ABSTRACT

DEMARCATION OF CODING AND NON-CODING REGIONS OF DNA USING LINEAR TRANSFORMS

by
Krithika Venkat

Deoxyribonucleic Acid (DNA) strand carries genetic information in the cell. A strand of DNA consists of nitrogenous molecules called nucleotides. Nucleotides triplets, or the codons, code for amino acids. There are two distinct regions in DNA, the gene and the intergenic DNA, or the junk DNA. Two regions can be distinguished in the gene- the exons, or the regions that code for amino acid, and the introns, or the regions that do not code for amino acid. The main aim of the thesis is to study signal processing techniques that help distinguish between the regions of the exons and the introns. Previous research has shown the fact that the exons can be considered as a sequence of signal and noise, whereas introns are noise-like sequences. Fourier Transform of an exonic sequence exhibits a peak at frequency sample value $k = N/3$ where N is the length of the FFT transform. This property is referred to as the period -3 property. Unlike exons, introns have a noise-like spectrum. The factor that determines the performance efficiency of a transform is the figure of merit, defined as the ratio of the peak value to the arithmetic mean of all the values. A comparative study was conducted for the application of the Discrete Fourier Transform and the Karhunen Loeve Transform. Though both DFT and KLT of an exon sequence produce a higher figure of merit than that for an intron sequence, it is interesting to note that the difference in the figure of merits of exons and introns was higher when the KLT was applied to the sequence than when the DFT was applied. The two transforms were also applied on entire sequences in a sliding window fashion. Finally, the two transforms were applied on a large number of sequences from a variety of organisms. A Neyman Pearson based detector was used to obtain receiver operating curves, i.e., probability of detection versus probability of false alarm. When a transform is applied as a sliding window, the values for exons and introns are taken separately. The exons and the introns served as the two hypotheses of the detector. The Neyman Pearson detector helped indicate the fact the KLT worked better on a variety of organisms than the DFT.

DEMARCATION OF CODING AND NON-CODING REGIONS OF DNA USING LINEAR TRANSFORMS

by
Krithika Venkat

A Thesis
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Electrical Engineering

Department of Electrical and Computer Engineering

January 2006

APPROVAL PAGE

DEMARCATION OF CODING AND NON-CODING REGIONS OF DNA USING LINEAR TRANSFORMS

Krithika Venkat

12/7/2005

Dr. Alexander M. Haimovich, Thesis Advisor
Professor of Electrical and Computer Engineering, NJIT
Date

11/29/05

Dr. Ali N. Akansu, Committee Member
Professor of Electrical and Computer Engineering, NJIT
Date

11/29/05

Dr. Yun - Qing Shi, Committee Member
Professor of Electrical and Computer Engineering, NJIT
Date

BIOGRAPHICAL SKETCH

Author: Krithika Venkat

Degree: Master of Science

Date: January, 2006

Graduate and Undergraduate Education:

- Master of Science in Electrical Engineering
  New Jersey Institute of Technology, Newark, NJ, 2006

- Bachelor of Science in Electrical and Communication Engineering
  Sri Siddhartha Institute of Technology, Tumkur, India, 2003

Major: Electrical Engineering

To You, Charles Dickens!!

Dear Charles,

Oh! The writer of the Victorian age!!!

Thy words of wisdom and knowledge have flowed into flawless works. Thy characters, so unique in themselves art so poignant. Attired in the cloak of misery and sorrow, the *Boy Who Wanted More* taught me more than any one could ever imagine.

Your works have shown me the path of selflessness and altruism. I shall, forever, work towards making a difference to at least one child, so that I have the satisfaction of being elevated to the position of an angel to that one innocent life.

To this, I shall always be indebted to you, Charles!!! I want to thank you for giving me an insight, a talisman, a goal, a purpose, an objective in life – to be a good person!!!

Thanks,
KV

ACKNOWLEDGMENT

A considerable quantity of my gratitude is tilted in favour of my Mom and Dad, Mrs. and Mr. K.G.Venkitachalam. Their love, trust and inspiring words during the painstaking time of my thesis work make me believe that I have the 'bestest' parents.

It is with deep pleasure that I thank my advisor Dr. Alexander M. Haimovich for his encouragement, support and priceless advice. He has put up with all my mistakes and my eccentric ideas and has made the different ramifications of Signal Processing so easy to me by the depth of knowledge he possesses of them. His intuition and keen discernment have always kindled my mind to enlightenment. His memorable advice, given on 21 January, 2005 at 3:15 PM, "On the road of life, you may trip and fall, just rise again", shall forever be etched in my memory.

I am indebted to my Thesis Defense Committee Members, Dr. Ali N. Akansu, and Dr. Yun - Qing Shi, for their time, patience and understanding. They adjusted their schedule to suit mine. I really appreciate their consideration.

Dr. Ramaswamy Ramakrishnan, the author of the paper I based my thesis work on, has put me on the right track with his inestimable suggestions and counsel.

I extend special gratitude to all my friends at the Center for Wireless Communication and Signal Processing Research Lab (CWCSPR lab) who have contributed directly or indirectly in shaping my thesis. Mr. Nikolaus Lehmann deserves a special acknowledgment for his helpful suggestions, his clear perception, his wonderful teaching caliber and his ability to devote time to an area least related to his area of research. I am grateful to Dr. Osvaldo Simeone for answering all my doubt calls and to Mr.Vlad Mihai Chiriac for his timely encouragement and wise council on efficient programming.

TABLE OF CONTENTS

Chapter **Page**

TABLE OF CONTENTS
(Continued)

LIST OF TABLES

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INTRODUCTION

The year 1953 recorded a revolution in the history of human discoveries, when J.D Watson and F.H.C Crick published a paper on the structure of DNA [1].The unsolved mystery of the hereditary units of life has now been decoded. But, though Watson and Crick detailed out the structure of DNA, they planted seed of extensive research and development for many more years to come. To determine the sequence of DNA of each organism as remained a challenging issue to all computation biologists. This area of research has unlimited vistas that has instigated many a keen mind to devote time and enterprise.

This thesis work is essentially the application of signal processing tools in the prediction of coding and non-coding regions of DNA But it is superfluous to talk about the application of signal processing tools on a few terms that not many are aware of. The first priority is to understand few details about DNA itself. This thesis later describes the applications of different linear transforms on the strand of DNA, comparison of the different techniques and go on to say why one method is better than the other.

In biology, genome of an organism is the whole hereditary information of an organism that is encoded in the DNA. This includes both the genes and the non-coding sequences. The term 'Genomics' [33] was introduced in 1986 by Thomas Roderick to describe the scientific discipline of mapping, sequencing and analyzing genomes. It is now important to understand the basic definition of DNA, since this is essentially what the whole thesis is all about. DNA, abbreviation for Deoxyribonucleic Acid , is a nucleic acid that contains the genetic instructions specifying the biological development of all cellular forms of life. (For the structure of nucleic acid, refer to Appendix A).

Watson and Crick [1] establish the structure of DNA as a very long chain, the backbone of which is made up of alternate sugar and phosphate groups, joined together in a regular 3'-5'phophate di-ester linkages. To each sugar is attached a nitrogenous base, four of which are very commonly found. They are classified in two – the purines and the pyrimidines. The two bases that come under the category of purines are adenine and guanine and the two pyrimidines are thymine and cytocine. Each strand has polarity, such that the 5'-hydroxyl group of the first nucleotide begins the strand and the 3'-hydroxyl group of the final nucleotide ends strand implying that this strand runs 5' to 3'. It is also essential to know that the two strands of DNA run antiparallel, such that one strand runs 5' -> 3' while the other one runs 3' -> 5'. At each nucleotide residue along the double-stranded DNA molecule, the nucleotides are complementary. That is, **A** forms two hydrogen-bonds with **T**; **C** forms three hydrogen bonds with **G**. This is to say that if the main DNA strand has a string of ATCGATCGATGC… the complementary strand of DNA will have TAGCTAGCTACG… In most cases the two-stranded, antiparallel, complementary DNA molecule folds to form a helical structure which resembles a spiral staircase. This is the reason why DNA has been referred to as the "Double Helix" [36]

