DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF
NATURAL AND APPLIED SCIENCES

AN APPLICATION OF INFORMATION THEORY FOR DNA STRUCTURE

by
Ömer DURSUN

September, 2009
İZMİR

AN APPLICATION OF INFORMATION THEORY FOR DNA STRUCTURE

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Degree of Master of Science
in Statistics Program

by
Ömer DURSUN

September, 2009
İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"AN APPLICATION OF INFORMATION THEORY FOR DNA STRUCTURE"** completed by **ÖMER DURSUN** under supervision of **Assist. Prof. Dr. ÖZLEM EGE ORUÇ** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Assist. Prof. Dr. Özlem EGE ORUÇ

Supervisor

Prof. Dr. Serkan ERYILMAZ

(Jury Member)

Assist. Prof. Dr. Emel KURUOĞLU

(Jury Member)

Prof. Dr. Cahit HELVACI
Director
Graduate School of Natural and Applied Sciences

ACKNOWLEDGMENTS

I would like to thank my advisor Assist. Prof. Dr. Özlem EGE ORUÇ, who has guided and inspired me with her close interest in all phases of this study.

I would also like to thank Prof. Dr. Serdar KURT and other valued academic staff in Dokuz Eylul University Statistics Department, who had supported me during my graduate education as well as post-graduate education and who had great contribution in my personal development.

I would like to express my sincere gratitude to my parents Erdoğan and Feyhan DURSUN, my sister Burcu and our family friend Memnune Fulden DÖNMEZ who has not withhold his moral and material support and who helped me to accomplish this task. I am indebted to them for their patience and abiding trust in me.

I would like to thank all my friends, especially Aslı SUNER with whom I treasured all precious moments of life, who has supported and encouraged me, for standing by me.

The generous material support from TÜBİTAK under the "TÜBİTAK – BİDEB 2210 Domestic Postgraduate Scholarship Program" is appreciated.

Ömer DURSUN

AN APPLICATION OF INFORMATION THEORY FOR DNA STRUCTURE

ABSTRACT

In this study, first the definitions of the concepts of entropy and information which express the measurement of the indeterminacy existing in a probabilistic system are addressed. The features of these two concepts and their variety are presented in detail, and the relation and differences between these two are touched upon. For both concepts, applications had been conducted on DNA structures belonging to human genome taken from the NCBI (National Center for Biotechnology information) website.

The results on the relative entropy (Kullback-Leibler divergence) and Bhattacharyya distance values calculated from the probability distributions, formed by taking the bases of the regions that encode proteins (exon) and that do not encode proteins (intron) of the DNA structures of the eukaryote cells, and the results on the similarity of the probability distribution of these two structures were presented contrastively. A similar study is conducted for the splice site regions where exons and introns separated, and an alternative way is proposed for estimating this region. Also, via the probability distributions of the amino acids, the Entropy, Kullback-Leibler distance and Mutual Information values is calculated and interpreted.

Key Words: Entropy, Relative Entropy (Kullback-Leibler Divergence), Joint Entropy, Conditional Entropy, Mutual Information, Molecular Biology, DNA, RNA

DNA YAPISI İÇİN BİLGİ TEORİSİ UYGULAMASI

ÖZ

Bu çalışmada, öncelikle olasılıksal bir sistemde var olan belirsizliğin ölçümünü ifade eden entropi kavramı ve bilgi kavramı tanımlarına yer verilmiştir. Bu iki kavramın özellikleri ve çeşitleri ayrıntılarıyla gösterilmiş, aralarındaki ilişki ve farklılıklara değinilmiştir. Her iki kavram için, NCBI (National Center for Biotechnology Information) internet sitesinden alınan insan genomuna ait DNA yapılarında uygulamalar yapılmıştır.

Ökaryot hücrelerin DNA yapılarında bulunan protein kodlayan bölüm (exon) ve kodlamayan bölümlerin (intron) bazları temel alınarak oluşturulan olasılık dağılımlarından hesaplanan Relative Entropy (Kullback-Leibler uzaklığı) ve Bhattacharyya uzaklığı değerleri ile bu iki yapının olasılılık dağılımlarının benzerliği hakkında sonuçlar karşılaştırılarak sunulmuştur. Exon ve intronların ayrıldığı yer olan Splice site bölgesi içinde benzer bir uygulama yapılmış ve bu bölgenin önceden belirlenebilmesi için alternatif bir yol önerilmiştir. Ayrıca, aminoasitlerin bazlarının olasılık dağılımları ile de Entropi, Kullback-Leibler uzaklığı ve Karşılıklı Bilgi (Mutual Information) değerleri hesaplanmış ve yorumlanmıştır.

Anahtar Kelimeler: Entropi, Göreli Entropi (Kullback-Leibler Uzaklığı), Koşullu Entropi, Bileşik Entropi, Karşılıklı Bilgi, Moleküler Biyoloji, DNA, RNA

CONTENTS

Page

CHAPTER ONE
INTRODUCTION

The goal of Information Theory is to examine the qualitative rules related to acquisition, transfer, computation and retention of information. Information Theory has a quite widespread area of use. The randomness in transferring information has made the use of statistical methods in the theory necessary.

Communication is the process of transferring particular messages through a channel from a source to a receiver after encoding them. Formulating the communication system as a stochastic process comprises the departure point of the Information Theory. The publication "The Mathematical Theory of Communication" by Claude Shannon in 1948, laid the foundations of the relation between the communication system and information theory. Shannon, in his publication, touched upon communication system, information theory and the concept of entropy which is the amount of information in this theory (Shannon, 1948).

The concept of entropy, a word of Greek origin, first appeared as the second law of thermodynamics. The first law of thermodynamics is about conservation of energy in the universe. According to this law, the amount of energy in the universe is constant. The second law of thermodynamics, namely entropy, states that these energies are irreversible. For instance, ice will not freeze again by itself after it melts. At this point, entropy represents the mixedupness of the order after conversion of the energy of a system.

The concept of entropy, in time, became meaningful in subjects other than thermodynamics in which it represented the consumption and conversion of energy. Ludwig Boltzmann mentioned that in statistical physics, the chaos in an organisation in which the event take place, would increase as a result of the increase in entropy. After statistical physics, the use of the concept of entropy has become common via the studies of Claude Shannon.

Shannon's use of entropy in Information Theory in 1948 and his formulating this concept for stochastic states caused the studies on this subject to increase. The formula which Shannon formed using the effectuation probabilities of events has been called "Shannon Entropy" in literature. The concept of entropy is calculated differently when it is used in different areas. This situation may be exemplified by common entropy calculation methods such as "Gibbs Entropy" in thermodynamics, "Kolmogorov-Sinai Entropy" in mathematics, and "Renyi Entropy" in Information Technology.

Today, in every area in which statistics is used and indeterminacy exists, entropy is used. Molecular biology is one of the areas where entropy is used. After the discovery of the Deoxyribonucleic Acids (DNA) and protein sequences of the living things carry important information and these sequences have important roles in the formation of some illnesses, interest in this topic has increased visibly. Many entropy studies on the DNA structure on the base level have been conducted (Chun & Wang, 2004; Herzel, Ebeling & Schmitt, 1994; Mantegna, et al 1994; Schmitt & Herzel, 1997).

When the recent studies are to be examined; it is suspected that introns, which exist in the DNA structures and which discarded during RNA translation process and separated from exons, include important information with regard to living organisms. In the studies conducted, the similarities of the distributions of exons to the distributions of introns are being investigated. In this respect, entropy and information theory are used widespreadly (Cover & Thomas, 2006).

In this study, an application of information theory to the base sequences that comprise the DNA, one of the basic notions of molecular biology, was conducted. The study comprises of five chapters.

