# Analysis of Proteomics and Genomics Based on Signal Processing and Communication Theory

BY

NIDHAL BOUAYNAYA
B.S., Ecole Nationale Supérieure de l'Electronique et de ses Applications, France, 2002
M.S., Illinois Institute of Technology, Chicago, IL, 2002
M.S., University of Illinois at Chicago, Chicago, IL, 2007

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Chicago, 2007

Chicago, Illinois

# ACKNOWLEDGMENT

I would like to thank my advisor and mentor, Dr. Dan Schonfeld for his guidance, illuminating discussions related to this work and beyond, encouragement and moral and financial support in this research. I also would like to extend my gratitude to Dr. Rashid Ansari, Dr. Richard Magin, Dr. William O'Neill and Dr. Aggelos Katsaggelos for being part of my committee and for their insights and interest in my work. I would especially like to thank my referees: Dr. Dan Schonfeld, Dr. Rashid Ansari, Dr. Richard Magin, Dr. Ashfaq Khokhar and Dr. Nikolas P. Galatsanos for spending so much time and effort to write numerous reference letters for me.

I continue to be extremely grateful to my family who have been my greatest support. This accomplishment is not mine alone. Thank you for sharing my struggles and my victories.

Thank you to my friends and colleagues for sharing my pauses and supporting me during my ups and downs; thank you to Dr. Vivek Nigam, Mr. Murat Bagci, Mr. Pan Pan, Miss Julia Young and all graduate students of the multimedia communications lab from 2003 to 2007.

## TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF FIGURES (Continued)

## LIST OF FIGURES (Continued)

# SUMMARY

*"A brief summary will here be sufficient to recall to the reader's mind the more salient points in this work. Many of the views which have been advanced are highly speculative, and some no doubt will prove erroneous; but I have in every case given the reasons which have led me to one view rather than to another."*

Charles Darwin, *The Descent of Man.*

We developed a mathematical model of the genetic information storage and transmission system and investigate its properties. Breaking with tradition, whereas the genetic information storage and transmission apparatus is conventionally modelled as an engineering communication system with the DNA sequence as the input and the amino acid chain as the output, in this thesis the genetic communication model is viewed as one between proteins. A connection in series of protein communication systems is equivalent to a channel through time: *The Channel of Evolution*. We investigated the dynamics of the channel of evolution in both cases of a constant and time-variant point mutation rates. We proved that the distribution of amino acids converges geometrically to a specific distribution which matches nearly perfectly an estimate of the natural abundance of amino acids in Nature today.

By modelling evolution as the iteration of a protein communication system over time, we were able to study it from an information theoretic perspective. Investigation of the protein communication channel capacity shows that an organism cannot reliably (channel capacity almost zero) transmit its genetic information to its offsprings of many generations. Equivalently, organisms with high mutation rates are less efficient, from an information theoretic perspective, than organisms with low mutation rates. A comparison of the channel capacity with the rate-distortion of the three groups of life, Archaea, Bacteria and Eukaryotes, reveals that the biological fidelity does not reach the Shannon optimum distortion in the three domains. This result is somehow expected given that the level of distortion should account for the evolutionary processes. We relied on these results to provide an evolutionary model of the three groups of life based on mutations and unequal crossovers.

We then investigated the structure of the genetic codeword, the DNA. we proved that the introns play the role of a decoy for mutations. It is important to emphasize that the role of introns is not to ensure a perfect (errorless) communication system, but to temper the effervescence of the ever-changing genome under the chemical, physical and environmental conditions. Perfect information transmission will spell stagnation and ultimately extinction. This is the major difference between an engineering communication system and the biological communication system. We also maintain that introns increase the rate of evolutionary adaptation by providing hot spots for genetic recombination. The proposed dual role of introns serves to provide a balance between stability and adaptability. It is interesting to note that the role of introns in protection against mutations is enhanced by increasing the size of the intron regions. On the other hand, the function of introns in encouraging recombination depends on the presence of long contiguous nucleotide sequences in introns. In order to moderate the adaptability

## SUMMARY (Continued)

rate of the genomic sequence, the length of contiguous nucleotide sequences must be limited. Indeed, most eukaryotes display multiple intron regions within a single gene. Introns therefore seem to control the balance between stability and adaptability of the genomic sequence.

Finally, we introduced new non-stationary methods to study the correlation properties in nucleotide sequences, and defined a quantitative measure of the degree of randomness. We found that both coding and non-coding DNA sequences exhibit long-range correlations as captured by an evolutionary $1/f$ spectrum. So, DNA correlations are much more complex than power laws with a single scaling exponent. An alternative approach to non-stationary processes based on the Hilbert transform is proposed to provide a quantitative measure of how far the process deviates from randomness (white noise). We found that coding segments are "closer", on average, to random sequences than non-coding segments. This observation is most likely the source of confusion and controversy in previous work based on stationary analysis of DNA correlations. Finally, we showed that the evolutionary rate, which is the derivative of the average power law scaling exponent, can be used to observe and possibly predict the dynamics of change in a lineage.

# CHAPTER 1

# INTRODUCTION

*"It is the glory of God to conceal a matter; to search out a matter is the glory of kings."*

Proverb 25:2.

The simultaneous existence of creativity, stability and decadence is an astonishing property of the storage, processing and transfer of information in the genetic system. Biological information is encoded in nucleic acids, DNA or RNA molecules. By decoding this information into proteins, organisms come into being. Investigation of the genomic structure has focused primarily on physicochemical and related issues. Yet protein synthesis is at its core an information transmission phenomenon. It therefore seems reasonable to postulate that the evolutionary pressures shaping the DNA sequence might not have been confined to physicochemical issues alone, and that considerations relating to informatics might have had a constraining evolutionary role guided by limitations imposed by molecular physics and chemistry.

With the increased availability of genomic data, the research emphasis has shifted from sequence compilation techniques to genomic sequence and system analysis. Various investigators have developed models that attempt to capture different information related aspects of the genetic system. Of particular interest is the development of a mathematical model of a communication system that captures the genetic information storage and transmission apparatus, during asexual and sexual reproduction.

## 1.1 Problem Statement

Over the past half a century we have undergone a revolution in our ability to archive, process and exchange information. Communication of biological systems took a head start 3.5 billion years ago. How have living systems evolved to handle the same problems with which we are confronted in this so-called Information Age: problems of information storage and processing, problems of transmission and reliability? What is the nature of biological information and the ways in which it is processed and transmitted?

Communication systems are used to study both transmission of information between remote locations and data storage for future retrieval. Information theoretic principals have been used to develop effective algorithms to successfully transmit information from a source to a receiver in engineered systems. Living systems also successfully transmit their biological information through genetic processes such as replication, transcription, and translation, where the genome of an organism is the contents of the transmission. Species evolution can be understood and described as a communication process through time. The genetic information storage and transmission apparatus resembles communication engineering systems in many ways: The genomic information is encoded in the DNA. By decoding genes into proteins, organisms come into being. However, unlike communication engineer's systems, the biological communication system is not designed to minimize transmission errors. In the absence of errors, evolution will not be possible. Furthermore, a perfect (i.e., errorless) communication of the genetic information spells stagnation and ultimately extinction. So, intuitively, there has to be a balance between maintaining the organism identity by reliable transmission of its genetic information (stability) and allowing errors to occur purposefully to encourage evolution (adaptability). Then, what is the right mathematical model to capture the genetic information storage and transmission apparatus? Moreover, can we mathematically quantify Nature's design specifications which balance stability and adaptability?

Several researchers have explored the *central dogma* of genetics from an information transmission viewpoint (38), (40), (75), (99), (122). Gatlin (40), Yockey (122) and Roman-Roldan et al. (99) model the biological information transfer as a communication system, where the input is the DNA sequence and the output is the amino acid chain in the protein. That is the channel of the communication system is assumed to be the translation process. On the other hand, May (75) and Rosen et al. (38) consider the channel to be the replication and transcription processes during which errors are introduced into the nucleotide sequence. However, both models are inconsistent with engineering communication systems, which model transmission and storage of the same messages at the source and destination (excluding errors due to channel degradation). As Shannon clearly states in his seminal paper "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point" (110). Consequently, the reductionist approach to the central dogma as often misused by "DNA → RNA → protein" cannot be modelled by a communication engineering system but rather by some non bijective mathematical transformation. So, a communication system, which attempts to capture the transmission of genetic information from parent(s) to offspring, should have some level of abstraction in modelling the biological flow of information. The development of such a model is important for many research areas such as intron research, aging theories and evolutionary studies.

A biological communication system framework should have the five classical parts of a communication engineering system: source; source-channel encoder; channel; channel-source decoder and destination. It is agreed upon today that the DNA contains the encoded information specifying the biological development of all cellular forms of life. Hence, a communication system modelling the genetic transmission of information during asexual and sexual reproduction must model the DNA as the encoded information. The encoder is unknown biologically. Nevertheless, studying the structure of the DNA sequence might shed light into the evolutionary constraints which shaped it. In particular, the encoded information should reflect in its structure the biological system design specifications: stability and adaptability. An amazing feature of the DNA is its phenomenal redundancy. In many species, only a small fraction of the total sequence of the genome appears to encode protein. For example, only about 1.5% of the human genome consists of protein-coding sequences. Some non-coding DNA is involved in regulating the activity of coding regions. However, much of this DNA has no known function and is sometimes referred to as "junk DNA". The great deal of extra energy required to sustain, process and conserve non-coding DNA during many millions of years of evolution may imply an essential function. Otherwise, most likely it would have been eliminated by natural selection long ago. It appears difficult to prove this via molecular biology. A better strategy would be to seek an answer outside of the traditional domain. From a communication engineering point of view, the so-called "junk DNA" may turn out to be just as important as the much sought-after genes.

### 1.1.1 Research Objectives

The ultimate goal of this work is to model and understand the genetic information processing system. After more than half century's revolution in genetics it seems increasingly possible to translate Darwin observation-based theory into the language of mathematics, in particular the language of information theory and constraint optimization. The nontrivial task, however, is to translate the problem of a species evolution into a well-defined optimization problem with all determining fitness objective functions and constraints. A mathematical framework of transmission of genetic information during asexual and sexual reproduction can explain the current structure of the biological information processing apparatus such as the presence of junk DNA. The goal of this research is realized through the following objectives:

1. Abstract a cell as a set of proteins and the process of cell division as an information communication system between protein sets. This protein communication model can be extended to develop a communication system, which models the transmission of information in sexual reproduction.
2. A series connection of time-dependent protein communication channels is equivalent to a channel through time: the channel of evolution. Investigate the asymptotic behavior of the protein communication system. Specifically, is there an equilibrium state? If yes, what is the rate of convergence to this equilibrium? What are the biological implications of such equilibrium?

3. Study the protein communication system from an information theoretic perspective. Specifically, address the questions: (i) at what rate can the genomic information be transmitted? And (ii) what is the average distortion between the transmitted message and the received message at this rate? Comparison of the genomic transmission rate with the channel capacity will reveal whether the genomic code is efficient from an information theoretic perspective. However, even if the channel capacity is not exceeded, we are assured that biological communication systems do not rely on codes that produce negligible errors since the level of distortion present must account for evolutionary processes. It is, therefore, interesting to ask ourselves whether biological communication systems maintain an optimal balance between the transmission rate and the desired distortion levels needed to support adaptive evolution.

4. The protein communication channel relies on encoding of the information source using the DNA. Based on the highly redundant structure of the DNA sequence (e.g., presence of a large percentage of non-coding segments), argue that the encoder models a source and channel encoder. Use probability of error analysis and optimization to show that the role of non-coding sequences in DNA is to provide a balance between stability and adaptability in the genome.

5. Since coding and non-coding segments in the genome serve different purposes, investigate whether their functional differences is reflected upon a statistical one. Specifically, can we derive statistical features, which discriminate between the coding and non-coding

sequences in the genome? If so, can we use them to discover interesting evolutionary patterns?

### 1.1.2 Motivation

Informational analysis of genetic sequences has provided significant insight into parallels between the genetic process and information processing systems used in the field of communication engineering (7), (8), (11), (33), (40), (54), (56), (75), (79), (90), (105), (107), (108), (122). The development of a mathematical model to capture the genetic information storage and transmission apparatus, during asexual and sexual reproduction, has a two-fold consequence: First, it sheds light on the evolutionary processes that fashioned living organisms as agents of information processing and transmission. Second, it provides communication engineering models of biological systems that can be used in practical bioinformatics tasks (e.g. genomic and proteomic database search and retrieval). In a communication context, Darwin's evolutionary theory can be formulated in the language of constraint optimization. This point is almost intuitive and conceptually trivial: every species, extinct or present, can be considered as a constrained optimal solution that maximizes a certain fitness function given the environmental and ecological conditions of the time. The difficulty, however, stems from practicality given the enormous number of fitness functions and constraints of a given species and the missing links between biology and mathematics. Fortunately, at the genetic and molecular level, the same practical challenge may not be as acute. All species share a surprisingly few fundamental commonalities: 4-base replication, 20 amino acid groups and 3-codonization for amino acids. We will take advantage of these quasi-universal molecular rules to propose a protein communication channel which models the transmission of genetic information during asexual and sexual reproduction. Using a mathematical model of protein communication during cell replication, the problem of a species' evolution will be represented as the iteration of a communication channel over time: The channel of evolution. Moreover, the protein communication channel formalism unifies under the same mathematical framework the study of previously unrelated problems, e.g., channel capacity of molecular machines, role of junk DNA in the genome, the choice of the size of the genetic alphabet, etc.

The results derived from the proposed project have the potential to dramatically impact our current understanding of evolution and the role of genetic information encoded in the DNA. For example, from a communication engineering point of view, the so-called "junk DNA" may turn out to be just as important as the much sought-after genes. The proposed research will allow us to better understand the genetic mechanisms for transmission of information and to use new models in bioinformatics tasks.

## 1.2 Research Contributions

This work contributes to the field of computational bioinformatics and biology through the application of information theory and communication theory to the study and analysis of genetic sequences. Our work shifts the focus of the genomic signal processing community from analyzing the decoder of genomic systems to considering the mathematical model of the proper communication channel between proteins. This theoretical framework has two advantages: First, it suggests new formulations that depart from the traditional problem definition, e.g., study of evolution as a communication channel through time. Investigating this problem from a communication engineering perspective will shed light on many open issues in evolution. Second, it unifies under the same mathematical framework the study of previously unrelated problems, e.g., channel capacity of molecular machines, role of junk DNA in the genome, the choice of the size of the genetic alphabet, etc. Specific contributions of this work include:

- The development of a mathematical model of protein communication during cell replication.
- The prediction of the distribution of amino acids in nature today using Markov chains and non-negative matrix theory.
- Investigation of the channel capacity and rate distortion functions of biological communication systems for the three branches of life: Archaea, Bacteria and Eukaryotes.
- Proposition of a theory for the role of introns, which claims that introns maintain a genius balance between stability and adaptability in eukaryotic genomes. The predictions of the theory are confirmed using the length distribution of exons and introns in the genome of various eukaryotic organisms.
- The discovery of a new statistical structure within the DNA using non-stationary analysis of nucleotide sequences.

## 1.3 Organization

This thesis is organized as follows.

In Chapter 2, we provide a cursory overview of the basic mechanics of protein production and present the definitions of various terms used in this work. We also summarize the error repair mechanisms in the genetic system. Grasping the essence of the biological inspirations of this work is crucial to understanding the motivation, assumptions and theoretical results of this work.

In Chapter 3, we develop a mathematical model to capture the genetic information storage and transmission apparatus. We first motivate the importance of information processing in biology. Then, we present a survey on the communication systems studied in the literature to model the genetic information processing apparatus. We argue that these biological communication models are inconsistent with engineering communication systems and we present the protein communication system, where the transmitted messages are protein sequences and the encoded message is the DNA. We further study the analogies and differences between the engineering communication system for video transmission and the protein communication system during cell replication or asexual reproduction.

In Chapter 4, we mathematically model evolution as a series connection of protein communication channels through time: the channel of evolution. We study the evolutionary dynamics of this channel in both cases of constant and time-varying point mutation rate. We establish, using matrix analysis, that stochastic messages sent through the channel of evolution are received according to a fixed probability distribution, which is independent of the original message.

In Chapter 5, we study the structure of the encoded biological message in the DNA. We propose that introns provide a dual contradictory and complementary role in the genomic code: (1) Introns reduce the probability of error from mutation errors by serving as decoys which absorb isolated mutations. According to this view, introns protect coding regions in the

DNA sequence from frequent errors in the same way hollow uninhabited structures are used by the military to protect important installations, such as aircraft hangars and missile launching facilities, from a bomb attack by serving as a 'dummy' target that resembles the protected structure. (2) Introns increase the rate of evolutionary adaptation by providing a mechanism for unequal crossovers . This process results in the introduction of new exons into the genomic sequence and thereby is responsible for its rapid evolution.

Our approach to prove the first hypothesis relies on a probability of error analysis. Errors are mutations, which occur in the coding sequences of the gene, called exons. We derive the distribution of the optimal exon length, which minimizes the probability of error, and show that it accurately fits the biological exon length distribution of most eukaryotic genomes. Furthermore, to understand how can Nature generate the optimal distribution, we propose a diffusive random walk model for exon generation throughout evolution. This model results in an alpha stable exon length distribution, which is asymptotically equivalent to the optimal distribution. It is interesting to note that the role of introns in protection against mutations is enhanced by increasing the size of the intron regions. On the other hand, the role of introns in increasing the rate of recombination between genes must be tempered in order to prevent excessive evolutionary adaptability. Rapid changes in the genomic code must not occur too frequently, or else we would experience evolutionary jumps in each generation.

In Chapter 6, we investigate the information theoretic bounds of the channel of evolution. We compute the capacity and the rate-distortion functions of the protein communication system for the three domains of life: Achaea, Bacteria and Eukaryotes. We analyze the tradeoff between the transmission rate and the distortion in noisy protein communication channels. As expected, comparison of the optimal transmission rate with the channel capacity indicates that the biological fidelity does not reach the Shannon optimal distortion. However, the relationship between the channel capacity and rate distortion achieved for different biological domains provides tremendous insight into the dynamics of the evolutionary processes. We rely on these results to provide a model of protein sequence evolution based on the two major evolutionary processes: mutations and unequal crossovers.

In Chapter 7, we will bring to bear new tools to analyze non-stationary signals that have emerged in the statistical and signal processing community over the past few years. The emergence of these new methods will be used to shed new light and help resolve the issues of (i) the existence of long-range correlations in DNA sequences and (ii) whether they are present in both coding and non-coding segments or only in the latter. It turns out that the statistical differences between coding and non-coding segments are much more subtle than previously thought using stationary analysis. In particular, both coding and non-coding sequences exhibit long-range correlations, as asserted by a $\mathbf{1/f^{\beta(n)}}$ evolutionary (i.e., time-dependent) spectrum. However, we will use an index of randomness, which we derive from the Hilbert-Huang Transform, to demonstrate that coding sequences, although not random as previously suspected, are often "more random" (i.e., whiter) than non-coding sequences. Moreover, the study of the evolution of the rate of change of these time-dependent parameters in homologous gene families shows a sudden jump around the rat, which might be related to the well-known supercharged evolution of this rodent.

Finally, we provide a brief statement about future research work.

# CHAPTER 2

# MOLECULAR BIOLOGY: GENOMICS AND PROTEOMICS

> "*Who are we? Where do we come from? Why are we this way and not some other? What does it mean to be human? Are we capable, if need be, of fundamental change, or do the dead hands of forgotten ancestors impel us in some direction, indiscriminately for good or ill, and beyond our control? Can we alter our character? Can we improve our societies? Can we leave our children a world better than the one that was left to us? Can we free them from the demons that torment us and haunt our civilization? In the long run, are we wise enough to know what changes to make? Can we be trusted with our own future?*"
>
> Carl Sagan, *Shadows of Forgotten Ancestors.*

This section provides a cursory overview of the basic mechanics of protein production and supplies the definitions of various terms used in this thesis. The material presented here is available in standard texts on molecular genetics (51), (63), (65).

## 2.1 Nucleic Acid Structure

There are two types of nucleic acid that are of key importance in cells: deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). DNA is found as a double strand. The backbone of the molecule is composed of deoxyribose sugars linked by phosphate groups in a repeating polymer chain. Each sugar is linked to a molecule known as a base. In DNA, there are four types of base, called adenine, tymine, guanine and cytosine, usually referred to simply as A, T, G and C. The two distinct ends of a DNA sequence are known under the name of the 5' end and the 3' end. The fundamental building block of a nucleic acid is called a nucleotide: this is the unit of one base plus one sugar plus one phosphate. We usually think of the "length" of a nucleic acid sequence as the number of nucleotides in the chain. The two strands of the DNA molecule are held together by hydrogen bonding between A and T and between C and G. The two strands run in opposite directions and are exactly complementary in sequence, so that where one has A, the other has T and where one has C the other has G. Therefore, naming the bases on the conventionally chosen side of the strand is enough to describe the entire double-strand sequence. The two strands are coiled around one another in the famous double helical structure elucidated by Watson and Crick 50 years ago. This is shown schematically in Figure 1. In contrast, RNA molecules are usually single stranded, and can form a variety of structures by base pairing between short regions of complementary sequences within the same strand. In RNA, the base uracil (U) occurs instead of T. The structure of U is similar to that of T but lacks the $CH_3$ group linked to the ring of the T molecule. In addition, a variety of bases of slightly different structures, called modified bases, can also be found in some types of RNA molecule. The base-pairing rules in RNA are more flexible than DNA. The principle pairs are GC and AU (which is equivalent to AT in DNA), but GU pairs are also relatively frequent, and a variety of unusual, so called, "non-canonical", pairs are also found in some RNA structures (e.g., GA pairs). Figure 2 shows an illustration of RNA and DNA sequences with their respective nitrogenous bases.

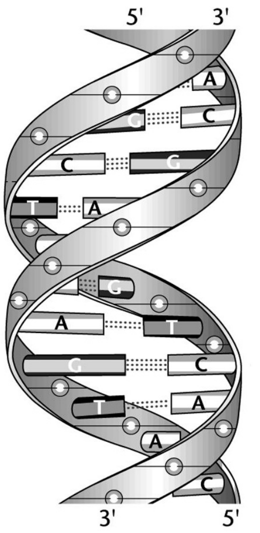


Figure 1. Schematic diagram of the DNA double helical structure.

### 2.1.1 DNA replication

DNA replicates every time a cell divides. In a multicellular organism, each cell contains a full copy of the genome of the organism (with the exception of certain cells without nuclei, such as red blood cells). The DNA is needed in every cell in order that protein synthesis can proceed in those cells. DNA replication is also essential for reproduction, because DNA contains the genetic information that ensures heredity.

DNA replication is semi-conservative. This means that the original double strand is replicated to give two double strands, each of which contains one of the original strands and one newly synthesized strand that is complementary to it. In the initiation step, several key factors are recruited to an origin of replication. This origin of replication is unwound with one

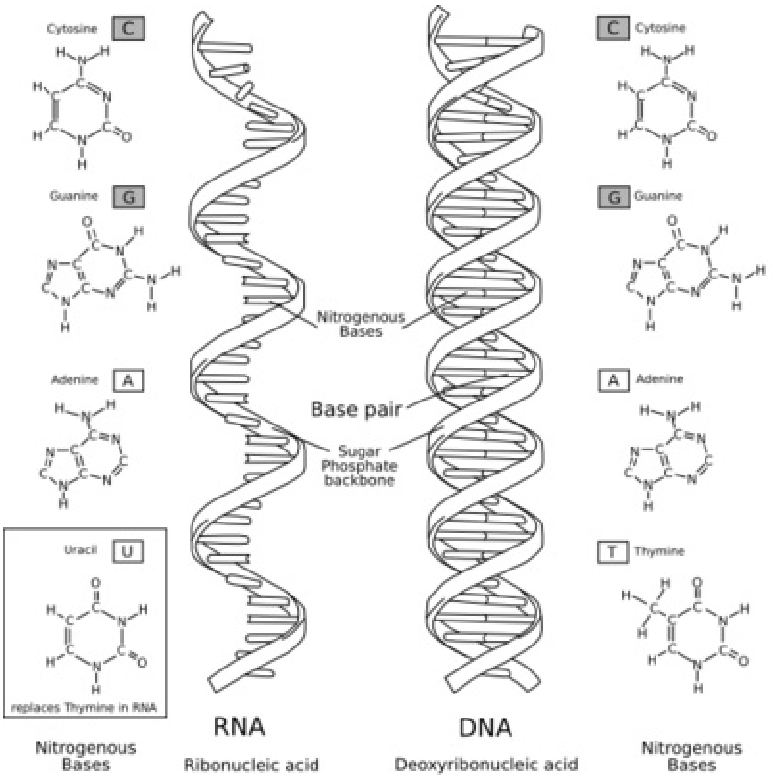


Figure 2. RNA with its nitrogenous bases to the left and DNA to the right.

replication fork on either end and the new strands are synthesized. The main enzyme that does this job is DNA polymerase III. The DNA polymerase III can only travel on one side of the original strand without any interruption. This original strand, which goes from 5’ to 3’, is called the leading strand. The complement of the leading strand, from 3’ to 5’, is the lagging strand where it is necessary to initiate synthesis independently many times. The new strand is therefore formed in pieces, which are known as Okazaki fragments.

DNA polymerase III is able to carry out the addition of new nucleotides but it cannot initiate a new strand. DNA polymerase therefore needs a short sequence called a primer, from which to begin. Primers are short sequences of RNA that are synthesized by an RNA called primase. The processus of DNA synthesis initiated by primers has been harnessed to become an important laboratory tool, the polymerase chain reaction.

Once the fragments on the lagging strand have been synthesized, it is necessary to connect them together. This is done by two enzymes. DNA polymerase I removes the RNA nucleotides of the primers and replicates them by DNA nucleotides. DNA ligase makes the final connection between the fragments. The process of DNA replication is schematically illustrated in Figure 3. Both DNA polymerase I and III have the ability to excise nucleotides if they do not match the template strand. This process of error correction is called proof-reading.

## 2.2 Protein Structure

Proteins are involved in practically every function performed by a cell, including regulation of cellular functions such as signal transduction and metabolism. Proteins control almost all the molecular processes of the body and are the actors that do everything that happens within us. The fundamental building block of proteins is the amino acid. There are 20 types of amino acid found in proteins. Proteins are linear polymers composed of chains of amino acids. Proteins, or “polypeptides”, are typically composed of several hundred amino acids. Each amino acid has a standard one-letter code (see Figure 6). A protein can be represented simply by a sequence of these letters, for example:

```
MAALDSLSF TSLGLSEQKA RETLKNSALS AQLREAATQA QQTLGSTIDK ATGILLYGLA
```

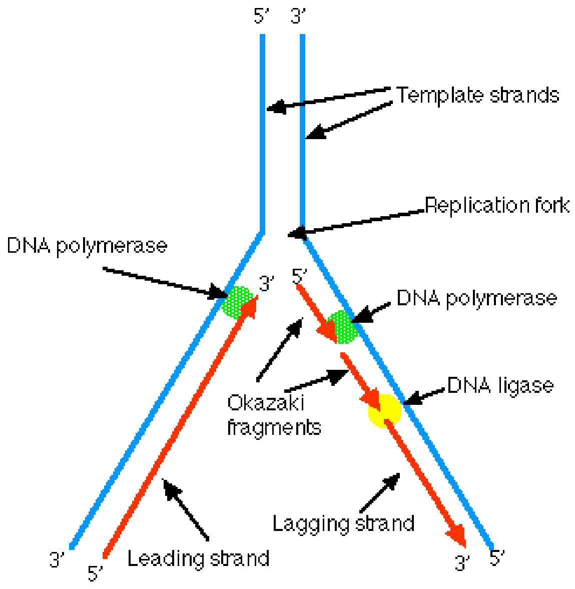


Figure 3. DNA Replication.

is the first part of a protein called glutaminyl-tRNA synthetase. Each protein has a structure that is specific to its sequence. The formation of this three-dimensional structure is called “protein folding”. The mechanism of protein folding is not entirely understood. The structure of a protein gives much more insight in the function of the protein than its sequence. Therefore, computational prediction of protein structure from its sequence is an active area of research. Many protein structures are known from X-ray crystallography. Around 90% of the protein structures available in the Protein Data Bank have been determined by X-ray crystallography. This method allows the exact 3D coordinates of all the atoms in the protein to be determined to within a certain resolution. Figure 4 illustrates the crystal structure of a biopolyester degrading

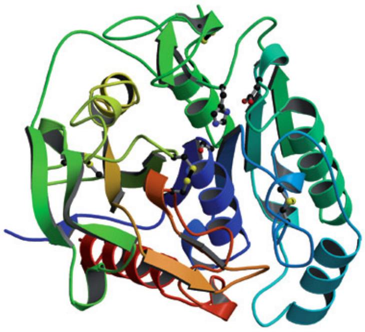


Figure 4. Crystal structure of a biopolyester-degrading enzym.

enzyme (78). Rather than drawing all the atomic positions, the figure is drawn at a coarse-grained level so that the most important features of the structure are visible.