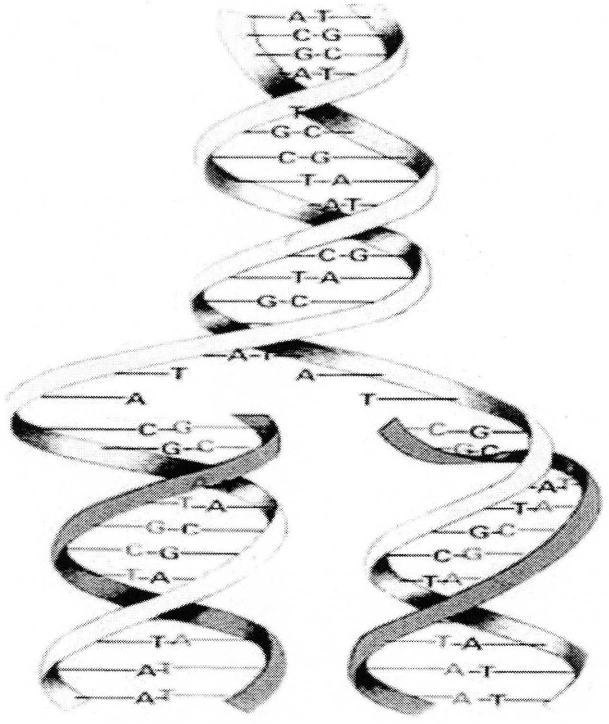


Figure 1.1 A strand of DNA.
Source [37]

As is mentioned in [2], a DNA sequence can be separated into two types of regions and intergenic spaces. Genes contain the information to code for proteins. Each gene is responsible for the production of a different protein. Each gene is further divided into two types of subregions- the exons and the introns. The central dogma of molecular biology refers to the creation of protein in this fashion:

gene in DNA → RNA → Protein

The gene is first copied into a single stranded chain called the messenger RNA or mRNA molecule. The introns are removed from the mRNA by the method of splicing. The spliced mRNA is then used by a large called the ribosome to produce the corresponding protein. The translation from mRNA to protein is aided by adapter molecules called the transfer RNA or tRNA molecules. When there is a conversion from the DNA to the RNA, the nucleotide thymine is replaced by uracil strand is AGCTGATGCTAAATG, the corresponding RNA strand will be AGCUGAUGCUAAAUG.

As seen earlier, the entire DNA strand is made of nucleotides. Now these nucleotides are grouped in trios to form a codon. Since there are four bases, the total number of possible combinations are $4^3 = 64$. Thus there are 64 different codons possible. These codons give rise to amino acid, which in turn produce proteins. There are totally 20 different amino acids. This gives an implicit meaning that the mapping between codons and the amino acids is a many to one mapping [2]. That is, there are many codons that must code for the same amino acid. These types of codons are called as synonymous codes [3]. Looking at Table 1.1 it can be observed that there are only two one codon – one amino acid (non degenerated) mappings for tryptophan and methionine, but ten double, three triple, six quadruple and two sextuple degeneracies [4].

When a gene is expressed, each codon in the mRNA produces an amino acid according to the genetic code, and the amino acids are bonded together into a chain. When all the codons in the mRNA are exhausted, a long chain of amino acids is obtained. This is the protein corresponding to the original gene. Referring to table 1.1, it can be seen that there is one start codon, corresponding to the nucleotide triplet ATG. This is the start. This signifies the beginning of the protein-coding part of the gene. If the start codon appears again, it produces the amino acid methionine. There are three stop codons, or the codon s that signify that the protein coding part of the gene has come to an end. They are TAA, TAG, and TGA. They do not code for any amino acid.

Table 1.1 List of Codons
Source [2]

AAA: K (Lys)	GAA: E (Glu)	TAA: STOP	CAA: Q (Gln)
AAG: K (Lys)	GAG: E (Glu)	TAG: STOP	CAG: Q (Gln)
AAT: N (Asn)	GAT: D (Asp)	TAT: Y (Tyr)	CAT: H (His)
AAC: N (Asn)	GAC: D (Asp)	TAC: Y (Tyr)	CAC: H (His)
AGA: R (Arg)	GGA: G (Gly)	TGA: STOP	CGA: R (Arg)
AGG: R (Arg)	GGG: G (Gly)	TGG: W (Trp)	CGG: R (Arg)
AGT: S (Ser)	GGT: G (Gly)	TGT: C (Cys)	CGT: R (Arg)
AGC: S (Ser)	GGC: G (Gly)	TGC: C (Cys)	CGC: R (Arg)
ATA: I (Ile)	GTA: V (Val)	TTA: L (Leu)	CTA: L (Leu)
ATG: M (Met)/START	GTG: V (Val)	TTG: L (Leu)	CTG: L (Leu)
ATT: I (Ile)	GTT: V (Val)	TTT: F (Phe)	CTT: L (Leu)
ATC: I (Ile)	GTC: V (Val)	TTC: F (Phe)	CTC: L (Leu)
ACA: T (Thr)	GCA: A (Ala)	TCA: S (Ser)	CCA: P (Pro)
ACG: T (Thr)	GCG: A (Ala)	TCG: S (Ser)	CCG: P (Pro)
ACT: T (Thr)	GCT: A (Ala)	TCT: S (Ser)	CCT: P (Pro)
ACC: T (Thr)	GCC: A (Ala)	TCC: S (Ser)	CCC: P (Pro)

It is clear by now that both the coding and the non-coding regions of DNA have the same set of codons in them. It is also obvious that the exons are, by and far, more important than the introns. But, as seen earlier, both the exons and introns together constitute the gene. It is implicit that each cannot exist without the other. In eukaryotes, the genes are made up of exons and introns, whereas the prokaryotic gene is made of only exons.

Over the years, many researchers have been successful in the prediction of exonic region in a gene.

CHAPTER 2

USE OF BIOINFORMATICS TOOLBOX

As is now understood, there are two different regions in a DNA gene – one the coding region, and the other the non-coding region. It is also obvious that the coding region is by far the most important the region in a DNA. But it does not mean that the non-coding region is not important. Since both introns and exons constitute the gene, the existence of one depends on the other. That is, the exons can be properly defined only when the reader knows the precise region as to where the exons begin from, or rather where the introns end.

Now given two sequences, it is usually easy to recognize if it is a coding region or not. It takes a simple count of the number of codons that are present in the sequence. Exons, as discussed earlier, will have a start and a stop codon But it so happens that, even the introns will be having the same codons too. So what is it that differentiates the exons from the introns? The answer is definitely not because certain codons are present in the exons that are not there in the introns. It is because of the way they are arranged.

Coming back to the codons present in the exons and the introns, exons will have a start codon ATG. If ATG occurs in the sequence more than as the start codon, it will produce the amino acid methionine [2]. Exons will also have one stop codon, either a TAA or TAG or TGA. None of them code for any amino acid. And, as the name suggests, the stop will occur at the end of the exon sequence. If it occurs in the middle, then it will result in a mutation. This thesis does not indulge itself in the mutated genes; it just tries to predict the coding region from the non coding region in a proper, unmutated gene. So in all the genes considered, there is a start codon and a stop codon. The difference between the codon count for exons and introns will basically lie in the fact that there will be only 1 stop codon in an exon, where as there will be many stop codons in an introns. So a simple count should be able to do the trick.