In Chapter One, a general introduction was presented; and the history of information theory and entropy, their areas of use and their importance were touched upon. In Chapter Two, basic information on the main concepts of molecular biology

was presented in order the interpretations in the application sections to be understood more clearly. In Chapter Three, the communication system, information theory, and the concept of entropy, known as the amount of information, were put forward as the theoretical background. In Chapter Four, the application of information theory on the DNA structures was given which comprised the main topic of the study. Various entropy applications were conducted on the genes HUMGALK1A, HUMCD19A and HSALADG, which belong to human genome, and which were selected randomly from the NCBI (National Center for Biotechnology Information) website. First, the distance of the probability distributions of exons and introns of the genes aforementioned to the uniform distribution and to each other were examined using Kullback-Leibler distance (Relative Entropy) and Bhattacharyya distance. Then, the distances of the probability distributions of the bases in the "Splice Site" regions, where exons and introns of these genes separated, to the probability distributions of exons and introns were examined; and the differences between the bases in the splice site regions were tried to be investigated. After these studies, conducted by taking the bases as basis, same procedures were carried out for the probability distributions, taking the amino acids as basis. In the last part of the application section, combined entropy and mutual information calculations were performed, taking the base sequences of the genes in question as basis. In the last chapter, particular interpretations were made using the tables, formed from the results of the application, and tried to shed light for the researcher who intend to do research in this area.

CHAPTER TWO
BASIC CONCEPTS OF MOLECULAR BIOLOGY

Molecular biology, the branch of biology that investigates the events in the world of living organisms on a molecular level, examines the systems that comprise the cells, and the relations between these systems. This field of science has gained great importance recently. Especially, the developments of genetics, biochemistry and biophysics caused molecular biology to be more and more important. The examining of proteins, amino acids and enzyme structures, and shedding light on the genetic structures of the living things are in the domain of molecular biology (Gates, 2000). Figure 2.1 shows the relations between biology, biochemistry and genetics.

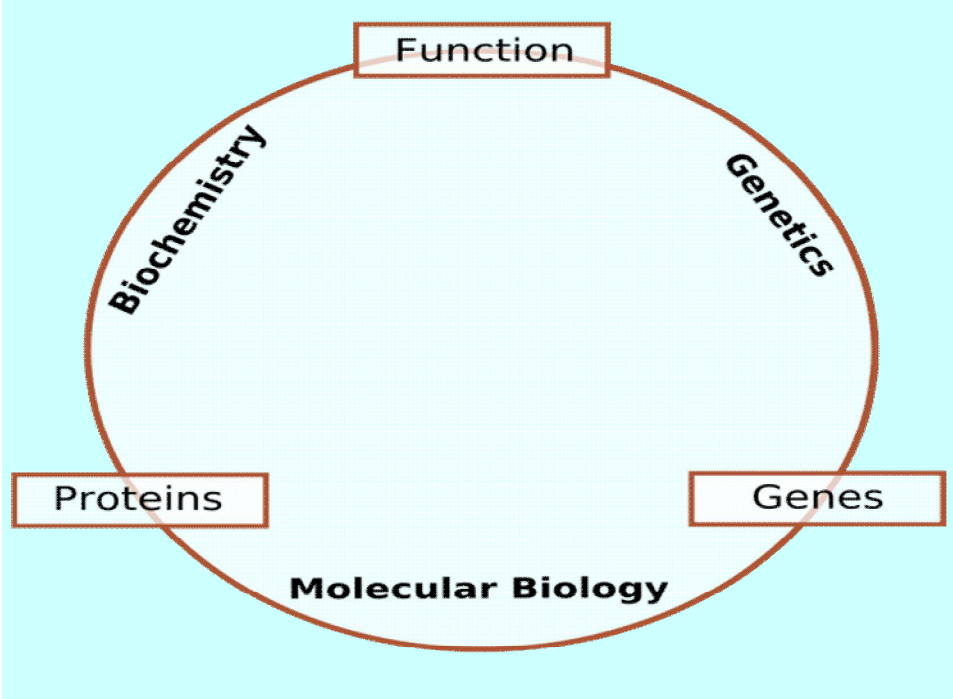

Figure 2.1 The relations between molecular biology, biochemistry and genetics

The organism of a living thing comprises of chemical substances such as proteins, carbon hydrates, fats, and DNA and RNA molecules. Each chemical structure has particular tasks in the cell structure. These tasks are performed in a particular systematic order, thus help sustaining the liveliness of the organism. In this chapter, proteins, DNA and RNA structures will be identified.

2.1 Proteins

Proteins are the most fundamental structures for the biological events in the cells of living things to occur. Proteins are formed by the combination of molecules of oxygen, hydrogen and nitrogen atoms, together with the structures called amino acids. The events occurring in the cells take place as a result of the tasks performed by particular proteins. For instance, haemoglobin proteins provide the transfer of oxygen needed by the cell and the insulin protein meet the need of sugar of our cells.

Some of the functions of proteins can be listed as below:

- Proteins are the building blocks of the cellular organs and soft tissues.
- They take part in forming new tissues.
- They function in repairing the tissues.
- They play role in transfer of neural stimuli
- They function in supporting the organism and in providing mobility.
- They help to protect the body against external factors.
- They take part in transfer of oxygen and other materials.
- They play a role in the coagulation of blood.
- They help keeping the equilibrium in the cell.

Amino acids, comprising of nucleotides, are the building blocks that form the protein. There are over 300 kinds of amino acids in the nature; but there are only 20 of them in mammals. It is possible to produce various proteins using the 20 different amino acids in the cell. The tasks of the proteins differ according to the numbers and kinds of amino acids they include. Proteins include an average of 350 amino acids. However, in the cell structure, there are small proteins comprising of 20 amino acids, as well as huge proteins that include 5000 amino acids (Gates, 2000).

The genes in the chromosomes of the living things, too, are formed of amino acids. Each of the bases entering the structure of the amino acids (Adenine, Thymine, Guanine and Cytosine) is used as a symbol for coding. The genetic code of the living

things first forms the amino acids using these four bases. These amino acids, then, form the proteins and the enzymes. Three bases sequenced successively represent a code. 64 codes can be formed using four different nucleotides. These 64 codes form the 20 amino acids. Thus, some amino acids are represented via more than one code. The changes in the sequences of these codes, causes different meanings to occur (Riyazuddin, 2006). The codes formed by the bases and the corresponding amino acids are presented in Table 2.1.

Table 2.1 Standard Genetic Code Table

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

As it is shown in the table of Standard Genetic Code, amino acids may generally be presented in a three letters or single letter way (Gates, 2000; Schmitt & Herzel, 1997). Amino acids show variety according to their type. The 20 different amino acids can be classified under four groups according to their chemical differences:

- **Positive Charged Amino Acid (Basic):**

  - Arginine Arg R
  - Histidine His H
  - Lysine Lys K

- **Negative Charged Amino Acids (Acidic):**

  - Aspertic Acid Asp D
  - Glutamic Acid Glu E

- **Polar Amino Acids:**

  - Asparagine Asn N
  - Cysteine Cys C
  - Glutamine Gln Q
  - Glycine Gly G
  - Serine Ser S
  - Threonine Thr T
  - Tyrosine Tyr Y

- **Non-Polar Amino Asitler:**

  - Alanine Ala A
  - Isoleucine Ile I
  - Leucine Leu L
  - Methionine Met M
  - Phenylalanine Phe F
  - Proline Pro P
  - Tryptophan Trp W
  - Valine Val V

2.2 Deoxyribonucleic Acid (DNA)

Deoxyribonucleic Acid (DNA) is a giant molecule that has important roles in all the vital functions of the cell which comprises of carbon, hydrogen, oxygen, nitrogen and phosphate atoms. Each gene, which is a part of the DNA molecule, controls a particular feature in the human body. The functions to sustain the liveliness of the

organism such as the body shape of the organism, the division of labour of the organs and the order of functioning of these organs, the genetic codes of the proteins to be produced in the cells, the control of the amount of the proteins to be produced (gene regulations) are planned and encoded on the DNA. Shortly, DNA is the genetic information store of living things. Genetic information is like a language. First the words are written by combining the letters in our alphabet, and then the words are combined in order to form sentences and then paragraphs and larger texts. In DNA however, the alphabet contains only four letters. Each letter represents a chemical molecule called base or nucleotide. Codons, the genetic words, are formed of these letters. As distinct from other languages, in genetic language all words (codons) are formed of three letters. These words combine and form sentences which are called genes. All sentences come together and form the book, namely the genome, which includes all the genetic information (Gates, 2000; Riyazuddin, 2006).

DNA molecules were first observed by A. F. Miescwer in the end of the 19th century. In 1953 Watson and Crick conducted studies to identify the structure of DNA. According to the results of these studies, DNA is a chain of two molecules of infinite length in theory, wrapped around each other as a double helix.

