Alma Mater Studiorum - Università di Bologna

Scuola di Dottorato in Scienze Economiche e Statistiche
Dottorato di ricerca in

Metodologia Statistica per la Ricerca Scientifica
XXV ciclo

Genome characterization through a mathematical model of the genetic code: an analysis of the whole chromosome 1 of *A. thaliana*

Enrico Properzi

Dipartimento di Scienze Statistiche "Paolo Fortunati"
Gennaio 2013

Alma Mater Studiorum - Università di Bologna

Scuola di Dottorato in Scienze Economiche e Statistiche
Dottorato di ricerca in

Metodologia Statistica per la Ricerca Scientifica
XXIV ciclo

Genome characterization through a mathematical model of the genetic code: an analysis of the whole chromosome 1 of *A. thaliana*

Enrico Properzi

Coordinatore:
Prof.ssa Angela Montanari

Tutor:
Prof.Rodolfo Rosa

Co-Tutor:
Dott.Simone Giannerini
Dott.Diego L. Gonzalez

Settore Disciplinare: SECS-S/02
Settore Concorsuale: 13/D1

Dipartimento di Scienze Statistiche “Paolo Fortunati”
Gennaio 2013

Indice

1 Introduction 5

2 Genetic information 8

2.1 Central dogma of molecular biology 12

2.2 The genetic code 16

2.3 Gene structure 20

2.4 *Arabidopsis thaliana* as a model organism 21

3 A mathematical model for the genetic code 23

3.1 Mathematical structure of the code 23

3.1.1 Degeneracy and redudancy 25

3.1.2 Non-power binary representation of the genetic code . 27

3.1.3 A hiearchy of symmetries 31

3.1.4 Dichotomic classes for codons 35

4 Descriptive analysis 40

4.1 Base distribution 44

4.2 Percentage of dichotomic classes 46

4.2.1 Normal sequences 46

4.2.2 Complementary sequences 50

4.2.3 Reverted sequences 51

4.2.4 Keto/Amino global transformation sequences 51

4.2.5 Purine/Pyrimidine global transformation sequences . . 52

4.2.6 Comments 53

4.3 Independence test 54

4.3.1 Comments 57

4.4 Conclusions 60

5 Discriminating different portions of the genome 62

5.1 Classification through Logistic Regression 62

5.2 Sequence classification of *A.thaliana* chromosome 1 65

5.2.1 Exons vs Introns 67
5.2.2 CDS vs. Long introns 70
5.2.3 Exons vs. CDS 72
5.2.4 Genes vs. Intergenes 74
5.2.5 Exons vs. UTRs 76
5.2.6 Introns vs. UTRs 78
5.3 Comments 80

6 Dependence analysis 81
6.1 Dependence measures 82
6.1.1 Chi-squared test for independence 84
6.1.2 Mutual information 87
6.1.3 S_ρ 89
6.2 Sequence analysis 92
6.2.1 Results 95
6.2.2 Comparison between indexes 102
6.2.3 Comments 102
6.3 Multiple testing problem 106
6.3.1 Statistical Solutions to the Multiple Testing Problem . 108
6.3.2 Related problems 110

7 Conclusions 112

A Dichotomic class tables 115
A.1 Normal sequences 116
A.1.1 Median values 116
A.1.2 Mean values 117
A.2 Complementary sequences 118
A.2.1 Median values 118
A.2.2 Mean values 119
A.3 Reverted sequences 120
A.3.1 Median values 120
A.3.2 Mean values 121
A.4 Global transformation 1 122
A.4.1 Median values 122
A.4.2 Mean values 123
A.5 Global transformation 2 124
A.5.1 Median values 124
A.5.2 Mean values 125

B Distributions **126**

B.1 Dichotomic classes distributions 127

B.2 Base distribution 143

Capitolo 1

Introduction

In this thesis work I would like to combine my different skills as a biotechnologist and as a statistician. I decided to analyze, from a mathematical point of view, the genome of the whole chromosome 1 of a simple plant organism, *Arabidopsis thaliana* (*A.thaliana*), that represents a model for molecular biology and genetic studies.

The discovery of the genetic code, a universal translation table that links the world of nucleic acids to the world of proteins, led scientists to focus on sequencing the entire genomes of different organisms. The Human Genome Project succeeded in sequencing the whole human genome in 2001 [38, 54] and triggered a strong hype on the possibility of diagnosing and treating many serious diseases. However, after ten years, it looks like the expectations have not been met. The recent article by S.S. Hall published on Scientific American: "*Revolution Postponed: Why the Human Genome Project Has Been Disappointing*" is emblematic: In fact its subtitle states: "*The Human Genome Project has failed so far to produce the medical miracles that scientists promised. Biologists are now divided over what, if anything, went wrong - and what needs to happen next*".

The whole genetic information is passed from a parent cell to two or more daughter cells through the process of cell division. The main concern of cell division is the maintenance of the genome of the original cell. Before division can occur, the genetic information must be replicated and the duplicated genome is separated cleanly between cells. During DNA replication several errors may occur. Some of these errors have no effect on the life of the cell, while others can result in growth defects, cell death or cancer. A permanent change in the DNA sequence of a gene is called mutation. [44, 45].

Mutations occasionally occur within cells as they divide and can affect the behaviour of cells, sometimes causing them to grow and divide more frequently. Several biological mechanisms can stop this process: biochemical signals can cause inappropriately dividing cells to die. Sometimes additional mutations make cells ignore these messages. Most dangerously, a mutation may give a cell a selective advantage, allowing it to divide more vigorously than its neighbours and to become a founder of a growing mutant clone. Since mutations may occur because of errors during DNA replication, the study of error detection/correction mechanism in such process could be of key importance for understanding the onset of different serious pathologies, among with there is cancer.

In biology, a reading frame is a way of breaking a sequence of nucleotides in DNA or RNA into three letter codons which can be translated in amino acids. There are 3 possible reading frames in an DNA strand: each reading frame corresponds to starting at a different alignment. Usually, there is only one correct reading frame. Moreover, error detection and correction mechanisms are strictly involved with frame recognition. [39]

In this work I study the features of different portions of the genome of *A.thaliana*, by using a recently developed mathematical model for the genetic code [16, 18, 17]. I use the information of dichotomic classes, binary variables naturally derived from the above mentioned model, in order to assess different behaviours between coding and non coding sequences. In particular I analyze the role of frame. So far, the mathematical model of the genetic code has been used to investigate only some proteins of different origin [19, 20, 21, 22, 23]. Now I apply it to a whole chromosome of a single (and well-known) organism: *A.thaliana*. It has many advantages for genome analysis: a small size, a short generation time and relatively small nuclear genome. These advantages promoted the growth of a scientific community that has investigated the biological processes of *A.thaliana* and has characterized many genes [50].

Finally, since the existence of a coding mechanism for error correction and detection implies some kind of dependence inside data, I want finally to study the presence of dependence structure, related to dichotomic classes, within different portion of the genome. It could be useful in order to develop alternative methods to understand error detection and correction mechanisms involved in the translation process.

The thesis is organized as follows: in chapter 1 I introduce the basic concept and terminology of genetics and the model organism *A.thaliana*. In chapter 2 I describe the salient features of the mathematical model. In chapter 3 I perform a descriptive statistical analysis on the data set; moreover, I implement and apply a test for independence based on dichotomic classes. In chapter 4

I describe logistic regression models built in order to discriminate between different portions of the genome. In chapter 5 I use three different measures of dependence (χ^2, S_ρ and mutual information) in order to assess if there are short-range dependence structure related to dichotomic class within different portions of the genome. Finally I discuss the results.

Capitolo 2

Genetic information

Genetics is the science of genes, heredity, and variation in living organisms [27]. It deals with the molecular structure and function of genes. Since genes are universal to living organisms, genetics can be applied to the study of all living systems, from viruses and bacteria, through plants and domestic animals, to humans (as in medical genetics).

The genetic information is carried by genes segments of DNA (deoxyribonucleic acid) located on chromosomes. They exist in alternative forms called alleles that determine distinct traits which can be passed on from parents to offspring. The process by which genes are transmitted was discovered by Gregor Mendel and formulated in what is known as Mendel's law of segregation.

DNA is a self-replicating nucleic acid which is present in nearly all living organisms as the main constituent of chromosomes. It is the carrier of genetic information.

DNA The molecular basis of genes is DNA. Each molecule of DNA consists of two strands coiled round each other to form a double helix, a structure like a spiral ladder. DNA is a polymer. The monomer units of DNA are nucleotides, and the polymer is known as a polynucleotide. Each nucleotide consists of a 5-carbon sugar (deoxyribose), a nitrogen containing base attached to the sugar, and a phosphate group (see Figure 2). There are four different types of nucleotides found in DNA, differing only in the nitrogenous base. The four nucleotides are given one letter abbreviations as shorthand for the four bases.

- A is for adenine
- G is for guanine
- C is for cytosine

- T is for thymine

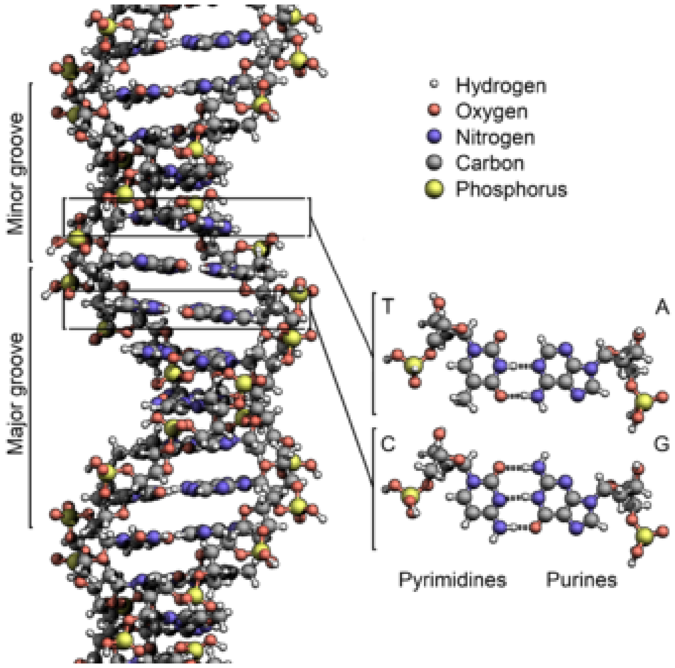


Figura 2.1: DNA molecule structure
(from http://commons.wikimedia.org, under Creative Commons)

Adenine and guanine are purines while cytosine and thymine are pyrimidines. Purines are the larger of the two types of bases found in DNA; they have two nitrogen-containing rings while pyrimidines have only one.

A **nucleoside** is one of the four DNA bases covalently attached to the C1' position of a sugar. The sugar in deoxynucleosides is 2'-deoxyribose. A **nucleotide** is a nucleoside with one or more phosphate groups covalently attached to the 3'- and/or 5'-hydroxyl group (see Figure 2).

The DNA backbone is a polymer with an alternating sugar-phosphate sequence. The deoxyribose sugars are joined at both the 3'-hydroxyl and 5'-hydroxyl groups to phosphate groups in ester links, also known as phosphodiester bonds. Chain has a direction (known as polarity), 5'- to 3'- from top to bottom and A, G, C, and T bases can extend away from chain, and stack atop each other.The bases combine in specific pairs (A/T and C/G)so that the sequence on one strand of the double helix is complementary to that on the other: it is the specific sequence of bases which constitutes the genetic information. [27]

Features of the DNA double helix:

- Two DNA strands form a helical spiral, winding around a helix axis in a right-handed spiral
- The two polynucleotide chains run in opposite directions
- The sugar-phosphate backbones of the two DNA strands wind around the helix axis like the railing of a spiral staircase
- The bases of the individual nucleotides are on the inside of the helix, stacked on top of each other like the steps of a spiral staircase.
- Within the DNA double helix, A forms 2 hydrogen bonds with T on the opposite strand, and G forms 3 hydrogen bonds with C on the opposite strand. For this reason and G are called Strong bases (S) while T and A are called Weak (W).

Genes are arranged linearly along long chains of DNA base-pair sequences. In bacteria, each cell usually contains a single circular genophore, while eukaryotic organisms (including plants and animals) have their DNA arranged in multiple linear chromosomes. These DNA strands are often extremely long; the largest human chromosome (chromosome 1), for example, is about 247 million base pairs in length.[26]

All living organisms can be sorted into one of two groups depending on the fundamental structure of their cells. These two groups are the prokaryotes and the eukaryotes:

- **Prokaryotes** are organisms made up of cells that lack a cell nucleus or any membrane-encased organelles. This means the genetic material DNA in prokaryotes is not bound within a nucleus. Additionally, the DNA is less structured in prokaryotes than in eukaryotes. In prokaryotes, DNA is a single loop. In eukaryotes, DNA is organized into chromosomes. Most prokaryotes are made up of just a single cell (unicellular) but there are a few that are made of collections of cells (multicellular). Scientists have divided the prokaryotes into two groups, the Bacteria and the Archaea.
- **Eukaryotes** are organisms made up of cells that possess a membrane-bound nucleus (that holds genetic material) as well as membrane-bound organelles. Genetic material in eukaryotes is contained within a nucleus within the cell and DNA is organized into chromosomes. Eukaryotic organisms may be multicellular or single-celled organisms. All animals are eukaryotes. Other eukaryotes include plants, fungi, and protists.

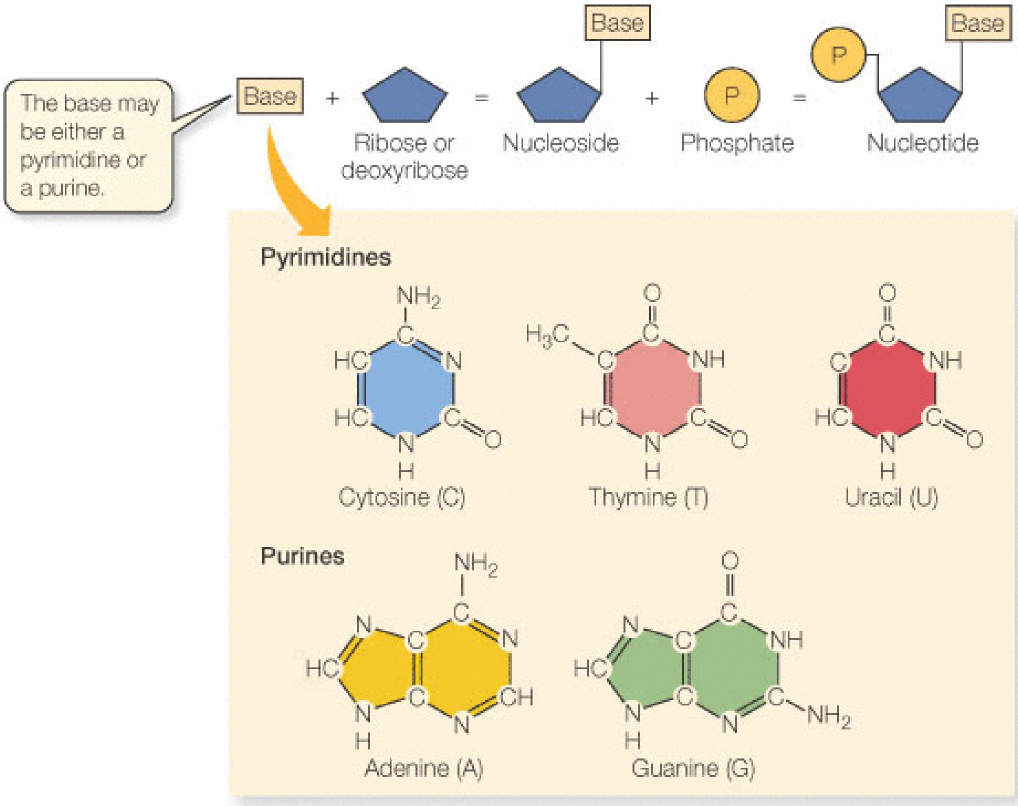


Figura 2.2: Components of a DNA molecule
(from: http://www.nature.com/scitable)

2.1 Central dogma of molecular biology

The central dogma of molecular biology describes the flow of genetic information within a biological system. It was first stated by Francis Crick in 1958 and re-stated in a Nature paper published in 1970. [7]

The central dogma of molecular biology deals with the detailed transfer of sequential information. It states, as Marshall Nirenberg said, *DNA makes RNA makes protein* meaning that information is transferred from DNA to RNA and from RNA to proteins and not in the opposite sense.

The dogma is a framework for understanding the transfer of sequence information between sequential information-carrying biopolymers, in the most common or general case, in living organisms. There are 3 major classes of such biopolymers: DNA and RNA and protein. DNA, deoxyribonucleic acid, and RNA, ribonucleic acid, are molecules that hold the genetic information of each cell. The DNA strands store information, while the RNA molecules take the information from the DNA, transfer it to different places in the cell, and decode or read the information. RNA molecules are similar to DNA ones except for:

- The sugar present in the backbone is ribose instead of deoxyribose
- The nucleobase thymine (T) is substituted by Uracil (U)
- RNA molecules are formed by a single strand

Proteins, instead, are large biological molecules consisting of one or more chains of amino acids. [42]

The flow of biological information is:

- DNA can be copied to DNA (DNA replication),
- DNA information can be copied into mRNA (transcription), and
- proteins can be synthesized using the information in mRNA as a template (translation).

Replication It is the process in which a cell makes an exact copy of its own DNA (copy DNA → DNA). Replication occurs in the step of cell division cycle during which the genetic information is transferred from the mother-cell to the daughter-cell. Replication begins with local decondensation and separation of the double DNA helices, so that the DNA molecule becomes accessible for enzymes that make a complementary copy of each strand (see Figure 2.1).During DNA replication, it takes place DNA errors control.

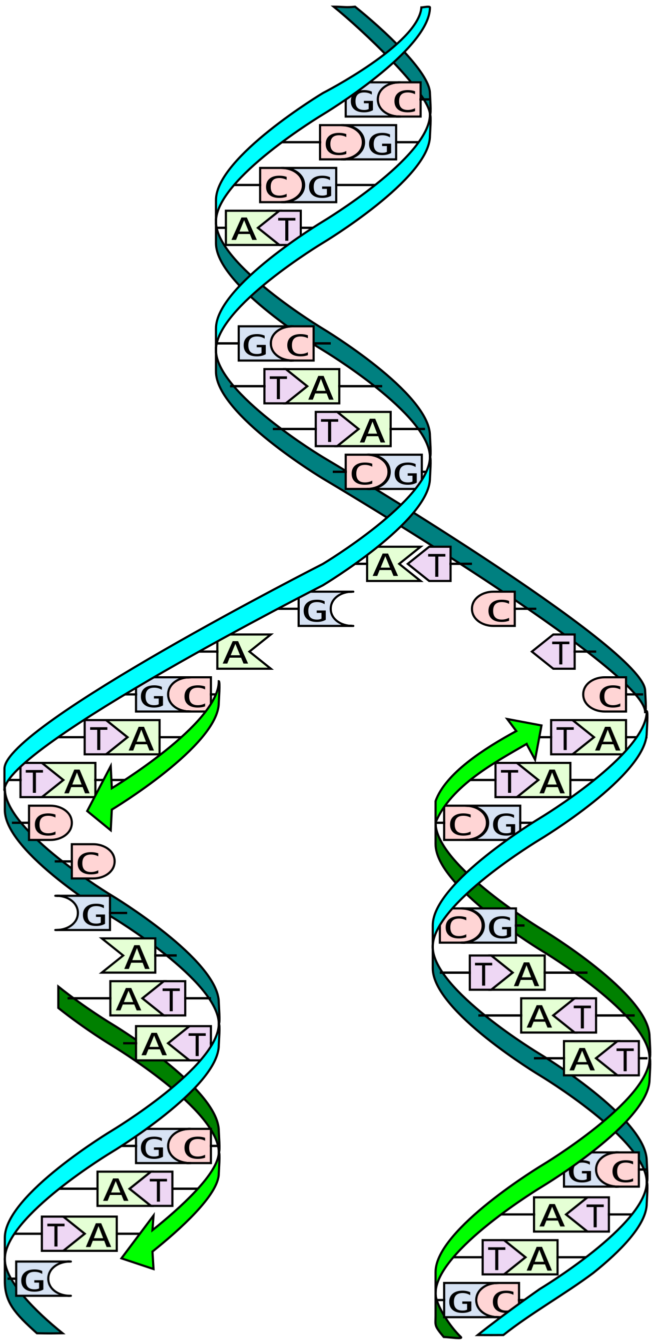


Figura 2.3: DNA replication
(from http://commons.wikimedia.org, under Creative Commons)

Transcription In the transcription step, DNA is copied to RNA in order to produce mRNA (messenger RNA), rRNA (ribosomal RNA) or tRNA (transport RNA). This happens in the nucleus by means of enzymatic complexes produced themselves by specific genes. The RNA is further transported outside the nucleus, to the cytoplasm, where it become active in the translation (the actual synthesis of proteins). During transcription the chromosomes are locally despiralized (decondensed), so that the genes present inside can be read. (see Figure 2.1)

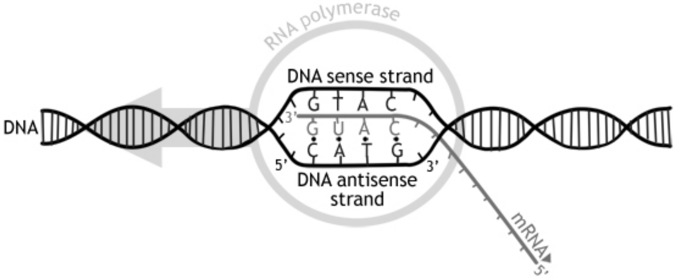
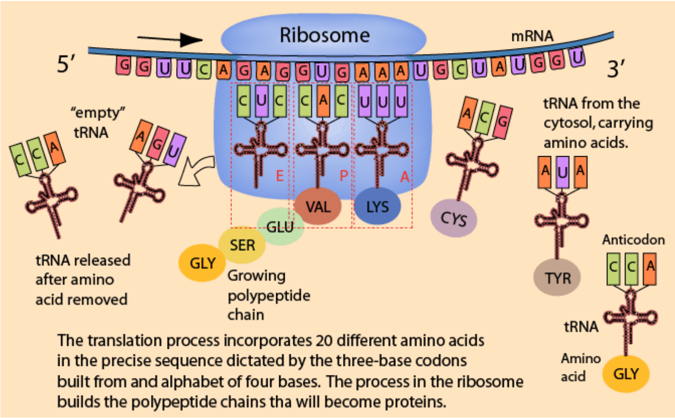


Figura 2.4: Transcription process
(from http://commons.wikimedia.org, under Creative Commons)

Translation The synthesis of new proteins occurs in the cytoplasm, more precisely in ribosomes located in polyribosomal complexes or in the rough endoplasmatic reticulum where a rRNA unit binds a single-strand mRNA chain, which enhosts the genetic code as mirror of the DNA template. tRNA units carry aminoacids (each tRNA binds specifically to one of the 20 different amninoacids) to the ribosomes where they are coupled to form a polypeptide (see Figure 2.1). [36]

There is an exception to the central dogma of molecular biology, and it is represented by the process of ***reverse transcription***. It is directly opposite to the process of transcription: an enzyme, called *reverse transcriptase* (RT) is able to generate a complementary DNA molecule (cDNA) from an RNA template. RT is needed for the replication of particular viral species (retroviruses - e.g. HIV), and its activity is also associated with the replication of chromosome ends (telomerase) and some mobile genetic elements (retrotransposons).


Ribosome
mRNA
5′
3′
G G U U C A G A G G U G A A A U G C U A U G G U
C U C
C A C
U U U
E
P
A
"empty" tRNA
C C A
A G U
A C G
tRNA from the cytosol, carrying amino acids.
A U A
VAL
LYS
GLU
SER
GLY
CYS
TYR
Anticodon
C C A
tRNA
Amino acid
GLY
tRNA released after amino acid removed
Growing polypeptide chain

The translation process incorporates 20 different amino acids in the precise sequence dictated by the three-base codons built from and alphabet of four bases. The process in the ribosome builds the polypeptide chains tha will become proteins.

Figura 2.5: Translation process
from: http://hyperphysics.phy-astr.gsu.edu/hbase/hframe.html

2.2 The genetic code

Genetic information, represented by genes, is used by organisms to create all the proteins necessary for their metabolism. The genetic code is the dictionary used by the cell to translate a sequence of codons (triplets or bases) of RNA in a sequence of amino acids during the translation process. Almost all living organisms use the same genetic code, called the standard genetic code (see Table 3.1), but many slight variants have been discovered. There are various alternative mitochondrial codes, and small variants in some members of bacteria and archaea.

Despite these differences, all known naturally-occurring codes are very similar to each other, and the coding mechanism is the same for all organisms: it implies three-base codons, tRNA, ribosomes, reading the code in the same direction and translating the code three letters at a time into sequences of amino acids.

The ribosome facilitates the decoding process by inducing the binding of tRNAs with complementary anticodon sequences to that of the mRNA. The tRNAs carry specific amino acids that are chained together into a polypeptide as the mRNA passes through and is read by the ribosome in a fashion reminiscent to that of a stock ticker and ticker tape. Each codon corresponds to a specific amino acid, then it is said that the codon encodes that specific aminoacid in the genetic code. [8] [43]

The RNA is made of four bases: adenine (A), guanine (G), cytosine (C) and uracil (U) (in DNA uracil is replaced by thymine (T)). There are therefore $4^3 = 64$ possible codons. 61 of them encode amino acids, while the remaining three (UAA, UAG, UGA) encode stop signals, that is, at what point the assembly of the polypeptide chain should be stopped. Because the amino acids that contribute to the formation of proteins are 20, they generally are encoded by more than one codon.

Therefore genetic code is said to be degenerate and different codons that encode the same amino acid are synonymous. For example, the sequence of RNA UUUACACAG consists of three codons, UUU, ACA, CAG, which correspond to the amino acids Phenylalanine (Phe), Threonine (Thr) and Glutamine (Gln). Protein synthesis applied to this sequence would then generate the tripeptide Phe-Thr-Gln. Therefore, we can say that the genetic code connects the language of nucleic acids and the language of proteins. In Table 3.1 codons are displayed into quartets: groups of codons sharing the first two bases.

Not all genetic information is stored using the genetic code. In all organisms, DNA contains regulatory sequences: intergenic segments, chromosomal structural areas, and other non-coding DNA that can contribute greatly to

UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
UUA	Leu	UCA	Ser	UAA	Stp	UGA	Stp
UUG	Leu	UCG	Ser	UAG	Stp	UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

Tabella 2.1: Standard genetic code

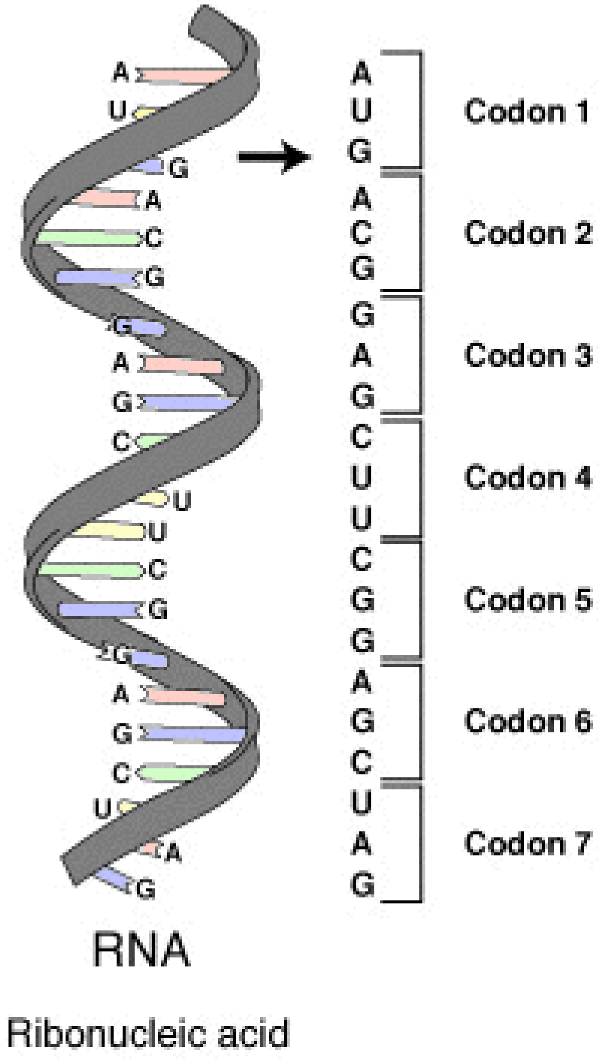


Figura 2.6: RNA codons (from http://commons.wikimedia.org, under Creative Commons)

phenotype. Those elements operate under sets of rules that are distinct from the codon-to-amino acid paradigm underlying the genetic code.[?]

Sequence reading frame A codon is defined by the initial nucleotide from which translation starts. For example, the previous sequence UUUACACAG, if read from the first position, contains the codons UUU, ACA, and CAG; and, if read from the second position, it contains the codons UUA and CAC; if read starting from the third position, UAC and ACA. Every sequence can, thus, be read in three reading frames, each of which will produce a different amino acid sequence (in the given example, Phe-Thr-Gln, Leu-His, or Tyr-Thr, respectively). With double-stranded DNA, there are six possible reading frames, three in the forward orientation on one strand and three reverse on the opposite strand. The actual frame in which a protein sequence is translated is defined by a start codon, usually the first AUG codon in the mRNA sequence. [43]

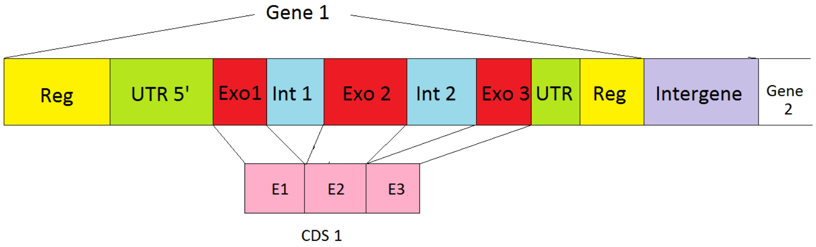
Start/stop codons Translation starts with a chain initiation codon (start codon). Unlike stop codons, the codon alone is not sufficient to begin the process. Nearby sequences (such as the Shine-Dalgarno sequence in E.coli) and initiation factors are also required to start translation. The most common start codon is AUG, which is read as methionine (Met) or, in bacteria, as formylmethionine. The three stop codons are: UAG, UGA and UAA. Stop codons are also called termination or nonsense codons. They signal release of the nascent polypeptide from the ribosome because there is no cognate tRNA that has anticodons complementary to these stop signals, and so a release factor binds to the ribosome instead.[?]

Mutations During the process of DNA replication, errors occasionally occur in the polymerization of the second strand. These errors, called mutations, can have an impact on the phenotype of an organism, especially if they occur within the protein coding sequence of a gene. Error rates are usually very low (1 error in every 10-100 million bases) due to the proofreading ability of DNA polymerases.[27][?] Missense mutations and nonsense mutations are examples of point mutations. Clinically important missense mutations generally change the properties of the coded amino acid residue between being basic, acidic polar or non-polar, whereas nonsense mutations result in a stop codon. Mutations that disrupt the reading frame sequence by *indels* (insertions or deletions) of a non-multiple of 3 nucleotide bases are known as frameshift mutations. These mutations usually result in a completely different translation from the original, and are also very likely to cause a stop codon to be read, which truncates the creation of the protein. These mutations may impair the function of the resulting protein. Although most mutations that change protein sequences are harmful or neutral, some mutations have a positive effect on an organism. These mutations may enable the mutant organism to withstand particular environmental stresses better than wild-type organisms, or reproduce more quickly. In these cases a mutation will tend to become more common in a population through natural selection.[43]

2.3 Gene structure

There are two general types of gene in the human genome: non-coding RNA genes and protein-coding genes. Non-coding RNA genes represent 2-5 per cent of the total and encode functional RNA molecules. Many of these RNAs are involved in the control of gene expression, particularly protein synthesis. They have no overall conserved structure. Protein-coding genes represent the majority of the total and are expressed in two stages: transcription and translation. They show incredible diversity in size and organisation and have no typical structure. There are, however, several conserved features. The boundaries of a protein-encoding gene are defined as the points at which transcription begins and ends. The core of the gene is the coding region, which contains the nucleotide sequence that is eventually translated into the sequence of amino acids in the protein. The coding region begins with the initiation codon, which is normally AUG. It ends with one of three termination codons: UAA, UAG or UGA. On either side of the coding region are DNA sequences that are transcribed but are not translated. These untranslated regions or non-coding regions often contain regulatory elements that control protein synthesis. Both the coding region and the untranslated regions may be interrupted by introns. Most human genes are divided into exons and introns. The exons are the sections that are found in the mature transcript (messenger RNA), while the introns are removed from the primary transcript by a process called splicing. [39]

Summarizing, eukaryotic gene structure is shown in the following figure:



We can recognize:

- **Genes**: regions of a genomic sequence corresponding to a unit of inheritance. They are formed by regulatory regions, transcribed regions, and/or other functional sequence regions.
- **Exons**: portions of a gene that are transcribed into mRNA and then translated into a protein. Each gene can contain one or more exons.
- **CDS**: portions of a gene that encode for a given protein. It is formed by joining exons (one or more) within a gene.
- **Introns**: portions of a gene that are transcribed but not translated.
- **Intergenes**: sequences between a gene and the following one.
- **UTR**: portions of mRNA that precede the codon that begins translation (AUG) (5'UTR) and follow the termination codon (3' UTR)
- **Regulatory regions**: portions of a gene, with regulatory function, that precede (upstream) and follow (downstream) the fragment transcripted into mRNA

2.4 *Arabidopsis thaliana* as a model organism

Although geneticists originally studied inheritance in a wide range of organisms, researchers began to specialize in studying the genetics of a particular subset of organisms. The fact that significant research already existed for a given organism would encourage new researchers to choose it for further study, and so eventually a few model organisms became the basis for most genetics research.Common research topics in model organism genetics include the study of gene regulation and the involvement of genes in development and cancer. Organisms were chosen, in part, for convenience-short generation times and easy genetic manipulation. Widely used model organisms include the gut bacterium *Escherichia coli*, the plant *Arabidopsis thaliana*, baker's yeast *Saccharomyces cerevisiae*, the nematode *Caenorhabditis elegans*, the common fruit fly *Drosophila melanogaster*, and the common house mouse *Mus musculus.*

Arabidopsis thaliana (*A. thaliana*) is a small flowering plant native to Europe, Asia, and northwestern Africa. A spring annual with a relatively short life cycle, *A. thaliana* is popular as a model organism in plant biology and genetics. *A. thaliana* has a rather small genome, only 157 megabase pairs (Mbp) and five chromosomes [50]. Arabidopsis was the first plant genome to be sequenced, and is a popular tool for understanding the molecular biology of many plant traits. By the beginning of 1900s, *A.thaliana* began to be used in some developmental studies. It plays the role in plant biology that mice and fruit flies (Drosophila) play in animal biology. Although *A.thaliana* has little direct significance for agriculture, it has several traits that make it a useful model for understanding the genetic, cellular, and molecular biology of flowering plants.

The small size of its genome makes *A. thaliana* useful for genetic mapping and sequencing. It was the first plant genome to be sequenced, completed in 2000 by the Arabidopsis Genome Initiative.[35] The most up-to-date version of the *A.thaliana* genome is maintained by the Arabidopsis Information Resource (TAIR). Much work has been done to assign functions to its 27,000 genes and the 35,000 proteins they encode.[50]

Capitolo 3

A mathematical model for the genetic code

The genetic code is the dictionary used by the cell to translate a sequence of codons (triplets or bases) of RNA in a sequence of amino acids during the translation process. In Table 3.1 codons are displayed into quartets: groups of codons sharing the first two bases. Since 64 codons encode for 20 amino acids and the stop signal, some amino acids are necessarily encoded by more than one codon. This fact determines the properties of redundancy and degeneracy typical of the genetic code.

Why amino acids are not represented in a unique way? And why the level of degeneracy is different between amino acids?

3.1 Mathematical structure of the code

In order to try to answer these questions we could resort to information theory, a branch of applied mathematics involving the quantification of information. A key measure of information is known as **entropy**.In information theory, entropy is a measure of the uncertainty associated with a random variable. In this context, the term usually refers to the Shannon entropy, which quantifies the expected value of the information contained in a message, usually in units such as bits. In this context, a "message" means a specific realization of the random variable and implies the presence of a term of uncertainty (error).

Considering genetic code as a communication system allow us to apply information theory concept. Therefore we can state that it is not a free-error system. Errors, which occur mainly during the transmission phase, can be detected and then corrected at the time of decoding the message. Gonzalez

Tabella 3.1: Standard genetic code

UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
UUA	Leu	UCA	Ser	UAA	Stp	UGA	Stp
UUG	Leu	UCG	Ser	UAG	Stp	UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

et al. [19] investigated the existence of error-detection and correction mechanisms in the genetic machinery based on a particular mathematical model of the genetic code. This model can encode each nucleotidic sequence into three binary strings of different meaning; these strings show some interesting correlation patterns that enforce the hypothesis of deterministic error and correction mechanisms. In the following pages we explain the features of this mathematical model.

3.1.1 Degeneracy and redudancy

The genetic code is a surjective (all amino acids are encoded by at least one codon) and non-injective (some amino acids are degenerate) function between two sets of different cardinality. We can say that the codon set is the domain and the amino acids set the codomain. This implies the degeneration of the code.

Gonzalez [16, 18, 17] proposed a model that explains the degeneracy of the genetic code based on a non-power number representation system. This approach describes the structure of the genetic code from a mathematical point of view and allows the analysis of degeneracy and redundancy properties on two related levels:

- the distribution of degeneration (the number of codons that code for each amino acid)
- the distribution of codons (codons assigned to each specific amino acid)

The code is degenerate from the amino acids point of view: a given amino acid can indeed be encoded by more than one codon. The redundancy, however, is a property that concerns the codons: a set of triplets that encode the same amino acid is said to be redundant. Degeneracy and redundancy are still described by the numerical quantities that define the respective subsets: tyrosine is a degeneracy-2 amino acid because it is encoded by a set of two redundant codons (UAU, UAC). Tables 3.2 and 3.3 show the degeneracy distribution of euplotid genetic code.

The main difference with the standard version concerns the UGA codon: here it encodes the amino acid Cysteine while in the standard version of the code it is one of the stop signal. The degenracy distribution inside quartets is obtained by taking into account that the 3 degeneracy-6 amino acids (Arginine, Leucine and Serine) are divided into two subsets of degeneracy-2 and 4.