There are four distinct aspects of a protein's structure:

- *Primary structure*: the amino acid sequence.
- *Secondary structure*: highly patterned sub-structures, e.g., alpha helix and beta sheet, or segments of chain that assume no stable shape. Secondary structures are locally defined so that there can be many different secondary motifs present in one single protein molecule.
- *Tertiary structure*: the overall shape of a single protein molecule including the spatial relationship of the secondary structural motifs to one another.
- *Quaternary structure*: the shape or structure that results from the union of more than one protein, usually called protein subunits, which function as part of the larger assembly or protein complex.

In addition to these levels of structure, proteins may shift between several similar structures in performing of their biological function. In the context of these functional rearrangements, these tertiary or quaternary structures are usually referred to as “conformations”, and transitions between them are called conformational changes.

## 2.3 Protein Synthesis

Protein synthesis is a multi-step process, beginning with transcription and ending with translation. Protein biosynthesis, although very similar, differs between prokaryotes and eukaryotes.

### 2.3.1 Transcription

Transcription is the process through which a DNA sequence is copied to produce a complementary RNA. Or, in other words, the transfer of genetic information from DNA into RNA. Typically sections of DNA a few thousand base pairs long are transcribed that correspond to single genes (or sometimes a small number of sequential genes). Transcription is carried out by an enzyme called RNA polymerase (RNAP). RNA polymerase recognizes and specifically binds to the promoter region on DNA. The DNA temporarily unwound and becomes single-stranded (”open”) in the vicinity of the initiation site. This strand is called the template strand. The polymerase catalyzes the assembly of individual ribonucleotides into an RNA strand that is complementary to the template DNA strand. When the template is a A, C, G or T, the base added to RNA is U, G, C or A. In contrast to the DNA polymerase III which can only add new nucleotides to a new strand but cannot initiate it, RNA polymerase is able to perform both initiation and addition. As the polymerase moves along, the RNA separates from the template and the two DNA strands close up again. Since the RNA is complementary to the template strand, it is actually the same as the non-template DNA strand, with the exception that Ts are converted to Us. The polymerase needs to know where to start and stop. This information is contained in the DNA sequence. A promoter is a short DNA sequence recognized as a strait signal by RNA polymerase. However, these sequences are not fixed and there is considerable variation between genes. Reliable location of promoter sequences is an active research topic. There two known termination mechanisms:

• *Intrinsic termination* (also called Rho-independent termination) involves terminator sequences within the RNA that signal the RNA polymerase to stop. The terminator sequence is usually a palindromic sequence that forms a stem-loop hairpin structure that leads to the dissociation of the RNAP from the DNA template. One such common termination motif is the palindromic sequence '`GCCGCCAG`'. The RNA polymerase fails to proceed beyond this point and consequently, the nascent DNA-RNA hybrid dissociates. The RNA polymerase then proceeds to look for a new initiation-region from which to start the initiation process again.

• *Rho-dependent termination* uses a termination factor called $\rho$ factor to stop RNA synthesis at specific sites. This protein binds and runs along the mRNA towards the RNAP. When $\rho$ factor reaches the RNAP, it causes RNAP to dissociate from the DNA, terminating transcription.

### 2.3.2 RNA Processing

An RNA strand that is transcribed from a protein-coding region of DNA is called a messenger RNA (mRNA). The mRNA is used as a template for protein synthesis in the translation process discussed below. In prokaryotes, mRNA consists of a central coding sequence that

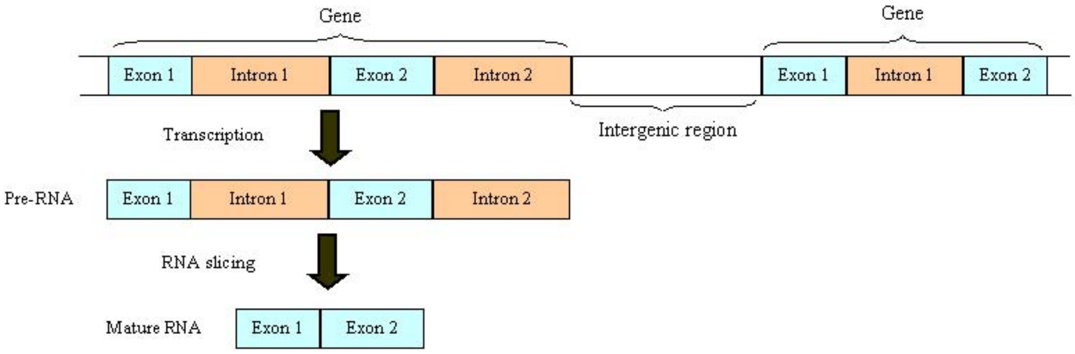


Figure 5. Structure and processing of eukaryotic mRNA.

contains the information for making the protein and short untranslated regions (UTRs). The UTRs are transcribed but not translated. In eukaryotes, the RNA transcript has a more complicated structure. When the RNA is newly synthesized, it is called a pre-mRNA. It must be processed in several ways before it comes a functional or “mature” mRNA. Eukaryotic gene sequences are composed of alternating sections called exons and introns. Exons are the pieces of the sequence that contain the information for protein coding. These pieces will be translated. Introns do not contain protein-coding information. The discovery of introns led to the Nobel Prize in Physiology or Medicine in 1993 for Phillip Allen Sharp and Richard J. Roberts. The introns are cut out of the pre-mRNA and are not present in the mRNA after processing. When an intron is removed the ends of the exons on either side of it are linked together to form a continuous strand. This is known as splicing. Figure 5 illustrates the mechanism of RNA splicing in eukaryotes.

Splicing is carried out by spliceosome, a complex of several types of RNA and proteins bound together and acting as a molecular machine. The spliceosome is able to recognize signals in the pre-mRNA sequence that tell it where the intron-exon boundaries are and hence which bits of the sequence to remove. As with promoter sequences for transcription, the signals for the splice sites are fairly short and variable, so that reliable identification of the intron-exon structure of a gene is a difficult problem in bioinformatics. Nevertheless, the spliceosome manages to do it. Introns that are spliced out by the spliceosome are called spliceosomal introns. This is the majority of introns in most organisms. In addition, there are some interesting, but fairly rare introns that are capable of catalyzing their own splicing out of the primary RNA transcript without the action of the spliceosome. There are surprisingly large numbers of introns in many eukaryotic genes: 10 or 20 in one gene is not uncommon. In contrast, most prokaryotic genes do not contain introns. It is still rather controversial where and when introns appeared, and what is the use, if any, of having them.

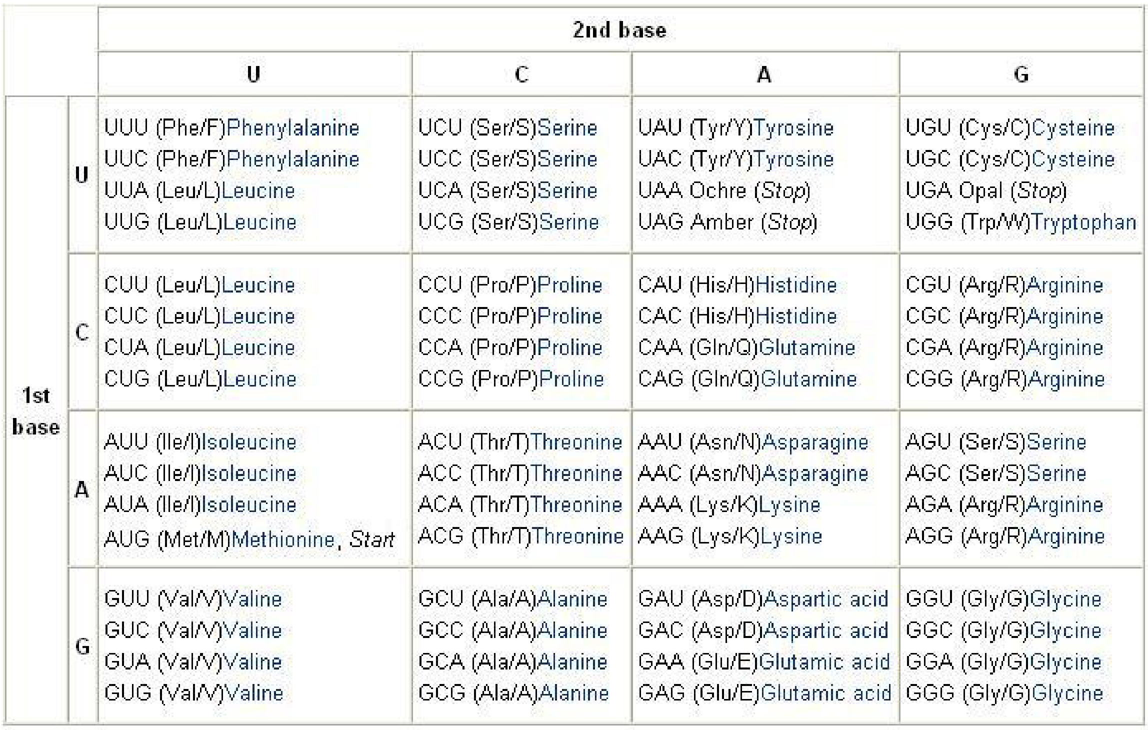
In eukaryotes, the DNA is contained in the nucleus, and transcription and RNA processing occur in the nucleus. The mRNA is then transported out of the nucleus through the pores in the nuclear membrane, and translation occurs in the cytoplasm.

### 2.3.3 The Genetic Code

We now need to consider how information in the form of sequences of four types of base is turned into information in the form of sequences of 20 types of amino acid. The mRNA is read in groups of three bases called codons. There are $4^3 = 64$ codons that can be made with four bases. Each of these codons codes for one type of amino acid, and since 64 is greater than

20, most amino acids have more than one codon that codes for them. The set of assignment of codons to amino acids is known as the genetic code, and is given in Figure 6. There are three codons that act as stop signals rather than coding for amino acids. These denote the end of the coding region of a gene. Because many codons are redundant, i.e., two or more codons can code for the same amino acid, the genetic code is degenerate. The degeneracy of the genetic code makes it more fault-tolerant for point mutations. A practical consequence of the degeneracy is that some errors in the genetic code only cause a silent mutation or an error that would not affect the amino acid's hydrophilic/hydrophobic property.

When Marshall W. Nirenberg and Heinrich J. Matthaei at the National Institutes of Health performed the experiments that first elucidated the genetic code in the 60's, it was thought to be universal, i.e., identical in all species. Now we realize that it is extremely widespread but not completely universal. The standard code shown in Figure 6 applies to almost all prokaryotic genomes (including both bacteria and archaea) and to the nuclear genomes of almost all eukaryotes. In mitochondrial genomes, there are several different genetic codes, all differing from the standard code in small respects (e.g., the assignment of the stop codon UGA to Trp, or the assignment of the Ile codon AUA to Met). There are also some changes in the nuclear genome codes for specific group of organisms, such as the ciliates (a group of unicellular eukaryotes). These changes are all quite small, and presumably they occurred at a relatively late stage in evolution. The main message is that the code is shared between all three domains of life (archaea, bacteria and eukaryotes) and hence must have evolved before the divergence of



| | | 2nd base | | | |
|---|---|---|---|---|---|
| | | U | C | A | G |
| 1st base | U | UUU (Phe/F)Phenylalanine<br>UUC (Phe/F)Phenylalanine<br>UUA (Leu/L)Leucine<br>UUG (Leu/L)Leucine | UCU (Ser/S)Serine<br>UCC (Ser/S)Serine<br>UCA (Ser/S)Serine<br>UCG (Ser/S)Serine | UAU (Tyr/Y)Tyrosine<br>UAC (Tyr/Y)Tyrosine<br>UAA Ochre (*Stop*)<br>UAG Amber (*Stop*) | UGU (Cys/C)Cysteine<br>UGC (Cys/C)Cysteine<br>UGA Opal (*Stop*)<br>UGG (Trp/W)Tryptophan |
| | C | CUU (Leu/L)Leucine<br>CUC (Leu/L)Leucine<br>CUA (Leu/L)Leucine<br>CUG (Leu/L)Leucine | CCU (Pro/P)Proline<br>CCC (Pro/P)Proline<br>CCA (Pro/P)Proline<br>CCG (Pro/P)Proline | CAU (His/H)Histidine<br>CAC (His/H)Histidine<br>CAA (Gln/Q)Glutamine<br>CAG (Gln/Q)Glutamine | CGU (Arg/R)Arginine<br>CGC (Arg/R)Arginine<br>CGA (Arg/R)Arginine<br>CGG (Arg/R)Arginine |
| | A | AUU (Ile/I)Isoleucine<br>AUC (Ile/I)Isoleucine<br>AUA (Ile/I)Isoleucine<br>AUG (Met/M)Methionine, *Start* | ACU (Thr/T)Threonine<br>ACC (Thr/T)Threonine<br>ACA (Thr/T)Threonine<br>ACG (Thr/T)Threonine | AAU (Asn/N)Asparagine<br>AAC (Asn/N)Asparagine<br>AAA (Lys/K)Lysine<br>AAG (Lys/K)Lysine | AGU (Ser/S)Serine<br>AGC (Ser/S)Serine<br>AGA (Arg/R)Arginine<br>AGG (Arg/R)Arginine |
| | G | GUU (Val/V)Valine<br>GUC (Val/V)Valine<br>GUA (Val/V)Valine<br>GUG (Val/V)Valine | GCU (Ala/A)Alanine<br>GCC (Ala/A)Alanine<br>GCA (Ala/A)Alanine<br>GCG (Ala/A)Alanine | GAU (Asp/D)Aspartic acid<br>GAC (Asp/D)Aspartic acid<br>GAA (Glu/E)Glutamic acid<br>GAG (Glu/E)Glutamic acid | GGU (Gly/G)Glycine<br>GGC (Gly/G)Glycine<br>GGA (Gly/G)Glycine<br>GGG (Gly/G)Glycine |

Figure 6. The standard genetic code.

these groups. Thus the last universal common ancestor of all current life must have used this genetic code.

### 2.3.4 Translation

Translation is the process of synthesis of a protein sequence using mRNA as a template. A key molecule in the process is transfer RNA (tRNA). The structure of tRNA is shown in Figure 7. The three bases in the central hairpin loop in the cloverleaf are called the anticodon. The sequence shown in Figure 7 is a tRNA-Ala, i.e., a tRNA for the amino acid alanine. Reading from 5' to 3' in the mRNA, the anticodon forms complementary base pairs with the codon sequence GCA, which codes for alanine in the genetic code. The anticodon end connects to the mRNA, and the other end connects to the growing protein chain. Organisms possess sets of tRNAs capable of base pairing with all 61 codons that denote amino acids. These tRNAs differ from one another not only in the anticodon identification but also in many other parts of their sequence, but they all have the same cloverleaf structure.

Protein synthesis is also carried out by another molecular machine called a ribosome. The ribosome plays an essential role in the catalysis of the process of peptide bond formation. “Ribosyme” is the term used for a catalytic RNA molecule, by analogy with “enzyme”, which is a catalytic protein. The ribosome is composed of a large and small subunit (represented by the two large ellipses in the cartoon in Figure 8(b). The small subunit contains the small subunit ribosomal RNA (SSU rRNA), together with ribosomal proteins. The large subunit contains large subunit ribosomal RNA (LSU rRNA), together with proteins and another smaller ribosomal RNA known as 5S rRNA. The ribosome binds to the mRNA and moves along it one codon at a time. tRNAs, charged with their appropriate amino acid, bind to the mRNA at a site inside the ribosome. The amino acid is them removed from the tRNA and attached to the end of a growing protein chain. The old tRNA then leaves and can be recharged with another molecule of the same type of amino acid and used again. The tRNA corresponding to the next codon then binds to the mRNA and the ribosome moves along one codon. This translocation continues on, and a long chain of amino acid (protein), is formed. When the ribosome reaches a stop codon, a protein known as a release factor enters the appropriate site in the ribosome instead of a tRNA. The release factor triggers the release of the completed protein from the

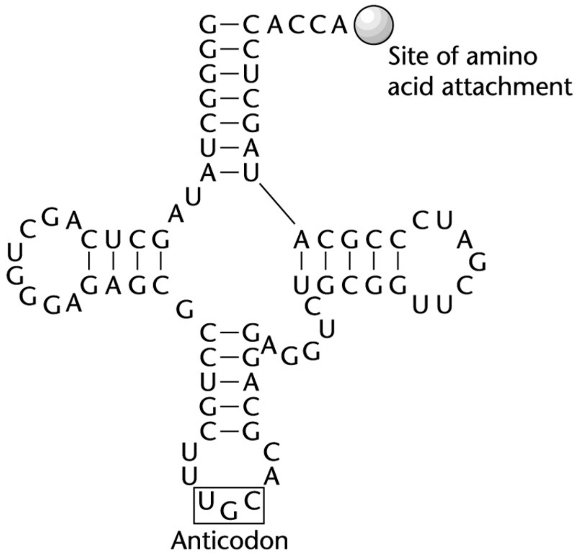


Figure 7. Structure of transfer RNA (tRNA).

ribosome. This ends the translation process. The protein synthesis mechanism (transcription and translation) is illustrated in Figure 8.

There is also a specific start codon, AUG, which codes for methionine. The ribosome begins protein synthesis at the first AUG codon it finds, which will be slightly downstream of the place where it initially binds to the mRNA. Other AUG codons occurring in the middle of a gene sequence lead to the usual form of a Met being added to the protein sequence. Other alternative start codons are common in prokaryotes. In bacteria, mRNAs contain a conserved sequence of about eight nucleotides, called the Shine-Dalgarno sequence, close to their 5' end. This sequence is complementary to part of SSU rRNA in the small subunit of the ribosome. This interaction triggers the binding of the ribosome to the mRNA.

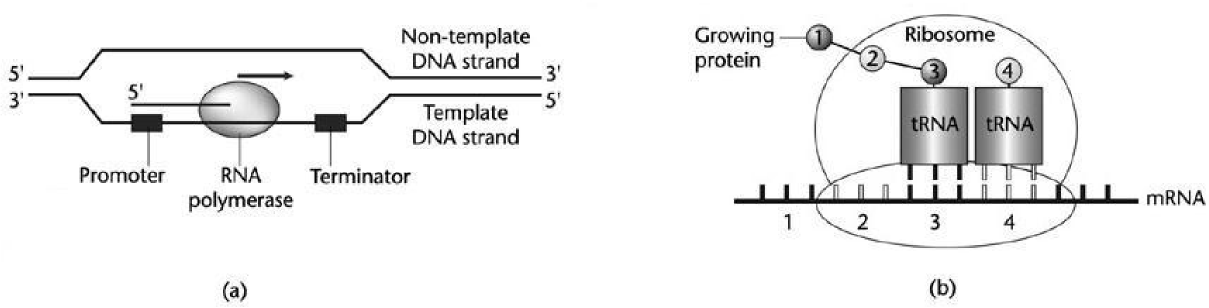


Figure 8. Protein synthesis: (a) Transcription; (b) Translation.

## 2.4 Error Repair Mechanisms in the Genetic System

### 2.4.1 DNA repair

DNA repair mechanisms are constantly operating in cells (18), (36). In human cells, both normal metabolic activities and environmental factors can result in as many as 1 million individual molecular lesions per cell each day (70), (115). Many of these lesions cause structural damage to the DNA molecule and can alter or eliminate the cell's ability to transcribe the gene that the affected DNA encodes. Other lesions induce potentially harmful mutations in the cell's genome, which will affect the survival of its daughter cells. The rate of DNA repair is dependent on many factors, including the cell type, the age of the cell, and the extracellular environment. A cell that has accumulated a large amount of DNA damage, or one that no longer effectively repairs damage incurred by its DNA, can enter one of three possible states:

1. an irreversible state of dormancy, known as senescence.

2. cell suicide, also known as apoptosis or programmed cell death.

3. unregulated cell division, which can lead to the formation of a tumor that is cancerous.

The DNA repair ability of a cell is vital to the integrity of its genome and thus to its normal functioning and that of the organism. Many genes that were initially shown to influence lifespan have turned out to be involved in DNA damage repair and protection.(20). Failure to correct molecular lesions in cells that form gametes can introduce mutations into the genomes of the offspring and thus influence the rate of evolution.

DNA damage can be subdivided into two main types:

1. endogenous damage such as attack by reactive oxygen species produced from normal metabolic byproducts (spontaneous mutation).

2. exogenous damage caused by external agents such as:

- ultraviolet [UV 200-300nm] radiation from the sun;
- other radiation frequencies, including x-rays and gamma rays;
- hydrolysis or thermal disruption;
- certain plant toxins human-made mutagenic chemicals, especially aromatic compounds that act as DNA intercalating agents;
- cancer chemotherapy and radiotherapy.

Mutations, or heritable alterations in the genetic material, can occur either as single nucleotide substitutions or as a frameshift mutation, i.e., insertion or deletion of one or more nucleotides.

Depending on the type of damage inflicted on the DNA's double helical structure, a variety of repair strategies have evolved to restore lost information.

### Direct reversal

Some types of DNA damage are so common that they have their own cellular subsystem dedicated to counteracting them. These mechanisms do not require a template. Such direct reversal mechanisms are specific to the type of damage incurred. Specific enzymes restore normal structure without breaking backbone.

### Single strand damage

When only one of the two strands of a double helix has a defect, the other strand can be used as a template to guide the correction of the damaged strand. In order to repair damage to one of the two paired molecules of DNA, there exist number of excision repair mechanisms that remove the damaged nucleotide and replace it with an undamaged nucleotide complementary to that found in the undamaged DNA strand:

- Base excision repair, which repairs damage due to a single nucleotide.
- Nucleotide excision repair, which repairs damage affecting longer strands of 2-30 bases
- Mismatch repair, which corrects errors of DNA replication and recombination that result in mispaired nucleotides following DNA replication

### Double strand damage

A type of DNA damage particularly hazardous is a break to both strands in the double-helix. Two mechanisms exist to repair this damage: non-homologous end-joining (NHEJ) and recombinational repair (also known as template-assisted repair or homologous recombination repair) (120).

The NHEJ pathway operates when the cell has not yet replicated the region of DNA on which the lesion has occurred. The process directly joins the two ends of the broken DNA strands without a template, losing sequence information in the process. Thus this repair mechanism is necessarily mutagenic. However, if the cell is not dividing and has not replicated its DNA, the NHEJ pathway is the cell's only option

Recombinational repair requires the presence of an identical or nearly identical sequence to be used as a template for repair of the break. For instance, this repair mechanism allows a damaged chromosome to be repaired using its sister chromatid as a template. The products of the human breast cancer susceptibility genes BRCA1 and BRCA2 may be involved in recombinational repair.

### Translesion synthesis

Not all DNA damage is or can be removed immediately. In some circumstances, the cell adopts damage tolerance, which is not truly repair but a way of coping with damage so that life can go on. Translesion synthesis is an error-prone (almost error-guaranteeing) last-resort method of repairing a DNA lesion that has not been repaired by any other mechanism. The DNA replication machinery cannot continue replicating past a site of DNA damage, so the advancing replication fork will stall on encountering a damaged base. DNA polymerases inserts extra bases at the site of damage and thus allow replication to bypass the damaged base to continue with replication. From the cell's perspective, it is "better" to introduce mutations around a single site than to continue the cell cycle with an incompletely replicated genome.

Despite their specialization and accuracy, these DNA repair mechanisms are not 100% efficient and many errors remain undetected or uncorrected in the genome. The error rate after the correction mechanisms is estimated at $10^{-9}$. When DNA damage is not repaired properly, or is repaired by an error-prone mechanism, mutations are introduced into the genomes of the cell's progeny. When this occurs in a germ line cell that will eventually produce a gamete, the mutation is passed on to the affected organism's offspring. The rate of evolution in a particular species (or, more narrowly, in a particular gene) is a function of the rate of mutation and thus of the accuracy and the rate of the DNA repair pathway and factors that can influence it (73).

### 2.4.2 RNA editing

The term RNA editing describes those molecular processes in which the information content is altered in an RNA molecule. To date such changes have been observed in tRNA, rRNA and mRNA molecules of eukaryotes, but not prokaryotes (19). The demonstration of RNA editing in prokaryotes may only be a matter of time, considering the range of species in which the various RNA editing processes have been found. The diversity of RNA editing mechanisms includes nucleoside modifications such as C to U as well as nucleotide additions and insertions. RNA editing in mRNAs effectively alters the amino acid sequence of the encoded protein so that it differs from that predicted by the genomic DNA sequence. Often the genomic information encoding an open reading frame or a tRNA is cryptic or incomplete and will not yield a functional product. Thus the genetic system involved is dependent on RNA editing for its biological optimization and eventual survival. In addition, RNA editing ensures that tRNAs can fold correctly.

From first impressions RNA editing is rather extravagant and costly, since without exception all of the RNA editing events described would be rendered unnecessary if the sequence of the mature mRNA was encoded in the genome. Advantages of RNA editing include the potential of synthesizing two or more distinct products from a single gene as exemplified by the viral editing systems or in the tissue-specific apoB editing, where two lipid carrier proteins with different properties are derived from one and the same genomic coding region (45).

### 2.4.3 Errors in the protein synthesis process

Translation is a complex process, entailing many steps and dozens of molecules. The potential for error exists at each step. The complexity of translation creates a conflict between two requirements: the process must be not only accurate, but also fast enough to meet a cell's needs. The accuracy of translation in protein synthesis is measured as the rate of misincorporation of a particular amino acid, different from that specified by an mRNA codon, into protein. The error frequencies depend on the amino acid being substituted and varies from $10^{-3}$ to $10^{-6}$. The observed values are close to $10^{-4}$. An error frequency of about $10^{-4}$ per amino acid residue was selected in the course of evolution to accurately produce proteins consisting of as many as 1000 amino acids while maintaining a remarkably rapid rate for protein synthesis (13).

# CHAPTER 3

# PROTEIN COMMUNICATION CHANNEL

*"It is the simple hypotheses of which one must be most wary; because these are the ones that have the most chances of passing unnoticed."*

Jules Henri Poincaré.

**Abstract**

*In this Chapter, we propose a protein communication system where the transmitted messages are protein sequences and the encoded message is the DNA. A series connection of the protein communication channel is equivalent to a channel through time: the channel of evolution.*

## 3.1 Introduction

The study of the information processing capabilities of living systems began in the 1970's (35), (40), (102) and was revived in the later part of the 1980's, due to the increase in genomic data which spurred a renewed interest in the use of information theory in the study of genomics. Information measures, such as entropy, have been used in recognition of DNA patterns, classification of genetic sequences, and other computational studies of genetic processes (2), (3), (8), (10), (11), (38), (75), (76), (83), (85), (89), (99), (101), (103), (104), (106), (107), (108), (113). The development of a mathematical model to capture the genetic information storage and transmission apparatus is at the heart of many information-theoretic problems in genomics and proteomics. For instance the structure of the DNA (e.g., presence of non-coding sequences), the size of the genetic alphabet, the codon length for amino acid coding, the distribution of amino acids in nature and many more, can be analyzed using a communication engineering framework modelling the transmission of genetic information during asexual and sexual replication.

### 3.1.1 Information in Biology

One cannot discuss biology without ever mentioning “information” or using any of its scientific concepts. Without it important insights would not be gained and, even more relevant, basic questions would never have been asked. For example, after the molecular structure of nucleic acids and proteins had been discovered by the pioneers of structural biology it was straightforward to associate DNA with a string of symbols encoding a message. The analogy to message processing in information technology immediately invited the idea of a code relating DNA and protein. Asking the right question initiated a true rush in research that ended with the successful deciphering of the genetic code (58).