The nucleotides forming DNA comprise of three sections:

- Base: Adenine (A), Thymine (T), Guanine (G), Cytosine (C)
- Sugar (Carbohydrate with five carbon atoms)
- Phosphate Group

A base is attached to the sugar sequence attached by the phosphate bonds. This bond presented in Figure 2.2 forms one of the chains of DNA.

C		T		G		A		...
Sugar	Phosphate	Sugar	Phosphate	Sugar	Phosphate	Sugar	Phosphate	...

Figure 2.2 Base, Sugar and Phosphate bonds in DNA chain

A second sequence of DNA with the same structure is attached to the first one by hydrogen bonds that exist between particular bases in the two sequences. DNA is formed by these two chains to wrap around each other to form a helix. The DNA structure is shown in Figure 2.3.

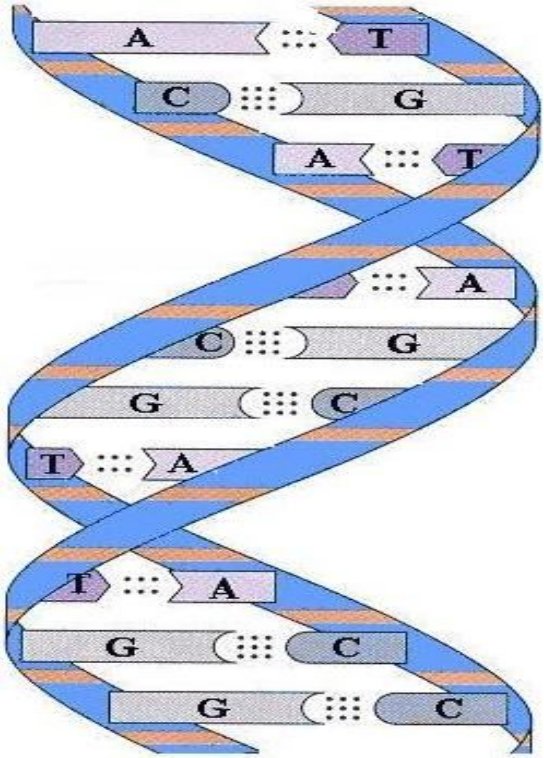

Figure 2.3 Double Helical Structure of DNA

DNA has two main functions. First of them is to copy itself during cell division. During the division of the chromosomes DNA creates a copy of itself. This copying process is called replication, put it differently, duplication. This event is necessary for the same features to occur in the new cell after cell division.

The second function of DNA is to transfer the information gathered on itself to RNA (Ribonucleic acid). This process is called transcription. Transcription is the synthesis of RNA over the DNA mould. Thus the information on DNA is transferred to RNA molecule. The information gathered on RNA is read in the ribosomes and used in protein and enzyme synthesis. This process is called “translation”. These to events are called Central Dogma (Farach, et al, 1995; McGrats, 2000).

In eukaryote cells, in DNA molecules, there are sections that include protein synthesis code (exon) and that do not include the code (intron). As proteins and enzymes are being synthesised a RNA copy sequence called mRNA is placed by taking a letter sequence of a gene in DNA as an example. This process is called Transcription. During transcription, while mRNA is being formed, the letter sequence of the gene is not read from the beginning to the end. Some section of the code is read and copied and then a long section is skipped and the reading process is then started from another section. The section in the gene that is not read is called intron. Transcription phase is when introns and exons are separated from each other. During this process the Thymine base changes its place with Urasil base. The sections of DNA that does not carry information (introns) are discarded during “translation” phase. In this phase, exons that are separated from introns are turned into amino acid and protein chains according to the Standard Genetic Code Table. Figure 2.4 shows the phases of transcription and translation. Introns occupy a great part of the total genome. The section where exons and introns separated is called “splice site”. The sections in DNA sequence that do not carry the code start with GT base pair and end with AG base pair. This situation is always like this, but not all GT base pairs are intron beginnings. Accordingly, not all AG base pairs are intron endings (Mantegna, et al, 1994; Sakharkar, Chow & Kangueane, 2004).

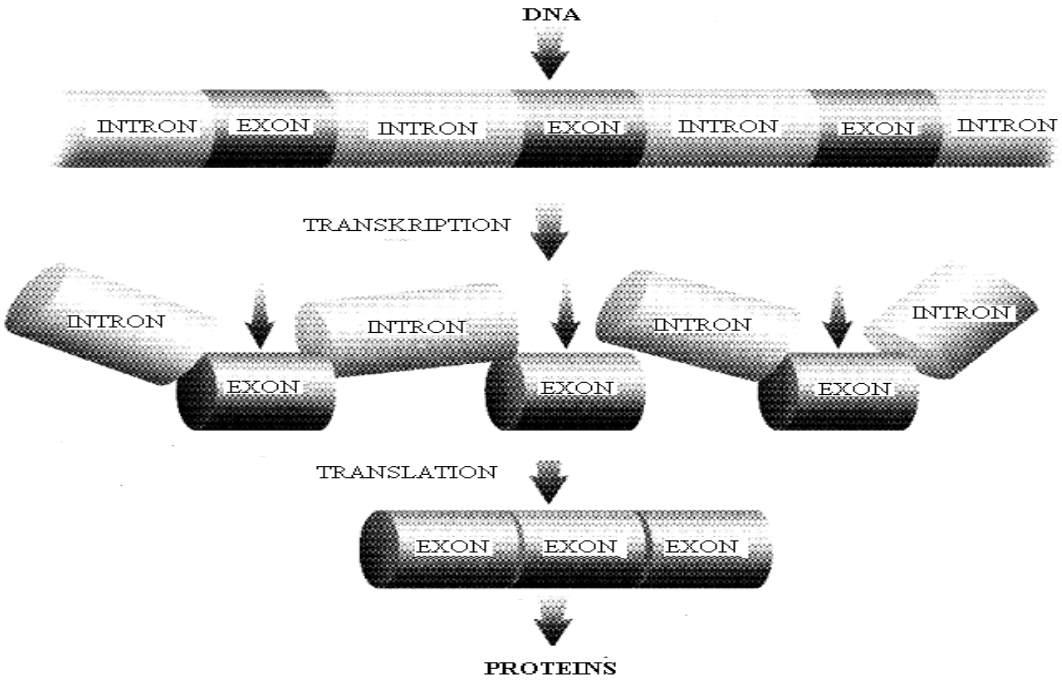

Figure 2.4 Transcription and Translation Process

DNA, by reduplicating itself, transfers the codes of life from one generation to another through germ cells. Encoding the body structures and characters of living things in a non-living molecule and this molecule's reduplicating itself are done by means of DNA molecules. Therefore, DNA takes an important task for living organisms.

2.3 Ribonucleic Acid (RNA)

Ribonucleic acid (RNA) is a nucleic acid in the form of a single helix formed by the combination of ribonucleotides successively. RNA, together with DNA functions in the cell, in protein synthesis. The length of RNA molecules is shorter than DNA molecules (Gates, 2000). Although some features of RNA molecules are similar to DNA molecules, there are some differences from them. These differences can be listed as below:

- RNA contains ribose sugar instead of deoxyribose sugar
- It is in single helix form by contrast with DNA
- It contains Urasil base instead of Thymine
- It is shorter than DNA molecules.

There are RNA molecules that are used in different tasks in the cell. In eukaryotic and prokaryotic cells, there are three types of RNA used in different tasks. These are called mRNA, rRNA and tRNA (Adami, 2004; Gates, 2000).

mRNA, also known as the messenger RNA, is a type of RNA that functions as a mould in transferring the hereditary information stored in the DNA to the protein structure. mRNA is synthesised by the RNA polymerase enzyme in the nucleus through a single chain of DNA and then it separates from the nucleus and attaches on ribosomes. It determines the amino acid sequence of the protein to be synthesised according to the genetic information it obtained from DNA. Each RNA molecule shows conjugation with a section on the DNA, namely the gene, mRNA plays an important role in protein synthesis (Adami, 2004; Gates, 2000).

rRNA, known as the ribosomal RNA, is a type of RNA which is a part of ribosomes. It constitutes 65% of the ribosome weight. It has important roles in ribosomes' structures and functions. As rRNA is preserved in every living organism, the evolutional relations between living things can be calculated by analysing the nucleotide sequences.

tRNA, know as the transferring RNA, is the type of RNA which functions in translation process. It is a single helix form as mRNA, but it is smaller as a molecule than mRNA. This RNA type fulfils the functions of selection and transfer. For each of the 20 amino acids there is one corresponding tRNA molecule. The amino acid molecules synthesised in the cell are found by the corresponding tRNA, and their free ends of the tRNA molecules are attached to these amino acids. tRNA molecules, align the amino acids on the polypeptide chain according to the codon carried by the mRNA. tRNA molecules, by their ends, called anti-codon, comprising of three bases,

attach to the codon section temporarily and provide the amino acids to be aligned correctly according to the code on the mRNA. There may be more than one tRNA molecule for each amino acid. The anti-codon sections of these molecules provide the identification of codon sections of mRNAs and thus they provide the translation of the RNA code into protein code.