After specifying the degeneracy distribution it is necessary to associate an amino acid to each codon. This distribution of codons represents a second level of complexity and defines uniquely the genetic code.

Tabella 3.2: Degeneracy distribution for the euplotid genetic code

number of amino acids sharing the same degeneracy	Degeneracy
2	1
9	2
2	3
5	4
3	6

Tabella 3.3: Degeneracy distribution inside quartets of euplotid nuclear version of genetic code

number of amino acids sharing the same degeneracy	Degeneracy
2	1
12=9+3	2
2	3
8=5+3	4

3.1.2 Non-power binary representation of the genetic code

The genetic code can be defined as a translation table that connect two finite sets of 64 codons and 20 amino acids. In this context, the binary system of representation is very interesting. Indeed, using a binary string of length n we can represent 2^n different objects. For example, the 4 nucleotides (A, C, G, U) can be represented by a binary string of length 2. Consequently codons (groups of 3 nucleotides) are represented by binary strings of length 6, in fact $2^6 = 64$. Gonzalez [16, 18, 17] showed that using a particular type of number positional representation, called non-power representation, we can fully describe the degeneracy distribution of the genetic code.

Usual number representation systems are positional power representation systems. In these systems numbers are represented by a combination of digits, from 0 to $n-1$, where n is the system base, that are weighted with values that grow following the power expansion of the base n. For example, if we want to represent number 476 in base 10 we have to use digit from 0 to 9 in this way:

$$476 = 4 * 10^2 + 7 * 10^1 + 6 * 10^0$$

while number 13 is obviously represented by

$$13 = 1 * 10^1 + 3 * 10^0$$

If we turn to the binary system, the power positional representation of number 13 is 1101:

$$13 = 1 * 2^3 + 1 * 2^2 + 0 * 2^1 + 1 * 2^0$$

In ***non-power representation systems*** the positional values grows more slowly than the powers of the system base. This implies that:

- it is possible to represent all the numbers from 0 to the sum of all the positional weights;
- the system is redundant: a given number can be represented by more than one string.

Hence, non-power representation systems can be used to describe degeneracy distributions. A typical example of non-power representation is Fibonacci representation [57] where positional weights are represented by successive Fibonacci numbers. These numbers form a series in which the n^{th} element of the series is the sum of its two predecessors, with the first two elements of the series = 1. For example, if we consider Fibonacci representation of order 6, we use as positional values the first 6 Fibonacci numbers: 1, 1, 2, 3, 5 and 8. Using this representation system we can describe number from 0 to 21 with degeneracy distribution as shown in Table 3.4.

Tabella 3.4: Degeneracy distribution for Fibonacci non-power representation

numbers sharing the same degeneracy	Degeneracy
2	1
4	2
8	3
5	4
2	5

We can notice, for instance, that in this system number 7 is represented by 3 different strings : 010011, 010100 and 001111.

There is a non-power binary representation that describes perfectly the degeneracy of the genetic code. This system is based on a specific sequence of positional wiegths: (1, 1, 2, 4, 7, 8) and this solution is unique up to trivial equivalence classes [16]. The solution is specific for the degeneracy inside quartets (presented in Table 3.3) because there is no solution for the degeneracy distribution at large (presented in Table 3.2).

First of all, we can observe that any non-power representation is palindromic: the represented number n and $N - n$ (where N, the sum of all weigths, is the maximum integer that can be represented) share the same degeneracy. Therefore the degeneracy distribution of Table 3.2 can't be represented in any way. On the contrary, degeneracy inside quartets can be ordered in a palindromic table.

Table 3.5 shows the non-power representation of the first 23 integers by length-6 binary strings and positional weights 1,1,2,4,7,8. Notice the same degeneracy distribution of euplotid genetic code (see Table 3.6).

We can state that each codon can be associated to a length-6 binary string representing a whole number. So the genetic code and this specific non-power binary representation are linked by a **structural isomorphism**: they share the same logical structure.

Two structures are isomorphic if they are indistinguishable given only a selection of their features. In our case the nuclear genetic code of the fla-

Tabella 3.5: Non power representation of whole numbers by length 6 binary strings

Number	Positional weights: [8,7,4,2,1,1]
0	000000
1	000001 000010
2	000011 000100
3	000101 000110
4	000111 001000
5	001001 001010
6	001011 001100
7	010000 001101 001110
8	100000 010001 010010 001111
9	100001 100010 010100 010011
10	100011 100100 010101 010110
11	100101 100110 011000 010111
12	101000 100111 011001 011010
13	101001 101010 011100 010111
14	101100 101011 011100 011011
15	110000 101101 101110 011111
16	110001 110010 101111
17	110100 110011
18	110101 110110
19	111000 110111
20	111001 111010
21	111100 111011
22	111101 111110
23	111111

Tabella 3.6: Palindromic representation of the euplotid version of the genetic code and non-power binary representation of whole numbers

Degeneracy	Amino acid	Coded whole number
1	T Trp	0
2	F Phe	1
2	Stop	2
2	Y Tyr	3
2	L Leu(2)	4
2	H His	5
2	Q Glu	6
3	C Cys	7
4	S Ser(4)	8
4	P Pro	9
4	V Val	10
4	L Leu(4)	11
4	R Arg(4)	12
4	G Gly	13
4	A Ala	14
4	T Thr	15
3	I Ile	16
2	E Glu	17
2	D Asp	18
2	R Arg(2)	19
2	N Asn	20
2	K Lys	21
2	S Ser(2)	22
1	M Met	23

gellate Euplotes and the non power binary representation of length 6 with bases 1,1,2,4,7,8, are isomorphic with regard to the cardinality of the sets and the cardinality of the respective applications. In fact both domains and both co-domains have the same cardinality: 64 codons and 64 binary non-power strings for the domains, and 24 amino acids (inside quartets) and 24 integer numbers for the co-domains. Moreover the two applications have the same degeneracy distribution (see table 3.6). However such properties do not suffice for establishing a correspondence between codons and binary strings, and between integer number and amino acids, that is, do not suffice for creating a mathematical model of the genetic code. However, studying the properties of both applications, we found that "all" symmetry properties are shared by both applications (for example we have 16 subsets of 2 codons each that are invariant under the $C \leftrightarrow T$ transformation of the last letter of the codon, and we have 16 subsets of 2 binary strings each that are invariant under the complement to one of the two last digits of the string). Of course, there is not any reason "a priori" for the sharing of symmetries between the applications. We can say that the structural isomorphism describing the global degeneracy properties of the genetic code, describes also its internal symmetries. In such a way some important organizational aspects of the genetic code can be uncovered and a true mathematical model of the genetic code can be constructed by assigning specific codons to specific non-power binary strings, and specific amino acids to specific whole numbers.

3.1.3 A hiearchy of symmetries

Pyrimidine ending codons

If we analyze the genetic code and the mathematical model we can notice many symmetry properties. First, if we make a pyrimidine (U vs. C) exchange in the last base of each codon, the meaning of the codon remains the same. This implies the definition of two groups of 16 codons that encode the same amino acid. So far we know 26 variants of the genetic code (10 nuclear and 16 mitochondrial) and all of these respect this symmetry. It's remarkable to observe that the non-power representation system shows this same symmetry. In fact the 6-digits binary strings xxxx01 and xxxx10 always encode the same whole number. This is a property of this specific representation system because it belongs to the positional weights chosen. This means that we can associate strings ending in 01 or 10 with pyrimidine ending codons. It must be underlined that there is no biological reason for this degeneracy as codons ending in C or U are recognized by different tRNA.

Purine ending codons

This aspect determines an immediate consequence since the remaining 32 codons have to be associated with the remaining 32 strings representing whole numbers. Thus strings ending in 00 or 11 are linked with purine ending codons. Since the only two degeneracy-1 strings (000000 and 111111) have to be associated with the only degeneracy-1 codons (AUG and UGG) a new concept raises naturally: the parity of a string, that is, the sum of its digits. So, we can assume that, in case of purine ending codons, strings with even parity are associated with G-ending codons, while strings with odd parity are associated with A-ending codons.

All these aspects are summarized in Table 3.7

Tabella 3.7: Equivalence between strings and purine/pyrimidine ending codons

Strings	Parity	Codons
x x x x 0 1	Even	N N C/U
x x x x 0 1	Odd	N N C/U
x x x x 1 0	Even	N N C/U
x x x x 1 0	Odd	N N C/U
x x x x 1 1	Even	N N G
x x x x 1 1	Odd	N N A
x x x x 0 0	Even	N N G
x x x x 0 0	Odd	N N A

Degeneracy-3 elements

We can notice that there are only two degeneracy-3 whole numbers (7 and 16) and amino acids (Cysteine and Isoleucine). Obviously these elements must be associated. So the group (AUU, AUC, AUA) and (UGU, UGC, UGA) are linked with binary strings representing numbers 7 and 16. These two groups of codons are linked by a degeneracy-preserving transformation: U <–> A in the first base and U <–> G in the second one. It is remarkable to notice that this transformation corresponds to a symmetry property from the model point of view: the palindromic simmetry. In fact the first group strings can be obtained by the 0 <–> 1 exchange of the digits of the second group strings. This palindromy is observed also for the degeneracy-1 string (000000 and

111111) coding for amino acid Methionine (AUG) and Tryptophan (UGG). Notice that the numbers encoded by a palindromic couple sum up to 23.

Summarizing we can say that degeneracy-3 and degeneracy-1 amino acid form two group of quartets that show a *palindromic simmetry*. Notice that, in the standard genetic code we have UGA codon encoding a stop signal and not Cysteine. Palindromic symmetry involves all the quartet of the genetic code. It connects quartets with the same degeneracy distribution and strings related by the complement to one operation.

Degeneracy-6 elements

So far we have noticed the following rules:

- pyrimidine ending codons are linked to strings ending in 01 or 10
- G ending codons are linked to strings ending in 00 or 11 and with even parity
- A ending codons are linked to strings ending in 00 or 11 and with odd parity
- pairs of quartets with the same degeneracy are linked by palindromic symmetry

By analalyzing Table 3.8 we can see that there are two degeneracy-2 numbers that correspond to A-ending codons (4 and 19); but there are no amino acid with degeneracy 2 encoded by two A-ending codons. Therefore these numbers must be associated with the degeneracy-2 part of degeneracy-6 amino acids encoded by at least two A-ending codons.

Looking at the Tables it is easy to recognize these amino acids in leucine (Leu) and arginine (Arg). Both of them are encoded also by two G-ending codons that necessarily belongs to their degeneracy-4 part. The only degeneracy-4 numbers showing this feature are 11 and 12: both of them display two even strings ending with 00 or 11. It can be observed once more that this couples of numbers (4 and 19) and (11 and 12) are palindromic (their sum equals 23). So we can state that there is a symmetry of the role of Leu and Arg.

Pyrimidine ending with odd parity

We succeded in linking binary strings with codons whose second letter is U or G. Moreover all the U or C ending codons so far associated show an odd parity. We could introduce another rule:

- Amino acids with pyrimidine ending codon and with a G or a U (keto base) in the second position are encoded by an odd string

We can find only two degeneracy-4 numbers (10 and 13) and only two degeneracy-2 numbers (1 and 22) satisfying this rule and, as a consequence, we can associate to them the amino acids valine (Val), glycine (Gly), phenylalanine (Phe) and the degeneracy-2 part of serine (Ser).

The last associations: second base A or C

Now it remains to associate only codons whose second base is an amino-base (A or C). It's quite simple because codons with A in second position share all degeneracy-2, while codons with C in the second position have degeneracy-4. Following the rules described above, we can try to give a place to all codons into the mathematical model. The result is shown in Table 3.8

Tabella 3.8: Non-power model of the euplotid nuclear genetic code

	U			C			A			G			
U	1	000001	Phe	15	101101	Ser	18	110110	Tyr	16	110010	Cys	U
	1	000010	Phe	15	101110	Ser	18	110101	Tyr	16	110001	Cys	C
	4	001000	Leu	15	011111	Ser	2	000100	Stp	16	101111	Cys	A
	11	011000	Leu	15	110000	Ser	2	000011	Stp	23	111111	Trp	G
C	11	100101	Leu	14	011110	Pro	3	000101	Tyr	12	011010	Arg	U
	11	100110	Leu	14	011101	Pro	3	000110	Tyr	12	011001	Arg	C
	4	000111	Leu	14	101100	Pro	17	110100	Stp	19	111000	Arg	A
	11	010111	Leu	14	101011	Pro	17	110011	Stp	19	101000	Arg	G
A	7	001101	Ile	8	010010	Thr	5	001001	Asn	22	111110	Ser	U
	7	001110	Ile	8	010001	Thr	5	001010	Asn	22	111101	Ser	C
	7	010000	Ile	8	100000	Thr	21	111011	Lys	19	110111	Arg	A
	0	000000	Met	8	001111	Thr	21	111100	Lys	12	100111	Arg	G
G	13	101001	Val	9	100001	Ala	20	111010	Asp	10	010110	Cys	U
	13	101010	Val	9	100010	Ala	20	111001	Asp	10	010101	Cys	C
	13	011100	Val	9	010011	Ala	6	001011	Glu	10	100011	Cys	A
	13	011011	Val	9	010100	Ala	6	001100	Glu	10	100100	Trp	G

It's easy to notice how palindromy preserve degeneracy within quartets. From a mathematical point of view palindromy is represented by the complement to one operation of all the binary digit of a given string. From a biochemical point of view palindromy is given by different base transformations depending on the quartet considered.

Looking at Table 3.8 we succeded in assigning a binary string to each codon but of course it is not the unique way to do it. In fact it is obvously possible to exchange the full set of a quartet'strings with the set assigned to the palindromic quartet. However this assignation is the most probable one, taking into account all the simmetry properties we have found.

3.1.4 Dichotomic classes for codons

Studying degeneracy properties of the genetic code we can classify couples of two nucleotides into three dichotomic classes:

- parity class
- Rumer's class
- hidden class

For the definition of these classes it is necessary to introduce the unique three possible chemical binary classification of the bases (U, C, A, G):

- Purine(R) vs Pyrimidine(Y): A,G vs C,U
- Keto(K) vs Amino(Am): G,U vs A,C
- Strong(S) vs Weak(W): C,G vs A,U

Parity class

According to the mathematical model described so far, each codon is associated with a binary string. The parity of a codon corresponds to the parity of the sum of all the digits of the associated string. We can observe that the parity of a binary string can be obtained simply by counting the number of ones: an even number of ones leads to an even string while an odd number of ones leads to an odd string. It's important to underline that palindromic symmetry preserves parity in fact the complement to one operation doesn't change the parity of the string (since the string length is even, the parity of one digit remains the same in the complement string). The parity bit of a string can be determined also by its biochemical composition: if we assume that a codon ending with A is represented by an odd string, then every codon ending with G is associated to an even string. If the codon ends with a pyrimidine (U or C) then we have to look at the second base of the codon: when it is an amino-base then the codon is even, while a keto-base in the second position leads to a odd codon. So we can underline that R-Y transformation changes the parity of a string. Now it is possible to build an algorithm as we did for Rumer's class in order to define the parity of a codon from its biochemical composition. This algorithm, obviously, takes into account the last two bases of the codon as it's shown in Figure 3.1

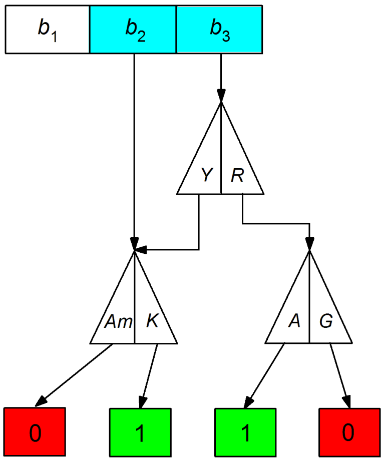


Figura 3.1: Algorithmic definition of the parity class.

Rumer's class

Y. B. Rumer was a theoretical physicist who first noticed a regularity of the degeneracy distribution within quartets in the standard genetic code. He observed that exactly one half of the quartets showed degeneracy-4 while the other half showed degeneracy 1, 2 or 3. So each codon can be assigned to a dichotomic class named Rumer's class depending on whether it belongs to a degeneracy-4 or degeneracy 1, 2 or 3 quartet. Moreover Rumer observed that a specific transformation links the two halves of the genetic code: U,C,A,G <-> G,A,C,U. This Rumer's transormation convert a codon of class 1 2 or 3 in a codon of class 4 and vice-versa; it breaks the degeneracy of the code since it reveals an antisymmetric property of the degeneracy distribution. Rumer's transformation is global. That means that it acts univocally on the 4 mRNA bases. Anyway, the same effect can be obtained if we apply this transformation only to the first two bases (remember that a quartet is a group of four codons sharing the first two letters).

Considering the chemical properties of the codon's bases we can create an algorithm in order to easily determine the Rumer's class which the codon belongs to (see Figure 3.2). First we can take into account the second base of a codon: if it's an amino-base we can immediatly determine the class (class 4 if it's C, class 1,2,3 if it's A). If the second base is a keto-type base (G or U) we have to make one more step considering the strong/weak character of the first base of the codon. If the first base is a strong type base (C or G) then the codon is a class 4 type. Otherwise it's a class 1,2 or 3 type.

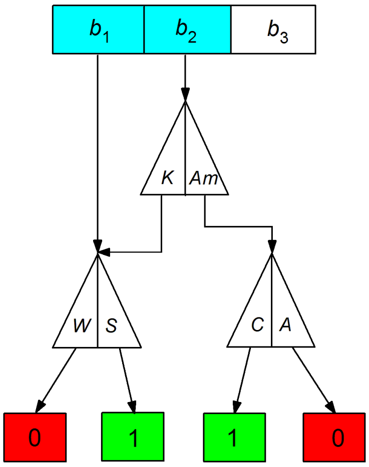


Figura 3.2: Algorithmic definition of the Rumer's class.

Hidden class

At this point we have two algorithms that allow us to generate parity and Rumer class by reading the biochemical properties of a dinucleotide within a codon (we observed that Y-R transformation changes the parity of a codon, while the K-Am transformation changes its the Rumer's class). The two algorithms are obtained by moving of one position the "reading-frame" of the dinucleotide within a codon.

Since the nucleotides present within a codon are three, it would seem logical moving of one more base within the codon in order to generate a new algorithm that will give rise to a new dichotomic class: the hidden class (see Figure 3.3.)

Although the hidden class does not have a specific meaning in relation to the properties of the codons or of the amino acids, it can be interpreted on the basis of the biochemical properties of the bases i.e. it should be antisymmetric with respect to the missing global transformation (S-W).

In this case we have to consider the bases of two different codons: the first base of a certain codon and the third base of the previous one. If the first base is a weak base (A or T) then the hidden class is arbitrarily determined: 0 for A and 1 for T. In case of strong first base (C or G) we have to consider the last base of the previuos codon: if it's a pyrimidine base the hidden class is 0 otherwise it is 1.

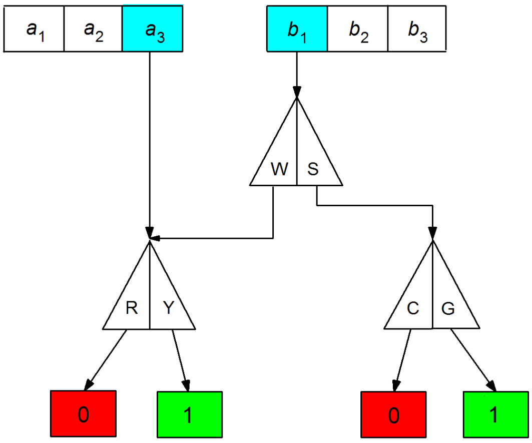


Figura 3.3: Algorithmic definition of the hidden class.

The three global transformations described above, together with the identity transformation, define a Klein V group structure as shown in Table 3.9.

Tabella 3.9: Product table of the Klein V group as implied by the three global transformations plus the identity.

	I	**K-Am**	**S-W**	**Y-R**
I	I	K-Am	S-W	Y-R
K-Am	K-Am	I	Y-R	S-W
S-W	S-W	Y-R	I	K-Am
Y-R	Y-R	S-W	K-Am	I

In fact if we consider the bases as four-dimensional column vectors:

$$U' = (1,0,0,0); \quad C' = (0,1,0,0); \quad A' = (0,0,1,0); \quad G' = (0,0,0,1)$$

the possible global transformation of the bases are defined by the matrix product of the following permutation matrices:

$$L = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \quad M = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \quad N = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

If we include in this set the identity matrix, I_4, we obtain the Klein V group. These matrices are orthogonal and the following identities hold:

$$LM = ML = N; \qquad LN = NL = M; \qquad MN = NM = L$$

Now, defining the infinite order matrix norm for a pXp matrix Q as:

$$\|Q\|_\infty = \max_{1 \le i \le p} \sum_{j=1}^{p} |q_{ij}|$$

we can obtain operators that acting on a 4x4 matrix made of four consecutive vector or bases computes the values of dichotomic classes. For example operators:

$$O_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 2 & 2 & 1 \\ 0 & 0 & 3 & 4 \end{pmatrix} \quad M = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 2 & 1 & 2 \\ 0 & 4 & 3 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

can compute the values of $c_1 = parity$ and $c_2 = Rumer$ classes through the following operation:

$$c_i = \|O_i \odot Q'\|_\infty mod2, \qquad i = 1, 2$$

where $\odot$ denotes the matrix Hadamard product.

Capitolo 4

Descriptive analysis

In this chapter I perform a statistical analysis on the dichotomic classes computed on the eight groups of sequences of the chromosome 1 of *A. thaliana* described before. I want to study whether the information conveyed by dichotomic classes can characterize different portions of the genome. In order to accomplish the task, I encode all the sequences into the three dichotomic classes and study the distributions of such binary sequences. In particular, we focus on their mean value, that is, the percentage of "ones". Thus, for each sequence, I obtain 22 variables as reported in Table 4.1:

Tabella 4.1: Variables included in each dataset

Name	**Description**
$p0, r0, h0$	mean value for parity, Rumer, hidden classes, in frame
$p1, r1, h1$	mean value for parity, Rumer, hidden classes, out of frame 1
$p2, r2, h2$	mean value for parity, Rumer, hidden classes, out of frame 2
$p0a, r0a, h0a$	mean value for parity, Rumer, hidden classes, antisense strand in frame
$p1a, r1a, h1a$	mean value for parity, Rumer, hidden classes, antisense strand out of frame 1
$p2a, r2a, h2a$	mean value for parity, Rumer, hidden classes, antisense strand out of frame 2

I consider eight groups of sequences (see Fig. 4.1) from the chromosome 1 of *A. thaliana*, that is composed by a long DNA sequence of 30.427.671 base pairs as follows:

1. **Genes**: regions of a genomic sequence corresponding to a unit of inheritance. They are formed by regulatory regions, transcribed regions, and/or other functional sequence regions.
2. **Exons**: portions of a gene that are transcribed into mRNA and then translated into a protein. Each gene can contain one or more exons.
3. **CDS**: portions of a gene that encode for a given protein. It is formed by joining exons (one or more) within a gene.
4. **Introns**: portions of a gene that are transcribed but not translated.
5. Long introns: sequences artificially built in order to be compared to CDS. They are composed by joining all the introns present within a gene.
6. **Intergenes**: sequences between a gene and the following one.
7. (**UTR**): portions of mRNA that precede the codon that begins translation (AUG) (5'UTR) and follow the termination codon (3' UTR)
8. **Regulatory regions**: portions of a gene, with regulatory function, that precede (upstream) and follow (downstream) the fragment transcripted into mRNA

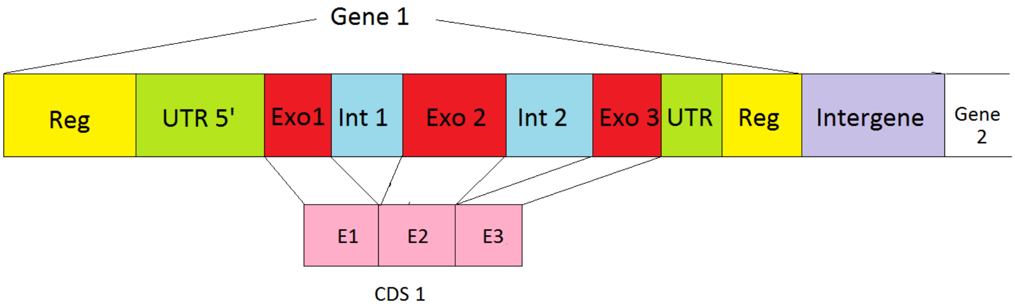


Figura 4.1: Definition of type of sequences within a fragment of DNA

First of all I have extracted the complete sequence of *A. thaliana* chromosome 1 from Genbank. This dataset allows to extract four kind of sequence data in fasta format: the complete sequence of the entire chromosome, a list of the genes sequences, a list of CDS, a list of mRNA sequences. I imported and processed the data by the means of R [51]. Then I created specific original routines that, using the information of annotation, allowed us to extract the remaining group of sequences of interest: exons, introns, intergenes, 5' and 3' untranslated regions (UTR), and upstream and downstream regulatory regions. The annotation file, in fact, contains useful information for this purpose such as the nucleotide position of the beginning and the end of each gene, CDS and mRNA. The procedure led to the creation of eight datasets, one for each sequence group.

Once the data have been imported, I removed from the datasets those sequences that display undefined bases (different from A, C, G, T) or that are shorter than 6 bases. The eight different dataset together with the number of records are shown in Table 4.2.

Tabella 4.2: Number of records and percentages of bases for each type of sequence analyzed from *A. thaliana* chromosome 1

Type	***Records***	A	C	G	T
Genes	8428	28.47	18.77	21.33	31.43
Exons	37549	29.00	19.94	23.73	27.33
CDS	9262	28.61	20.48	23.87	27.04
Introns	30663	26.93	15.72	16.68	40.68
Long introns	5532	27.73	15.15	16.83	40.29
Intergenes	8350	34.01	15.92	16.04	34.03
UTR	14427	30.42	17.76	16.78	35.10
Reg	2037	31.08	18.34	16.38	34.19

Sequence length I start by analyzing the sequence length to study the differences between the classes considered.

From Figure 4.2 and Table 4 it is evident that genes and intergenes, CDS and long introns are much longer than exons, introns, UTRs and regulatory sequences. I am going to compare genes with intergenes, introns with exons, CDS with long introns and finally UTRs with regulatory sequences.

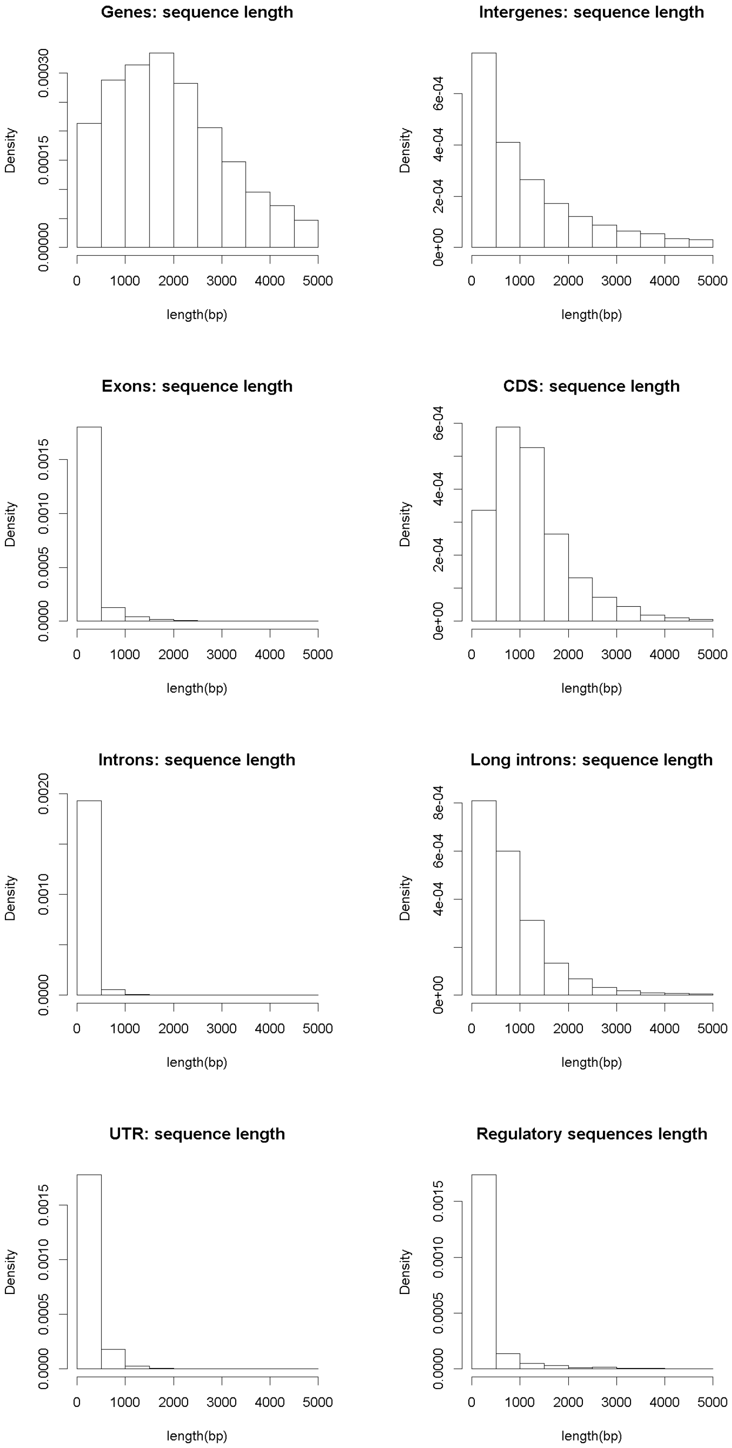


Figura 4.2: Length of the different classes of sequences

Tabella 4.3: Length statistics for each portion of the genome

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Genes	37.00	1041.00	1843.00	2160.00	2846.00	26440.00
Exons	6.00	85.00	133.00	236.40	243.00	7761.00
CDS	63.00	648.00	1065.00	1264.00	1587.00	16180.00
Introns	8.00	85.00	98.00	155.60	153.00	5610.00
Long introns	20.00	299.00	651.00	862.20	1123.00	13170.00
Intergenes	6.00	317.00	844.50	1499.00	1928.00	72640.00
UTR	6.00	109.00	190.00	262.10	299.00	13790.00
Regulatory	6.00	37.00	90.00	295.40	247.00	10280.00

4.1 Base distribution

The data reported in Table 4.4 and in Figures in appendix B.2 shows some interesting differences between the eight groups of sequences we considered. In fact while non coding sequences (such as introns, intergenes, UTR and regulatory sequences) have an higher prevalence of A and T bases (with median values), CDS and exons show an increase of C and G bases. The whole gene sequences obviously show intermediate features because they are composed by introns, exons, UTR and regulatory sequences together.

In particular we can see that:

- Itergenes display a clear prevalence of A nd T (median values of 34%).
- In UTR and regulatory sequences the proportions of T remain more or less the same while the presence of A decrease to a median value of 29-30% with a correspondent increase of C and G.
- CDS and exons show a further increase of strong bases, in particular G. In fact the median distribution of bases in CDS and exons sequences is the following:

	A	C	G	T
CDS	28,6	20,1	23,8	27,0
Exons	28,9	19,6	23,6	27,3

Genes

	A	C	G	T
Min.	8.78	6.45	6.69	13.58
1st Qu.	26.78	16.99	19.37	29.29
Median	28.45	18.34	20.86	31.93
Mean	28.47	18.77	21.33	31.43
3rd Qu.	30.38	20.07	22.60	34.00
Max.	44.48	41.75	50.29	45.45

Intergenes

	A	C	G	T
Min.	8.33	2.56	1.70	7.14
1st Qu.	31.79	13.97	14.04	31.68
Median	34.32	15.57	15.63	34.41
Mean	34.01	15.92	16.04	34.03
3rd Qu.	36.40	17.54	17.69	36.48
Max.	63.64	44.44	42.86	66.67

Exons

	A	C	G	T
Min.	2.63	2.56	1.27	2.63
1st Qu.	25.47	16.78	20.90	24.17
Median	28.93	19.64	23.66	27.27
Mean	29.01	19.94	23.73	27.33
3rd Qu.	32.39	22.73	26.44	30.46
Max.	64.29	50.63	60.00	62.79

CDS

	A	C	G	T
Min.	11.56	6.31	6.14	10.44
1st Qu.	26.59	18.54	22.30	25.23
Median	28.63	20.12	23.78	27.06
Mean	28.61	20.48	23.87	27.04
3rd Qu.	30.66	22.06	25.37	28.90
Max.	51.37	48.36	48.97	46.55

Introns

	A	C	G	T
Min.	4.40	1.21	2.73	8.22
1st Qu.	23.01	12.93	14.08	37.18
Median	26.67	15.62	16.67	40.65
Mean	26.93	15.72	16.68	40.68
3rd Qu.	30.61	18.29	19.19	44.21
Max.	53.49	45.65	53.40	65.52

Long introns

	A	C	G	T
Min.	7.77	3.09	4.04	12.73
1st Qu.	24.89	13.47	15.07	38.46
Median	26.96	15.26	16.90	40.37
Mean	27.73	15.15	16.83	40.29
3rd Qu.	30.30	16.76	18.45	42.42
Max.	50.00	39.77	53.40	62.50

UTR

	A	C	G	T
Min.	3.77	1.35	1.41	1.92
1st Qu.	25.59	13.87	13.92	31.43
Median	29.39	16.78	17.02	36.23
Mean	30.42	17.70	16.78	35.10
3rd Qu.	33.84	20.47	19.71	40.26
Max.	78.26	55.56	50.00	63.64

Regulatory sequences

	A	C	G	T
Min.	2.78	1.85	1.32	2.08
1st Qu.	25.72	13.99	12.68	29.45
Median	30.28	17.24	16.30	34.38
Mean	31.08	18.34	16.38	34.19
3rd Qu.	35.94	21.72	19.67	39.66
Max.	75.00	54.55	46.15	75.00

Tabella 4.4: Percentage of bases in sequences of different classes

4.2 Percentage of dichotomic classes

In this section I perform a statistical analysis on the dichotomic classes computed on the eight groups of sequences of the chromosome 1 of *A. thaliana* described in the previous section. As mentioned above, the aim is to study whether the information conveyed by dichotomic classes can characterize different portions of the genome. In order to accomplish the task, I code all the sequences into the three dichotomic classes and study the distributions of such binary sequences. In particular, I focus on their mean value, that is, the percentage of ones. Thus, for each sequence, we obtain 22 variables as reported in Table 4.1:

Here I report, as an example, mean percentage tables for normal sequences computed on both sense and antisense strand (median values are very similar and lead to the same conclusions). The whole set of tables are reported in appendix A . They show median and mean values of percentage of ”one” digits in dichotomic class binary string computed respectively on sense and antisense strands for

- normal sequences
- complementary sequences
- reverted sequences
- sequences undergone to Keto/Amino global transformation
- sequences undergone to Purine/Pyrimidine global transformation

4.2.1 Normal sequences

Sense strand

(see table 4.2.1) First of all we can notice that coding and non-coding sequences show a different behaviour:

- Non coding sequences (all the genome’s portions but Exons and CDS) show similar values for mean and median percentages of the same dichotomic class in and out of frame 1 and 2. That is, for example, $p0 = p1 = p2$. In some cases (such as UTR and regulatory sequences) this similarity is very very high.
- exons and CDS, the only sequence classes that undergo to transcription and translation processes, show different mean and median values in different frames. If we consider parity, for instance, we can see that $p0$ is similar to $p1$ but both of them are lower than $p2$

These observations can lead to the conclusion that the frame is important only for coding sequences

Tabella 4.5: Mean values of dichotomic class proportions computed in sense strand for each portion of the genome

	P0	R0	H0	P1	R1	H1	P2	R2	H2
Genes	55.19	38.87	48.00	55.86	38.66	48.18	56.84	38.56	47.17
CDS	50.13	44.09	51.88	51.63	41.74	54.78	58.67	38.81	45.90
Exons	52.13	42.17	50.58	52.71	39.93	51.44	56.26	39.47	47.41
Introns	61.21	34.85	38.74	60.40	33.06	38.33	59.29	32.82	37.83
Long Int	60.67	33.22	41.04	60.42	32.74	40.87	60.05	32.56	40.76
IG	60.62	31.48	49.19	60.53	31.43	49.07	60.42	31.49	49.13
UTR	60.22	35.27	44.55	60.01	35.24	44.23	60.15	35.25	44.31
Reg	59.67	35.86	44.08	59.72	35.84	43.56	60.19	36.53	43.45

Parity class:

- Parity percentage values for introns, intergenes, UTR and regulatory sequences are almost the same (the range is from 59% to 61%). They do not show differences in values in and out of frame.
- Genes show the same behaviour as non coding sequences but the values of their proportions are different (55-56%). This value can be considered as the weighted mean between coding and non-coding sequences of which it is made of.
- CDS and exons show a lower value in $p0$ and $p1$ (CDS: 50-51%; exons: 52%) and an higher value in $p2$ (CDS: 58,7%; exons:56,3%). CDS and exons percentages vary with frame in this way:

$$p0 < p1 < p2$$

Rumer class:

- Non-coding sequences (introns, intergenes, UTRs and regulatory) show almost the same percentage values in and out of frame.

- UTR and regulatory sequences show values between 35% and 36%, while introns percentages vary within 32,8% and 34,8%.

- Intergenes seems to have a specific Rumer percentage value that is around 31%

- Genes show the same behaviour as non coding sequences but the values of their percentages are around 39%). This value can be considered again as the weighted mean between coding and non-coding sequences of which it is made of.

- Coding sequences proportions vary with the frame in this way:

  $$r2 < r1 < r0$$

  Their values are remarkably higher than non-coding ones, in particular for what concerns $r0$ (CDS: 44%, exons: 42%).

Hidden class

- Non-coding sequences (introns, intergenes, UTRs and regulatory) show almost the same percentage values in and out of frame but with specific values for each sequence class: introns (38%) , intergenes (49,1%), UTRs and regulatory(44%)

- Once again coding sequences show a different pattern that take into account the frame:

  $$h2 < h0 < h1$$

  For what concerns exons, we can see again that $h0$ is very similar to $h1$ (50,6% and 51,4%) but higher than $h2$ (47,4%)

- Once again genes show the same behaviour as non coding sequences (they do not change with frame) but the values of their percentages (about 48%) can be considered as the weighted mean between coding and non-coding sequences of which it is made of.

- Finally it is interesting to underline that introns seem to minimize hidden class.