An organism was taken. The details of the gene can be found as follows:

Name of Organism:	Schizosaccharomyces Pombe
Total Length of sequence:	2784
Total Number of Amino Acids:	$\frac{2784}{3} = 928$
Coding Region:	From 891:1982
Length of Coding Region:	1092
Amino Acids in Coding Region:	$\frac{1092}{3} = 364$
Non-Coding Region:	From 1:890 and from 1983: 2784
Length of Non-Coding Region:	1692
Amino Acids in Non-Coding Region:	$\frac{1692}{3} = 564$

In the Bioinformatics Toolbox of MATLAB 7.0.1, each amino acid is depicted by an alphabet (Refer to Table 1.1). The output for the program to check the amino acids present in the DNA strand is given as follows:

Amino Acid Count in Coding Region:

A: 26	R: 23	N: 16	D: 18	C: 1	Q: 15	E: 32	G: 21
H: 12	I: 19	L: 32	K: 31	M: 7	F: 18	P: 15	S: 35
T: 11	W: 4	Y: 13	V: 14	Others: 1			

Amino Acid Count in Non-Coding Region:

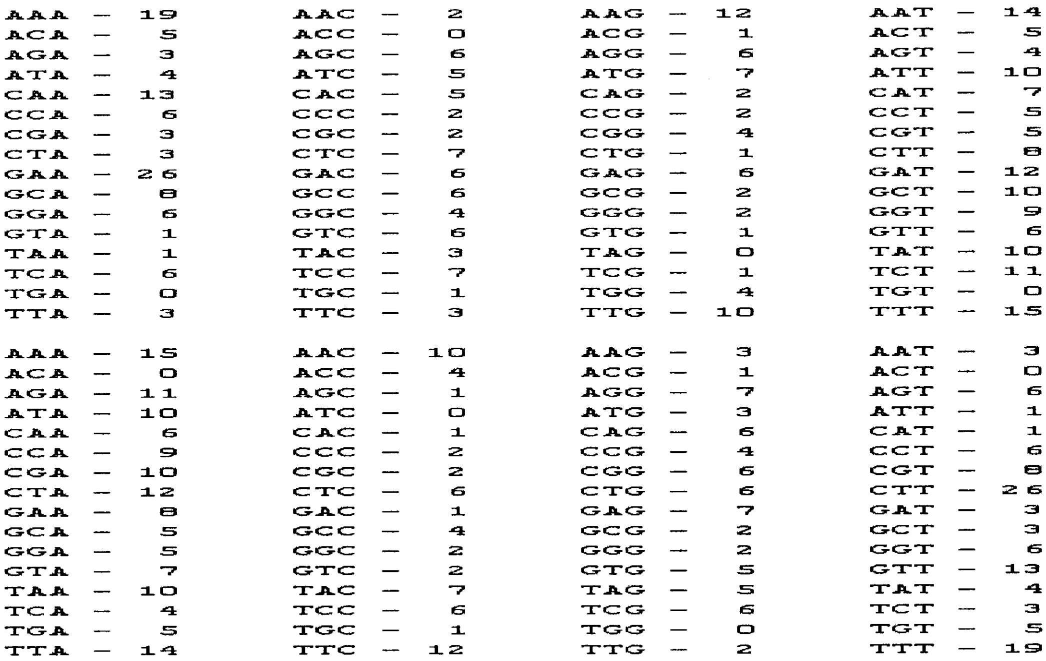
A: 9	R: 30	N: 21	D: 19	C: 20	Q: 19	E: 18	G: 25
H: 16	I: 45	L: 60	K: 34	M: 7	F: 68	P: 14	S: 47
T: 36	W: 4	Y: 18	V: 27	Others: 27			

Out of the above two lists, it is the list of M and others that are of significant importance. M is the start codon. The very fact that there are equal number of methionine producing codons (it is exclusive to this particular organism. This may or may not be the case with other organisms) proves that both the exons and the introns have the same codons and that the presence of start codon does not really signify anything. The thing of interest here is the number of the others. As we know, the stop codons do not code for any amino acid. So they are represented by an asterix* . If a sequence has just one stop codon, then it is an exon. Else, it is an intron.

The following is the heat diagram of the presence of codons in the main DNA strand and in the complementary DNA strand. Heat diagram is just another tool, thought out by the Bioinformatics toolbox. The heat diagram counts the number of codons present. The following is the heat diagram of the organism taken into consideration.

For the coding region,

Table 2.1 Codon Count for the Coding Region



```
AAA - 19     AAC -  2     AAG - 12     AAT - 14
ACA -  5     ACC -  0     ACG -  1     ACT -  5
AGA -  3     AGC -  6     AGG -  6     AGT -  4
ATA -  4     ATC -  5     ATG -  7     ATT - 10
CAA - 13     CAC -  5     CAG -  2     CAT -  7
CCA -  6     CCC -  2     CCG -  2     CCT -  5
CGA -  3     CGC -  2     CGG -  4     CGT -  5
CTA -  3     CTC -  7     CTG -  1     CTT -  8
GAA - 26     GAC -  6     GAG -  6     GAT - 12
GCA -  8     GCC -  6     GCG -  2     GCT - 10
GGA -  6     GGC -  4     GGG -  2     GGT -  9
GTA -  1     GTC -  6     GTG -  1     GTT -  6
TAA -  1     TAC -  3     TAG -  0     TAT - 10
TCA -  6     TCC -  7     TCG -  1     TCT - 11
TGA -  0     TGC -  1     TGG -  4     TGT -  0
TTA -  3     TTC -  3     TTG - 10     TTT - 15

AAA - 15     AAC - 10     AAG -  3     AAT -  3
ACA -  0     ACC -  4     ACG -  1     ACT -  0
AGA - 11     AGC -  1     AGG -  7     AGT -  6
ATA - 10     ATC -  0     ATG -  3     ATT -  1
CAA -  6     CAC -  1     CAG -  6     CAT -  1
CCA -  9     CCC -  2     CCG -  4     CCT -  6
CGA - 10     CGC -  2     CGG -  6     CGT -  8
CTA - 12     CTC -  6     CTG -  6     CTT - 26
GAA -  8     GAC -  1     GAG -  7     GAT -  3
GCA -  5     GCC -  4     GCG -  2     GCT -  3
GGA -  5     GGC -  2     GGG -  2     GGT -  6
GTA -  7     GTC -  2     GTG -  5     GTT - 13
TAA - 10     TAC -  7     TAG -  5     TAT -  4
TCA -  4     TCC -  6     TCG -  6     TCT -  3
TGA -  5     TGC -  1     TGG -  0     TGT -  5
TTA - 14     TTC - 12     TTG -  2     TTT - 19
```

The top set of 64 codes is the codon count in the main DNA strand. The latter half of the table has a set of codons that are present in the complementary strand, or in the reverse strand. Thus if there are 19 AAA's present in the main strand, one can observe 19 TTT present in the complementary strand, as A and T are complementary to each other. Another example would be the there are 13 CAA in the main strand and the complementary strand has 13 GTTs too.

The heat diagram of the same is given as follows.

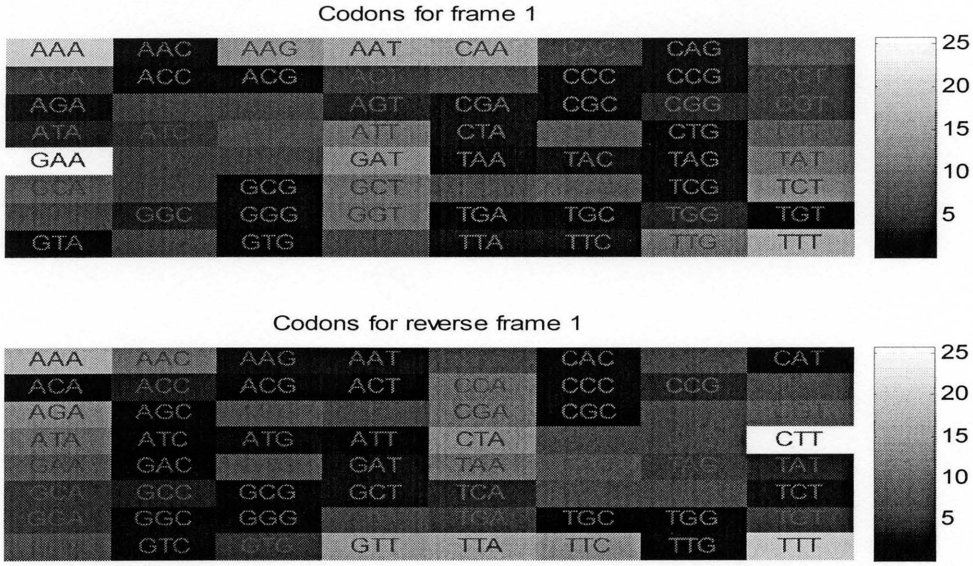


Figure 2.1 Heat diagram of a coding region.

Similarly, one can get out the codon count and the heat diagram of the introns for this organism.