CHAPTER THREE
BASIC CONCEPTS OF INFORMATION THEORY

The goal of Information Theory is to examine the qualitative rules related to acquisition (obtaining the message), transfer, computation and retention of information. In this section the basic concepts of information theory the concept of entropy, which is defined as the measure of variety on probability distributions and the concept of information will be indentified.

3.1 Information Theory

Information Theory had emerged during the analysis of problems related to the telecommunications in the 1940's. Improving the quality of communication and the encoding used during communication were the main concerns in those years. For this purpose, Shannon aimed at building a mathematical model of communication. Thus, the foundation of Information Theory had been laid. The study "A Mathematical Theory of Communication" by Shannon is the starting point of Information Theory (Giriftinoğlu, 2005; Yolaşan, 2005).

Communication is the process of transferring particular messages through a channel from a source to a receiver after encoding them. Formulating the communication system as a stochastic process comprises the departure point of the Information Theory. Information Theory deals with topics such as the amount of information of the message produced by the Source, the maximum amount of information that the channel can transmit, the correction of the errors occurring during communication, and encoding for a more efficient communication.

Information Theory has a widespread use in many fields of science. There are common areas of research of information theory and physics, mathematics, statistics, computer engineering, etc... Figure 3.1 shows the areas that are in relation with information theory and some common areas of research. According to this figure,

some of the common areas of research for information theory and statistics can be exemplified as Hypothesis Thesis and Fisher Informatics (Cover & Thomas, 2006).

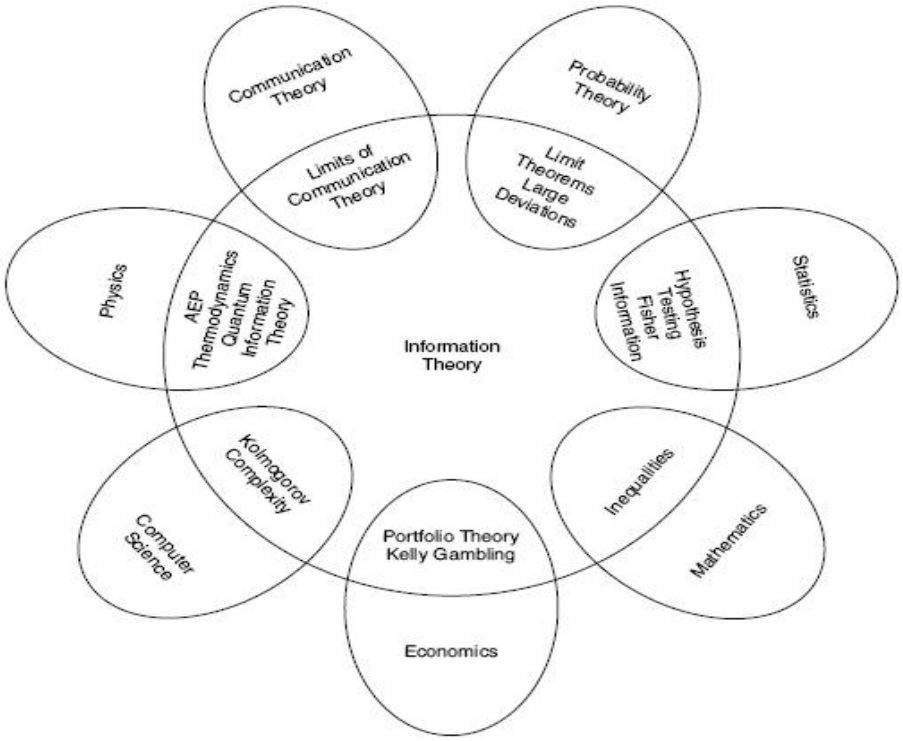

Figure 3.1 Interest Areas of Information Theory

It has been mentioned before that information is a concept that can be measured. When the occurrence probabilities of each event of finite number are defined, it is possible to calculate the amount of information manifested by the help of these probabilities. This amount is expressed via the concept of entropy in Information Theory. After briefly touching upon Information Theory, the concept of entropy will be dealt with.

3.2 Entropy

Entropy is a word of Greek origins with a lexical meaning of indeterminacy. Entropy, with its coarsest definition, is the measure of indeterminacy of a particular system. There are at least three ways, thermodynamics, statistical physics theory and

information theory, to define entropy. As only the definition of entropy within information theory was made use of in this study, other definitions were left out (Cover & Thomas, 2006).

Shannon argues that, the process of transmitting the message that has been produced in the source is one of a probabilistic one in information theory. Acquiring information about some particular event is valid only if there is indeterminacy on that event. The required information for a system's probable states accurately equals to the entropy of that system. By this approach entropy can be defined as the expected value of the states that an event can take.

The entropy value of the discrete random variable X which can take $x_1, x_2, \ldots, x_n$ values with $p_i=P(X=x_i)$ $\left[p_i \geq 0, i = 1,2,3,...,n \text{ and } \sum_{i=1}^{n} p_i = 1 \right]$ probability is calculated by the equation;

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i) = -E(\log_2(x_i)) \quad (1)$$

Entropy is calculated with the formula, if the random variable X is a continuous variable;

$$H(X) = -\int f(x)\log_2 f(x)dx = -E(\log_2 f(x)) \quad (2)$$

An example for calculating entropy is given below;

Assume that asking a person to think of a number between 1 and 16 and let's try to reach the correct answer by requesting him/her to answer the question we ask by saying just "yes/no" (binary answer). Let's calculate the entropy value as below;

$$H(X) = -\sum_{i=1}^{16} p(x_i)\log_2 p(x_i)$$

$$H(X) = -\sum_{i=1}^{16} \frac{1}{16}\log_2(1/16) + \frac{1}{16}\log_2(1/16).... .. + \frac{1}{16}\log_2(1/16)$$

$$= -16(\frac{1}{16}\log_2 2^{-4})$$

$$= 4 \text{ bit.}$$

This means; that it can be obtained the correct answer by asking an average of four questions. As it can be understood from this example, as a result of an event with p=1/16 probability occur an information of $\left(-\log_2 \frac{1}{16}\right)$ has formed.

In our example, the answers in found are in bits. The only reason for this is our own choice. We can calculate the result by taking the logarithm base to a different radices different than 2, and we can get different values of different units. These are called;

- bits, if the logarithm is taken to the base 2 (binary) ,
- trits, if the logarithm is taken to the base 3 (trinary),
- nats, if the logarithm is taken to the base e (natural logarithm),
- hartleys, if the logarithm is taken to the base 10.

The easiest logarithm value is $\log_2(p)$ in entropy calculations. Therefore, this base is preferred in the literature when entropy (information value) is calculated (Cover & Thomas, 2006). The below inferences can be made about entropy according to the occurrence probabilities of different states;

- No information occurs with a state, having a occurrence probability 1, to arise. In this case, the entropy value is zero. For instance; we always know that the resulting number will be “4” when we throw a tricky dice with each side having the value “4”. The result of the throw does not make any difference in our knowledge.

- The entropy value of a system increases as the probable states of that systems increase.

- Occurrence of a state with lower occurrence probability accumulates more amount of information than the one with higher probability. For instance, instead of knowing that the result of flipping a coin will be heads or tails, knowing all six numbers in 6/49 lottery, which has a lower estimation probability, will contain more information.

- As entropy of a system increases, the estimation or knowing of the results beforehand gets more difficult. The power of estimation will decrease since indeterminacy increases.

Entropy has the following features;

- Nonstorage
- Static
- Statistically independent
- Constant
- Symmetrical
- Summable

Entropy is at maximum where all probabilities are equal. The entropy graph of an event with two possible equal results $(p_1 = p_2 = 0.5)$ is given in Figure 3.2 (Cover & Thomas, 2006).