Antisense strand

Analyzing dichotomic classes percentages in antisense strand (see Table 4.2.1), we can observe a similar behaviour for coding and non-coding sequences with respect to what is described in sense strand:

- Genes and non coding sequences do not seem to consider frame, while exons and CDS do.
- Percentage values are different from the ones observed in sense strand. This is valid for all the sequence classes except intergenes.

  Surprisingly, intergenes show the same percentage values that we observed analysing sense strand
- For the other non-coding sequences (introns, long introns, UTR and regulatory) we can see that parity percentage values are similar to those computed in sense strand, while Rumer and hidden percentage values differ from sense strand ones.
- CDS and exons show again a frame-related behaviour:

$$p2a < p0a < p1a$$

$$r0a < r2a < r1a$$

$$h2a < h0a < h1a$$

Tabella 4.6: Mean values of dichotomic class proportions computed in antisense strand for each portion of the genome

	P0a	R0a	H0a	P1a	R1a	H1a	P2a	R2a	H2a
Genes	56.29	38.60	51.04	56.33	39.33	51.01	56.01	39.09	50.59
CDS	54.61	39.05	48.66	56.33	49.32	49.39	51.06	43.64	44.14
Exons	54.85	40.75	49.09	55.92	45.28	49.27	53.29	42.76	47.06
Introns	60.45	29.38	61.01	61.58	26.75	61.68	61.74	28.64	61.36
Long Int	60.57	29.15	58.66	60.92	28.43	58.85	60.94	29.04	58.78
IG	60.56	31.56	49.12	60.47	31.56	49.09	60.54	31.53	49.14
UTR	60.12	32.16	52.20	59.70	32.40	52.30	59.71	32.19	52.65
Reg	60.39	32.61	49.50	59.74	33.28	48.97	59.18	33.15	49.13

4.2.2 Complementary sequences

Complementary sequences are those sequences which undergo to strong/weak global transformation that is A is converted to T while C is converted to G (and viceversa). We applied this transformation to all the sequences considered before and then we computed dichotomic classes proportions again on sense and antisense strand (the relative tables are shown in Appendix A).

Sense strand

Also in complementary sequences we can see that coding and non-coding sequences show the different behaviour observed in normal sequences: in fact only exons and CDS show different mean and median values in different frames.

Proportions values for introns, long introns, intergenes, UTRs and regulatory sequences are almost the same in and out of frame for each dichotomic class.

Genes show the same behaviour as non coding sequences but the values of their proportions are different. This value can be considered as the weighted mean between coding and non-coding sequences of which it is made of.

CDS and exons proportions vary with frame in this way:

$$p0 < p2 < p1$$

$$r1 < r0 < r2$$

$$h1 < h0 < h2$$

Finally we can observe that proportion values are different from the one computed in normal sequences for all classes except for intergenes.

Antisense strand

We can make the same considerations as in sense strand. The only difference is the proportions pattern of CDS and exons:

$$p0a < p1a < p2a$$

$$r1a < r0a < r2a$$

$$h1a < h0a < h2a$$

We have to underline that proportions in intergenes sequences are exactly the same as in sense strand and in sense and antisense strand of normal sequence!!

$$p = 60\%, r = 31\%, h = 50\%$$

4.2.3 Reverted sequences

Computing dichotomic classes on a reverted sequence is like computing dichotomic class on the antisense strand of the complementary sequence. Therefore in this section we have a copy of the tables shown in the previous one (i.e: complementary sense strand is identical to reverted antisense strand and viceversa).

4.2.4 Keto/Amino global transformation sequences

Sequences that undergo to keto/amino global transformation convert A to C and G to C (and viceversa). We applied this transformation to all the normal sequences and then we computed dichotomic classes percentages again on sense and antisense strand (the relative tables are shown in Appendix A).

Sense strand

Also in this case we can see that only exons and CDS show different mean and median values in different frames.

Proportions values for introns, long introns, intergenes, UTRs and regulatory sequences are almost the same in and out of frame for each dichotomic class. In particular, percentages of UTRs and regulatory sequences are almost the same between them

Genes show the same behaviour as non coding sequences but the values of their percentages are different. This value can be considered as the weighted mean between coding and non-coding sequences of which it is made of.

CDS and exons percentages vary with frame in this way:

$$p0 < p2 < p1$$

$$r0 < r1 < r2$$

$$h0 < h1 < h2$$

Finally we can observe that percentage values are different from the one computed in normal and complementary sequences for all classes (also intergenes).

Antisense strand

We can make the same considerations as in sense strand. The only difference is the percentages pattern of CDS and exons:

$$p2a < p1a < p0a$$

$$r1a < r2a < r0a$$

$$h1a < h2a < h0a$$

Moreover, we can notice that intergenes percentages are the same computed in sense strand.

4.2.5 Purine/Pyrimidine global transformation sequences

Sequences that undergo to purine/pyrimidine global transformation convert A to G and C to T (and viceversa). We applied this transformation to all the normal sequences and then we computed dichotomic classes percentages again on sense and antisense strand (the relative tables are shown in Appendix A).

Sense strand

We can make the same general considerations done for keto/amino global transformation.

CDS and exons proportions vary with frame in this way:

$$p2 < p1 < p0$$

$$r2 < r0 < r1$$

$$h2 < h1 < h0$$

It is important to underline that percentages of intergenes are the same observed in sense and antisense strand of keto/amino global transformation sequences.

Antisense strand

The proportions pattern of CDS and exons is:

$$p1a < p0a < p2a$$

$$r2a < r0a < r1a$$

$$h0a < h2a < h1a$$

As expected the percentages pattern of intergenes is the same as in sense strand!!

$$p = 40\%, r = 67\%, h = 50\%$$

4.2.6 Comments

By these observations we could speculate that:

- Since dichotomic classes percentages vary with frame only for coding sequences, the frame is important only for coding sequences and dichotomic calsses can be useful to recognize it.
- All dichotomic classes can distinguish between coding and non-coding sequences, in fact their percentage values are always different. In particular, if we consider, for example, normal sequences we can see that:
  - Parity could discriminate also between sense and antisense coding sequences. In fact it is always around 60% for non-coding sequences (in both strands), while it is remarkably lower for CDS and exons and, moreover, it varies with strand: $p0$ in sense strand seem to be a bit lower (50-52%) than in antisense strand (54%). Similar differences can be found for $p1$ and $p2$.
  - Since Rumer and hidden percentage values vary between non-coding sequences and, moreover, between sense and antisense strand, they are useful to discriminate between the different classes of non-coding sequences (i.e: introns, long introns, intergenes, UTRs and regulatory)

  Similar considerations can be done for complementary, reverted, keto/amino and purine/pyrimidine transformated sequences.
- Since intergenes sequences show the same pattern of dichotomic classes percentages in sense and antisense strand, they could be considered as the expression of a random sequence (they don't carry any kind of information)
- Finally, for what concerns coding sequences, we can see that CDS show proportions values that differ with frame a bit more than exons ones. For example, if we consider parity, we can see that:

- $p0$ and $p1$ values are almost the same for exons ($p0 = 52,1\%$, $p1 = 52,7\%$) while they differ a bit more for CDS ($p0 = 50,1, p1 = 51,6$)
- $p2$ value is higher for CDS than for exons (58,7% and 56,3% respectively)

This kind of difference can be observed in Rumer and hidden proportions too. This could suggest that there is a sort of "union effect" that occurs when fragments of coding sequences (exons) join together to form a CDS.

4.3 Independence test

Now I want to check if the information provided by these properties is specific for dichotomic classes or simply comes from the proportion of bases in the original sequence. In the latter case, the information content should be always the same, regardless of the order of bases within the sequence if their proportions are kept constant. If, given a nucleotide sequence A, I perform a permutation of the basis without changing their number, I get a new sequence B with the same proportion of the four bases (A,C,G,T). However, the dichotomic classes binary strings computed from the two original sequences (A and B) will be different (for what concerns parity class we will get P_A and P_B. If the proportion of 1 in each binary string (P_A and P_B) is the same, then the information carried by dichotomic classes depends only on the proportion of bases. Otherwise we can state that dichotomic classes carry an additional information content, with respect to that carried by the proportion of bases in the original sequence.

For instance, define the random variable X as the parity of a dinucleotide for a given sequence. Then, X follows a Bernoulli distribution with parameter $\pi = P(X = 1)$, that is: $E(X) = \pi$ and $V(X) = \pi(1 - \pi)$. Then, we have the following null hypothesis:

$$H_0 : \pi = \pi_0$$

$$H_1 : \pi \neq \pi_0$$

where $\pi_0 = P(X = 1)$ under the assumption that the DNA sequence is an expression af an i.i.d. process, that is its sequence is randomly derived on the basis of a given proportion of the four bases. Now, we can see that:

$$\pi_0 = P(X = 1) = P(D) \in S_1 = \sum_{i=1}^{8} P(b_1)_i * P(b_2)_i \qquad (4.1)$$

where:

- D is the dinucleotide considered
- S_1 and S_0 are, respectively the group of dinucleotides that correspond to parity class 1 (S_1) and parity class 0 (S_0)
- b_1 and b_2 are, respectively, the bases in the first and in the second position of the dinucleotide.
- $P(b_1)$ and $P(b_2)$ are the probability of occurrences of the 4 nucleotides (T,C,A,G) in the first and second base, respectively.

The association scheme for the parity is presented in Table 4.7. Therefore, the possible differences observed between the original and *i.i.d.* sequences are not due to the proportion of bases and all the quantities derived from it (e.g. the GC content and the like).
For example, if we have the following probability distribution for the nucleotides: [1]

base	P
A	0.20
C	0.25
G	0.20
T	0.35

then

$$\begin{aligned}\pi_0 &= 0.20 \times 0.20 + 0.25 \times 0.20 + 0.20 \times 0.20 + 0.35 \times 0.20 + 0.20 \times 0.35 + \\ &+ 0.35 \times 0.35 + 0.20 \times 0.25 + 0.35 \times 0.25 = 0.53\end{aligned}$$

Now, if we take the usual sample mean $\hat{\pi}$ as the estimator of π we have that under the null hypothesis, $E(\hat{\pi}) = \pi_0$ and $V(\hat{\pi}) = \frac{\pi_0(1-\pi_0)}{n}$ where n is the length of binary sequence. Thus, we can use the test statistic Z

$$Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} .$$

[1] They are global (not position-dependent) proportions

Tabella 4.7: Partition of the 16 dinucleotides into parity groups.

i	1^{st} base (b_1)	2^{nd} base (b_2)	Dinucleotide group (D)	Parity
1	A	A	S_1	1
2	C	A		
3	G	A		
4	T	A		
5	G	T		
6	T	T		
7	G	C		
8	T	C		
9	A	G	S_0	0
10	C	G		
11	G	G		
12	T	G		
13	A	T		
14	C	T		
15	A	C		
16	C	C		

Z converges in distribution to a standard normal random variable so that the usual critical values can be used.

I have computed the p-values associated to the test for each sequence and for each dichotomic class. The results are shown in Figures 4.3 and 4.4, where we present the histograms of the p-values for the sequences analyzed.

Tabella 4.8: Percentages of p-values lower than 0.05

	Parity	Rumer	Hidden
Genes	17.94	13.24	8.93
CDS	28.28	14.96	5.80
Exons	11.45	5.80	4.00
Introns	4.31	1.50	1.20
Long introns	5.71	1.84	1.43
Intergenes	9.09	3.13	1.70
UTR	6.27	1.98	1.43
Regulatory	5.65	2.85	1.96

4.3.1 Comments

By looking at the histograms in Figures 4.3 and 4.4, we can see that:

- only coding sequences and genes show a pattern which indicates the presence of differences between the proportions values computed on original sequences and those computed on sequences whose properties belongs only to proportions of bases. Only these classes, in fact, show an important rate of p-value lower than 0.05. That could mean that only genes, exons and CDS are non-random sequences with an informative content.

- parity sequences seem to be more informative. They show, in fact, an higher rate of low p-values. This trend can be observed also in non-coding sequences where parity p-values seem to follow an uniform distribution while Rumer and hidden show fewer low p-values.

- If we compare graphs relative to exons and CDS we can see that the latter show an higher rate of low p-values. This effect could be related to the sample size. However, this observation could be also a further indication of the presence of a “union effect” that occurs when fragments of coding sequences (exons) join together to form a CDS.

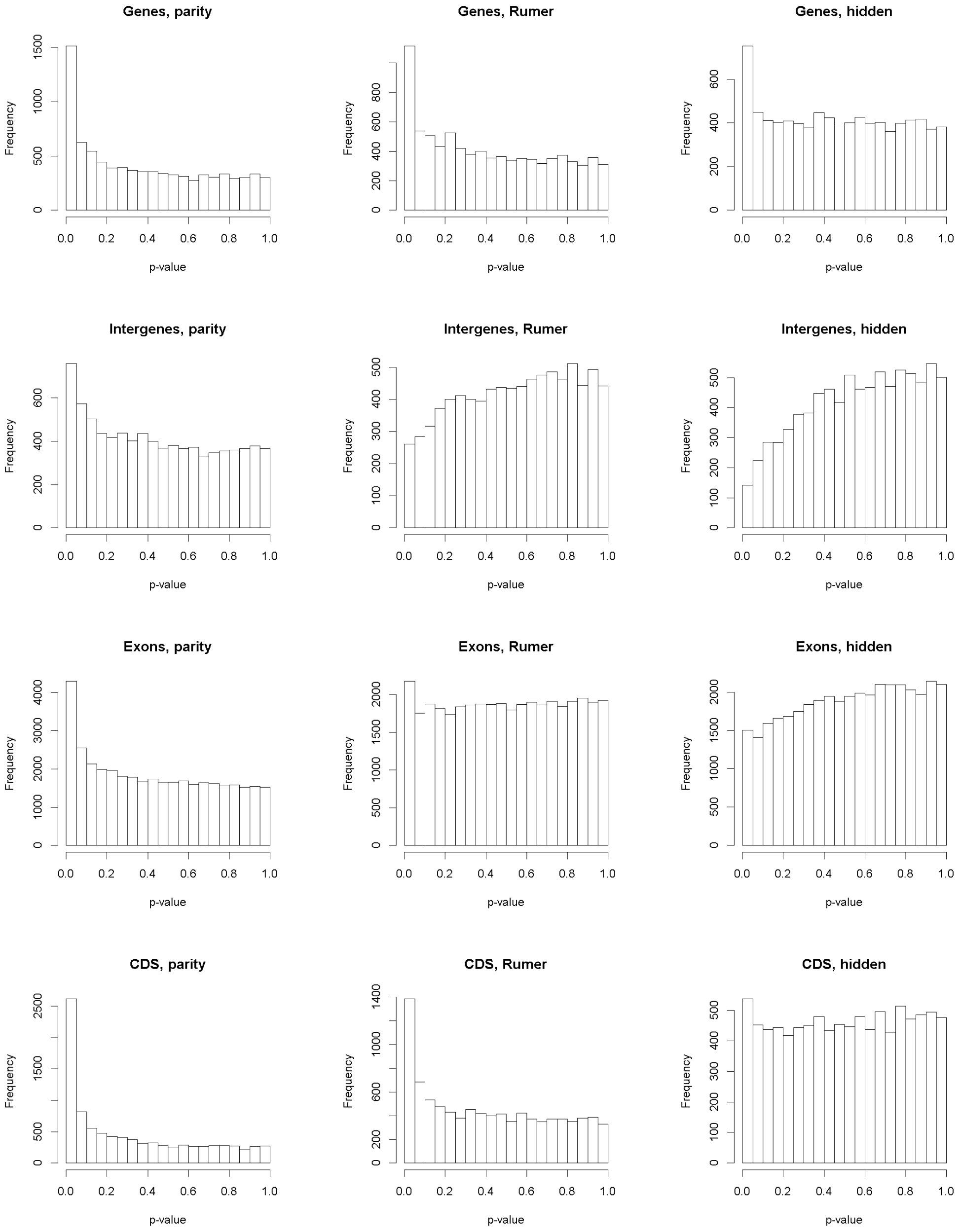


Figura 4.3: Histograms of the *p*-values associated to the independence test for genes, intergenes, exons and CDS

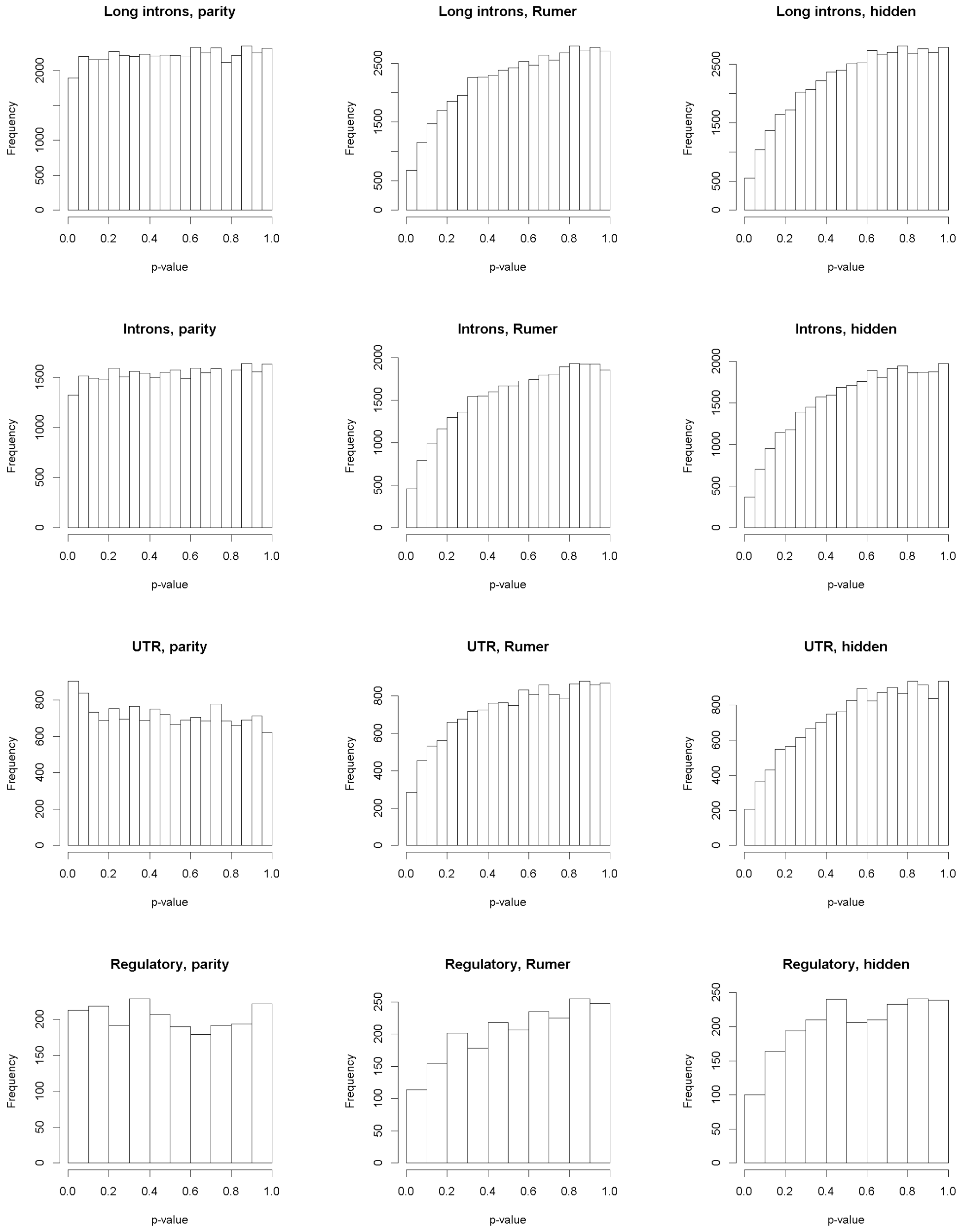


Figura 4.4: Histograms of the p-values associated to the independence test for introns,long introns, UTRs and regulatory sequences

The latter issue might be also studied by resorting to adjusted p-values and to false discovery rate estimation (see e.g. [11]). The overall results confirm the findings of [20] regarding the presence of correlations between dichotomic classes in coding sequences.

4.4 Conclusions

The analysis of bases distributions and percentage of dichotomic classes distributions is helpful in order to characterize different portions of the genome in *A.thaliana* chromosome 1 and it leads to the following considerations:

- Coding and non-coding sequences show different patterns of distribution. Since the percentages of dichotomic classes vary with frame only for coding sequences, we can conjecture that frame is important only for these kind of sequences (CDS and exons).
- Dichotomic classes seem to be useful in order to recognize coding sequences. In fact all the dichotomic classes can distinguish between coding and non-coding sequences: their mean values are always different.
- Parity class could discriminate also between sense and antisense coding sequences: while the mean percentage for non-coding sequences is always around 60% (in both strands), it is remarkably lower and strand dependent for CDS and exons.
- Rumer and hidden percentages vary in both strand between coding and non-coding sequences. Therefore they are useful in order to discriminate between the different groups of non-coding sequences (i.e: introns, long introns, intergenes, UTR and regulatory sequences). In particular, introns seem to minimize hidden class.
- Intergenes show a constant distribution pattern that vary neither with frame nor with strand. They always show:

$$p \simeq 60\%, \qquad r \simeq 31\%, \qquad h \simeq 49\%$$

- Finally, we can identify a sort of “union effect” that occurs when short fragments of coding sequences (exons) join together to form a CDS. In fact mean percentage values of CDS in each dichotomic class differ vary with frame more than those of exons. For example, if we consider parity, we can see that $p0$ and $p1$ are almost the same for exons ($p0$=52.1%, $p1$=52.7%) while for CDS we have $p0 = 50.1\%, p1 = 51.6\%$. Moreover the mean for $p2$ is higher for CDS than for exons (58.7% and 56.3% respectively). This kind of difference can be observed in Rumer and hidden percentages too.

- The results obtained from the independence tests show that the framework suggested by dichotomic classes is able to uncover the existence of significant correlations in those sequences that are involved in protein synthesis.

Capitolo 5

Discriminating different portions of the genome

In the previous chapter I showed how dichotomic classes can characterize the different portions of *A.thaliana* genome. It means they may carry a quantity of information higher than that explained by the proportion of bases. Therefore, I want to answer the following question: is the information carried by dichotomic classes useful in order to discriminate between two different portions of the genome?

5.1 Classification through Logistic Regression

I try to answer this question with the help of logistic regression models that is part of generalized linear models. Logistic regression allows to predict a discrete outcome, such as group membership, from a set of variables that may be continuous, discrete, dichotomous, or a mix of any of these. ([32] [28])

An explanation of logistic regression begins with an explanation of the logistic function, which, like probabilities, always takes on values between zero and one:

$$\pi(x) = \frac{e^{(\beta_0+\beta_1 x_1)}}{e^{(\beta_0+\beta_1 x_1)}+1} = \frac{1}{e^{-(\beta_0+\beta_1 x_1)}+1}$$

The dependent variable in logistic regression is usually dichotomous, that is, the dependent variable can take the value 1 with a probability of success π, or the value 0 with probability of failure $(1-\pi)$. This type of variable is called a Bernoulli (or binary) variable where:

$$Y \sim Ber(\pi) \qquad E(Y) = \pi \qquad V(Y) = \pi(1-\pi)$$

As mentioned previously, the independent or predictor variables, in logistic regression, can take any form, that is, logistic regression makes no assumption about the distribution of the independent variables. They do not have to be normally distributed, linearly related or of equal variance within each group. The relationship between the predictor and the response variables is not a linear function in logistic regression. It is used a link function, called "logit transformation of π":

$$\pi = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k}}$$

$$logit[\pi] = log\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

The goal of logistic regression is to correctly predict the category of outcome for individual cases using the most parsimonious model. To accomplish this goal, a model is created that includes all predictor variables that are useful in predicting the response variable.

The logistic regression method is used for example to make a classification. The response variable Y can assume two values that, for expository convenience, we could call A and B. Let's suppose that a certain observation belongs to class A with probability π.

Starting from a set of data, namely a set of observations that are known, the coefficients of the model are calculated. Then, examining new observations that we want to classify on the basis of mere knowledge of the predictors, we calculate the value of π and assign the observation to the class A if the probability π exceeds a certain threshold (s).

Classification Classification models are tested by comparing the predicted values to known target values in a set of test data. The test data must be compatible with the data used to build the model and must be prepared in the same way that the build data was prepared. Typically the build data and test data come from the same historical data set. A percentage of the records is used to build the model; the remaining records are used to test the model.Test metrics are used to assess how accurately the model predicts the known values. If the model performs well, it can then be applied to new data to predict the future.

In order to assess the prediction ability of the model it is useful to build a confusion matrix that displays the number of correct and incorrect predictions made by the model compared with the actual classifications in the test data. The matrix is $n \times n$, where n is the number of classes.

The following table shows a confusion matrix for a binary classification model. The rows present the number of predicted classifications in the test data. The columns present the number of observed classifications made by the model.

		observed		
		0	1	
predicted	0	n_{00}	n_{01}	$n_{0\cdot}$
	1	n_{10}	n_{11}	$n_{1\cdot}$
		$n_{\cdot 0}$	$n_{\cdot 1}$	N

where $N = n_{00} + n_{01} + n_{10} + n_{11}$ is the sample size of the test set.

$$W = \frac{n_{10} + n_{01}}{N} 100$$

I will use W, as the misclassification rate.

5.2 Sequence classification of *A.thaliana* chromosome 1

Sequence prediction can be seen as a classification problem:

$$Y = f(X_1, ..., X_p) + \epsilon$$

where Y is a dichotomic response variable and $(X_1, ..., X_p)$ is a set of predictors.

I would like to predict if a given sequence belongs $(Y = 1)$ or not $(Y = 0)$ to a specific portion of the genome (i.g: exons, introns, etc.) using dichotomic classes percentages as predictors.

- $Y \sim \text{Ber}(\pi)$;

$$g\left(E[Y|X]\right) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i \ . \tag{5.1}$$

where $g(\pi) = \log \frac{\pi}{1-\pi}$ is the link function.

I create logistic regression model in order to verify if some combination of the variables in the datasets can discriminate the different sequence classes. I take into account the following pairwise comparisons:

1. Exons vs Introns
2. CDS vs Long Introns
3. Exons vs CDS
4. Genes vs intergenes
5. Exons vs UTR
6. Introns vs UTRs

For each of these couple I created six logistic regression models that differs in number and kind of explanatory variables considered, as shown in Table 5.1 and 5.2.

For each model I joined the two classes dataset. Then I divided the records into two random groups: the first, containing 80% of records, is used to fit the model, while the other (containing the remaining 20% of records) is used to make predictions and build the misclassification table (out of sample analysis).

Tabella 5.1: Logistic regression models created for each couple of sequence classes

Model	Regressors
P	p0, p1, p2
R	r0, r1, r2
H	h0, h1, h2
Sen	h0, h1, h2, p0, p1, p2, r0, r1, r2
Anti	h0a, h1a, h2a, p0a, p1a, p2a, r0a, r1a, r2a
Tot	h0, h1, h2, p0, p1, p2, r0, r1, r2, h0a, h1a, h2a, p0a, p1a, p2a, r0a, r1a, r2a
B	A, C, G, T

Tabella 5.2: Variables included in each dataset

Name	**Description**
$p0, r0, h0$	mean value for parity, Rumer, hidden classes, in frame
$p1, r1, h1$	mean value for parity, Rumer, hidden classes, out of frame 1
$p2, r2, h2$	mean value for parity, Rumer, hidden classes, out of frame 2
$p0a, r0a, h0a$	mean value for parity, Rumer, hidden classes, antisense strand in frame
$p1a, r1a, h1a$	mean value for parity, Rumer, hidden classes, antisense strand out of frame 1
$p2a, r2a, h2a$	mean value for parity, Rumer, hidden classes, antisense strand out of frame 2

5.2.1 Exons vs Introns

First, I want to assess if I can discriminate between intron and exons using, as regressors, the sets of variable defined in Table 5.1. I proceeded as follows:

- I create a dataset with both the sequences Then I define a new binary variable (named ”class”) that takes value 1 if the sequence is an intron and 0 if the sequence is an exon.
- I randomly divided th 68212 records into two groups:
  - The first group composed by 53212 sequences was used in order to build a logistic regression model
  - The second group composed by 15000 sequences was used for out of sample analysis:
    - I compute the estimation of π
    - I set a threshold value ($s = 0.5$)
    - If $\pi > s$ then I consider the sequence as an intron, otherwise I consider it as an exon.
    - Finally I build the confusion matrix and compute the misclassification rate.
- I repeated this scheme for all the seven models defined in Table 5.1 and obtained the results as shown in Table 5.3 and Figure 5.1

Tabella 5.3: Misclassification rate for each model comparing exons and introns

Model	Misclassification rate (%)
P	15.72
R	29.34
H	21.39
Sen	8.53
Anti	7.17
Tot	5.7
B	5.81

We can see that:

- Models Sen, Anti, Tot and B discriminate with a good accuracy between intron and exon sequences. We can see that each of these model show a misclassification rate lower than 10%. The best models are tot and B with an error rate around 5.75%
- P, H and R models show higher misclassification rates. However parity class percentages seem to discriminate better between intron and exon sequences than other dichotomic classes

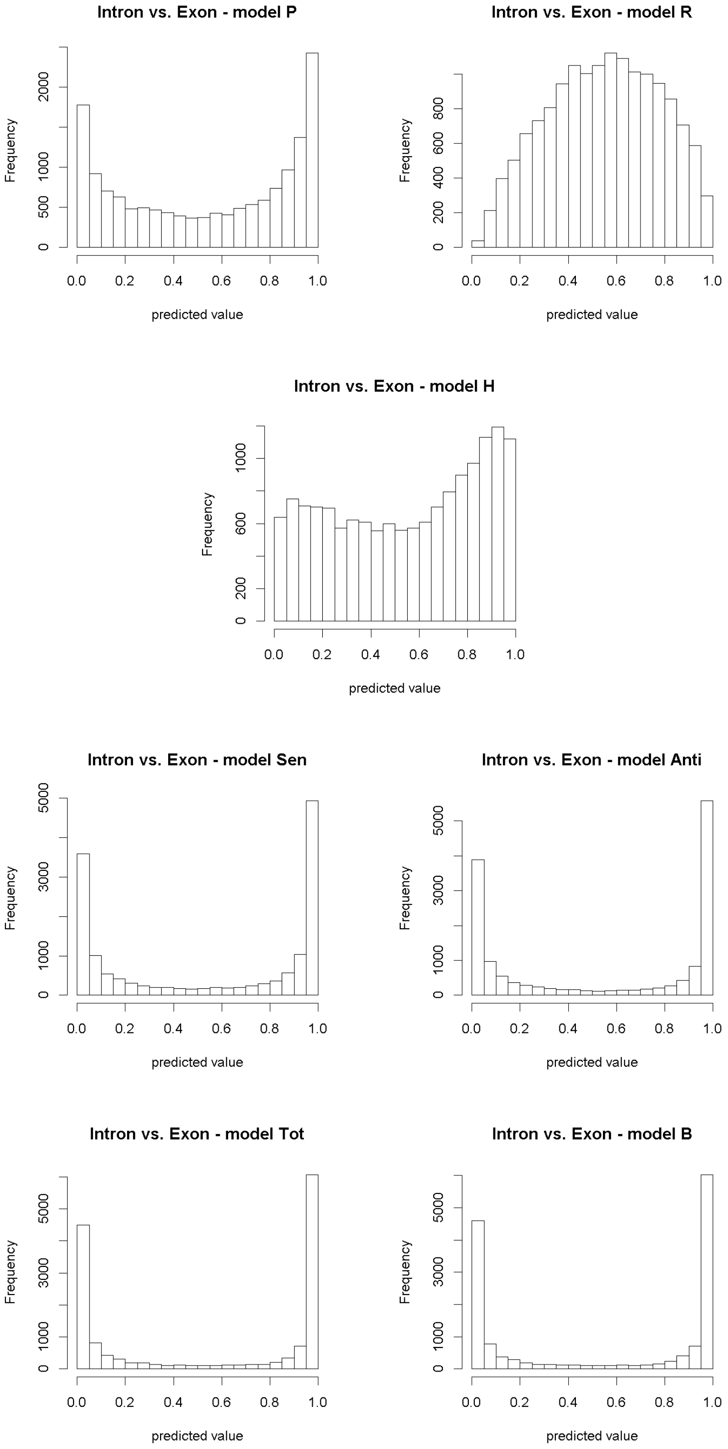


Figura 5.1: Sequence prediction: introns vs. exons

5.2.2 CDS vs. Long introns

I randomly divided the 14794 records into two groups and used 3000 sequences for out of sample analysis. The results are summarized in Table 5.4 and Figure 5.2.

Tabella 5.4: Misclassification rate for each model comparing CDS and long intron sequences

Model	Misclassification rate (%)
P	3.9
R	12.37
H	7.63
Sen	1.93
Anti	1.33
Tot	1.27
B	1.77

- All the models can discriminate with a good accuracy between the two groups of sequences. The worst model is R-model, with a misclassification rate equal to 12.37%. All the other models show a rate of error lower than 10%.
- In particular sen, anti, tot and B models discriminate with an excellent accuracy (lower than 2%). The best model is tot (m.r.=1.27%).

Once again we can see that comparison between CDS and long introns presents enhances the features observed in the comparison between exons and introns.

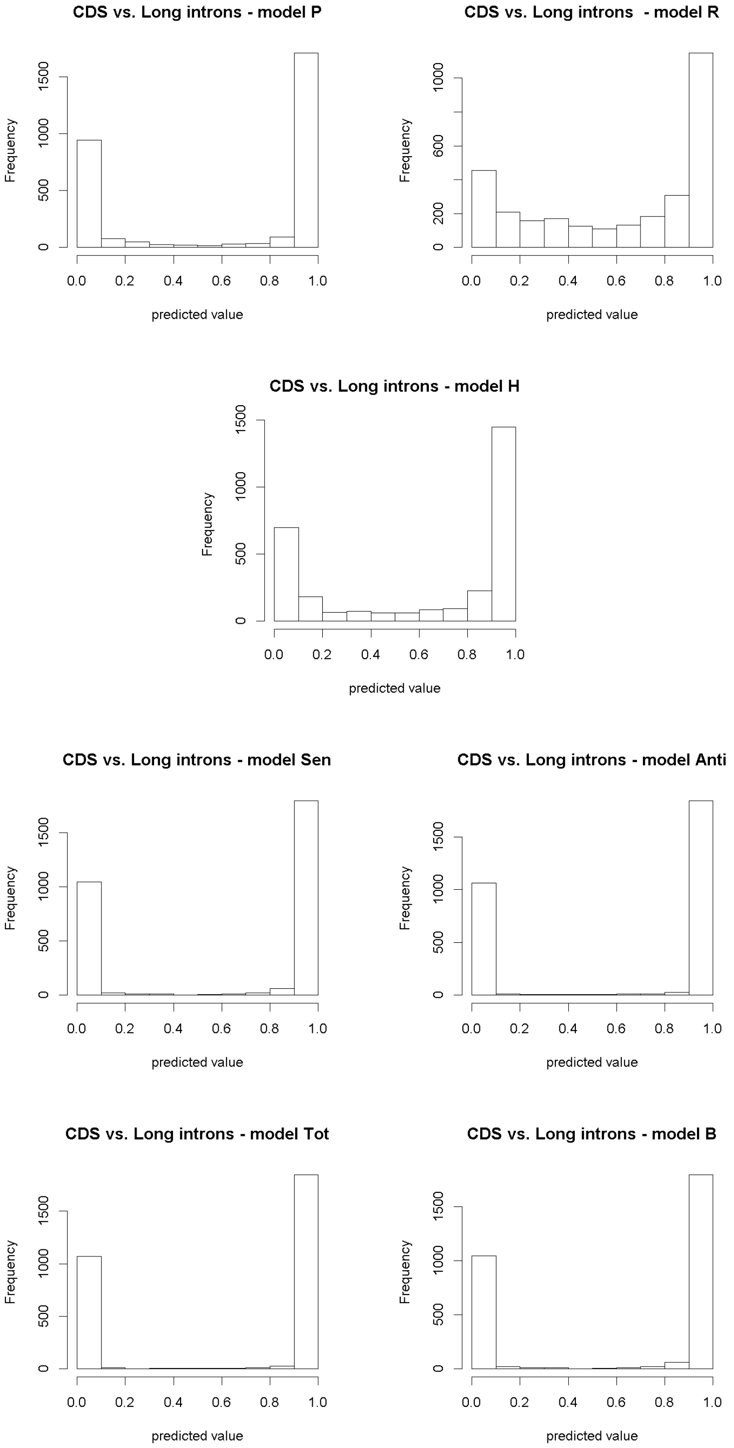


Figura 5.2: Sequence prediction: CDS vs. long introns

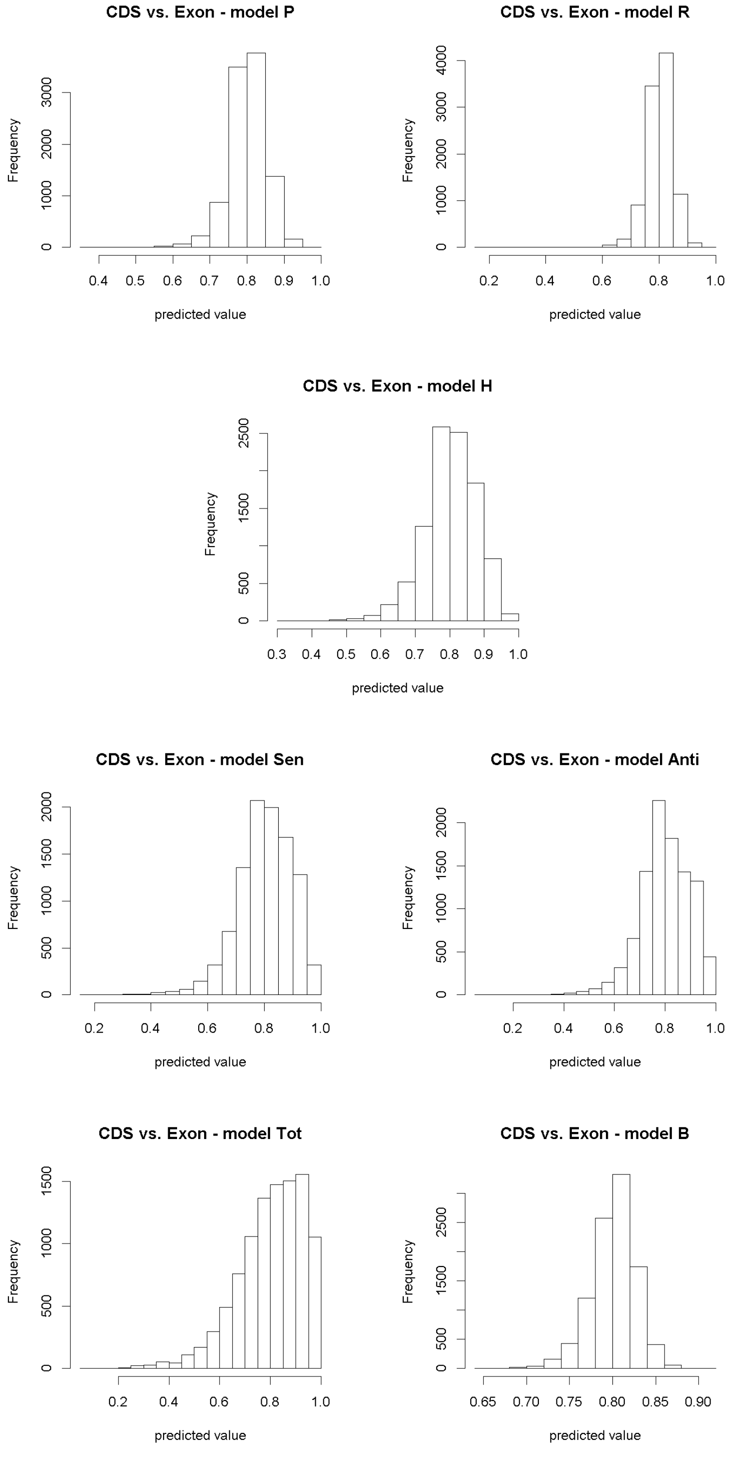
5.2.3 Exons vs. CDS

I randomly divided th 46811 records into two groups and used 10000 sequences for out of sample analysis. The results are summarized in Table 5.5 and Figure 5.3

Tabella 5.5: Misclassification rate for each model comparing CDS and exons

Model	Misclassification rate (%)
P	19.67
R	19.57
H	19.89
Sen	20.23
Anti	20.27
Tot	21.71
B	80.45

All the models show a similar misclassification rate (around 20%). As expected, none of the models is good in order to discriminate between CDS and exon sequences (remember that CDS are obtained by joining different exons present within the same gene).