Table 2.2 Codon Count for the Non-Coding Region

AAA	—	24	AAC	—	15	AAG	—	10	AAT	—	6
ACA	—	10	ACC	—	6	ACG	—	7	ACT	—	13
AGA	—	8	AGC	—	7	AGG	—	4	AGT	—	7
ATA	—	12	ATC	—	6	ATG	—	7	ATT	—	27
CAA	—	8	CAC	—	9	CAG	—	11	CAT	—	7
CCA	—	4	CCC	—	3	CCG	—	1	CCT	—	6
CGA	—	6	CGC	—	3	CGG	—	3	CGT	—	6
CTA	—	7	CTC	—	2	CTG	—	8	CTT	—	11
GAA	—	15	GAC	—	6	GAG	—	3	GAT	—	13
GCA	—	2	GCC	—	1	GCG	—	1	GCT	—	5
GGA	—	9	GGC	—	3	GGG	—	1	GGT	—	12
GTA	—	6	GTC	—	6	GTG	—	3	GTT	—	12
TAA	—	17	TAC	—	6	TAG	—	8	TAT	—	12
TCA	—	7	TCC	—	8	TCG	—	5	TCT	—	13
TGA	—	2	TGC	—	8	TGG	—	4	TGT	—	12
TTA	—	18	TTC	—	25	TTG	—	14	TTT	—	43
AAA	—	43	AAC	—	12	AAG	—	11	AAT	—	27
ACA	—	12	ACC	—	12	ACG	—	6	ACT	—	7
AGA	—	13	AGC	—	5	AGG	—	6	AGT	—	13
ATA	—	12	ATC	—	13	ATG	—	7	ATT	—	6
CAA	—	14	CAC	—	3	CAG	—	8	CAT	—	7
CCA	—	4	CCC	—	1	CCG	—	3	CCT	—	4
CGA	—	5	CGC	—	1	CGG	—	1	CGT	—	7
CTA	—	8	CTC	—	3	CTG	—	11	CTT	—	10
GAA	—	25	GAC	—	6	GAG	—	2	GAT	—	6
GCA	—	8	GCC	—	3	GCG	—	3	GCT	—	7
GGA	—	8	GGC	—	1	GGG	—	3	GGT	—	6
GTA	—	6	GTC	—	6	GTG	—	9	GTT	—	15
TAA	—	18	TAC	—	6	TAG	—	7	TAT	—	12
TCA	—	2	TCC	—	9	TCG	—	6	TCT	—	8
TGA	—	7	TGC	—	2	TGG	—	4	TGT	—	10
TTA	—	17	TTC	—	15	TTG	—	8	TTT	—	24

The heat diagram of the same is given as follows

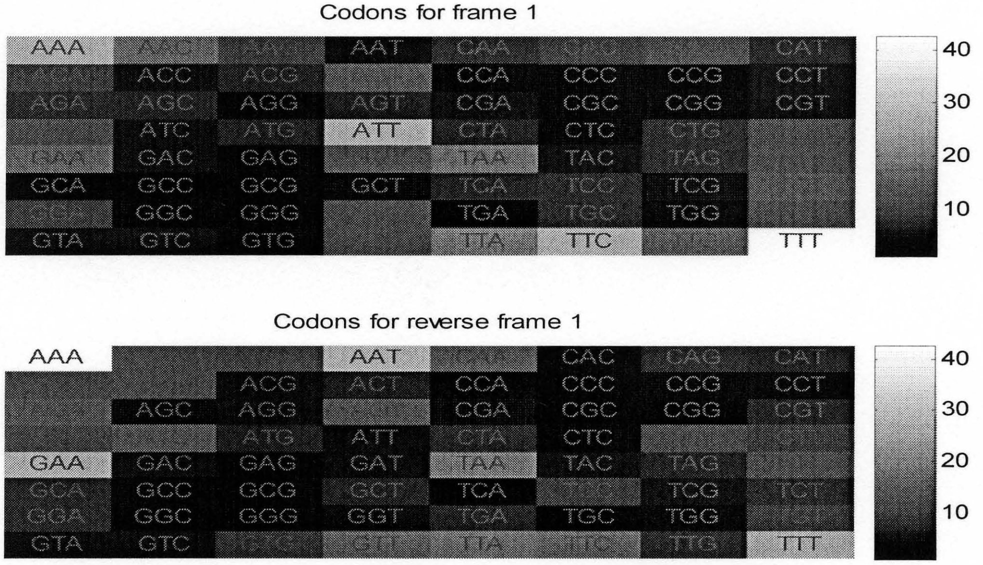


Figure 2.2 Heat diagram of a non-coding region.

Another was of checking, is the order of occurrence of the two codons. The start codon should obviously be before the stop codon, in an exon, whereas, there is no such rule in the intron.

Now the question arises as to why is it necessary to look into some other ways and means to predict the coding and the -coding regions of the DNA, when it is just enough to check out the start and stop codons. Well, two important reasons lie behind all this. The first reason is that, the initiation signals may not be present in a few organisms [15]. One example is the lysis gene of phage MS2, which is only translated upon read through of the stop codon of the previous gene, and the yeast mitochondrial introns which code for protein.

The second reason requires a peek into a genbank. The database used here is National Centre for Biotechnology Information (NCBI) [36]. Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information – all for the better understanding of molecular processes affecting, human health and disease.

Taking a look into the gene structure can see that NCBI is a well catalogued database. It includes the organism name, its hierarchy and much more background information. It goes on to detail the gene, and the coding region. The coding region is denoted by CDS. It tells accurately that the coding region exists from the sequence number say 1234 to sequence number 2345. This means that the region from 1234 to 2345 is a coding region and it is going to show certain properties that shall be discussed in the subsequent chapters. But, biology is not as simple as it seems to be. It is only in certain organisms that the coding region exists on the main DNA strand. There are three different ways in which one can find the coding region in a DNA strand to be present:

1. This is common. It states a direct reference to the DNA strand – like for example, when it states that the coding region exists from 1234 to 2345.

2. Another form is, when the coding region is not in the same place, whereas it is spread out. In such cases, the CDS is denoted by
   join (1..2330 , 2590..3400 , 3500..4950)
   This implies that there are three different sequences, ranging from 1 to 2330, from 2590 to 3400, from 3500 to 4950. The concatenation of all the three different sequences will form the coding region.

3. The next form is a complement. This means that the coding region is not in the main DNA strand, but in the complementary stand. If CDS is denoted by
   complement (2300 : 3400)
   This implies that we are to take the sequence from 2300 to 3400 in the main strand, reverse it, and find the complementary base pair. i.e., if the sequence is ATAGTAGCCAGTGCAGCGATGC, then the coding region would be GCATCGTGCACTGGCTACTAT

4. Another condition that can exist is the join complement. If the CDS is denoted as
   join(complement (1..2330 , 2590..3400 , 3500..4950)
   This implies that the complementary strands from 1 to 2330, from 2590 to 3400, and from 3500 to 4950 are to be taken. The concatenated sequence of the three is the coding region.

5. The next condition that can exist is the complement join. If the CDS is denoted as
   complement (join (1..2330 , 2590..3400 , 3500..4950)
   This implies that the coding regions from 1 to 2330, from 2590 to 3400 and from 3500 to 4950 are taken, concatenated, and the complementary strand of the resultant strand is the coding region.

6. This is by and far the most complex one. In certain cases, the segment of the coding region would be in this DNA strand, and another segment would be in another DNA strand. The coding region would be the concatenation of a section from this DNA and to another DNA.

In the thesis, only the first 5 conditions were taken into consideration. The 6th was left out do to computational issues.