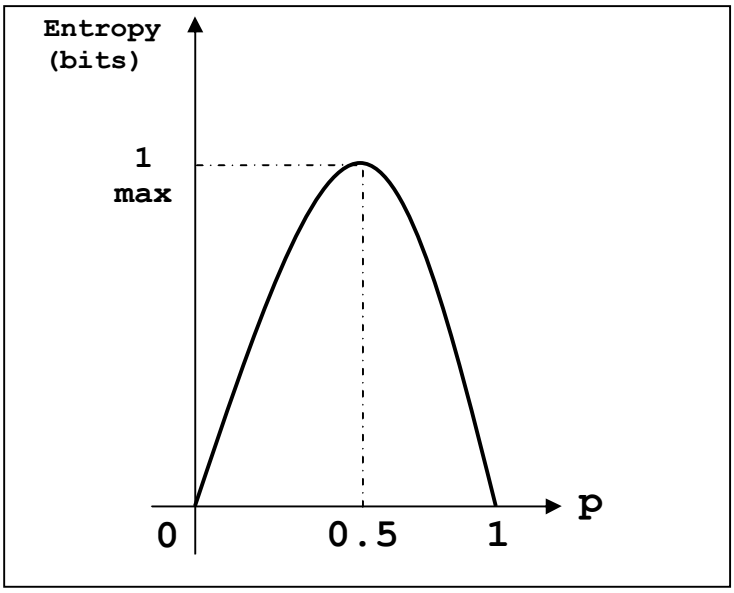

Figure 3.2 Maximum Entropy in case of Equal Probability

Various entropies can be calculated for the states under interest. In this section some explanations on some entropy types will be given.

3.2.1 Joint Entropy

Let X and Y be two discrete random variables taking values $\{x_1,\ldots,x_n\}$ and $\{y_1,\ldots,y_n\}$, respectively. If P(X=x, Y=y) denote the joint probability mass function of X and, then the joint entropy of these random variables is defined by;

$$H(X,Y) = -\sum_{j}\sum_{i} p(X = x_i, Y = y_j)\log p(X = x_i, Y = y_j) \qquad (3)$$

If X and Y are continuous and have the joint probability density function f (x,y), then;

$$H(X,Y) = -\int\int f(x,y)\log f(x,y)dxdy \qquad (4)$$

Joint entropy is also called the common information measure.

If X and Y are independent, then the joint entropy equals to the sum of the entropies of each random variable.

$$H(X, Y) = H(X) + H(Y) \tag{5}$$

Let X and Y be two random variables taking values X:{1, 2, 3, 4} and Y:{1, 2, 3, 4}. The marginal distribution of X is {0.5, 0.25, 0.125, 0.125} and the marginal distribution of Y is {0.25, 0.25, 0.25, 0.25} hence X and Y have the joint distribution;

Y \ X	1	2	3	4
1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
4	$\frac{1}{4}$	0	0	0

The joint entropy is calculated from this joint distribution. $H(X,Y) = \frac{27}{8}$ bit.

3.2.2 Conditional Entropy

Let the random variables X and Y are random variables that have joint probability distributions. When the values of the random variable Y are given, the measurement of the indeterminacy in the random variable X is the conditional entropy of X dependent on Y. Knowing Y always decreases the indeterminacy of X. It is shown as H(X|Y) and can be calculated as follows (Cover & Thomas, 2006).

For discrete random variables;

$$H(X|Y) = -\sum\sum p(x_i, y_j) \log\left(p(x_i|y_j)\right) \tag{6}$$

For continuous random variables**;**

$$H(X|Y)=-\int\int f(x,y)\log\big(f(x|y)\big)dxdy \tag{7}$$

If the variables X and Y are independent of each other, the chain rule that shows the combination of the joint entropy and conditional entropy explained above is given in Eq. 8.

$$H(X,Y)=H(X)+H(Y\mid X) \tag{8}$$

This rule can be applied for the joint entropy of the variables of X and Y if the value of the third variable Z is known.

$$H(X,Y\mid Z)=H(X\mid Z)+H(Y\mid X,Z) \tag{9}$$

Let X and Y be two random variables taking values X:{1, 2, 3, 4} and Y:{1, 2, 3, 4}. The marginal distribution of X is {0.5, 0.25, 0.125, 0.125} and the marginal distribution of Y is {0.25, 0.25, 0.25, 0.25}. And hence $H(X)=\frac{7}{4}$ bits and $H(Y)=2$ bits. Also;

$$H(X|Y)=\sum_{i=1}^{4}p(Y=i)H(X\mid Y=i)$$

$$H(X|Y)=\frac{11}{8}bits$$

3.2.3 Relative Entropy

The Kullback-Leibler divergence ($KL = D(p||q)$) is a non-commutative measure of the divergence between two probability distributions p and q (Kullback, 1987). *KL* is also sometimes called the information gain about X if p is used instead of q. It is also called the relative entropy in using q in the place of p. The relative entropy is an

appropriate measure of the similarity of the underlying distribution. It may be calculated as given in Eq. 10 and 11.

$$D(p||q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \tag{10}$$

$$D(p||q) = \int_{x \in X} f(x) \log \frac{f(x)}{g(x)} dx \tag{11}$$

The properties of the relative entropy equation make it non-negative, non-symmetric and it is zero if both distributions are equivalent namely $p = q$. The smaller the relative entropy is the more similar the distribution of the two variables and vice versa (Kullback, 1987; Leutenneger, 2000).

Let X={0,1} and consider two distributions p and q on X. Let $p(0) = 1 - r$, $p(1) = r$ and let $q(0) = 1 - s$, $q(1) = s$. Then;

$$D(p||q) = (1-r) \log \frac{1-r}{1-s} + r \log \frac{r}{s}$$

If r = s, then $D(p||q) = D(q||p) = 0$, If $r = \frac{1}{2}, s = \frac{1}{4}$, the relative entropy calculated as follows;

$$D(p||q) = \frac{1}{2} \log \frac{\frac{1}{2}}{\frac{3}{4}} + \frac{1}{2} \log \frac{\frac{1}{2}}{\frac{1}{4}} = 0.2075 bit$$

and whereas;

$$D(q||p) = \frac{3}{4} \log \frac{\frac{3}{4}}{\frac{1}{2}} + \frac{1}{4} \log \frac{\frac{1}{4}}{\frac{1}{2}} = 0.1887 bit$$

Note that $D(p||q) \neq D(q||p)$ in general.

3.3 Bhattacharyya Distance

In the application section of our study the Bhattacharyya distance which is a metric distance value was calculated as well as the non-metric Kullback-Leibler distance. In this section the Bhattacharyya distance will be touched upon briefly.

Bhattacharyya distance is the measure that shows the similarities of two different probability distributions (Bhattacharyya, 1943). This measure is also used to classify different groups. Bhattacharyya distance is calculates as below for two different distributions.

$$D_B(p,q) = -\ln(BC(p,q)) \tag{12}$$

In the formula above, BC(p,q) is expressed as the Bhattacharyya coefficient. This coefficient takes values between 0 and 1. For this reason Bhattacharyya, too, is non-negative. For discrete distribution BC(p,q) is calculates as Eq. 13;

$$BC(p,q) = \sum_{x \in X} \sqrt{p(x)q(x)} \tag{13}$$

For continuous distribution BC(p,q) is calculated as;

$$BC(p,q) = \int \sqrt{p(x)q(x)}dx \tag{14}$$

3.4 Concept of Information and Its Features

In this section the concept of information used in the framework of information theory and its features will be briefly explained. Information is the processed state of data and facts related to objects, events or people.

3.4.1 Concept of Information

Information is a complex concept, which gains various meanings according to the content in question and the perspective taken. Therefore, it is hard to provide a general definition. The development of researchers is devoted to finding the key to the problem. Researchers come to common grounds in this point and develop mathematical terms for information systems analysis. Information is storable, visible, transferable, re-obtainable, observable and interpretable The randomness in the transfer of information requires the use of statistical methods in examining these processes (Adami, 2004).

