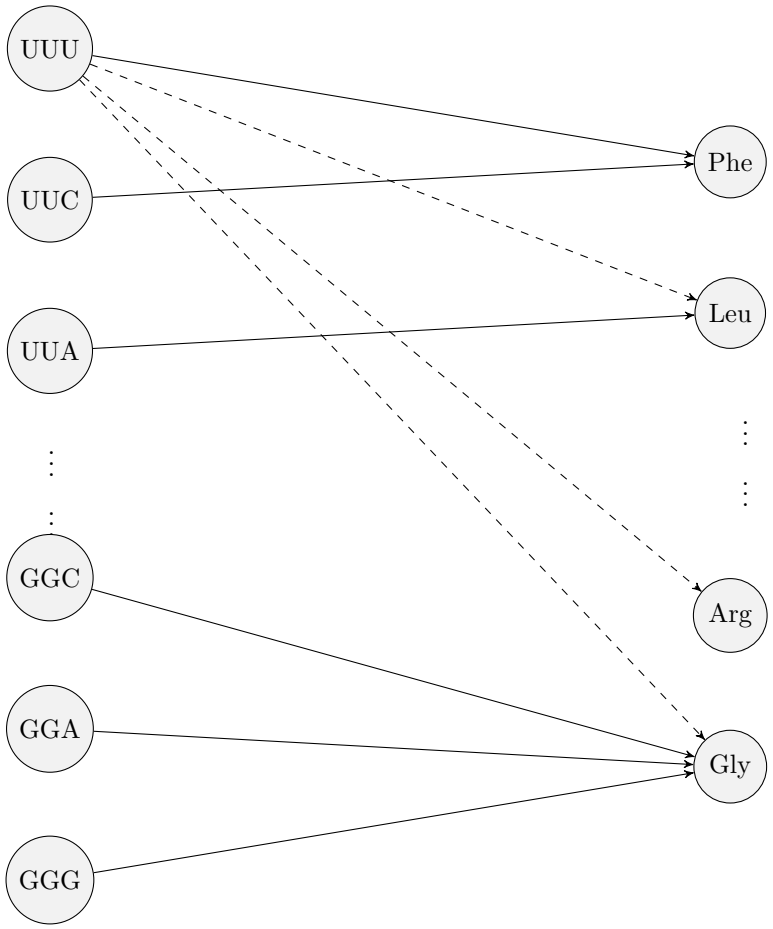

Figure 4.2: Transmission diagram for the ribosome. There are 64 input symbols (codons, left) and 21 output symbols (amino acids plus the "Stop" symbol, right), although most symbols are hidden for clarity. Solid arrows indicate "correct" transmissions (i.e., $y = \mathcal{G}(x)$), and dashed arrows represent "incorrect" transmissions (i.e., $y \neq \mathcal{G}(x)$). Only UUU's incorrect transmissions are depicted here. Synonymous codons are mapped to the same amino acid—e.g., codons GGC, GGA, and GGG are all mapped by $\mathcal{G}$ to the amino acid glycine (Gly). Synonymous codons represent the degeneracy of $\mathcal{G}$, which contributes to the ribosomal channel's asymmetry.

According to Eq. (4.1), the ribosome correctly matches a codon $x \in \mathcal{X}$ to its corresponding amino acid

$y \in \mathcal{Y}$ according to $\mathcal{G}$ (i.e., $y = \mathcal{G}(x)$) with some (preferably large) probability $1 - r$. Conversely, there is a (preferably small) probability r that the ribosome performs an incorrect match (i.e., $y \neq \mathcal{G}(x)$). In this case, our model assumes that the error probability r is distributed equally over all 20 possible incorrect outputs in $\mathcal{Y}$, i.e., for a fixed codon x, each $y \neq \mathcal{G}(x)$ has a probability $\frac{r}{20}$. These conditions ensure that $p(y|x)$ is properly normalized.

It is worth noting that the channel given by Eq. (4.1) resembles the q-ary symmetric channel, a generalization of the binary symmetric channel from 2 to q symbols and whose capacity is known [64]. Unlike the q-ary symmetric channel, in which the input and output alphabets are the same, the input and output alphabets here are different (codons vs. amino acids). Furthermore, the ribosome channel in Eq. (4.1) has a dependence on the external function $\mathcal{G}$, which the q-ary symmetric channel lacks. These additional features require us to perform new calculations to characterize the ribosome's channel capacity.

In general, explicit formulas for channel capacities are difficult to obtain. Eq. (4.1) can be written as a large 64×21 channel matrix and contains a sufficiently large degree of asymmetry such that a closed form of the capacity $\mathcal{C}$ is difficult to obtain. However, one can still bound $\mathcal{C}$, and we do so in the following section.

4.3 Bounding the capacity

An upper bound is easy: by using a basic fact that the capacity is always bounded above by the maximum entropies of both the input and the output [18], we have that

$$0 \leq \mathcal{C} \leq \min\{\log|\mathcal{X}|, \log|\mathcal{Y}|\} = \log 21 = 4.3923\ldots, \tag{4.2}$$

A lower bound takes a little more work to obtain. We obtain the following theorem.

Theorem 4.3.1. *The channel capacity $\mathcal{C}(r)$ of the ribosome in bits/use satisfies $g(r) \leq \mathcal{C} \leq \log 21$, where $g(r)$ is given by*

$$\begin{aligned} g(r) := \frac{1}{64}\Bigg\{ & 2q\log\frac{1280q}{43r+20} + \frac{63r}{10}\log\frac{64r}{43r+20} + 18q\log\frac{640q}{11r+20} + \frac{279r}{10}\log\frac{32r}{11r+20} \\ & + 6q\log\frac{1280q}{r+60} + \frac{61r}{10}\log\frac{64r}{r+60} + 20q\log\frac{64q}{4-r} + 15r\log\frac{16r}{5(4-r)} \\ & + 18q\log\frac{640q}{60-31r} + \frac{87r}{10}\log\frac{32r}{60-31r}\Bigg\} \end{aligned} \tag{4.3}$$

and $q := 1 - r$.

Proof. Let X and Y be (discrete) random variables in $\mathcal{X}$ and $\mathcal{Y}$. We substitute the definition of mutual information into the defininition of channel capacity to obtain

$$\begin{aligned}\mathcal{C} &= \sup_{p_X} I(X;Y) \\ &= \sup_{p_X} \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p_X(x)p_Y(y)} \\ &= \sup_{p_X} \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(y|x)p(x) \log \frac{p(y|x)}{p(y)},\end{aligned} \tag{4.4}$$

where $p(x) := p_X(x)$ and $p(y) := p_Y(y)$ are the respective marginal distributions for X and Y, and $p(x,y)$ is their corresponding joint distribution.

We continue by choosing a particular marginal probability distribution p_X, namely the uniform distribution over the 64 possible codons, which gives a lower bound on the supremum:

$$\begin{aligned}\mathcal{C} &= \sup_{p_X} \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(y|x)p(x) \log \left(\frac{p(y|x)}{\sum_{x'} p(y|x')p(x')} \right) \\ &\geq \frac{1}{64} \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(y|x) \log \left(\frac{p(y|x)}{\frac{1}{64}\sum_{x'} p(y|x')} \right) \\ &=: g(r).\end{aligned} \tag{4.5}$$

To simplify notation, we define $f(y,r)$ by

$$f(y,r) := \sum_{x\in\mathcal{X}} p(y|x) \log \left(\frac{64p(y|x)}{\sum_{x'} p(y|x')} \right), \tag{4.6}$$

so that

$$g(r) = \frac{1}{64} \sum_{y\in\mathcal{Y}} f(y,r). \tag{4.7}$$

Using Eq. (4.1), for Met we have

$$\sum_{x'\in\mathcal{X}} p(\text{Met}|x') = (1-r) + \frac{63r}{20} = \frac{43r+20}{20}. \tag{4.8}$$

Here, the term $1-r$ corresponds to the case $y = \mathcal{G}(x)$, and the term $\frac{63r}{20}$ corresponds to the cases where

Codon #	1	2	3	4	5	6
Amino acid y	Met Trp	Phe, Tyr, His, Gln, Asn, Lys, Asp, Glu, Cys	Stop Ile	Val, Pro, Thr, Ala, Gly	-	Leu, Ser, Arg
$f(y,r)$	$q\log\frac{1280q}{43r+20}+\frac{63r}{20}\log\frac{64r}{43r+20}$	$2q\log\frac{640q}{11r+20}+\frac{31r}{10}\log\frac{32r}{11r+20}$	$3q\log\frac{1280q}{r+60}+\frac{61r}{20}\log\frac{64r}{r+60}$	$4q\log\frac{64q}{4-r}+3r\log\frac{16r}{5(4-r)}$	-	$6q\log\frac{640q}{60-31r}+\frac{29r}{10}\log\frac{32r}{60-31r}$

Table 4.1: Tabulation of $f(y,r)$ for different values of $y \in \mathcal{Y}$. For instance, Eq. (4.9) is given in the last row under the column labelled "1."

$y \neq \mathcal{G}(x)$. Substituting Eq. (4.8) into Eq. (4.6), we have

$$\begin{aligned} f(\text{Met},r) &= \sum_{x\in\mathcal{X}} p(\text{Met}|x)\log\left(\frac{1280p(\text{Met}|x)}{43p+20}\right) \\ &= q\log\frac{1280q}{43r+20}+\frac{63r}{20}\log\frac{64r}{43r+20}, \end{aligned} \tag{4.9}$$

where $q := 1-r$. Similar to Eq. (4.8), the first term in Eq. (4.9) corresponds to the case $\text{Met} = \mathcal{G}(x)$ and the second term corresponds to $\text{Met} \neq \mathcal{G}(x)$.

Notice that because the genetic code $\mathcal{G}$ is degenerate, $\mathcal{G}$ is not injective. By the standard genetic code, tryptophan (Trp) is the only other amino acid whose preimage is a singleton: $\mathcal{G}^{-1}(\{\text{Trp}\}) = \{\text{UGG}\}$. Thus, $f(\text{Met}) = f(\text{Trp})$. This value of f is listed in Table 4.1 under the column labelled "1." Proceeding in this way, we see that all amino acids whose preimage under $\mathcal{G}$ has two elements have the same value of f, i.e., $f(\text{Phe}) = f(\text{Tyr}) = f(\text{His}) = f(\text{Gln}) = f(\text{Asn}) = f(\text{Lys}) = f(\text{Asp}) = f(\text{Glu}) = f(\text{Cys})$. Continuing in a similar way, the values $f(y,r)$ for each of the other amino acids are also listed in Table 4.1.

By Eq. (4.7), $g(r)$ becomes

$$g(r) = 2f(\text{Met},r) + 9f(\text{Phe},r) + 2f(\text{Stop},r) + 5f(\text{Val},r) + 3f(\text{Leu},r). \tag{4.10}$$

Using Table and Eq. (4.10), we obtain Eq. (4.3). Finally, by combining Eq. (4.3) with Eq. (4.2), we arrive at the desired result.

□

We plot $g(r)$ as a function of error probability in Fig. 4.3. Note that $g(r)$ is a lower bound for an upper bound, namely, the capacity $\mathcal{C}(r)$. Thus, our model predicts that $\mathcal{C}(r)$ lies in the region between $g(r)$ and the

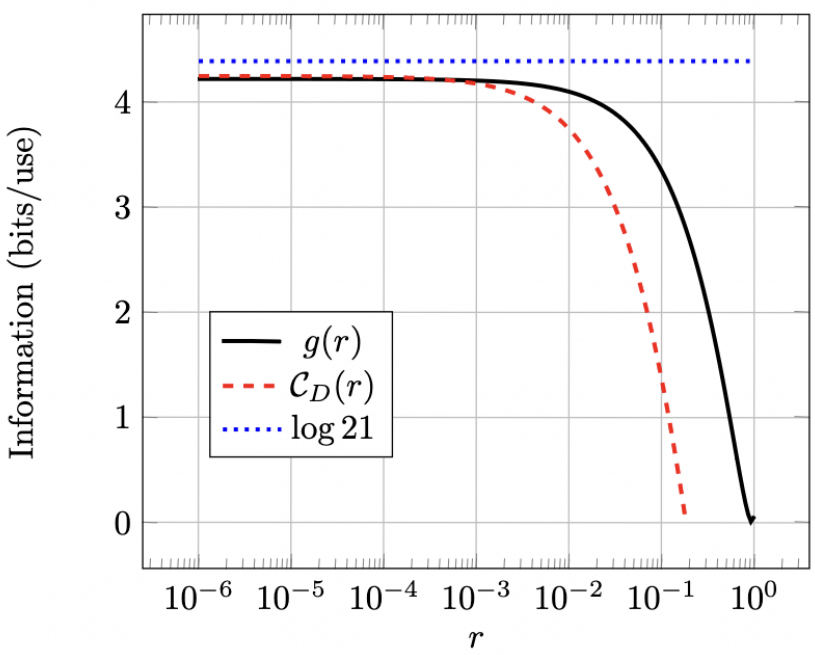

Figure 4.3: A linear-log plot showing the ribosome's capacity's lower bound $g(r)$ (black, solid), Djordjevic's maximization $\mathcal{C}_D(r)$ of Yockey's mutual information (red, dashed), and the capacity's upper bound in Theorem 4.2.1 (blue, dotted) as functions of error probability r. The space between the black/solid and blue/dotted curves represents the region where our model predicts that the capacity lies in.

constant curve $\log 21$, as stated in Theorem 4.3.1. It is easy to show that $g(r)$ is decreasing on $(0, \frac{20}{21})$, has a root at $r = \frac{20}{21}$, is increasing on $(\frac{20}{21}, 1)$, and $\lim_{r\to 0^+} = \log 21$.