It is important to distinguish between a statistical notion of information linked to uncertainty and a semantic notion of information referring to the content of a message and the consequences it elicits.“Meaning” and “purpose” are difficult notions in evolutionary biology since they can only be defined and discussed a posteriori. We shall find, however, that meaning is not entirely unrelated to information, uncertainly and specificity. The study of information theory begun with the revolutionary work of Claude Shannon in the early years of World War II (110). One of Shannons great contributions to the field of information theory is the separation of the semantic content of a message from the dynamic channel that transmits the message:

"The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the message have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this unknown at the time of design". One need only change the one word *engineering* to *biological* to make this paragraph apply to the application of communication theory in molecular biology. A particular information source may be, for example in the context of the Internet, an audio file, a video clip, or an email. Similarly, in the context of the biological communication channel, an information source can be the protein set of a bacterium, a person, or any organism.

Historically, the application of information theory to genetic analysis began in the 70's (35), (40), (99), (102). Between 1970 and 1977, in an attempt to quantify and convey the complexity of DNA, methods were developed for estimating information, redundancy or divergence parameters for DNA sequences (40). These efforts did not prove completely successful. After a ten year hiatus, the increase in genomic data encouraged renewed interest in the use of information theory in the study of genomics. This second research period began in 1987 and continues to the present. In this present period, techniques from the field of signal processing (such as auto correlation analysis, Fourier transform, and random walks) have been used in the informational analysis of genetic sequences (99). Discovering the existence of long-range correlations in DNA sequences proved to be a significant result of information-based analysis of genetic sequences (90). Mutual information, an information theory measure, has also been used to detect long-range correlations in nucleotide sequences (99). Other information measures, such as entropy based measures, have been used in recognition of DNA patterns, classification of genetic sequences, and various other computational studies of genetic sequences (2), (4), (5), (8), (10), (11), (38), (76), (83), (89), (99), (101), (103), (104), (105), (106), (108), (114), (123).

### 3.1.2 The Biological Communication Channel: Literature Review

The design of a communication channel is not for a particular message or type of messages; rather, the transmission machinery is designed for all possible messages, regardless of their semantic meanings. This explains, from an information theoretic point of view, why the biological information storage and transmission system is common (with rare variations, which happened later in evolution) to the three domains of life (archae, bacteria and eukaryotes). A communication process is characterized by three main sub-processes: the coding of the information to be communicated, the transmission of the information along the communication channel, and the decoding of the information at the receiver. Usually, it is in the communication channel that unwanted errors are introduced.

**Gatlin's communication channel (40)**

Gatlin first extended Shannon and Weaver (111) theory of information, developed for the transmission of electronic signals, to biology. Gatlin argued that the genetic system is the source, reproduction and ontogeny are the channel, and the environment is the receiver. Genetic possibilities thus become phenotypic signals as a result of reproduction and ontogeny

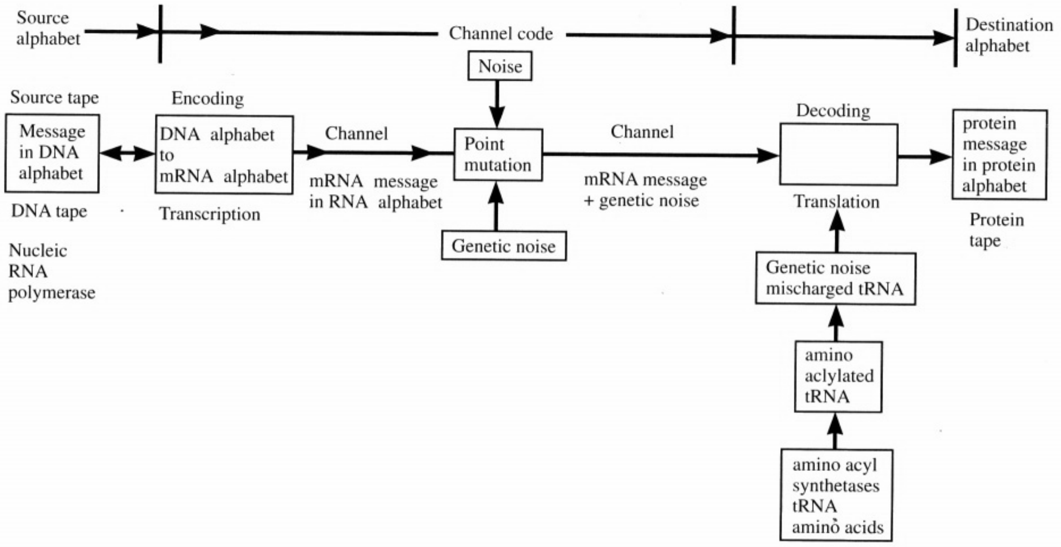


Figure 5.2. The transmission of genetic messages from the DNA tape to the protein tape as conceived in molecular biology. Genetic noise occurs in all stages but is lumped in the figure to fix the idea.

Figure 9. Hubert Yockey's DNA-protein communication channel model. (122)

and become meaningful biological information as a result of environmental selection on the phenotypes.

**Yockey's communication channel (122)**

Figure 9 shows the DNA-protein communication system proposed by Yockey (122). In Yockey's model, the genome or ensemble of genetic messages is generated by a stationary Markov process and recorded in the DNA sequence, which is isomorphic with the tape in a Turing machine (116). The DNA source is then encoded in the mRNA sequence. mRNA plays the role of the channel that communicates the genetic message to the ribosomes, which serve as the decoder. A second stochastic Markov process operates on the message and introduces point mutations. The genetic message is decoded by the ribosomes from the 64-letter mRNA alphabet to the 20-letter protein alphabet. The decoding is accomplished by the genetic code.

**Roman-Roldan's communication channel (99)**

Roman-Raldan's models the transfer of biological information as a communication channel with the DNA sequence as the input and the amino acid sequence which forms the protein as the output (99). The communication channel view suggested by Roman-Roldan et al. is depicted in Figure 10. In this model, the information source, the DNA, is viewed as a stationary and ergodic Markov source. The transmission channel is the genetic code and is assumed to be stationary and memoryless. The correspondence between the codon set $\mathbf{B^3} = \{B_1, B_2, B_3\}$ and the amino acid set $\mathbf{A} = \{A_i\}$ is given by $\mathbf{B^3} \xrightarrow{p(A_i|B_1,B_2,B_3)} \mathbf{A}$. If the genetic channel is noiseless, or free of genetic mutations, the input/output probabilities are specified as follows (99):

$$p(A_i|B_1, B_2, B_3) = \begin{cases} 1, & \text{if } (A_i|B_1, B_2, B_3) \text{ is part of the genetic code;} \\ 0, & \text{otherwise.} \end{cases}$$

**May's communication channel (77)**

Like Yockey's model, May's communication channel model has a well-defined encoder and decoder. However, the encoder/decoder pair of the genetic system proposed by May differs from Yockey's encoder/decoder pair. In May's model, The DNA sequence is the output of a genetic encoder and the input into an error-introducing channel or the replication process. The

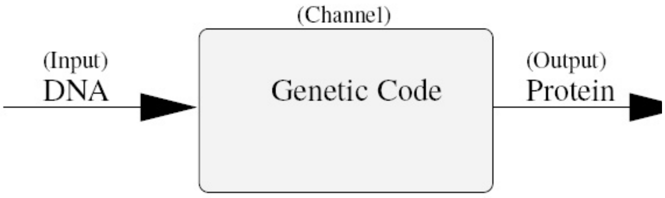


Figure 10. Roman-Raldan's communication channel model.
(99)

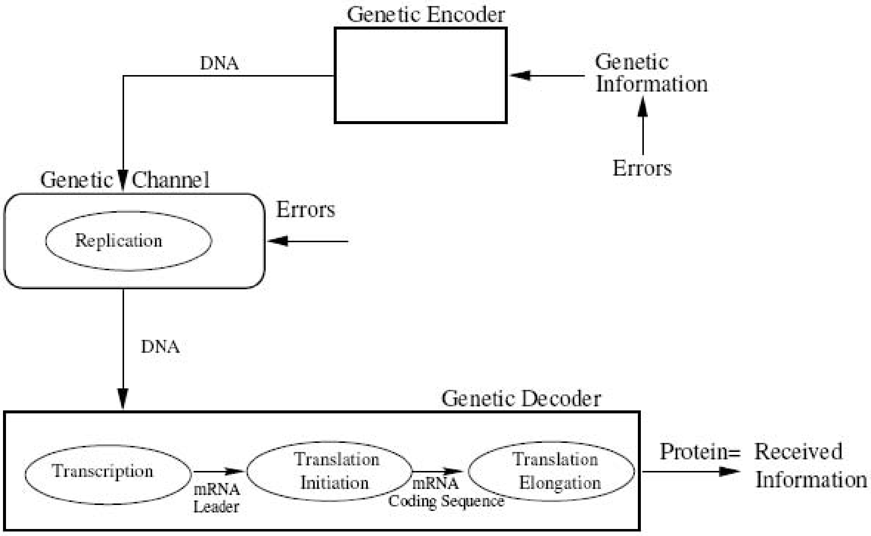


Figure 11. May's biological information transmission model.
(119)

biological process that corresponds to the encoder is unknown. The replicated DNA is decoded by the two-level genetic decoder: transcription and translation. Figure 11 illustrates May's communication view of the genetic process.

All the above communication views of the genetic information storage and transmission apparatus are based on the central dogma of genetics. Even though the described communication systems faithfully reproduce the biological flow of information, they fail to explain the basic elements in a proper biological communication system. Specifically:

- A DNA-Protein system is inconsistent with engineering communication systems, which model transmission and storage of the same messages at the source and destination (excluding errors due to channel degradation). It is, therefore, incorrect to view the translation between DNA sequences and proteins as a communication system. The DNA-protein system is a transformation between the 4-letter alphabet message in the DNA and the 20-letter alphabet message in the amino acid polypeptide. The genetic code dictates this transformation. Thus, from a communication point of view, the DNA-Protein system corresponds to a decoding system and not a communication system.

- Yockey's and Roman-Raldan communication systems view the DNA as the message source and hence completely neglects the true nature of the DNA sequence as the encoded information, which is well established in molecular biology, even though there is no encoding process in biology.

- In Yockey's and Roman-Raldan DNA-Protein system, the source DNA generates the genome according to a specific stochastic process, which uses a 4-letter alphabet. Hence, the DNA-Protein system cannot explain the current structure of DNA, e.g. presence of non-coding DNA and the size of the genetic alphabet.

## 3.2 The Protein Communication System

By abstracting organisms as protein sets, we propose to view the genetic communication system as one between proteins. Like any communication system, the mathematical model of the genetic communication system is composed of an encoder and a decoder. The encoded information is the DNA sequence. The encoding process does not take place in biology since proteins cannot be used to generate DNA. It is only a mathematical model of the protein information captured by DNA. To clarify this idea, assume a computer that maintains an MPEG code while decoding a video for display. Copies of the video to other computers only require sending the MPEG code. Assume further that the first MPEG code was created by chance. This system never encodes a video into MPEG. It only decodes MPEG to display a video. The proper communication model is, however, "video → MPEG → MPEG → video", even though the process "video → MPEG" never occurs. The decoder of the genetic communication system is the biological translation apparatus, which decodes the 4-letter alphabet DNA sequence to the 20-letter alphabet amino acid chain. DNA replication, transcription, translation and external conditions, such as thermal noise, radioactivity and cosmic rays, are sources of errors in this communication channel. Even though proofreading and repair mechanisms are known to detect and correct errors during DNA replication, transcription and translation, they are not perfect and errors still occur. The protein communication system is depicted in Figure 12 The protein communication channel is composed of five elements:

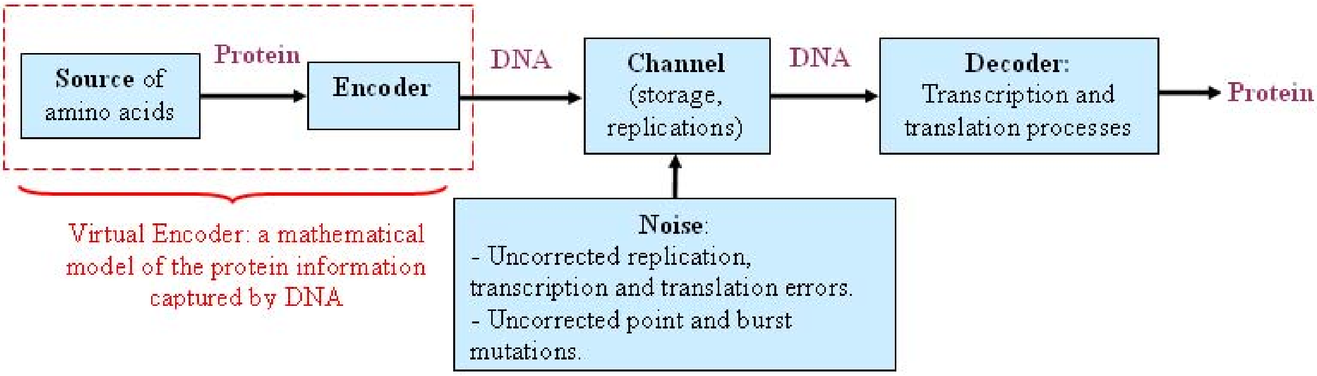

Source of amino acids
Protein
Encoder
DNA
Channel (storage, replications)
DNA
Decoder: Transcription and translation processes
Protein
Virtual Encoder: a mathematical model of the protein information captured by DNA
Noise:
- Uncorrected replication, transcription and translation errors.
- Uncorrected point and burst mutations.

Figure 12. Protein Communication Channel during cell replication.

- *Source:* The source generates the message, i.e., the amino-acid sequence in the protein. It chooses the successive amino acids in our model according to a particular stochastic process.
- *Source and channel encoder:* The encoding process does not take place in biology. It is only a mathematical model of the protein information captured by DNA. Biological organisms have resolved the real communication problem, i.e. "protein → protein", by ensuring that organisms maintain both proteins and DNA. Therefore, the "protein → DNA" encoder is not required biologically. Biological processes only decode DNA into proteins via the translation process. However, the method employed by biological organisms to represent protein information in a DNA sequence does not prevent us from using a nonexistent biological process "protein → DNA" in our information model. In the communication system, in Figure 12, the output of the encoder is the DNA sequence. Hence, we view the DNA sequence as the encoded message and not the information source as was previously argued in the literature (122), (99), (40).

Furthermore, based on the highly redundant structure of the DNA sequence (e.g., presence of a large percentage of noncoding segments), we argue that the encoder models a source and channel encoder. The source encoder efficiently represents the information source in a message, called the "information message". This message is made of symbols belonging to an alphabet, namely {A, T, C, G}. The channel encoder protects the message from channel errors by adding redundancy to the information message. The newly generated message is called a codeword. In this project, a single stranded DNA sequence is regarded as a codeword.

- *Physical Channel:* The physical channel models the transmission and storage medium and the source of errors. Chemical mutagens cause errors in DNA replication. UV and X-rays cause damage to DNA. Many of these errors are corrected by repair enzymes. The overall error rate in DNA ranges from $10^{-7}$ to $10^{-12}$. For a more comprehensive review of DNA damage and repair mechanisms, refer to Section 2.4.1. Therefore, DNA storage and replication is part of the biological communication medium, or physical channel, which introduces errors to the system.

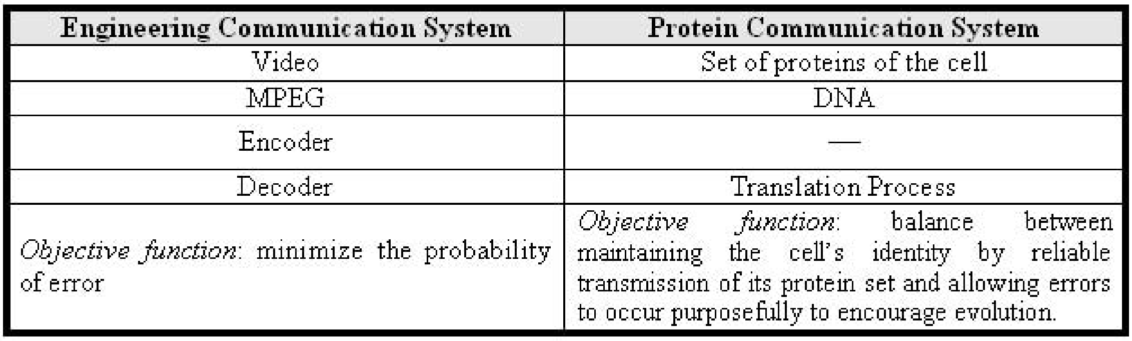
- *Source and channel decoder:* The encoded DNA sequence is transcribed into mRNA; then decoded by the ribosomes from the 4-letter alphabet mRNA sequence to the 20-letter alphabet amino-acid chain in the protein. The decoding process is accomplished based on the well-known genetic code (see Section 2.3). The translation process can make errors; for example, it has been proven experimentally that the protein synthesis system might have difficulty in discriminating between similar amino acids (71). To simplify the communication model, these errors are incorporated as part of the physical channel.

- *Destination:* : The decoded one-dimensional amino acid sequence is the received message, which folds up to become a three-dimensional active protein molecule.

In sexual reproduction with k offsprings, a protein communication model accounting for the k offsprings is equivalent to k independent communication systems in parallel, each modelling 1 offspring. Therefore, it is sufficient to consider a protein communication system modelling sexual reproduction with one offspring. However, a protein communication system, which models the transmission of information in sexual reproduction, is much more involved mathematically than the single source communication system in cell replication. From an information theoretic perspective, we have two sources; each source is a parent containing two homologous protein sets. The output of this communication model consists of two proteins, randomly selected from each parent, received after transmission over the communication channel. Analysis of this communication system requires the use of multi-user information theory and distributed coding. For analytical simplicity, we decompose this complex system into two parallel communication systems. Each communication system consists of a source (a single parent) generating two homologous protein sets. A stochastic process selects one protein from each homologous pair. The selected protein is transmitted through an identical communication system to the single source protein channel depicted in Fig. 1. The received message is formed by the union of the two proteins received from each parent.

Even though May introduced a communication model with a virtual genetic encoder, where the DNA is the encoded information and the proteins are the decoded information, she somehow failed to model the information source as a source of amino acid alphabets. The taboo of using a mathematical model of a virtual genetic encoder of proteins traces its roots, in our opinion, to the Central Dogma introduced by Crick (28), which states that the transfer of information from protein to DNA or mRNA is strictly forbidden. In his DNA-Protein system, Yockey (122) pointed out that the Central Dogma is a property of any code in which the source alphabet is larger than the destination alphabet. He further emphasizes that "The Central Dogma is not regarded as a first principle or axiom of molecular biology" and it does not forbid the protein-protein transfer of information" (122). Bernstein (100) points out that the question of the existence of the protein-protein transfer of information is important to the problem of scrapie, a protein that appears to have a self-replication mechanism, and to the scenarios of the origin of life. If, at the outset of life, proteins came first, then there is a requirement for replication through a protein communication channel. It is important to emphasize however that, in this thesis, we are not supporting the theory of a biological protein-protein genetic code. The proposed protein communication system is a mathematical model of information transmission during cell division. This model does not support either the theories of proteins-first or nucleotides-first at the origin of life. It is merely an abstraction, which models a cell as a set of proteins and the process of cell division as an information communication system between protein sets. In fact, the proposed biological communication model could be used to explain the transmission of information in both the proteins-first and nucleotides-first theories.

There are two main differences between the genetic information processing system and the communication engineer's system: The first is that biology does not encode proteins into DNA. It only decodes genes into proteins. Biological organisms have resolved the basic communication



| Engineering Communication System | Protein Communication System |
|---|---|
| Video | Set of proteins of the cell |
| MPEG | DNA |
| Encoder | — |
| Decoder | Translation Process |
| *Objective function*: minimize the probability of error | *Objective function*: balance between maintaining the cell's identity by reliable transmission of its protein set and allowing errors to occur purposefully to encourage evolution. |

Figure 13. Comparison between the engineering communication system for video transmission and the protein communication system during cell replication.

problem, i.e. "protein → protein", by ensuring that organisms maintain both proteins and DNA. Therefore, the "protein → DNA" encoder is not required in biology. The second is that, unlike the communication engineer's system, the biological communication system is not designed to minimize transmission errors. In the absence of errors, evolution will not be possible. Intuitively, there has to be a balance between maintaining the cell's identity by reliable transmission of its protein set and allowing errors to occur purposefully to encourage evolution. Figure 13 summarizes the analogy between an engineering communication system for video transmission and the protein communication system.

Using the mathematical model of protein communication during cell replication, we will translate the problem of a species' evolution into the language of mathematics, in particular the language of communication theory. The problem of a species' evolution will be represented as the iteration of a communication channel over time. We will investigate the structure of the genetic code; in particular the presence of “junk DNA” and the dynamics of the channel over time. The mathematical model of the protein communication system provides a unified framework to study the above questions and more from a communication theory perspective.

# CHAPTER 4

# DISTRIBUTION AND CONVERGENCE ANALYSIS OF AMINO ACIDS

> *“Ah, but my Computations, People say,*
>
> *Reduced the year to better reckoning? Nay,*
>
> *T’was only striking from the Calendar*
>
> *Unborn tomorrow and dead Yesterday.”*
>
> Omar Khayyam, The Rubaiyat.

## Abstract

*In this Chapter, we study the evolutionary dynamics of the protein communication channel in both cases of constant and time-varying point mutation rate. We establish, using matrix analysis, that stochastic messages sent through the channel of evolution are received according to a fixed probability distribution, which is independent of the original message.*

## 4.1 Introduction

The protein communication channel is time-dependent: thermal noise, radioactivity and cosmic rays are sources of errors and they occur with a probability that is a function of time regardless of the number of replications of the DNA. A series connection of the protein communication channel is equivalent to a channel through time: “the channel of evolution”. In this chapter, we will investigate the behavior of this channel. Specifically, we will address the following questions:

1. Given an infinitely small probability of error at each generation of cell replication, how are the cell offsprings related to their ancestral mother cell after a large number of generations?
2. Given an initial distribution of amino acids, how does this distribution evolve with time? Is there an equilibrium distribution? If yes, what is the rate of convergence to this equilibrium distribution and what are the biological implications of such equilibrium?

### 4.1.1 The PAM Model of Protein Sequence Evolution

The original substitutions rate matrices for amino acids were termed PAM matrices, where PAM stands for "point accepted mutation". An accepted mutation is one that spreads through the population and goes to fixation. The PAM model of amino acid substitutions was developed by Dayhoff, Schwartz and Orcutt (30). At that time, relatively few protein sequences were available. Several groups have since used much larger data sets to derive improved PAM models, but the methods used have remained similar. Dayhoff, Schwartz and Orcutt made alignments of 71 families of closely related proteins. Sequences in the same family were more than 85% similar to one another. They constructed an evolutionary tree for each family using the parsimony method: The selected tree is the one that minimizes the total number of amino acid substitutions required. Having determined all the substitutions, the obtained a matrix whose elements , $A_{i,j}$, are the number of times that amino acid $i$ is substituted by amino acid $j$. The $A_{i,j}$ matrix obtained by Dayhoff *et al.* (30) involved all 20 amino acids and had 1572 substitutions. An estimate of the substitution rate is proportional to the number of observed substitutions, $A_{i,j}$, divided by the total number of times $N_i$ that amino acid $i$ is seen in the data. Hence

$$M_{i,j} = \lambda \frac{A_{i,j}}{N_i} \quad (\text{for } i \neq j). \tag{4.1}$$

where $\lambda$ is a constant of proportionality to be determined. The frequency of amino acid $i$ in the data is $\pi_i = N_i/N_{tot}$, where $N_{tot}$ is the total number of amino acids in the data set. To determine $\lambda$, Dayhoff *et al.* adopted the convention that 1 PAM unit is the time such that an average of 1% of amino acids have changed. The fraction of sites that have changed is

$$\sum_i \pi_i \sum_{j \neq i} M_{ij} = \sum_i \pi_i \sum_{j \neq i} \pi_i \lambda \frac{A_{ij}}{N_{tot} \pi_i} = \frac{\lambda A_{tot}}{N_{tot}} = 0.01 \tag{4.2}$$

where $A_{tot}$ is the total of all the elements in the $A_{ij}$ matrix. Hence

$$\lambda = 0.01 \frac{N_{tot}}{A_{tot}} \tag{4.3}$$

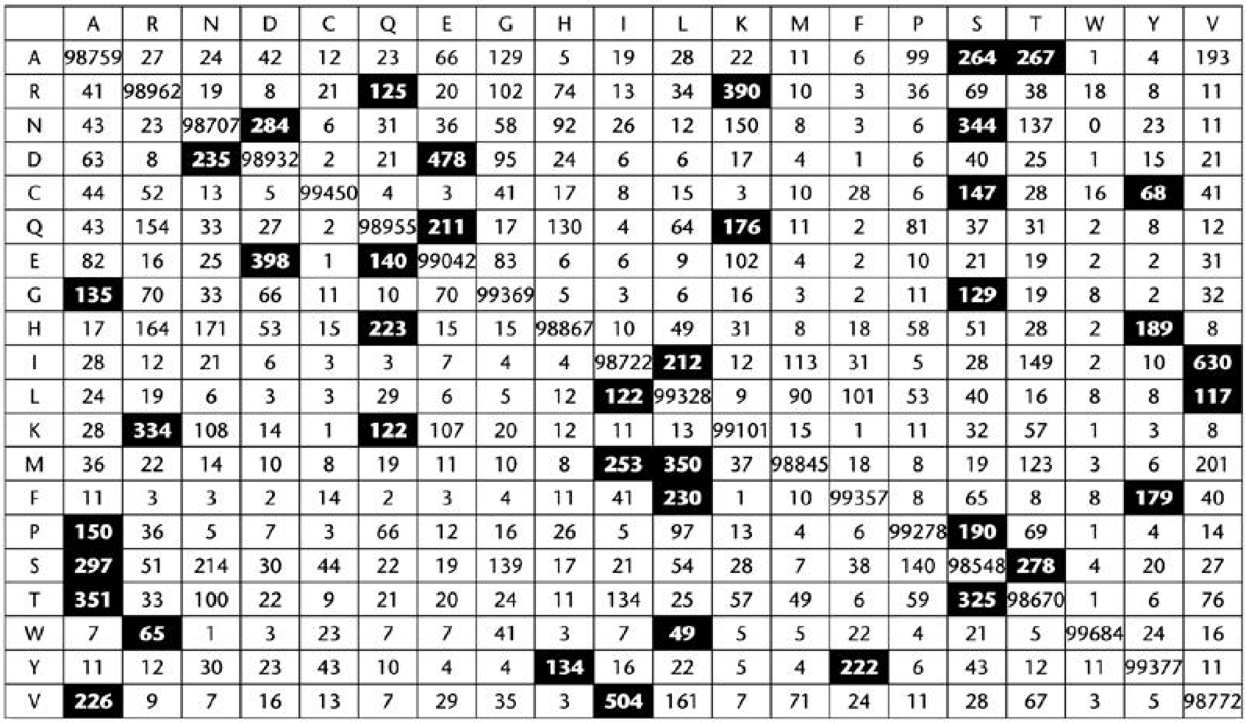
The $\mathbf{PAM}_1$ matrix obtained by Jones *et al.* (57) is shown in Figure 14. Values have been multiplied by $10^5$ for convenience. Since $\mathbf{PAM}_1$ corresponds to an average probability of 0.01 of amino acids changing, we would expect all diagonal elements to be 0.99 if each amino acid changed at the same rate. However, some amino acids are more likely to undergo substitutions than others. Amino acids that change more rapidly than average will have a probability lower than 0.99 of remaining unchanged after a time of 1 PAM, whereas those that change more slowly than average will have a probability higher than 0.99 of remaining unchanged. The two highest non-diagonal elements in each row have been highlighted in black. These are the two highest substitution rates for each amino acid, e.g., A is more likely to change to S and T than the other amino acids. The $\mathbf{PAM}_1$ matrix is used as the basis for calculating other matrices by assuming that repeated mutations would follow the same pattern as those in the $\mathbf{PAM}_1$ matrix, and multiple substitutions can occur at the same site so that $\mathbf{PAM}_n = \mathbf{PAM}_1^n$. Using this logic, Dayhoff derived matrices as high as $\mathbf{PAM_{250}}$. The number of times we multiply the matrix is called the PAM distance. PAM substitution matrices are crucial in evaluating the quality of a pairwise sequence alignment, which assigns a score for aligning any possible pair of residues. $\mathbf{PAM_{250}}$ is the default in BLAST (Basic Local Alignment Search Tool) search.