Now that one knows how exactly the strands look like, it is necessary to go back to the discussion on how one cannot simply trace out for the presence of one stop codon in a sequence and jump to the 'Eureka' stage of having found a coding region. It is understood that DNA has a lots of parts that don't mean anything.[] Introns consist of large stretches of DNA whose biological functions are only beginning to be elucidated. All genes begin with exons, but most have a variable number of introns within them that alternate with the exons. Introns were discovered in 1977 as a result of observing that the mRNA used to code for proteins was almost always shorter than the DNA from which it has been transcribed. The mRNA was eventually found to be shorter because it lacked the non-coding regions sequences between the coding regions on the DNA. It was discovered that the introns were normally removed by splicing enzymes before the three different types of RNA (mRNA, rRNA, tRNA) can complete their functions. Because introns interrupt the nucleotide sequences, they were first called as the interrupted genes. The 'int' in introns refers to intervening because introns always exist between exons. Sequences that code for protein are called exons because they travel outside the nucleus to code for proteins, and thus are the DNA sequences that are expressed. It has also been found that more primitive and simpler the organism, fewer the introns. A new view of intron states that introns are a complex mix of different DNA, much of which are vital to the life of the cell. As their functions are being determined, the relationship of introns to cancer and their use as tumor markers is also being explored. Several functions for introns have already been identified, and evidence for a role for them is indicated by the finding that some intron alterations are directly related to the development of cancer.

CHAPTER 3

A FEW EXISTING METHODS FOR THE PREDICTION OF EXONS AND INTRONS IN A GENE

It now becomes necessary to determine the coding and the non-coding region in a gene. The following chapter looks into a few methods that have been previously implemented. Many papers were referred to before it was decided what method was to be implemented in the thesis. Inspirations were drawn from various research papers. This chapter later goes on to talk about one paper that has contributed a lot to many research papers, including this thesis.

[5] Enumerates a few of the many methods that are used to predict the intron-exon region in a DNA. The abstract of each are as enlisted:

1. Linear Discriminant Analysis and Quadratic Discriminant Analysis [6]

   This is a statistical pattern – recognition method that is used to categorize the samples into two classes. Once samples have been represented as points in space, linear discriminant analysis (LDA) finds an optimal plane surface that best separates points that belong to two classes. Quadratic discriminant analysis (QDA) finds an optimal curved quadratic surface instead. Both the methods seek to minimize some form of classification error.

2. Perceptron Method

   It is a machine learning algorithm for pattern recognition or classification. A perceptron method is based in a simple neural network that begins with an arbitrary initial place and then iteratively moves the plane in a way that tries to reduce the classification error at each step.

3. Hidden Markov Models

   Hidden Markov Models (HMMs) represent a system as a set of discrete states and as transitions between those states, each if the possible transitions having an associated probability. Markov models are 'hidden' when one or more of the states cannot be observed directly. HMMs are valuable in bioinformatics because they allow a search or alignment algorithm to be built on firm probability bases, and it is straightforward to train the parameters (transition probabilities) with known data.

4. Hexamer – coding measures [7]

   Some methods interpret sequences as successions of words (so-called because nucleotides are not independent of each other, but tend to occur together as if in a word) of length k (k- tuples); 6-tuples are called hexamers. In-frame hexamer frequencies in a region of DNA have traditionally been used as a powerful way of discrimination coding regions from non-coding regions, as some words are more likely to be present in DNA.

5. Weight Matrix Method and Weight Array Method [8]

   This is used for scoring a signal motif site. In the weight matrix method (WMM), a score *s(x,b)* is assigned to each position c for each base pair b, such that the total score of a motif site can be calculated as the sum of scores at all positions in the site. In the weight array method (WAM), a score *s(x,w)* is assigned to each position *x* for each word *w* of length *k.* (when k = 1, the two methods are the same).

6. Maximal – Dependence Decomposition (MDD) donor matrices [8]
   It is a set of donor splice-site weight matrices that are generated using the WMM, each of which is built for a different class of splicing donor sites in such a way that the dependence between nucleotide positions is determined.

7. Decision Tree [9]

   This is a classification scheme, which can be used, for example, to split a sample into two subsamples according to some rule (feature variable threshold). Each subspace can be further split, and so on.

8. Artificial Neural Networks [10]

   The key element of the artificial neural network (ANN) model is the novel structure of the information processing system. It is composed of many highly interconnected processing elements that are analogous to neurons and are tied together with weighted connections that are analogous to synapses. Once it is trained to known exon or introns sample sequences, it will be able to predict exons or introns in a query sequence automatically.

9. Genetic signal analysis [4]

   The tetrahedral representation of the genetic code adequately grasps its basic features and degeneracy. Optimal symbolic-to-digital mappings of the genetic information contained in nucleic acid molecules, as well as in the primary structure of the corresponding proteins are derived at nucleotide, codon and amino acid levels. Nitrogenous basis and/or amino acid sequences are converted into genetic signals so that a large variety of signal processing methods can be used for their analysis. The use of Independent Component Analysis (ICA) to search for control sequences in the intergenic DNA, i.e., the part of the genome that does not encode for proteins, is proposed.

10. Statistical correlation of nucleotides in a DNA sequence [11]

    This paper demonstrates two basic properties of the correlation through statistical analysis, namely, the short – range dominance of nucleotide correlation in most DNA sequences and the coarse – grained evolutionary dependence of the short – range correlation in coding sequences. A corresponding evolutionary mechanism is suggested. By the use of spectral analysis a large inhomogeneity in long-range correlations for different sequences is indicated. Some results on three dimensional DNA walks are reported. The linguistic differences between coding and noncoding sequences are also indicated.

11. DFT based DNA splicing algorithms for prediction of protein - coding regions [12]
    This involves the use of the DFT based approach that explots the empirical observation that the spectrum of a protein coding region DNA of length N nucleotides has a peak at frequency k = N/3 corresponding to the length of a DNA codon.

12. Detect redundant coding structure [13]

    The coined term genetic code maps nucleotide triplets to amino acids. This is in the computer coding sense because a codon instruction is performed to an output of an amino acid sequence. Here, methods have been formed to detect redundant coding structure in DNA. First, a finite field framework for a nucleotide symbolic sequence is presented. Then approaches to find sequence structure associated with error detecting codes are examined. Subspace partitioning algorithm is a general approach to finding any linear coding redundancy.

13. Autoregressive Modeling and Feature Analysis of DNA sequences [21]

    A parametric signal processing approach for DNA sequence analysis based in autoregressive modeling is presented. AR model residual errors and AR model parameters are used as features. The AR residual error analysis indicates a high specificity of coding DNA sequences, while AR feature based analysis helps distinguish between coding and non-coding DNA sequences. An AE model – based string searching algorithm is also proposed. The effect of several types of numerical mapping rules in the proposed method is demonstrated.

14. Species independence of mutual information in coding and non-coding DNA [22]

    There could exist certain universal statistical patterns that are different in coding and non-coding DNA and can be found in all living organisms, regardless of their phylogenetic origin. Mutual information function I has significantly different functional form in coding and non-coding DNA. The probability distributions of the average mutual information are also significantly different for the two regions. But both these parameters are almost the same for organisms of all taxonomic classes. Mutual information is also capable of predicting coding regions as accurately as organism –specific coding measures.

15. Spectral analysis of DNA sequences [23]

    The analysis of DNA sequences through spectral based methods is reviewed. The issues include the nature of biomolecular sequences, the representation of the DNA as numerical sequences amenable to digital signal processing tools, and the spectral methods for gene finding and DNA feature extraction.

16. Gene and exon prediction using an all-pass based filter [16]

    This paper introduces a simple and efficient scheme for identifying the coding regions of DNA sequences based on anti-notch IIR filters. These filtered can be implemented very efficiently using the one- multiplier Gray and Markel lattice structure.

The above were just a few methods in which the exonic and intronic regions can be predicted. The thesis was originally based on the paper [14] this paper talks about the prediction of DNA coding and non-coding region with the help of Fourier Transform Techniques. Since the thesis gained inspiration from this paper, the description of this paper shall have a greater predominance over that of others.

3.1 A Brief Summary of the Paper: Prediction of Probable genes by Fourier Transform Analysis of Genomic Sequences [14]

The major signal in coding regions of genomic sequences is a three-base periodicity. This paper uses Fourier Transform to analyze this periodicity. The three – based periodicity in the nucleotide arrangement is evidenced as a sharp peak at the frequency f = N/3 in the Fourier domain. It is found that the relative height of the peak at f = N/3 in the Fourier spectrum if a good discriminator of coding potential Local Signal to Noise ratio is examined within a sliding window.

Basically the way to distinguish the region of coding and non-coding regions of DNA is based on a variety of contrasting characteristics of protein-coding sequences and DNA sequences that do not encode for protein. Fourier Transform technique is based in the existence of short range correlation in the nucleotide arrangement. The most prominent of these is the 1/3 periodicity which has been shown to be present in coding sequences. The method applied here does not involve the *a priori* knowledge of the signal.