To explain the slight increase for $r > \frac{20}{21}$, notice that when $r = \frac{20}{21}$, Eq. (4.1) is maximally symmetric, that is, each amino acid (including the stop symbol) is equally likely. In this case, no information is transmitted. However, once $r > \frac{20}{21}$, this symmetry is broken and asymmetry is reintroduced into the channel. Thus, $g(r)$ can only increase beyond this point.

It is worth noting that in our calculation of a lower bound $g(r)$ for the capacity, we chose the uniform distribution over the codons. In principle, we can choose any distribution—perhaps even one that more closely bounds $\mathcal{C}$ from below—but the uniform distribution was chosen because it is straightforward to use and (as we will see) yields a sufficiently good estimate to explain experimental observations.

Whereas our channel matrix is Eq. (4.1), Yockey in Refs. [76, 77] models protein synthesis using an alternative matrix that incorporates point mutations. He calls protein synthesis the "genetic communication system" and through several approximations obtains the following expression for its mutual information:

$$I(X;Y) = H(X) - 1.68 + 6.509r \log 0.4594r, \tag{4.11}$$

where $H(X)$ is the entropy of the input random variable X and $r \in [0,1]$ is the probability of error due to noise. To obtain a channel capacity, Djordjevic maximizes Eq. (4.11) using the input probability distribution of X that is uniform over the sense codons and zero over the stop codons [23], i.e.,

$$p(x) = \frac{1}{61}, \quad \forall x \in \mathcal{X} \backslash \{\mathrm{UAA}, \mathrm{UAG}, \mathrm{UGA}\}, \tag{4.12}$$

$$p(\mathrm{UAA}) = p(\mathrm{UAG}) = p(\mathrm{UGA}) = 0. \tag{4.13}$$

The distributions Eqs. (4.12) and (4.13) yields $H(X) = \log 61$ so that by Eq. (4.11) the capacity $\mathcal{C}_D$ becomes

$$\mathcal{C}_D(r) = \log 61 - 1.68 + 6.509r \log 0.4594r, \tag{4.14}$$

which we plot in Fig. 4.3 alongside our lower bound $g(r)$ of Theorem 4.3.1.

Evaluating $g(r)$ at $r = 1 \times 10^{-4}$, a typical value for the ribosome's error probability [27, 25, 55, 52, 44, 2], yields

$$0.7027... \leq \mathcal{C} \leq 0.7321... \ \frac{\text{codons}}{\text{use}}. \tag{4.15}$$

4.4 Numerically approximating the capacity

To approximate the capacity numerically, we employ the well-known Blahut-Arimoto (BA) algorithm, which is commonly used to compute the capacities of arbitrary channels.

By the definition of channel capacity, the problem of computing the capacity of a channel amounts to optimizing the mutual information $I(X;Y)$ over all input probability distributions p_X. One way to perform such an optimization is to calculate the gradient of $I(X;Y)$. Because there are $|\mathcal{X}|$ different parameters to we are left to contend with a large $|\mathcal{X}|$-dimensional space. For the ribosome, we have a 64-dimensional space, a computationally difficult problem.

Fortunately, there exists an alternative, iterative algorithm due to Arimoto and Blahut that provides an efficient method for calculating channel capacities [3, 6]. In general, the mutual information is a function of both the conditional probability distribution $p(y|x)$ and the input distribution p_X: $I(X;Y) = I(p(y|x), p_X)$. Because we fix the channel $p(y|x)$, we can think of the mutual information of the channel as a function of only the input probability distribution p_X. That is to say, $I(X;Y) = I(p_X)$. The idea behind the BA algorithm is to generate a sequence $\{Q_n\}_{n \in \mathbb{N}}$ of input probability distributions such that $I(Q_n)$ converges to the desired

capacity. More precisely,

1. starting from an arbitrary initial distribution Q_1, calculate for all $x \in \mathcal{X}$ the quantity

$$T_n(x) := \sum_{y \in \mathcal{Y}} p(y|x) \log \left(\frac{Q_n(x)p(y|x)}{R_n(y)} \right), \tag{4.16}$$

   where $R_n(y)$ is the marginal distribution of the output:

$$R_n(y) := \sum_{x \in \mathcal{X}} p(y|x)Q_n(x). \tag{4.17}$$

2. For $n \geq 2$, update Q_n according to the rule

$$Q_{n+1}(x) = \frac{e^{T_n(x)}}{\sum_{x' \in \mathcal{X}} e^{T_n(x')}}. \tag{4.18}$$

Theorem 4.4.1 (Convergence of the BA algorithm to the capacity, Theorem 3 in [6])**.** *Let $\{Q_n\}_{n \in \mathbb{N}}$ be the sequence of probability distributions generated by the BA algorithm. Then*

$$\lim_{n \to \infty} I(Q_n) = \mathcal{C}. \tag{4.19}$$

Moreover, the convergence is monotonic from below.

Of course, the BA algorithm can be carried out for as large an n as desired, i.e., up to arbitrary precision. However, since the convergence is monotonic from below, we would like a condition that stops the algorithm once our computed $I(Q_n)$ is "good enough." Such a condition is provided by the following theorem. (For a short derivation, please see Appendix A.)

Theorem 4.4.2 (Termination criterion, [31])**.** *The channel capacity $\mathcal{C}$ satisfies*

$$m_n \leq \mathcal{C} \leq M_n, \tag{4.20}$$

where $m_n := \min_{x \in \mathcal{X}} T_n(x) - \log Q_n(x)$ and $M_n := \max_{x \in \mathcal{X}} T_n(x) - \log Q_n(x)$.

By this theorem, we can allow the algorithm to run until the difference $M_n - m_n$ is as small as we desire.

In our implementation of the BA algorithm, we initialize $\{Q_n\}_{n \in \mathbb{N}}$ with the uniform distribution, i.e.,

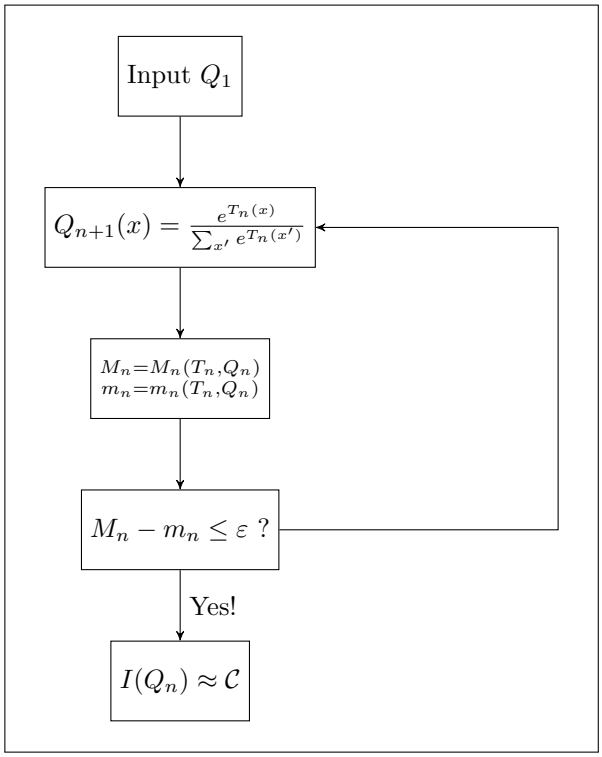

Figure 4.4: Flowchart of the Blahut-Arimoto (BA) algorithm. The BA algorithm outputs the terms of a sequence $\{Q_n\}_{n\in\mathbb{N}}$. We initialize the algorithm using an arbitrarily chosen distribution Q_1, which is used as the input to calculate the next distribution in the sequence, Q_{n+1}. The algorithm then calculates the quantities M_n and m_n and checks to see if the condition $M_n - m_n \leq \varepsilon$ holds. If yes, the algorithm stops. If no, the next term in the sequence $\{Q_n\}_{n\in\mathbb{N}}$ is calculated. In our implementation, Q_1 is the uniform distribution, and $\varepsilon = 1 \times 10^{-35}$.

$$Q_1(x) = \frac{1}{64}, \ \ \forall x \in \mathcal{X}. \tag{4.21}$$

Starting with this initialization Q_1, we iteratively generate subsequent terms of the sequence using the BA algorithm (Eqs. (4.16)-(4.18)) for $r = 1 \times 10^{-4}$. We continue the algorithm until we reach the termination criterion $M_n - m_n \leq 10^{-35}$. A flowchart of the BA algorithm is shown in Fig. 4.4.

We plot the results of the BA algorithm in Fig. 4.5 and observe that the $I(Q_n)$ indeed converges to the approximated capacity monotonically from below. For a probability of error $r = 1 \times 10^{-4}$, the calculated value of the capacity is found to be $\mathcal{C} \approx 4.3904$ bits/use $= 0.7317$ codons/use. This value falls with the range Eq. (4.15) , verifying our analytic results. The BA algorithm also outputs a capacity-achieving distribution, i.e., a distribution Q^* for which $\mathcal{C} = I(Q^*)$. This distribution is plotted in Fig. 4.6.

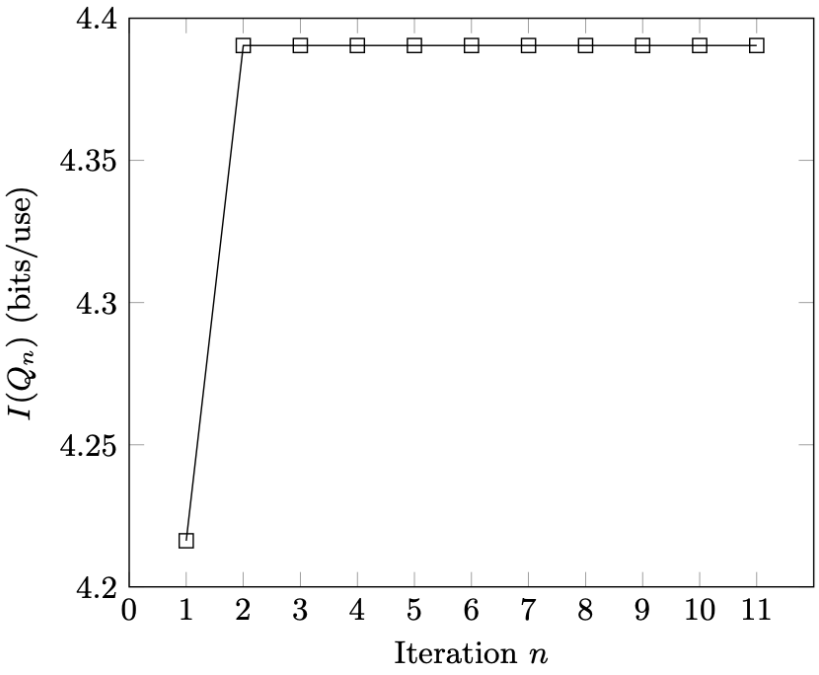

Figure 4.5: Results of the Blahut-Arimoto algorithm with $r = 1 \times 10^{-4}$. The mutual information $I(Q_n)$ converges to the capacity $\mathcal{C}$ monotonically from below.

4.5 Explaining experimental observations

We would like to compare both our analytical estimate in Theorem 4.3.1 and our numerical estimate from the BA algorithm to experimentally determined translation rates. We first recognize that the channel capacity is measured in bits per use of the channel. Notice that time is not included in this quantity. To incorporate

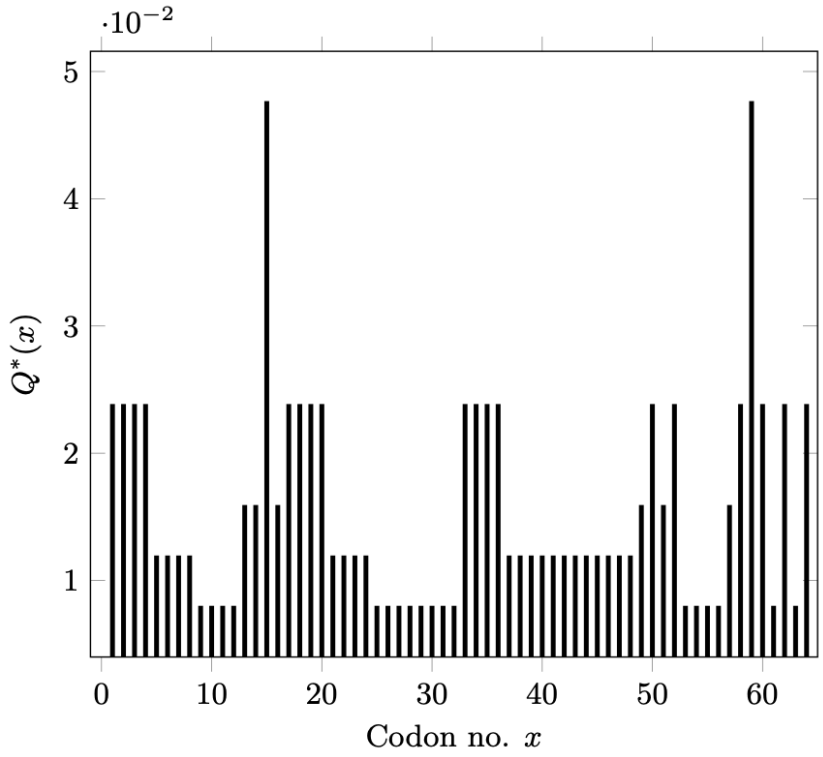

Figure 4.6: A capacity-achieving input distribution Q^* obtained from the BA algorithm. Codons are ordered alphabetically. The codons having the largest probabilities are the start codon AUG, which codes for methionine, and UGU, which codes for cysteine.

time, we must divide our by estimates by the time it takes for a single operation of the channel—that is, the translation time of a single codon. This time interval must be obtained by examining the intrinsic kinetics of the ribosome. We utilize an estimate on this time interval from a model developed by Fluitt *et al.* [29].