In our work, we will use the PAM substitution matrices as probability transition matrices for the protein communication channel. The proofs of all theoretical results that are new contributions are presented in the appendix at the end of the chapter.

## 4.2 Mathematical Characterization of the Protein Communication Channel

The protein communication channel is uniquely characterized by its probability transition matrix. The $(i, j)$ entry of this matrix, Pr $(P_j|P_i)$, is the probability of receiving protein $P_j = (a_1^j, \cdots, a_N^j)$ given that protein $P_i = (a_1^i, \cdots, a_N^i)$ was transmitted. We assume that the protein channel is memoryless. Hence, we have

$$\text{Pr } (P_j|P_i) = \prod_{k=1}^{N} \text{Pr } (a_k^j|a_k^i), \tag{4.4}$$



| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 98759 | 27 | 24 | 42 | 12 | 23 | 66 | 129 | 5 | 19 | 28 | 22 | 11 | 6 | 99 | **264** | **267** | 1 | 4 | 193 |
| R | 41 | 98962 | 19 | 8 | 21 | **125** | 20 | 102 | 74 | 13 | 34 | **390** | 10 | 3 | 36 | 69 | 38 | 18 | 8 | 11 |
| N | 43 | 23 | 98707 | **284** | 6 | 31 | 36 | 58 | 92 | 26 | 12 | 150 | 8 | 3 | 6 | **344** | 137 | 0 | 23 | 11 |
| D | 63 | 8 | **235** | 98932 | 2 | 21 | **478** | 95 | 24 | 6 | 6 | 17 | 4 | 1 | 6 | 40 | 25 | 1 | 15 | 21 |
| C | 44 | 52 | 13 | 5 | 99450 | 4 | 3 | 41 | 17 | 8 | 15 | 3 | 10 | 28 | 6 | **147** | 28 | 16 | **68** | 41 |
| Q | 43 | 154 | 33 | 27 | 2 | 98955 | **211** | 17 | 130 | 4 | 64 | **176** | 11 | 2 | 81 | 37 | 31 | 2 | 8 | 12 |
| E | 82 | 16 | 25 | **398** | 1 | **140** | 99042 | 83 | 6 | 6 | 9 | 102 | 4 | 2 | 10 | 21 | 19 | 2 | 2 | 31 |
| G | **135** | 70 | 33 | 66 | 11 | 10 | 70 | 99369 | 5 | 3 | 6 | 16 | 3 | 2 | 11 | **129** | 19 | 8 | 2 | 32 |
| H | 17 | 164 | 171 | 53 | 15 | **223** | 15 | 15 | 98867 | 10 | 49 | 31 | 8 | 18 | 58 | 51 | 28 | 2 | **189** | 8 |
| I | 28 | 12 | 21 | 6 | 3 | 3 | 7 | 4 | 4 | 98722 | **212** | 12 | 113 | 31 | 5 | 28 | 149 | 2 | 10 | **630** |
| L | 24 | 19 | 6 | 3 | 3 | 29 | 6 | 5 | 12 | **122** | 99328 | 9 | 90 | 101 | 53 | 40 | 16 | 8 | 8 | **117** |
| K | 28 | **334** | 108 | 14 | 1 | **122** | 107 | 20 | 12 | 11 | 13 | 99101 | 15 | 1 | 11 | 32 | 57 | 1 | 3 | 8 |
| M | 36 | 22 | 14 | 10 | 8 | 19 | 11 | 10 | 8 | **253** | **350** | 37 | 98845 | 18 | 8 | 19 | 123 | 3 | 6 | 201 |
| F | 11 | 3 | 3 | 2 | 14 | 2 | 3 | 4 | 11 | 41 | **230** | 1 | 10 | 99357 | 8 | 65 | 8 | 8 | **179** | 40 |
| P | **150** | 36 | 5 | 7 | 3 | 66 | 12 | 16 | 26 | 5 | 97 | 13 | 4 | 6 | 99278 | **190** | 69 | 1 | 4 | 14 |
| S | **297** | 51 | 214 | 30 | 44 | 22 | 19 | 139 | 17 | 21 | 54 | 28 | 7 | 38 | 140 | 98548 | **278** | 4 | 20 | 27 |
| T | **351** | 33 | 100 | 22 | 9 | 21 | 20 | 24 | 11 | 134 | 25 | 57 | 49 | 6 | 59 | **325** | 98670 | 1 | 6 | 76 |
| W | 7 | **65** | 1 | 3 | 23 | 7 | 7 | 41 | 3 | 7 | **49** | 5 | 5 | 22 | 4 | 21 | 5 | 99684 | 24 | 16 |
| Y | 11 | 12 | 30 | 23 | 43 | 10 | 4 | 4 | **134** | 16 | 22 | 5 | 4 | **222** | 6 | 43 | 12 | 11 | 99377 | 11 |
| V | **226** | 9 | 7 | 16 | 13 | 7 | 29 | 35 | 3 | **504** | 161 | 7 | 71 | 24 | 11 | 28 | 67 | 3 | 5 | 98772 |

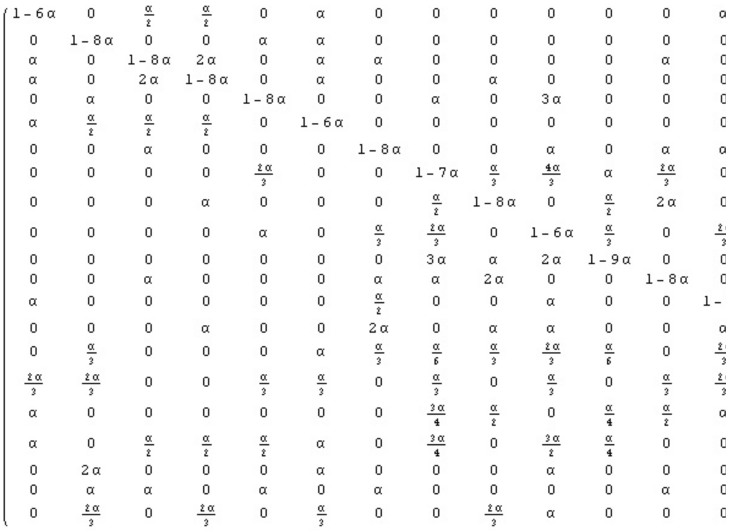
Figure 14. Mutation probability matrix for the evolutionary distance of 1 PAM calculated by Jones *et al.* [56]. For display clarity, values are multiplied by $10^5$. $M_{i,j}$ is the probability that amino acid in row $i$ changes to the amino acid in column $j$ in a 1 PAM time interval. The two highest non-diagonal elements in each row are highlighted in black. These are the two highest substitution rates for each amino acid.

From the above equation, we see that it is sufficient to study the probability transition matrix, $\mathbf{Q}(k) = \{q_{i,j}(k)\}_{1\leq i,j\leq 20}$, at time $k$, of the amino acids.

In our work, we use two different probability transition matrices: $\mathbf{PAM_{250}}$ probability transition matrix (30) and a first-order Markov transition probability matrix, **P**. **P** is constructed from the genetic code as follows: Let $\alpha(k)$ be the probability of a base interchange of any one nucleotide at time $k$, all interchanges being equally probable. Assuming that the 64 codons are equally probable and from Baye's rule, we obtain the following formula for the probability of a transition from amino acid $a$ to amino acid $\hat{a}$,

$$\begin{aligned} \Pr(\hat{a}|a) &= \Pr(\{c_1, \cdots, c_n\}|\{b_1, \cdots, b_m\}) \\ &= \frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha(k)^{h(b_j,c_i)}(1-3\alpha(k))^{3-h(b_j,c_i)}, \end{aligned} \tag{4.5}$$

where $\{c_1, \cdots, c_n\}$, (resp., $\{b_1, \cdots, b_m\}$), are the codons of the received, (resp., transmitted), amino acid and $h(b_j, c_i)$ is the hamming distance between codon $b_j$ and codon $c_i$. For computational efficiency and since burst mutations are less likely to happen than 1 point mutations, we retain only the terms of the first degree in $\alpha(k)$. The probability transition matrix **P** is displayed in Figure 15. The amino acids are alphabetically ordered by their one-letter standard abbreviations, e.g., $p_{1,1} = \Pr(A|A)$, $p_{1,2} = \Pr(A|C)$, etc.



$$\left(\begin{array}{ccccccccccccc}
1-6\alpha & 0 & \frac{\alpha}{2} & \frac{\alpha}{2} & 0 & \alpha & 0 & 0 & 0 & 0 & 0 & 0 & \alpha \\
0 & 1-8\alpha & 0 & 0 & \alpha & \alpha & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\alpha & 0 & 1-8\alpha & 2\alpha & 0 & \alpha & \alpha & 0 & 0 & 0 & 0 & \alpha & 0 \\
\alpha & 0 & 2\alpha & 1-8\alpha & 0 & \alpha & 0 & 0 & \alpha & 0 & 0 & 0 & 0 \\
0 & \alpha & 0 & 0 & 1-8\alpha & 0 & 0 & \alpha & 0 & 3\alpha & 0 & 0 & 0 \\
\alpha & \frac{\alpha}{2} & \frac{\alpha}{2} & \frac{\alpha}{2} & 0 & 1-6\alpha & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \alpha & 0 & 0 & 0 & 1-8\alpha & 0 & 0 & \alpha & 0 & \alpha & \alpha \\
0 & 0 & 0 & 0 & \frac{2\alpha}{3} & 0 & 0 & 1-7\alpha & \frac{\alpha}{3} & \frac{4\alpha}{3} & \alpha & \frac{2\alpha}{3} & 0 \\
0 & 0 & 0 & \alpha & 0 & 0 & 0 & \frac{\alpha}{2} & 1-8\alpha & 0 & \frac{\alpha}{2} & 2\alpha & 0 \\
0 & 0 & 0 & 0 & \alpha & 0 & \frac{\alpha}{3} & \frac{2\alpha}{3} & 0 & 1-6\alpha & \frac{\alpha}{3} & 0 & \frac{2}{3} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 3\alpha & \alpha & 2\alpha & 1-9\alpha & 0 & 0 \\
0 & 0 & \alpha & 0 & 0 & 0 & \alpha & \alpha & 2\alpha & 0 & 0 & 1-8\alpha & 0 \\
\alpha & 0 & 0 & 0 & 0 & 0 & \frac{\alpha}{2} & 0 & 0 & \alpha & 0 & 0 & 1- \\
0 & 0 & 0 & \alpha & 0 & 0 & 2\alpha & 0 & \alpha & \alpha & 0 & 0 & \alpha \\
0 & \frac{\alpha}{3} & 0 & 0 & 0 & \alpha & \frac{\alpha}{3} & \frac{\alpha}{6} & \frac{\alpha}{3} & \frac{2\alpha}{3} & \frac{\alpha}{6} & 0 & \frac{2}{3} \\
\frac{2\alpha}{3} & \frac{2\alpha}{3} & 0 & 0 & \frac{\alpha}{3} & \frac{\alpha}{3} & 0 & \frac{\alpha}{3} & 0 & \frac{\alpha}{3} & 0 & \frac{\alpha}{3} & \frac{2}{3} \\
\alpha & 0 & 0 & 0 & 0 & 0 & 0 & \frac{3\alpha}{4} & \frac{\alpha}{2} & 0 & \frac{\alpha}{4} & \frac{\alpha}{2} & \alpha \\
\alpha & 0 & \frac{\alpha}{2} & \frac{\alpha}{2} & \frac{\alpha}{2} & \alpha & 0 & \frac{3\alpha}{4} & 0 & \frac{3\alpha}{2} & \frac{\alpha}{4} & 0 & 0 \\
0 & 2\alpha & 0 & 0 & 0 & \alpha & 0 & 0 & 0 & \alpha & 0 & 0 & 0 \\
0 & \alpha & \alpha & 0 & \alpha & 0 & \alpha & 0 & 0 & 0 & 0 & \alpha & 0 \\
0 & \frac{2\alpha}{3} & 0 & \frac{2\alpha}{3} & 0 & \frac{\alpha}{3} & 0 & 0 & \frac{2\alpha}{3} & \alpha & 0 & 0 & 0
\end{array}\right.$$

Figure 15. **P**: a first-order Markov probability transition matrix between amino acids. For display clarity, the dependence of $\alpha$ on the time $k$ has been omitted. Only the terms of the first degree in $\alpha(k)$ are retained.

Let $\mathbf{p}_0$ be the row probability vector of the initial distribution of the amino acids (at time 0). It is straightforward to show that the row probability vector of the amino acids at time $k$ is given by

$$\mathbf{p}_k = \mathbf{p}_0 \ \mathbf{Q}(1) \ \mathbf{Q}(2) \ \cdots \mathbf{Q}(k), \tag{4.6}$$

where $\mathbf{Q} \in \{\mathbf{PAM_{250}}, \mathbf{P}\}$. Observe that $\mathbf{P}$ takes into account all possible mutations between amino acids whether they are accepted or rejected by natural selection whereas the PAM transition matrix is estimated from phylogenetic trees of protein sequences and hence takes into account the accepted mutations only.

### 4.3 Constant Point Mutation Rate

In this section, we assume that the point mutation rate is constant over time, i.e., $\alpha(k) = \alpha$, for all $k \geq 0$. Hence, Equation 4.6 becomes

$$\mathbf{p}_k = \mathbf{p}_0 \ \mathbf{Q}^k. \tag{4.7}$$

**Proposition 1** *Consider an initial probability distribution of the amino acids at time 0,* $\mathbf{p}_0$*. Then, the probability distribution of the amino acids converges, over time, towards the stationary distribution given by*

$$\begin{cases} \mathbf{s_1}, & \textit{if } \mathbf{Q} = \mathbf{P}; \\ \mathbf{s_2}, & \textit{if } \mathbf{Q} = \mathbf{PAM_{250}}, \end{cases}$$

*where*

$$\mathbf{s_1} = (\tfrac{4}{61}, \tfrac{2}{61}, \tfrac{2}{61}, \tfrac{2}{61}, \tfrac{2}{61}, \tfrac{4}{61}, \tfrac{2}{61}, \tfrac{3}{61}, \tfrac{2}{61}, \tfrac{6}{61}, \tfrac{1}{61}, \tfrac{2}{61}, \tfrac{4}{61}, \tfrac{2}{61}, \tfrac{6}{61}, \tfrac{6}{61}, \tfrac{4}{61}, \tfrac{4}{61}, \tfrac{1}{61}, \tfrac{2}{61}). \tag{4.8}$$

*and*

$$\mathbf{s_2} = \begin{array}{l} (0.0873,\ 0.0338,\ 0.0479,\ \ 0.05,\ \ 0.0383,\ 0.0909,\ 0.0330,\ 0.0375,\ 0.0808,\ 0.0844,\ 0.0143,\ 0.0411,\ 0.0522,\ 0.0390,\ 0.0406,\ 0.0704, \\ 0.0594,\ \ 0.0651,\ 0.0075,\ 0.0294). \end{array} \tag{4.9}$$

In order to make biological sense of the limiting distribution vectors $\mathbf{s_1}$ and $\mathbf{s_2}$, we compare them to the experimental distribution of amino acids computed in the literature (17), (46), (57), (59). We found that there are some fluctuations between the different experimental

TABLE I. Correlation between the experimental frequencies of amino acids and the limiting distributions $\mathbf{s_1}$ and $\mathbf{s_2}$.

| Experimental distribution $\mathbf{r}$ | Correlation coeff. between $\mathbf{r}$ and $\mathbf{s_2}$ | Correlation coeff. between $\mathbf{r}$ and $\mathbf{s_1}$ |
|---|---|---|
| $\mathbf{r}$ in (46) | 0.96 | 0.66 |
| $\mathbf{r}$ in (59) | 0.937 | 0.632 |
| $\mathbf{r}$ in (17) Eukaryotes | 0.824 | 0.74 |
| $\mathbf{r}$ in (17) Bacteria | 0.836 | 0.701 |
| $\mathbf{r}$ in (17) Archaea | 0.76 | 0.602 |
| $\mathbf{r}$ in (17) all taxa | 0.834 | 0.7 |

distributions. The reason behind this disparity is that different experiments use different sets of organisms and different protein families. Let us denote by $\mathbf{r}$ the experimental probability vector of the amino acids. Table V displays the correlation coefficients between the different experimental distributions and the limiting distributions $\mathbf{s_1}$ and $\mathbf{s_2}$. Since $\mathbf{PAM_{250}}$ estimates the rate of accepted mutations only, we find that the limiting distribution $\mathbf{s_2}$ has a higher correlation with the experimental distribution $\mathbf{r}$ than the limiting distribution $\mathbf{s_1}$. Moreover, the highest correlation was obtained between $\mathbf{s_2}$ and the experimental distribution computed in (46). Figure 16 shows the plot of $\mathbf{r}$ in (46) versus $\mathbf{s_2}$. Another interesting observation is that the limiting distributions have higher correlations with the experimental frequencies of amino acids calculated from Eukaryotes and Bacteria than the experimental frequency calculated from Archaea (see rows 3, 4 and 5 of Table V).

Notice that $\mathbf{s_1}$ is proportional to the number of codon assignment for the amino acids. So, $\mathbf{s_1}$ is the distribution of the amino acids if the codons were randomly distributed in the genome.

TABLE II. Mean experimental and limiting distributions of the amino acid classes.

| Classes | Mean experimental prob. in (46) | Mean limiting prob. (**P**) | Mean limiting prob. (PAM) |
|---|---|---|---|
| $C_1$ | 0.0155 | 0.0163 | 0.01075 |
| $C_2$ | 0.045 | 0.0327 | 0.044 |
| $C_3$ | 0.04843 | 0.0492 | 0.0508 |
| $C_4$ | 0.0656 | 0.0656 | 0.0709 |
| $C_6$ | 0.0663 | 0.0983 | 0.0665 |

Equivalently, we can view $\mathbf{s_1}$ as the distribution of amino acids if all randomly distributed point mutations were accepted by Nature (i.e., survived). According to this view, the discrepancy between $\mathbf{s_1}$ and $\mathbf{s_2}$ can be related to the relative probability of survival of the amino acids after mutations. We shall divide the amino acids into classes $C_1, C_2, C_3, C_4$ and $C_6$ , the subscripts indicating the number of codons for each class. For example, the class $C_1$ contains two amino acids: Met (M) and Trp (W), i.e., $C_1 = \{M, W\}$. The mean experimental in (46) and limiting distributions using both matrices **P** and $\mathbf{PAM_{250}}$ are displayed in Table VI. The mean experimental and limiting distributions, for each class, are very close except for the class of amino acids corresponding to 6 codons obtained from the limiting distribution using the probability transition matrix **P**. The reason is that Arginine, which is coded by 6 codons, appears with a much lower frequency than $\frac{6}{61}$. This has been ascribed to the rare appearance of the CG base doublet so that, in fact, in most observed proteins, arginine is coded only by AGA and AGG (122).

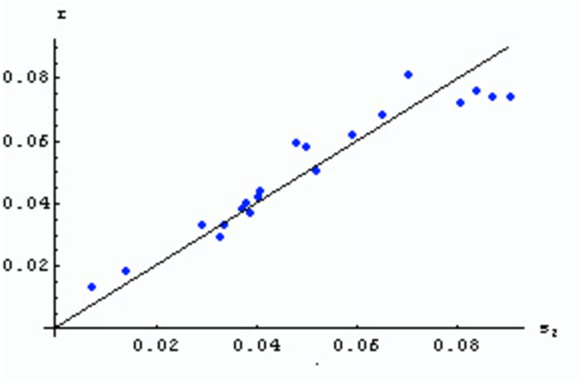


Figure 16. The experimental distribution of amino acids $\mathbf{r}$ in v.s. the limiting distribution $\mathbf{s_2}$ given by the $\mathbf{PAM_{250}}$ probability transition matrix.

A question naturally arises now: what is the rate of convergence? And how is this rate related to the rate of point mutation $\alpha$ ? The answer is provided in the following proposition:

**Proposition 2** $\{\mathbf{p}_0\mathbf{Q}^k\}_{k\geq 1}$ *converges at a geometric rate with parameter* $|\lambda_2|$*, where*

$$\begin{cases} |\lambda_2| = 0.53, & \textit{if} \;\; \mathbf{Q} = \boldsymbol{PAM}_{250}; \\ |\lambda_2| \leq 1 - \frac{1}{2}\alpha, & \textit{if} \;\; \mathbf{Q} = \mathbf{P}. \end{cases} \tag{4.10}$$

Thus, the convergence rate for $\mathbf{P}$ is no slower than $\mathsf{O}((1-\frac{1}{2}\alpha)^k)$. Moreover, when $\alpha$ decreases, the convergence is slower and vice versa. This result is somehow intuitive and, as a consequence, proves that no evolution is possible if $\alpha = 0$.

## 4.4 Time-Varying Point Mutation Rate

In this section, we consider a rate of point mutation, $\alpha(k)$, which varies in time. Consider the products $\mathbf{T}_{p,k} = \{t_{i,j}^{(p,k)}\} = \mathbf{Q}_{p+1}\mathbf{Q}_{p+2}\cdots\mathbf{Q}_{p+k}$ for every $p \geq 0$. For a fixed $p$, let $t$ be the smallest integer satisfying $\mathbf{T}_{p,t} > 0$, in the sense that all its entries are strictly positive.

**Definition 1 (Weak and Strong Ergodicity)** *(109) The forward products* $\mathbf{T}_{p,k}$ *are said to be* weakly ergodic *if*

$$t_{i,s}^{p,k} - t_{j,s}^{p,k} \xrightarrow{k\to\infty} 0 \quad \textit{for each } i,j,s,p. \tag{4.11}$$

*If weak ergodicity is obtained and the* $t_{i,s}^{p,k}$ *themselves tend to a limit for all* $i,s,p$*, i.e.,* $t_{i,j}^{(p,k)} \xrightarrow{k\to\infty} v_j^{(p)}$*, then we say* strong ergodicity *is obtained.*

Moreover, if strong ergodicity obtains, then the limit row vector $\mathbf{v}_p = \{v_j^{(p)}\}$ is a probability vector and is independent of $p \geq 0$, i.e., $\mathbf{v}_p = \mathbf{v}$ (109). Hence, strong ergodicity is equivalent to the existence of the limit of $\mathbf{T}_{p,k}$ as $k \to \infty$, for all $p \geq 0$.

**Definition 2** *(109) A matrix* $\mathbf{Q} = \{q_{i,j}\}$ *is called a* scrambling *matrix if given any two rows* $\beta$ *and* $\delta$*, there is at least one column* $\rho$ *such that* $q_{\beta,\rho} > 0$ *and* $q_{\delta,\rho} > 0$.

It is easy to show, that since every transition matrix at time $k$, $\mathbf{Q}(k)$, is scrambling, then so is $\mathbf{T}_{p,k}, p \geq 0$.

**Theorem 3** *Consider a finite number of PAM matrices denoted by* $\mathbf{PAM(1)}$, $\cdots$, $\mathbf{PAM(N)}$*, where* $\mathbf{PAM(i)}$ *can be* $\mathbf{PAM_1}$ *or* $\mathbf{PAM_{160}}$ *or* $\mathbf{PAM_{250}}$*, etc, for all* $i = 1,\cdots,N$*. Consider the sequence:* $\mathbf{T}_{p,k} = \mathbf{t}_{p+1}\mathbf{t}_{p+2}\cdots\mathbf{t}_{p+k}$*, where each* $\mathbf{t}_i \in \{\mathbf{PAM(1)},\cdots\mathbf{PAM(N)}\}$*. That is at* *each time $k$, the probability transition matrix is some PAM matrix (the evolutionary time of the PAM matrix and the time $k$ are not necessarily equal). Then, $\mathbf{T}_{p,k}$ is weakly ergodic at a uniform geometric rate for all $p \geq 0$. So the sequence $\{\mathbf{p}_k\}_{k\geq 1}$, in Equation 4.6, tends to a sequence of distributions independently of $\mathbf{p_0}$.*

If we approximate the matrices $\mathbf{PAM_k}$ by $\mathbf{PAM_1^k}$, the sequence $\mathbf{T}_{p,k} = \mathbf{PAM^{p+1}}\, \mathbf{PAM^{p+2}} \cdots \mathbf{PAM^{p+k}}$ becomes strongly ergodic. In particular, the sequence $\{\mathbf{p}_k\}_{k\geq 1}$, in Equation 4.6, converges to the limiting distribution $\mathbf{s_2}$ given in Equation 4.9.

**Theorem 4** *Consider a point mutation rate, $\alpha(k)$, which is bounded uniformly on $k$, i.e., $0 < a \leq \alpha(k) \leq b < 1$. Then the products $\mathbf{T}_{p,k} = \mathbf{P}_{p+1} \cdots \mathbf{P}_{p+k}$ are strongly ergodic. Thus, the sequence $\{\mathbf{p}_k\}_{k\geq 1}$, in Equation 4.6, converges towards the stationary distribution $\mathbf{s_1}$, in Equation 4.9, independently of the initial distribution $\mathbf{p}_0$. Moreover, the convergence rate is at least geometric with parameter $(1-\gamma^t)^{\frac{1}{t}}$, where $\gamma = \min\{\frac{a}{6}, 1-9b\}$.*

The convergence result, in both the constant and time-varying cases, implies that, after a sufficiently long time, the channel characteristics will determine the final distribution which will be independent of the initial distribution. This conclusion has very different ramifications on bioinformatics than on communication engineering: The convergence analysis in engineering is interpreted as a loss of information after an infinite number of transmissions. The reason is that, in communications, only the initial distribution (i.e. the message) is used to convey information and not the channel. In bioinformatics, on the other hand, the final distribution of amino acids captures the information of the channel (i.e. the mutations) regardless of the initial distribution. The critical information in modelling the channel of evolution is therefore the representation of the channel and not the starting point of the evolutionary process. These implications are verified experimentally (59).

We can obtain similar results with other amino acid substitution matrices, e.g., the BLOSUM (50) probability transition matrix constructed from the log-odds BLOSUM matrix. The convergence of the probability transition matrix shows that a parent organism will be unrelated to its offsprings after infinitely many generations no matter how small the initial point mutation rate is as long as it is non-zero. The rate of convergence quantifies the speed of this divergence. The limiting distribution $\mathbf{s_1}$ shows that, if all mutations were accepted, the asymptotic abundance of amino acids in nature would be proportional to their codon assignment. The discrepancy between this limiting distribution and the natural abundance can be related to the relative survival of the amino acids after they mutate. More accurate transition matrices can be built, where for example, zero is substituted for $\alpha$ if the replacement amino acid is functionally acceptable at a given site in the protein sequence. The mathematical tools used will apply, under some general conditions on the matrices.