The origin of the period-3 signal in protein-coding sequences derives from the triplet nature of the codon. There are two specific biases that are spoken about here:

1. codon bias
2. triplet bias

These terms can be better explained by the look at table 1.1. As said earlier, since there are 64 codons and 20 different amino acids, the mapping between codons and amino acid is many-to-one. If there exist two or more codons that can code for the same amino acid, a no bias in codons would mean that these can all be used interchangeably and will occur with the same frequency. There is a preferential usage of one or more codons, depending on the organism and the gene, in a way that is not totally clear, except that it is non-random. This occurs due to the unequal usage of codons corresponding to a given amino acid. To call something a codon, it has to be given a frame. Triplets, on the other hand, are just groups of three nucleotides. A triplet non-bias would mean that each triplet would occur with equal probability of $\frac{1}{64}$. And such a condition does not exactly occur. So this is the triplet bias. The triplet bias comes from the unusual usage of amino acids in naturally occurring proteins and it is universally present. Experimental results show that though the two compositional biases, especially the codon – bias do play an important role in generating the peak at N/3 of the frequency, it is not the primary reason for the periodicity. Experiments were conducted to prove the fact that when the codon bias was removed (by using all codons corresponding to a given amino acid with equal frequency) the resulting genomic sequences continue to show a sharp peak at f = N/3. The peak remains prominent even when the nucleotide was changed completely by assigning arbitrary codons to a given amino acid. These experiments suggest that the periodicity may be a consequence of the amino acid sequence in naturally occurring proteins which manifests itself as a bias in the use if certain triplets in the coding regions of genomic DNA.

The idea conceived is very simple and yet is very unique. As there are four bases, four indicator sequences were taken. An indicator sequence, as was observed in this paper, is a string of binary numbers, where 1 indicates the presence of the base and 0 indicates the absence. This is also called a projection vector.

Thus for a given strand of DNA, the indicator vectors would be as follows

(3.1)

Given Strand Of DNA		A	C	G	T	A	C	G	T	A	C	G	T
Indicator Sequence of A	X_A	1	0	0	0	1	0	0	0	1	0	0	0
Indicator Sequence of C	X_C	0	1	0	0	0	1	0	0	0	1	0	0
Indicator Sequence of G	X_G	0	0	1	0	0	0	1	0	0	0	1	0
Indicator Sequence of T	X_T	0	0	0	1	0	0	0	1	0	0	0	1

The sum of the individual spectra is taken

$$S(f) = \sum_{\alpha} S_{\alpha}(f) = \sum_{\alpha} \frac{1}{N^2} | \sum_{j=1}^{N} X_{\alpha}(x_j) \exp(2\pi i f j)|^2 \qquad (3.2)$$

where N is the length of the sequence

x_j is one of the four nucleotides and denotes the occurrence of that particular nucleotide in position j

$f = \frac{k}{N}$ is the discrete frequency with k = 1,2….N/2

α is one of the four symbols, A,C, G, T

Average is calculated by using the formula

$$\bar{S} = \frac{2}{N} \sum_{k=1}^{N/2} S(k/N) = \frac{1}{N}\left(1 + \frac{1}{N} - \sum_{\alpha} \rho_{\alpha}^{2}\right) \qquad (3.3)$$

where ρ_{α} denotes the frequency of occurrence of each base

For protein coding region, the max is a distinct peak at f = N/3. This implies that the signal to noise ratio is maximum at f = N/3

$$P = \frac{S(\frac{N}{3})}{\bar{S}} \tag{3.4}$$

A survey on large number of organisms shows that the value of P = 4 is a good discriminating factor, i.e. when the value of P > 4, then the region is an exon, whereas if the value of P < 4, then the sequence is intronic.

The same method can be applied on the sequence as a sliding window. The window size used is equal to 351 and the hop size used is either 1 or 3. The window size is so used because there are very few introns-containing genes in these sequences, and the open reading frames (ORFs) of length less than 300 bp are not frequently encountered. A window length n the range 250 – 400 gives similar results. Windows of length less than 250 have increased noise and poor statistics, while those greater than 400 tend to miss the ORFs due to numerous overlaps.

This is approximately what the paper has explained about finding the coding and non-coding region in a DNA strand by the reference paper [14].

This thesis works on similar levels to obtain similar results. Of course, a few modifications are made. And they are enumerated as follows:-

1. If the indicator sequence is taken as binary, then there will be a DC component in the Fourier Spectrum. This means that the first value is gone. Of course, it can be made equal to zero, but this will result in the shifting of the data by one. Thus the peak will not be at f = 1/3 but it will be shifted to the right by one. Instead of having confusion with the DC value, it is easier to choose the indicator sequence in such a way that the DC component is eliminated. One such easy way is to have the indicator sequences as this

Given DNA Strand		A	C	G	T	A	C	G	T	A	C	G	T
Indicator Sequence of A	X_A	$+\frac{3}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$+\frac{3}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$+\frac{3}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$
Indicator Sequence of C	X_C	$-\frac{1}{4}$	$+\frac{3}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$+\frac{3}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$+\frac{3}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$
Indicator Sequence of G	X_G	$-\frac{1}{4}$	$-\frac{1}{4}$	$+\frac{3}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$+\frac{3}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$+\frac{3}{4}$	$-\frac{1}{4}$
Indicator Sequence of T	X_T	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$+\frac{3}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$+\frac{3}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$+\frac{3}{4}$

(3.5)

i.e. a $\frac{3}{4}$ to indicate the presence of a base and $-\frac{1}{4}$ to indicate the absence. The total addition of this is going to be equal to zero. Thus the DC component is totally eliminated.

2. The paper [14] has assumed that the maximum of the spectrum in the window length will be only at f = 1/3. Strictly speaking, this should be the case. But it is better to assume that the peak lies somewhere else. It is now known that the peak of the whole spectrum should be at f = 1/3. All other values are really small when compared to it. This peak should be observed for each window length. Thus, assuming that the max value is at f = 1/3, it is better to put the SNR as the ratio of maximum of all values and the mean of all values. This thesis now defines the Signal to Noise or the Figure of Merit as

$$\text{Signal to Noise Ratio} = \frac{\text{Maximum Value}}{\text{Arithmetic Mean of all values}}$$

3. The paper goes in for a normalization of the frequency. The thesis work does not do that, as it is essential to see the value of the frequency at N/3

4. Since Fourier Transform is symmetrical, only half the spectrum is taken into consideration. But this thesis deals not only with Fourier, but with a few other transforms too, which are not symmetric. The thesis considers the entire spectrum and not just half.

It was decided to experimentally prove the codon bias and the triplet bias. For this a sequence of DNA was taken into consideration. The first column of Table 1.1 was taken. The column consists of all the codons that start with A. These codons were put as a sequence and made to repeat. The end sequence looked something like: AAAAAGAACAAT..ACCAAAAAG...ACCAAAAAG... repeating itself a 650 times (say).