A critical observation is that the translation rate depends on a number of biophysical parameters, including (a) the arrival times of aa-tRNAs to the ribosome's A-site and (b) the kinetics of peptide bond formation and ribosome translocation. (See Chapter 3, Section 3.2.3 for a review of translation.) From these parameters, it is plausible that one could extract a rough estimate of a single codon's translation time.

Given a particular codon to be translated, charged tRNAs (aa-tRNAs) can be classified according to whether their anticodon is a correct base pair match (cognate tRNA), a nearly correct base pair match (near-cognate tRNA), or a completely incorrect base pair match (non-cognate tRNA). (To be clear, given a codon to be translated, a near-cognate tRNA is defined as an aa-tRNA having an anticodon that has at most a single incorrect base pair match with this codon. Any aa-tRNA having two or three base pair mismatches is a non-cognate tRNA for that codon.) It has been suggested that competition between cognate and near-cognate tRNAs is a major determining factor in the translation time. In particular, Varenne *et*

al. showed that the period that a ribosome waits for an incoming aa-tRNA is the rate-limiting step during ribosome elongation [73].

Following these assumptions, Fluitt *et al.* introduce a probabilistic, chemical kinetic model capturing both peptide bond formation and ribosome translocation along an mRNA transcript [29]. Their model consists of a series of 18 individual steps that the ribosome makes during the translation of a single codon. The rates of these steps are set by a series of kinetic rate constants.

More precisely, for a given codon i, an aa-tRNA is randomly chosen. This aa-tRNA will follow the ribosome's reaction pathway defined by the model; and depending on whether the chosen aa-tRNA is a cognate, near-cognate, or non-cognate tRNA, the rates at which these steps proceed will vary. Some steps require the calculation to return to the start of the translation process (e.g., an aa-tRNA is rejected in the proofreading step). In any case, as the simulation proceeds, all the time intervals between the reaction steps are added up, yielding a final total codon insertion time after peptide bond formation and ribosome translocation has occurred. These steps were repeated until the total insertion times converged to some average, typically taking about 10^5–10^6 runs.

To make the model biologically realistic, the rate constants chosen were taken from experimental kinetic rate constants at 20°C determined by Gromadski and Rodnina [34] and reevaluated for 37°C. Running these Monte Carlo simulations using these experimentally determined and reevaluated rate constants, the authors derived the following equation for the average translation time $\tau(i)$ of the ith codon at 37° C in milliseconds:

$$\tau(i) = 9.06 + 1.45[10.48C(i) + 0.5R(i)], \tag{4.22}$$

where $C(i)$ is a number quantifying the degree of competition between cognate and near-cognate tRNAs for the ith codon, and $R(i)$ is a number quantifying the degree of competition between cognate and non-cognate tRNAs for the ith codon.

We are interested in a lower bound for the translation time, that is, the smallest time interval possible that a ribosome can translate. Under the assumption that each aa-tRNA that arrives via diffusion is a cognate aa-tRNA, we find that the average translation time for a single codon is $\tau_{ribo} = 9.06$ ms. This value arises because when there is no competition from either near- or non-cognate tRNAs, we have that $C(i) = R(i) = 0$ in Eq. (4.22). Therefore, we interpret τ_{ribo} as a good estimate of the theoretically minimum (i.e., fastest) translation time possible.

Dividing Eq. (4.15) by τ_{ribo}, we arrive at

$$77.5964... \leq \mathcal{C}^*_{ribo} \leq 80.7655... \ \frac{\text{codons}}{\text{second}}, \tag{4.23}$$

where we define $\mathcal{C}^*_{ribo} := \mathcal{C}/\tau_{ribo}$.

Translation rates in prokaryotes lie roughly within 13-22 codons/second [9]. (Eukaryotic translation rates are even slower at ~ 5 codons/s [37, 54].) This range lies below the range of Eq. (4.23) by a large margin, which implies, by Shannon's Noisy Channel Coding theorem, that the ribosome is able to translate at its observed speeds without sacrificing accuracy.

It should be noted that if our assumption that τ_{ribo} is incorrect, then the ribosome's translation time should in principle be even faster, yielding an even higher range than that found in Eq. (4.23).

4.6 Summary

In this chapter, we have shown that the accuracy of the ribosome can be explained through purely information-theoretic means by introducing a new model that views the ribosome as a discrete memoryless channel. The ribosomal channel operates at rates below its capacity in time, allowing it to reliably transmit information with an arbitrary degree of error. We have shown this result by analytically bounding and numerically computing the ribosome's channel capacity and verifying that these values lie above the ribosome's experimentally observed operation rate. Our study is, as far as we know, the first to compare experimentally determined translation rates with a calculated capacity, showing that these rates lie safely below the ribosome's channel capacity.

To summarize, our result successfully explains, from an information-theoretic perspective, existing observations that the ribosome translates accurately at experimentally measured translation rates.

It is worth noting that Shannon's theorem is a nonconstructive theorem. In other words, although the theorem guarantees the existence of a coding scheme that achieves information transmission having an arbitrary degree of error, such a scheme is not specified.

It is well-known that there are many other alternative, naturally occurring genetic codes, with the standard genetic code the most prevalent [53]. For example, vertebral mitochondria utilize a genetic code that maps the codon AUA to the amino acid methionine, whereas the standard genetic code maps AUA to isoleucine. Our method can be extended to other genetic codes by changing the function $\mathcal{G}$ appropriately, and we anticipate

this accommodation may be accomplished at a later time.

Several other questions naturally arise when considering alternative genetic codes in the context of our model. Can the channel capacity be further optimized by choosing a different genetic code? And if so, which one? Is it the standard genetic code? And as we mention above, it is currently unknown whether or not the numerically computed capacity-achieving distribution Q^* is unique. These are some questions that we hope will be addressed in a future study.

The ribosome is found universally across all domains of life, albeit with some variations across these domains. Taken together, our results for the ribosome may serve as a case study of a more general feature of biological machines, namely, that biomolecules, when viewed as information channels, have evolved ways to process information quickly while minimizing errors. One such class of machines may include other enzymes such as DNA polymerases during DNA replication and RNA polymerases during transcription. In fact, these enzymes will be the subject of study in the next chapter.

Chapter 5

The Channel Capacities of DNA Polymerase and RNA Polymerase

In the previous chapter, we showed that the ribosome, the principal enzymatic mediator of translation, can be understood using information theory. In particular, we introduced an information-theoretic model for the ribosome that views it as a communication channel. We calculated bounds for the ribosome's channel capacity, approximated it numerically, and showed that the ribosome translates at speeds far below the capacity.

In this chapter, we turn our attention to the two other important processes in protein synthesis: DNA replication and transcription. From an information-theoretic viewpoint, both processes are similar. Both processes take letters from a nucleotide base alphabet and output a string also taken from a nucleotide base alphabet (although transcription takes adenine to uracil rather than to thymine).

We conduct an information-theoretic analysis for DNA polymerase and RNA polymerase—the enzymatic mediators of replication and transcription, respectively—that is similar to the analysis conducted for the ribosome in Chapter 4. The work conducted in this chapter was performed by the author with supervisory input from Kirkpatrick.

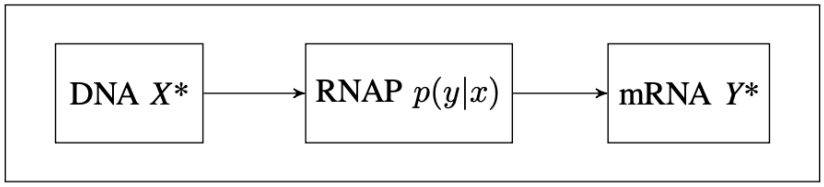

Figure 5.1: RNAP as an information-theoretic channel: input DNA string X^* is transcribed into an output mRNA string Y^*

5.1 DNA & RNA Polymerases as information channels

We model both replication and transcription as a discrete memoryless channel from an input nucleotide alphabet to an output nucleotide alphabet. Our model accommodates both replication and transcription simultaneously because DNA and RNA both use nucleotide alphabets, and distinguishing between the two processes simply amounts to using thymine (T) for replication and uracil (U) for transcription.

Unlike the ribosome, which has a degenerate genetic code, standard base-pairing dictates that outputs have unique inputs. This yields a much simpler information channel model. Our input and output alphabet for replication is $\mathcal{X} := \mathcal{Y} := \{\text{A,C,G,T}\}$. For transcription, the input alphabet is identical, but the output alphabet is $\mathcal{Y} := \{\text{A,C,G,U}\}$. As a channel, DNAP/RNAP has the following conditional distribution:

$$p(y|x) = \begin{cases} 1-r & , \quad y = B(x) \\ \frac{r}{3} & , \quad y \neq B(x), \end{cases} \tag{5.1}$$

where $B : \mathcal{X} \to \mathcal{Y}$ is the standard base-pairing function defined by

$$B(\text{A}) = \text{T/U}, \quad B(\text{C}) = \text{G}, \quad B(\text{G}) = \text{C}, \quad B(\text{T}) = \text{A}, \tag{5.2}$$

where $B(A) = \text{T}$ when considering replication and $B(A) = \text{U}$ when considering transcription. RNAP's probability transition matrix is therefore

$$p(y|x) = \begin{pmatrix} r/3 & r/3 & r/3 & 1-r \\ r/3 & r/3 & 1-r & r/3 \\ r/3 & 1-r & r/3 & r/3 \\ 1-r & r/3 & r/3 & r/3 \end{pmatrix}, \tag{5.3}$$

a symmetric matrix. (See Fig. 5.2 for a transmission diagram of RNAP.) More generally, $p(y|x)$ belongs to a

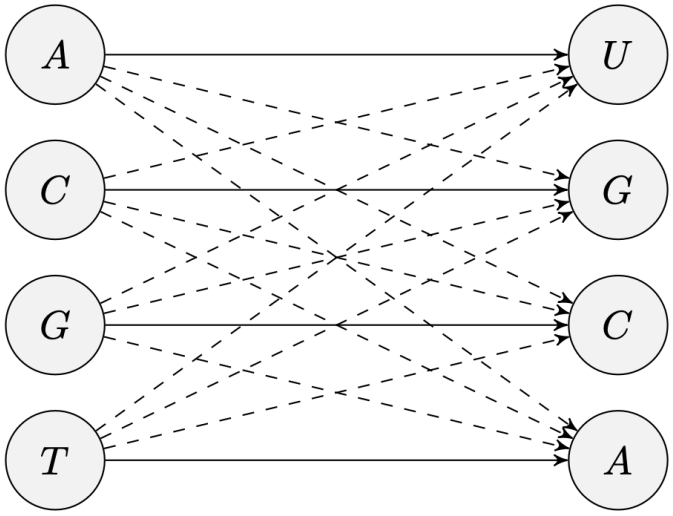

Figure 5.2: Transmission diagram for RNAP, with input symbols on the left and transmitted output symbols on the right. Each horizontal transmission (bold arrows representing correct base pairs) has probability $1-p$. Each diagonal transmission (dashed arrows representing incorrect base pairs) has probability $p/3$. The transmission diagram for DNAP is identical to this one, except that the output uracil (U) on the right is replaced with thymine (T).

class of matrices called weakly symmetric.

Definition 5.1.1. *A channel matrix is called weakly symmetric if each row is a permutation of every other row and the column sums* $\sum_{x\in\mathcal{X}} p(y|x)$ *are equal.*

Example 5.1.2. The matrix

$$p(y|x) = \begin{pmatrix} \frac{1}{4} & \frac{1}{6} & \frac{7}{12} \\ \frac{1}{4} & \frac{7}{12} & \frac{1}{6} \end{pmatrix} \tag{5.4}$$

is a weakly symmetric matrix.

Example 5.1.3 (The ribosome channel is not weakly symmetric)**.** Consider the channel matrix Eq. (4.1). The rows are permutations of each other because $\mathcal{G}$ is a function and is therefore single-valued. However, the column sums are not all equal because $\mathcal{G}$ is not injective (i.e., it is degenerate). For example, the column $p(\text{Met}|x)$ has one entry whose value is $1-r$ because AUG is the only codon mapped to Met by $\mathcal{G}$, and each of the rest of the entries of this column is $\frac{r}{20}$; but the column $p(\text{Leu}|x)$ has six entries whose value is $1-r$ because UUA, UUG, CUU, CUC, CUA, and CUG are all mapped to Leu by $\mathcal{G}$, and each of the rest of the entries of this column is $\frac{r}{20}$. Therefore,

$$\sum_{x\in\mathcal{X}} p(\text{Met}|x) = (1-r) + \frac{63r}{20} \neq \sum_{x\in\mathcal{X}} p(\text{Leu}|x) = 6(1-r) + \frac{58r}{20}. \tag{5.5}$$

Therefore, the ribosome channel is not weakly symmetric.

The following theorem turns out to be a useful theorem for weakly symmetric matrices.

Theorem 5.1.4 (Theorem 7.2.1 in [18])**.** *For a weakly symmetric channel,*

$$\mathcal{C} = \log |\mathcal{Y}| - H(\text{any row of } p(y|x)), \tag{5.6}$$

where $p(y|x)$ is the channel's conditional distribution as a matrix and H is the entropy.

Turning back to DNAP/RNAP's matrix Eq. (5.3), we obtain the following theorem.