**Appendix**

**Proof 1 (Proof of Proposition 1)** *The probability transition matrices* $\mathbf{P}$ *and* $\boldsymbol{PAM}_{250}$ *are irreducible and aperiodic. Therefore, from the Perron-Frobenius theorem (37), there exists a unique stationary probability row vector* $\mathbf{s_1}$ *(resp.,* $\mathbf{s_2}$*) such that the sequence of powers* $\{\mathbf{p}_0\mathbf{P}^k\}_{k\in\mathbb{N}}$ *(resp.,* $\{\mathbf{p}_0\ \boldsymbol{PAM}_{250}^k\}_{k\in\mathbb{N}}$*) approaches the fixed probability vector* $\mathbf{s_1}$ *(resp.,* $\mathbf{s_2}$*) as* $k \to \infty$*. Moreover,* $\mathbf{s_1}$ *and* $\mathbf{s_2}$ *are independent of the initial distribution* $\mathbf{p}_0$*. The stationary* *probability vector* $\mathbf{s_1}$ *(resp.,* $\mathbf{s_2}$*) is the unique solution of the linear system* $\mathbf{s_1 P} = \mathbf{s_1}$ *(resp.,* $\mathbf{s_2}$ $PAM_{250} = \mathbf{s_2}$*), subject to* $\mathbf{s_1 1} = 1$ *(resp.,* $\mathbf{s_2 1} = 1$*), where* $\mathbf{1}$ *is the column vector with all its entries equal to 1.*

**Proof 2 (Proof of Proposition 2)** *The matrix* $\mathbf{Q} \in \{\mathbf{P}, \boldsymbol{PAM}_{250}\}$ *is an irreducible, aperiodic and stochastic matrix. Therefore, the eigenvalues of* $\mathbf{Q}$ *can be ordered by* $1 > |\lambda_2| \geq \cdots \geq |\lambda_t|$ *. As* $k \to \infty$*,* $\mathbf{Q}^k = \mathbf{Q}_\infty + \mathsf{O}(k^{m_2-1}|\lambda_2|^k)$*, elementwise, where* $m_2$ *is the algebraic multiplicity of* $\lambda_2$ *and* $\mathbf{Q}_\infty$ *is the matrix whose rows are equal to the limiting distribution (109, Theorem 1.2). Thus the convergence is geometric with rate* $|\lambda_2|$*. For* $\boldsymbol{PAM}_{250}$*, we numerically compute* $|\lambda_2| = 0.53$*. However, Finding the eigenvalues of* $\mathbf{P}$*, other than 1, amounts to analytically finding the roots of a polynomial of degree 19. Since there is no algebraic way to find the roots of such a polynomial, the following inequality, due to Deutsch & Zenger, gives an upper bound for* $\lambda_2$ *(31):*

$$|\lambda_2| \leq \frac{1}{2} \max_{i,j} \{p_{i,i} + p_{j,j} - p_{i,j} - p_{j,i} + \sum_{\substack{k \\ k \neq i,j}} |p_{i,k} - p_{j,k}|\}. \tag{4.12}$$

*Applying Equation 4.12 to the probability transition matrix* $\mathbf{P}$*, in Figure 15, leads to* $|\lambda_2| \leq 1 - \frac{1}{2}\alpha$*.*

**Proof 3 (Proof of Theorem 3)** *Denote by* $\min^+ I$ *the minimum of the strictly positive elements of the set* $I$*. Theorem 3 follows from (109, Theorem 4.10), which states that if the* *sequence* $\mathbf{T}_{p,k}$ *is scrambling, for all* $k \geq 1$*, and* $\min^{+}_{i,j} q(k)_{i,j} \geq \gamma > 0$ *uniformly for all* $k \geq 1$*, then weak ergodicity obtains at a uniform geometric rate for all* $p \geq 1$*. Let*

$$\gamma = \min_{1\leq k\leq N} \{ \overset{+}{\min_{i,j}} \ \boldsymbol{PAM}(k)_{i,j} \}.$$

*Then we have* $\min^{+}_{i,j} \boldsymbol{PAM}(k)_{i,j} \geq \gamma > 0$ *uniformly for all* $k \geq 1$*. Observe that the main assumption in Theorem 3 is the finite number of PAM matrices. From the proof of (109, Theorem 4.10), it follows that the convergence rate is geometric with parameter* $(1-\gamma^t)^{\frac{1}{t}}$*.*

**Proof 4 (Proof of Theorem 4)** *From the probability transition matrix* $\mathbf{P}(k)$*, depicted in Figure 15, we have*

$$\overset{+}{\min_{i,j}} p_{i,j}(k) = \min\{1-9\alpha(k), \frac{1}{6}\alpha(k)\} \tag{4.13}$$

*From the boundedness of the mutation rate* $\alpha(k)$ *,* $(0 < a \leq \alpha(k) \leq b < 1)$*, we obtain*

$$\overset{+}{\min_{i,j}} p_{i,j}(k) \geq \min\{\frac{a}{6}, 1-9b\} = \gamma, \tag{4.14}$$

*uniformly on* $k$*. Let* $\mathbf{e}_k$ *be the unique stationary distribution of* $\mathbf{P}(k)$*. We have,* $\mathbf{e}_k = \mathbf{s_1}$*, in Equation 4.9, for all* $k \geq 1$*. In particular, the sequence of vectors* $\{\mathbf{e}_k\}_{k\geq 1}$ *converges to* $\mathbf{s_1}$*. Since* $\mathbf{T}_{p,k}$ *have no zero column, the strong ergodicity property follows from (109, Theorem 4.15). The rate of convergence follows from (109, Theorem 4.10).*

# CHAPTER 5

# GENOMIC STRUCTURE

*"It is through science that we prove, but through intuition that we discover."*

Jules Henri Poincaré.

## Abstract

*In this Chapter, we prove that the introns play the role of a decoy in absorbing mutations in the same way hollow uninhabited structures are used by the military to protect important installations. Our approach is based on a probability of error analysis, where errors are mutations which occur in the exon sequences. We derive the optimal exon length distribution, which minimizes the probability of error in the genome. Furthermore, to understand how can Nature generate the optimal distribution, we propose a diffusive random walk model for exon generation throughout evolution. This model results in an alpha stable exon length distribution, which is asymptotically equivalent to the optimal distribution. Experimental results show that both distributions accurately fit the real data. Given that introns also drive biological evolution by increasing the rate of unequal crossover between genes, we conclude that the role of introns is to maintain a genius balance between stability and adaptability in eukaryotic genomes.*

## 5.1 Introduction

The unexpected discovery of the intron-exon structure of eukaryotic genomes in 1977 struck the molecular biology community (121). The genes of eukaryotic genomes contain protein-coding sequences, called *exons*, separated by non-coding sequences, called *introns*. Thus, introns are excluded from the main gene function: making proteins (see Section 2.3.2). What is more intriguing is that introns make up a large portion of eukaryotic DNA. In humans, for example, approximately 30% of the human genome is made up of introns (12). Only about 3% consists of coding DNA and the rest of the genome consists of other non-coding DNA, repetitive segments and regulatory regions. Questions and speculations about the evolutionary origins and function of introns appeared immediately after their discovery. Finding the role of introns is critical in understanding the function and evolution of genomes. More than 25 years later, the subject is still an active area of research and the same question remains: "What function(s), if any, did introns have?". The extra energy needed to maintain and process the introns, throughout evolution, seems to defy evolutionary logic; "The cell puts a huge amount of its energy into the creation of these introns, then discards them ... Nature would not go to all that trouble without a reason" (64). In our work we consider only the main type of introns, known as 'spliceosomal'. This type consists of all introns found in nuclear protein-coding genes transcribed by RNA polymerase II. Two other types of introns, group I and group II, are very small in number and mainly restricted to the genomes of cellular organelles.

### 5.1.1 The Exon Theory of genes

On year after their discovery, Gilbert (43) advanced that recombination in intronic regions of genes increases the rate of creation of new genes by forming novel combinations of exons. The large amounts of non-coding DNA provides chromosomal regions where recombination between homologous portions of chromosomes can take place without disrupting the function of genes. Such shuffling must have speeded up evolution by accelerating the diversity of proteins and so of living things. The Exon Theory of genes (41) is “a specific statement of the idea that genes were made of small pieces. The crucial elements of that theory are that the very first genes and exons represented small polypeptide chains ≈ 15-20 amino acids long, that the basic method used by evolution to make new genes is to shuffle the exons and that a major trend of evolution was then to loose introns and to fuse small exons together to make complicated exons.” Since introns were used to assemble the first genes, the Exon Theory of genes propound an “introns-early” view. The proponents of the introns-early position were added during evolution to break up previously continuous genes (23). The origin of introns is still the subject of debate between the introns-early camp and the introns-late proponents. In our work, we are interested in the function, if any, of introns and not in their origin, i.e., if they were present during the early life forms of were added later on during evolution. I want to point out, however, that there is a clear distinction between the Exon Theory of genes and exon shuffling, the latter of which has clearly occurred in the recent evolution of some proteins and is frequently exploited in differential splicing, but does not necessarily explain the origin of introns. Known cases of exon shuffling detected in the genes of present-day eukaryotes are reviewed by (72) and many examples of exon shuffling are surveyed by Patty in (88).

### 5.1.2 Introns as potential error detecting sequences

Modern methods of encoding information into digital form include error check digits that are functions of the other information digits. When digital information is transmitted, the values of the error check digits can be computed from the information digits to determine whether the information has been received accurately. These error correcting codes make it possible to detect and correct common errors in transmission. The sequence of bases in DNA is also a digital code consisting of four symbols: A, C, G, and T. Does DNA also contain an error correcting code?

Many authors hypothesized that error-correcting codes are used in the replication process of the genome (10), (11), (34), (48), (69). A consequence of this hypothesis is the existence of redundant DNA. The genes in the DNA are viewed as the encoded messages composed of the information symbols (i.e., exons) and the redundant symbols (i.e., introns) needed by the error-correction process. Forsdyke (34) speculated that error correction codes in genetic sequences might exist in series, where the DNA could be arranged in cells to produce geometrical arrangements of bases which could be read against checking sequences. His model does not require that an exon be checked by contiguous introns. Liebovitch *et al.* (69) developed a procedure to check for the existence of a linear block code in genetic sequences. If a linear block error correcting code is present in DNA then some bases would be a linear function of the other bases in each set of bases. However, their experimental results on the lac operon and the gene for cytochrome c revealed that these two genes do not appear to contain such a simple error correcting code. Battail (10), (11) suggested that genetic information undergoes nested encoding, where the result of a previous encoding process is combined with new information and encoded again. The more important genetic information is assumed to be in the primary coded message. Battail makes a plea for increased research for the purpose of identifying possible nested error-correcting codes. It is well known that DNA replication and protein synthesis involve error repair mechanisms (see Section 2.4). However, no linkage has been found between these repair mechanisms and the intron sequences in the genes. So, either there is no error correction mechanisms encoded in the introns or the genetic error correcting mechanisms are algorithmically different from what has been proposed in the literature so far.

### 5.1.3 Introns stabilize an unstable genome

Using a branch of mathematics called sequence algebra, Huen (53) mathematically showed that introns stabilize an unstable genome without ever bringing it to absolute equilibrium. He argues that all genomes have a degree of instability, for otherwise, life would be impossible. Absolute equilibrium spells stagnation and ultimately death. Introns and junk DNA help stabilize the unstable genome by a mechanism of negative feedback, which is imperfect and hence can never achieve full equilibrium. The equilibrium state, in Huen's mathematical model, corresponds to perfect (errorless) transmission of information in our protein communication channel. A closely related idea is the one proposed by Doolittle (32) and further extended by Matsuo and coworkers (74). Doolittle suggested that the principal selective value of introns concerns the prevention of pairing between the sequences of duplicate genes. By preventing pairing, potential recombination events would be discouraged. Since duplicate genes have similar sequences, and recombination can occur when sequences are similar, then the introduction of introns that are more able to accept mutations (i.e. to diversify), would serve to preserve the duplicate genes (i.e. prevent them blending by recombination). The common principal idea in Huen's stability calculations and Doolittle's genetic maintenance is that introns act as a buffer for mutations.

## 5.2 Proposed Approach

We propose that introns maintain a genius balance between stability and adaptability in eukaryotic genomic sequences as follows:

- Introns reduce the probability of mutation error in the coding regions (i.e., exons) by serving as decoys which absorb isolated mutations. According to this view, introns protect coding regions in the DNA sequence from frequent errors in the same way hollow uninhabited structures are used by the military to protect important installations, such as aircraft hangars and missile launching facilities, from a bomb attack by serving as a 'dummy' target that resembles the protected structure. It is important to emphasize that the role of introns is not to ensure a perfect (errorless) communication system, but to temper the effervescence of the ever-changing genome under the chemical, physical and environmental conditions. Perfect information transmission will spell stagnation and ultimately extinction. This is the major difference between an engineering communication system and the biological communication system.
- Introns drive biological evolution by increasing the rate of recombination of exons and consequently participate in the creation of new genes. This process results in the introduction of new exons into the genomic sequence and thereby is responsible for its rapid evolution. The role of introns in driving evolution by increasing the rate of recombination of exons is inspired by Gilbert's exon shuffling hypothesis. However, unlike Gilbert, we do not necessarily claim that exons represent functionally and/or structurally important subunits of proteins nor do we adopt his intron-early view. It is clear that a process of

> genomic recombination, leading to gene rearrangements and the assembly of genes from 'pieces' played a crucial role in gene evolution. However, the details of these genomic recombinations at the time of the origin of life are unclear (was it 'classical' exon-shuffling or another type of nucleic acids rearrangements). All we claim is that the long sequences of introns make them hot spots for genetic recombination via unequal crossover. In other words, we view introns as analogous to shelters in a community or country which ignores the presence of isolated individuals who may come and go as they please; however, once an entire family has moved into the shelter, it is immediately absorbed into the community and granted citizenship by the host country.

The proposed dual role of introns serves to provide a balance between two competing biological evolutionary functions: stability and adaptability. We point out that the role of introns in increasing the rate of unequal crossovers must be tempered in order to prevent excessive evolutionary adaptability. Rapid changes in genomic sequences must not occur too frequently, or else we would experience evolutionary jumps in each generation. It is interesting to note that the role of introns in protection against mutations is enhanced by increasing the size of the intron regions (and, as we will prove later, regardless of the location and structure of the introns in the gene). On the other hand, the function of introns in encouraging unequal crossovers depends on the presence of long contiguous nucleotide sequences in introns. In order to moderate the adaptability rate of the genomic sequence, the length of contiguous nucleotide sequences must be limited. Indeed, most eukaryotes display multiple intron regions within a single gene. Introns therefore seem to control the balance between stability and adaptability of the genomic sequence. Our approach to prove this hypothesis relies on a probability of error analysis. Errors are mutations, which occur in the coding sequences of the gene (i.e., exons).

The stability role attributed to introns accounts for at least two biological facts:

- The absence of introns in prokaryotic genomes translates, according to our view, to a high mutability rate of these primitive organisms. It is widely known today that many bacteria and viruses rely on mutations for diversification.
- The decoy role for introns predicts that coding sequences should be more conserved among organisms than non-coding sequences. Studies in comparative genomics showed that functional DNA sequences tend to undergo mutation at a slower rate than nonfunctional sequences (27). For example, the coding sequence of a human protein-coding gene is typically about 80% identical to its mouse ortholog, while their genomes as a whole are much more widely divergent.

Crossover is known to be an important driving force in biological evolution (82). Evidence of the origin of homologous gene clusters suggests that they arose by either unequal crossover or gene conversion events (68). To understand the role of introns in the assembly of new genes, we found no better explanation than Gilbert's words: "Consider a new gene made by a new combination of regions of earlier genes by an unequal crossover, a rare event at the DNA level, that matches small, similar sequences between two DNAs. To make a new protein that contains the first part of one protein with the second part of another requires such a rare, and in frame, event. However, if the regions that encode parts of the protein are separated by 1,000-10,000 base long introns along the DNA, a process of unequal crossing-over occurring anywhere within that intron between the exons will create a new combination of exons." (42). Hence, the presence of introns can speed up the process of evolution. Recently, it has been experimentally proved that introns length are negatively correlated with the rate of recombination in Drosophila melanogaster and humans (26). That is the advantage of longer introns is expected to decrease inversely with the rate of recombination. Hence, in the chromosomal regions where crossing over is infrequent, introns tend to be larger to increase the rate of recombination between exons. Whether introns were used to assemble the first genes or not is not relevant to our investigation as long as we have biological evidence that new genes were and are currently created through the mechanism of unequal crossover. It is important to notice that the crossover does not necessarily lead to a loss of introns. Depending on the locus of crossover and the intron length of both genes to be recombined, the newly created gene might have a longer intron than both of the original genes.

### 5.2.1 The efficiency of the proof-reading mechanisms in the DNA

DNA repair mechanisms are constantly operating in cells. In human cells, both normal metabolic activities and environmental factors can result in as many as a million molecular lesions per cell each day (see Section 2.4). Consequently, DNA repair mechanisms are essential for the survival of the organism. However, it is also known that these DNA repair mechanisms are not 100% efficient and many errors remain undetected or uncorrected in the genome. Intuitively, nature maintains a balance between keeping the identity of a particular organism by reliable transmission of its protein set and, at the same time, allowing for errors to occur. To see this, let us anecdotically compare the efficiency of the Reed-Solomon code and the biological error correction mechanisms. The potential efficiency of a code is a function of the number of redundant bits. A commonly used Reed-Solomon code, in CD players for instance, uses a codeword length of 255 bytes, of which 223 bytes are data and 32 bytes are parity. The redundancy rate of the Reed-Solomon code is then $\frac{32}{255} = 13\%$. The human genome contains about 30,000 genes, of which about 130 code for DNA repair enzymes (98). Assuming that the genes have roughly the same number of nucleotides, the redundancy rate of the human error correction mechanism is $\frac{130}{30{,}000} = 0.43\%$! Here, we refer to redundancy in the genome only when it is used in regulatory genes responsible for DNA repair and ignore the presence of non-coding genes. Hence, despite all the excitement that the discovery of DNA repair mechanisms brought (especially to creationists), this simple argument indicates that the repair mechanism of the human genome is unlikely to be very efficient. We argue that, in order to achieve a lower error rate, nature introduces introns to temper the effervescence of the ever-changing genome. One can ask: Why wouldn't nature invest in more error correction mechanisms rather than carrying this enormous decoy luggage? Several reasons lie behind this choice: First, if nature had to design error correction codes to control the exact rate of mutation required to simultaneously maintain life and encourage evolution, it would need to know the exact distribution and form of all possible mutations which occurred in the past and will occur in the future. Designing complex error correcting codes for a given noise model might be completely useless in the face of dynamic noise characteristics. Second, a reduction in the error rate comes at the price of an increase in complexity. Nature might have preferred to spend more energy in carrying the decoy sequences rather than investing in complex and costly error repair enzymes.

## 5.3 Genomic Structure: Deterministic Analysis

### 5.3.1 A Poisson mutation model

**Proposition 5** *Consider a genome of length $T$. Assume that the point mutation rate is randomly distributed in the genome, i.e., the occurrence of mutations is independent and identically distributed in all regions of the genome. Then, the probability of error is a decreasing function of the length of introns and is independent of the distribution of introns in the genome.*

Hence, we see that a binomial error model does not account for the biological exon (or intron) length distribution inside the genome. In other words, the biological intron-exon distribution would be equivalent, from an error robustness criterion, to the distribution which groups all exons in the beginning of the gene and all introns at its end. Therefore, we need to consider a different mutation model, which can account for the observed intron-exon structure in eukaryotic genomes. We propose a Poisson mutation model. This choice is justified by numerous arguments. First, the Poisson distribution is the limiting distribution of the binomial when the probability of error is small and the genome size is large such that the rate of point mutation in a unit interval is held constant (De Moivre-Laplace theorem (87)). Second, many rare random phenomena in nature follow a Poisson distribution, e.g., the number of winning tickets in a large lottery, the number of printing errors in a book, etc. In the remainder of this paper, we assume that the mutations are Poisson distributed in the genome.

### 5.3.2 Error robustness analysis

Assume that there are $K$ exons of total length $M$ in a gene of $T$ nucleotides. Let $l_k$ be the length of exon $k$. In this subsection, we answer the question: "What are the optimal exon lengths, $l_k^*,\ k = 1, \cdots, K$, which minimize the probability of error in the gene?".

**Proposition 6** *Assume that the mutations are Poisson distributed with rate $\lambda$. Consider a genome of length $T$ nucleotides including $K$ exons having total length $M$. Let $l_k$ be the length of the $k^{th}$ exon. Then, the probability of error is given by*

$$P_e = 1 - e^{-\lambda KT} \prod_{k=1}^{K} \sum_{n=0}^{T-l_k} \frac{\lambda^n (T - l_k)^n}{n!}. \tag{5.1}$$

Since $l_k \leq M$ for all $k = 1, \cdots, K$, we obtain an upper bound on the probability of error by truncating the summation in Equation 5.1 to $T - M$ instead of $T - l_k$. Minimizing the maximum probability of error, $P_e^{\max}$, is more tractable analytically than minimizing the probability of error in Equation 5.1. Using the Lagrange multiplier technique, with constraint $\sum_{k=1}^{K} l_k = M$, and taking the derivative of $P_e^{\max}$ with respect to $l_k$, we obtain the following coupled system for the optimal exon lengths:

$$l_{i_0} = M \frac{[\prod_{k \neq i_0} \sum_{n=0}^{T-M} \frac{\lambda^n (T-l_k)^n}{n!}][\frac{\sum_{n=1}^{T-M} \lambda^n (T-l_{i_0})^{n-1}}{(n-1)!}]}{\sum_{j=1}^{K} [\prod_{k \neq j} \sum_{n=0}^{T-M} \frac{\lambda^n (T-l_k)^n}{n!}][\frac{\sum_{n=1}^{T-M} \lambda^n (T-l_j)^{n-1}}{(n-1)!}]}. \tag{5.2}$$

An obvious solution to the system in Equation 5.2 is obtained when $l_k^* = \frac{M}{K}$ for all $k = 1, \cdots, K$. This surprising simple result states that the optimal exon lengths are distributed according to a delta function centered at the mean value $\frac{M}{K}$. But, in nature, the exon lengths are not uniformly distributed in the genome (see Figure 17). The reason this deterministic analysis fails in capturing the intron-exon distribution is that the genome is not a deterministic entity but rather a continuously evolving one. Therefore, a stochastic model for the exon lengths would be more appropriate to correctly describe the genome's dynamic nature. The deterministic analysis does, however, capture some characteristics of the biological data in the following sense:

**Proposition 7** *Let $\delta_{\frac{M}{K}}$ be the delta function centered at $\frac{M}{K}$. For every $\rho > 0$, consider the measure $d_\rho$ between a continuous unimodal probability density function $f_X$ and $\delta_{\frac{M}{K}}$ given by*

$$d_\rho(\delta_{\frac{M}{K}}, f_X) = 1 - Pr(X \in [\frac{M}{K} - \rho, \frac{M}{K} + \rho]). \tag{5.3}$$

*Let $x_0$ be the mode of $f_X$. Then, $argmin_{x_0} d_\rho = \frac{M}{K}$. That is the mode of $f_X$, which minimizes the measure $d_\rho$, is equal to $\frac{M}{K}$.*

The biological exon distribution is asymmetric given that its support is $[0, \infty]$. The mode of asymmetric distributions is always less or equal than their mean. From proposition 7, the distribution, which best approximates $\delta_{\frac{M}{K}}$ in the $d_\rho$ measure sense, would have its mode very close to its mean. Amazingly, the exon length distribution of the human genome has its mode almost equal to its mean obtained at about 170 nucleotides (see Figure 17)!

Even though the deterministic analysis gave some insights on the optimality of the biological exon length distribution from an error minimization criterion, a stochastic model for the exon distribution is needed to capture the dynamics of the evolving genome.

## 5.4 Genomic Structure: Stochastic Analysis

### 5.4.1 Error Robustness Analysis

In this subsection, we readdress the probability of error optimization problem formulated in Section 5.3 assuming a stochastic distribution of the exon lengths. The following proposition establishes the new expression for the probability of error assuming an infinite genome length, i.e., $T = \infty$.

**Proposition 8** *Let $p(l)$ be the continuous distribution of the length of exons. Assume that there are $K$ exons in a genome infinitely long. The mutations are assumed to be Poisson with parameter $\lambda$. Then the probability of error is given by*

$$P_e = 1 - (\int_0^\infty e^{-\lambda l} p(l)\ dl)^K. \tag{5.4}$$

We want to determine the optimal exon length distribution, $p^*(l)$, which minimizes the probability of error subject to $\int_0^\infty p^*(l)\ dl = 1$. It can be easily shown that the delta function centered at 0, $\delta_0$, satisfies this optimization problem. This solution is somehow intuitive: no exons implies no error! In order to get a meaningful solution to this optimization problem, we need to impose more constraints on the exon length distribution. For instance, the mean exon length should be larger than a pre-specified number $l_0$ or, in general, the $\alpha^{\text{th}}$ moment of $p(l)$

should be larger than $l_0$. Consequently, the stochastic optimization problem is reformulated as follows:

$$p^*(l) = \underset{p(l)}{\operatorname{argmax}} \int_0^\infty e^{-\lambda l} p(l)\ dl, \tag{5.5}$$

subject to

$$1) \int_0^\infty p(l)\ dl = 1;$$

$$2) \int_0^\infty l^{1+\alpha} p(l)\ dl \geq l_0, \text{ for some } \alpha \geq 0.$$

The optimization problem formulated in Equation 5.5 is solved using the Euler-Lagrange equation. We obtain:

$$p^*(l) = \frac{p_0(1+\mu)}{e^{-\lambda l} + \gamma l^{1+\alpha} + \mu}, \tag{5.6}$$

where $\mu$ and $\gamma$ are the Lagrange multipliers, which are determined numerically. Taking the derivative of $p^*$, it is easy to show that it has a unique maximum. Observe that the $(1+\alpha)^{\text{th}}$ moment of $p(l)$ is infinite; thus satisfying condition 2) in Equation 5.5. This infinite moment agrees with the heavy tail characteristic of the biological exon length distribution (see Figure 17). The parameter $\alpha$ determines the tail decay of the distribution for a given mutation rate $\lambda$.

At this point, it is interesting to ask ourselves: "How can Nature generate such a distribution? Is there a simple enough model for exon generation, which leads to the distribution $p^*$?" The answer is investigated in the next subsection.

### 5.4.2 A Random Walk model

Insertion and deletion of exon nucleotides have been confirmed biologically for many primitive organisms. If, during evolution, exons were formed by insertion and deletion mechanisms, their lengths would follow some kind of a random walk. The length of the exon at any time corresponds to the position of the random walk. We assume that the sub-exons are formed independently by a stochastic process according to a distribution $f(l)$. So, the length of the final exon after $N$ steps, $X_N$, is the sum of $N$ independent displacements distributed according to $f(l)$, i.e., $X_N = \sum_{i=1}^{N} l_i$. Given the heavy tail characteristic of the biological exon length distribution, we assume that the sub-exons are generated by a distribution of the form:

$$f(l) = \alpha \; l^{-(\alpha+1)}, \quad l \geq 1. \tag{5.7}$$

where $0 < \alpha < 2$. We want to determine the limiting distribution of $X_N$ as $N \to \infty$ or as the time $t \to \infty$. By the Generalized Central Limit Theorem (44), the density of $X_N$ tends towards an alpha-stable distribution $S_\alpha(l|\beta, \sigma, \xi)$. Since Paul Levy found the class of $\alpha$-stable distributions, in 1925, as simple exceptions to the Central Limit Theorem, a vast amount of knowledge has been accumulated about the properties of these probability distributions especially infinite moments, elegant scaling properties and the inherent self-similarity property. They have been found to provide useful models in the study of physical and economic systems, especially phenomena with large fluctuations and high variability that are not compatible with the Gaussian models. Except the Gaussian , the Cauchy and the Levy distributions which are special cases of the stable class, there is no exact expression of the probability density function of an $\alpha$-stable distribution. $\alpha$-stable distributions are defined by their characteristic function. Four parameters are needed: an index of stability $\alpha \in (0, 2]$, a skewness parameter $\beta \in [-1, 1]$, a scale parameter $\sigma > 0$ and a location parameter $\xi \in (-\infty, +\infty)$. There are multiple parameterizations of $\alpha$-stable distributions. For numerical purpose, we will use a variant of the M-parametrization of Zolotarev (124) with the following characteristic function (81):

$$\exp^{\imath\omega X} = \begin{cases} \exp^{-\sigma^{\alpha}|\omega^{\alpha}|[1+\imath\beta\tan(\frac{\alpha\pi}{2})sign(\omega)((\sigma|\omega|)^{1-\alpha}-1))+\imath\xi\omega]}, & \text{if } \alpha \neq 1; \\ \exp^{\sigma|\omega|[1+\imath\beta\frac{2}{\pi}sign(\omega)\ln(\sigma|\omega|)]+\imath\xi\omega}, & \text{if } \alpha = 1. \end{cases} \tag{5.8}$$

The above parametrization is a scale and location family of distributions: if $Y \sim S_{\alpha}(\sigma, \beta, \xi)$, then for any $a, b, aY + b \sim S_{\alpha}(|a|\sigma, (sign(a)\beta, a\xi + b)$. Other related issues of stable distributions are discussed in (81). Some of the prominent properties of $\alpha$-stable distributions are: heavy tail, skewness (when $\beta \neq 0$), and smooth unimodal density. Their asymptotic behavior is described by:

$$\lim_{|x|\to\infty} S_{\alpha}(x|\beta, \sigma, \xi) = \frac{C}{|x|^{1+\alpha}}, \tag{5.9}$$

where $C$ is some constant (44). Hence, from Equation 5.6, we see that the optimal distribution $p^*$ is asymptotically equivalent to an alpha-stable distribution. Nature would prefer to generate a simple random walk rather than solve the Euler-Lagrange equation!