CDS vs. Exon - model P

CDS vs. Exon - model R

CDS vs. Exon - model H

CDS vs. Exon - model Sen

CDS vs. Exon - model Anti

CDS vs. Exon - model Tot

CDS vs. Exon - model B

Figura 5.3: Sequence prediction: exons vs. CDS

5.2.4 Genes vs. Intergenes

I randomly divided th 16778 records into two groups and used 3000 sequences for out of sample analysis. The results are summarized in Table 5.6 and Figure 5.4

Tabella 5.6: Misclassification rate for each model comparing genes and inergenes

Model	Misclassification rate (%)
P	15.27
R	19.4
H	39.47
Sen	15
Anti	14.43
Tot	14.2
B	13.23

The misclassification rates in these model are quite high, in fact they vary from 13,23% (B-model) to 19,4% (R-model). It is interesting to notice that H-model is very bad, in fact it shows a really high misclassification rate (39,47%)

Genes vs. Intergenes - model P

Genes vs. Intergenes - model R

Genes vs. Intergenes - model H

Genes vs. Intergenes - model Sen

Genes vs. Intergenes - model Anti

Genes vs. Intergenes - model Tot

Genes vs. Intergenes - model B

Figura 5.4: Sequence prediction: Genes vs. Intergenes

5.2.5 Exons vs. UTRs

I randomly divided the 51976 records into two groups and used 10000 sequences for out of sample analysis. The results are summarized in Table 5.7 and Figure 5.5

Tabella 5.7: Misclassification rate for each model comparing exons and UTRs

Model	Misclassification rate (%)
P	14.98
R	25.8
H	27.16
Sen	12.6
Anti	13.63
Tot	11.9
B	10.66

Misclassification rates vary from 10% to 15% except for R and H-models that high shows misclassification rates (25,8 and 27,16% respectively). One again parity class percentages seem to be better than Rumer and hidden's ones when coding sequences (exon and CDS) are involved.

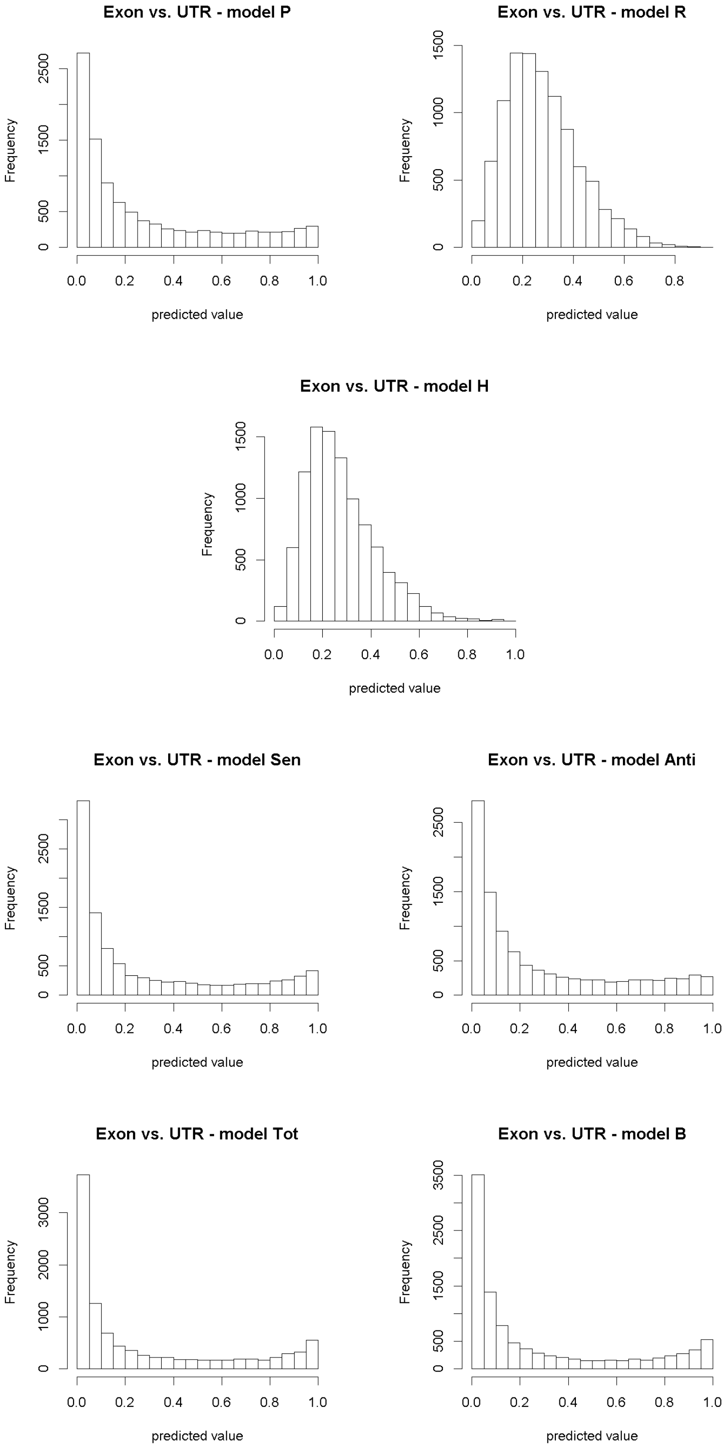


Figura 5.5: Sequence prediction: exons vs. UTRs

5.2.6 Introns vs. UTRs

I randomly divided the 45090 records into two groups and used 10000 sequences for out of sample analysis. The results are summarized in Table 5.8 and Figure 5.6

Tabella 5.8: Misclassification rate for each model comparing UTR and Intron sequences

Model	Misclassification rate (%)
P	31.78
R	31.27
H	28.62
Sen	26.09
Anti	24.05
Tot	20.2
B	26.67

In these models misclassification rates are very high, varying from 20,2 % (Tot-model) to 31,78 (P-model). It is interesting to underline that, in this case, H-model is better than P and R ones. In fact it happens every time introns are considered.

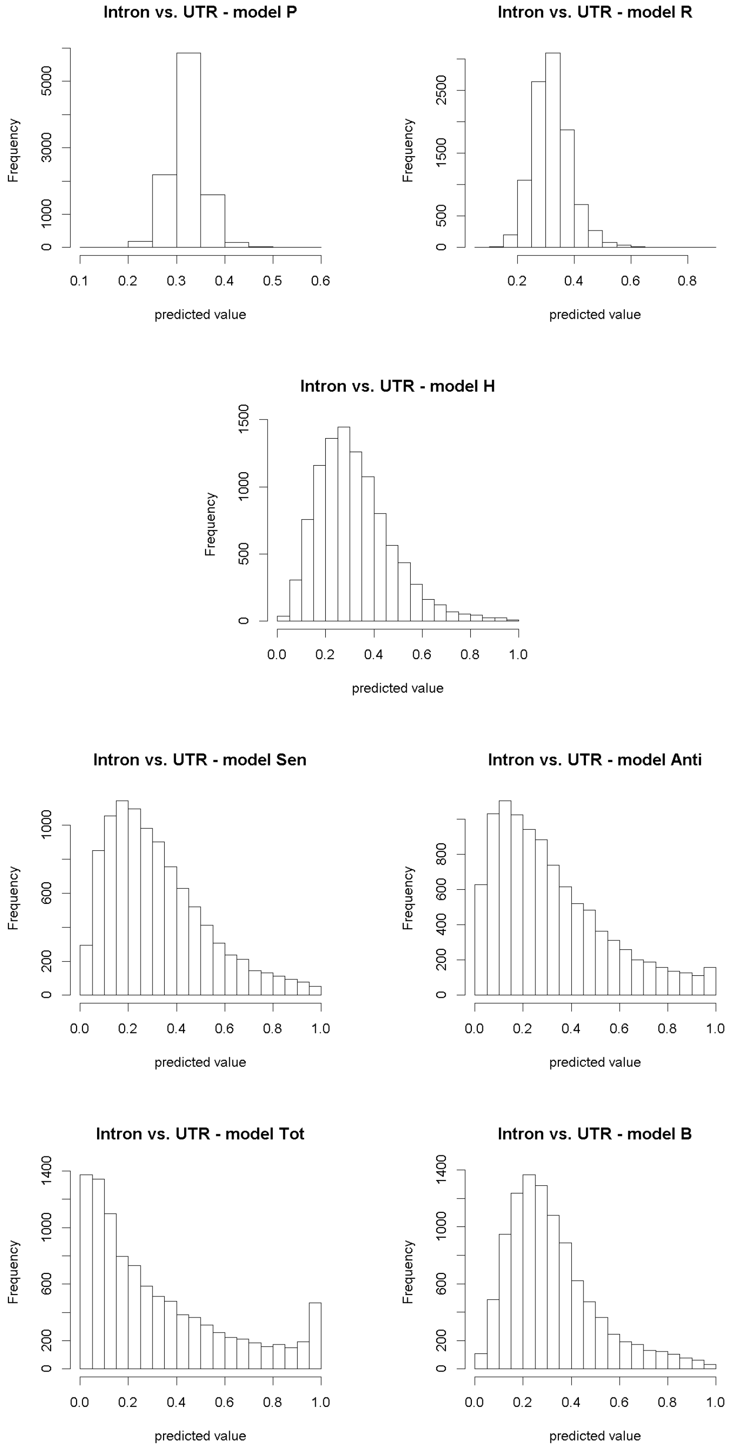


Figura 5.6: Sequence prediction: introns vs. UTRs

5.3 Comments

- Models involving exons and CDS (sequences within a gene that are transcripted and translated) show an good discrimination rate. In fact, it is possible to discriminate with a good sensitivity exons from introns and, overall, CDS from Long introns.
  - This could mean that *A.thaliana* cell has to distinguish between coding and non coding region within a DNA sequence. The discrimination power grows when we consider CDS instead of exons indicating, as we have already seen before, a sort of “joined effect” occuring when different exons join to form a CDS
  - It is evident how the percentage of bases is, itself, sufficient in order to achieve the best discrimination rate.
    - The only exception is the *Tot*-model that uses as regressors all the percentages of dichotomic classes computed both in sense and antisense strand.
    - Anyway, the gain of sensitivity is not so much to overcome the parsimony of the B-model (3 vs. 18 regressors)

- P-models are usually better than R and H-models. This could mean that parity class could discriminate better between the different portions of a gene.

- H-model (usually better than R-model) applied in order to discriminate between genes and intergenes seems to be very unhelpful. On the contrary it is more sensitive in discriminating between CDS and long introns. This could mean that hidden class represent a mathematical structure meaningful only to recognize different portions of coding sequences

- Because the model B is almost always the best model, the computation of dichotomic classes seems not to be useful for the discrimination between the different portions of the genome, since it seems to be sufficient the computation of the proportion of bases.

Capitolo 6

Dependence analysis

The analysis conducted in the previous chapter showed that the dichotomic classes carry no informational advantage for classification purposes if compared to proportions of bases. However, the information content of the dichotomic classes could detect the presence of short-range dependence structures within genome sequences. Therefore, in this chapter, I want to check if there is any dependence structure in the dichotomic classes within different portions of a gene.

In information theory and coding theory with applications in computer science and telecommunication, error detection and correction or error control are techniques that enable reliable delivery of digital data over unreliable communication channels. Many communication channels are subject to channel noise, and thus errors may be introduced during transmission from the source to a receiver. Error detection techniques allow detecting such errors, while error correction enables reconstruction of the original data.

The general idea for achieving error detection and correction is to add some redundancy (i.e., some extra data) to a message, which receivers can use to check consistency of the delivered message, and to recover data determined to be corrupted. Error-detection and correction schemes can be either systematic or non-systematic: In a systematic scheme, the transmitter sends the original data, and attaches a fixed number of check bits (or parity data), which are derived from the data bits by some deterministic algorithm. If only error detection is required, a receiver can simply apply the same algorithm to the received data bits and compare its output with the received check bits; if the values do not match, an error has occurred at some point during the transmission. In a system that uses a non-systematic code, the original message is transformed into an encoded message that has at least as many bits as the original message.

Redundancy and degeneracy are two main features of a communication code with an error detection and correction system. Since the genetic code is redundant and degenerate it is supposed to have such a system: it is a still open challenge linking information theory and biology. Moreover it is known that DNA-polymerase shows a proofreading ability in order to detect and correct some point mutations during DNA-replication process. When an error is detected, these polymerases halt the process of DNA replication, work backward to remove nucleotides from the daughter DNA chain until it is apparent that the improper nucleotide is gone, and then reinitiate the forward replication process.

The existence of a coding mechanism for error correction/detection implies some kind of dependence inside data and several studies have highlighted the presence of fractal long-range correlations in nucleotide sequences. However, error detection and correction should act at a local level. The existence of a dependence structure in the dichotomic classes has been demonstrated [19, 20].

The objective of my analysis is to assess if there is a dependence structure in the genome sequence of *A.thaliana* and if it is present within all the different portions of the genome. Moreover I would like to assess if dichotomic classes can detect such dependence structure.

In order to try to answer this question I will perform the same analysis as in [19, 20] applied to different portions of the genome.

6.1 Dependence measures

In statistics, dependence refers to any statistical relationship between two or more random variables. Formally, dependence refers to any situation in which random variables do not satisfy a mathematical condition of probabilistic independence:

Two events A and B are independent if and only if their joint probability equals the product of their probabilities:

$$P(A \cap B) = P(A)P(B)$$

Similarly, Two random variables X and Y are independent if and only if for every a and b, the events $\{X \leq a\}$ and $\{Y \leq b\}$ are independent events (as defined above). That is, X and Y with cumulative distribution functions $F_X(x)$ and $F_Y(y)$, and probability densities $f_X(x)$ and $f_Y(y)$, are independent if and only if the joint random variable (X,Y) has a cumulative distribution function

$$F_{X,Y}(x,y) = F_X(x)F_Y(y),$$

or equivalently, a joint density

$$f_{X,Y}(x,y) = f_X(x)f_Y(y).$$

There is an extensive literature on how to measure dependence, mainly on the basis of the distance between the joint distribution of the data and the product of the marginal distributions, where the latter distribution assumes the property of independence. [30, 2, 52, 37, 46, 25, 13].

The most commonly used measures of dependence and test statistics belongs to ”correlation” function. This is motivated by linear relations involving continuous variables and/or Gaussian processes. These measures tend to fail when variables are discrete, or in detection, as when they face nonlinear, or non-Gaussian processes. The currently dominant measures tend to be functions of only one or two moments of the underlying processes. While this has the advantage of simplicity, it can mislead when distinctions between the tail areas and higher order moments are germane. Thus, it is clearly desirable for measures of association and dependence to be robust towards possible (but unknown) nonlinearities and non-Gaussian processes.

In recent years there have been developed several different methods and indicators statistics of dependence, using the concepts of entropy borrowed from information theory. Some of them are finalized to measure the level of dependence, autocorrelation and irregularities of the fluctuations of a single series or of multiple series between them (dependence,correlation and synchronization).

Examples of other “well-informed” measures include the moment generating and characteristic functions, as well as many entropy functionals developed in information theory. Entropies are defined over the space of distributions which form the bases of independence/dependence concepts in both continuous and discrete cases. Entropy is also *dimensionless* as it applies seamlessly to univariate and multivariate contexts.

In information theory, entropy is a measure of the uncertainty associated with a random variable [34]. In this context, the term usually refers to the Shannon entropy, which quantifies the expected value of the information contained in a message. Equivalently, the Shannon entropy is a measure of the average information content one is missing when one does not know the value of the random variable. The concept was introduced by Claude E. Shannon in his 1948 paper *“A Mathematical Theory of Communication”* [47]. The Shannon entropy of a random variable (r.v) X is defined as:

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

In this chapter consider three measures of dependence

- Chi-squared test
- S_ρ (a normalized variant of the Bhattacharya-Hellinger-Matusita distance)
- Average mutual information

6.1.1 Chi-squared test for independence

Chi-squared test allows us to study the relationship between two categorical variables. In particular, the chi-square test is used to test the null hypothesis that the variables, indicated with X and Y are independent.

Given the double random variable (X, Y), consider the following joint probability distribution:

	y_1	$\cdots$	y_k	$\cdots$	y_c	
x_1	π_{11}	$\cdots$	π_{1k}	$\cdots$	π_{1c}	π_{10}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_j	π_{j1}	$\cdots$	π_{jk}	$\cdots$	π_{jc}	π_{j0}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_r	π_{r1}	$\cdots$	π_{rk}	$\cdots$	π_{rc}	π_{r0}
	π_{01}	$\cdots$	π_{0k}	$\cdots$	π_{0c}	1

where:

- $x_1, x_2, ..., x_j, ..., x_r$ and $y_1, y_2, ..., y_k, ..., y_c$, are, respectively, the possible realizations of X and Y;
- π_{jk} is the joint probability that the v.a. X takes the value x_j and v.a. Y takes the value y_k:

$$\pi_{jk} = P(X = x_j, Y = y_k), \qquad\qquad j = 1, ..., r, \;\; k = 1, ..., c;$$

- π_{j0} is the marginal probability that the v.a. X takes the value x_j (for any value of Y), or

$$\pi_{j0} = P(X = x_j) = \sum_{k=1}^{c} \pi_{jk} \qquad\qquad j = 1, ..., r;$$

- π_{0k} is the marginal probability that the v.a. Y takes the value y_k ((for any value of X), or

$$\pi_{0k} = P(Y = y_k) = \sum_{j=1}^{r} \pi_{jk} \qquad\qquad k = 1, ..., c.$$

The null hypothesis is that X and Y are independent, namely:

$$P(X = x_j, Y = y_k) = P(X = x_j)P(Y = y_k)$$

or, equivalently, that:

$$\pi_{jk} = \pi_{j0} \cdot \pi_{0k}$$

with $j = 1, ..., r$ $\qquad\qquad$ $k = 1, ..., c.$
Therefore, the problem can be formalized as follows:

$$\begin{aligned} H_0 : \forall j, k &\qquad \pi_{jk} = \pi_{j0} \cdot \pi_{0k}, \\ H_1 : \forall j, k &\qquad \pi_{jk} \neq \pi_{j0} \cdot \pi_{0k}. \end{aligned}$$

Given a random sample from a. double v.a (X, Y), consider the following contingency Table ($r \times c$):

	y_1	$\cdots$	y_k	$\cdots$	y_c	
x_1	n_{11}	$\cdots$	n_{1k}	$\cdots$	n_{1c}	n_{10}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_j	n_{j1}	$\cdots$	n_{jk}	$\cdots$	n_{jc}	n_{j0}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_r	n_{r1}	$\cdots$	n_{rk}	$\cdots$	n_{rc}	n_{r0}
	n_{01}	$\cdots$	n_{0k}	$\cdots$	n_{0c}	n

where :

- $x_1, x_2, ..., x_j, ..., x_r$ and $y_1, y_2, ..., y_k, ..., y_c$, are, respectively, the possible realizations of X and Y;

- n_{jk} is the frequency with which it is presented in the sample pair (x_j, y_k);
- n_{j0} is the marginal frequency of the sample realization x_j, for any value of Y:

$$n_{j0} = \sum_{k=1}^{c} n_{jk} \qquad\qquad j = 1, ..., r$$

- n_{0k} is the marginal frequency of the sample realization y_j, for any value of X:

$$n_{0k} = \sum_{j=1}^{r} n_{jk} \qquad\qquad k = 1, ..., c$$

The maximum likelihood estimators $\hat{\pi}_{jk}, \hat{\pi}_{j0}, \hat{\pi}_{0k}$ for the probability π_{jk}, π_{j0} andπ_{0k}, coincide with the corresponding sample relative frequency, as follows:

$$\hat{\pi}_{jk} = \frac{n_{jk}}{n}; \qquad\qquad \hat{\pi}_{j0} = \frac{n_{j0}}{n}; \qquad\qquad \hat{\pi}_{0k} = \frac{n_{0k}}{n}$$

Moreover, in case the hypothesis of independence between X and Y is true, the following relationship is expected to hold:

$$\hat{\pi}^0_{jk} = \hat{\pi}_{j0} \cdot \hat{\pi}_{0k}$$

or, multiplying both sides by n:

$$n^0_{jk} = \frac{n_{j0} \cdot n_{0k}}{n}, \qquad\qquad j = 1, ..., r; k = 1, ..., c.$$

N.B. "0" symbol in the apex means that we are considering H_0 as true

Therefore, the test of independence between X and Y can be conducted on the quantities:

$$(n_{jk} - n^0_{jk})^2, \qquad\qquad j = 1, ..., r; k = 1, ..., c$$

that is, on the squared distances between the sampling frequency and the corresponding expected frequencies in the case of independence.

If the differences between n_{jk} and n^0_{jk} are not too high, we will accept the H_0 hypothesis of independence between X and Y, otherwise we will have to reject it.. In particular, the test statistics is as follows:

$$Y_0 = \sum_{j=1}^{r} \sum_{k=1}^{c} \frac{(n_{jk} - n^0_{jk})^2}{n^0_{jk}},$$

with

$$n^0_{jk} = \frac{n_{j0} \cdot n_{0k}}{n}$$

If H_0 is true Y_0 converges in distribution to a χ^2 r.v. with $(r-1)(c-1)$ degrees of freedom.

$$Y_0 \to \chi^2_{(r-1)(c-1)}$$

Once the significance level of the test is fixed, we have the following decision rule:

$$A : Y_0 < y_{(g;\alpha)}, \qquad\qquad R : Y_0 \leq y_{(g;\alpha)},$$

where $y_{(g;\alpha)}$ is the above centile of the χ^2 distribution with $g = (r{-}1)(c{-}1)$ d.f.

6.1.2 Mutual information

A second measure of dependence between two random variables X and Y is given by the mutual information [6]:

$$I(X,Y) = \sum_i \sum_j p(x_i, y_j) log_2 \frac{p(x_i, y_j)}{p(x_i)p(y_j)}$$

If the two variables are independent, the mutual information between them is zero. If the two are strongly dependent, e.g., one is a function of another, the mutual information between them is large. There are other interpretations of the mutual information; for example, the stored information in one variable about another variable, and the degree of the predictability of the second variable by knowing the first. Clearly, all these interpretations are related to the same notion of dependence and correlation [40].

The previous equation can be rewritten as follows:

$$
\begin{aligned}
I(X,Y) &= \sum_i \sum_j p(x_i,y_j) log_2 \frac{p(x_i,y_j)}{p(x_i)p(y_j)} \\
&= \sum_i \sum_j p(x_i,y_j) log_2 p(x_i,y_j) + \sum_i \sum_j p(x_i,y_j) log_2 \frac{1}{p(x_i)} + \\
&+ \sum_i \sum_j p(x_i,y_j) log_2 \frac{1}{p(y_j)} + \\
&= \sum_i \sum_j p(x_i,y_j) log_2 p(x_i,y_j) - \sum_i p(x_i) log_2 p(x_i) - \sum_j p(y_j) log_2 p(y_j) \\
&= -H(X,Y) + H(X) + H(Y)
\end{aligned}
$$

ovvero

$$
\begin{aligned}
I(X,Y) &= H(X) + H(Y) - H(X,Y) \\
&= H(X) - H(X|Y) \\
&= H(Y) - H(Y|X)
\end{aligned}
$$

Therefore, it is clear that the mutual information between two random variables is nothing more than the reduction of uncertainty of a variable due to knowledge of the other. If the knowledge of Y reduces our uncertainty about X, then we say that Y carries information about X.

If X and Y are independently distributed, ie if $p(x,y) = p(x)p(y)$,the mutual information between the two variables is zero. In addition, this measure is symmetric, that is

$$I(X,Y) = I(Y,X)$$

and is always non-negative.

Mutual information provides an indication of the link between two variables random X and Y. In particular it:

- assumes the value zero if only if $p(x,y) = p(x)p(y)$, that is, if the variables X and Y are independent;
- is always non-negative, then $I(X,Y) > 0$;
- in the continuous case we have that $I(X,Y) = +\infty$ if $Y = g(X)$ that is the indicator tends to positive infinity if there is a perfect relation, although not linear, between X and Y.

Mutual information has been proposed, for example, as a criterion on which base test of independence and for the study of the level of dependency (not linear) in time series [10], [9].

In the present work I will compute mutual information between two dichotomic class sequences such that:

$$I(X_t, Y_{t+k}) = \sum_i \sum_j p(x_t, y_{t+k}) log_2 \frac{p(x_t, y_{t+k})}{p(x_t)p(y_{t+k})}$$

where

- $X_t(\cdot)$ is a random process that measures which digit appears at position t of the first dichotomic class sequence
- $Y_{t+k}(\cdot)$ is a random process that measures which digit appears at position $t + k$ of the second dichotomic class sequence
- (X_t, Y_{t+k}) is the bivariate random process that measures the joint appearance of $X_t(\cdot)$ at position t and $Y_{t+k}(\cdot)$ at position $t + k$

6.1.3 S_ρ

Shannon's relative entropy and almost all other entropies fail to be "metrics", as they violate either symmetry, or the triangularity rule, or both. This means that they are measures of divergence, not distance. A metric measure would have the additional advantage of allowing multiple comparisons of departures/distances. It is also desirable to provide the framework for assessing statistical significance of any proposed measure. [49], [24]

Among the various indices based on the concept of entropy cited in the literature, it is worth mentioning that proposed by Granger, Maasoumi and Racine [25], more often referred to as S_ρ and defined as:

$$S_\rho(k) = \frac{1}{2} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \left(f^{\frac{1}{2}} - f_1^{\frac{1}{2}} f_2^{\frac{1}{2}} \right)^2 dxdy$$

where f is the joint density distribution of the random variables X and Y, while f_1 and f_2 are their respective distributions of marginal densities. If X and Y are independent, $S_\rho = 0$ otherwise, S_ρ is positive and greater than zero. Granger, Maasoumi Racine and developed the S_ρ based on Hellinger-Battacharya-Matusita distance measure [25] [41]. In particular, the Battacharya coefficient on the generic distributions f_a and f_b is defined as in [5] [1]:

$$\rho^*(f_a, f_b) = \int_{-\infty}^{+\infty} (f_a f_b)^{\frac{1}{2}} dx$$

and is a measure of divergence between distributions. This coefficient does not have a metrical structure because it does not meet all three axioms. On the contrary, its modified version is a metric distance, known as the Battacharya distance:

$$B(f_a, f_b) = \sqrt{1 - \rho^*(f_a, f_b)}$$

Matusita measure, has the following form [47] [26]:

$$M(f - a, f_b) = \int_{-\infty}^{+\infty} (f_a^{\frac{1}{2}} - f_b^{\frac{1}{2}})^2 dx$$

Between these two indicators there is the following relation:

$$M(f_a, f_b) = 2B(f_a, f_b)^2$$

Both $M(f_a, f_b)$ and $B(f_a, f_b)$, a rare case between the measures of divergence, have the peculiar characteristic to meet, among other things, the triangular inequality and therefore can be considered as metric measures

Moreover, if we replace f_a with the joint density distribution of the random variables X and Y ($f = f(X, Y)$) and f_b with the product of the respective marginal density distributions ($f_1 f_2$) the we will obtain S_ρ:

$$S_\rho = 1 - \rho^*(f_a, f_b) = B(f_a, f_b)^2$$

S_ρ is therefore a metric too and has a close relation with the Havrda and Charvat entropies family of k-order:

$$I_k(f_1, f_2) = \frac{1}{k-1}\left[\int (f_1^k / f_2^k) dF_2 - 1\right], \qquad k \neq 1$$

S_ρ is an indicator that can measure the degree of deviation from the condition of independence and is robust against possible (but unknown) nonlinear and non-Gaussian processes. In fact, the entropy of Granger, Maasoumi and Racine normalized to its maximum formally meets the following six properties:

1. It is well defined for both continuous and discrete variables.

2. It is normalized to zero if X and Y are independent, and lies between 0 and $+1$.

3. The modulus of the measure is equal to unity (or a maximum) if there is a measurable exact (nonlinear) relationship, $Y = m(X)$ say, between the random variables.

4. It is equal to or has a simple relationship with the (linear) correlation coefficient in the case of a bivariate normal distribution.

5. It is *metric*, i.e., it is a true measure of ”distance” and not just of divergence.

6. The measure is invariant under continuous and strictly increasing transformations $\psi(\cdot)$. This is useful since X and Y are independent if and only if $\psi(X)$ and $\psi(Y)$ are independent. Invariance is important since otherwise clever or inadvertent transformations would produce different levels of dependence.

For a detailed discussion on the definition, implementation and estimation issues of S_ρ, see [25].

In the present work, the statistical analysis of binary sequences obtained by means of our coding framework is based on the implementation of a bivariate version of the metric entropy measure S_ρ. This version corresponds to a two-dimensional implementation of the methods employed in [19], as follows:

$$S_\rho(k) = \frac{1}{2}\int\int\left[\sqrt{f_{(X_t,Y_{t+k})}(x,y)} - \sqrt{f_{X_t}(x)f_{Y_{t+k}}(y)}\right]^2 dxdy$$

where

- $X_t(\cdot)$ is a random process that measures which nucleotide appears at position t

- (X_t, Y_{t+k}) is the bivariate random process that measures the joint appearance of $X_t(\cdot)$ at position t and $Y_{t+k}(\cdot)$ at position $t+k$

The measure has been proven to have impressive and robust power for characterizing nonlinear processes. In particular, it has been shown that tests based upon S_ρ have very good performances in terms of power and size [15]. In the binary case the double integral reduces to summation and probabilities are estimated through relative frequencies:

$$S_\rho(k) = \frac{1}{2}\sum_{i=0}^{1}\sum_{j=0}^{1}\left[\sqrt{Pr(X_t = i, Y_{t+k} = j)} - \sqrt{Pr(X_t = i)Pr(Y_{t+k} = j)}\right]^2$$

Tabella 6.1: Legend

class	frame	anticodon

p = parity
r = Rumer
h = hidden
	0 = in frame
	1 = out of frame 1
	2 = out of frame 2
		a = reversed complement

6.2 Sequence analysis

In previous works [19, 20] the authors analyzed both the univariate dependence structure of parity sequences generated from protein coding DNA regions, and bivariate dependence structure comparing pairwise chemical or dichotomic codon classes.

In this work, I would like to assess the short-range dependence stuctures between dichotomic classes computed on nucleotide sequences of different portions of the genome. In order to do that, I have analyzed the following set of sequences:

Group	# of sequences
Exons	500
CDS	100
Introns	500
Long introns	100
Intergens	250
UTR	100

For each sequence, I computed all the possible nontrivial combinations of dichotomic classes (an overall set of 153 different cases): Then I computed the dependence measure between the two binary strings relative to each combination at lag -1, 0 and +1.

Given two binary sequences X_t and Y_t the null hypothesis tested is:

$$\begin{cases} H_0 : X_t \text{ and } Y_{t+k} \text{ are independent} \\ H_1 : X_t \text{ and } Y_{t+k} \text{ are not independent} \end{cases} \qquad \text{for } k \in \mathbb{Z}$$

and

$$\begin{cases} p_{X_t}(x) \text{ is the relative frequency of } X_t; \\ p_{Y_{t+k}}(y) \text{ is the relative frequency} Y_{t+k}; \\ p_{(X_t,Y_{t+k})}(x,y) \text{ is the joint relative frequency } (X_t, Y_{t+k}). \end{cases}$$

In order to build a valid test, I need a suitable measure of dependence and a scheme for testing H_0.

Measures I use the three dependence measures (D) described in the previuos section:

1. Chi-squared test
2. S_ρ (a normalized variant of the Bhattacharya-Hellinger-Matusita distance)
3. Mutual information

We must consider that:

- the null hypothesis we test is that of independence between binary sequences, that is, the absence of an informational organization between codons
- the dichotomic classes are naturally correlated because they can be computed on the same bases
- spurious correlations due to nonstationarity/different GC content
- when comparing dichotomic classes the test does not have to depend on correlations induced by their definition; in fact, some specific combinations of dichotomic classes and reading frames induce nonzero spurious correlations even in random sequences.

Because of such issues simple nonparametric bootstrap schemes that resample the binary sequences are not appropriate.

Testing scheme The above requirements can be satisfied by resorting to suitable nonparametric bootstrap or permutation schemes proposed in [19, 20]. The original DNA base sequence is randomly permuted. On this new sequence, the chemical (or dichotomic) classes are computed and the dependence measure D is estimated. The procedure is repeated B times (say B=5000) as to obtain the bootstrap distribution of S_ρ under the null hypothesis. Clearly, each permutation of the original data preserves the original proportion of bases and this fulfils requirement.

Given a nucleotide sequence Z_t

1. compute the two dichotomic classes X_t and Y_t on Z_t
2. compute the measure on X_t and Y_{t+k}: $\hat{D}_k$
3. draw Z_t^*, a random permutation of Z_t
4. compute the two dichotomic classes X_t^* and Y_t^* on Z_t^*
5. compute the measure on X_t^* and Y_{t+k}^*: $\hat{D}_k^*$
6. repeat steps 3 – 5 B times.
7. compare $\hat{D}_k$ with the quantiles of the distribution of $\hat{D}_k^*$.

I applied the testing scheme to all the three dependence measures considered.

6.2.1 Results

CDS

Looking at Table 6.2 we can see that:

- There is a strong dependence between dichotomic classes h0a and r1a at lag 0: in fact it is present in 60% of the sequences analyzed.
- Strong dependence seems to involve also:
  - h1a and r2 at lag 0
  - h0 and r2 at lag 0
  - h2 and r1a at lag1

  All of these couples of dichotomic classes are dependent in at least the 40% of the sequences I analyzed. Table 6.2.1 display the differences between CDS and the other portions of the genome: we can observe a clear prevalence of dependence structures in sequences CDS.
- As far as S_ρ and mutual information (MI) concern, there are 10 couple of dichotomic classes that show a dependence structure at least in 30 % of sequences. These structures involves only Rumer and hidden classes.
- S_ρ and MI shows a very similar pattern. They seem to recognize the same dependence structures with the same level of accuracy.
- $\chi^2 - test$ recognize the same kind of dependence structure as S_ρ and MI, but it shows a lower sensibility

Exons

We can see the same dependence structures observed for CDS sequences but, in this case, they are present in a lower percentage the sequences considered (see Table 6.3). For example the couple h0a-r1a is dependent at lag 0 in only 36.5% of the exon sequences while it was in 62.5% of CDS (as far as S_ρ index is concerned). This could be another clue supporting the hypothesis that hidden class represent a mathematical structure meaningful to recognize different portions of coding sequences

None of the dichtomic class combination show a frequency of dependence higher than 40% and only the couple h0a-r1a exceeds 30%. S_ρ and MI shows the similar frequencies in each dichotomic class combination, while χ^2 seems to be less sensitive.

Tabella 6.2: Dependence measures computed on CDS sequences. Combinations that show dependence in the highest percentage of sequences.

Combination	Lag	Freq S_ρ	Freq MI	Freq χ^2	Bases
h0a-r1a	0	62.5	62.0	50.5	$\overline{34} - \overline{45}$
h1a-r2	0	48.5	48.5	47.5	$\overline{56} - 34$
h0-r2	0	47.0	49.0	46.5	$34 - 67$
h2-r1a	+1	46.0	46.0	41.5	$23 - \overline{78}$
r1-r1a	0	38.0	38.5	10.0	$23 - \overline{45}$
h1-r2	0	37.5	36.0	38.0	$23 - 34$
h0a-r2a	0	36.0	36.5	28.5	$\overline{34} - \overline{34}$
r0-r0a	-1	35.0	35.5	4.5	$45 - \overline{23}$
h0-r1	0	34.5	34.0	28.5	$34 - 45$
r2-r2a	0	32.0	31.5	24.5	$34 - \overline{34}$

Tabella 6.3: Dependence measures computed on Exon sequences. Combinations that show dependence in the highest percentage of sequences.

Combination	Lag	Freq S_ρ	Freq MI	Freq χ^2	Bases
h0a-r1a	0	36.5	36.5	25.5	$\overline{34} - \overline{45}$
h0-r2	0	26.0	23.5	20.5	$34 - 67$
h1a-r2	0	23.5	23.0	17.5	$\overline{56} - 34$
h2-r1a	+1	22.5	22.5	12.0	$23 - \overline{78}$
h1-r1a	0	22.0	23.0	12.5	$12 - \overline{45}$
h0-r1a	0	19.0	18.0	15.5	$34 - \overline{78}$
h0-h1	0	19.0	19.0	15.5	$34 - 45$
r1-r2a	0	18.0	18.5	5.0	$23 - \overline{34}$
p0-p1	0	17.0	18.5	9.0	$23 - 34$
h1-r2	0	16.5	16.5	18.0	$23 - 34$

Introns

The analysis conducted on intron sequences indicates that no dichotomic class combination shows a dependence in a number of sequences higher than 20% (see Table 6.4). This indicates that there are no short-range dependence structure as far as introns are concerned.