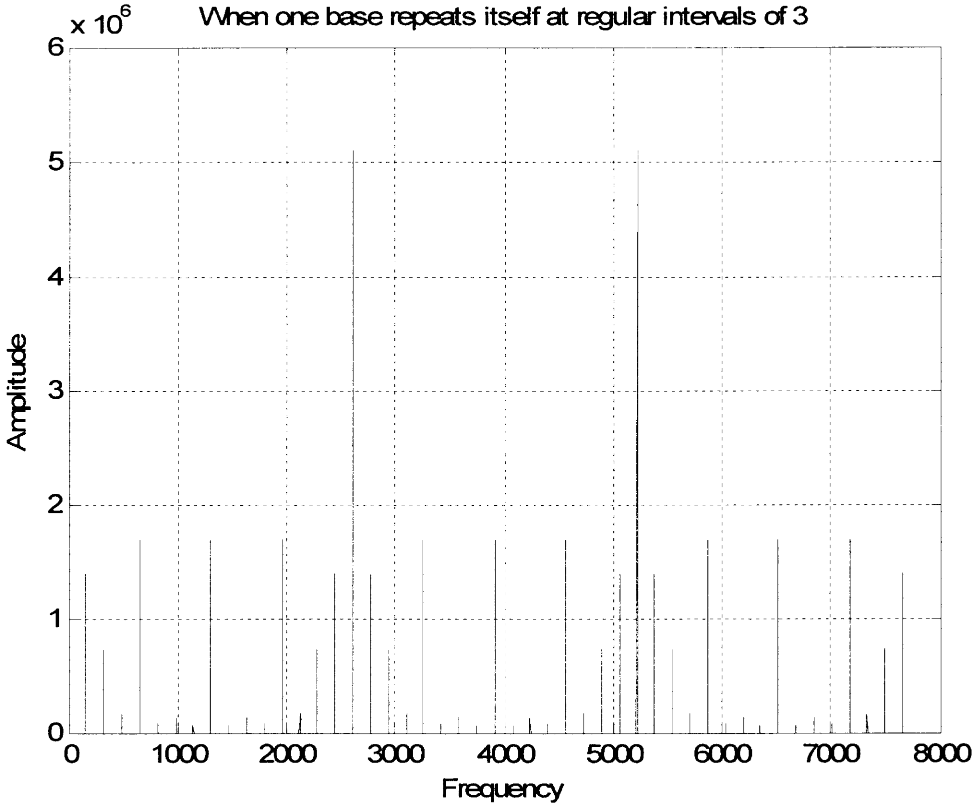


Figure 3.1 DFT for repetitive base sequence.

The Figure of Merit, defined by the ration of the maximum and the mean of all the values is calculated to be 978.000.

Now an artificial codon bias is inserted. This codon bias is the repetition of the same base sequence as above. But it is observed that there are a few codons that are synonymous, or that they code for the same amino acid. As per the definition of codon bias, the DFT of this sequence will still produce a high at N/3 and with symmetry, at 2N/3. The figure of merit is 476.5751. This proves the part that codon bias produces a peak at N/3, but with a lower peak.

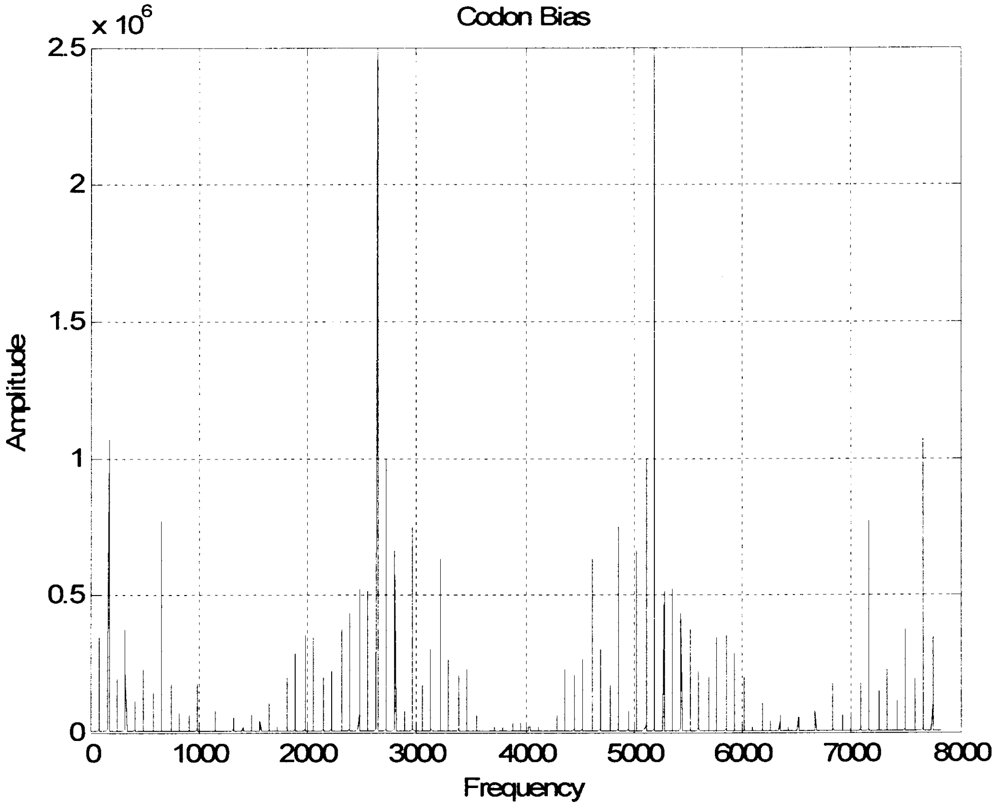


Figure 3.2 Codon bias.

CHAPTER 4

APPLICATION OF DISCRETE FOURIER TRANSFORM AND KARHUNEN LOEVE TRANSFORM ON SYNTHETIC GENOMIC SEQUENCE

The thesis mentions the fact that the codons exist in both the exons and the introns. The codon and the triplet bias help in the existence of certain properties in the exons that are not found in the introns. The exons or the coding regions of a DNA gene exhibit the *period – 3* property. Past research works show that when DFT is applied to the coding sequence, there is a peak at sample frequency f = N/3 whereas, this is not exhibited by the intronic region.

It is indeed a good idea to check out how exactly the two transforms discussed in the thesis – DFT and KLT work on the synthetic genetic sequence.

The question now arises as to what is the meaning of a synthetic exon sequence and a synthetic intron sequence. The two concepts shall be explained as follows:

4.1 Synthetic Exon Sequence and the Application of Transforms

As is seen in the clearly previous results, the spectrum of the exonic region indicates a peak at the frequency of N/3. Since Fourier Transform is symmetric, it follows that the peak should be at f = 2N/3 also. Thus the spectrum of an exon will have two peaks, at N/3 and 2N/3 of the frequency. [24] talks about the fact that the exon should typically behave as if it were a sequence in which one base repeats itself at the interval of 3. This would also explain the period-3 property. But the strength of the peak will depend on the organism. [2]. For a while, if the concentration is devoted to the fact to determine a sequence in which the number repeats itself at regular intervals of 3, it would mostly be a sequence of binary string of 1 and 0 arranged as 1 0 0 1 0 0 1 0 0 1 0 0 .. This is also in accordance to the projection vector or the indicator sequence that the reference paper [14] talks about. But this sequence has a DC component. The approach now would be to remove the DC component from this binary string of numbers, such that the DC component is zero. Such a string would be a series of $\frac{3}{4}\ -\frac{1}{4}\ -\frac{1}{4}\ -\frac{1}{4}\ \frac{3}{4}\ -\frac{1}{4}\ -\frac{1}{4}\ -\frac{1}{4}\ \frac{3}{4}\ -\frac{1}{4}\ -\frac{1}{4}\ -\frac{1}{4}\ \frac{3}{4}\ -\frac{1}{4}\ \ldots$

4.1.1 Application of DFT on Synthetic Exon Sequence

DFT is applied to the signal. The plot is as shown below.