Theorem 5.1.5. *The channel capacity of DNAP/RNAP is $\mathcal{C}(r) = 2 + (1-r)\log(1-r) + r\log\frac{r}{3}$.*

Proof. It is clear that Eq. (5.3) is a weakly symmetric matrix. By Theorem 5.2.4, we have that

$$\begin{aligned}
\mathcal{C}(r) &= \log 4 - h\left(\frac{r}{3}, \frac{r}{3}, \frac{r}{3}, 1-r\right) \\
&= 2 + \left\{\frac{r}{3}\log\frac{r}{3} + \frac{r}{3}\log\frac{r}{3} + \frac{r}{3}\log\frac{r}{3} + (1-r)\log(1-r)\right\} \\
&= 2 + r\log\frac{r}{3} + (1-r)\log(1-r)
\end{aligned} \tag{5.7}$$

□

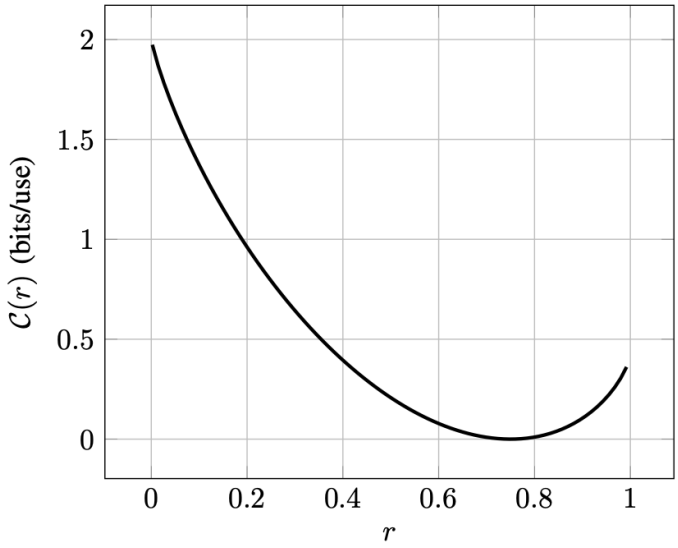

Figure 5.3: Capacity of RNAP as a function of error probability r.

This result recapitulates a result by Djordjevic [24], although we arrive at it by an alternative route.

A plot of $\mathcal{C}(r)$ as a function of the error probability r is given in Fig. 5.3. It is straightforward to show that $\mathcal{C}(r)$ is decreasing on $(0, 3/4)$, is increasing on $(3/4, 1)$, and has a root at $r = 3/4$. As with the ribosome, the root at $r = 3/4$ corresponds to the special case where B maps each DNA base x uniformly to any RNA base and so transmits no information. Average error probabilities p of transcription fall approximately within $10^{-6} \leq r \leq 10^{-5}$ [33, 7, 70], showing that RNAP is nearly optimal as a channel since its capacity is near its maximum. Using these values of r, we see that $\mathcal{C} \approx 0.999804 - 0.999977$ bits/use.

5.2 Explaining experimental observations

Average transcription rates for RNA polymerase during elongation fall within approximately 30-90 nucleotides/second (nt/s) [75]. Monte Carlo simulations similar to those done by the ribosome have suggested that elongation rates can reach speeds greater than 400 nt/s [28].

We can take $\tau_u = 9.06$ ms—the operation time of the ribosome—as a first approximation for an upper bound on the operation time of a single use of RNAP as a channel, τ_{RNAP}. This estimation is justified by noting that incoming free nucleotides diffuse faster than free aa-tRNA molecules due to nucleotides' smaller molecular masses. Thus, for a lower bound τ_l of τ_{RNAP}, we take $\tau_l = 0.5$ ms as a rough estimate. Calculating $\mathcal{C}(r)$ for an error probability of $p = 1 \times 10^{-5}$ using Theorem 5.1.5, we divide this resulting value by τ_u and τ_l to obtain a likely capacity range for RNAP:

$$110.36... \leq \mathcal{C}^*_{RNAP} \leq 999.90... \ \frac{\text{nucleotides}}{\text{second}}, \tag{5.8}$$

where we have defined $\mathcal{C}^*_{RNAP} := \mathcal{C}/\tau_{RNAP}$ and used the fact that there are $\log 4 = 2$ bits/nucleotide.

The expression Eq. (5.8) shows that $\mathcal{C}^*_{RNAP}$ upper bounds the rate of 30-90 nt/s, demonstrating that RNAP can transcribe with an arbitrarily small error probability. Experimental fluorescence studies of a type of RNAP known as RNA polymerase II in eukaryotic cells have found that elongation rates can reach speeds in excess of approximately 1000 nt/s, at least in short bursts [47]. This suggests that $\mathcal{C}^*_{RNAP}$ may be even higher, providing RNAP with a considerably wide range of speeds where it can still transcribe reliably. This result could be accommodated by taking a smaller value for τ_l, once there are tighter experimental results for the speed of nucleotides passing through the reading frame.

One should also observe that the range in Eq. (5.8) gives bounds on an upper bound, namely the capacity $\mathcal{C}^*_{RNAP}$. Thus, observed rates faster than, for example, ~ 110.36 nt/s (yet lower than ~ 999.90 nt/s) might

lie below the capacity and may therefore still satisfy the hypotheses of Shannon's Noisy Channel Coding Theorem.

In this section, we have only explained experimental observations for RNAP during transcription. A next step would be to obtain an analogous result for DNAP and replication, which we hope to accomplish in future work.

5.3 Summary

In this chapter, we have introduced a model of the DNA polymerase and RNA polymerase as discrete memoryless channels. We have calculated explicit formulas for their channel capacities and converted RNAP's capacity to a capacity in time.

By comparing the capacity in time of RNAP to experimentally observed transcription rates, we have estimated that the observed rates fall below the capacity. **This result suggests that RNA polymerase safely operates below its capacity, enabling RNAP to transmit information with arbitrarily low error, as stated in Shannon's Noisy Channel Coding theorem.**

Chapter 6

Mereological measures on a Finite Space

6.1 Introduction

Mereology (derived from the ancient Greek word $\mu\varepsilon\rho o\varsigma$, meaning "part") is the philosophical study of parthood and the relations of constituent parts to their wholes [74]. Traditionally, mereology is concerned with abstract concepts such as the famed "Ship of Theseus," which attempts to determine whether or not a system, in this case a ship, remains the same whole even if some of its individual parts are replaced, such as the ship's sail [10].

Mereology is a subdiscipline of ontology, the philosophical study of being and reality [65]. Ontology studies the existence of different entities and, ultimately, how these entities may be categorized, their properties, and how they are related to each other. For example, a philosopher may ask, "What is the atmosphere? What is it made of, and how does it connect to both living and non-living entities on the earth?" One way to visualize an ontological perspective is a Porphyrian Tree (Fig. 6.1), a diagram that classifies different entities and separates them into distinct subclasses according to distinguishing properties. A Porphyrian Tree clarifies how different entities are related to each other.

In informatics, ontology acquires a more concrete meaning, although like the ontology of philosophy, there are many ways to interpret the ontology of informatics. In general, ontology refers to a systematic way of organizing data to facilitate the aggregation, annotation, storage, and retrieval of these data for practical use

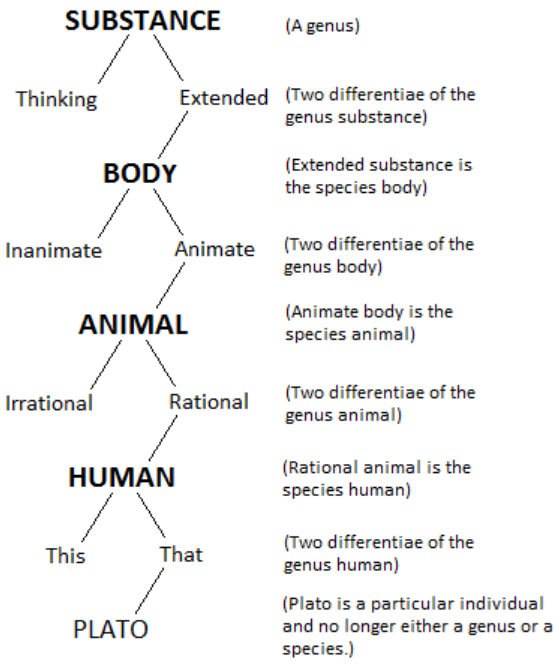

Figure 6.1: Diagram of a Porphyrian Tree. Different classes are represented in bold and distinguishing features, or differentiae, are separated out (diagonal lines). As one descends the tree, classes become more narrow and specific. Here, the entity "Plato" belongs to the class "Human," which belongs to the class "Animal," and so on. Courtesy of *Voice of the Commons* under a CC BY-SA 3.0 License.

[49]. This interpretation of ontology—sometimes called *applied ontology* to distinguish it from philosophical ontology—is useful in knowledge representation and reasoning, a field of computer science that aims to symbolize information in ways that a computer can use to solve complex tasks. Just as philosophical ontology describes the natures of various realities and their relationships, applied ontology captures these natures in terms of data. That is, ontology describes the real world using symbolic data. An ontology can also refer to a controlled vocabulary, a set of keywords or index terms that "point to" or "label" specific files (e.g., a document) in a search engine. In short, an ontology is a way to specify a set of ideas and the relationships between them.

Ontologies facilitate data processing by creating relevant associations between different entities and the processes they undergo across all different levels of granularity, or scales. In a biological context, one might define entities such as "cell" or "cell nucleus," as well as "part of" relations to declare that a cell nucleus is part of a cell. By associating these entities with certain biological processes, one could deduce, say, a particular diagnosis given certain symptoms or observations—a kind of automated biomedical reasoning.

Mereological and ontological applications in biology

Living systems are often characterized by complex networks or hierarchies such as phylogenetic trees and taxonomies (e.g., species → genus → family → order → class → phylum → kingdom → domain) or systems hierarchies (e.g., organelle → cell → tissue → organ → organ systems → organisms → populations → communities → ecosystems → biospheres). It is only natural to import ontological principles into biology to deal with these complex hierarchies. Documenting the number of parts (e.g., organs, enzymes)—let alone all of their interactions—of biological systems is becoming increasingly computationally expensive to sort through. Thus, as the amount of biological data increases, ontology-based formalisms are finding increased use in biological and biomedical informatics and electronic healthcare records to uncover relationships between biological processes, diseases, symptoms, and more [13, 58]. A well-known example of an ontology is the Gene Ontology, which categorizes genes and their interactions [16]. A key goal in these fields is to bridge the gap between molecular biological data and clinical medicine, such as in diagnostics, treatment, and drug design. It would be useful to develop a consistent, logical ontology that—in addition to being user-friendly for both biologists and clinicians—makes inferences satisfying biomedical realities, consensus, and intuition.

Inspired by mereological formalisms in biology, we expand upon a model of parthood on a finite space first introduced by Schumm *et al.* [61], who defined a family of measures assigning a weight to different blocks of sets when this finite space is partitioned. Furthermore, we establish a Radon-Nikodym-type theorem that gives an explicit relationship between different measures. Finally, we conclude by proposing a way that this mereological model could be applied in biomedical knowledge representation.

This chapter builds upon the work "Composition and Trans-Scalar Identity," by A. Schumm, W. Rohloff, and G. Piccinini [61]. The author (Inafuku) refined some proofs found in the above paper and proved some original results, including Theorem 6.4.3, Proposition 2, and Theorem 6.4.8. Kay L. Kirkpatrick contributed Examples 6.4.4 and 6.4.5.

6.2 Mathematical preliminaries

In Section 6.4, we will introduce a family of measures on a finite space and use these measures to assign a "size" to different sets that represent different parts of our whole through the subset relation. We begin with a brief survey of the relevant measure-theoretic tools that we will use. For a comprehensive treatment of measure theory, we refer the interested reader to [30]. For the remainder of this chapter, we denote any

nonempty set by Ω and its power set of Ω (i.e., the set of all subsets of Ω) by 2^Ω.

Definition 6.2.1. *A* ***σ-algebra*** *is a set $\mathcal{F} \subseteq 2^\Omega$ such that*

1. *$\Omega \in \mathcal{F}$,*
2. *if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$, and*
3. *if $\{E_n\}_{n=1}^\infty \subseteq \mathcal{F}$, then $\bigcup_{n=1}^\infty E_n \in \mathcal{F}$.*

We call the elements of a σ-algebra **measurable sets**, and we call the pair $(\Omega, \mathcal{F})$ a **measurable space**.

Definition 6.2.2. *A* ***measure*** *is a function $\mu : \mathcal{F} \to [0, \infty]$ such that*

1. *$\mu(\varnothing) = 0$, and*
2. *if $\{E_n\}_{n=1}^\infty \subseteq \mathcal{F}$ is a sequence of pairwise disjoint sets, then*

$$\mu\left(\bigcup_{n=1}^{\infty} E_n \right) = \sum_{n=1}^{\infty} \mu(E_n).$$

Given a nonempty set Ω, a σ-algebra $\mathcal{F}$, and a measure $\mu : \mathcal{F} \to [0, \infty]$, we call the triple $(\Omega, \mathcal{F}, \mu)$ a **measure space**.

Example 6.2.3. In probability theory, one often considers a measure space $(\Omega, \mathcal{F}, \mathbb{P})$. Ω is called the sample space, whose elements are called outcomes; $\mathcal{F}$ is a σ-algebra of events; and $\mathbb{P}$ is a measure satisfying $\mathbb{P}(\Omega) = 1$, which we call a probability measure.

Example 6.2.4 (Counting and Dirac measures)**.** Let Ω be nonempty, $\mathcal{F} = 2^\Omega$ be the power set of Ω, and $f : \Omega \to [0, \infty]$. Define $\mu : \mathcal{F} \to [0, \infty]$ by

$$\mu(A) := \sum_{x \in A} f(x). \tag{6.1}$$

Then μ is a measure on $\mathcal{F}$. There are two interesting special cases.

- Define f by $f(x) = 1$ for all $x \in \Omega$. Then

$$\mu(A) = \begin{cases} |A| \ , & A \text{ is finite} \\ \infty \ , & \text{otherwise.} \end{cases} \tag{6.2}$$

This measure is called the counting measure.