## 5.5 Experimental Results

All exon lengths for each of the Homo sapiens (Human), Rattus Norvegicus (Rat), Mus Musculus (mouse), Apis Mellifera (Honey bee), Schizosaccharomyces Pombe (fission yeast), Plasmodium Falciparum (malarial parasite) and Arabidopsis Thaliana (thale cress) genomes were studied. The data files used were obtained from the NCBI web site: `ftp://ftp.ncbi.nih.gov/genomes`. Exons tagged as CDS were included in the analysis. The NCBI handbook makes clear that CDS refers to the portion of a genomic DNA sequence that is translated. Alternative spliced variants were kept in the data, so some exons can be recorded several times from a given gene.

An initial data analysis is presented in Table VI. Of the seven different organisms examined, H. sapiens contained the greatest number of exons, 281,975. S. pombe has the least number of exons of the organisms analyzed here. The descriptive statistics for H. Sapiens, M. Musculus, R. Norvegicus, A. Mellifera and A. Thaliana are similar. The two single cellular organisms, S. Pombe and P. Falciparum, have considerably higher average exon lengths as well as greater exon length variation than all the other organisms. For all the organisms the mean exon length is greater than the median exon length, indicating a right-skewed distribution. Figure 17 shows the biological data, the optimal density and the alpha-stable distribution of the analyzed organisms. For alpha-stable density fitting, we used the Mathematica package for stable distributions available from J. P. Nolan's website: `academic2.american.edu/~jpnolan`. The parameter $\alpha$ was estimated by plotting the data on a log-log scale and estimating the slope: If we order the data $X(1) \geq X(2) \geq \cdots \geq X(n)$ (the order statistics of the empirical data) then we can estimate $y = P(X > t)$ by taking $y = \frac{i}{n}$ and $t = X(i)$. A plot of the points $(t, y) = (\ln(X(i)), \ln(\frac{i}{n}))$ should fit a straight line with slope $-\alpha$. The stable distributions $S_{1.5}(l|0.9, 35, 135)$, $S_{1.5}(l|0.85, 35, 140)$, $S_{1.5}(l|0.9, 60, 190)$, $S_{1.5}(l|0.9, 35, 143)$,

TABLE III. Descriptive statistics of exon lengths for the seven organisms

| | Nb exons | Mean | Stdev | Min | Max |
|---|---|---|---|---|---|
| Homo Sapiens | 281975 | 167 | 233 | 1 | 17105 |
| R norvegicus | 185769 | 177 | 378 | 1 | 9820 |
| M musculus | 226498 | 178 | 326 | 1 | 16625 |
| A mellifera | 32753 | 234 | 320 | 1 | 7241 |
| S pombe | 9772 | 698 | 1038 | 1 | 11099 |
| P falciparum | 12660 | 943 | 1957 | 2 | 27815 |
| A thaliana | 164986 | 228 | 722 | 1 | 6040 |

$S_{1.5}(l|0.9, 60, 130)$, $S_{1.5}(l|0.9, 46, 135)$ and $S_1(l|0.85, 45, 332)$ fit the exon length distributions of Homo Sapiens, Rat Norvegicus, Apis Mellifera, M. Musculus, S. Pombe, P. Falciparum and A. Thaliana respectively. The same $\alpha$ was used to display the optimal density $p^*(l)$ for these organisms. The mutation rate $\lambda$ can be interpreted as the average rate of accepted mutations since the beginning of life on earth.

**Appendix**

**Proof 5 (Proof of Proposition 5)** *Write $T = M + S$, where $S$ is the total number of nucleotide introns in the gene. Then, assuming a total of $n \geq 1$ mutations in the gene, the probability of error $P_e$ is given by*

$$P_e(S) = \sum_{k=1}^{n} \binom{n}{k} \frac{M^k S^{n-k}}{(M+S)^n} = 1 - (\frac{S}{M+S})^n. \tag{5.10}$$

*The derivative of $P_e$ with respect to the variable $S$ is*

$$P_e^{'}(S) = -\frac{nMS^{n-1}}{(M+S)^{n+1}} < 0, \textit{ for all } n \geq 1. \tag{5.11}$$

*Hence $P_e$ is a decreasing function of the intron length for all $n \geq 1$. Moreover, Eqs (Equation 5.10) is independent of the intron-exon structure in the gene.*

**Proof 6 (Proof of Proposition 6)** *Let $x_k$ denote the start position of the $k^{th}$ exon in the genome. We have*

$$P_e = 1 - \prod_{k=1}^{K} Pr \ (\text{“}0 \quad error \ in \ exon \ k\text{”}), \tag{5.12}$$

*where*

$$\begin{aligned}
& Pr \ (\text{“}0 \quad error \ in \ exon \ k\text{”}) \\
= \ & \sum_{n=0}^{T-l_k} e^{-\lambda l_k} \ Pr \ (\text{“}n \ errors \ outside \ l_k) \\
= \ & \sum_{n=0}^{T-l_k} e^{-\lambda l_k} \{ \sum_{i=0}^{n} Pr \ (\text{“}i \ errors \ \in [1, x_k - 1]\text{”}) \ Pr \ (\text{“}(n-i) \ errors \ \in [x_k + l_k, T]\text{”}) \} \\
= \ & \sum_{n=0}^{T-l_k} e^{-\lambda l_k} ( \sum_{i=0}^{n} e^{-\lambda(x_k - 1)} \frac{(\lambda(x_k-1))^i}{i!} e^{-\lambda(T - x_k - l_k + 1)} \frac{(\lambda(T - x_k - l_k + 1))^i}{(n-i)!} ) \\
= \ & \sum_{n=0}^{T-l_k} e^{-\lambda T} \sum_{i=0}^{n} \frac{\lambda^n}{n!} \binom{n}{i} (x_k - 1)^i (T - x_k - l_k + 1)^{n-i} \\
= \ & e^{-\lambda T} \sum_{n=0}^{T-l_k} \frac{\lambda^n}{n!} (T - l_k)^n.
\end{aligned} \tag{5.13}$$

*From Equation 5.13 and Equation 5.12, we obtain*

$$P_e = 1 - e^{-\lambda K T} \prod_{k=1}^{K} \sum_{n=0}^{T-l_k} \frac{\lambda^n (T - l_k)^n}{n!}. \tag{5.14}$$

**Proof 7 (Proof of Proposition 7)** *Let $f_X$ be a unimodal density which reaches its mode at $x_0$. Then $f_X(x - x_0)$ reaches its mode at 0. We have*

$$x_0^* \ = \ \underset{x_0}{argmax} \int_{\frac{M}{K} - \rho}^{\frac{M}{K} + \rho} f_X(x - x_0) dx. \tag{5.15}$$

*By continuity of $f_X$, we have*

$$|(x-x_0)-(\frac{M}{K}-x_0)|<\rho \Rightarrow |f_X(x-x_0)-f_X(\frac{M}{K}-x_0)|<\epsilon, \tag{5.16}$$

*for some $\epsilon > 0$. So,*

$$|x-\frac{M}{K}|<\rho \Rightarrow f_X(\frac{M}{K}-x_0)-\epsilon < f_X(x-x_0) < f_X(\frac{M}{K}-x_0)+\epsilon. \tag{5.17}$$

*So,*

$$\underset{x_0}{argmax}\ 2\rho(f_X(\frac{M}{K}-x_0)-\epsilon) \leq x_0^* \leq \underset{x_0}{argmax}\ 2\rho(f_X(\frac{M}{K}-x_0)+\epsilon).$$

*Since $f_X(x-x_0)$ reaches its mode at $0$, we obtain $x_0^* = \frac{M}{K}$.*

**Proof 8 (Proof of Proposition 8)**

$$\begin{aligned}
P_e &= 1-\prod_{k=1}^{K} Pr(\text{“}\mathit{0\ error\ in\ exon\ k}\text{”}) \\
&= 1-\prod_{k=1}^{K}\int_0^\infty Pr(\text{“}\mathit{0\ error\ in\ exon\ k|\ its\ length\ is\ l}\text{”})p(l)dl \\
&= 1-\prod_{k=1}^{K}\int_0^\infty e^{-\lambda l}p(l)\ dl \\
&= 1-(\int_0^\infty e^{-\lambda l}p(l)\ dl)^K.
\end{aligned}$$

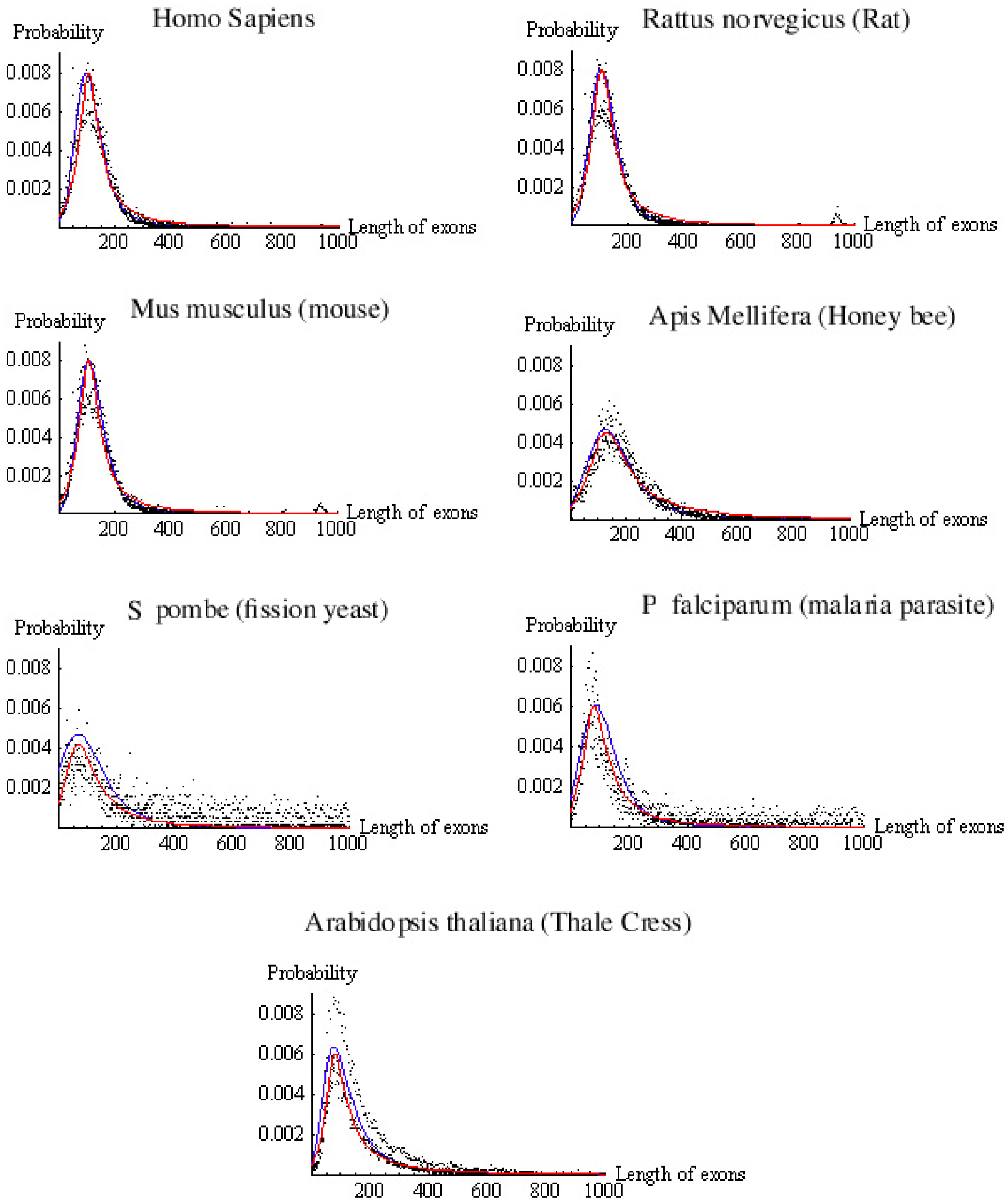


Figure 17. Exon length distribution: The data points represent the biological data; the red curve is the optimal density, which minimizes the probability of error; and the blue curve is the fitted alpha-stable distribution. The graphs of the densities are truncated at exon lengths of 1000 nucleotides.

# CHAPTER 6

# INFORMATION-THEORETIC BOUNDS OF EVOLUTIONARY PROCESSES

*"Come, fill the Cup, and in the fire of Spring*
*Your Winter-garment of Repentance fling*
*The Bird of Time has but a little way*
*To flutter–and the Bird is on the Wing*
*Whether at Naishápúr or Babylon,*
*Whether the Cup with sweet or bitter run,*
*The Wine of Life keeps oozing drop by drop,*
*The Leaves of Life keep falling one by one. "*

Omar Khayyam, The Rubaiyat.

## Abstract

*In this Chapter, we investigate the information theoretic bounds of the channel of evolution. We compute the capacity and the rate-distortion functions of the protein communication system for the three domains of life: Achaea, Bacteria and Eukaryotes. We analyze the tradeoff between the transmission rate and the distortion in noisy protein communication channels. As expected, comparison of the optimal transmission rate with the channel capacity indicates that the biological fidelity does not reach the Shannon optimal distortion. However, the relationship between the channel capacity and rate*

*distortion achieved for different biological domains provides tremendous insight into the dynamics of the evolutionary processes. We rely on these results to provide a model of protein sequence evolution based on the two major evolutionary processes: mutations and unequal crossovers.*

## 6.1 Introduction

The protein communication system, proposed in Chapter 3 (see Figure 12), is a communication model of the genetic information storage and transmission apparatus. The genome is viewed as the joint source-channel encoded message of the protein communication system and hence can be investigated in the context of engineering communication codes. In particular, it is legitimate to ask at what rate can the genomic information be transmitted? And what is the average distortion between the transmitted message and the received message at this rate? Shannon's channel capacity theorem (110) states that, by properly encoding the source, a communication system can transmit information at a rate that is as close to the channel capacity as one desires with an arbitrarily small transmission error. This goal is sought, in communication engineering, by incorporating redundancy using an error correcting code. Conversely, it is not possible to reliably transmit at a rate greater than the channel capacity. Therefore, an information source whose transmission rate matches the capacity rate takes advantage of the channel to the fullest. The theorem, however, is not constructive and does not provide any help in designing such codes. In the case of biological communication systems, however, evolution has already designed the code for us. The encoded message is the DNA sequence. Comparison of the genomic transmission rate with the channel capacity will reveal whether the genomic code is efficient from an information theoretic perspective. However, even if the channel capacity is not exceeded, we are assured that biological communication systems do not rely on codes that produce negligible errors since the level of distortion present must account for evolutionary processes. It is, therefore, interesting to ask ourselves whether biological communication systems maintain an optimal balance between the transmission rate and the desired distortion levels needed to support adaptive evolution.

An extension of the early work in information theory, called rate-distortion theory (14), has been developed to analyze the optimal tradeoff between the transmission rate and distortion in noisy communication channels. In this theory, one can associate with a source-destination pair a function $R(D)$, called the rate distortion function, which has the following significance: A communication system can be designed that achieves fidelity $D$, or equivalently that produces the source output at the destination with an average distortion of $D$ or less, if and only if the capacity of the channel that connects the source to the destination exceeds $R(D)$. The rate at which a source produces information subject to a requirement of perfect reproduction at the destination is the Shannon entropy of the source, i.e., $R(0) = H$. Given the fidelity $D$ present in biological communication systems, comparison of the rate $R(D)$ with the channel capacity $C$ can be used to determine whether or not a code exists that can achieve the optimal rate distortion criteria. Moreover, by comparing the transmission rate and distortion of biological communication systems, we can verify if the genomic code has realized the optimal rate-distortion criteria. Furthermore, it will be interesting to compare the channel capacity and rate distortion functions of single and two source protein communication systems modelling, respectively, asexual and sexual reproduction.

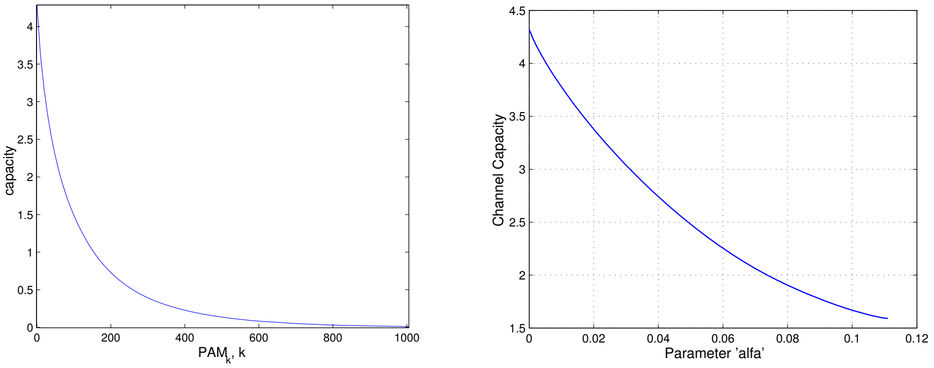
### 6.2 Protein Channel Capacity

In this Chapter, we use the probability transition matrices $\mathbf{Q} = \{q_{i,j}\}_{1\leq i,j\leq 20}$, of the protein communication channel, introduced in Section 4.2, where $\mathbf{Q} \in \{\mathbf{PAM_{250}}, \mathbf{P}\}$. Recall that $\mathbf{PAM_{250}}$ is Dayhoff's Point Accepted Mutation (PAM) matrix at the evolutionary distance 250, and $\mathbf{P}$ is a first-order Markov transition probability matrix constructed from the genetic code using a time-dependent point mutation rate $\alpha(k)$.

The capacity of a channel is the maximum rate at which information can be reliably conveyed by the channel. It is defined as

$$C = \max_{p\in P^n} I(p, \mathbf{Q}) = \max_{p\in P^n} \Sigma_j \Sigma_k p_j q_{jk} \log \frac{q_{jk}}{\Sigma_k p_j q_{jk}}, \tag{6.1}$$

where $P^n$ is the set of all probability distributions on the channel input and $\mathbf{Q}$ is the probability transition matrix of the channel. Evaluation of the channel capacity involves solution of a convex programming problem. In most cases, analytic solutions cannot be found. However, we rely on Blahut's iterative algorithm (16) to compute the channel capacity. Figure 18(a) (resp., Figure 18(b)) shows the capacity of the protein communication system as a function of the evolutionary distance of PAM matrices (resp., point mutation rate $\alpha$). As expected, the channel capacity decreases to zero as the evolutionary distance or the point mutation rate $\alpha$ increases. This result has different ramifications on bioinformatics than on communication engineering: In engineering, it is interpreted as a loss of information after an infinite number of transmissions. The reason is that, in communications, only the initial message is used to convey information and not the channel. In bioinformatics, on the other hand, the output message



(a) (b)

Figure 18. Channel Capacity: (a) Channel capacity v.s. the evolutionary distance of PAM matrices; (b) Channel capacity v.s. the point mutation rate $\alpha$

captures the information of the channel (i.e. the mutations) regardless of the initial message. In particular, a parent organism cannot transmit reliably (channel capacity zero) its genetic information to its offspring of many generations no matter how small the point mutation rate is as long as it is not zero. It is interesting to observe that organisms with lower mutation rates have higher channel capacity, and therefore their genetic information can be reliably transmitted at a higher rate.

Having computed the capacity of the protein communication channel, a comparison of the genomic transmission rate with the channel capacity will reveal whether the genomic code is efficient from an information theoretic perspective.

## 6.3 Protein Rate Distortion

The rate distortion function, $R(D)$, is the effective rate at which the source produces information subject to the constraint that the receiver can tolerate an average distortion $D$. A distortion matrix with elements $\rho_{i,j}$ specifies the distortion associated with reproducing the $i^{\text{th}}$ source letter by the $j^{\text{th}}$ reproducing letter. The rate-distortion function is defined as

$$R(D) = \min_{Q \in Q_D} I(p, Q) = \min_{Q \in Q_D} \Sigma_j \Sigma_k p_j Q_{jk} \log \frac{Q_{jk}}{\Sigma_k p_j Q_{jk}}, \qquad (6.2)$$

where $Q_D = \{Q \in \mathbb{R}^n \times \mathbb{R}^n : \Sigma_k Q_{jk} = 1, Q_{jk} \geq 0, d(Q) \leq D\}$, $d(Q) = \Sigma_j \Sigma_k p_j Q_{jk} \rho_{jk}$, and $p = \{p_j\}$ is the probability vector of the channel input.

We define the distortion between a pair of amino acids as their distance in the Principal Component Analysis (PCA) plane obtained from 7 physico-chemical properties (volume, bulkiness, polarity, PH index, hydrophobicity and surface area). The amino acid data was obtained from [Chapter 2] (51). The result of PCA analysis is shown in Figure 19.

Figure 6.3 shows the rate-distortion curves for Archaea, Bacteria and Eukaryotes, where the amino acid probability distributions of Archaea, Bacteria and Eukaryote were obtained in (17). The *R*-*D* curves of the three branches of life reveal two distinct regions: a low distortion region ($0 \leq D \leq 1.4$) and a high distortion region ($1.4 \leq D \leq 7.5$). In the low-distortion region, the R-D curve of Eukaryotes is the highest followed by Bacteria, then Archaea, i.e., we have

$$R(D)_{Ar} < R(D)_{Ba} < R(D)_{Eu}, \quad \forall\ D < 1.4, \qquad (6.3)$$

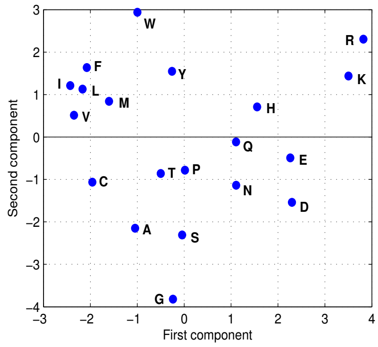
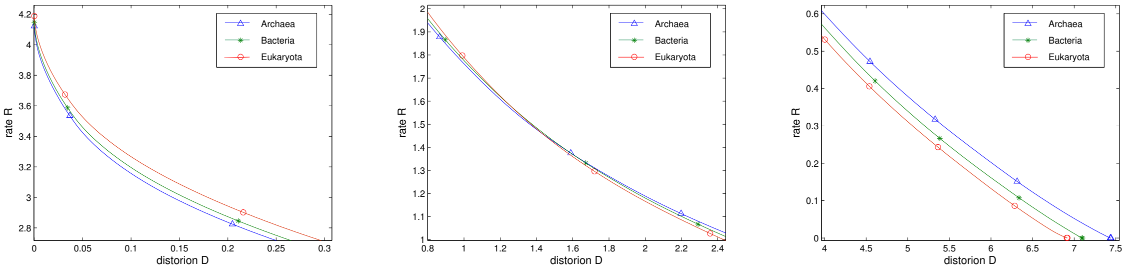


Figure 19. Plot of the amino acids on the first two components of the PCA analysis. The amino acids are labelled by their one-letter standard abbreviations (see Figure 6).



(a) (b) (c)

Figure 20. Rate-distortion curves for Archaea, Bacteria and Eukaryotes: (a) low distortion region; (b) intersection region; (c) high distortion region.

where $R(D)_{Ar}, R(D)_{Ba}$ and $R(D)_{Eu}$ denote the rate-distortion curves of Archaea, Bacteria and Eukaryotes, respectively. At about $D \approx 1.4$, the above order switches to

$$R(D)_{Eu} < R(D)_{Ba} < R(D)_{Ar}, \quad \forall\ 1.4 < D < 7.5. \tag{6.4}$$

The distortion can be associated with the evolutionary distance. That is a low distortion region would correspond to small evolutionary distances, whereas the high distortion region corresponds to larger evolutionary distances. It is quite interesting to observe that for small evolutionary distances (or at the beginning of life), Archaea was the most efficient organism from an information theoretic perspective, followed by Bacteria then Eukaryotes. Specifically, given a fixed transmission rate (of the genetic information), Archaea would have the least distortion. At about $D \approx 1.4$, the three $R$-$D$ curves intersect and reverse orders. So, for longer evolutionary distances, Eukaryotes maintain the most biological fidelity among the three domains.

The actual average distortion over the protein communication channel is defined as

$$D = \textstyle\sum_j \sum_k p_j q_{jk} \rho_{jk}, \tag{6.5}$$

where $\mathbf{Q} = \{q_{i,j}\}$ is the probability transition matrix of the channel, $p = \{p_j\}$ is the distribution of the channel input and $\rho_{i,j}$ is the distortion between amino acids $i$ and $j$. By trial and error Dayhoff et al. (30) found that the matrix $\mathbf{PAM_{250}}$ works well for scoring of actual protein sequences. At this evolutionary distance (250 substitutions per hundred residues) only one amino acid in five remains unchanged. Table VI displays the actual average distortion for Archaea,

TABLE IV. Average distortion for the three domains of life

| | Archaea | Bacteria | Eukaryotes |
|---|---|---|---|
| Distortion | 9.1491 | 8.9964 | 8.8979 |

Bacteria and Eukaryotes, where $\mathbf{PAM_{250}}$ was used as the probability transition matrix of the channel. Observe that the biological rate-distortion values $R(D)$, corresponding to the average distortions given in Table VI, are less than the Shannon channel capacity ($C = 0.8197 > R(D)$). So, from the rate-distortion theory, we can ascertain that the genetic information is encoded so that the system reproduces the initial input with fidelity $D$. In particular, the biological communication system does not rely on codes that produce negligible errors since the level of distortion present must account for evolutionary processes.

### 6.3.1 Evolutionary Model: Amino Acid Distribution

It is well known from information theory that the Gaussian input maximizes the mutual information in an additive Gaussian noise, i.e.,

$$I(X; X + Z^*) \leq I(X^*; X^* + Z^*), \tag{6.6}$$

where $I(a, b)$ is the mutual information between input $a$ and output $b$, $X$ is the input, $Z$ is the channel noise and $^*$ denotes Gaussianity. We will show that the amino acid distribution in Eukaryotes is "more Gaussian" than Bacteria and Archaea. Since a distribution is uniquely characterized by the set of its moments and given that the odd moments of the Gaussian

TABLE V. Scaled norm of odd moments

| | Archaea | Bacteria | Eukaryote |
|---|---|---|---|
| Scaled norm of the $3^{rd}, 5^{\text{th}}, 7^{\text{th}}$ and $9^{\text{th}}$ moments | 5.0653 | 73.5401 | 1.0000 |

distribution are identically zero, we compute the odd moments of the amino acid distribution for the three branches of life (17). Table V displays the scaled norm of the first 4 odd moments ($3^{rd}, 5^{\text{th}}, 7^{\text{th}}$ and $9^{\text{th}}$) for Archaea, Bacteria and Eukaryotes. The odd moments norm of Archaea (resp., Bacteria) are 5 (resp., 73) times higher than Eukaryotes, asserting that the amino acid distribution of Eukaryotes is "more Gaussian" than the two other groups of life. This explains the $R$-$D$ curve in the high distortion region (see Figure 6.3(c)).

To explain the low-distortion region in 6.3 and the switching-over of the R-D curves, we have to dig deeper into the evolutionary processes, which shaped the three groups of life.