Once again S_ρ and MI measures shows similar outcomes, while χ^2 seems to be less sensitive.

Tabella 6.4: Dependence measures computed on intron sequences. Combinations that show dependence in the highest percentage of sequences.

Combination	Lag	Freq S_ρ	Freq MI	Freq χ^2	Bases
r2-r2a	0	18.5	17.5	4.0	$12 - \overline{45}$
h0a-r2a	0	17.5	17.0	12.0	$\overline{34} - \overline{34}$
r1- r2a	0	16.0	15.5	4.0	$23 - \overline{34}$
h0a-h1a	0	15.0	14.0	14.5	$\overline{34} - \overline{56}$
h0-r2	0	12.5	12.5	5.0	$34 - 67$
r0a-r2a	0	12.0	12.5	3.0	$\overline{23} - \overline{34}$
r0-r0a	-1	11.5	11.5	2.5	$45 - \overline{23}$
h1a-r2	0	11.5	11.5	6.5	$\overline{56} - 34$
h0a-r1a	0	11.5	11.5	13.5	$\overline{34} - \overline{45}$
h0-r1	0	11.5	11.5	2.5	$34 - \overline{45}$

Long introns

Table 6.5 shows the dichotomic class combinations with the highest dependence frequency. We can see that no combination exceed a 30% frequency. Dependence is more frequent in the couples r2-r2a and r1-r2a at lag 0 (around 28% when computed with S_rho and MI indexes). It could be argued that Rumer class is relevant in order to recognize different introns within the same gene.

It is interesting to notice that if we use χ^2 as measure of dependence, no combination shows a dependece frequency higher than 20%.

Tabella 6.5: Dependence measures computed on long intron sequences. Combinations that show dependence in the highest percentage of sequences.

Combination	Lag	Freq S_ρ	Freq MI	Freq χ^2	Bases
r2-r2a	0	28.5	28.5	7.5	$12 - \overline{45}$
r1- r2a	0	26.5	28.0	5.5	$23 - \overline{34}$
h0a-r2a	0	24.0	24.5	13.5	$\overline{34} - \overline{34}$
r1-r1a	0	22.5	21.0	13.0	$23 - \overline{45}$
h0-r1	0	22.0	22.0	11.0	$34 - 45$
h1a-r2	0	22.0	22.0	6.5	$\overline{56} - 34$
h0a-h1a	0	21.0	20.5	18.5	$\overline{34} - \overline{56}$
r0-r1	-1	19.0	19.0	6.5	$45 - 23$
p0a-r1	-1	18.5	18.0	0	$\overline{45} - 23$
h1-r2	0	18.0	18.5	16.0	$23 - 34$

UTRs

Dependence analysis on UTR sequences shows interesting results. In fact we can see that only two combinations are dependent in more than 20% of the sequences: h0a-r1a and h0-h1, both at lag 0 (see Table 6.6) While the first combination is typical of coding sequences (it is indeed the more frequent both in CDS and exons), the second one is specific of UTRs, since we have not met it previously. Notice that both these two combinations involve bases 3 and 4 on the first binary sting and bases 4 and 5 on the second binary string. It is also interesting to underline that χ^2 detects $h0 - h1$ dependence structure in 26% of the sequences, while it does not detect h0a-r1a with a rate higher than 20%.

Tabella 6.6: Dependence measures computed on UTR sequences. Combinations that show dependence in the highest percentage of sequences.

Combination	Lag	Freq S_ρ	Freq MI	Freq χ^2	Bases
h0-h1	0	29.0	29.0	26.0	$34 - 45$
h0a-r1a	0	28.5	30.0	17.0	$\overline{34} - \overline{45}$
h0-r2	0	19.0	20.0	6.0	$34 - 67$
h1-r1a	0	18.5	19.0	7.5	$12 - \overline{45}$
p1-p2	0	17.5	16.5	10.5	$34 - 445$
r0-r0a	-1	16.5	16.0	0.5	$45 - \overline{23}$
h1a-r1a	0	15.5	16.0	7.0	$\overline{23} - \overline{12}$
h0a-r2a	0	15.0	14.0	5.5	$\overline{34} - \overline{34}$
p0-p2	0	13.5	14.0	1.5	$23 - 45$
h0-r2a	0	13.0	14.0	2.05	$34 - \overline{67}$

Intergenes

Finally we can see that dependence analysis conducted on intergenes shows similar results as UTRs. In fact all the dependence measures used detect a dependence structure in more than 20% of sequences only for the combination h0-h1 at lag 0 (see Table 6.7). This could mean that there is a short-range dependence involving hidden class that is relevant in order to recognize portions of DNA that are not genes or that do not undergo to translation process.

It is interesting to underline that the difference between UTRs and intergenes is in the combination h0a-r1a, detected as relevant in UTRs but not in intergenes. This short-range dependence could discriminate between non coding region present within (UTRs) or outside (intergenes) genes, remembering that it is present with the highest frequency both in CDS and exons.

Tabella 6.7: Dependence measures computed on intergenes sequences. Combinations that show dependence in the highest percentage of sequences.

Combination	Lag	Freq S_ρ	Freq MI	Freq χ^2	Bases
h0-h1	0	26.0	28.0	23.0	$34 - 45$
h0-r2a	0	18.5	17.5	3.0	$34 - \overline{67}$
h0-h1a	0	18.0	18.0	15.5	$34 - \overline{89}$
h1-h2a	-1	18.0	17.5	3.5	$45 - \overline{45}$
h0a-h2a	-1	17.0	15.5	14.5	$\overline{67} - \overline{45}$
h1a-h2a	+1	16.5	17.0	10.5	$\overline{23} - \overline{45}$
h1a-h2a	-1	16.5	15.0	13.0	$\overline{56} - \overline{12}$
h0a-h2	+1	15.5	15.0	0.5	$\overline{34} - 56$
h0a-r2a	0	15.5	15.5	4.0	$\overline{34} - \overline{34}$
h2-r2a	+1	14.5	14.5	0	$23 - \overline{67}$

Tabella 6.8: Comparison between the dependence measures of the four main combinations computed in all the genome portions

		CDS			Exon			Intron		
Combination	Lag	S_ρ	MI	χ^2	S_ρ	MI	χ^2	S_ρ	MI	χ^2
h0a-r1a	0	62.5	62.0	50.5	36.5	36.5	25.5	11.5	10.0	13.5
h1a-r2	0	48.5	48.5	47.5	23.5	23.0	17.5	11.5	13.0	6.5
h0-r2	0	47.0	49.0	46.5	26.0	23.5	20.5	12.5	12.5	5.0
h2-r1a	+1	46.0	46.0	41.5	22.5	22.5	12.0	7.0	8.5	2.5

		Long introns			UTR			Intergenes		
Combination	Lag	S_ρ	MI	χ^2	S_ρ	MI	χ^2	S_ρ	MI	χ^2
h0a-r1a	0	16.0	15.5	18.5	28.5	30.0	17.0	3.0	3.5	1.0
h1a-r2	0	22.0	22.0	11.0	11.5	11.0	8.0	8.0	8.0	4.5
h0-r2	0	16.5	17.0	4.0	19.0	20.0	6.0	2.0	2.5	2.5
h2-r1a	+1	16.0	16.5	6.0	7.5	7.0	2.0	4.0	4.5	1.5

6.2.2 Comparison between indexes

S_ρ vs. Mutual information

Figure 6.1 shows the boxplots of the differences between the frequency of rejections computed with S_ρ and MI. We can see that the median is around 0 for all the portion of the genome analyzed.

Moreover only few dichotomic class combinations display a difference greater than 2.

S_ρ vs. χ^2

Looking at Figure 6.2 it is clear how S_ρ detects dependence structure better than χ^2.

A big group of dichotomic class combinations display differences greater than 10%, in particular when CDS are considered.

It is evident (and interesting) how χ^2 index has difficulty in detecting dependence structures involving Rumer class variables.

Mutual information vs. χ^2

If we compare MI and χ^2 indexes we can observe big differences in detecting dependence structures as shown in Figure 6.3. Mutual information seems to be more sensitive than χ^2. Once again, a big group of dichotomic class combinations display differences greater than 10%, in particular when CDS are considered.

6.2.3 Comments

The analysis shows that there is an important correlation between h0a and r1a at lag 0 in exon and, overall, in CDS.

- involves adjacent bases of the same codon.(12 and 23)
- the percentage of sequences that shows this correlation is remarkably higher in CDS than in exons.

The fact that we can find this short-range dependence only in coding sequences could represent a clue supporting the hypothesis of the importance of dichotomic class in error detection and correction mechanism.

There are at least other three combinations of dichotomic classes that seem to show a relevant dependence in CDS:

- h1a and r2 at lag 0

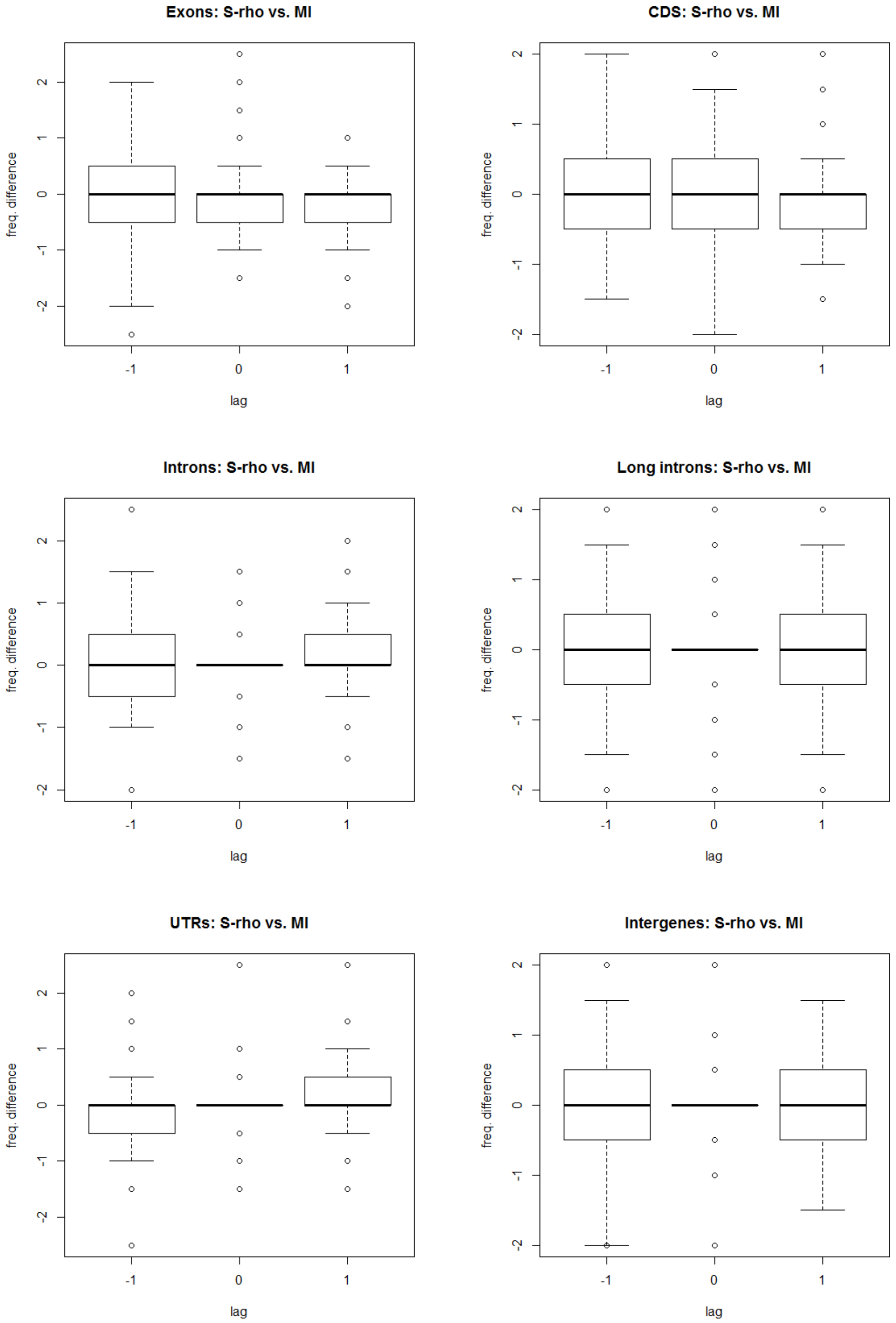


Figura 6.1: Comparison between S_ρ and MI ability of detecting dependence structures.

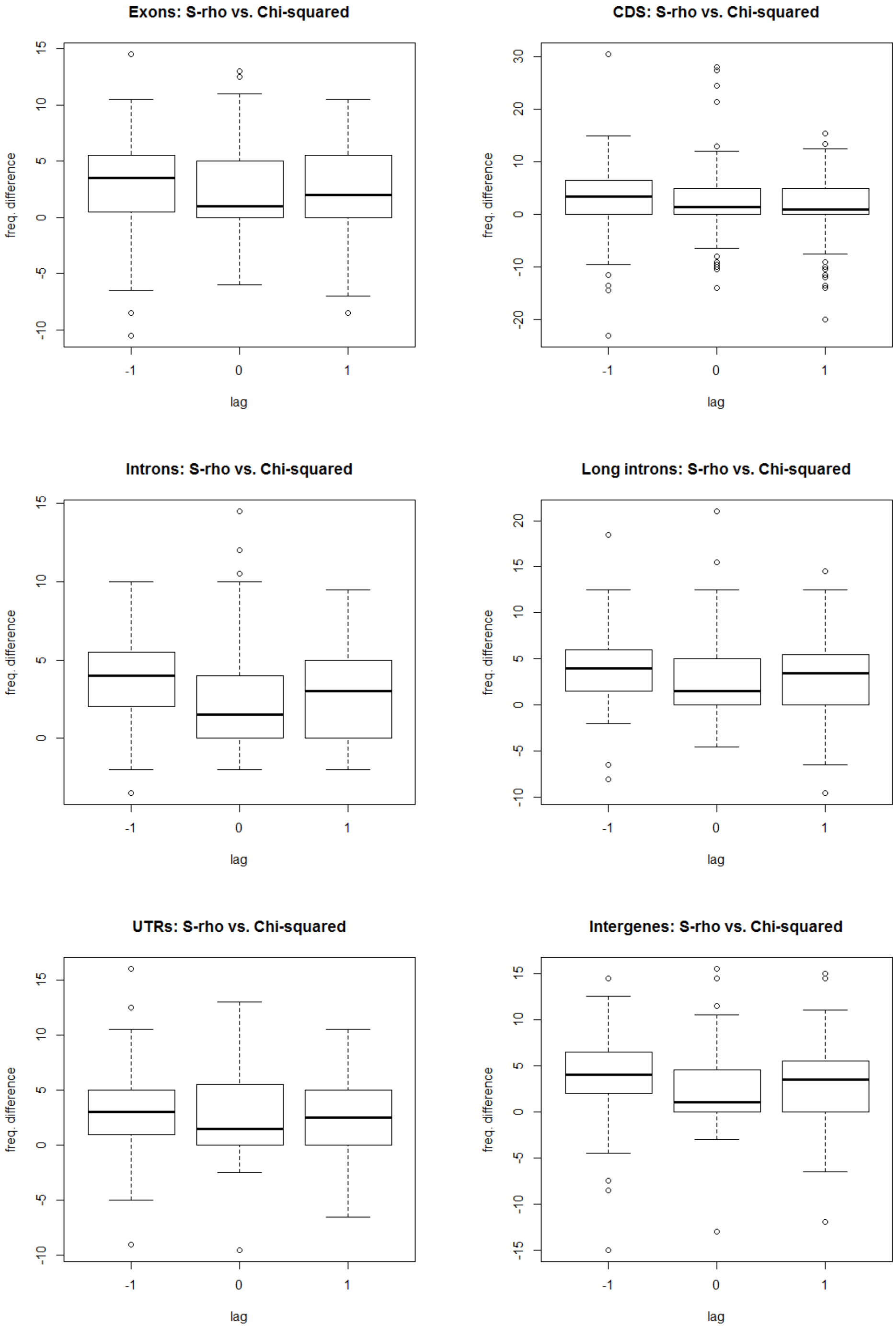


Figura 6.2: Comparison between S_ρ and χ^2 ability of detecting dependence structures.

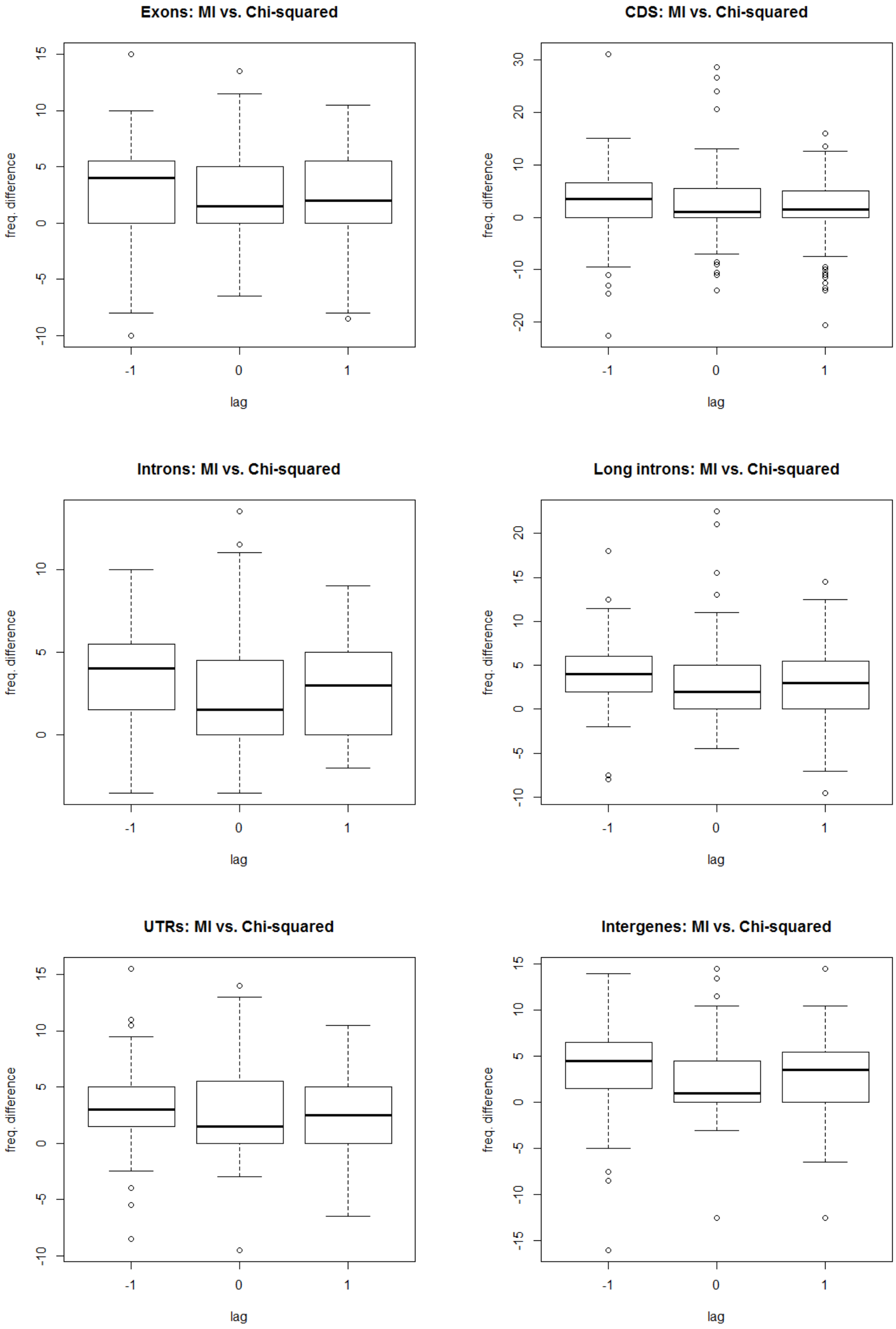


Figura 6.3: Comparison between MI and χ^2 ability of detecting dependence structures.

- h0 and r2 at lag 0
- h2 and r1a at lag 1

All of these involve a Rumer and a hidden binary string as to indicate that it is necessary an interaction between the two dichotomic class in order to detect a dependence structure.

We can also see an important even though less frequent, dependence between h0 and h1 binary strings in UTRs and intergenes. This combination seems to be specific for non coding regions external to the genome portions that undergo to transcription and translation processes.

Finally we can state that S_ρ and MI indexes show a similar behaviour in detecting dependence structure in couples of binary sequences. χ^2 index, instead, seem to be less sensitive: it, in fact, detect the same kind of dependence but in a lower number of sequences.

6.3 Multiple testing problem

In this section I introduce the problem of multiple testing and describe some methods that I could (and should) apply in order to adjust the type I error rate. This is one of the main future developments of this research work.

Notice that Giannerini et al. [14] has applied such adjustment in a similar analysis on different data. Their results show that there are dependence structures similar to those presented here.

In this study I conducted a series of repeated tests on different groups of sequences in order to assess the presence of dependence structures between dichotomic classes binary strings. The use of repeated tests may lead to errors in results interpretation. In fact, once the type I error rate is set at $\alpha = 5\%$ for each test, the chance of erroneously finding a statistically significant impact is 5%. However, when the "family" of hypothesis tests are considered together, the "combined" type I error rate could be considerably larger than 5%.

For example, suppose that the null hypothesis is true for each test and that tests are independent. Then, the chance of finding at least one spurious impact is $1-(1-\alpha)^N$, where N is the number of tests. Thus, the probability of making at least one type I error is 23 % if 5 tests are conducted, 64% for 20 tests, and so on as shown in Table 6.3.

The two most common definitions of the combined type I error rate found in literature are the family-wise error rate (FWER) and the false discovery rate (FDR):

Tabella 6.9: Chances of findings spurious impacts for independent tests

Number of independent tests with true null hypotheses	Probabilty that at least one test is statistically significant
5	0.23
10	0.40
20	0.64
50	0.92

- The FWER, defined by Tukey [53] is the probability that at least one null hypothesis will be rejected when all null hypotheses are true. As discussed, the FWER is $1 - (1 - \alpha)^N$, for independent tests, where N is the number of tests.
- The FDR, defined by Benjamini and Hochberg [3], is a more recent approach for assessing how errors in multiple testing could be considered. The FDR is the expected proportion of all rejected null hypotheses that are rejected erroneously.

Table 6.10 helps clarify these two error rates. Suppose that multiple tests are conducted to assess intervention effects on N study outcomes and that M null hypotheses are true (M is unobservable). Suppose further that based on t-tests, Q null hypotheses are rejected and that A, B, C, and D signify cell counts when t-test results are compared to the truth. The counts Q and A to D are random variables.

Tabella 6.10: The number of errors when testing multiple hypotheses

	Results from hypothesis tests (Observed)		
Truth(Unobserved)	H_0 is not rejected	H_0 is rejected	Total
H_0 is true	A	B	M
H_0 is false	C	D	(N-M)
Total	(N-Q)	Q	N

In Table 6.10, the FWER is the probability that the random variable B is at least 1 among the M null hypotheses that are true. The FDR equals the expected value of B /Q , where B /Q is defined to equal 0 if Q = 0.3

If all null hypotheses are true, then B = Q and the FDR and FWER are equivalent, otherwise the FDR is smaller than or equal to the FWER.

The two error rates have a different philosophical basis. The FWER measures the likelihood of a single erroneous rejection of the null hypothesis across the family of tests. FWER is concerned with mistakenly reporting any statistically significant findings. Unwarranted scientific conclusions about evaluation findings could be made as a result of even one mistake and that researchers may select erroneous significant findings for emphasis when reporting and publishing results.

The rationale behind the FDR is that a few erroneous rejections may not be as problematic for drawing conclusions about the family tested when many null hypotheses are rejected as they would be if only a few null hypotheses are rejected. The rejection of many null hypotheses is a signal that there are real differences across the contrasted groups.

FDR is a less conservative measure than the FWER, especially if a considerable fraction of all null hypotheses are false. Thus, as reported below, methods that control the FDR could yield tests with greater statistical power than those that control the FWER. The choice of which error criterion to control is important and must be made prior to the data analysis.

6.3.1 Statistical Solutions to the Multiple Testing Problem

A large body of literature describes statistical methods to adjust type I errors for multiple testing ([?, 55, 56, 33]). The literature suggests that there is not one method that is preferred in all instances. Rather, the appropriate measure will depend on the study design, the primary research questions that are to be addressed, and the strength of inferences that are required.

Methods for FWER Control

Until recently, most of the literature on multiple testing focused on methods to control the FWER at a given α level (that is, methods to ensure that the FWER $\leq \alpha$). The most well-known method is the Bonferroni procedure, which sets the significance level for individual tests at α/N where N is the number of tests.

The **Bonferroni procedure** controls the FWER when all null hypotheses are true or when some are true and some are false (that is, it provides *strong* control of the FWER). The Bonferroni method applies to both continuous and discrete data, controls the FWER when the tests are correlated, and provides adjusted confidence bounds (by using α/N rather than α in the calculations). Furthermore, it is flexible because it controls the FWER for tests of joint hypotheses about any subset of N separate hypotheses (including individual contrasts). The procedure will reject a joint hypothesis H_0 if any p-value for the individual hypotheses included in H_0 is less than α/N. The Bonferroni method, however, yields conservative bounds on type I error and, hence, has low power.

Many modified and sometimes more powerful versions of the Bonferroni method have been developed that provide strong control of the FWER. Here some examples are provided:

- **Sidák** (1967) [48] developed a slightly less conservative bound where the significance level for individual tests is set at $1-(1-\alpha)^{1/N}$ rather than α/N. This method has properties similar to those of the Bonferroni method and is slightly more powerful, although it does not control the FWER in all situations in which test statistics are dependent.

- **Holm** (1979) [31] developed a sequential *step-down method*: (1) order the p-values from the individual tests from smallest to largest,$p_1 \leq p_2 ... \leq p_N$, and order the corresponding null hypotheses $H_{0(1)}, H_{0(2)}, ..., H_{0(N)}$; (2) define k as the minimum j such that $p_j > \alpha/(N-j+1)$; and (3) reject all $H_{0(j)}$ for $j = 1, ..., (k-1)$. This procedure is more powerful than the Bonferroni method because the bound for this method sequentially increases whereas the Bonferroni bound remains fixed. The Holm method controls the FWER in the strong sense, but cannot be used to obtain confidence intervals.

- **Hochberg** (1988) [29] developed a *step-up procedure* that involves sequential testing where p-values are ordered from largest to smallest (rather than vice versa as for the Holm test). The method first defines k as the maximum j such that $p_j \leq \alpha/(N-j+1)$, and then rejects all $H_{0(j)}$ for $j = 1, ..., k$. This procedure is more powerful than the Holm method, but the control of the FWER is not guaranteed for all situations in which the test statistics are dependent (although simulation studies have shown that it is conservative under many dependency structures).

- Bootstrap and permutation resampling methods are alternative, computer-intensive methods that provide strong control of the FWER ([56]). These methods incorporate distributional and correlational structures across tests, so they tend to be less conservative than the other general-purpose methods and, hence, may have more power. Furthermore, they are applicable in many testing situations.

Methods for FDR Control

Benjamini and Hochberg ([3]) showed that when conducting N tests, the following four-step procedure will control the FDR at the α level:

1. Conduct Nseparate t-tests, each at the common significance level α.
2. Order the p-values of the N tests from smallest to largest, where $p_{1*} \leq p_2 \leq ... \leq p_N$ are the ordered p-values.
3. Define k as the maximum j for which $p_j \leq \dfrac{j}{N}\alpha$
4. Reject all null hypotheses $H_{0(j)}j = 1, 2, ..., k$.. If no such $k$exists, then no hypotheses are rejected.

This step-up sequential procedure, which has become increasingly popular in the literature, is easy to use because it is based solely on p-values from the individual tests. Benjamini and Hochberg [3] first proved that this procedure (BH procedure) controls the FDR for continuous test statistics and Benjamini and then proved that this procedure also controls the FDR for discrete test statistics [4].

The original result was proved assuming independent tests corresponding to the true null hypotheses (although independence was not required for test statistics corresponding to the false null hypotheses). More research is needed to assess whether the BH procedure is robust when independence and positive regression dependency are violated.

6.3.2 Related problems

There are two related concerns with the adjustment procedures discussed above:

1. they result in tests with reduced statistical power
2. they could result in tests with even less power when the test statistics are correlated (dependent).

Loss in Statistical Power

The statistical procedures that control for multiplicity reduce type I error rates for individual tests. Consequently, these adjustment procedures result in tests with reduced statistical power. the probability of rejecting the null hypothesis given that the null hypothesis is false. Stated differently, these adjustment methods reduce the likelihood that the tests will identify true differences between the contrasted groups. The more conservative the multiple testing strategy, the greater the power loss.

Multiplicity adjustments involve a trade-off between type I and type II error rates. Conservative testing strategies, such as the Bonferroni and similar methods, can result in considerable losses in the statistical power of the tests, even if only a small number of tests are performed. The less conservative BH test has noticeably more power if a high percentage of all null hypotheses are false.

Dependent Test Statistics

Individual test statistics are likely to be related in many situations.

Some of the adjustment methods discussed above (such as the Bonferroni and Holm methods) control the FWER at a given α level when tests are correlated. However, for some forms of dependency, these methods may adjust significance levels for individual tests by more than is necessary to control the FWER. This could lead to further reductions in the statistical power of the tests. For example, if test correlations are positive and large, each test statistic is providing similar information about intervention effects, and thus, would likely produce similar p-values. Consequently, in these situations, fewer adjustments to type I error rates are needed to control the FWER.

Finally, the BH method controls the FDR under certain forms of dependency and for certain test statistics, but not for others.

Capitolo 7

Conclusions

The objective of this thesis was to characterize the genome of the entire chromosome 1 of *A.thaliana*, a small flowering plants used as a model organism in studies of biology and genetics, on the basis of a recent mathematical model of the genetic code.

First I have reviewed the main mathematical features of the model and its symmetry properties. Then, I have analyzed the whole chromosome 1 of *A.thaliana* by creating specific routines that, by using genome annotations, extract and build seven groups of sequences: genes, exons, introns, coding sequences (CDS), intergenes, untranslated regions (UTR) and regulatory sequences. Then I created fictitious sequences called long introns in order to have a type of sequence to be compared with CDS.

The nucleotide sequences were then transformed into binary sequences based on the definition of the three different dichotomic classes: Rumer, hidden and parity. So I generated 18 random variables corresponding to the proportion of 1 in each one of the binary string computed in and out of frame 1 and 2 on both the sense and the antisense strand . These 18 variables, along with the 4 variables relating to the proportions of A, C, G and T of each sequence, were used to conduct three different types of analysis.

First, it is carried out a descriptive analysis with the aim to characterize the different portions of the genome on the basis of 22 variables described above. Results indicate the presence of regularities in each portion of the genome considered. In particular intergenes, DNA sequences present between two successive genes that have no apparent biological or regulating function, show an impressive regularity given by:

$$p \simeq 60\% \qquad\qquad r \simeq 31\% \qquad\qquad h \simeq 49\%$$

in all binary sequences analyzed (computed on both the sense and the antisense strand, and undergone to all the transformations).

Moreover we can observe the presence of remarkable differences between coding sequences (CDS and exons) and non-coding sequences. Since mean values of dichotomic classes variables vary with frame only for coding sequences, frame seems to be important only for them. This effect is higher in CDS than exons, suggesting a sort of "*union effect*" that occurs when fragments of coding sequences (exons) join together to form a CDS.

For what concerns the specific role of each dichotomic class, we could assume that parity class is important in order to distinguish between coding sequences in sense and antisense strand while Rumer and hidden classes seem to be useful in order to discriminate different kinds of non-coding sequences. In particular introns seem to minimize hidden class mean values.

In a second moment I wanted to verify if the dichotomic classes were able to discriminate between the different portions of the genome. So I created a set of logistic regression models in order to discriminate between pairs of portions of the genome using, as regressors, different combinations of the variables previously created. The results obtained show that dichotomic classes seem to be useful in order to discriminate between coding and non-coding sequences, as to underline the importance of frame. Once again there seems to be a sort of "union effect" that enhance, when we consider CDS, the features observed in exons. However I realized the use of dichotomic classes does not improve, almost always, the ability of discriminating different portions of the genome respect to what is obtained using only proportion of bases.

Finally, I wanted to check the existence of short-range dependence between binary sequences computed on the basis of the different dichotomic classes. I used three different measures of dependence: the well-known $\chi^2 - test$ and two indices derived from the concept of entropy i.e. Mutual Information (MI) and S_ρ, a normalized form of "Bhattacharya Hellinger Matusita distance". The results obtained show that there is a significant short-range dependence structure only for the coding sequences. This dependence involves in particular CDS, highlighting once again the existence of a sort of "*union effect*" when the exons join together to form a CDS. The existence of such a dependence structure is a clue of an underlying error detection nd correction mechanism, whose biological bases are partly known (for example the proofreading process by DNA polymerase).

I have also compared the results obtained using the different indices of dependence. It is clear that both S_ρ and mutual information are better than

χ^2 in detecting dependence structures between binary sequences. All three indices, in fact, reveal a dependence between the same combinations of dichotomic classes. However, MI and S_ρ do it in a significantly higher percentage of sequences analyzed with respect to what χ^2 do. However, it is not possible to define which is the best between MI and S_ρ as they seem to have the same ability to detect dependence structures .

In conclusion of my thesis work, I can say that the mathematical model here described is a useful tool for interpreting the different portions of the genetic code of *A.thaliana*. Dichotomic classes seem in fact useful in discriminating the coding portions of the genome (sequences which are transcribed and translated) from the non-coding ones. They also seem to characterize the different non-coding portions of the genome. The presence of a short-range dependence structure and the enhanced effect that is observed in CDS compared to exons suggest that this mathematical structure of the genetic code can be the basis of both the recognition of the different portions of coding sequences present within a same gene and the error detection/correction mechanism.

No doubt, further studies are needed in order to assess how the information carried by dichotomic classes could discriminate between coding and noncoding sequence and, therefore, contribute to unveil the role of the mathematical structure in error detection and correction mechanisms. Still, I have shown the potential of the approach presented for the understanding the management of genetic information.

It will be interesting to perform the same type of analysis on the genome of other organisms considered as models for studies of molecular biology and genetics such as the nematode *Caenorabditis elegans*, the fruitfly *Drosphila melanogaster*, the bacterium *Escherichia coli* and the mouse *Mus musculus*.

The main long term objective of this research is the understanding, in informational terms, of the coding/decoding strategies that govern the accuracy of genetic processing, with particular emphasis on the error detection and correction mechanisms and its possible implementation in terms of dynamical molecular machines. I believe that this approach could help to keep the promises and hopes related to molecular biology and the Human Genome Project.