The value of;

$$I(x_i) = -\log_2(p_i) = \log_2(1/p_i) \tag{15}$$

calculated for the $\{x_1, x_2, ..., x_i\}$ values of the discrete random variable X in the state of $i = 1, 2, ..., n$, is called the information content of x_i state. The information value of the random variable X is calculated as below:

$$I(x_i) = \sum_i p_i I(x_i) \tag{16}$$

This value is the weighted average of the information contents of the values that X has taken, and the probability of taking these values; and at the same time it is called entropy. The information content that the random variable takes is only dependent on the random variable's probability of the taking that value. As lower this probability is so the bigger is the information content.

There are four basic axioms for information:

- Information is a value that is not negative.

$$I(p) \geq 0$$

- The information value of an accurate event is zero.

$$I(1) = 0$$

- For two independent event, the information is obtained from observations equals to the sum of two informations.

$$I(p_1 * p_2) = I(p_1) + I(p_2)$$

- I(p) is monotonous and constant.

3.4.2 Mutual Information

In Information Theory, knowing the mutual information of two variables is great importance. Mutual information is the amount of information that a random variable contains about other random variable. According to this definition, the mutual information between the random variables X and Y can be found by comparing entropy of X, which is the amount of information about X; and entropy of X while the value of Y is given. Mutual information is always equal to or greater than zero (Cover & Thomas, 2006).

In case $i = 1, 2, ..., n$ and $j = 1, 2, ..., m$, the mutual information value is calculated as follows.

$$I(x_i, y_j) = \log(P(x_i \mid y_j) \mid P(x_i)) \tag{17}$$

Essential features of mutual information are presented below:

- If X and Y are independent random variables; $P(X;Y) = P(X)P(Y)$ therefore $I(X;Y) = 0$.
- If random variable X is not independent from random variable Y; $H(X \mid Y) = 0$ and $I(X;Y) = H(X)$.

- $I(X; Y) = I(Y; X)$ is a symmetrical function.

- $I(X;Y) = H(X) + H(Y) - H(X,Y)$
  $I(X;Y) = H(X) - H(X|Y)$
  $I(X;Y) = H(Y) - H(Y|X)$

- $I(X;Y) \geq 0$

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

3.5 Relationship between Entropy and Information

The definition of information is done by taking the definition of entropy, which measures the randomness in a system, as a model. Therefore, information and entropy may be considered as mingled concepts. When a question with yes or no as an answer: accuracy is not in question, there is in determinacy for the answer. Here the question carries an information value. If the answer is known accurately, asking the question will be unnecessary. For instance, when the kick-off of a match is seen by everyone, declaring this fact will not carry an information value for other people. Therefore, it can be argued that "information is the source that decreases the indeterminacy about the subject". Increase of information causes entropy to decrease by decreasing indeterminacy. Thus, minimum indeterminacy is obtained by maximum information.

$$I(X; Y) = \log (P(X \mid Y)) + I(X) = H(X) - H(X \mid Y)$$
$$H(X,Y) = H(X) + H(Y|X)$$
$$H(X,Y) = H(Y) + H(X|Y)$$

The relations between the conditional entropies and joint entropies of the random variables X and Y can be defined as above (Cover & Thomas, 2006).

Information is one of the main components of the communication process. The most important feature of information, which has a very important place in the theory, is its being measurable. We may think that an information transfer process toward the destination starts as the information source gives out one of his/her finite numbered states he owned (message). If we define the probabilities of occurrence (giving out) of each finite state of the source, we can calculate the amount of information by the help of these probabilities.

CHAPTER FOUR
APPLICATION

The data sets pertaining to the human genome genes used in this study are obtained via the web sites of the NCBI (National Center for Biotechnology Information) and BDGP (Berkeley Drosophila Genome Project). The human genes HUMGALK1A, HUMCD19A and HSALADG are used as data sets. The Human Galactokinase (HUMGALK1A) gene extracted from the 17th chromosome of the human genome comprises of 8 exons and 7 introns. The Human CD19 (HUMCD19A) gene extracted from the 16th chromosome is composed of 14 exons and 13 introns. As for the Homosapiens ALAD (HSALADG) gene, it has 14 exons and 13 introns.

This study consists of two separate applications concerning the ways of using information theory in investigating the DNA structure. The purpose of the first application is to show that the probability distributions of the exons and the introns of genes are the same and to emphasize that introns also include information as exons do, and to conduct an analysis for the splice site regions of exons and introns, considering that exons always begin with a GT base pair. The second application aims at providing an example for information theory on the amino acid sequences in the genes. The Kullback-Leibler and Bhattacharyya distances are used in order to measure the similarity among the probability distributions that are obtained using the bases in the exons, introns and the amino acids of the genes. Moreover, various entropy values are computed from the probability distributions. Also, the distance of the probability distributions of the base sequences of the amino acids in the exons and introns to the uniform distribution, where each base has the same chance to be seen and the entropy is at the maximum value, is examined by calculating the positional Relative entropy values. Besides, some interpretations on the randomness in the sequences are made with respect to the distances.

The application of this study comprises of two sections. In the first section of the application, the results of the analysis, on exon-intron structures of the three genes belonging to human genome which we used as data set, will be presented. In the second section, though, the results of the analysis on the amino acid structures belonging to these three genes will be presented.

4.1 Results of Exon and Intron Structure

In this paper, in order to have knowledge of the base lengths of the exons and introns in the genes that are under examination, the descriptive statistics are calculated. The resulting values are presented in Table 4.1. These values show that, for the genes under examination, the introns are longer than the exons and the variations in the exon base lengths are smaller compared to the introns'.

Table 4.1 Descriptive statistics for exons and introns

	MEAN		STDEV	
GENES	**EXON**	**INTRONS**	**EXONS**	**INTRONS**
HUMGALK1A	147,375	843,857	38,067	1468,331
HUMCD19A	119,357	407,153	74,635	513,093
HSALADG	90,272	426,100	27,935	374,668

In the first application, the distances between the uniform distribution and the probability distributions obtained from the bases in the each exon of the three genes under inspection are examined separately. According to the results presented in Table 4.2, it is observed that the most distant exons of the eight exons in the HUMGALK1A gene are the first and the sixth ones. Within the HUMCD19A gene, it is observed that the most distant exons from the uniform distribution are the sixth and the eighth one of 14 exons. In the HSALADG gene, the first and the fifth ones of 14 exons are monitored as having the most distant probability distribution from the uniform distribution. In each three genes examined in this study, the probability distributions of the base sequences of the first exons come off as distant from the uniform distribution. This shows that, the randomness within the base sequences in the first exons of the genes is less compared to other exons. This randomness,

comparingly being less, indicates that these exons have lower entropy values than others. Put another way, this shows that the sequences within the first exons are more easily predictable.

Table 4.2 Distance from the uniform distribution of every exons distribution

EXONS	K-L DIVERGENCE		
	HUMGALK1A	HUMCD19A	HSALADG
EXON 1	0.117250121	0.063496262	0.133480575
EXON 2	0.061215271	0.038485734	0.020604997
EXON 3	0.094559815	0.058549359	0.048171142
EXON 4	0.032627667	0.033028449	0.041251677
EXON 5	0.059138844	0.057697685	0.095835948
EXON 6	0.128324164	0.106080273	0.054035293
EXON 7	0.087534554	0.092462677	0.056512438
EXON 8	0.028026239	0.176920683	0.075973206
EXON 9	-	0.073596307	0.082013944
EXON 10	-	0.032008555	0.051441677
EXON 11	-	0.050367251	0.013437345
EXON 12	-	0.094410013	-
EXON 13	-	0.028781143	-
EXON 14	-	0.070509481	-
TOTAL	0.068914955	0.035600501	0.025785072

When the introns are examined with the same analysis, the values obtained are presented in Table 4.3, it is observed that the most distant introns from the uniform distribution are the sixth, the seventh and the fourth of the 7 introns within the HUMGALK1A gene. Within the HUMCD19 gene, the seventh, the ninth and the thirteenth introns are monitored as the most distant of the 13 introns from the uniform distribution. Within the HSALADG gene, the fourth and the eighth introns are observed as having the most distant probability distributions from the uniform distribution. In each of the three genes, the distance - from the uniform distribution -, coming off from different introns, indicates that a generalisation cannot be made as it can be with the exons.