- Let $x_0 \in \Omega$. Define f by

$$f(x) = \begin{cases} 1 \ , & x = x_0 \\ 0 \ , & x \neq x_0. \end{cases} \tag{6.3}$$

Then if we denote the measure in Eq. (6.1) by $\delta_{x_0} := \mu$, then

$$\delta_{x_0}(A) = \begin{cases} 1 \ , & x_0 \in A \\ 0 \ , & x_0 \notin A. \end{cases} \tag{6.4}$$

This measure is called the Dirac measure, and is a rigorous way to define the Dirac delta function, which is often used in physics.

Example 6.2.5 (Lebesgue measure)**.** Among the most well-known measures is the Lebesgue measure m, which generalizes the notion of length. For an interval $(a, b) \subseteq \mathbb{R}$, where $a < b$, the Lebesgue measures gives $m((a, b)) = b - a$.

There are many types of interesting classes of measures. Given a measure space $(\Omega, \mathcal{F}, \mu)$, μ is said to be finite if $\mu(\Omega) < \infty$. Another interesting class is given in the next definition.

Definition 6.2.6. *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. We call the measure μ σ-finite if Ω is a countable union of measurable sets, each of which has finite measure. That is, μ is σ-finite if there exists a sequence $\{A_n\}_{n=1}^{\infty} \subseteq \mathcal{F}$ such that $\Omega = \bigcup_{n=1}^{\infty} A_n$ and $\mu(A_n) < \infty$ for each $n \in \mathbb{N}$.*

Example 6.2.7 (Lebesgue measure is σ-finite but not finite)**.** It is a standard result in measure theory that the Lebesgue measure m is σ-finite. Notice that

$$\mathbb{R} = \bigcup_{n \in \mathbb{Z}} [n, n + 1). \tag{6.5}$$

Although $m(\mathbb{R}) = \infty$, we have that $m([n, n + 1)) = 1 < \infty$ for each $n \in \mathbb{Z}$.

6.2.1 Integrating measurable functions

Given a measure space $(\Omega, \mathcal{F}, \mu)$, we can define an integral. Such an integral will allow us to compare two measures with each other.

We first recall that for $E \subseteq \Omega$, the characteristic (or indicator) function of E, $\chi_E : \Omega \to \mathbb{R}$, is given by

$$\chi_E(x) = \begin{cases} 1 \ , & x \in E \\ 0 \ , & x \notin E. \end{cases} \tag{6.6}$$

We now make the following definition towards the construction of an integral on the measure space.

Definition 6.2.8. *Let $(\Omega, \mathcal{F})$ be a measurable space. A simple function ϕ on Ω is a finite linear combination of characteristic (or indicator) functions χ_{E_j} having complex coefficients, where $E_j \in \mathcal{F}$, i.e.,*

$$\phi = \sum_{j=1}^{n} a_j \chi_{E_j} \tag{6.7}$$

for some $n \in \mathbb{N}$ and where $a_j \in \mathbb{R}$ for all j.

We define the integral of a simple function ϕ with respect to the measure μ as

$$\int_\Omega \phi \, d\mu = \int \phi \, d\mu := \sum_{j=1}^{n} a_j \mu(E_j). \tag{6.8}$$

We can now generalize the definition of the integral for simple functions to the class of all measurable functions[1] from Ω to $[0, \infty]$. For a measurable function $f : \Omega \to [0, \infty]$, we define

$$\int_\Omega f \, d\mu = \int f \, d\mu := \sup \left\{ \int \phi \, d\mu : 0 \leq \phi \leq f, \, \phi \text{ is simple} \right\}. \tag{6.9}$$

Moreover, we can extend this definition to any measurable function $f : \Omega \to [-\infty, \infty]$ by writing $f = f^+ - f^-$, where f^+ is the positive part of f and f^- is the negative part of f:

$$f^+(x) := \begin{cases} f(x) \ , & f(x) > 0 \\ 0 \ , & \text{otherwise,} \end{cases} \qquad f^-(x) := \begin{cases} -f(x) \ , & f(x) < 0 \\ 0 \ , & \text{otherwise.} \end{cases} \tag{6.10}$$

Notice that both f^+ and f^- are non-negative. We can also define the integral for any other measurable set $A \in \mathcal{F}$:

$$\int_A f \, d\mu := \int_\Omega f \chi_A \, d\mu = \int f \chi_A \, d\mu. \tag{6.11}$$

[1] For the sake of brevity, we leave it to the reader to recall the definition of a measurable function. A suggested reference is Ref. [30].

Later, we will recall a mereological model of parthood from [61], in whose setting we will define a family of finite measures. We will then use the integral definition above to relate different measures to each other.

Example 6.2.9. Consider the measure space $(\mathbb{N}, 2^{\Omega}, \mu)$, where μ is the counting measure introduced in Example 6.2.5. Given a measurable function $f : \mathbb{N} \to \mathbb{R}$, the integral with respect to the counting measure yields the sum

$$\int_{\mathbb{N}} f \, d\mu = \sum_{n=1}^{\infty} f(x). \tag{6.12}$$

Example 6.2.10. Let $x \in \Omega$. Recall Eq. (6.4), the Dirac measure, from Example 6.2.4. One can show that for a measurable function f, one has

$$\int f \, d\delta_x = f(x). \tag{6.13}$$

Example 6.2.11. A standard result in measure theory says that if a bounded function $f : [a, b] \to \mathbb{R}$ is Riemann integrable, then it is Lebesgue integrable. In particular, that integral is given by

$$\int_a^b f(x) \, dx = \int_{[a,b]} f \, dm, \tag{6.14}$$

where m is the Lebesgue measure. Thus, the Lebesgue integral subsumes the Riemann integral.

One of our main results is a Radon-Nikodym-type theorem for a family of measures on a finite measure space, so we restate the classic theorem here. First, we recall that for two measures μ and ν on a measurable space $(\Omega, \mathcal{F})$, μ is called **absolutely continuous with respect to** ν (denoted $\mu \ll \nu$) if $\mu(A) = 0$ whenever $\nu(A) = 0$ for all $A \in \mathcal{F}$. Moreover, we call μ and ν **equivalent** if both $\mu \ll \nu$ and $\nu \ll \mu$. Finally, it is useful to remind ourselves that, given a measure space $(\Omega, \mathcal{F}, \mu)$, a statement is true **almost everywhere** (denoted μ-a.e.) if the statement holds for all elements in Ω except for those elements in a set $N \in \mathcal{F}$ satisfying $\mu(N) = 0$. (Recall that any measurable set $N \in \mathcal{F}$ satisfying $\mu(N) = 0$ is called a null set.)

Theorem 6.2.12 (Radon-Nikodym)**.** *Let μ, ν be two σ-finite measures on a measurable space $(\Omega, \mathcal{F})$. If $\mu \ll \nu$, then there exists a function $f : \Omega \to [0, \infty)$ such that*

$$\nu(E) = \int_E f \, d\mu, \quad \forall E \in \mathcal{F}. \tag{6.15}$$

Moreover, f is unique up to a null set, i.e., if $g : \Omega \to [0, \infty)$ is another function such that $\nu(E) = \int_E g \, d\mu$, then $f = g$ μ-a.e.

We call f the **Radon-Nikodym derivative of ν with respect to μ** and denote it by $\frac{d\nu}{d\mu} := f$.

6.3 Model of Parthood

Before we state a model of parthood, we first recall the standard definition of a partition in set theory.

Definition 6.3.1. *A partition $P \subseteq 2^\Omega$ of Ω is a collection of subsets of Ω such that*

1. $\varnothing \notin P$,
2. *for all $B_i, B_j \in P$, $B_i = B_j$ or $B_i \cap B_j = \varnothing$ for all $i \neq j$, and*
3. $\bigcup_{B \in P} B = \Omega$.

We call the elements of P **blocks**. Furthermore, to compare two different partitions P and Q of a nonempty set Ω, we say that Q is **finer** than P if $\forall D \in Q$, $\exists B \in P$ such that $D \subseteq B$. Equivalently, we say that P is **coarser** than Q and that Q is a **refinement** of P.

When defining composition as identity (CAI), Schumm *et al.* write

$$w = p_1 \cdots p_n$$

to mean that a whole w is made up of parts $p_1, p_2, \ldots$ all the way up to and including p_n. We identify the "parts" of Ref. [61] with the blocks of a partition given by Definition 6.3.1. In particular, the union of all blocks recovers the original set Ω:

$$\bigcup_{i=1}^{k} B_i = \Omega \tag{6.16}$$

for a partition $P = \{B_1, \ldots, B_k\}$. This identification is natural since we would like to assign a size to each block, which can be accomplished by a measure when the cells are elements of a σ-algebra, i.e., measurable sets.

Schumm *et al.* explain in Ref. [61] that blocks are measurement scales in worlds with finitely many atoms. To demonstrate this, they work with the following (modified) definition.

Definition 6.3.2. *Fix an integer $m \geq 1$. A world is a 4-tuple $\mathbf{G} := (\Omega, \hat{\mathcal{P}}, \mathcal{C}, \mathcal{R})$, where*

1. *Ω is some nonempty universal set whose elements are called atoms,*

2. $\hat{\mathcal{P}} := \{P_1, \ldots, P_m\}$ *is a set of partitions of* Ω,

3. $\mathcal{C} := \{C_1, \ldots, C_m\}$ *is a collection where each* C_i*—called the combination set of the partition* P_i*—is defined by*

$$C_i := \left\{ \bigcup_{B \in \mathcal{A}} B \,\middle|\, \mathcal{A} \subseteq P_i \right\}$$

where P_i *is a partition, and*

4. $\mathcal{R} := \{|\cdot|_1, \ldots, |\cdot|_m\}$ *is a set of functions where each* $|\cdot|_i : C_i \to \mathbb{N} \cup \{0\}$ *assigns to each* $U \in C_i$ *the value*

$$|U|_i = \begin{cases} 0, & \text{if } U = \varnothing \\ q, & \text{if } U = B_{l_1} \cup \cdots \cup B_{l_q} \text{ for distinct } B_{l_1}, \ldots, B_{l_q} \in P_i. \end{cases}$$

Given $U \in C_i$*, we call* $|U|_i$ *the length of* U *relative to the partition* P_i*.*

For clarity, for a given combination set, C, and an element $U \in C$, we call

$$U = \bigcup_{j=1}^{q} B_{l_j} \tag{6.17}$$

the *block representation of* U. Since the blocks are pairwise disjoint by definition of partition, each $U \in C$ has a unique block representation (modulo the order of the blocks). So, each $|\cdot|_i$ is well-defined.

Since we consider finite partitions, the following proposition immediately follows.

Proposition 1. *Let* P *be a partition and* C *its combination set. Then* $C = \sigma(P)$.

A corollary to Proposition 1 is the following.

Corollary 1. $|\cdot|_i$ *is a measure on* C_i *for each* i.

It is worth noting that $|\cdot|_i$ makes precise the notion of "has more proper parts than" **relative to a given partition** P_i. **Each partition counts each of its members as a single object**.

In mereology, a common way to build bigger parts from smaller (atomic) parts is by an operation called "fusion," sometimes denoted by $\circ$. In this article, we identify the fusion operation on atoms with the union operation over sets, $\cup$. Moreover, given a partition P and its combination set C, we call any set $U \in C \backslash P$ a

composite or *composite object*. Composites are what Schumm *et al.* call "a fusion of atoms with atoms, a fusion of atoms with composite objects, or a fusion of multiple composite objects."

Since we are dealing with a world having a finite set of atoms, it is worth recognizing two special partitions. For a world having n atoms, so that $\Omega = \{a_1, \ldots, a_n\}$, the *finest partition* P_f is the partition whose elements are singleton blocks, one for each atom. That is to say, $P_f := \{\{a_1\}, \ldots, \{a_n\}\}$. We also define the *coarsest partition* P_c as the partition whose only element is the union of all the atoms, that is, $P_c := \{\{a_1, \ldots, a_n\}\}$.

Example 6.3.3. Let $\Omega := \{a_1, a_2, a_3\}$ be the universe having three atoms. The finest partition and its associated combination set are, respectively, $P_f = \{\{a_1\}, \{a_2\}, \{a_3\}\}$ and $C_f = 2^\Omega$. Similarly, the coarsest partition and its associated combination set are, respectively, $P_c = \{\{a_1, a_2, a_3\}\}$ and $C_c = \{\varnothing, \{a_1, a_2, a_3\}\} = \{\varnothing, \Omega\}$, the trivial σ-algebra.

It is easy to see that $|\{a_1, a_2\}|_f = 2$ and $|\{a_1\}|_f = 1$. So, $|\{a_1, a_2\}|_f > |\{a_1\}|_f$. Additionally, $|\{a_1, a_2, a_3\}|_f = 3$ and $|\{a_1, a_3\}|_f = 2$. So, $|\{a_1, a_2, a_3\}|_f > |\{a_1, a_3\}|_f$.

Now, define a partition $P_2 := \{\{a_1, a_2\}, \{a_3\}\}$. So, $C_2 = \{\varnothing, \{a_1, a_2\}, \{a_3\}, \{a_1, a_2, a_3\}\}$. It follows that $|\{a_1, a_2, a_3\}|_2 = 2$ and $|\{a_1, a_2\}|_2 = 1$. Thus, $|\{a_1, a_2, a_3\}|_2 > |\{a_1, a_2\}|_2$. Notice, however, that $|\{a_3\}|_2 = 1$. Therefore, $|\{a_1, a_2\}|_2 \not> |\{a_3\}|_2$.

Recall that for any subset $A \in \Omega$, we denote by $\sigma(A)$ the smallest σ-algebra containing A, which we call the σ-algebra generated by A. Generalizing Example 6.3.3 to the case of n atoms, we obtain the following fact.