Appendice A

Dichotomic class tables

A.1 Normal sequences

A.1.1 Median values

	$p0$	$r0$	$h0$	$p1$	$r1$	$h1$	$p2$	$r2$	$h2$
Genes	56.05	38.04	47.66	56.25	37.97	47.76	56.88	37.79	46.94
CDS	50.25	43.88	51.85	51.57	41.04	54.84	58.78	38.23	45.94
Exons	52.00	41.94	50.82	52.63	39.62	52.00	56.52	38.98	47.54
Introns	61.11	34.48	38.89	60.50	32.99	38.46	59.38	32.56	37.93
Long Int	60.44	33.42	41.07	60.29	33.16	40.94	60.00	33.13	40.78
IG	60.85	30.59	49.42	60.82	30.56	49.40	60.84	30.54	49.35
UTR	60.00	34.58	43.61	60.00	34.55	43.57	60.00	34.55	43.66
Reg	60.00	34.55	45.00	60.00	34.52	44.65	60.00	34.97	45.10

Tabella A.1: Median values of dichotomic class percentages computed in sense strand for each sequence class

	$p0a$	$r0a$	$h0a$	$p1a$	$r1a$	$h1a$	$p2a$	$r2a$	$h2a$
Genes	56.80	37.55	51.36	56.88	38.15	51.32	56.52	37.85	51.15
CDS	54.62	38.24	48.87	56.90	49.33	49.35	51.30	43.10	44.08
Exons	54.84	40.00	48.94	56.00	45.78	49.12	53.23	42.58	46.67
Introns	60.61	29.17	60.82	61.54	26.76	61.54	61.71	28.57	61.29
Long Int	60.45	29.12	58.91	60.69	28.62	59.05	60.71	29.17	59.05
IG	60.86	30.67	49.40	60.78	30.75	49.41	60.85	30.65	49.44
UTR	60.00	31.58	53.73	60.00	31.65	53.95	59.81	31.58	53.95
Reg	60.00	32.69	50.46	60.00	33.33	50.23	59.00	32.35	50.00

Tabella A.2: Median values of dichotomic class percentages computed in antisense strand for each sequence class

A.1.2 Mean values

	$p0$	$r0$	$h0$	$p1$	$r1$	$h1$	$p2$	$r2$	$h2$
Genes	55.19	38.87	48.00	55.86	38.66	48.18	56.84	38.56	47.17
CDS	50.13	44.09	51.88	51.63	41.74	54.78	58.67	38.81	45.90
Exons	52.13	42.17	50.58	52.71	39.93	51.44	56.26	39.47	47.41
Introns	61.21	34.85	38.74	60.40	33.06	38.33	59.29	32.82	37.83
Long Int	60.67	33.22	41.04	60.42	32.74	40.87	60.05	32.56	40.76
IG	60.62	31.48	49.19	60.53	31.43	49.07	60.42	31.49	49.13
UTR	60.22	35.27	44.55	60.01	35.24	44.23	60.15	35.25	44.31
Reg	59.67	35.86	44.08	59.72	35.84	43.56	60.19	36.53	43.45

Tabella A.3: Mean values of dichotomic class percentages computed in sense strand for each sequence class

	$p0a$	$r0a$	$h0a$	$p1a$	$r1a$	$h1a$	$p2a$	$r2a$	$h2a$
Genes	56.29	38.60	51.04	56.33	39.33	51.01	56.01	39.09	50.59
CDS	54.61	39.05	48.66	56.33	49.32	49.39	51.06	43.64	44.14
Exons	54.85	40.75	49.09	55.92	45.28	49.27	53.29	42.76	47.06
Introns	60.45	29.38	61.01	61.58	26.75	61.68	61.74	28.64	61.36
Long Int	60.57	29.15	58.66	60.92	28.43	58.85	60.94	29.04	58.78
IG	60.56	31.56	49.12	60.47	31.56	49.09	60.54	31.53	49.14
UTR	60.12	32.16	52.20	59.70	32.40	52.30	59.71	32.19	52.65
Reg	60.39	32.61	49.50	59.74	33.28	48.97	59.18	33.15	49.13

Tabella A.4: Mean values of dichotomic class percentages computed in antisense strand for each sequence class

A.2 Complementary sequences

A.2.1 Median values

	$p0$	$r0$	$h0$	$p1$	$r1$	$h1$	$p2$	$r2$	$h2$
Genes	56.67	40.54	52.16	55.96	40.07	52.10	56.37	41.00	52.88
CDS	59.76	46.42	47.81	48.21	41.08	44.82	52.63	53.85	53.74
Exons	56.99	46.15	48.15	50.54	43.48	47.00	53.57	50.00	51.43
Introns	59.38	30.43	61.11	60.71	30.30	61.54	60.87	30.30	62.07
Long Int	60.00	31.25	58.93	60.36	31.13	59.06	60.38	31.05	59.22
IG	60.79	31.96	50.22	60.80	32.08	50.28	60.85	32.00	50.29
UTR	60.00	33.33	55.19	60.00	33.33	55.22	60.00	33.52	55.10
Reg	60.00	33.33	52.09	60.00	33.64	51.82	60.00	33.45	51.69

Tabella A.5: Median values of dichotomic class percentages computed in sense strand for each complement sequence class

	$p0a$	$r0a$	$h0a$	$p1a$	$r1a$	$h1a$	$p2a$	$r2a$	$h2a$
Genes	56.43	40.69	48.49	56.57	40.51	48.48	57.13	40.99	48.71
CDS	48.77	43.46	50.90	53.31	41.28	50.41	60.64	49.01	55.67
Exons	51.35	43.33	50.00	54.05	42.86	50.00	57.89	46.81	52.63
Introns	60.00	34.78	39.18	60.39	36.00	38.46	61.11	35.71	38.71
Long Int	60.29	35.25	41.09	60.46	35.48	40.95	60.66	35.42	40.95
IG	60.77	31.89	50.25	60.87	31.95	50.22	60.77	31.92	50.19
UTR	60.00	36.64	45.24	60.00	36.36	45.00	60.00	36.54	44.93
Reg	60.00	36.59	46.99	60.00	36.67	46.83	60.00	36.36	46.76

Tabella A.6: Median values of dichotomic class percentages computed in antisense strand for each complement sequence class

A.2.2 Mean values

	$p0$	$r0$	$h0$	$p1$	$r1$	$h1$	$p2$	$r2$	$h2$
Genes	56.43	41.27	51.78	55.33	40.97	51.61	56.07	42.43	52.62
CDS	59.40	46.45	47.72	48.28	41.51	44.82	52.40	54.02	53.70
Exons	56.68	46.36	48.08	50.94	44.24	47.10	53.60	50.05	51.13
Introns	59.38	30.37	61.25	60.66	30.20	61.67	60.88	30.25	62.17
Long Int	60.09	30.84	58.96	60.52	30.78	59.13	60.53	30.84	59.24
IG	60.56	32.64	50.00	60.46	32.70	50.03	60.68	32.67	49.97
UTR	59.98	33.95	53.82	60.09	33.99	54.00	60.10	34.11	53.92
Reg	59.69	33.91	51.29	60.01	34.71	50.60	59.81	34.40	50.71

Tabella A.7: Mean values of dichotomic class percentages computed in sense strand for each complement sequence class

	$p0a$	$r0a$	$h0a$	$p1a$	$r1a$	$h1a$	$p2a$	$r2a$	$h2a$
Genes	55.90	41.21	48.74	55.97	40.90	48.77	56.74	41.62	49.19
CDS	48.77	43.97	51.02	52.98	41.58	50.29	60.44	49.19	55.55
Exons	51.72	43.59	50.08	54.12	43.32	49.83	57.55	46.78	52.04
Introns	59.69	34.92	38.99	60.21	36.10	38.32	61.06	35.82	38.64
Long Int	60.19	34.58	41.34	60.52	34.93	41.15	60.83	34.86	41.22
IG	60.64	32.51	50.05	60.61	32.62	49.97	60.44	32.62	49.93
UTR	59.63	37.18	46.23	59.95	36.96	45.98	59.82	36.97	45.62
Reg	59.96	38.05	46.34	59.86	37.64	45.85	60.23	37.15	45.69

Tabella A.8: Mean values of dichotomic class percentages computed in antisense strand for each complement sequence class

A.3 Reverted sequences

A.3.1 Median values

	$p0$	$r0$	$h0$	$p1$	$r1$	$h1$	$p2$	$r2$	$h2$
Genes	56.43	40.69	48.49	56.57	40.51	48.48	57.13	40.99	48.71
CDS	48.77	43.46	50.90	53.31	41.28	50.41	60.64	49.01	55.67
Exons	51.35	43.33	50.00	54.05	42.86	50.00	57.89	46.81	52.63
Introns	60.00	34.78	39.18	60.39	36.00	38.46	61.11	35.71	38.71
Long Int	60.29	35.25	41.09	60.46	35.48	40.95	60.66	35.42	40.95
IG	60.77	31.89	50.25	60.87	31.95	50.22	60.77	31.92	50.19
UTR	60.00	36.64	45.24	60.00	36.36	45.00	60.00	36.54	44.93
Reg	60.00	36.59	46.99	60.00	36.67	46.83	60.00	36.36	46.76

Tabella A.9: Median values of dichotomic class percentages computed in sense strand for each reverted sequence class

	$p0a$	$r0a$	$h0a$	$p1a$	$r1a$	$h1a$	$p2a$	$r2a$	$h2a$
Genes	56.67	40.54	52.16	55.96	40.07	52.10	56.37	41.00	52.88
CDS	59.76	46.42	47.81	48.21	41.08	44.82	52.63	53.85	53.74
Exons	56.99	46.15	48.15	50.54	43.48	47.00	53.57	50.00	51.43
Introns	59.38	30.43	61.11	60.71	30.30	61.54	60.87	30.30	62.07
Long Int	60.00	31.25	58.93	60.36	31.13	59.06	60.38	31.05	59.22
IG	60.79	31.96	50.22	60.80	32.08	50.28	60.85	32.00	50.29
UTR	60.00	33.33	55.19	60.00	33.33	55.22	60.00	33.52	55.10
Reg	60.00	33.33	52.09	60.00	33.64	51.82	60.00	33.45	51.69

Tabella A.10: Median values of dichotomic class percentages computed in antisense strand for each reverted sequence class

A.3.2 Mean values

	$p0$	$r0$	$h0$	$p1$	$r1$	$h1$	$p2$	$r2$	$h2$
Genes	55.90	41.21	48.74	55.97	40.90	48.77	56.74	41.62	49.19
CDS	48.77	43.97	51.02	52.98	41.58	50.29	60.44	49.19	55.55
Exons	51.72	43.59	50.08	54.12	43.32	49.83	57.55	46.78	52.04
Introns	59.69	34.92	38.99	60.21	36.10	38.32	61.06	35.82	38.64
Long Int	60.19	34.58	41.34	60.52	34.93	41.15	60.83	34.86	41.22
IG	60.64	32.51	50.05	60.61	32.62	49.97	60.44	32.62	49.93
UTR	59.63	37.18	46.23	59.95	36.96	45.98	59.82	36.97	45.62
Reg	59.96	38.05	46.34	59.86	37.64	45.85	60.23	37.15	45.69

Tabella A.11: Mean values of dichotomic class percentages computed in sense strand for each reverted sequence class

	$p0a$	$r0a$	$h0a$	$p1a$	$r1a$	$h1a$	$p2a$	$r2a$	$h2a$
Genes	56.43	41.27	51.78	55.33	40.97	51.61	56.07	42.43	52.62
CDS	59.40	46.45	47.72	48.28	41.51	44.82	52.40	54.02	53.70
Exons	56.68	46.36	48.08	50.94	44.24	47.10	53.60	50.05	51.13
Introns	59.38	30.37	61.25	60.66	30.20	61.67	60.88	30.25	62.17
Long Int	60.09	30.84	58.96	60.52	30.78	59.13	60.53	30.84	59.24
IG	60.56	32.64	50.00	60.46	32.70	50.03	60.68	32.67	49.97
UTR	59.98	33.95	53.82	60.09	33.99	54.00	60.10	34.11	53.92
Reg	59.69	33.91	51.29	60.01	34.71	50.60	59.81	34.40	50.71

Tabella A.12: Mean values of dichotomic class percentages computed in antisense strand for each reverted sequence class

A.4 Global transformation 1

Here is median values tables for sequences undergone to keto/amino global transformation :

A.4.1 Median values

	$p0$	$r0$	$h0$	$p1$	$r1$	$h1$	$p2$	$r2$	$h2$
Genes	43.33	61.99	48.63	44.04	62.08	48.59	43.63	62.23	48.94
CDS	40.24	56.12	44.23	51.79	59.10	45.25	47.37	61.77	49.77
Exons	43.01	58.33	45.16	49.46	60.87	45.83	46.43	61.29	47.83
Introns	40.62	65.71	53.12	39.29	67.57	52.78	39.13	67.74	52.94
Long Int	40.00	66.67	52.81	39.64	66.91	52.63	39.62	66.95	52.68
IG	39.21	69.49	49.52	39.20	69.57	49.54	39.15	69.57	49.49
UTR	40.00	65.74	51.16	40.00	65.71	51.35	40.00	65.79	51.35
Reg	40.00	66.15	50.59	40.00	66.28	50.73	40.00	65.71	50.91

Tabella A.13: Median values of dichotomic class percentages computed in sense strand for each gt1 sequence class

	$p0a$	$r0a$	$h0a$	$p1a$	$r1a$	$h1a$	$p2a$	$r2a$	$h2a$
Genes	43.57	62.48	50.26	43.43	61.89	49.65	42.87	62.19	50.24
CDS	51.23	61.76	56.83	46.69	50.84	45.76	39.36	56.90	53.88
Exons	48.65	60.09	53.85	45.95	54.79	47.83	42.11	57.75	52.11
Introns	40.00	71.43	43.75	39.61	73.91	44.00	38.89	72.00	43.40
Long Int	39.71	71.01	46.58	39.54	71.49	46.67	39.34	70.97	46.43
IG	39.23	69.43	49.51	39.13	69.34	49.53	39.23	69.45	49.61
UTR	40.00	68.75	47.06	40.00	68.57	46.75	40.00	68.67	46.99
Reg	40.00	68.18	47.06	40.00	67.69	47.34	40.00	68.33	47.37

Tabella A.14: Median values of dichotomic class percentages computed in antisense strand for each gt1 sequence class

A.4.2 Mean values

	$p0$	$r0$	$h0$	$p1$	$r1$	$h1$	$p2$	$r2$	$h2$
Genes	43.57	61.17	48.05	44.67	61.42	48.27	43.93	61.51	48.72
CDS	40.60	55.92	44.39	51.72	58.43	45.40	47.60	61.19	49.85
Exons	43.32	58.13	45.47	49.06	60.61	46.13	46.40	60.86	48.08
Introns	40.62	65.47	53.21	39.34	67.23	52.84	39.12	67.47	52.95
Long Int	39.91	66.89	52.43	39.48	67.36	52.33	39.47	67.54	52.40
IG	39.44	68.68	49.46	39.54	68.75	49.52	39.32	68.68	49.55
UTR	40.02	65.11	51.35	39.91	65.15	51.53	39.90	65.15	51.49
Reg	40.31	65.00	51.49	39.99	65.26	51.75	40.19	64.71	51.62

Tabella A.15: Mean values of dichotomic class percentages computed in sense strand for each gt1 sequence class

	$p0a$	$r0a$	$h0a$	$p1a$	$r1a$	$h1a$	$p2a$	$r2a$	$h2a$
Genes	44.10	61.43	50.60	44.03	60.75	49.80	43.26	60.97	50.50
CDS	51.23	60.95	56.47	47.02	50.87	45.76	39.56	56.36	53.84
Exons	48.28	59.60	53.11	45.88	55.36	47.84	42.45	57.64	51.37
Introns	40.31	71.27	43.31	39.79	73.91	43.58	38.94	72.02	42.90
Long Int	39.81	71.06	46.24	39.48	71.79	46.36	39.17	71.17	46.09
IG	39.36	68.60	49.52	39.39	68.60	49.68	39.56	68.65	49.68
UTR	40.37	68.21	46.58	40.05	67.98	46.16	40.18	68.21	46.32
Reg	40.04	68.15	46.20	40.14	67.74	45.87	39.77	68.05	45.66

Tabella A.16: Mean values of dichotomic class percentages computed in antisense strand for each gt1 sequence class

A.5 Global transformation 2

Here is median values tables for sequences undergone to purine/pyrimidine global transformation :

A.5.1 Median values

	$p0$	$r0$	$h0$	$p1$	$r1$	$h1$	$p2$	$r2$	$h2$
Genes	43.95	59.50	51.29	43.75	60.00	51.32	43.12	59.03	50.88
CDS	49.75	53.58	55.77	48.43	59.08	54.75	41.22	46.15	50.23
Exons	48.00	54.10	53.57	47.37	57.14	52.94	43.48	50.00	50.82
Introns	38.89	70.21	44.44	39.50	70.37	44.68	40.62	70.37	44.44
Long Int	39.56	68.89	46.64	39.71	69.05	46.78	40.00	69.07	46.72
IG	39.15	68.14	50.34	39.18	68.00	50.33	39.16	68.10	50.35
UTR	40.00	66.67	48.18	40.00	66.67	47.95	40.00	66.67	47.95
Reg	40.00	66.67	48.15	40.00	66.67	47.51	40.00	66.67	47.46

Tabella A.17: Median values of dichotomic class percentages computed in sense strand for each gt2 sequence class

	$p0a$	$r0a$	$h0a$	$p1a$	$r1a$	$h1a$	$p2a$	$r2a$	$h2a$
Genes	43.20	59.33	49.67	43.12	59.53	50.28	43.48	59.05	49.68
CDS	45.38	56.54	43.08	43.10	58.87	54.16	48.70	50.99	46.05
Exons	45.16	56.97	44.44	44.00	57.69	50.00	46.77	53.49	46.21
Introns	39.39	65.45	53.68	38.46	64.29	53.57	38.29	64.52	53.91
Long Int	39.55	64.85	52.90	39.31	64.60	52.78	39.29	64.63	53.02
IG	39.14	68.18	50.33	39.22	68.13	50.34	39.15	68.14	50.23
UTR	40.00	63.64	52.24	40.00	63.92	52.55	40.19	63.64	52.17
Reg	40.00	63.71	51.06	40.00	63.93	50.53	41.00	64.29	50.87

Tabella A.18: Median values of dichotomic class percentages computed in antisense strand for each gt2 sequence class

A.5.2 Mean values

	$p0$	$r0$	$h0$	$p1$	$r1$	$h1$	$p2$	$r2$	$h2$
Genes	44.81	58.77	51.75	44.14	59.11	51.52	43.16	57.62	51.07
CDS	49.87	53.55	55.61	48.37	58.70	54.60	41.33	45.98	50.15
Exons	47.87	54.00	53.08	47.29	56.46	52.33	43.74	50.35	50.37
Introns	38.79	70.28	44.02	39.60	70.44	44.30	40.71	70.40	44.19
Long Int	39.33	69.37	46.68	39.58	69.43	46.76	39.95	69.36	46.69
IG	39.38	67.51	50.21	39.47	67.46	50.12	39.58	67.50	50.08
UTR	39.78	66.44	47.67	39.99	66.40	47.39	39.85	66.30	47.43
Reg	40.33	66.90	46.41	40.28	66.26	45.58	39.81	66.65	45.71

Tabella A.19: Mean values of dichotomic class percentages computed in sense strand for each gt2 sequence class

	$p0a$	$r0a$	$h0a$	$p1a$	$r1a$	$h1a$	$p2a$	$r2a$	$h2a$
Genes	43.71	58.83	49.20	43.67	59.18	50.00	43.99	58.46	49.30
CDS	45.39	56.03	43.45	43.67	58.59	54.16	48.94	50.81	46.08
Exons	45.15	56.73	44.93	44.08	57.28	50.07	46.71	53.59	46.54
Introns	39.55	65.40	53.92	38.42	64.23	53.55	38.26	64.52	54.24
Long Int	39.43	65.52	52.87	39.08	65.18	52.73	39.06	65.24	52.99
IG	39.44	67.64	50.18	39.53	67.53	50.00	39.46	67.56	50.00
UTR	39.88	63.20	52.39	40.30	63.40	52.72	40.29	63.41	52.56
Reg	39.61	62.77	51.22	40.26	63.44	50.81	40.82	63.78	51.02

Tabella A.20: Mean values of dichotomic class percentages computed in antisense strand for each gt2 sequence class