Table 4.3 Distance from the uniform distribution of every introns distribution

INTRONS	K-L DIVERGENCE		
	HUMGALK1A	HUMCD19A	HSALADG
INTRON 1	0.067691744	0.014564122	0.002178067
INTRON 2	0.011122326	0.052822401	0.028066403
INTRON 3	0.067169914	0.052032269	0.012038870
INTRON 4	0.140393540	0.024313423	0.097834239
INTRON 5	0.015542284	0.055113999	0.085191660
INTRON 6	0.183141927	0.047315644	0.019199807
INTRON 7	0.220744700	0.108425144	0.058101863
INTRON 8	-	0.072933480	0.104100615
INTRON 9	-	0.164555834	0.022360678
INTRON 10	-	0.017970836	0.000015407
INTRON 11	-	0.051355752	-
INTRON 12	-	0.063764765	-
INTRON 13	-	0.188886978	-
TOTAL	0.017947962	0.029262792	0.004980854

The distances of probability distribution of the amino-acid sequences to the uniform distribution are analysed by calculating the positional entropy values. The values obtained are presented in Table 4.4. According to these values, it is observed that the probability distributions of the base sequences on the third raw of the amino-acid sequences are the most distant ones to the uniform distribution within the exons of the three genes. Therefore, the bases with the lowest entropy values are those on the third row. Put another way, to estimate the sequencing of the bases on the third row of the amino-acids in the exons within the genes is easier compared to the other bases.

Table 4.4 Positional relative entropies of exons of HSALADG, HUMCD19A, HUMGALK1A genes

POSITIONS	K-L DIVERGENCE		
	HUMGALK1A	HUMCD19A	HSALADG
FIRST	0.125890400	0.045496858	0.083164998
SECOND	0.005349354	0.004222578	0.012030802
THIRD	0.233643492	0.133183485	0.100216763
TOTAL	0.068914955	0.035600501	0.025785072

When the same analysis is performed on the introns within the three genes, the results obtained are presented in Table 4.5, it is seen that the probability distributions of the base sequences in the introns' amino-acid sequences are rather close to the uniform distribution. This situation indicates that it is very hard to estimate the probability distribution of the base sequences in amino-acids in the introns.

Table 4.5 Positional relative entropies of introns of HSALADG, HUMCD19A, HUMGALK1A genes

POSITIONS	K-L DIVERGENCE		
	HUMGALK1A	HUMCD19A	HSALADG
FIRST	0.022305415	0.03608056	0.005256144
SECOND	0.011689591	0.027389099	0.004140420
THIRD	0.022069205	0.025926105	0.006069155
TOTAL	0.017947962	0.029262792	0.004980854

The distances between the uniform distribution and the probability distributions obtained from the bases in the exons and introns of the three genes under inspection are examined separately. According to the results presented in Table 4.6, the probability distributions of the exons and introns in each gene are proximate to the uniform distribution. The similarity to the uniform distribution shows that the bases in the exons and introns occur in equal probability which indicates that the sequences are random.

Table 4.6 Distance from the uniform distribution of exons and introns distribution

		EXON&INTRON vs UNIFORM	
GENES		K-L DIVERGENCE	BHATTACHARYYA
HUMGALK1A	EXON	0.068914955	0.012187062
	INTRON	0.017947962	0.003159570
HUMCD19A	EXON	0.035600501	0.006244280
	INTRON	0.029262792	0.005045766
HSALADG	EXON	0.025785072	0.004502606
	INTRON	0.004980854	0.000873460

The similarities between the probability distributions of the exons and introns, grounded on the bases, are examined via the calculation of the Kullback-Leibler and Bhattacharyya distance values. The values obtained are presented in Table 4.7. These

values are close to zero in each of the genes. This closeness shows that the probability distributions of exons and introns are similar to each other.

Table 4.7 Distance between the distribution of exons and distribution of introns

	EXON vs INTRON	
GENES	**K-L DIVERGENCE**	**BHATTACHARYYA**
HUMGALK1A	0.036551825	0.006435956
HUMCD19A	0.037031963	0.006403048
HSALADG	0.012455115	0.002192561

In the last part of the first application, an analysis is carried out for the splice site regions of exons and introns, considering that introns always begin with a GT base pair. It is known that introns always begin with a GT base pair and end with an AG base pair. We observed this condition in the splice site regions of exons and introns. When the protein sequence is examined, each GT base pair observed is not always an intron beginning. Similarly, each AG base pair observed is not always an intron ending. In various studies on determining how introns begin, it is seen that the last bases of the exon before the GT pair and the first bases of the intron after, is important. As we examine the probability distributions of the sequences of the nine last bases of the exon before the GT pair and the sequences of the first nine bases of the intron after the GT pair in the genes in question, it is observed that exons ending before introns beginning with a GT pair most probably end with a Guanine base. Using Kullback- Leibler and Bhattacharyya distance scale one may deduce that the probability distributions of the splice site region bases of exons and introns are different from the probability distributions of the bases in exons and introns in the whole sequence. Table 4.8 demonstrates this result. Information on splice site region can be obtained by analyzing the probability distributions of the last nine bases of exons before the GT pair and of the first nine bases of introns after the GT pair.

Table 4.8 Distance from the splice site region bases distributions of exons to all exons distributions

		EXON&INTRON	
GENES		**K-L DIVERGENCE**	**BHATTACHARYYA**
HUMGALK1A	**EXON**	0.041399401	0.007148660
	INTRON	0.160774519	0.028743930
HUMCD19A	**EXON**	0.131393025	0.025274692
	INTRON	0.133212327	0.023236735
HSALADG	**EXON**	0.084580981	0.015868723
	INTRON	0.127826935	0.024140222

4.2 Results of Amino acid Structure

For the second application, we acquired the probability distributions of the amino acids belonging to the three genes under examination. The results obtained are presented in Table 4.9. The entropy values calculated using these probability distributions are 4.119339 bits for HSALADG, 4.071260 bits for HUMCD19A and 3.992027 bits for HUMGALK1A. Because the HUMGALK1A gene has the smallest entropy value according to the values found, it can be said that the estimation of the amino acids of this gene is easier.

Table 4.9 Probability distributions of amino acid in HSALADG, HUMCD19A, HUMGALK1A genes

AMINO ACID	HSALADG	HUMCD19A	HUMGALK1A
Phenylalanine (F)	0.03371	0.02551	0.02657
Leucine (L)	0.09831	0.10714	0.11353
Isoleucine (I)	0.03371	0.01361	0.02174
Methionine (M)	0.03090	0.02381	0.02174
Valine (V)	0.07022	0.04252	0.07488
Serine (S)	0.05618	0.09184	0.07488
Proline (P)	0.06461	0.09864	0.05556
Threonine (T)	0.06461	0.06803	0.06522
Alanine (A)	0.11236	0.04592	0.11594
Tyrosine (Y)	0.03371	0.02381	0.02657
STOP	0.00281	0.00170	0.00242
Histidine (H)	0.02528	0.01531	0.02174
Glutamine (Q)	0.02809	0.03741	0.04831
Asparagine (N)	0.01404	0.03061	0.01208
Lysine (K)	0.03371	0.02891	0.01691
Aspartic Acid (D)	0.05056	0.05272	0.03140
Glutamic Acid(E)	0.08708	0.09524	0.09420
Cysteine (C)	0.02247	0.01361	0.01932
Tryptophan (W)	0.00843	0.02891	0.00242
Arginine (R)	0.06461	0.05442	0.07488
Glycine (G)	0.06461	0.10034	0.07971
Total	1	1	1

The entropy value decreases when new information is added. We checked this condition for the genes we analyzed in our application. The results obtained are presented in Table 4.10. The entropy value that calculated when we do not know the first base of the amino acids can be seen in the Genes row of the table. The values when the first base is known are shown in the other rows. Any additional base information leads to a decrease in the entropy value. When the additional information confirms the realization of the amino acid, the entropy value is found zero as expected. We may implement these calculations for other base sequences. Among the three genes under examination, HUMGALK1A is the easiest one to estimate the amino acids when new information is added.

Table 4.10 Entropy values for additional information

ENTROPY	HSALADG	HUMCD19A	HUMGALK1A
GENES	4.1193386	4.07126012	3.99202660
Initial T	2.5185482	2.40382757	2.36292184
Initial C	2.1732272	2.01671637	2.16791716
Initial A	2.7419248	2.74608344	2.69412214
Initial G	2.2589254	2.24306345	2.22170036

The similarity between the probability distributions based on the amino acids comprising the three genes in our application is analyzed by computing Kullback-Leibler value. According to the results presented in Table 4.11, the genes where the amino acid distributions are the furthest are HUMCD19A and HUMGALK1A. It is seen that the amino acid distributions of HUMGALK1A and HSALADG are very similar. Therefore the amino acid sequences of genes with similar distribution may be estimated easily by examining other genes' distributions.