Fact 1. *$C_f = 2^\Omega$ and $C_c = \{\varnothing, \Omega\}$, the trivial σ-algebra.*

According to Schumm *et al.*, "Measurement is the assignment of a numerical scale to some objects in a way that preserves (certain) relations on the objects." And a given partition can define the scale used to measure a "portion of reality." Thus, given a partition P, it seems natural to define a measure on $\sigma(P)$. Our challenge is to define a finite measure(s) $\mu_P : 2^\Omega \to [0, \infty)$ that preserves some (certain) relations. For simplicity, we begin in the finite setting.

6.4 Main results

Definition 6.4.1. *Let P be a partition of Ω. Define $f_P : 2^\Omega \to [0, \infty)$ by*

$$f_P(A) := \begin{cases} 0, & \text{if } A = \varnothing \\ \frac{1}{|B_1|} + \cdots + \frac{1}{|B_k|}, & \text{if } A = \{a_1, \ldots, a_k\}, \text{and } \{a_i\} \subseteq B_i \text{ for some } B_i \in P \end{cases} \tag{6.18}$$

We call f_P **the representation function induced by the partition** P. Notice in Definition 6.4.1 that because each atom belongs to exactly one block of the partition P, the blocks B_i exist so that f_P is well-defined.

The following lemma is useful in showing that each representation function is a finite measure.

Lemma 6.4.2. *Let P be a partition. Then $f_P(B) = 1$ for all $B \in P$.*

Proof. Let $B \in P$ be an arbitrary partition. Then $B = \{a_1, \ldots, a_k\}$ for some $k \in \mathbb{N}$. So, $|B| = k$. We have

$$f_P(B) = f_P(\{a_1, \ldots, a_k\}) = \underbrace{\frac{1}{|B|} + \cdots + \frac{1}{|B|}}_{k \text{ terms}} = \underbrace{\frac{1}{k} + \cdots + \frac{1}{k}}_{k \text{ terms}} = k \cdot \frac{1}{k} = 1.$$

□

Theorem 6.4.3. *Let P be a partition of Ω. Then f_P is a finite measure on 2^Ω.*

Proof. It follows immediately from the definition of f_P that $f_P(\varnothing) = 0$. It remains to show that f_P is countably additive.

Let $\{A_n\}_{n=1}^{\infty} \subseteq 2^\Omega$ be a collection of disjoint subsets. Since Ω is finite, we may assume without loss of generality that $\exists N \in \mathbb{N}$ such that $n \geq N \Rightarrow A_n = \varnothing$. In particular, since each A_n is finite for $1 \leq n \leq N - 1$,

we may assume that $A_n = \{a_{n,1}, \ldots, a_{n,k_n}\}$ for some $k_n \in \mathbb{N}$ and $1 \leq n \leq N-1$. So, we have

$$\begin{aligned}
f_P\Big(\bigcup_{n=1}^{\infty} A_n\Big) &= f_P\Big(\bigcup_{n=1}^{N-1} A_n\Big) \\
&= f_P(\{a_{1,1}, \ldots, a_{1,k_1}, a_{2,1}, \ldots, a_{2,k_2}, \ldots, a_{N-1,1}, \ldots, a_{N-1,k_{N-1}}\}) \\
&= f_P(\{a_{1,1}\}) + \cdots + f_P(\{a_{1,k_1}\}) + \cdots + f_P(\{a_{N-1,1}\}) + \cdots + f_P(\{a_{N-1,k_{N-1}}\}) \\
&= f_P(\{a_{1,1}, \ldots, a_{1,k_1}\}) + \cdots + f_P(\{a_{N-1,1}, \ldots, a_{N-1,k_{N-1}}\}) \\
&= f_P(A_1) + \cdots + f_P(A_{N-1}) \\
&= \sum_{n=1}^{N-1} f_P(A_n) + \sum_{n \geq N} f_P(A_n) \\
&= \sum_{n=1}^{\infty} f_P(A_n).
\end{aligned}$$

So f_P is a measure. For finiteness, let $P = \{B_1, \ldots, B_n\}$ be an arbitrary partition. By definition of partition, $\Omega = \bigcup_{i=1}^{n} B_i$, and all the B_is are mutually disjoint. Using the fact that f_P is a measure and Lemma 6.4.2 yields

$$f_P(\Omega) = f_P\Big(\bigcup_{i=1}^{n} B_i\Big) = \sum_{i=1}^{n} f_P(B_i) = \sum_{i=1}^{n} 1 = n < \infty.$$

□

The next two examples illustrate specific representation functions for two special partitions, the finest and the coarsest partitions.

Example 6.4.4 (Finest partition for N atoms)**.** The finest partition for an N-atom world is $P_f = \{\{a_1\}, \ldots, \{a_N\}\}$. For an atomic singleton $A = \{a_j\}$, $B = \{a_j\}$ is the only block in P_f such that $A \subseteq B$, so $f_{P_f}(\{a_j\}) = 1$. Also, $f_{P_f}(\{a_{j_1}, a_{j_2}\}) = 1 + 1 = 2$. More generally, for any $A \in 2^{\Omega}$, we have $f_{P_f}(A) = |A|$. Therefore, f_{P_f} is the counting measure.

Example 6.4.5 (Coarsest partition for N atoms)**.** The coarsest partition for an N-atom world is $P_c = \{\{a_1, \ldots, a_N\}\}$. For an atomic singleton $A = \{a_j\}$, $B = \{a_1, \ldots, a_N\}$ is the only block in P_c and thus the only block such that $A \subseteq B$. So, $f_{P_c}(\{a_j\}) = \frac{1}{N}$. Also, $f_{P_c}(\{a_{j_1}, a_{j_2}\}) = \frac{1}{N} + \frac{1}{N} = \frac{2}{N}$. More generally, for $1 \leq n \leq N$, we have

$$f_{P_c}(\{a_{j_1}, \ldots, a_{j_n}\}) = \frac{n}{N}.$$

Thus, $f_{P_c}(A) = \frac{|A|}{N}$. Therefore, f_{P_c} is the uniform measure.

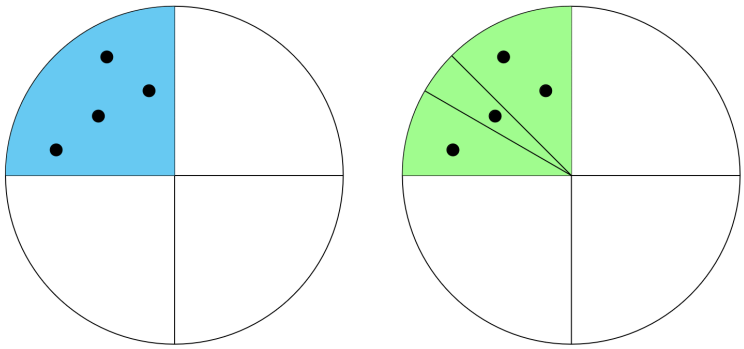

Figure 6.2: Two different partitions of the same universal set. The partition on the right is a finer partition than the one on the left.

Remark 1. By Theorem 6.4.3, it is easy to see in Example 6.4.4 (resp. 6.4.5) that the partition that yields the maximum (resp. minimum) size of $\Omega = \{a_1, \ldots, a_N\}$ is the finest (resp. coarsest) partition: $f_{P_f}(\Omega) = N$ (resp. $f_{P_c}(\Omega) = 1$).

Proposition 2. *Let P, Q be partitions of Ω with Q finer than P. Then $f_P(E) \leq f_Q(E)$ for all $E \in 2^\Omega$.*

Proof. Let $E \in 2^\Omega$ be arbitrary. If $E = \varnothing$, then $f_Q(E) = f_P(E) = 0$ by definition of measure. If $E \neq \varnothing$, then $E = \{a_1, \ldots, a_k\}$ for some $k \in \mathbb{N}$. By definition of partition, $\exists D_i \in Q$ such that $\{a_i\} \subseteq D_i, \forall i$. Since Q is finer than P, $\exists B_i \in P$ such that $D_i \subseteq B_i, \forall i$. So, $|D_i| \leq |B_i|$. Then

$$f_Q(E) = f_Q(\{a_1, \ldots, a_k\}) = \frac{1}{|D_1|} + \cdots + \frac{1}{|D_k|} \geq \frac{1}{|B_1|} + \cdots + \frac{1}{|B_k|} = f_P(E).$$

□

Essentially, this proposition says that the "size" of a set with respect to a coarser partition is smaller than the "size" of the same set with respect to the finer partition. For example, if we are measuring a rod using two different scales—say, meters and millimeters—the value for the magnitude of the length of the rod in meters will be less than the corresponding value in millimeters. Of course, this setting is in the continuous, uncountable case, but the intuition still holds for the finite case. The next example illustrates this concept.

Example 6.4.6. Let $\Omega = \{a_1, a_2, a_3, a_4, \ldots, a_n\}$. We break Ω into two partitions $P = \{B_1, B_2, B_3, B_4\}$, where $B_1 = \{a_1, a_2, a_3, a_4\}$. (see Fig. 1(A)) and $Q = \{D_1, D_2, D_3, D_4, D_5, D_6\}$, where $D_1 = \{a_1\}, D_2 = \{a_2\}, D_3 = D_4 = \{a_3, a_4\}$ (see Fig. 1(B)). Notice that Q is finer than P. For $E = \{a_1, a_2, a_3, a_4\} = B_1$, we have

$$f_P(E) = \frac{1}{|B|} + \frac{1}{|B|} + \frac{1}{|B|} + \frac{1}{|B|} = 1,$$

$$f_Q(E) = \frac{1}{|D_1|} + \frac{1}{|D_2|} + \frac{1}{|D_3|} + \frac{1}{|D_4|} = 1 + 1 + \frac{1}{2} + \frac{1}{2} = 3.$$

We see that $f_P(E) = 1 < f_Q(E) = 3$, which agrees with Proposition 4.4.

Theorem 6.4.7 (Representation Theorem)**.** *Let P_i be a partition and C_i its combination set. Then for all $A_1, A_2 \in C$, $|A_1|_i \leq |A_2|_i \iff f_{P_i}(A_1) \leq f_{P_i}(A_2)$ with equality iff $|A_1|_i = |A_2|_i$.*

Proof. The statement clearly holds for $A_1 = \varnothing, A_2 \neq \varnothing$. So, without loss of generality we may assume that A_1 and A_2 are both nonempty. We have that $A_1 = B_1 \cup \cdots \cup B_n$ and $A_2 = D_1 \cup \cdots \cup D_m$ for some blocks $B_1, \ldots, B_n, D_1, \ldots, D_m \in P$, $n, m \in \mathbb{N}$. That is, $|A_1|_i = n$ and $|A_2|_i = m$. Since $f_P(B) = 1$ for all $B \in P_i$ and f_{P_i} is a measure, we have

$$|A_1|_i < |A_2|_i \iff f_{P_i}(A_1) = \sum_{j=1}^{n} f_{P_i}(B_j) = n < m = \sum_{j=1}^{m} f_{P_i}(D_j) = f_{P_i}(A_2).$$

By a nearly identical argument, we also obtain $|A_1|_i = |A_2|_i \iff f_{P_i}(A_1) = f_{P_i}(A_2)$. □

Theorem 6.4.8. *Let P, Q be two partitions of $\Omega = \{a_1, \ldots, a_n\}$, and let f_P, f_Q be their representation functions. Then the Radon-Nikodym derivative of f_P with respect to f_Q exists. Moreover, the derivative is a simple function and is given by*

$$\frac{df_P}{df_Q} = \sum_{j=1}^{n} \frac{|D_j|}{|B_j|} \chi_{\{a_j\}},$$

where B_j and D_j are (the unique) blocks of the partitions P, Q, respectively, such that $a_j \in B_j \cap D_j$, and $\chi_{\{a_j\}}$ denotes the indicator function of $\{a_j\}$.

Proof. Let P and Q be two arbitrary partitions of a nonempty, finite set Ω. Clearly, f_P, f_Q are σ-finite since they are finite. Since $f_P \ll f_Q$ by Lemma 4.10, $\exists g : \Omega \to [0, \infty)$ such that

$$f_P(E) = \int_E g \, df_Q = \int_E g(\omega) \, df_Q(\omega), \ \ \forall E \in 2^\Omega$$

by the Radon-Nikodym theorem.