Appendice B

Distributions

B.1 Dichotomic classes distributions

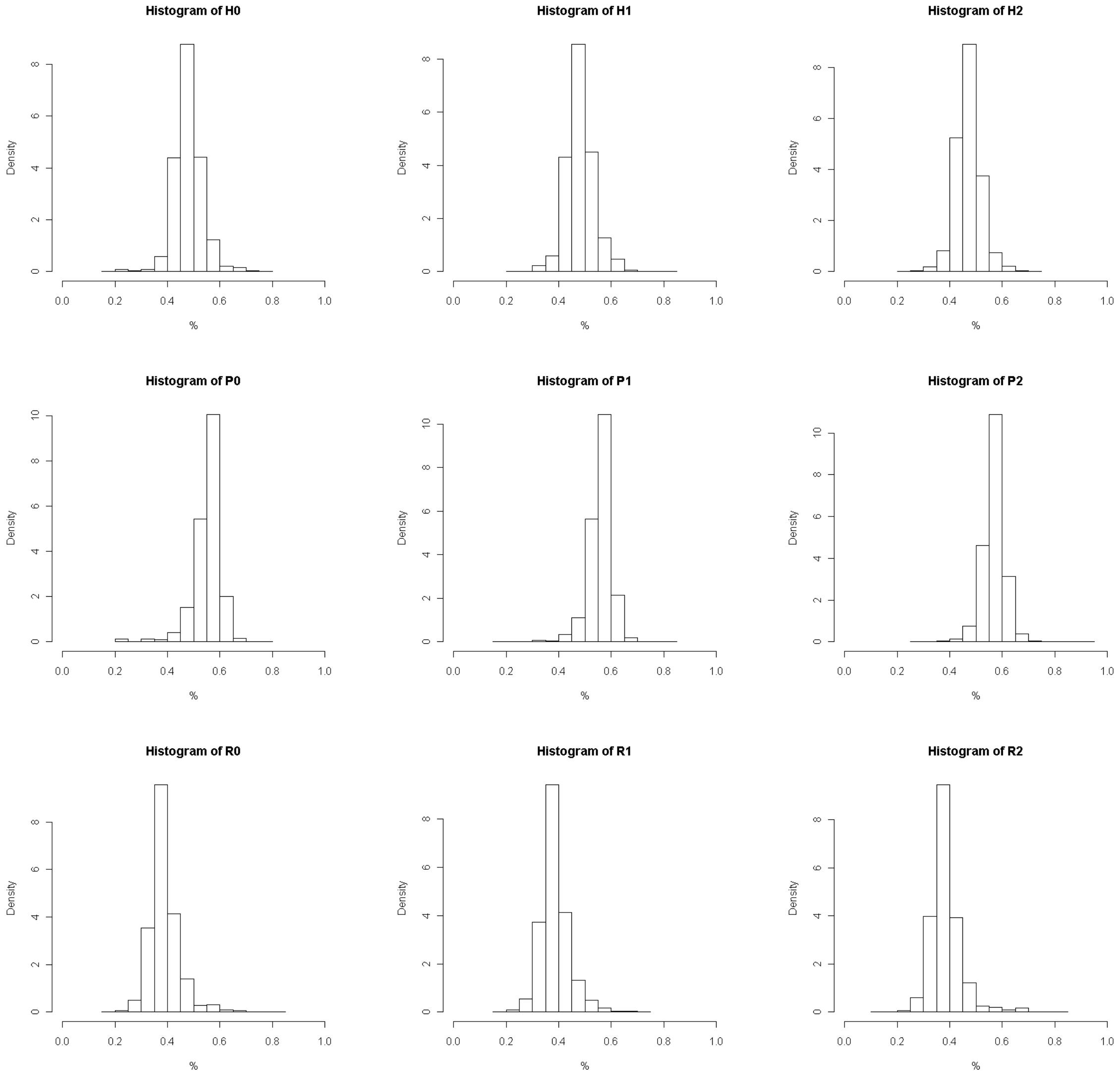


Figura B.1: proportions of dichotomic classes (sense) in gene sequences

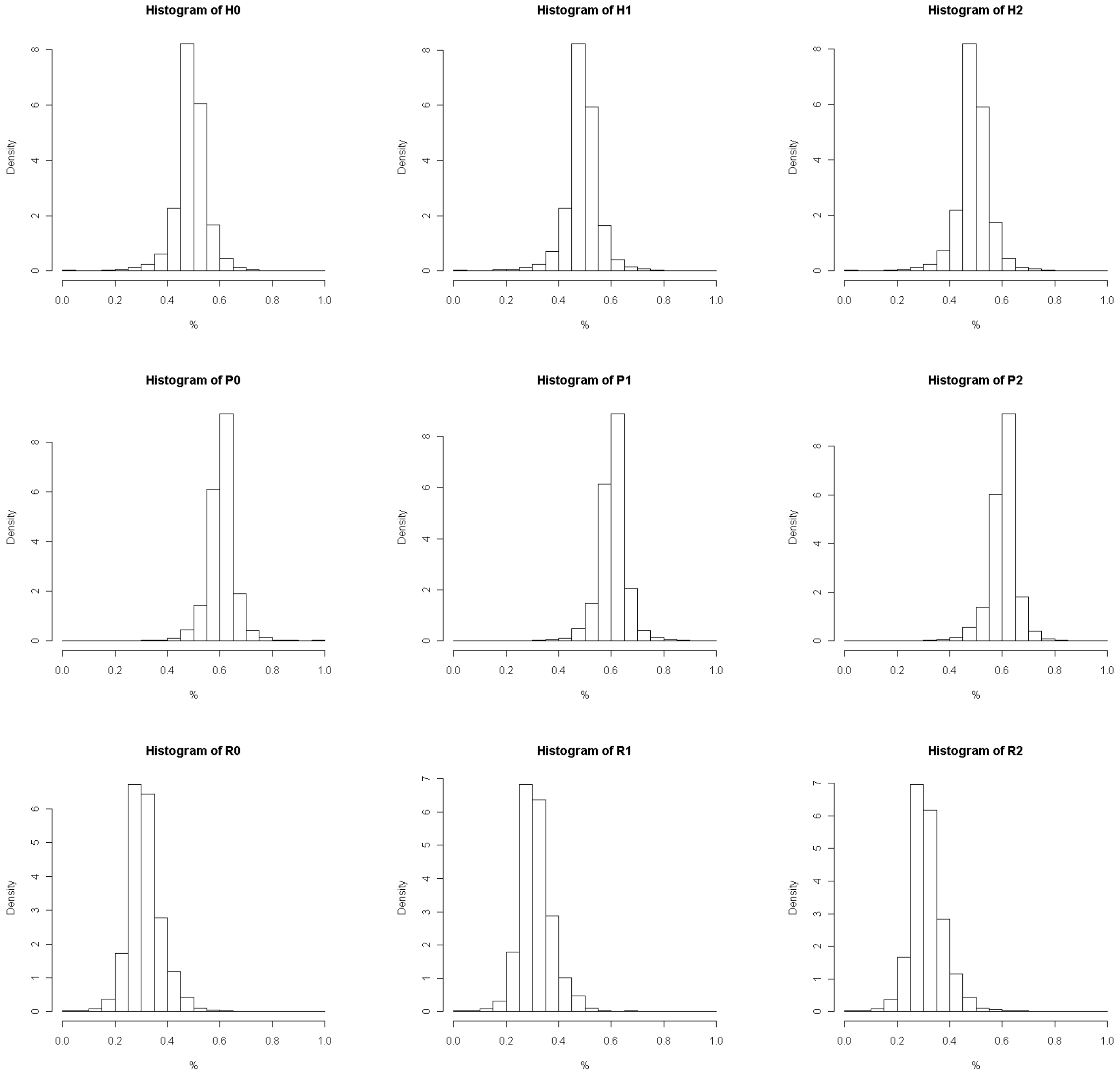


Figura B.2: proportions of dichotomic classes (sense) in intergene sequences

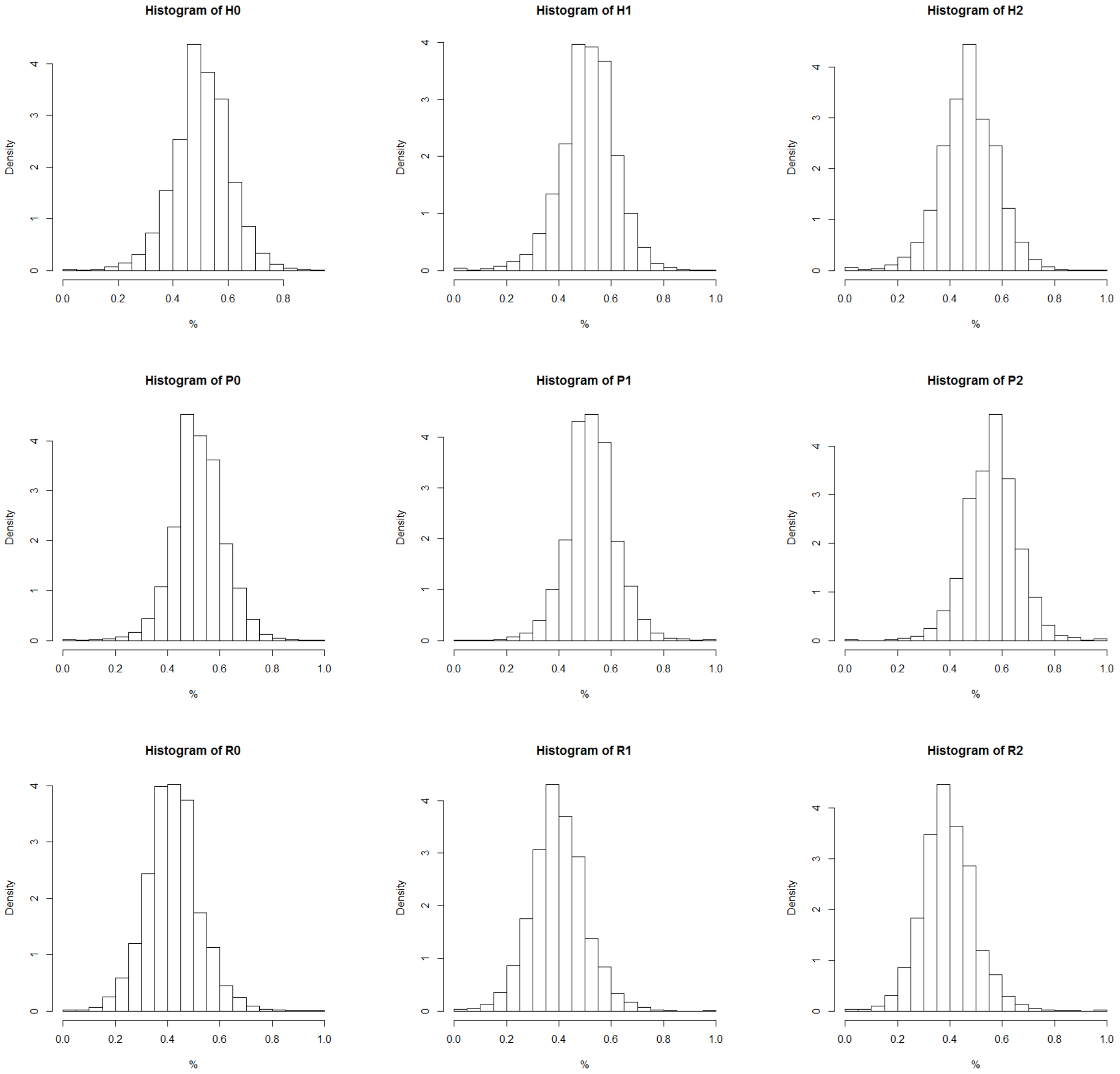


Figura B.3: proportions of dichotomic classes (sense) in exon sequences

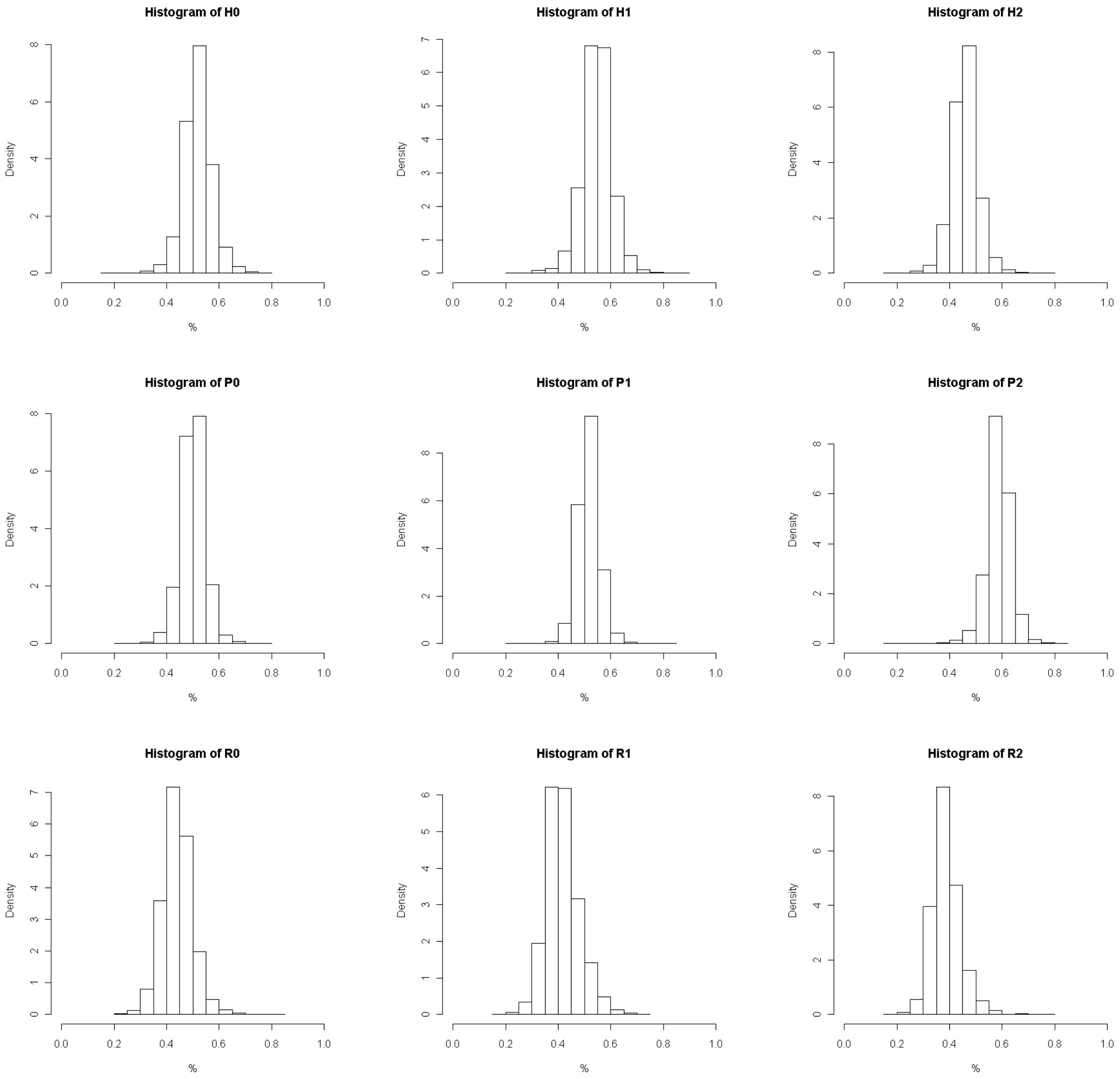


Figura B.4: proportions of dichotomic classes (sense) in CDS sequences

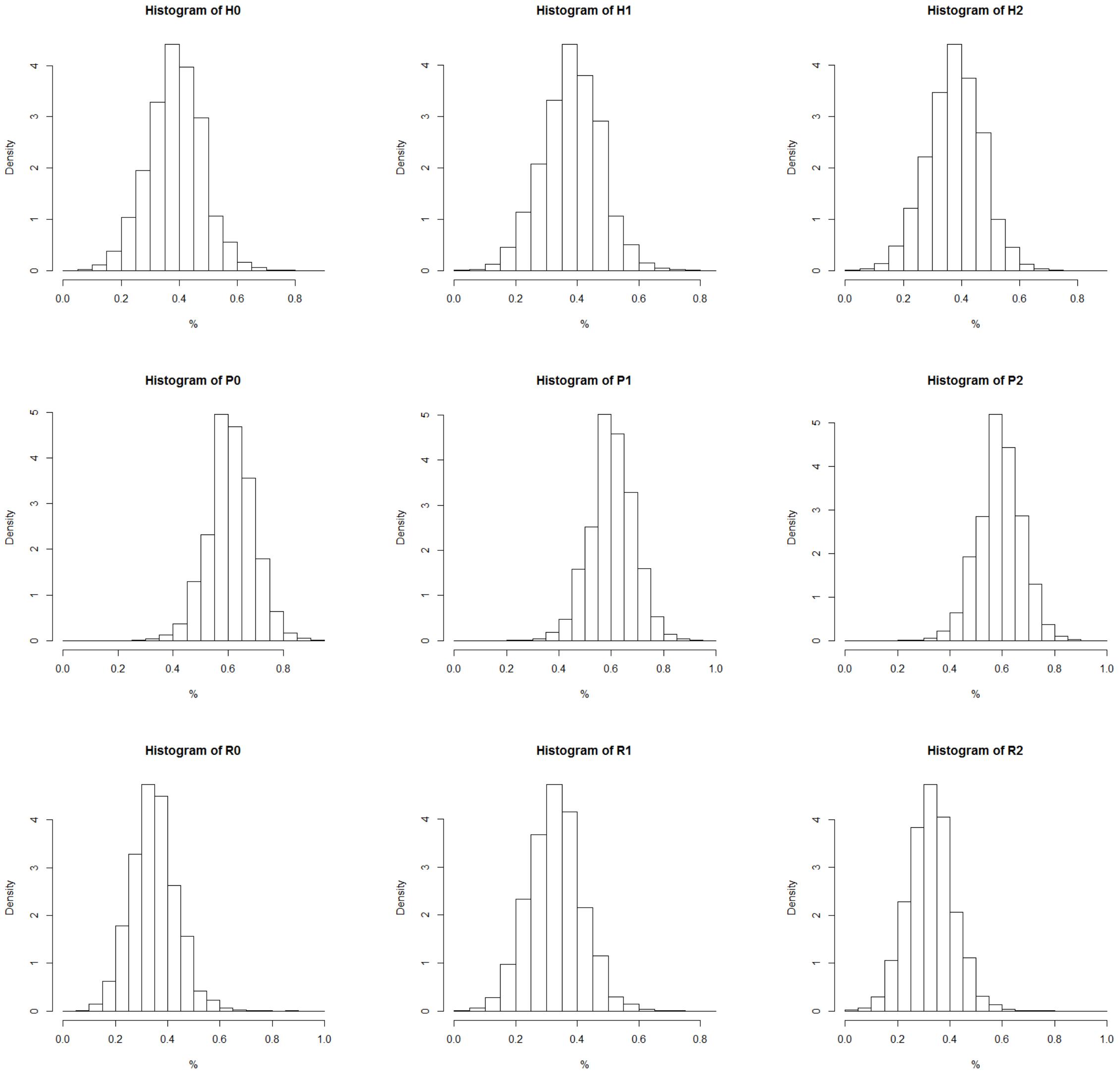


Figura B.5: proportions of dichotomic classes (sense) in intron sequences

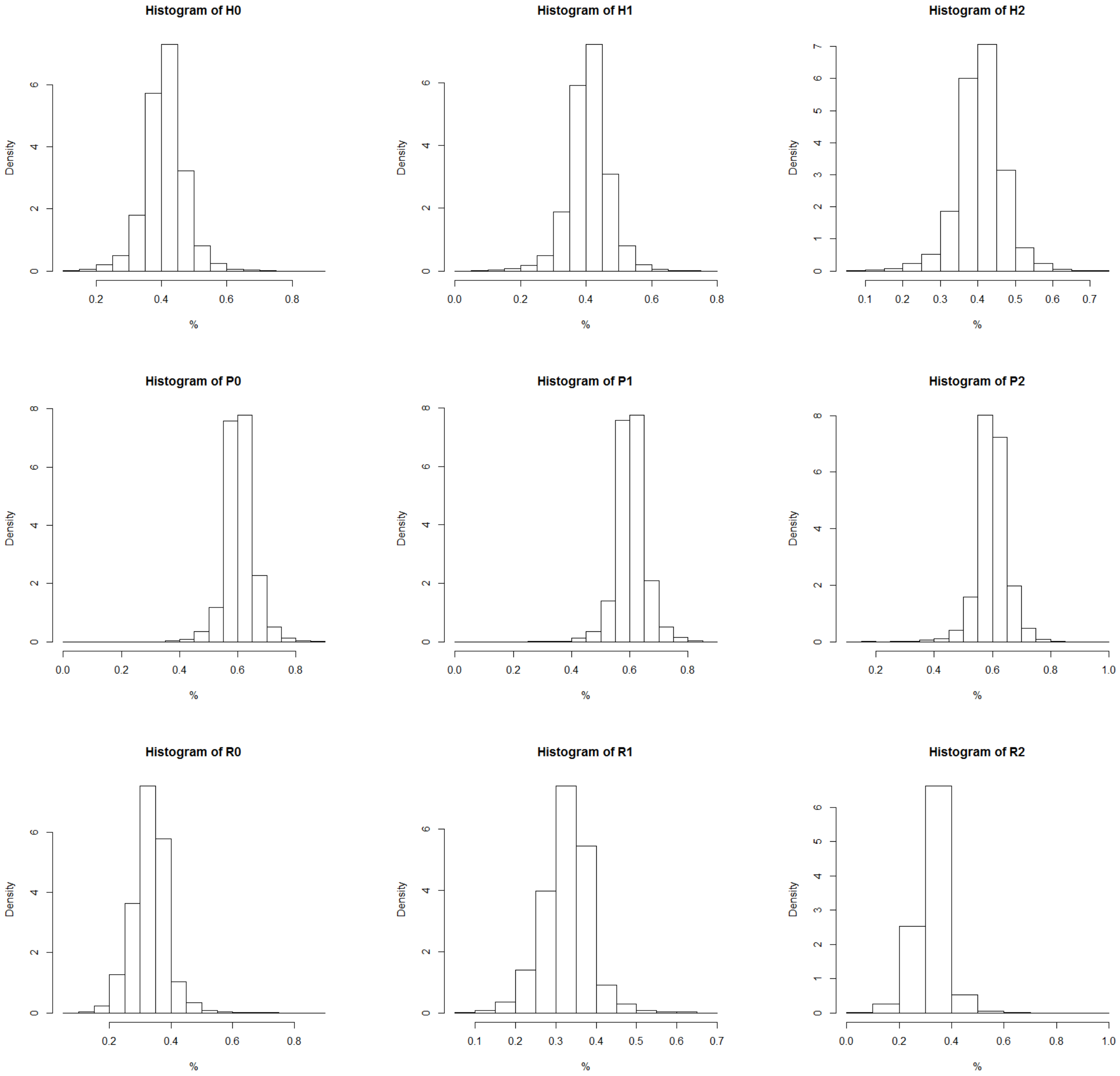


Figura B.6: proportions of dichotomic classes (sense) in long intron sequences

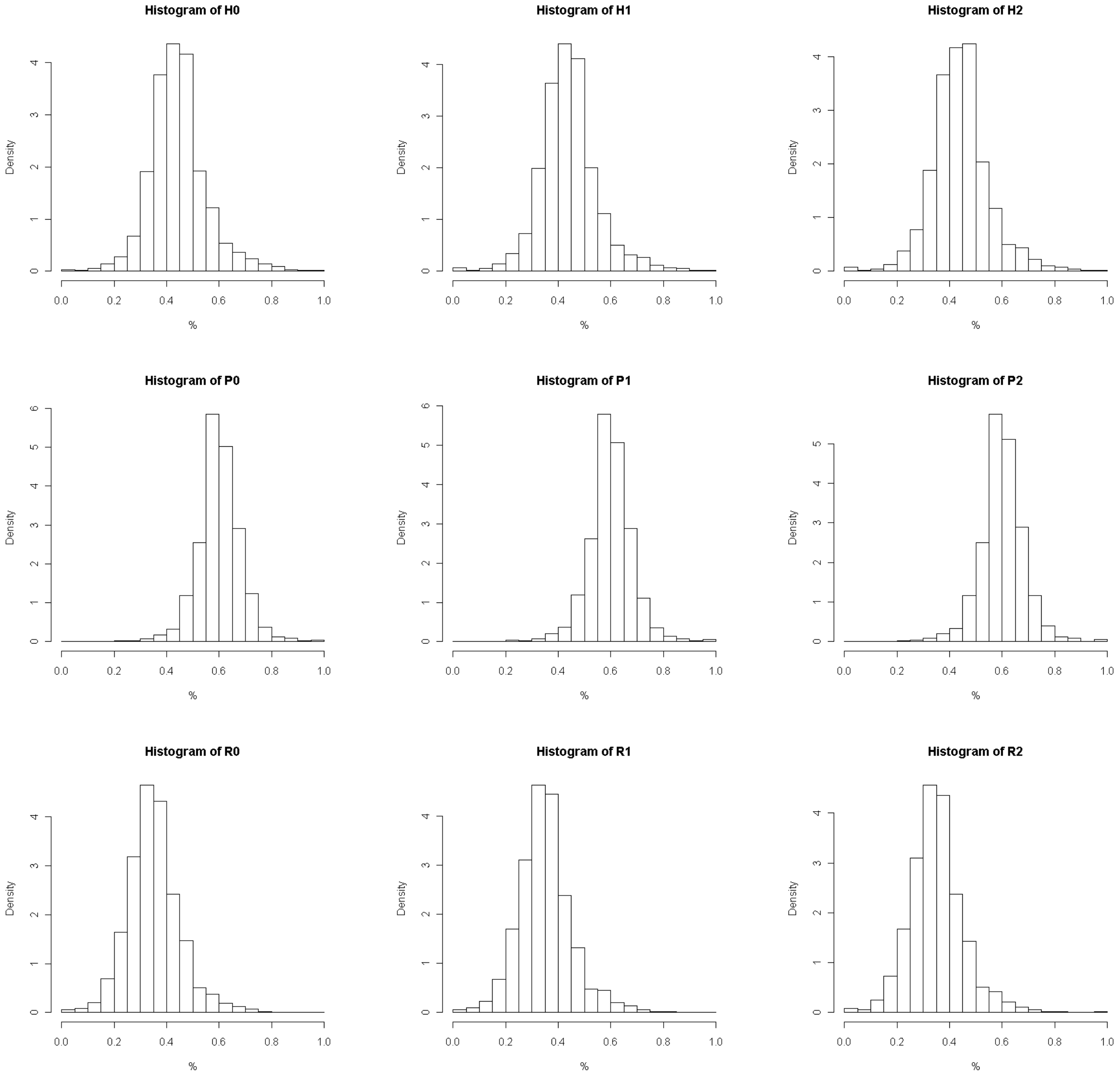


Figura B.7: proportions of dichotomic classes (sense) in UTR sequences

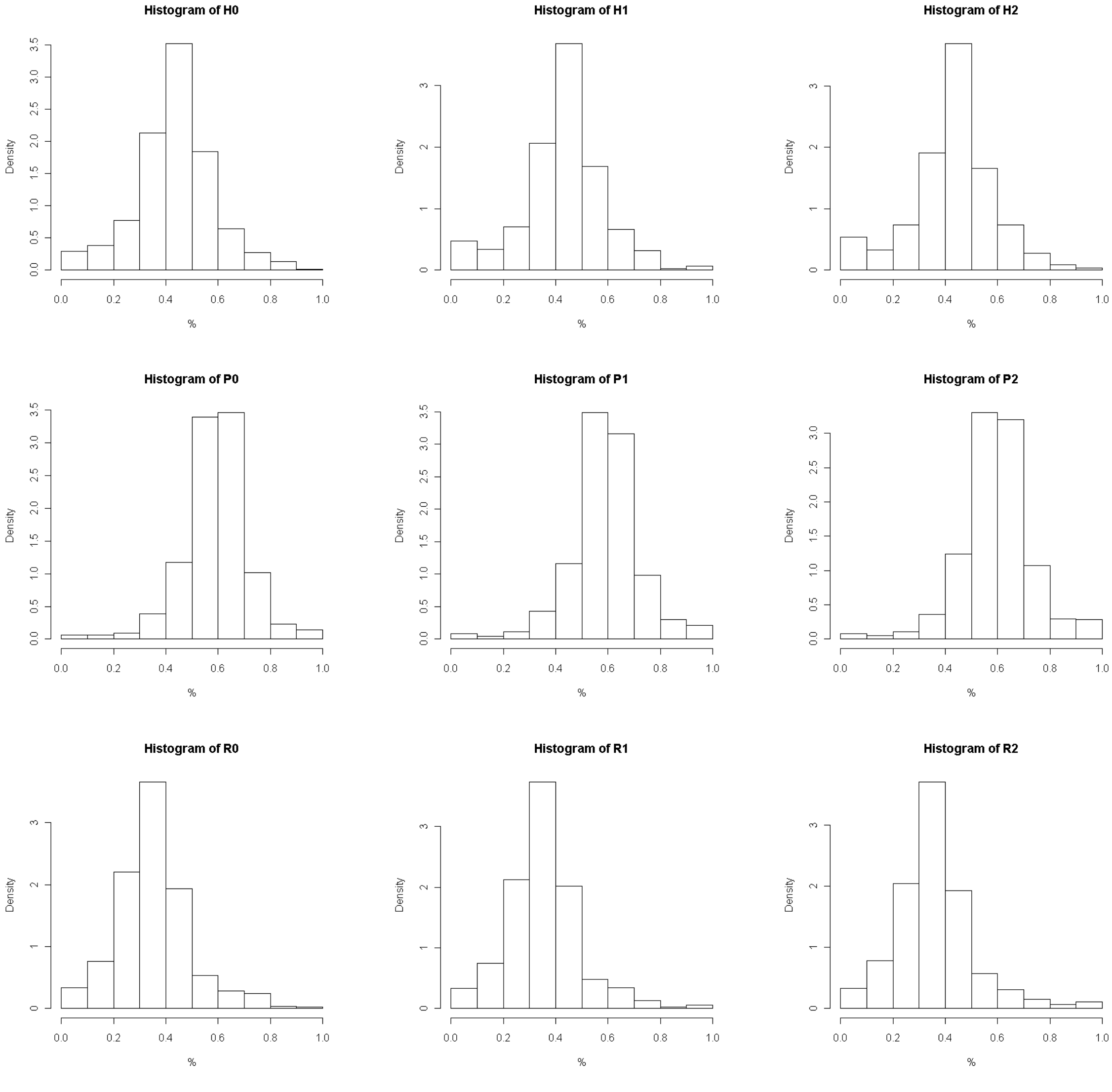


Figura B.8: proportions of dichotomic classes (sense) in regulatory sequences

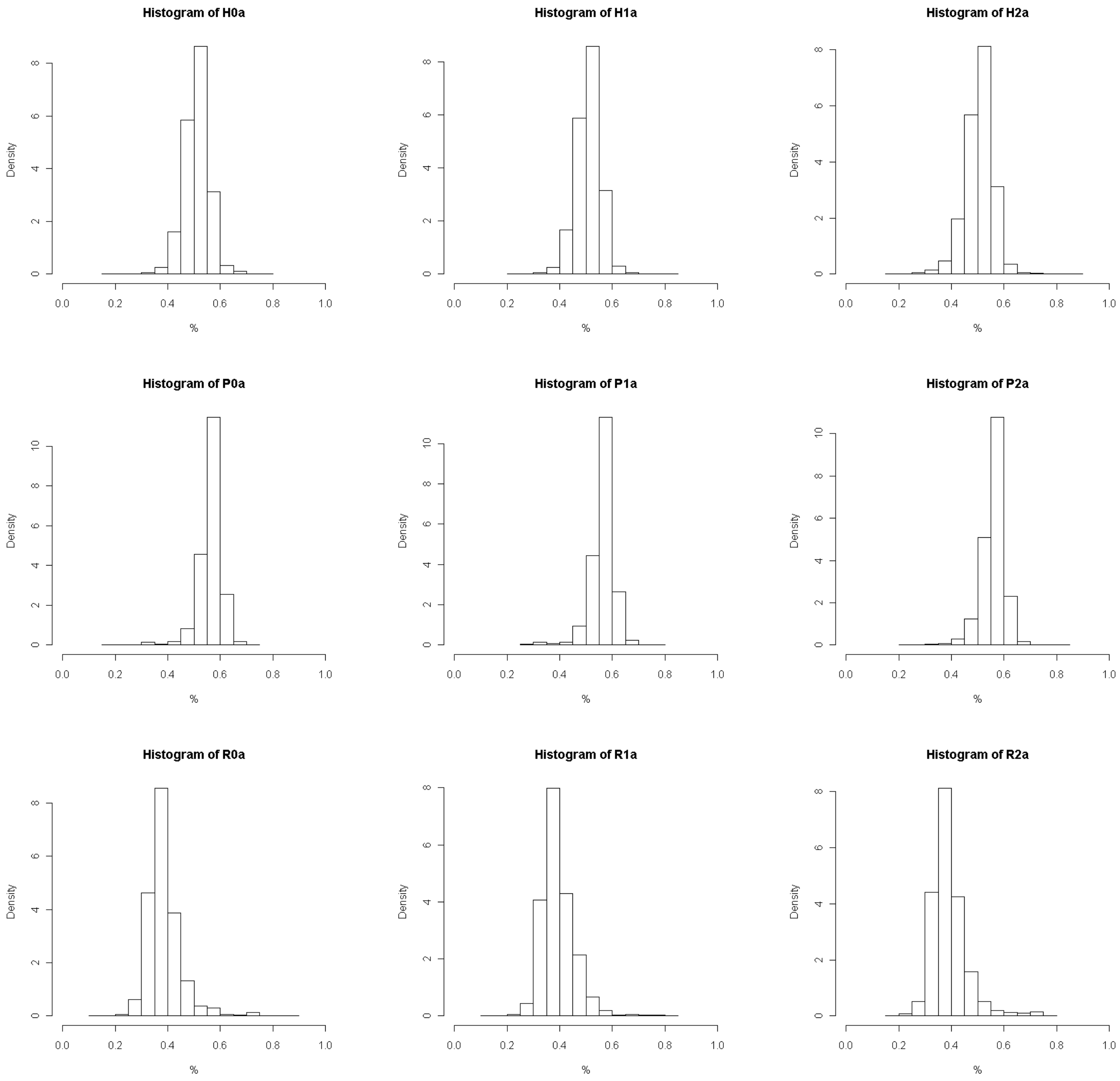


Figura B.9: proportions of dichotomic classes (antisense) in gene sequences

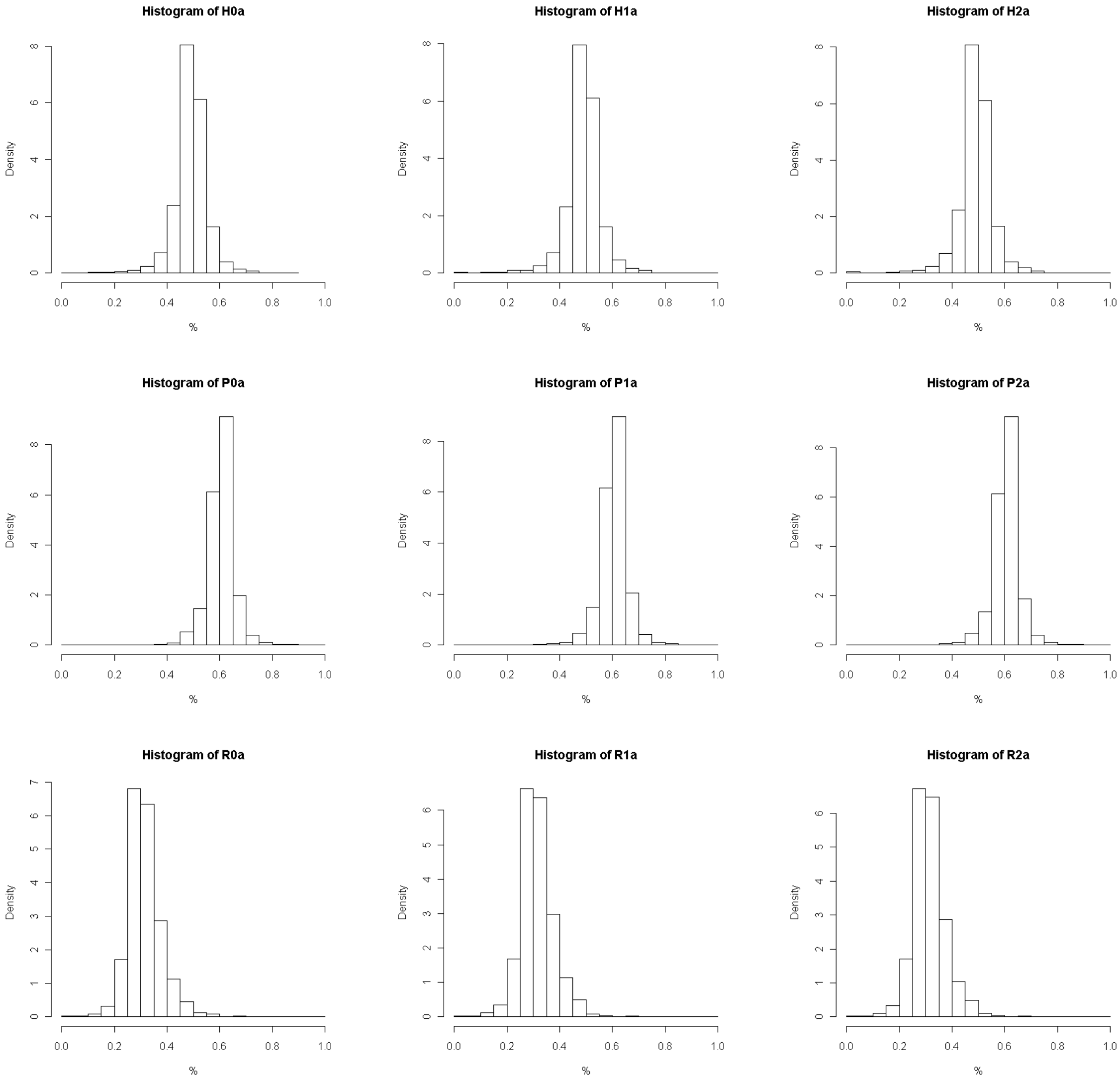


Figura B.10: proportions of dichotomic classes (antisense) in intergene sequences

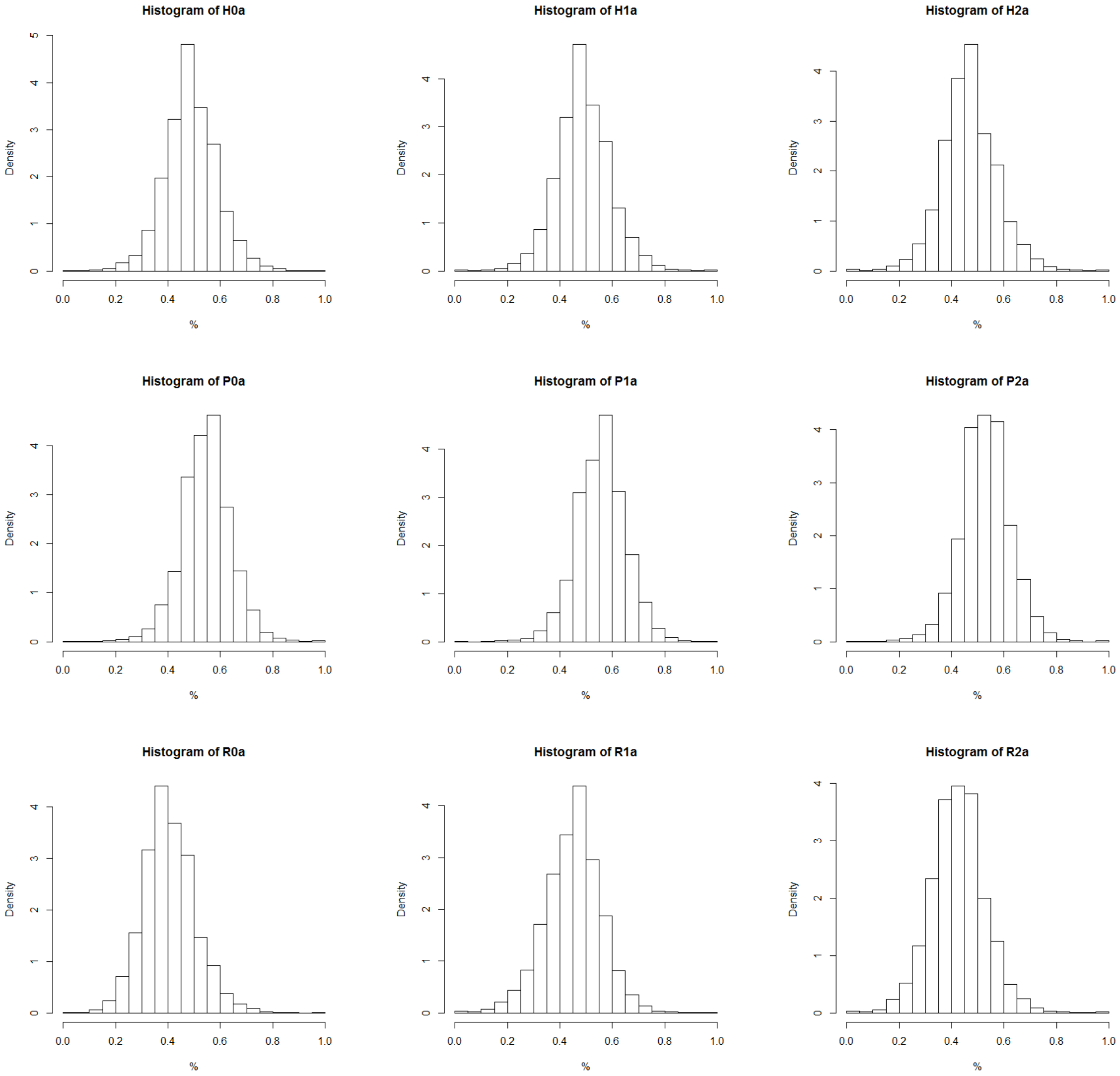


Figura B.11: proportions of dichotomic classes (antisense) in exon sequences

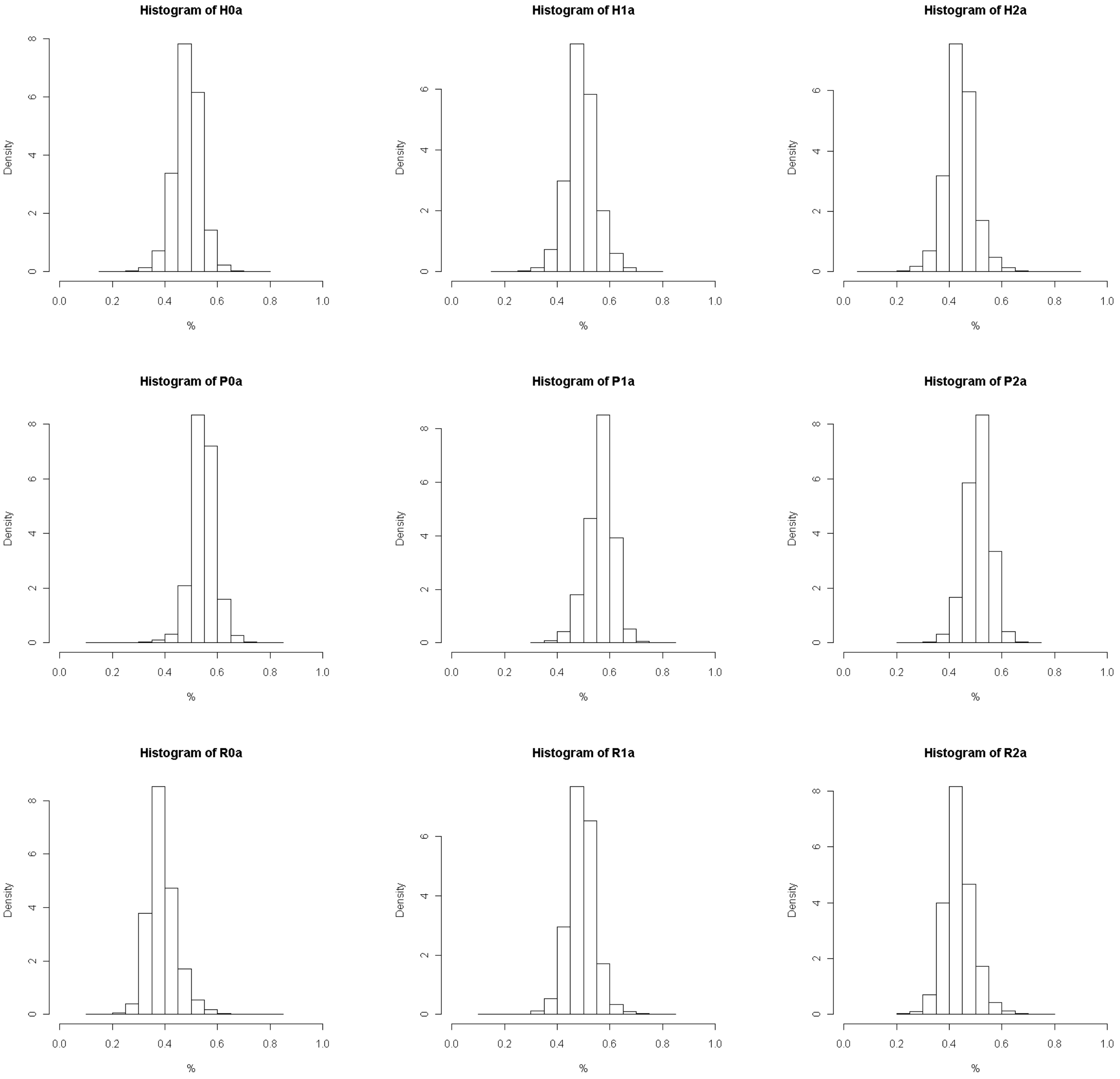


Figura B.12: proportions of dichotomic classes (antisense) in CDS sequences

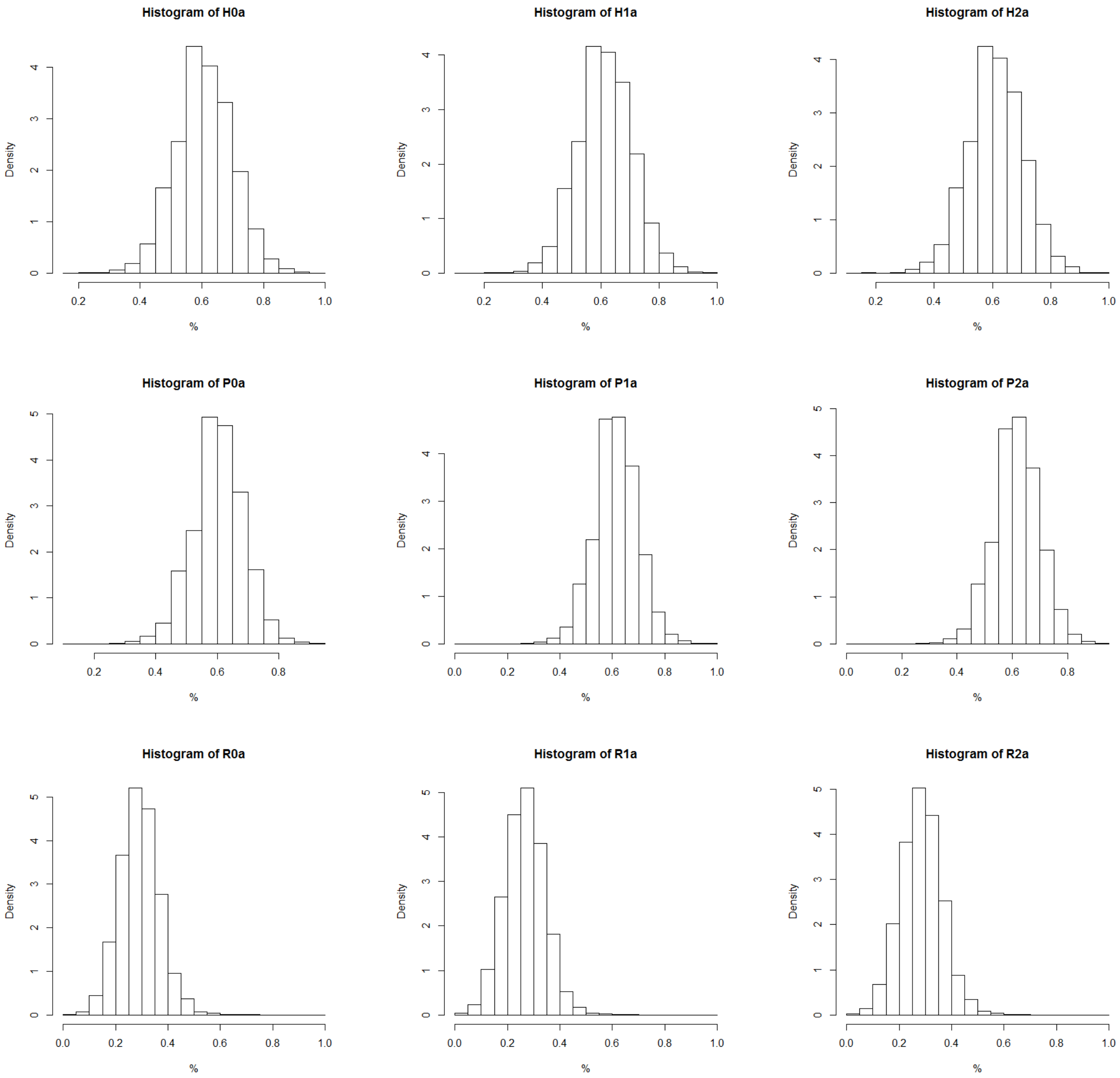


Figura B.13: proportions of dichotomic classes (antisense) in intron sequences

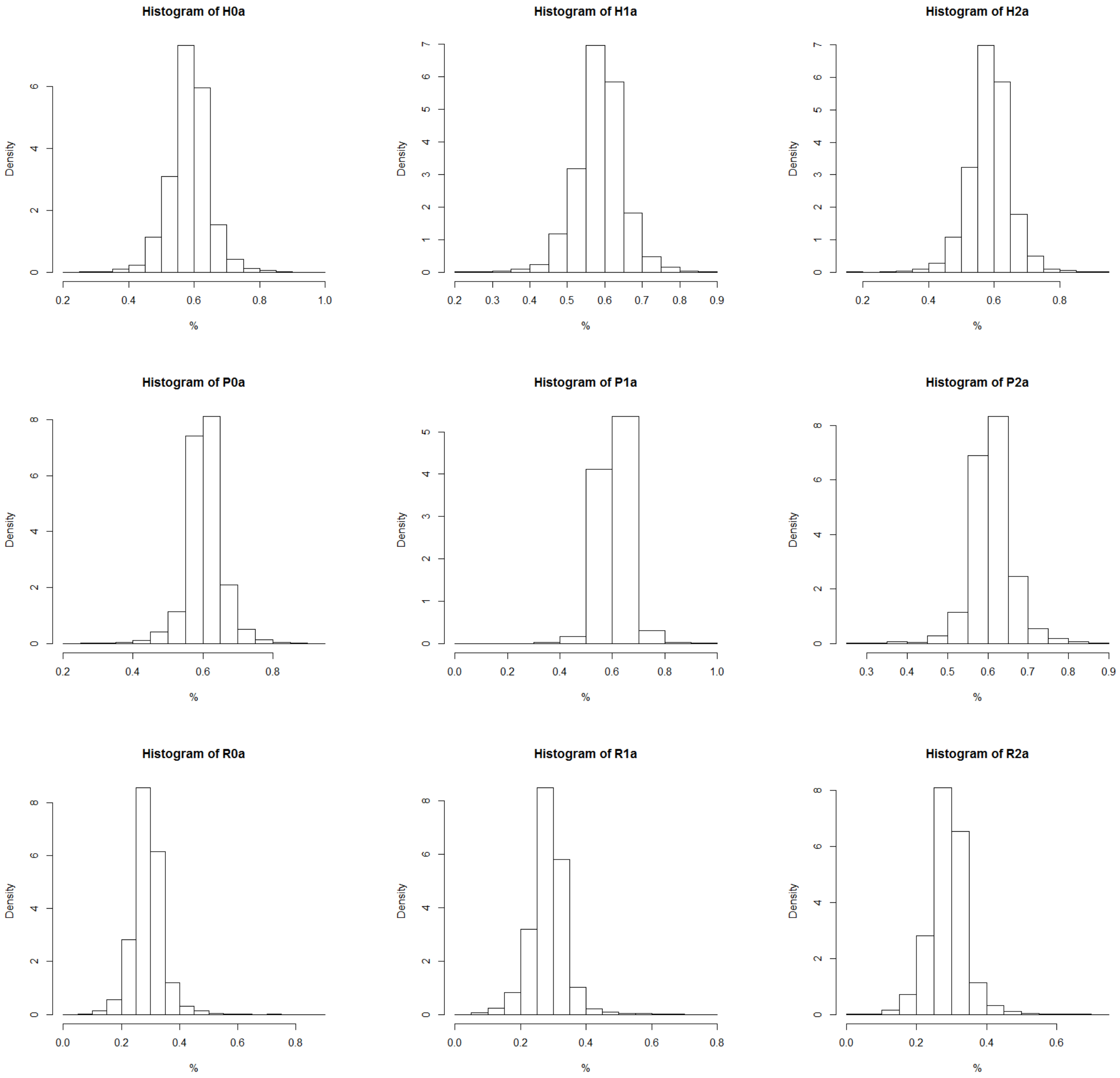


Figura B.14: proportions of dichotomic classes (antisense) in long intron sequences

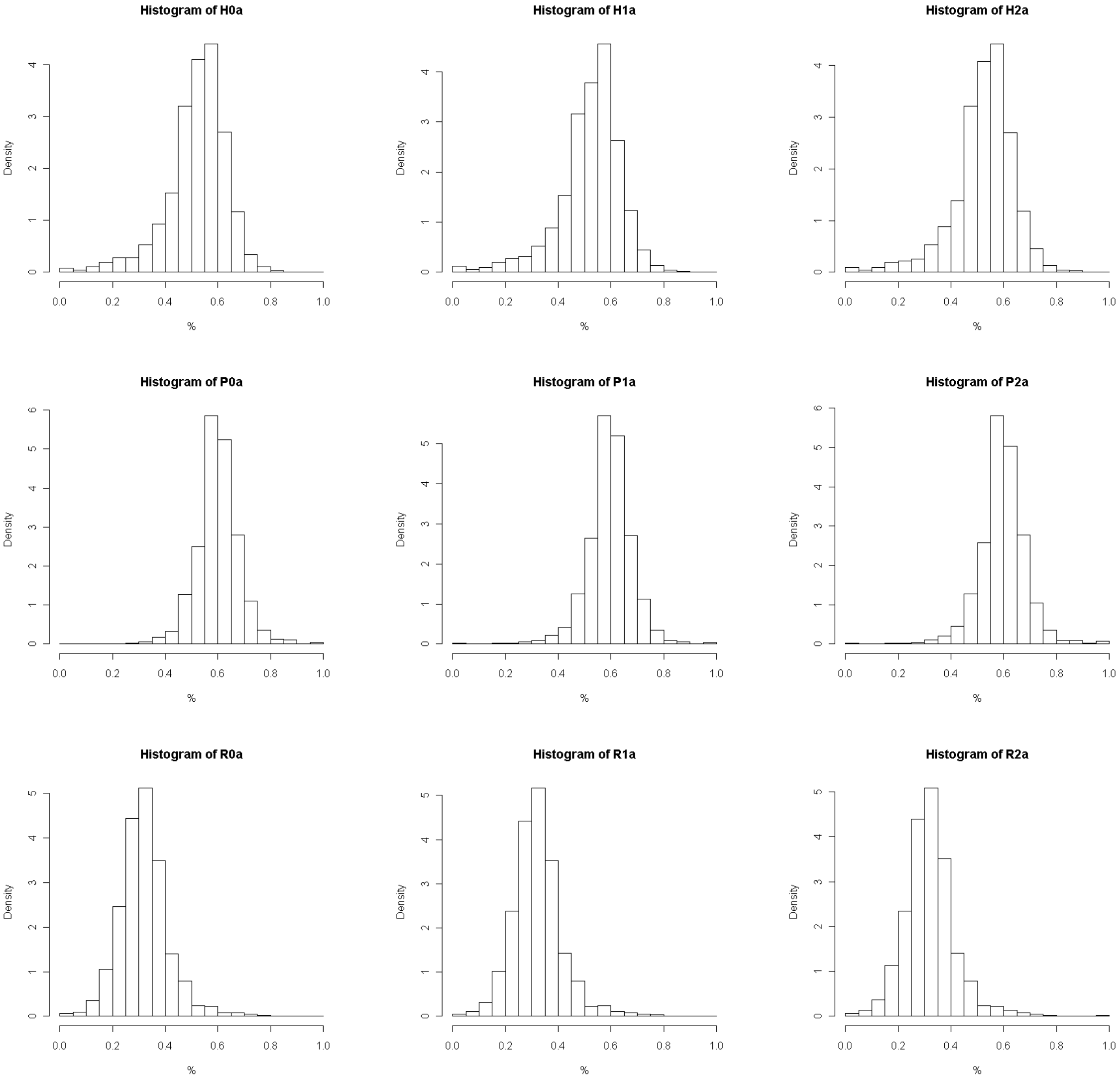


Figura B.15: proportions of dichotomic classes (antisense) in UTR sequences

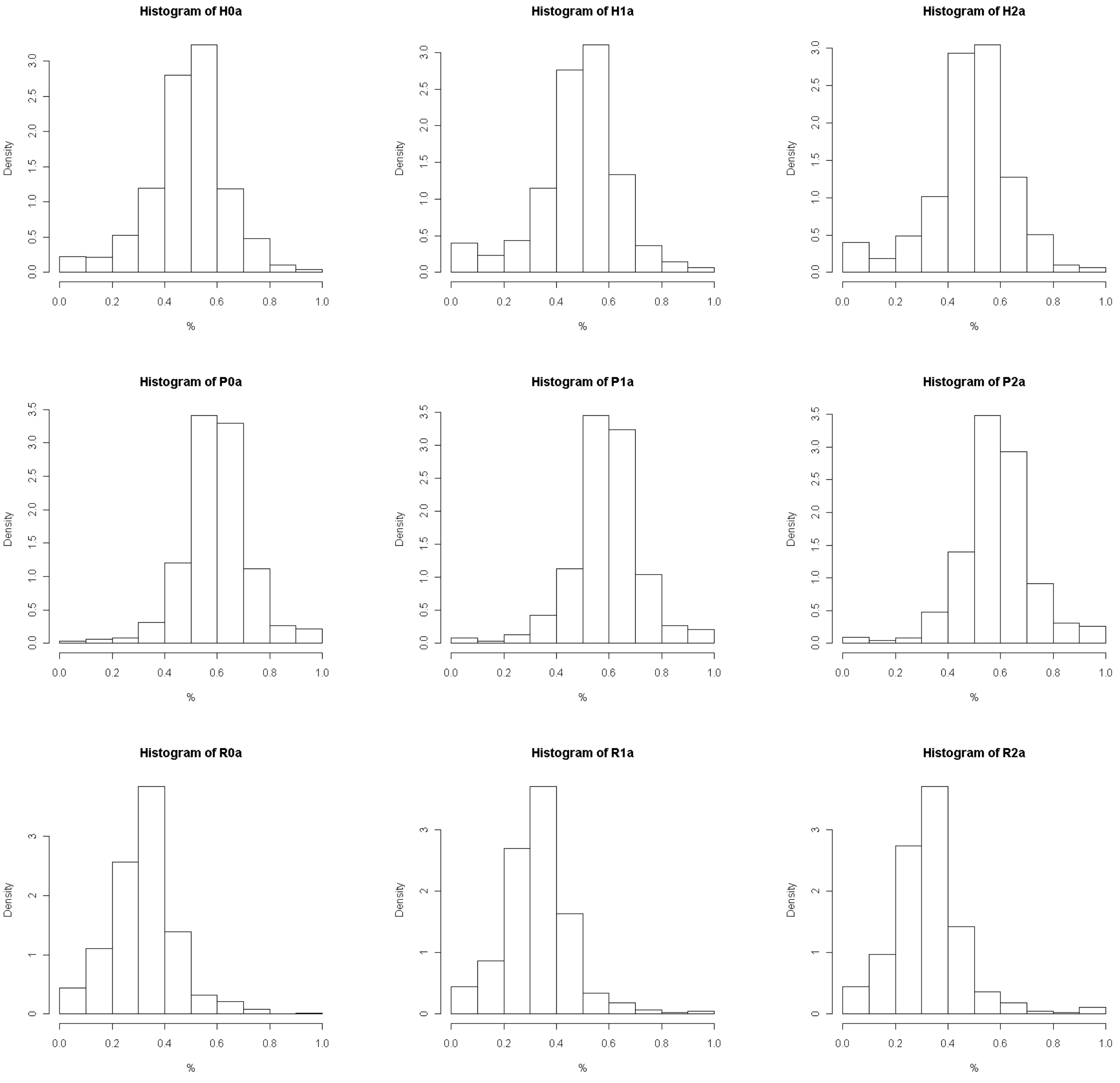
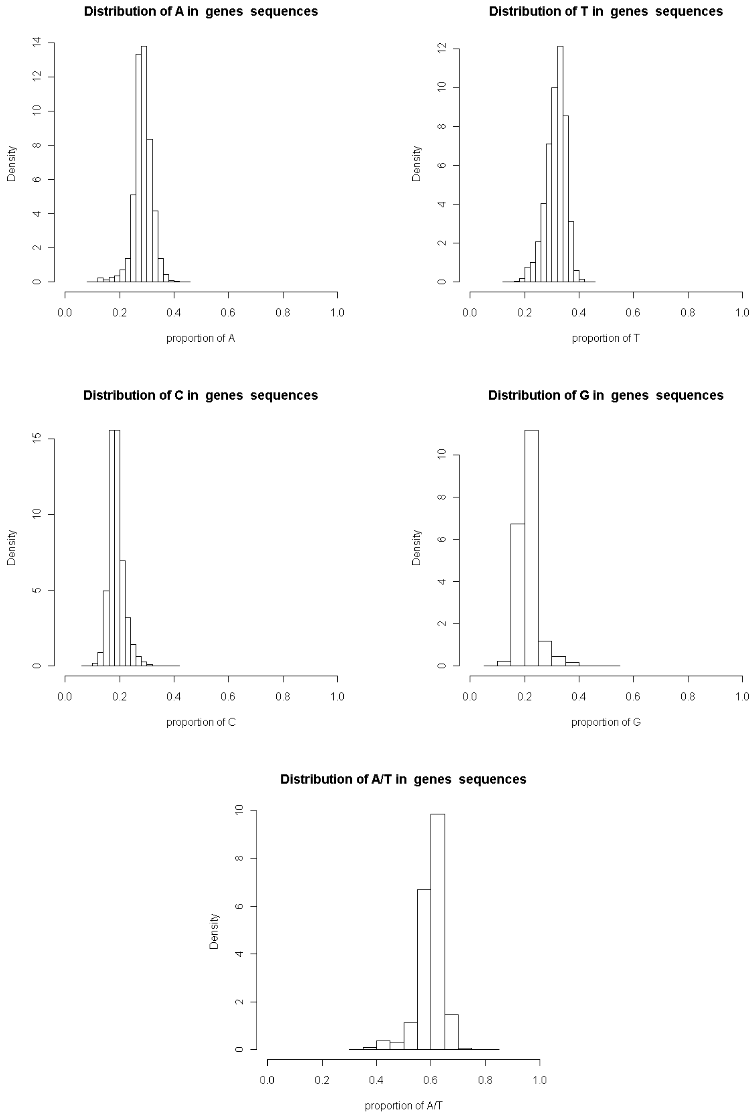