### 6.3.2 Evolutionary Process: Mutation and Crossover

It is accepted today that the main driving forces of evolution are mutations and unequal crossover [1]. Furthermore, Archaea and Bacteria rely mostly on mutations for adaptability and survival. So, we can fairly postulate that mutations drive the evolution of Archaea and Bacteria whereas unequal crossovers drive the evolution of Eukaryotes. A mutation involves 1 nucleotide or a very short sequence of nucleotides. Therefore, it induces much less modifications to the

[1] Unequal crossover is a crossover between homologous chromosomes that are not perfectly aligned. This can result in a duplication of genes on one chromosome and a deletion of these on the other. Unequal crossover is a rare phenomenon.

genome sequence than any unequal crossover. So at the beginning of evolution, the distortion caused by mutations is small compared to the distortion caused by unequal crossovers. Hence, the rate-distortion curve of Archaea whose main driving force is mutation stays lower than Eukaryotes. However, with time, mutations accumulate much faster than the rare unequal crossovers. So, the distortion caused by mutations exceeds, over time, the distortion caused by unequal crossovers. This implies higher fidelity in Eukaryotes than Bacteria and Archaea. For example, assume that mutations and unequal crossovers follow Poisson point processes within the genome with parameters $\lambda_{\text{mutation}} >> \lambda_{\text{unequal crossover}}$. Notice that this assumption does not contradict the Gaussianity of the noise but rather provides a likelihood of mutation and crossover within the genome. That is, each nucleotide can mutate to any other nucleotide following a Gaussian distribution whereas the Poisson points control the variance of the Gaussian (a small variance would correspond to a mutation and a large variance would correspond to a crossover). Then, it can be shown that the *R*-*D* curves of Archaea, Bacteria and Eukaryotes follow the trend observed in Figure 6.3.

# CHAPTER 7

# NON-STATIONARY ANALYSIS OF NUCLEOTIDE SEQUENCES

*“We are born by accident into a purely random universe. Our lives are determined by entirely fortuitous combinations of genes. Whatever happens happens by chance. The concepts of cause and effect are fallacies. There is only ‘seeming’ causes leading to ’apparent’ effects. Since nothing truly follows from anything else, we swim each day through seas of chaos, and nothing is predictable, not even the events of the very next instant.”*

Robert Silverberg , *The Stochastic Man.*

## Abstract

*Previous statistical analysis efforts of DNA sequences revealed that non-coding regions exhibit long-range power law correlations, whereas coding regions behave like random sequences or sustain short-range correlations. A great deal of debate on the presence or absence of long-range correlations in nucleotide sequences, and more specifically in coding regions, has ensued. These results were obtained using signal processing techniques for stationary signals and statistical tools for signals with slowly-varying trends superimposed on stationary signals. However, it can be verified using statistical tests that genomic sequences are non-stationary and the nature of their non-stationarity varies and is often much more complex than a simple trend. In this Chapter, we will bring to bear new tools to analyze non-stationary* *signals that have emerged in the statistical and signal processing community over the past few years. The emergence of these new methods will be used to shed new light and help resolve the issues of (i) the existence of long-range correlations in DNA sequences and (ii) whether they are present in both coding and non-coding segments or only in the latter. It turns out that the statistical differences between coding and non-coding segments are much more subtle than previously thought using stationary analysis. In particular, both coding and non-coding sequences exhibit long-range correlations, as asserted by a* $\mathbf{1/f}^{\beta(\mathbf{n})}$ *evolutionary (i.e., time-dependent) spectrum. However, we will use an index of randomness, which we derive from the Hilbert-Huang Transform, to demonstrate that coding sequences, although not random as previously suspected, are often "more random" (i.e., whiter) than non-coding sequences. Moreover, the study of the evolution of the rate of change of these time-dependent parameters in homologous gene families shows a sudden jump around the rat, which might be related to the well-known supercharged evolution of this rodent.*

## 7.1 Introduction

In 1992, Peng et al. (90) studied the stochastic properties of DNA sequences by constructing a map of the nucleotide sequences onto a walk, $u(i)$, which they termed a "DNA walk." The DNA walk is defined by the rule that the walker steps up $(u(i) = +1)$ (resp., down $(u(i) = -1)$) if a pyrimidine (resp., purine) resides at position $i$. In our analysis, we will rely on the same mapping of the nucleotides since our experiments have shown that the statistical properties remain unchanged even when we adopt a more complex multi-dimensional representation (29), (92). Peng et al. found that non-coding sequences exhibit long-range correlations; whereas coding sequences behave like random sequences or sustain at most short-range correlations.

Similar observations were reported independently by Li et al. in (67), who applied standard Fourier analysis to a sample of genes. This prompted a sequence of controversial papers, some affirming (6), (15), (21), (22), (39), (49), (66), (84), (91), (93), (112), (117), and others disputing (1), (9), (24), (47), (60), (80), (86), (94), (118) the existence of long-range correlations in DNA sequences or the statistical difference between coding and non-coding segments. This debate continues till today and consequently impedes further progress to explain the origins and functions of these correlations and their effect on the evolution of the DNA. We believe that such contradictory results are an artifact of using stationary signal processing and statistics tools to study non-stationary genomic signals. A time-series $\{X(n)\}$ (here $n$ denotes the position in the DNA sequence) is called *stationary* if, loosely speaking, its statistical properties do not change with time. This means that the processes $X(n)$ and $X(n+c)$ have the same statistics for any $c$. The Detrended Fluctuation Analysis (DFA) technique (91) constructs a stationary process from the non-stationary DNA walk by subtracting the non-stationary trend from the sequence. However, the DFA method is limited to the very special case of non-stationary signals consisting of stationary signals with embedded (polynomial) trends, i.e.,

$$X(t) = c(t) + X_0(t), \tag{7.1}$$

where $c(t)$ is a deterministic (polynomial) function and $X_0(t)$ is a stationary process (25). Moreover, even if a stationary process were embedded in some trend, then one has to know or estimate the form of the trend (polynomial, logarithmic, exponential, sinusoidal, etc) in order to subtract it.

## 7.2 The Evolutionary Spectrum and Test for Stationarity

Various methods have been proposed for testing whether or not a given series may be regarded as stationary. Some of these are designed to detect non-stationary "trends" in a particular characteristic of the series, such as the mean or variance (97) while others are designed to test whether the covariance or spectral properties of two sections of a series are compatible (55). In the latter case the two sections of the series have to be specified, a priori, and it is assumed that within each section the series is stationary. Priestly and Rao (96), proposed a method to test the overall stationarity of the complete second-order properties of a series. The basis of the method is to estimate its evolutionary (or time-dependent) spectrum over a range of time points, and then test these spectra for uniformity over time.

### 7.2.1 The Evolutionary Spectrum

Suppose we are given observations on a (possibly) non-stationary zero-mean continuous process $\{X(t)\}$. Then, an estimate of its evolutionary spectral density function, $\{h_t(\omega_0)\}$ at frequency $\omega_0$, is performed in two stages (95): (i) pass the data through a linear filter centered at frequency $\omega_0$, say, yielding output $U(t)$; (ii) compute a weighted average of $|U(t)|^2$ in the neighborhood of the time point $t$ to provide an estimate of the local power density at frequency $\omega_0$. Thus, given observations $\{X(t)\},\ 0 \leq t \leq T$, we set

$$U(t) = \int_{t-T}^{t} g(u)X(t-u)e^{-i\omega_0(t-u)}du, \tag{7.2}$$

$$\hat{h}_t(\omega_0) = \int_{t-T}^{T} w(v)|U(t-v)|^2 dv, \tag{7.3}$$

where $\{g(u)\}$ is a filter whose transfer function

$$\Gamma(\omega) = \int_{-\infty}^{\infty} g(u)\ e^{-i\omega u} du \tag{7.4}$$

is peaked in the neighborhood of $\omega = 0$, and is normalized; and $\{w(u)\}$ is a normalized function so that $\int_{-\infty}^{\infty} w(v) dv = 1$. Here, we use

$$g(u) = \begin{cases} \frac{1}{2\sqrt{\pi h}}, & |u| \leq h; \\ 0, & |u| > h. \end{cases} \tag{7.5}$$

and

$$w(v) = \begin{cases} \frac{1}{T'}, & -\frac{1}{2}T' \leq v \leq \frac{1}{2}T'; \\ 0, & \text{otherwise.} \end{cases} \tag{7.6}$$

### 7.2.2 Test for Stationarity

Suppose now that we have evaluated the estimated evolutionary spectra over the interval $(0, T)$. We choose a set of times $t_1, t_2, \cdots, t_I$ and a set of frequencies $\omega_1, \omega_2, \cdots, \omega_J$ which cover the range of times and frequencies of interest. If now we write

$$Y_{t,\omega} = \log\{\hat{h}_t(\omega)\}, \tag{7.7}$$

and

$$Y_{i,j} = Y(t_i, \omega_j), \quad h_{ij} = log\{h_{t_i}(\omega_j)\}, \quad e_{i,j} = e(t_i, \omega_j), \quad i = 1, \cdots, I; j = 1, \cdots, J. \tag{7.8}$$

then we obtain the model

$$Y_{ij} = h_{ij} + e_{ij}, , \quad i = 1, \cdots , I; j = 1, \cdots , J. \tag{7.9}$$

$\mathrm{var}\{e_{ij}\} = \mathrm{var}[Y_{ij}] = \sigma^2$ is given by

$$\sigma^2 = 2\pi\{\int_{-\infty}^{\infty} |\Gamma(\omega)|^4 d\omega\}/T^{'}, \tag{7.10}$$

If the $\{t_i\}$ and $\{\omega_j\}$ are spaced "sufficiently wide apart", then the $\{e_{ij}\}$ will be approximately uncorrelated. Priestley (95) showed that, in order to obtain approximately uncorrelated estimates, the points $\{\omega_j\}$, $\{t_i\}$ should be chosen so that the spacings between the $\{\omega_j\}$ are at least $\pi/h$ and the spacings between the $\{t_i\}$ are at least $T^{'}$. The test for stationarity can then be written in the form

$$H_0 : Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ij}.$$

$$H_1 : Y_{ij} = \mu + \beta_j + e_{ij}.$$

The parameters $\{\alpha_i\}$, $\{\beta_j\}$ may be interpreted as the "main effects" of the time and frequency "factors", respectively, and the $\{\gamma_{ij}\}$ represent an "interaction" term between these two factors. If all the $\{\gamma_{ij}\}$ are zero, then $log\{h_t(\omega)\}$ is additive in terms of time and frequency, so that $h_t(\omega)$ is multiplicative, i.e., may be written in the form

$$h_t(\omega) = c^2(t)\ h(\omega), \tag{7.11}$$

TABLE VI. Analysis of variance for a two-factor design

| Item | Degrees of freedom | Sum of squares |
|---|---|---|
| Between times | $I-1$ | $S_T = J\sum_{i=1}^{I}(Y_{i.} - Y_{..})^2$ |
| Between frequencies | $J-1$ | $S_F = I\sum_{j=1}^{J}(Y_{.j} - Y_{..})^2$ |
| Interaction + residual | $(I-1)(J-1)$ | $S_{I+R} = \sum_{i=1}^{I}\sum_{i=1}^{I}(Y_{ij} - Y_{i.} - Y_{.j} + Y_{..})^2$ |

for some functions $c(t), h(\omega)$. It is then not difficult to show that $\{X(t)\}$ must be of the form

$$X(t) = c(t)\ X_0(t), \tag{7.12}$$

where $\{X_0(t)\}$ is a stationary process with spectral density function $h(\omega)$. Processes of the form of Equation 7.12 are called *uniformly modulated processes.* Thus a test for the presence of interaction is equivalent to testing whether or not $\{X(t)\}$ is a uniformly modulated process. Given the computed values of $Y_{ij}$, we construct the standard analysis of variance table for a two-factor design, which with the usual notation, is set out in Table VI.

1. In testing for stationarity, the first step is to test for the interaction sum of squares, using the result, $S_{I+R}/\sigma^2 \sim \chi^2_{(I-1)(J-1)}$ (since we are assuming that $\sigma^2$ is known, all comparisons are based on $\chi^2$ rather than $F$-tests.)

2. If the interaction is not significant, we conclude that $\{X(t)\}$ is a uniformly modulated process, and proceed to test for stationarity by testing $S_T$ using $S_T/\sigma^2 \sim \chi^2_{(I-1)}$.

3. If, however, the interaction turns out to be significant, we conclude that $\{X(t)\}$ is non-stationary, and non-uniformly modulated.

4. Reversing the roles of "times" and "frequencies", the above procedure may be used in exactly the same way to test for "complete randomness".

Figure 21 shows the DNA walk of the Human gene TXNDC9. Using the same statistical parameters in (95, Chapter 6), we applied the above test to this gene with 95% confidence. We obtain the following statistics for the exponential signal of the Human gene TXNDC9: $S_{I+R}/\sigma^2 = 1284.5 > \chi^2_{336}(0.05) = 379.74; S_T/\sigma^2 = 9.7 \times 10^7 > \chi^2_{56}(0.05) = 74.46; S_F/\sigma^2 = 6912.4 > \chi^2_6(0.05) = 12.59$. The interaction, the between times sum of squares and the between frequencies sum of squares are highly significant confirming that the exponential signal is non-stationary, non-uniformly modulated and non-random. In particular, this genomic signal is non-stationary and the nature of its non-stationarity is not associated with a deterministic trend as in Equation 7.1.

### 7.3 The Evolutionary Periodogram and the Evolutionary $1/f$ Process

Much of the current evidence for long-range correlations in DNA sequences stems from the experimentally observed $1/f$ spectrum (21), (47), (66), (67), (117), (118). The $1/f$ spectrum assumes the existence of a stationary process with a fixed spectral exponent $\beta$. This assumption, however, is in contradiction to our assertion that nucleotide sequences are non-stationary. We therefore propose a new evolutionary (time-dependent) $1/f$ spectrum whose spectral exponent $\beta(n)$ varies in time. This approach also resolves the classical paradox of $1/f$ processes, namely,

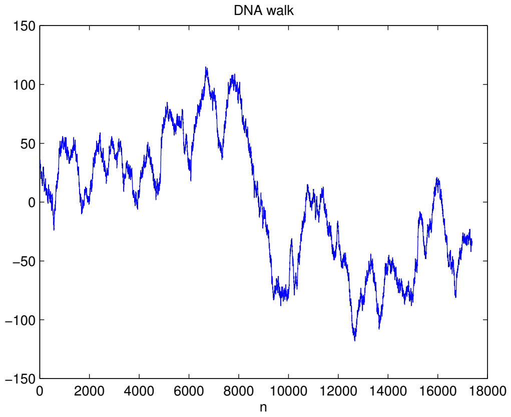


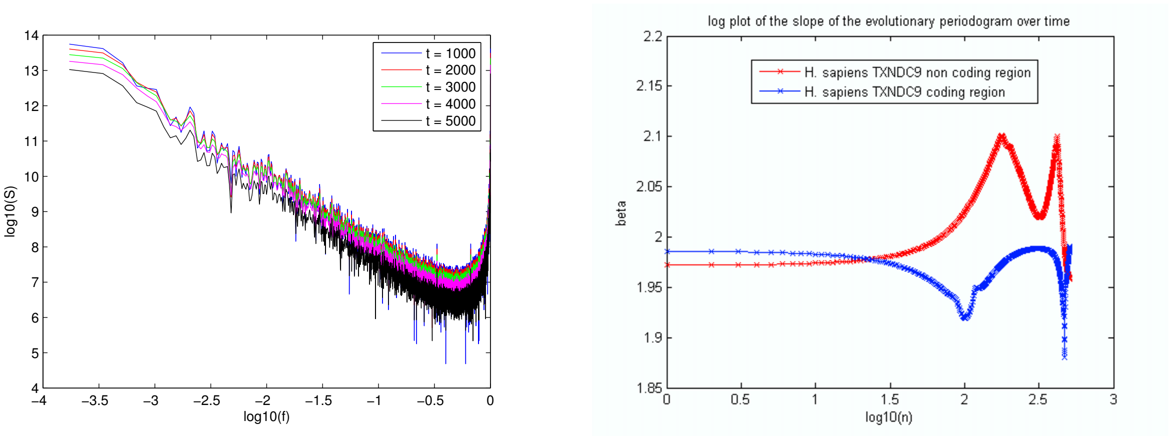
Figure 21. DNA walk of the Human gene TXNDC9 using the purine-pyrimidine rule.

the variance of a $1/f$ process with a spectral exponent $\beta$, $1 < \beta < 2$, obtained by integration of the power spectral density, is infinite (21), (62).

A generalization of the periodogram for estimating the power spectrum of non-stationary signals is given by the *evolutionary periodogram* (EP) (61). The EP of a non-stationary signal $x(n), n = 0, \cdots, N-1$, is defined as

$$\begin{aligned} S(n,f) &= \frac{N}{M}|A(n,f)|^2 \\ &= \frac{N}{M}\Big| \sum_{i=0}^{M-1} P_i^*(n) \sum_{k=0}^{N-1} P_i(k)x(k)e^{-2\pi jfk}\Big|^2, \end{aligned} \tag{7.13}$$

where $^*$ denotes complex conjugate, and $\{P_i(n)\}_{i=0}^{M-1}$ is an orthonormal basis. The number $M(\leq N)$ may depend on the frequency $f$, and indicates the degree to which $A(n,f)$ varies with time. For small values of $M$, $A(n,f)$ is slowly varying, and for large values of $M$, it is rapidly varying. In our simulations, we use the discrete Legendre polynomials with $M = 3$. Observe that Equation 7.13 can be interpreted as the magnitude squared of the Fourier transform of $x(k)$ windowed by the sequence $v(n, k) = \sum_{i=0}^{M-1} \beta_i^*(n)\beta_i(k)$. Therefore, the evolutionary periodogram can be efficiently implemented using the Fast Fourier Transform (fft) algorithm. Like the stationary power spectrum, the evolutionary periodogram is noisy for large $f$. To smooth it, we apply the same averaging procedure done for the stationary spectrum: We divide the sequence into non-overlapping subsequences of equal size (usually a power of 2 for the fft), then we compute the EP of each subsequence, and average the results from the individual subsequences together. The EP of the coding region of the Human MHY6 gene is shown in Figure 22(a) for $n = 1000, 2000, 3000, 4000, 5000$. Note that the two peaks, corresponding to the frequencies 1/3 and 2/3, are known to be related to the codon structure in DNA coding regions. Also, note that the scaling exponent $\beta$ is not constant, but rather varies for different values of $n$. This shows that DNA correlations are much more complex than power laws with a single scaling exponent. Thus, the proposed time-varying or "evolutionary $1/f$" process, where the exponent $\beta(n)$ is a function of time, provides a far superior model of the correlation structure of DNA sequences. We estimate the function $\beta(n)$ by a linear least-squares fit of the slope of the EP at each time instant $n$. White noise corresponds to $\beta(n) = 0$. Figure 22(b) depicts a plot of $\beta(n)$ versus $\log_{10}(n)$ for the coding and non-coding regions of the Human gene TXNDC9. Observe that, for this gene, both the coding and non-coding regions exhibit long-range correlations. Moreover, the average exponent function of the non-coding region is higher than the corresponding value in the coding region. Next, we will demonstrate that our



(a)

(b)

Figure 22. (a) Evolutionary Periodogram of the coding region of the Human MHY6 gene for $n = 1000, 2000, 3000, 4000, 5000$. The length of the gene is $N = 5820$. (b) The scaling exponent $\beta(n)$ for the coding and non-coding regions of the Human gene TXNDC9 as a function of $\log_{10}(n)$

conclusion that (i) neither the coding or non-coding regions are random and (ii) the "degree of randomness" of the coding regions is higher than non-coding regions, is not an artifact of the evolutionary $1/f$ model.

### 7.4 Empirical Mode Decomposition and Index of Randomness

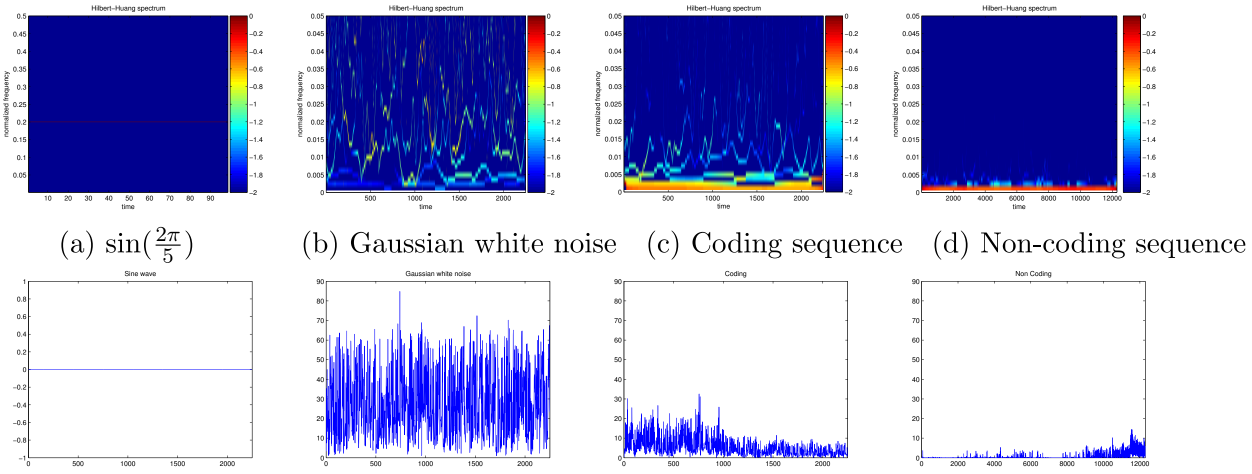
To quantify the statistical processes further, a more sensitive index is needed to give a quantitative measure of how far the process deviates from white noise; a prerequisite for such a definition is a method to present the data in the frequency-time space. There are many methods to obtain such a 3D distribution, e.g., the spectrogram, the wavelet analysis and the Wigner-Ville distribution. These techniques have been reviewed and assessed in (52), where the authors introduced a new non-linear technique, called *Empirical Mode Decomposition* (EMD), to represent non-stationary signals as sums of AM-FM components by decomposing them into mode functions and then applying the Hilbert Transform to each mode. The analytic process $Z(t)$ can then be expressed as (52)

$$Z(t) = \sum_{j=1}^{N} a_j(t) e^{i2\pi \int f_j(t)dt}. \tag{7.14}$$

Equation 7.14 enables us to represent the amplitude, $a_j(t)$, and the instantaneous frequency, $f_j(t)$, as functions of time in a three-dimensional plot, in which the amplitude can be contoured on the frequency-time plane. This frequency-time distribution of the amplitude is designated as the Hilbert spectrum. Figure 23 shows the Hilbert amplitude spectrum of a pure sine wave, a Gaussian random noise, the Human gene NOC2L and its coding and non-coding sequences. Visually, the coding segment looks “whiter” that the non-coding one. We propose to quantify the notion of “how far is a process from a white noise” by defining the index of randomness at time instant $t$, $\mathrm{IR}(t)$, as the weighted variance or spread of the spectrum at time $t$. So, for a pure sine wave, the spectrum is a delta function and the variance is zero; whereas for a white noise, the spectrum is flat and the variance is infinite. Analytically,

$$\mathrm{IR}(t) = \frac{1}{N} \sum_{f=1}^{N} \frac{\mathrm{a}(f,t)}{\max\limits_{f}\{\mathrm{a}(f,t)\}} (f - \mu(t))^2, \tag{7.15}$$

where $a(f,t)$ is the amplitude of the Hilbert spectrum at frequency $f$ and time $t$, $N$ is the maximum number of frequency cells, and $\mu(t) = \mathrm{mean}_{f \in I(t)}\ \{f\}$, where $I(t) = \{f : a(f,t) \neq 0\}$. Figure 23 depicts the Hilbert spectrum and the index of randomness for different signals. We



(a) $\sin(\frac{2\pi}{5})$ (b) Gaussian white noise (c) Coding sequence (d) Non-coding sequence

(e) IR of (a) (f) IR of (b) (g) IR of (c) (h) IR of (d)

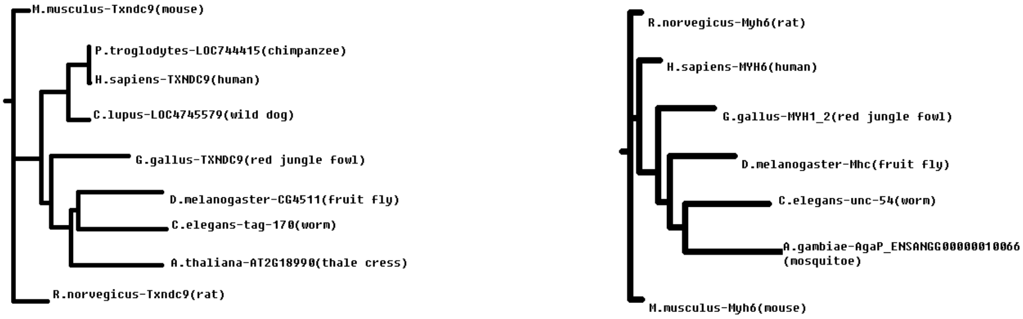
Figure 23. Row 1: Amplitude-frequency-time distribution using the Hilbert transform (amplitudes depicted in a logarithmic scale). Row 2: Index of randomness of the signals in row 1. (c) and (d) display the Hilbert spectrum of the coding and non-coding segments of the Human gene NOC2L (GI:89161185), respectively.

once again observe that (i) the coding and non-coding regions are not random, and (ii) the coding regions are more random than the non-coding regions. This observation is most likely the source of confusion and controversy in previous work related to DNA correlations.

## 7.5 Evolutionary trends

We now apply the non-stationary tools presented to two homologous gene families: the myosin heavy cardiac muscle gene and the thioredoxin domain containing 9 gene. Both homologous groups were identified using the online NCBI HomoloGene system for automated detection of homologs among annotated genes (`http://www.ncbi.nlm.nih.gov/sites/entrez?db=homologene`). Using the PHYLIP package (`http://evolution.genetics.washington.edu/`

`phylip.html`), we plotted the inferred phylogenies of both families in Figure 24. The evolutionary periodogram was computed by dividing each sequence into subsequences of length $L = 512$ and averaging the EP of the subsequences. The averaging procedure smoothes the evolutionary periodogram and renders better estimates of the exponent curve $\beta(n)$. In addition, the evolutionary periodograms and exponent curves of all considered genes will have length 512  and therefore can be compared. The exponent curves $\beta(n)$ of the coding and non-coding segments of each gene are displayed in Figure 25, along with the average index of randomness of the coding and non-coding segments of each gene group. Notice that the exponent curve $\beta(n)$ is more conserved across evolution in exons than in introns. This result is consistent with the findings that functional DNA sequences tend to undergo mutation at a slower rate than non-functional sequences (27). Moreover, the average index of randomness in coding sequences is higher than its counterpart in non-coding sequences. Finally, even though the exponent curves $\beta(n)$ do not seem to follow a particular evolutionary trend, we will show that some statistical features derived from $\beta(n)$ exhibit very interesting evolutionary patterns. For each gene, we consider the average exponent $\beta_a$ given by the mean of the coding curve. We define the evolutionary rate, $r_g$, at a node gene $g$ as the derivative of $\beta_a$ along the tree branch between the gene, $g$, and its ancestor $G$, i.e., $r_g = \frac{\beta_a(g)-\beta_a(G)}{t_g-t_G}$, where $\beta_a(g)$, $\beta_a(G)$ are the values of $\beta_a$ for the genes $g$ and $G$, respectively; and $t_g, t_G$ are the relative evolutionary times of genes $g$ and $G$, respectively. The evolutionary distance $t_g - t_G$ was computed as the distance between the aligned gene sequences $g$ and $G$. Table VII provides the evolutionary rates of both gene groups and shows a clear jump in the evolutionary rate around the mouse in both gene groups. This

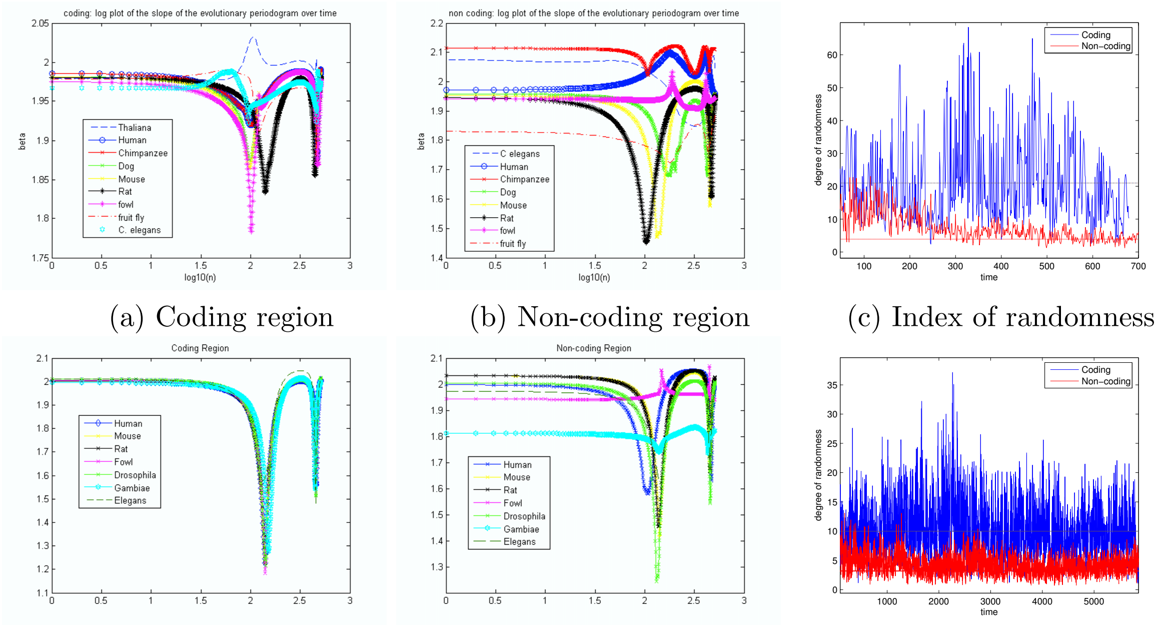

M.musculus-Txndc9(mouse)
P.troglodytes-LOC744415(chimpanzee)
H.sapiens-TXNDC9(human)
C.lupus-LOC4745579(wild dog)
G.gallus-TXNDC9(red jungle fowl)
D.melanogaster-CG4511(fruit fly)
C.elegans-tag-170(worm)
A.thaliana-AT2G18990(thale cress)
R.norvegicus-Txndc9(rat)
R.norvegicus-Myh6(rat)
H.sapiens-MYH6(human)
G.gallus-MYH1_2(red jungle fowl)
D.melanogaster-Mhc(fruit fly)
C.elegans-unc-54(worm)
A.gambiae-AgaP_ENSANGG00000010066 (mosquitoe)
M.musculus-Myh6(mouse)

(a) Homologous TXNDC9 genes

(b) Homologous MYH6 genes

Figure 24. The Phylogenetic trees of the gene groups: TXNDC9 and MYH6. They were plotted using the PHYLIP package.

observation is quite remarkable given the well-known explosive evolution of this rodent. Furthermore, the variance of the evolutionary rates, using a window of size 3, shows an increasing trend throughout evolution. The evolutionary rate could therefore possibly be used to observe and predict the dynamics of change in a lineage.