Note that $f_P(\varnothing) = \int_\varnothing h \, df_Q = 0$ for any measurable function h. WLOG we may assume that (upon sufficient reordering) $E = \{a_1, \ldots, a_k\}$ is an arbitrary element of 2^Ω for some $k \in \mathbb{N}$. As in the statement of the theorem, take

$$g = \sum_{j=1}^{n} \frac{|D_j|}{|B_j|} \chi_{\{a_j\}}. \tag{6.19}$$

We have

$$
\begin{aligned}
f_P(E) &= \int_E g\, df_Q = \int_E \left(\sum_{j=1}^{n} \frac{|D_j|}{|B_j|} \chi_{\{a_j\}} \right) df_Q \\
&= \int_E \left(\frac{|D_1|}{|B_1|} \chi_{\{a_1\}} + \cdots + \frac{|D_n|}{|B_n|} \chi_{\{a_n\}} \right) df_Q \\
&= \frac{|D_1|}{|B_1|} \int_E \chi_{\{a_1\}}\, df_Q + \cdots + \frac{|D_k|}{|B_k|} \int_E \chi_{\{a_k\}}\, df_Q + \frac{|D_{k+1}|}{|B_{k+1}|} \int_E \chi_{\{a_{k+1}\}}\, df_Q \\
&\quad + \cdots + \frac{|D_n|}{|B_n|} \int_E \chi_{\{a_n\}}\, df_Q \\
&= \frac{|D_1|}{|B_1|} \int \chi_{\{a_1\} \cap E}\, df_Q + \cdots + \frac{|D_k|}{|B_k|} \int \chi_{\{a_k\} \cap E}\, df_Q \\
&\quad + \frac{|D_{k+1}|}{|B_{k+1}|} \int \chi_{\{a_{k+1}\} \cap E}\, df_Q + \cdots + \frac{|D_n|}{|B_n|} \int \chi_{\{a_n\} \cap E}\, df_Q \\
&= \frac{|D_1|}{|B_1|} \int \chi_{\{a_1\}}\, df_Q + \cdots + \frac{|D_k|}{|B_k|} \int \chi_{\{a_k\}}\, df_Q \\
&\quad + \cdots + \frac{|D_{k+1}|}{|B_{k+1}|} \int \chi_{\varnothing}\, df_Q + \cdots + \frac{|D_n|}{|B_n|} \int \chi_{\varnothing}\, df_Q \\
&= \frac{|D_1|}{|B_1|} f_Q(\{a_1\}) + \cdots + \frac{|D_k|}{|B_k|} f_Q(\{a_k\}) + 0 + 0 \\
&= \frac{|D_1|}{|B_1|} \cdot \frac{1}{|D_1|} + \cdots + \frac{|D_k|}{|B_k|} \cdot \frac{1}{|D_k|} \\
&= \frac{1}{|B_1|} + \cdots + \frac{1}{|B_k|},
\end{aligned}
$$

which agrees with the definition of f_P. By uniqueness (up to a null set) of the Radon-Nikodym derivative, this shows that $df_P/df_Q = g$, which is given by Eq. (6.19). □

This theorem allows us to quantify the relationship between two different representation functions on the same space. In particular, it gives us a concrete formula, the Radon-Nikodym derivative, that relates these two measures.

6.5 Application of the mereological parthood model to gene ontology

There are several areas where we believe that our model could be implemented. The Gene Ontology (GO) Consortium is a ontological vocabulary bringing together genomic data of across a wide range of organisms [16]. The GO is composed of three subontologies/controlled vocabularies—"Cellular Component," "Molecular Function," and "Biological Process"—each annotating a gene, gene sequence, or gene product's

role. Such controlled vocabularies can be modeled by graphs, where the vertices are specific index terms or keywords in a subontology—for example, a particular cellular component or biological process (e.g., nucleus, phosphorylation)—and whose edges are relations between the terms, such as those denoting space (e.g., one type of cell being adjacent to another type of cell), time (e.g., an insect larva transforming into a pupa), participation (e.g., an ion channel conducting ions across a membrane) [66].

We propose that our model be used to enhance the interpretation of human language inputs in the GO by enabling ranking of different terms across different scales. Consider a set of biochemical pathways sharing a number of different molecules and similar reactions. By applying a specific representation function for an appropriate partition, one could conceivably assign weights to sets of genes or gene products within the GO, which form the vertices in a graph. If one is searching for a particular biochemical pathway involving known chemical reactants and products, one could conceivably reduce the space of possible pathways according to predetermined rankings set by a chosen partition. Perhaps there is a way of choosing appropriate partitions under certain chemical conditions. These are some goals towards which we propose our model be aimed towards.

6.6 Summary

In this chapter, we have provided a mereological model for parthood using partitions and have defined a class of so-called representation functions that assign a number to a set . We have presented four principal statements, three of which (Theorem 6.4.3, Proposition 2, and Theorem 6.4.8) are original:

- Theorem 6.4.3: This theorem says that any representation function is a finite measure on its domain.
- Proposition 2: This proposition tells us that for two partitions, one of which is finer than the other, the "size" of a set E with respect to the coarser partition is smaller than the "size" of E with respect to the finer partition.
- Theorem 6.4.8: For a given nonempty set, two partitions on this set, and two representation functions induced by these partitions, this theorem compares these representation functions and gives an explicit formula telling one how they are related to each other.

Finally, we have mentioned a potential application of our model in biomedical informatics.

Chapter 7

Summary & Conclusion

In this thesis, we have described three original research projects, one of which is published in the literature, whereas the other two are not published (excepting here in this thesis).

The published work is presented in Chapter 4 and is largely based on the following journal article:

- **Daniel A. Inafuku**, Kay L. Kirkpatrick, Onyema Osuagwu, Qier An, David A. Brewster, and Mayisha Zeb Nakib, "Channel capacity of the ribosome," *Physical Review E* 108(4), 044404 (2023) [39].

In Chapter 5, we conduct a similar analysis to the work found in Chapter 4, except that we examine a different system. In particular, Chapter 5 studies DNA and RNA polymerases, whereas Chapter 4 studies the ribosome.

The material outlined in Chapter 6 is a combination of both previous and original work. We expand upon the work presented in the article "Composition and Trans-Scalar Identity," by A. Schumm, W. Rohloff, and G. Piccinini [61] and make potential connections to the field of gene ontology.

7.1 Summary of Chapters 1-3

Chapter 1 motivates our topics of study in both biological information theory and mereology.

In Chapter 2, we introduce a number of information-theoretic definitions, including entropy, mutual information, discrete memoryless channel, and channel capacity. We use these definitions to describe the basic communication scheme—that is, the problem of a sender reliably transmitting information over a noisy medium to a receiver. This chapter culminates in the statement of a famous theorem in information theory:

Shannon's Noisy Channel Coding theorem. The theorem gives conditions that must be satisfied in order for a channel to transmit information accurately. More precisely, if the rate of the channel does not exceed its capacity, then symbols can be sent across the channel with an arbitrary degree of error. Conversely, if the rate exceeds the capacity, then

Chapter 3 introduces the biological systems that we study in Chapters 4 and 5. We first introduce the principal information-carrying biomolecules DNA, RNA, and protein. Then, we discuss the processes that transfer information between these molecules, namely, DNA replication, transcription, and translation.

7.2 Summary of Chapter 4

As mentioned above, the work in Chapter 4 is based on the journal article **Daniel A. Inafuku**, Kay L. Kirkpatrick, Onyema Osuagwu, Qier An, David A. Brewster, and Mayisha Zeb Nakib, "Channel capacity of the ribosome," *Physical Review E* 108(4), 044404 (2023) [39], which can be found here: https://link.aps.org/doi/10.1103/PhysRevE.108.044404.

In this paper, we accomplish the following.

- We introduced a model of the ribosome during translation as a discrete memoryless channel.
- We derived an upper bound and a lower bound for the channel's capacity (Theorem 4.3.1): $g(r) \leq \mathcal{C} \leq \log 21$, where $r \in [0, 1]$ is the error probability, and $g(r)$ is given by Eq. (4.3). For a typical error probability $r = 1 \times 10^{-4}$,
$$0.7027... \leq \mathcal{C} \leq 0.7321... \ \frac{\text{codons}}{\text{use}}.$$
- We numerically approximated the capacity using the Blahut-Arimoto (BA) algorithm. For a typical error probability $r = 1 \times 10^{-4}$, we obtain $\mathcal{C} \approx 0.7317$ codons/use.
- Converting the bounds to calculate the bounds of the capacity in time:
$$77.5964... \leq \mathcal{C}^*_{ribo} \leq 80.7655... \ \frac{\text{codons}}{\text{second}},$$
- We compared our analytically and numerically approximated capacity with observed translation rates

We find that the observed translation rates fall below the both our analytical and numerically approximated capacities. By Shannon's Noisy Channel Coding theorem, this finding shows that the ribosome operates safely

below its channel capacity, allowing it to translate accurately and quickly with an arbitrary degree of error.

7.3 Summary of Chapter 5

The work in this chapter is similar in content to Chapter 4. Whereas we study the ribosome in Chapter, we study DNA polymerase (DNAP) and RNA polymerase (RNAP). We have accomplished the following

- We introduced a model for RNAP as a discrete memoryless channel.
- We calculated an explicit formula for both DNAP and RNAP:

$$\mathcal{C}(r) = 2 + (1-r)\log(1-r) + r\log\frac{r}{3},$$

  where $r \in [0,1]$, recapitulating a result found in Ref. [24] by Djordjevic. For a typical error probability of $r = 1 \times 10^{-5}$, $\mathcal{C} \approx 0.99980$ bits/use.

- Using transcription rates obtained from Ref. [75], we find that RNAP's channel capacity in time is given by

$$110.36... \leq \mathcal{C}^*_{RNAP} \leq 999.90... \ \frac{\text{nucleotides}}{\text{second}}.$$

Typical transcription rates fall within the range , Thus, by Shannon's Noisy Channel Coding theorem, we find that RNAP, like the ribosome, safely operates below its channel capcacity, allowing it to transcribe accurately and quickly with an arbitrary degree of error.

7.4 Summary of Chapter 6

The work in this chapter builds upon the publication "Composition and Trans-Scalar Identity," by A. Schumm, W. Rohloff, and G. Piccinini [61]. We have accomplished the following.

- Built upon a mereological model of parthood for a finite space.
- Expanded upon a family of finite measures (called representation functions) indexed by the partitions of a finite space. (Theorem 6.4.3)
- Explicitly characterized the relationship between two representation functions. (Theorem 6.4.8)
- Suggested applications for our model in the context of biomedical ontologies.

7.5 Conclusion

In thesis, we have presented three major projects: two in biological information theory (Chapters 4 and 5) and one in mereology (Chapter 6).

The work in Chapter 4 is published in the literature and introduces an information-theoretic model of the ribosome that explains why ribosomes translate so quickly yet so accurately.

The work that follows in Chapter 5 is similar in spirit to Chapter 4 in that we introduce an information-theoretic model of biomolecules. Instead of the ribosome, however, we turn our attention to DNA polymerase and RNA polymerase. Our results suggest that, like the ribosome, RNAP can respectively transcribe information both quickly and accurately in line with information-theoretic means.

This thesis' original works conclude with Chapter 6, wherein we use measure theory to study a mereological model of parthood. This work is inspired by and builds upon work first introduced by Schumm et al. We present several original results on different ways of quantifying the sizes of sets related to each other through the subset parthood relation. Using our results, we propose applications in the classification of biomedical data in clinical settings.

Bibliography

[1] A.V. Aho, J.E. Hopcroft, and J.D. Ullman. A general theory of translation. *Mathematical Systems Theory*, 3:193–221, 1969.

[2] B. Alberts, R. Heald, A. Johnson, D. Morgan, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. W.W. Norton & Company, 2022.

[3] S. Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, 18(1):14–20, 1972.

[4] M. Bang, T. Centner, F. Fornoff, A.J. Geach, M. Gotthardt, M. McNabb, C.C. Witt, D. Labeit, C.C. Gregorio, H. Granzier, and S. Labeit. The complete gene sequence of titin, expression of an unusual ≈700-kDa titin isoform, and its interaction with obscurin identify a novel Z-line to I-band linking system. *Circulation Research*, 89(11):1065–1072, 2001.

[5] M. Berzell. *Electronic Healthcare Ontologies Philosophy, the real world and IT structures*. PhD thesis, Linköping University, 2010.

[6] R. Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, 18(4):460–473, 1972.

[7] A. Blank, J.A. Gallant, R.R. Burgess, and L.A. Loeb. An rna polymerase mutant with reduced accuracy of chain elongation. *Biochemistry*, 25(20):5920–5928, 1986.

[8] N. Block. Psychologism and behaviorism. *The Philosophical Review*, 90(1):5–43, 1981.

[9] H. Bremer and P.P. Dennis. Modulation of chemical composition and other parameters of the cell at different exponential growth rates. *EcoSal Plus*, 3(1), 2008.

[10] C.M. Brown. *Aquinas and the Ship of Theseus: Solving Puzzles about Material Objects*. Continuum, 2005.

[11] R. Cao. A teleosemantic approach to information in the brain. *Biology & Philosophy*, 27:49–71, 2011.

[12] M.C. Chibucos, A.E. Zweifel, J.C. Herrera, W. Meza, S. Eslamfam, P. Uetz, D.A. Siegele, J.C. Hu, and M.G. Giglio. An ontology for microbial phenotypes. *BMC Microbiology*, 14:294, 2014.

[13] J.J. Cimino. From data to knowledge through concept-oriented terminologies: Experience with the medical entities dictionary. *Journal of the American Medical Informatics Association*, 7(3):288–297, 2000.

[14] D.P. Clark, N.J. Pazdernik, and M.R. McGehee. *Molecular Biology*. Academic Press, 2019.

[15] M. Colombo and G. Piccinini. *The Computational Theory of Mind*. Cambridge University Press, 2023.

[16] The Gene Consortium. The gene ontology project in 2008. *Nucleic Acids Research*, 36:D440–D444, 2007.

[17] B. Corominas-Mutra, J. Fortuny, and R.V. Solé. Towards a mathematical theory of meaningful communication. *Scientific Reports*, 4:4587, 2014.

[18] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2006.

[19] A. Dashti, P. Schwander, R. Langlois, R. Fung, W. Li, A. Hosseinizadeh, H.Y. Liao, J. Pallesen, G. Sharma, V.A. Stupina, A.E. Simon, J.D. Dinman, J. Frank, and A. Ourmazd. Trajectories of the ribosome as a brownian nanomachine. *Proceedings of the National Academy of Sciences*, 111(49):17492–17497, 2014.

[20] K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, and M. Ashburner. Chebi: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 36:D344–D350, 2008.

[21] T.G. Dewey and H. Herzel. Applications of information theory to biology. *Pacific Symposium on Biocomputing*, 12:600–601, 2000.

[22] A. Dimitrov, J.P. Miller, T. Gedeon, Z. Aldworth, and A.E. Parker. Analysis of neural coding through quantization with an information-based distortion measure. *Network: Computation in Neural Systems*, 14(1):151–176, 2003.

[23] I.B. Djordjevic. Quantum biological channel modeling and capacity calculation. *Life*, 2(4):377–391, 2012.

[24] I.B. Djordjevic. *Quantum biological information theory*. Springer, 2016.