Table 4.11 Relative entropy values for HSALADG, HUMCD19A and HUMGALK1A genes.

GENES	HSALADG	HUMCD19A	HUMGALK1A
HSALADG	0	0.143158427	0.046260314
HUMCD19A	0.148137697	0	0.156046159
HUMGALK1A	0.050428046	0.183402414	0

The joint entropy values of all genes and all base position are calculated separately from the joint probability distribution. The joint entropy values are given in Table 4.12. The same result is also valid for the other variables.

Table 4.12 Joint entropy for amino acid variables and adenine, cytosine, guanine, thymine base position variables for three genes

AMINO ACID	JOINT ENTOPY			
	ADENINE	GUANINE	THYMINE	CYTOSINE
HSALADG	4.371859	4.094576	4.152196	4.283639
HUMCD19A	4.298459	4.168234	4.028408	4.193326
HUMGALK1A	4.206562	4.083357	3.867612	4.267013

The consequence is the joint entropy: $H(X; Y) = 4.371859$ where X is Adenine is and Y is HSALADG. It shows how much entropy is contained in a joint system of Adenine position and HSALADG amino acid. In probability theory and information theory, the mutual information, or transformation, of two random variables is a quantity that measures the mutual dependence of the two variables. Intuitively, mutual information measures the information that X and Y share: It measures how much we know about one of these variables and hence reduces the uncertainty about the other. For example, if X and Y are independent, then knowing X does not give any information about Y and vice versa, so their mutual information is zero. In this study, the mutual information value calculated for the Urasil base position- HSALADG amino acid, can be interpreted as follows. Those two variables seem to have a lot of information in common, 0.899653 bits of information. The mutual information values also found for the Urasil base position- HUMCD19A amino acid variables and the Urasil base position- HUMGALK1A amino acid variables are interpreted in the same way. It was observed that the variable of the Urasil base position, among the mutual information values obtained in this work, was able to restrict the uncertainty on other variables so much. Table 4.13 exhibits the shared information between pairs of all base and amino acid variables. The pair sharing the most information is Adenine base position - HSALADG, while the least is Cytosine base position - HUMCD19A amino acid variables.

Table 4.13 Mutual information for amino acid variables and adenine, cytosine, guanine, thymine base position variables for three genes

	MUTUAL INFORMATION			
AMINO ACID	**ADENINE**	**GUANINE**	**THYMINE**	**CYTOSINE**
HSALADG	1.126493	0.893481	0.899653	0.765229
HUMCD19A	0.985267	0.794084	1.009066	0.717221
HUMGALK1A	1.040720	0.784268	0.908829	0.773480

CHAPTER FIVE
CONCLUSIONS

By courtesy of studies in Information Theory, put forward since second half of the 20th century, the communication devices we use now have been attained. With the technology, improving by the second half of the 20th century, we have seen that living cells contain telecommunication techniques on a much more developed level. In the DNA of living organisms, there is an "information bank" that describes all the physical details of the body. Moreover, there is a system that reads, interprets this information and that makes productions according to this information. In all living organisms' cells, the information contained in DNA is "read" by various enzymes and proteins are synthesises according to this information. The production of millions of proteins each second, for the necessary place and in the necessary types is realised by this system. The system's containing this much information attracted the interest of researchers working in the field of information theory, and the number of studies on this subject has recently increased. In this study an application of Information Theory on DNA has been conducted.

Initially, the probability distributions of the bases in exons and introns of three genes belonging to human genome are examined. As a result, it is observed that the base sequences of both exons and introns are equally random and it is found that the probability distributions of exons are very similar to probability distributions of introns. Hence it is shown that introns can also carry information as exons do, in contrast to general agreement. If the study is repeated for the other data sets belonging to the human genome, we may obtain results concerning the similarity of the probability distributions of base sequences of exons and introns. Our work suggests that Relative entropy (Kullback-Leibler distance) is useful tool in exploring the distribution of intron and exons.

In analyzing the splice site regions of exons and introns, it is observed that the probability distributions of the bases are very different than the probability distributions of all the bases of exons and introns. It may be said that the last base of

exons, before the GT base pair in the splice site region of genes in data set, are most probably guanine. And the first base after the GT base pair is most probably Adenine or Thymine. We may claim that one may obtain information on the splice site region of the genes by examining the probability distributions of the last bases of exons before the GT pair and the first bases of introns after the GT pair.

Furthermore, when the entropy values calculated using the probability distributions of the amino acid sequences in each three genes, it is observed that HUMGALK1A has the smallest entropy value and this makes the estimation of this gene's amino acids easier. When the similarity of the amino acid distributions of the genes examined it is seen that some of them are quite close. These analyses using this method can be applied to different genes, and the amino acid sequences of genes with similar distribution may be estimated easily by examining other genes' distributions. Finally, the computation of the mutual information value between the amino acids in the genes and the sequence of bases reveals how much information does the knowledge on the base sequence value provides to acknowledge the amino acids in the genes.

REFERENCES

Adami, C. (2004). Information Theory in Molecular Biology. *Physics of Life Reviews, 1 (1), 3-22.*

Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by probability distributions. *Bulletin of the Calcutta Mathematical Society, (35), 99-109.*

Chun, L., & Wang J. (2004), Relative entropy of DNA and its application, *Phsica A: Statistical Mechanics and Its Applications (347), 465-471.*

Cover, T. M., & Thomas, J.A. (2006). *Elements of Information Theory (2nd ed.).* John Wiley & Sons: New Jersey.

Farach, M., Noordewier, M., Savari, S., Shepp, L., Wyner, A.J., & Ziv, J. (1995). On the Entropy of DNA: Algorithms and Measurements Based on Memory and Rapid Convergence. *In Proceedings of the Sixth Annual ACM-SIAM Symposium on Discrete Algorithms. Philadelphia: Society for Industrial and Applied Mathematics (SIAM), 48-57.*

Gates, M. (2000). Basics of Molecular Biology. M. Tompa (Ed.). *Lecture Notes Biological Sequence Analysis*. (1-8). Washington: University of Washington.

Giriftinoğlu, Ç. (2005). *Kesikli Rassal Değişkenler İçin Entropi Optimizasyon Prensipleri ve Uygulamaları*. Anadolu Üniversitesi Fen Bilimleri Enstitüsü İstatistik Bölümü. Eskişehir: Yüksek Lisans Tezi.

Herzel, H., Ebeling, W., & Schmitt, A.O. (1994). Entropies of biosequences: the role of repeats. *Physical Review. E 50, 5061–5071.*

Kullback S. (1987). The Kullback-Leibler distance. *The American Statistician (41), 340-341.*

Leutenneger, A.L. (2000). Relative Entropy. M. Tompa (Ed.). *Lecture Notes Biological Sequence Analysis*. (39-43). Washington: University of Washington.

Mantegna, R.N., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.K., Simons, M., et al. (1994). Linguistic features of non-coding DNA sequences. *Physical Review Letters, 73, 3169–3172.*

McGrats, T. (2000). Basics of Molecular Biology (continued), M. Tompa (Ed.). *Lecture Notes Biological Sequence Analysis*. (9-12). Washington: University of Washington.

Riyazuddin, M. (2006). Information Analysis of DNA Sequences. Louisiana State University the Department of Electrical and Computer Engineering. Louisiana: Master of Science Thesis.

Sakharkar, M.K., Chow V.T.K. & Kangueane P. (2004). Distributions of exons and introns in the human genome. *Silico Biology 4 (0032), Bioinformation Systems e.V., 387-393.*

Schmitt, A.O., & Herzel, H. (1997). Estimating the entropy of DNA sequences. *Journal of Theoretical Biol. 188, 369–377.*

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal (27), 379–423, 623–656.*

Yolaçan, Ş. (2005). *Farklı Dillerin Entropi ve İnformasyon Teorisi Açısından İstatistiksel Özelllikleri*. Anadolu Üniversitesi Fen Bilimleri Enstitüsü İstatistik Bölümü. Eskişehir: Yüksek Lisans Tezi.