Figura B.16: proportions of dichotomic classes (antisense) in regulatory sequences

B.2 Base distribution


Distribution of A in genes sequences

Distribution of T in genes sequences

Distribution of C in genes sequences

Distribution of G in genes sequences

Distribution of A/T in genes sequences

Figura B.17: proportions of bases in gene sequences

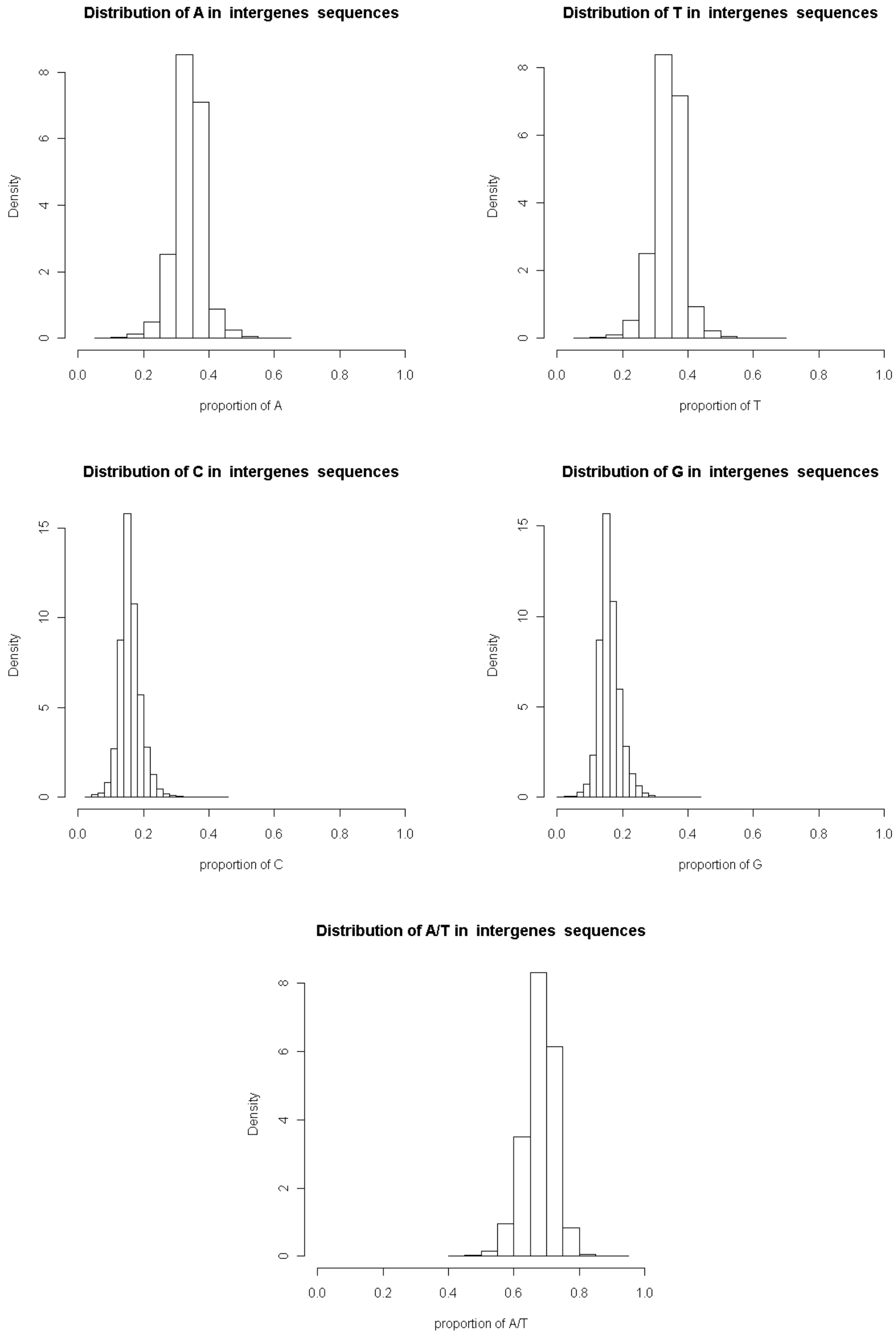


Figura B.18: proportions of bases in intergene sequences

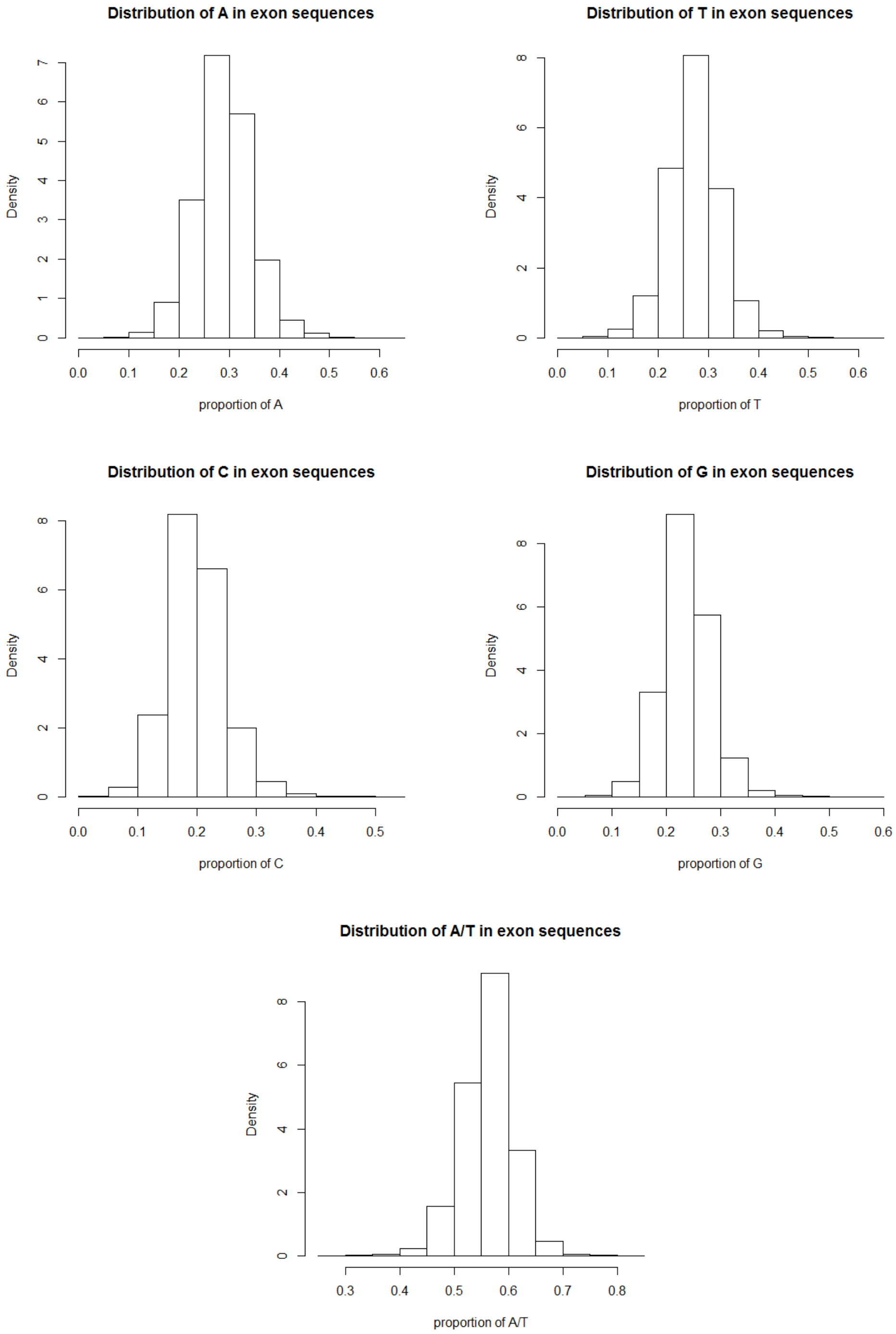


Figura B.19: proportions of bases in exon sequences

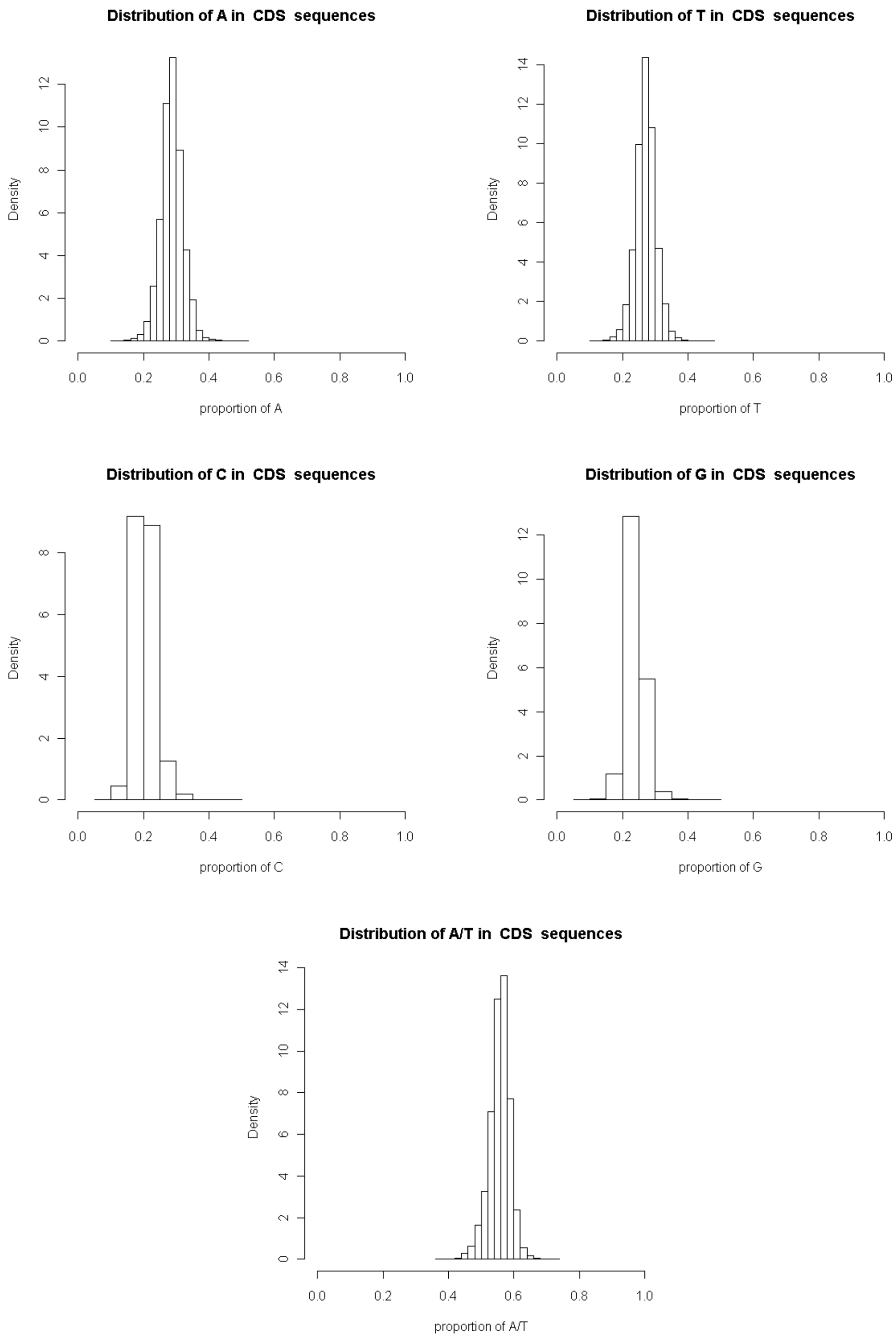


Figura B.20: proportions of bases in coding sequences (CDS)

Distribution of A in intron sequences

Distribution of T in intron sequences

Distribution of C in intron sequences

Distribution of G in intron sequences

Distribution of A/T in intron sequences

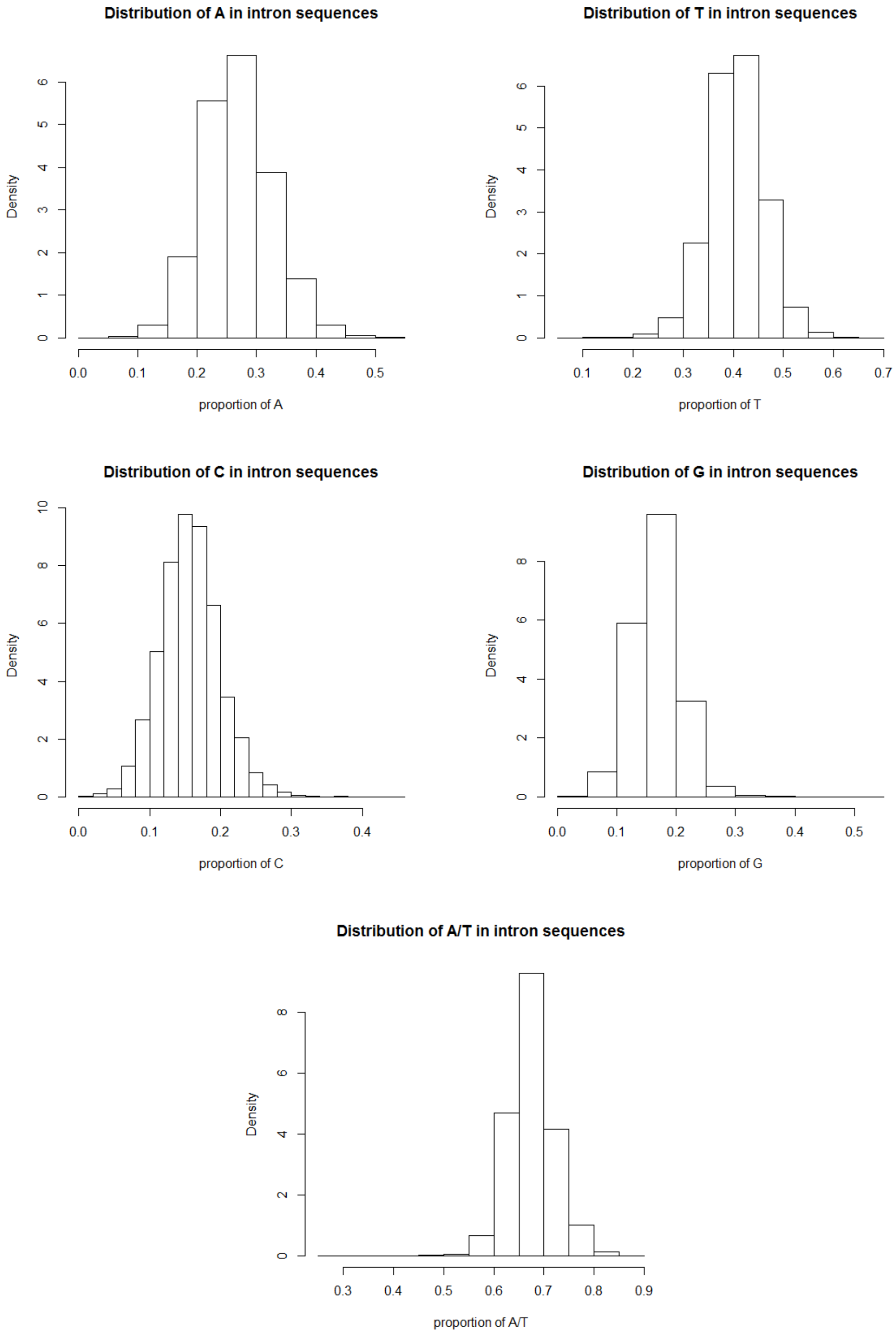
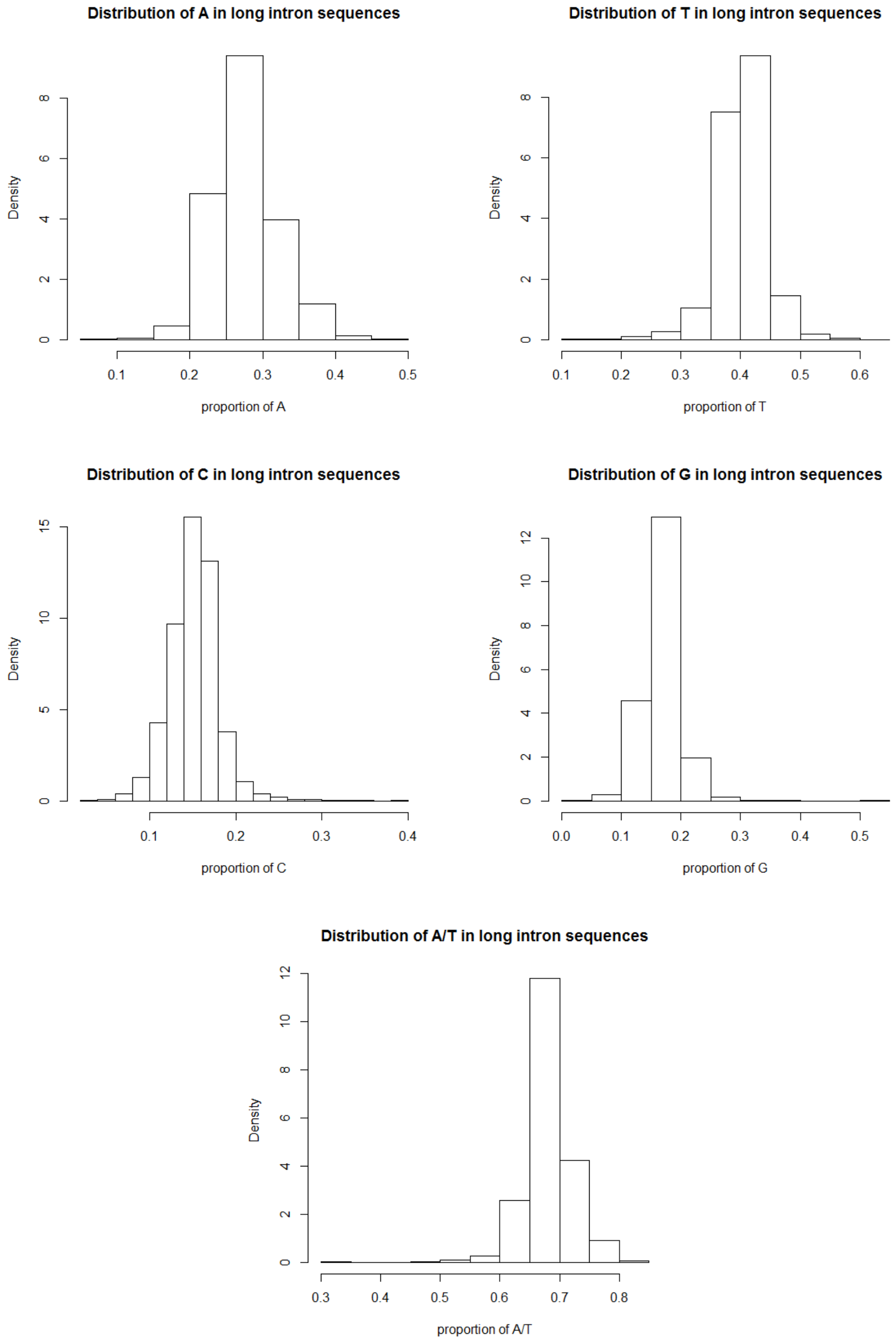


Figura B.21: proportions of bases in intron sequences


Distribution of A in long intron sequences

Distribution of T in long intron sequences

Distribution of C in long intron sequences

Distribution of G in long intron sequences

Distribution of A/T in long intron sequences

Figura B.22: proportions of bases in long intron sequences

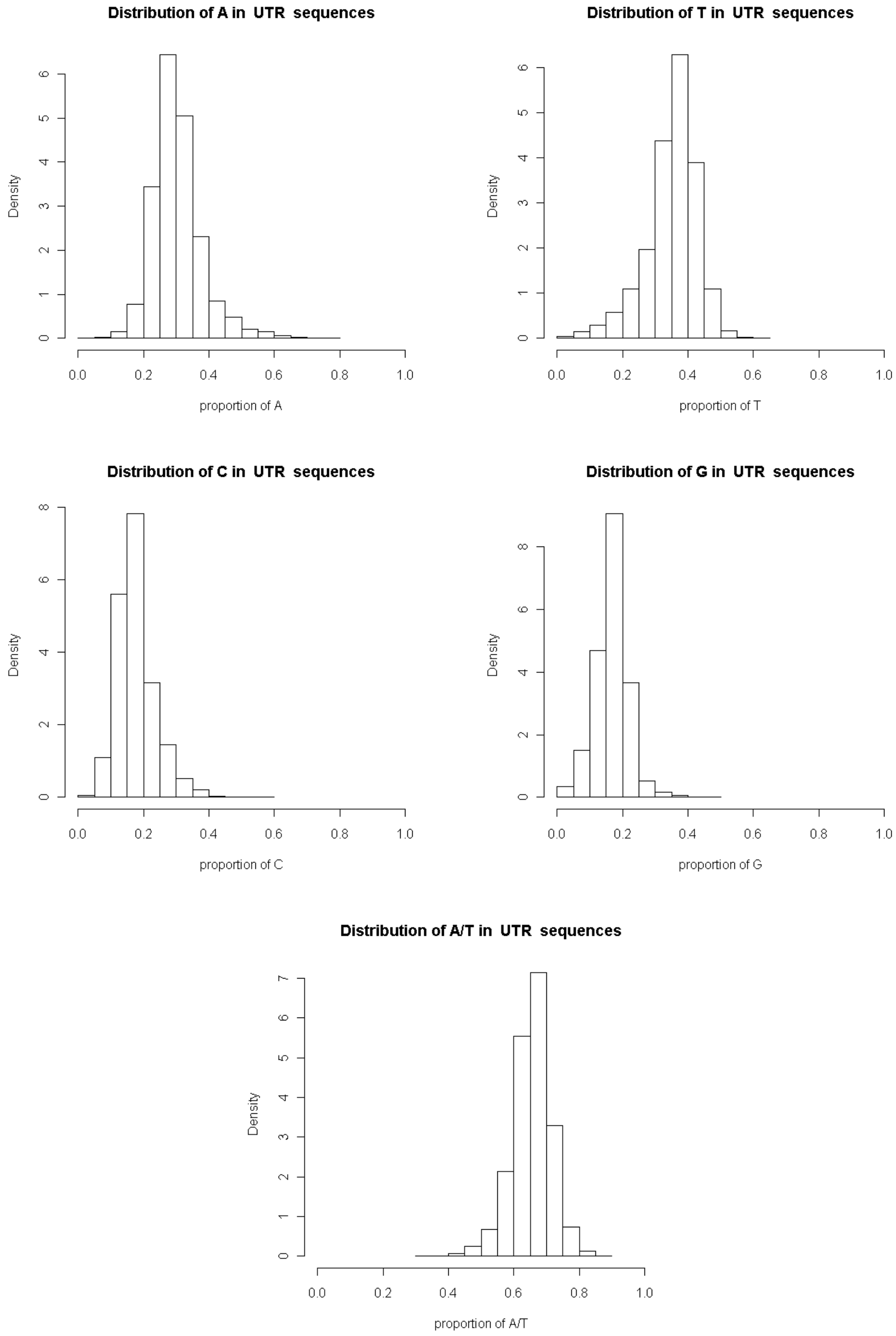


Figura B.23: proportions of bases in UTR sequences

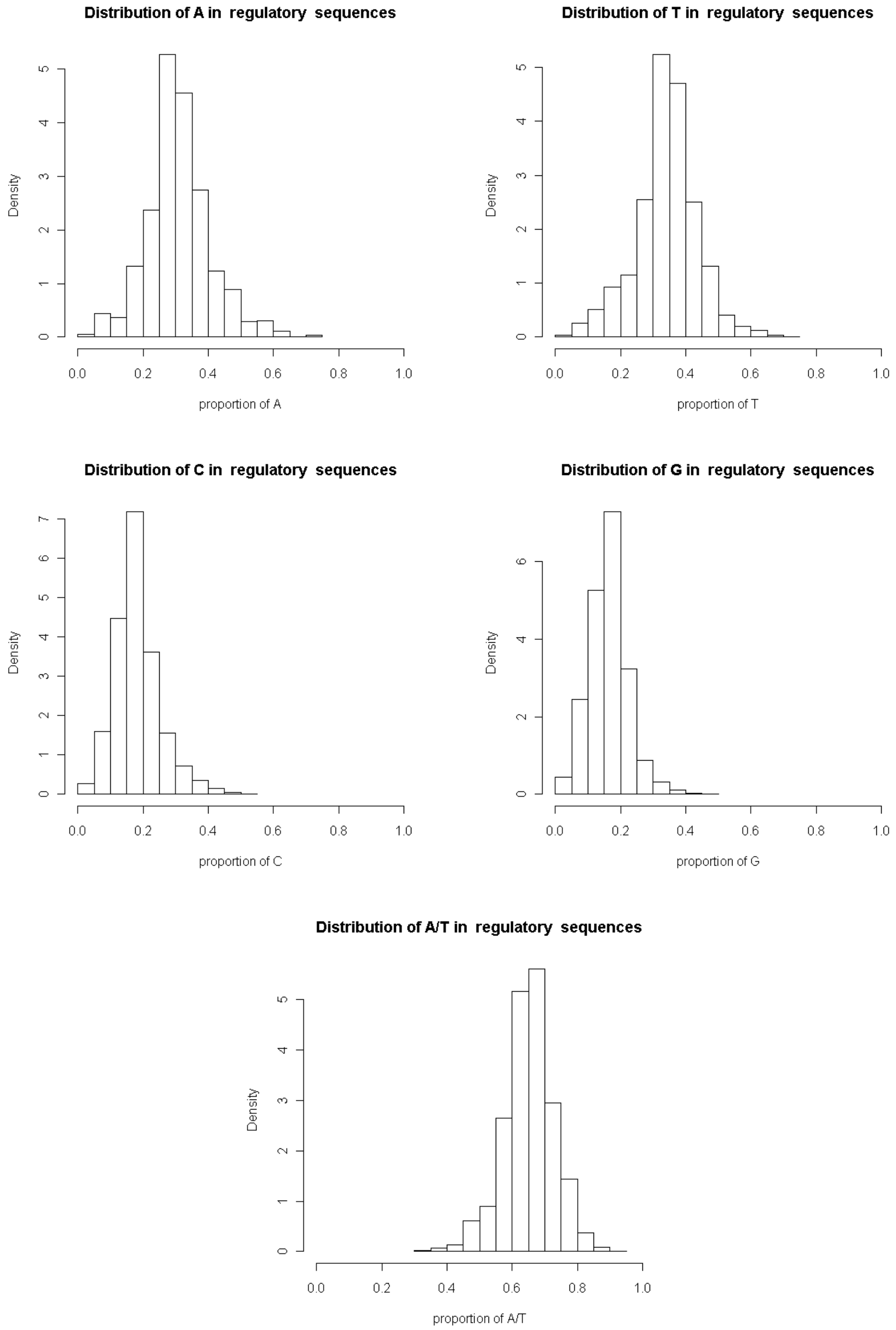


Figura B.24: proportions of bases in regulatory sequences

Bibliografia

[1] F.J. Aherne, N.A. Thacker, and P.I. Rockett. The bhattacharyya metric as an absolute similarity measure for frequency coded data. *Kybernetika*, 34(4):363–368, 1998.

[2] E.G. Baek and W.A. Brock. A non parametric test for independence of a multivariate time series. *Statistica Sinica*, 1992.

[3] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[4] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

[5] B.K. Bhattacharya and G.T. Toussaint. An upper bound on the probability of misclassification in terms of matusita's measure of affinity. *Annals of the Institute of Statistical Mathematics*, 34(1):161–165, 1982.

[6] T.M. Cover and J.A. Thomas. *Elements of information theory.* Wiley-interscience, 2006.

[7] F. Crick. Central dogma of molecular biology. *Nature*, 1970.

[8] Francis Crick. *"Chapter 8: The genetic code". What mad pursuit: a personal view of scientific discovery.* New York: Basic Books, 1988.

[9] G.A. Darbellay and D. Wuertz. The entropy as a tool for analysing statistical dependences in financial time series. *Physica A: Statistical Mechanics and its Applications*, 287(3):429–439, 2000.

[10] A. Dionísio, R. Menezes, and D.A. Mendes. Entropy-based independence test. *Nonlinear Dynamics*, 44(1):351–357, 2006.

[11] B. Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2010.

[12] M. Fernandes and B. Neri. Nonparametric entropy-based tests of independence between stochastic processes. *Econometric Reviews*, 2010.

[13] S. Giannerini, D. L. Gonzalez, and R. Rosa. Dichotomic classes and short range dependence in coding sequences. In *Second International Workshop on "Mathematical models for the genetic information"*.

[14] S. Giannerini, E. Maasoumi, and E.B. Dagum. *Bulletin of the International Statistical Institute, 56th Session*, 2007.

[15] D. L. Gonzalez. Can the genetic code be mathematically described? *Medical Science Monitor*, 10(4):11–17, 2004.

[16] D. L. Gonzalez. Error detection and correction codes. In M. Barbieri and J. Hoffmeyer, editors, *The Codes of Life: The Rules of Macroevolution*, volume 1 of *Biosemiotics*, chapter 17, pages 379–394. Springer Netherlands, 2008.

[17] D. L. Gonzalez. The mathematical structure of the genetic code. In M. Barbieri and J. Hoffmeyer, editors, *The Codes of Life: The Rules of Macroevolution*, volume 1 of *Biosemiotics*, chapter 8, pages 111–152. Springer Netherlands, 2008.

[18] D. L. Gonzalez, S. Giannerini, and R. Rosa. Detecting structure in parity binary sequences: Error correction and detection in DNA. *IEEE Engineering in Medicine and Biology Magazine*, 25:69–81, 2006.

[19] D. L. Gonzalez, S. Giannerini, and R. Rosa. Strong short-range correlations and dichotomic codon classes in coding DNA sequences. *Physical Review E*, 78(5):051918, 2008.

[20] D. L. Gonzalez, S. Giannerini, and R. Rosa. The mathematical structure of the genetic code: a tool for inquiring on the origin of life. *Statistica*, LXIX(3–4):143–157, 2009.

[21] D.L. Gonzalez, S. Giannerini, and R. Rosa. Circular codes revisited: A statistical approach. *Journal of Theoretical Biology*, 275(1):21–28, 2011.

[22] D.L. Gonzalez, S. Giannerini, and R. Rosa. On the origin of the mitochondrial genetic code. Working paper, 2012.

[23] C.W. Granger and J.L. Lin. Using the mutual information coefficient to identify lags in nonlinear models. *Journal of Time Series Analysis*, 15(4):371–384, 1994.

[24] C.W. Granger, E. Maasoumi, and J.C. Racine. A dependence metric for possibly nonlinear processes. *Journal of Time Series Analysis*, 2004.

[25] S.G. Gregory, K.F. Barlow, and et al. The dna sequence and biological annotation of human chromosome 1. *Nature*, 2006.

[26] W.M. Griffiths, J.H. Miller, D.T. Suzuki, and al. *An Introduction to Genetic Analysis (7th ed.).* W. H. Freeman., 2000.

[27] Joseph M. Hilbe. *Logistic Regression Models.* Chapman & Hall/CRC Press, 2009.

[28] Y. Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988.

[29] W. Hoeffding. A non parametric test of independence. *Annals of mathematical statistics*, 1948.

[30] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

[31] D.W. Hosmer and S. Lemeshow. *Applied Logistic Regression.* Wiley, 2000.

[32] J. Hsu. *Multiple comparisons: theory and methods.* Chapman & Hall/CRC, 1996.

[33] S. Ihara. *Information theory for continuous systems*, volume 2. World Scientific Publishing Company Incorporated, 1993.

[34] The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant arabidopsis thaliana. *Nature*, 2000.

[35] Berg J, Tymoczko J.L., and Stryer L. *Biochemistry (6th ed.).* W. H. Freeman, 2006.

[36] D. Johnson and R. McClelland. A general dependence test and applications. *Journal of Applied Econometrics*, 1998.

[37] E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[38] B. Lewin. *Genes Ix*. Jones and Bartlett Publishers Sudbury, MA, USA:, 2008.

[39] W. Li. Mutual information functions versus correlation functions. *Journal of statistical physics*, 60(5):823–837, 1990.

[40] E. Maasoumi and J. Racine. Entropy and predictability of stock market returns. *Journal of Econometrics*, 107(1):291–312, 2002.

[41] M.Nirenberg. *Deciphering the genetic code*. Office of NIH History, 2010.

[42] P.K. Mulligan, R. C. King, and W.D. Stansfield. *A dictionary of genetics*. Oxford University Press, 2006.

[43] H. Pearson. Genetics: What is a gene? *Nature*, 441(7092):398–401, 2006.

[44] E. Pennisi. Dna study forces rethink of what it means to be a gene. *Science*, 316(5831):1556–1557, 2007.

[45] J. Pinske. A consistent nonparametric test for serial independenc. *Journl of Econometrics*, 1998.

[46] E. Shannon. A mathematical theory of evidence: Bellsyt. *Techn. Journal*, 27:379–423, 1948.

[47] Z. Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967.

[48] H.J. Skaug and D. Tjostheim. Testing for serial independence using measures of distance between densities. *Lecture notes in statistics - New York - Springer Verlag-*, pages 363–377, 1996.

[49] D. Swarbreck, C. Wilks, P. Lamesch, T.Z. Berardini, M. Garcia-Hernandez, H. Foerster, D. Li, T. Meyer, R. Muller, L. Ploetz, et al. The arabidopsis information resource (tair): gene structure and function annotation. *Nucleic acids research*, 36(suppl 1):D1009–D1014, 2008.

[50] R.D.C. Team et al. R: A language and environment for statistical computing. *R Foundation Statistical Computing*, 2008.

[51] D. Tjostheim. Measues of dependence and tests for independence. *Statistics*, 1996.

[52] J.W. Tukey. The problem of multiple comparisons. *Mimeographed Notes.*

[53] J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, C.A. Evans, R.A. Holt, et al. The sequence of the human genome. *Science Signalling*, 291(5507):1304, 2001.

[54] P.H. Westfall, R.D. Tobias, and R.D. Wolfinger. *Multiple comparisons and multiple tests using SAS.* Sas Inst, 2011.

[55] P.H. Westfall and S.S. Young. *Resampling-based multiple testing: Examples and methods for p-value adjustment*, volume 279. Wiley-Interscience, 1993.

[56] E. Zeckendorf. Représentation des nombres naturels par une somme de nombres de fibonacci ou de nombres de lucas. *Bull. Soc. Roy. Sci. Liege*, 41:179–182, 1972.