[25] D.A. Drummond and C.O. Wilke. The evolutionary consequences of erroneous protein synthesis. *Nature Reviews Genetics*, 10:715–724, 2009.

[26] J.O. Dubuis, G. Tkačik, E.F. Wieschaus, and W. Bialek. Positional information, in bits. *Proceedings of the National Academy of Sciences*, 110(41):16301–16308, 2013.

[27] P. Edelmann and J. Gallant. Mistranslation in e. coli. *Cell*, 10(1):131–137, 1977.

[28] D. Fange, H. Mellenius, P.P. Dennis, and M. Ehrenberg. Thermodynamic modeling of variations in the rate of rna chain elongation of e.coli rrn operons. *Biophysical Journal*, 106(1):55–64, 2014.

[29] A. Fluitt, E. Pienaar, and H. Viljoen. Ribosome kinetics and aa-trna competition determine rate and fidelity of peptide synthesis. *Computational Biology and Chemistry*, 31:335–346, 2007.

[30] G.B. Folland. *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons, 1999.

[31] R.G. Gallager. *Information Theory and Reliable Communication*. John Wiley & Sons, 1968.

[32] I. Goodfellow, J. Schlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[33] J.F. Gout, W. Li, C. Fritsch, A. Li, S. Haroon, L. Singh, D. Hua, H. Fazelinia, Z. Smith, S. Seeholzer, K. Thomas, M. Lynch, and M. Vermulst. The landscape of transcription errors in eukaryotic cells. *Science Advances*, 3(10):1701484, 2017.

[34] K.B. Gromadski and M.V. Rodnina. Kinetic determinants of high-fidelity trna discrimination on the ribosome. *Molecular Cell*, 13(2):191–200, 2004.

[35] M. Haendel, G. Gkoutos, S. Lewis, and C. Mungall. Uberon: towards a comprehensive multi-species anatomy ontology. In *Nature Precedings, International Conference on Biomedical Ontology*, 2009.

[36] R.V.L. Hartley. Transmission of information. *The Bell System Technical Journal*, 7(3):535–563, 1928.

[37] J. W. B. Hershey, N. Sonenberg, and M. B. Matthews. Principles of translational control. *Cold Spring Harb. Perspect*, 11(9), 2019.

[38] E.P. Hoel. When the map is better than the territory. *Entropy*, 19(5):188, 2017.

[39] D. A. Inafuku, K.L. Kirkpatrick, O. Osuagwu, Q. An, and M. Zeb Nakib. Channel capacity of the ribosome. *Physical Review E*, 108(4):044404, 2023.

[40] R. Ishimura, G. Nagy, I. Dotu, H. Zhou, X. Yang, P. Schimmel, S. Senju, Y. Nishimura, J.H. Chuang, and S.L. Ackerman. Ribosome stalling induced by mutation of a cns-specific trna causes neurodegeneration. *Science*, 345(6195):455–459, 2014.

[41] N.R. James, A. Brown, Y. Gordiyenko, and V. Ramakrishnan. Translational termination without a stop codon. *Science*, 354(6318):1437–1440, 2016.

[42] A. Juarrero. *Dynamics in Action: Intentional Behavior as a Complex System*. The MIT Press, 1999.

[43] E.F. Keller. Rethinking the meaning of biological information. *Biological Theory*, 4:155–166, 2009.

[44] E.B. Kramer and P.J. Farabaugh. The frequency of translational misreading errors in e. coli is largely determined by trna competition. *RNA*, 13:87–96, 2007.

[45] M. Krüger and W.A. Linke. The giant protein titin: A regulatory node that integrates myocyte signaling pathways. *Journal of Biological Chemistry*, 286(12):9905–9912, 2011.

[46] J. Kubelka, J. Hofrichter, and W.A. Eaton. The protein folding 'speed limit'. *Current Opinion in Structural Biology*, 14(1):76–88, 2011.

[47] P. Maiuri, A. Knezevich, A. De Marco, D. Mazza, A. Kula, J.G. McNally, and A. Marcello. Fast transcription rates of rna polymerase ii in human cells,. *European Molecular Biology Organization Reports*, 12(12):1280–1285, 2011.

[48] W.S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133, 1943.

[49] G.H. Mealy. Another look at data. In *Proceedings of the Fall Joint Computer Conference*, 1969.

[50] I. Müller. *A History of Thermodynamics: The Doctrine of Energy and Entropy*. Springer, 2007.

[51] H. Nyquist. Certain factors affecting telegraph speed. *The Bell System Technical Journal*, 3(2):324–336, 1924.

[52] J.V. Ogle and V. Ramakrishnan. Structural insights into translational fidelity. *Annual Review of Biochemistry*, 74(1):129–177, 2005.

[53] S. Osawa, T.H. Jukes, K. Watanabe, and A. Muto. Recent evidence for evolution of the genetic code. *Microbiological Reviews*, 56(1):229–264, 1992.

[54] R.D. Palmiter. Quantitation of parameters that determine the rate of ovalbumin synthesis. *Cell*, 4(3):189–197, 1975.

[55] J. Parker. Errors and alternatives in reading the universal genetic code. *Microbiological Reviews*, 53(3):273–298, 1989.

[56] M.D. Petkova, G. Tkačik, W. Bialek, E.F. Wieschaus, and T. Gregor. Optimal decoding of cellular identities in a genetic network. *Cell*, 176(4):844–855, 2019.

[57] A. Prabhakar, J. Choi, J. Wang, A. Petrov, and J.D. Puglisi. Dynamic basis of fidelity and speed in translation: Coordinated multistep mechanisms of elongation and termination. *Protein Science*, 26(7):1352–1362, 2017.

[58] A. Rector. Medical informatics. In F. Baader, D. Calvanese, D.L. McGuinness, D. Nardi, and P.F. Patel-Schneider, editors, *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, 2003.

[59] S. Roman. *Introduction to Coding and Information Theory*. Springer, 1996.

[60] Y. Savir and T. Tlusty. The ribosome as an optimal decoder: A lesson in molecular recognition. *Cell*, 153(2):471–479, 2013.

[61] A. Schumm, W. Rohloff, and G. Piccinini. Composition and trans-scalar identity. http://philsci-archive.pitt.edu/18253/.

[62] J.R. Searle. Can computers think? In D. Chalmers, editor, *Philosophy of Mind: Classical and Contemporary Reading*. Oxford University Press, 1983.

[63] C.E. Shannon. A mathematical theory of information. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[64] A. Shokrollahi. Capacity-approaching codes on the q-ary symmetric channel for large q. In *Proceedings of the Information Theory Workshop*, 2004.

[65] B. Smith. *Parts and Moments Studies in Logic and Formal Ontology*. Philosophia Verlag, 1982.

[66] B. Smith, W. Ceusters, B. Klagges, J. Köhler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A.L. Rector, and C. Rosse. Relations in biomedical ontologies. *Genome Biology*, 6(5):R46, 2005.

[67] T.R. Sokolowski. Information theory entering soils and tissues. *Cell Systems*, 13(7):511–513, 2022.

[68] N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. In *The 37th Annual Allerton Conference on Communication, Control, and Computing*, 1999.

[69] G. Tkačik, J.O. Dubuis, M.D. Petkova, and T. Gregor. Positional information, positional error, and readout precision in morphogenesis: A mathematical framework. *Genetics*, 199(1):39–59, 2015.

[70] C.C. Traverse and H. Ochman. Conserved rates and patterns of transcription errors across bacterial growth states and lifestyles. *Proceedings of the National Academy of Sciences*, 113(12):3311–3316, 2016.

[71] A.M. Turing. Intelligent machinery. In B.J. Copeland, editor, *The Essential Turing*. Oxford University Press, 1948.

[72] A.M. Turing. Computable machinery and intelligence. *Mind*, 236:433–460, 1950.

[73] S. Varenne, J. Buc, R. Lloubes, and C. Lazdunski. Translation is a non-uniform process. effect of trna availability on the rate of elongation of nascent polypeptide chains. *Journal of Molecular Biology*, 180(3):549–576, 1984.

[74] A.C. Varzi. Spatial reasoning and ontology: Parts, wholes, and locations. In M. Aiello, I. Pratt-Hartmann, and J. Van Benthem, editors, *Handbook of Spatial Logics*. Springer, Dordrecht, 2007.

[75] U. Vogel and K.F. Jensen. The rna chain elongation rate in escherichia coli depends on the growth rate. *Journal of Bacteriology*, 176(10):2807–2813, 1994.

[76] H.P. Yockey. An application of information theory to the central dogma and the sequence hypothesis. *Journal of Theoretical Biology*, 46(2):369–406, 1974.

[77] H.P. Yockey. *Information Theory, Evolution and the Origin of Life*. Cambridge University Press, 2005.

[78] F. Zeng, Y. Chen, J. Remis, M. Shekhar, J.C. Phillips, E. Tajkhorshid, and H. Jin. Structural basis of co-translational quality control by arfa and rf2 bound to ribosome. *Nature*, 541:554–557, 2017.

Appendix A

Derivation of Eq. (4.20)

We want to show that

$$\min_{x\in\mathcal{X}} T_n(x) - \log Q_n(x) \leq \mathcal{C} \leq \max_{x\in\mathcal{X}} T_n(x) - \log Q_n(x). \tag{A.1}$$

Define

$$I(x;Y) := \sum_{y\in\mathcal{Y}} p(y|x) \log \frac{p(y|x)}{\sum_{x'\in\mathcal{X}} Q(x')p(y|x')}, \quad \forall x \in \mathcal{X} \tag{A.2}$$

and

$$c_n(x) := \exp\left[\sum_{y\in\mathcal{Y}} p(y|x) \log \frac{p(y|x)}{\sum_{x'\in\mathcal{X}} p(y|x')Q(x')}\right], \quad \forall x \in \mathcal{X} \tag{A.3}$$

so that $I(x;Y) = \log c_n(x)$. It is shown in Ref. [31] that

$$\sum_{x\in\mathcal{X}} Q(x)I(x;Y) \leq \mathcal{C} \leq \max_{x\in\mathcal{X}} I(x;Y), \tag{A.4}$$

where Q is an arbitrary probability distribution over $\mathcal{X}$. Notice that

$$\begin{aligned}
e^{T_n(x)} &= \exp\left[\sum_{y\in\mathcal{Y}} p(y|x)\log\frac{Q_n(x)p(y|x)}{\sum_{x'\in\mathcal{X}} p(y|x')Q(x')}\right] \\
&= \exp\left[\sum_{y\in\mathcal{Y}} p(y|x)\left\{\log\frac{p(y|x)}{\sum_{x'\in\mathcal{X}} p(y|x')Q(x')} + \log Q_n(x)\right\}\right] \\
&= \exp\left[\sum_{y\in\mathcal{Y}} p(y|x)\log\frac{p(y|x)}{\sum_{x'\in\mathcal{X}} p(y|x')Q(x')} + \sum_{y\in\mathcal{Y}} p(y|x)\log Q_n(x)\right] \\
&= c_n(x)\exp\left[\log Q_n(x)\sum_{y\in\mathcal{Y}} p(y|x)\right] \\
&= Q_n(x)c_n(x), \qquad \text{(A.5)}
\end{aligned}$$

where the last equality follows from the fact that $p(y|x)$ is a conditional probability distribution. Thus, $e^{T_n(x)} = Q_n(x)c_n(x)$. So we have

$$I(x;Y) = \log c_n(x) = T_n(x) - \log Q_n(x). \qquad \text{(A.6)}$$

Taking the maximum over $\mathcal{X}$ of Eq. (A.6) and using Eq. (A.4), we obtain the upper bound of Eq. (A.1). For the lower bound of Eq. (A.1), we have

$$\begin{aligned}
\sum_{x\in\mathcal{X}} Q(x)I(x;Y) &= \sum_{x\in\mathcal{X}} Q(x)[T_n(x) - Q_n(x)] \\
&\geq \sum_{x\in\mathcal{X}} Q(x)\min_{x'\in\mathcal{X}}[T_n(x') - \log Q_n(x')] \\
&= \min_{x'\in\mathcal{X}}[T_n(x') - \log Q_n(x')]\sum_{x\in\mathcal{X}} Q(x) \\
&= \min_{x\in\mathcal{X}}[T_n(x) - \log Q_n(x)], \qquad \text{(A.7)}
\end{aligned}$$

where the last equality follows from the fact that Q is a probability distribution. By Eq. (A.4), the inequality follows.

Appendix B

Mereological Dictionary

In Chapter 6, we expanded upon a model of mereological parthood first introduced in Ref. [61]. The authors make some mereological definitions, some of which are conceptually similar to definitions often encountered in the mathematical literature. Thus, we have devised the following dictionary, which proposes several updated terminologies for different mathematical objects. These new terminologies also include recommended notation.

We hope that this dictionary will clarify any inconsistencies and identify possible equivalences of definitions in language among philosophers and mathematicians and, more generally, facilitate discussions between philosophers and mathematicians studying mereology.

Table B.1: Proposed dictionary

Piccinini *et al.*'s terminology and notation	Proposed terminology/interpretation and notation
Domain/universe U	Universe Ω
Model/structure of partitions $\mathfrak{S}$	World $\mathbf{G}$
Fusion operator $\circ$	Union operator $\cup$
(Proper) Parts	Blocks $B_1, B_2, \ldots$
Partitions $P_1, P_2, \ldots$	Partitions $P_1, P_2, \ldots$
"Collection of partitions" P	Collection of partitions $\hat{\mathcal{P}}$
Atoms $a_1, \ldots, a_n$	Atoms $a_1, \ldots, a_n$
—	Block representation
Composite object	Composite/Composite object
Maximally fine-grained partition	Finest partition P_f
Maximally coarse-grained partition	Coarsest partition P_c