(a) Coding region (b) Non-coding region (c) Index of randomness

(d) Coding region (e) Non-coding region (f) Index of randomness

Figure 25. Exponent curves and index of randomness. Row one: Gene TXNDC9 (a) The exponent curves $\beta(n)$ of the coding region of gene TXNDC9; (b) The exponent curves $\beta(n)$ of the non coding region of gene TXNDC9; (c) Index of randomness of the coding (blue) and non-coding (red) segments of the TXNDC9 gene group. The plot of the non-coding graph was truncated to the length of the coding segment. The lower (upper) horizontal line is the average index of randomness of the non-coding (coding) regions. Row 2: same as Row 1 for gene MYH6.

TABLE VII. Evolutionary rates and their variances

| Gene TXNDC9 | Evolutionary Rate | Variance | Gene MYH6 | Evolutionary Rate | Variance |
|---|---|---|---|---|---|
| Thaliana | | | Gambiae | | |
| Elegans | -0.04 | 0.00 | Elegans | 0.08 | 0.02 |
| Drosophila | 0.00 | 0.00 | Drosophila | -0.09 | 0.01 |
| Fowl | -0.06 | 0.00 | Fowl | 0.00 | 0.03 |
| Rat | 0.03 | 0.01 | Rat | 0.26 | 0.58 |
| Mouse | 0.12 | 0.18 | Mouse | -1.16 | 0.58 |
| Dog | -0.67 | 0.21 | Human | 0.00 | |
| Chimpanzee | 0.15 | 0.19 | | | |
| Human | 0.00 | | | | |

# FUTURE WORK

My future objective is to discover, model and study the dynamics of a transcriptional network (gene regulatory network) that transitions cells from one phenotype to another. Specifically, I plan to represent complex biochemical processes within the genetic transcription system by relatively simple mathematical models in order to pursue questions about network dynamics. Given a genetic network with its negative and positive feedback loops, I plan to investigate stable equilibria and other dynamical aspects. In contrast to classical biological experiments in which conclusions about functionality are based on rough phenotypical and genetic input/output behavior, the dynamical system will allow one to more thoroughly explore biological hypotheses. Within this context, one can discover favorable equilibrium states, how different perturbations might affect gene dynamics, make deductions about the "dynamical functions" of the genes, and discover which of the network genes play critical roles in either creating or suppressing phenotypes.

# GLOSSARY

**archaea:** one of three principal domains of life. Archaeal cells can be distinguished from bacterial cells based on gene sequence and gene content. Many archaea are specialized to conditions of extreme heat or salinity.

**bacteria:** one of three principal domains of life. Bacterial cells contain no nucleus and usually possess a cell wall of petidoglycan.

**BLAST:** (Basic Local Alignment Search Tool) a widely used program for searching sequence databases for entries that are similar to a specified query sequence.

**bioinformatics:** the use of computational methods to study biological data.

**cDNA:** DNA strand that is complementary to an RNA stand and synthesized from it by a reverse transcriptase.

**central dogma:** the central dogma of molecular biology was first enunciated by Francis Crick in 1958 and re-stated in a Nature paper published in 1970: "*The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred from protein to either protein or nucleic acid.*" In other words, once information gets into protein, it can't flow back to nucleic acid. The central dogma is often misunderstood. It is frequently confused with the standard pathway of information flow from "DNA to RNA to protein". There are notable exceptions to the normal pathway of information flow and these are often mistakenly referred to as exceptions to the central dogma.

**codon:** triplet of three nucleotides that specifies one type of amino acid during the translation process.

**eukaryotes:** one of three principal domains of life. Eukaryotic cells possess a nucleus, and usually contain other organelles like, e.g., mitochondria. Eukaryotes comprise animals, plants, and fungi which are mostly multicellular as well as various other groups that are collectively classified as protists (many of which are unicellular).

**exon:** part of a gene sequence that is transcribed and translated to give rise to a protein (cf. intron).

**genetic code:** set of assignments of the 64 codons to the 20 amino acids.

## GLOSSARY (Continued)

**genome:** the complete sequence of heritable DNA of an organism.

**genomics:** the study of genomes. Usually applies to studies that deal with very large sets of genes using high-throughput experimental techniques.

**homologs:** sequences that are evolutionary related by descent from a common ancestor.

**intron:** part of the DNA sequence of a gene that is transcribed, but it cut out of the mRNA (i.e., spliced) prior to translation. Introns do not code for protein sequences.

**nucleic acid:** a polymerase molecule composed of nucleotides. May be either DNA (deoxyribonucleic acid) or RNA (ribonucleic acid).

**nucleotide:** chemical unit that forms the building block for nucleic acids. Composed of a nitrogenous base, a ribose or deoxyribose suagr, and a phosphate group (cf. nucleic acid, purines, pyrimidines).

**PAM**(point accepted mutation) **matrix:** a matrix describing the rate of substitution of one type of amino acid by another during protein evolution.

**phylogeny:** an evolutionary tree showing the relationship between sequences or species.

**promoter:** region of DNA upstream of a gene that acts as a binding site for a transcription factor and ensures that the gene is transcribed.

**prokaryotes:** organisms, such as bacteria, whose cells lack nuclei and other complex cell structures.

**proteome:** the complete set of proteins present in a cell.

**proteomics:** the study of the proteome, usually using two-dimensional gel electrophoresis and mass spectrometry to separate and identify the proteins.

**purines:** the bases A (adenine) and G (guanine) that are present in nucleic acid sequences.

**pyrimidines:** the bases C (cytosine), T (thymine), and U (uracil) that are present in nucleic acid sequences.

**transcription:** the synthesis of an RNA strand using a complementary DNA strand as a template.

# GLOSSARY (Continued)

**translation:** the process by which the ribosome decodes the gene sequence specified on a mRNA and synthesizes the corresponding protein.

**unequal crossover:** a crossover between homologous chromosome that are not perfectly aligned.

# BIBLIOGRAPHY

1. Abramson, G., Cerdeira, H. A., and Bruschi, C.: Fractal properties of DNA walks. Biosystems, 49(1):63–70, 1999.

2. Almagor, H.: Nucleotide distribution and the recognition of coding regions in DNA sequences: an information theory approach. Journal of Theoretical Biology, 117:127–136, 1985.

3. Altschul, S. F.: Amino acid substitution matrices from an information theoretic perspective. Journal of Molecular Biology, 219:555–565, 1991.

4. Altschul, S. F.: Amino acid substitution matrices from an information theoretic perspective. Journal of Molecular Biology, 219:555–665, 1991.

5. Anastassiou, D.: Genomic signal processing. IEEE Signal Processing Magazine, 18(4):8–20, July 2001.

6. Arneodo, A., Bacry, E., Graves, P. V., and Muzy, J. F.: Characterizing long-range correlations in DNA sequences from wavelet analysis. Physical Review Letters, 74(16):3293–3296, April 1995.

7. Avery, J.: Information Theory and Evolution. Singapore, World Scientific, 2003.

8. Aydin, Z. and Altunbasak, Y.: A signal processing application in genomic research: protein secondary structure prediction. IEEE Signal Processing Magazine, 23(4):128–131, July 2006.

9. Azbel, M. Y.: Universality in a DNA statistical structure. Physical Review Letters, 75(1):168–171, July 1995.

10. Battail, G.: Does information theory explain biological evolution? Europhysics Letters, 40(3):343–348, 1997.

11. Battail, G.: Should genetics get an information-theoretic education? IEEE Engineering in Medicine and Biology Magazine, 25(1):34–45, January 2006.

12. eds, G. I. Bell and T. G. Marr Computers and DNA. Addison-Wesley, 1988.

13. Berg, J. M., Tymoczko, J. L., and Stryer, L.: Biochemistry. Michelle Julet, 2002.

14. Berger, T.: Rate Distortion Theory: A Mathematical Basis for Data Compression. Englewood Cliffs, New Jersey, Prentice-Hall, Inc., 1971.

15. Berthelsen, C. L., Glazier, J. A., and Skolnick, M. H.: Global fractal dimension of human DNA sequences treated as pseudorandom walks. Physical Review A (Atomic, Molecular, and Optical Physics), 45(12):8902–8913, June 1992.

16. Blahut, R.: Computation of channel capacity and rate-distortion functions. IEEE Transactions on Information Theory, 18:460–472, July 1972.

17. Bogatryreva, N., Finkelstein, A., and Galzitskaya, O.: Trend of amino acid composition of proteins of different taxa. Journal of Bioinformatics and Computational Biology, 4(2):597–608, 2006.

18. Bohr, V. A., Wassermann, K., and Kraemer, K. H.: DNA repair mechanisms. In Alfred Benzon Symposium, volume 35, pages 1–428, 1993.

19. Brennicke, A., Marchfelder, A., and Binder, S.: Rna editing. FEMS Microbiol Rev, 23(3):297–316, June 1999.

20. Browner, W. S., Kahn, A. J., Ziv, E., Reiner, A. P., Oshima, J., Cawthon, R. M., Hsueh, W. C., and Cummings, S. R.: The genetics of human longevity. American Journal of Medicine, 117(11):851–860, 2004.

21. Buldyrev, S. V., Goldberger, A. L., Havlin, S., Mantegna, R. N., Matsa, M. E., Peng, C. K., Simons, M., and Stanley, H. E.: Long-range correlation properties of coding and noncoding DNA sequences: Genbank analysis. Physical Review E, 51:5084 – 5091, 1995.

22. Carpena, P., Bernaola-Galvan, P., Coronado, A. V., Hackenberg, M., and Oliver, J. L.: Identifying chracteristic scales in the human genome. Physical Review E, 75:032903, 2007.

23. Cavalier-Smith, T.: Selfish DNA and the origin of introns. Nature, 315:283284, 1978.

24. Chatzidimitriou-Dreismann, C. A. and Larhammar, D.: Long-range correlations in DNA. Nature, 361:212, January 1993.

25. Chen, Z., Ivanov, P. C., Hu, K., and Stanley, H. E.: Effect of nonstationarities on detrended fluctuation analysis. Physical Review E, 65:041107, 2002.

26. Comeron, J. M. and Kreitman, M.: Negative correlation between intron length and recombination rate. Genetics, 156(3):1175–1190, 2000.

27. Consortium, M. G. S.: Initial sequencing and comparative analysis of the mouse genome. Nature, 420(6915):520–62, December 2002.

28. Crick, F.: Central dogma of molecular biology. Nature, 227:561–563, 1970.

29. Cristea, P. D.: Large scale features in DNA genomic signals. Signal Processing, 83(4):871 – 888, April 2003.

30. Dayhoff, M., Schwartz, R., , and Orcutt, B.: A model of evolutionary change in proteins. In Atlas of Protein Sequence and Structure, ed. M. Dayhoff, volume 5, pages 345–352, 1978.

31. Deutsch, E. and Zenger, C.: Inclusion domains for the eigenvalues of stochastic matrices. Numerische Math., 18:182–192, 1971.

32. Doolittle, R. F.: The genealogy of some recently evolved vertebrate proteins. Trends in Biochemical Sciences, 10(6):233–237, 1985.

33. Eigen, M.: The origin of genetic information: viruses as models. Gene, 135:37–47, 1993.

34. Forsdyke, D. R.: Are introns in-series error-detecting sequences? Journal of Theoretical Biology, 93:861–866, 1981.

35. Fowler, T. B.: Computation as a thermodynamic process applied to biological systems. International Journal of Bio-Medical Computing, 10(6):477–489, 1979.

36. Friedberg, E., Walker, G., and Siede, W.: DNA Repair and Mutagenesis. Washington, D.C., ASM Press, 1995.

37. Frobenius, G.: Uber matrizen aus nicht negativen elementen. S.B. Preuss. Akad. Wiss., pages 456–477, 1912.

38. G. Rosen, J. M.: Investigation of coding structure in DNA. In IEEE International Conference on Acoustics, Speech, and Signal Processing, pages 361–364, Hong Kong, April 2003.

39. Gao, J., Qi, Y., Cao, Y., and Tung, W.-W.: Protein coding sequence identification by simultaneously characterizing the periodic and random features of DNA sequences. Journal of Biomedicine and Biotechnology, 2:139C146, 2005.

40. Gatlin, L. L.: Information Theory and the Living System. New York, Columbia University Press, 1972.

41. Gilbert, W.: The exon theory of genes. In Cold Spring Harbor Symposia on Quantitative Biology, volume 52, pages 901–905, 1987.

42. Gilbert, W., Souza, S. D., and Long, M.: Origin of genes. Proceedings of the National Academy of Sciences, 94:7698–7703, July 1994.

43. Gilbert, W.: Why genes in pieces? Nature, 271:501, February 1978.

44. Gnedenko, B. and Kolmogorov, A.: Limit Distributions for Sums of Independent Random Variables. Addison-Wesley, 1954.

45. Grosjean, H. and Benne, R.: Modification and Editing of RNA. Washington, DC, ASM Press, 1998.

46. Gross, L. J., Mullin, B. C., and Riechert, S. E.: Alternative routes to quantitative literacy for the life sciences. July 2000.

47. Guharay, S., Hunt, B. R., York, J. A., and White, O. R.: Correlations in DNA sequences across the three domains of life. Physica D, 146(1-4):388–396, 2000.

48. Gupta, M. K.: The quest for error correction in biology. IEEE Engineering in Medicine and Biology Magazine, 25(1):46–53, January 2006.

49. Haimovich, A. D., Byrne, B., Ramaswamy, R., and Welsh, W. J.: Wavelet analysis of DNA walks. Journal of Computational Biology, 13(7):1289–1298, 2006.

50. Henikoff, S. and Henikoff, J. G.: Amino acid substitution matrices from protein blocks. In Proc. Natl. Acad. Sci. USA., volume 89, pages 10915 – 10919, 1992.

51. Higgs, P. G. and Attwood, T. K.: Bioinformatics and Molecular Evolution. Blackwell publishing, 2005.

52. Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N. C., Tung, C. C., and Liu, H. H.: The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society, 454(1971):903–995, March 1998.

53. Huen, Y.: Brief comments on junk DNA: is it really junk? Complexity International, 9, February 2002.

54. Jeffrey, H. J.: Chaos game representation of gene structure. Nucleic Acids Research, 18(8):2163–2170, April 1990.

55. Jenkins, G. M.: General considerations in the estimation of spectra. Technometrics, 3:133–166, 1961.

56. Ji, S.: Molecular information theory: Solving the mysteries of DNA. In Modeling in Molecular Biology (Natural Computing Series), eds, G. Ciobanu and G. Rozenberg, page 141150., Berlin, 2004. Springer.

57. Jones, D. T., Taylor, W. R., and Thornton, J. M.: The rapid generation of mutation data matrices from protein sequences. Bioinformatics, 8:275–282, 1992.

58. Judson, H. F.: The Eighth Day of Creation. Makers of the Revolution in Biology. London, Cape, 1979.

59. Jukes, T. H., Holmquist, R., and Moise, H.: Amino acid composition of proteins: Selection against the genetic code. Science, 189:50–51, 1975.

60. Karlin, S. and Brendel, V.: Patchiness and correlations in DNA sequences. Science, 259(5095):677–680, 1993.

61. Kayhan, A. S., El-Jaroudi, A., and Chaparro, L. F.: Evolutionary periodogram for nonstationary signals. IEEE Transactions on Signal Processing, 42(6):1527–1536, June 1994.

62. Keshner, M. S.: 1/f noise. Proceedings of the IEEE, 70(3):212–218, March 1982.

63. Klug and Cummings: Concepts of Genetics. NJ, Prentice Hall, 1997.

64. Kopezynski, C. C. and Muskavitch, M. A. T.: Introns excised from the delta primarv transcript are localized near sites of delta transcription. The Journal of Cell Biology, 119:503, 1992.

65. Lewin, B.: Genes. New York, NY, Oxford University Press, 1995.

66. Li, W. and Holste, D.: Universal 1/f noise, crossovers of scaling exponents, and chromosome-specific patterns of guanine-cytosine content in DNA sequences of the human genome. Physical Review E, 71:041910, 2005.

67. Li, W. and Kaneko, K.: Long-range correlation and partial 1/f spectrum in a noncoding DNA sequence. Europhysics Letters, 17:655, February 1992.

68. Li, W. and Grauer, D.: Fundamentals of Molecular Biology. Sunderland, MA, Sinauer Assoc, 1994.

69. Liebovitch, L., Tao, Y., Todorov, A., and Levine, L.: Is there an error correcting code in the base sequence in DNA? Biophysical journal, 71(3):15391544, 1996.

70. Lodish, H., Berk, A., Matsudaira, P., Kaiser, C. A., Krieger, M., Scott, M. P., Zipursky, S. L., and Darnell, J.: Molecular Biology of the Cell. New York, NY, WH Freeman, 2004.

71. Loftfield, R. and Vanderjagt, D.: The frequency of errors in protein biosynthesis. Biochemical Journal, 128:1353–1356, 1972.

72. Long, M.: Evolution of novel genes. Curr. Opin. Genet. Dev., 11:673–680, 2001.

73. Maresca, B. and Schwartz, J. H.: Sudden origins: a general mechanism of evolution based on stress protein concentration and rapid environmental change. Anat Rec B New Anat, 289(1):38–46, 2006.

74. Matsuo, K., Kunzler, P., Georgiev, O., Urbanek, P., and Schaffner, W.: Short introns interrupting the oct-2 pou domain may prevent recombination between the pou family genes without interfering with potential pou domain shuffling in evolution. Journal of Biological Chemistry, 375:675–683, 1994.

75. May, E. E.: Analysis of coding theory based models for initiating protein translation in prokaryotic organisms. Doctoral dissertation, North Carolina State University, Raleigh, NC, March 2002.

76. May, E. E., Vouk, M. A., and Bitzer, D. L.: Classification of escherichia coli k-12 ribosome binding sites. IEEE Engineering in Medicine and Biology, 25(1):90–97, January 2006.

77. May, E. E., Vouk, M. A., Bitzer, D. L., and Rosnick, D. I.: A coding theory framework for genetic sequence analysis. In Workshop on Genomic Signal Processing and Statistics (GENSIPS), Raleigh. North Carolina, USA., October 2002.

78. Miki, K. and Hisano, T.: Determining the basic mechanisms of protein interactions. Riken News, (278), August 2004.

79. Milenkovic, O. and Vasic, B.: Information theory and coding problems in genetics. In Proc. IEEE Information Theory Workshop, October 2004.

80. Nee, S.: Uncorrelated DNA walks. Nature, 357:450, 1992.

81. Nolan, J. P.: Parametrizations and modes of stable distributions. Statistics and probability Letters, 38:187–195, 1998.

82. Ohno, S.: Evolution by Gene Duplication. Berlin, Springer, 1970.

83. Oliver, J. L., Bernaola-Galvan, P., Guerrero-Garcia, J., and Roman-Roldan, R.: Entropic profiles of DNA sequences through chaos-game-derived images. Journal of Theoretical Biology, 160:457–470, 1992.

84. Ossadnik, S. M., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Mantegna, R. N., Peng, C. K., Simons, M., and Stanley, H. E.: Correlation approach to identify coding regions in DNA sequences. Biophysical Journal, 67(1):64–70, 1994.

85. Palaniappan, K. and Jernigan, M. E.: Pattern analysis of biological sequences. In Proceedings of the 1984 IEEE International Conference on Systems, Man, and Cybernetics., 1984.

86. Pande, V. S., Grosberg, A. Y., and Tanaka, T.: Nonrandomness in protein sequences - evidence for a physically driven stage of evolution. Proceedings of the National Academy of Sciences, 91(26):12972–12975, 1994.

87. Papoulis, A. and Pillai, S. U.: Probability, Random Variables and Stochastic Processes. New York, McGraw-Hill, 1991.

88. Patthy, L.: Genome evolution and the evolution of exon shuffling: a review. Gene, 238:103–114, 1999.

89. Pavesi, A., Iaco, B. D., Granero, M. I., and Porati, A.: On the informational content of overlapping genes in prokaryotic and eukaryotic viruses. Journal of Molecular Evolution, 44(6):625–631, 1997.

90. Peng, C. K., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Sciortino, F., Simons, M., and Stanley, H. E.: Long-range correlations in nucleotide sequences. Nature, 356:168–170, 1992.

91. Peng, C. K., Buldyrev, S. V., Havlin, S., Simons, M., Stanley, H. E., and Goldberger., A. L.: Mosaic organization of DNA nucleotides. Phys. Rev. E, 49:16851689, 1994.

92. Podobnik, B., Shao, J., Dokholyan, N. V., Zlatic, V., Eugene, S. H., and Grossed, I.: Similarity and dissimilarity in correlations of genomic DNA. Physica A, 373:497–502, 2007.

93. Podobnik, B., Shao, J., Dokholyan, N. V., Zlatic, V., Stanley, H. E., and Grosse, I.: Similarity and dissimilarity in correlations of genomic DNA. Physica A, 373:497–502, 2006.

94. Prabhu, V. V. and Claverie, J. M.: Correlations in intronless DNA. Nature, pages 359–782, 1992.

95. Priestley, M. B.: Non-linear and Non-stationary time series analysis. Academic Press, 1988.

96. Priestley, M. B. and Rao, T. S.: A test for non-stationarity of time-series. Journal of the Royal Statistical Society. Series B (Methodological), 31(1):140–149, 1969.

97. Rao, S.: A note on the asymptotic relative efficiency of cox and stuart's tests for testing trend in dispersion of p-dependent time series. Biometrika, 55:381–385, 1968.

98. Richard, D. W., Mitchell, M., Sgouros, J., and Lindahl, T.: Human DNA repair genes. Science, 291(5507):1284–1289, 2001.

99. Roman-Roldan, R., Bernaola-Galvan, P., and Oliver, J. L.: Application of information theory to DNA sequence analysis: a review. Pattern Recognition, 29(7):11871194, 1996.

100. Root-Bernstein, R.: Protein replication by amino acid pairing. Journal of Theoretical Biology, 100:99–106, 1983.

101. Salamon, P. and Konopka, A. K.: A maximum entropy principle for the distribution of local complexity in naturally occuring nucleotide sequences. Computers and Chemistry, 16(2):117–124, 1992.

102. Sarkar, R., Roy, A. B., and Sarkar, P. K.: Topological information content of genetic molecules. Mathematical Biosciences, 39:299–312, 1978.

103. Schneider, T. D.: Theory of molecular machines. i. channel capacity of molecular machines. Journal of Theoretical Biology, 148:83–123, 1991.

104. Schneider, T. D.: Theory of molecular machines. ii. energy dissipation from molecular machines. Journal of Theoretical Biology, 148:125–137, 1991.

105. Schneider, T. D.: Measuring molecular information. Journal of Theoretical Biology, 201:87–92, 1999.

106. Schneider, T. D. and Mastronarde, D. N.: Fast multiple alignment of ungapped DNA sequences using information theory and a relaxation method. Discrete Applied Mathematics, 71:259–268, 1996.

107. Schneider, T. D., Stormo, G. D., Gold, L., and Ehrenfeucht, A.: Information content of binding sites on nucleotide sequences. Journal of Molecular Biology, 188:415–431, 1986.

108. Schneider, T. S.: Claude shannon: biologist (information theory used in biology). IEEE Engineering in Medicine and Biology, 25(1):30–33, January 2006.

109. Seneta, E.: Non-Negative Matrices and Markov Chains. Springer-Verlag, 1981.

110. Shannon, C.: A mathematical theory of communication. Bell System Technical Journal, 27:379–423, 1948.

111. Shannon, C. E. and Weaver, W.: The Mathematical Theory of Communication. Urbana, University of Illinois Press, 1949.

112. Stanley, H. E., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Peng, C. K., and Simons, M.: Scaling features of noncoding DNA. Physica A, 273(1):1–18, 1999.

113. Strait, B. J. and Dewey, T. G.: The shannon information entropy of protein sequences. Biophysical Journal, 71:148–155, 1996.

114. Strait, B. J. and Dewey, T. G.: The shannon information entropy of protein sequences. Biophysical Journal, 71:148–155, 1996.

115. Tornaletti, S. and Pfeiffer, G. P.: UV damage and repair mechanisms in mammalian cells. Bioessays, 18:221–228, 1996.

116. Turing, A. M.: On computable numbers, with an application to the entscheidungsproblem. Proceesings of the London Mathematical Society, 1937(42):230–265, 1937.

117. Viswanathan, G. M., Buldyrev, S. V., Havlin, S., and Stanley, H. E.: Quantification of DNA patchiness using long-range correlation measures. Biophysical Journal, 72:866–875, February 1997.

118. Voss, R. F.: Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. Physical Review Letters, 68:3805 – 3808, 1992.

119. Wang, X. H., Istepanian, R. S. H., Song, Y. H., and May, E. E.: Review of application of coding theory in genetic sequence analysis. In Proceedings 5th International Workshop on Enterprise Networking and Computing in Healthcare Industry, pages 5–9, June 2003.

120. Watson, J. D., Baker, T. A., Bell, S. P., Gann, A., Levine, M., and Losick, R.: Molecular Biology of the Gene. CSHL Press, 2004.

121. Williamson, B.: DNA insertions and gene structure. Nature, 270:295–297, 1977.

122. Yockey, H. P.: Information Theory and Molecular Biology. Cambridge, UK, Cambridge Univ. Press, 1992.

123. Zhou, X., Wang, X., Pal, R., Ivanov, I., Bittner, M. L., and Dougherty, E. R.: A bayesian connectivity-based approach to constructing probabilistic gene regulatory networks. Bioinformatics. to appear.

124. Zolotarev, V. M.: One-dimensional stable distributions. American Mathematical Society, 1986.