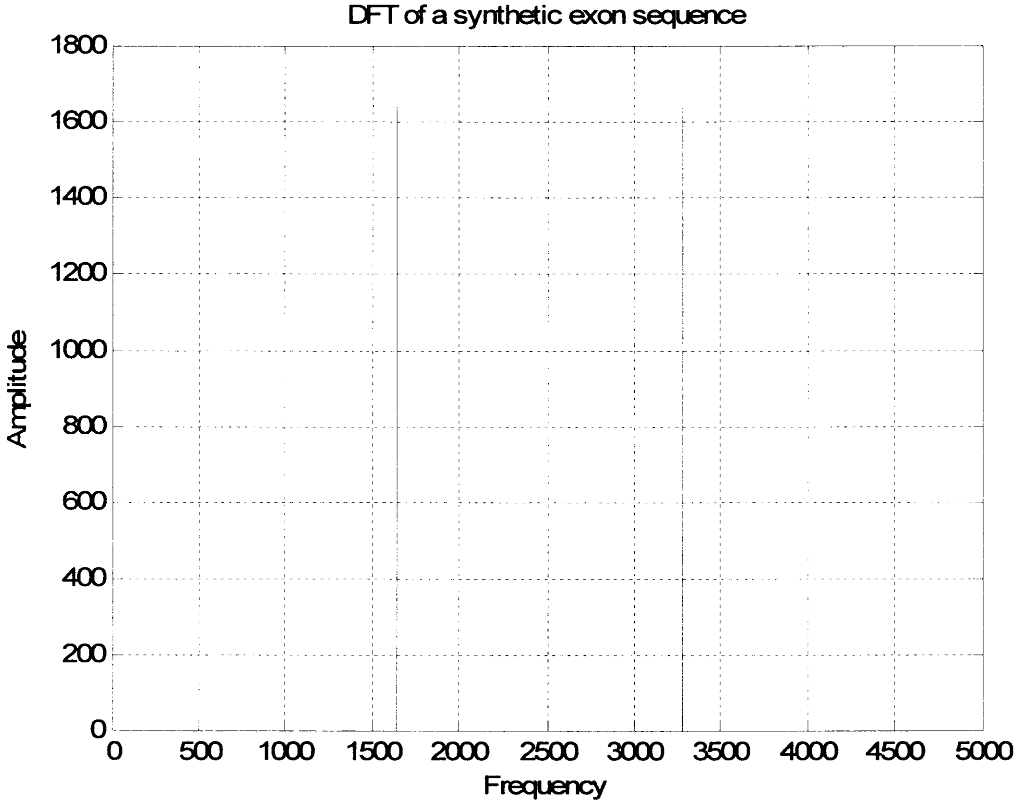


Figure 4.1 DFT applied on synthetic exon sequence.

It can be clearly observed that there are two distinctive peaks at sample frequencies f = N/3 and 2N/3. All other values go to a zero, as this is a noiseless deterministic sequence.

4.1.2 Application of KLT on Synthetic Exon Sequence

KLT is the eigen decomposition of the autocorrelation or the covariance matrix of the sequence. If N is the length of the sequence, a $\frac{N}{2}x\frac{N}{2}$ Toeplitz matrix is formed. The eigen of this matrix is plotted out.

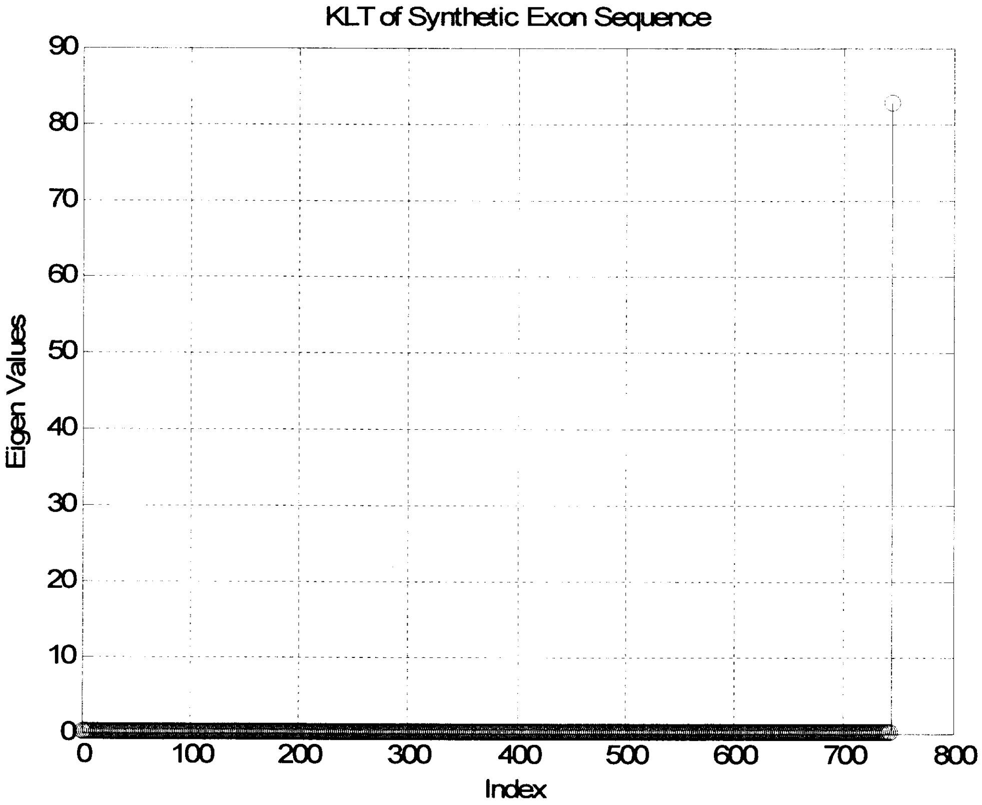


Figure 4.2 KLT applied on synthetic exon sequence.

One distinctive peak indicates the presence of a large signal in the first half of the spectrum. This is similar to what the Fourier Transform has to say about the exon region

4.2 Synthetic Intron Sequence and the Application of the Transforms

Now that the exons are dealt with, it is now necessary that the introns are looked into. The introns are basically described as a noisy sequence. The synthetic sequence that is required to be found here is a random sequence that has a probability of occurrence equal to ¾.

It is easy to compute a random binary string that fill in this requirement. But, as it is necessary to cancel out the DC component, the string is now modified into a string of ¾ and -¼ , such that the overall mean is equal to ¼. The probability of a quarter is because there are totally four different bases and all four of them can occur with an equal probability of ¼.

4.2.1. Application of DFT on a Synthetic Intron Sequence

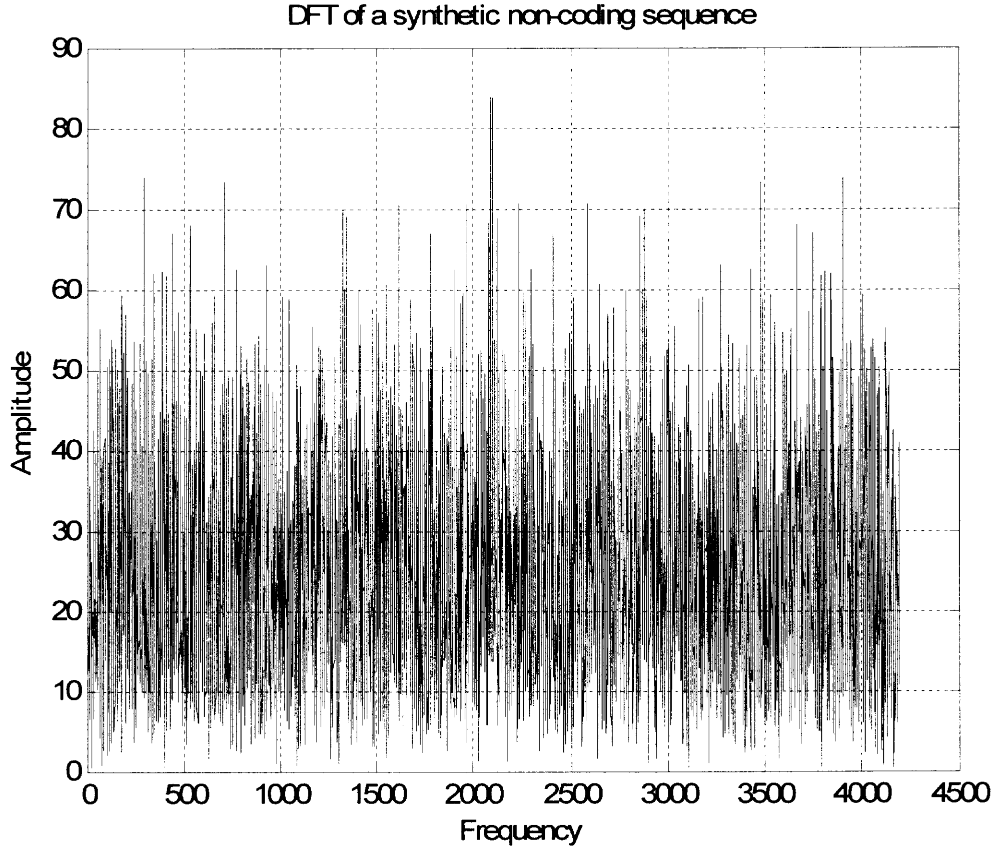


Figure 4.3 DFT applied on a synthetic intron sequence.

The DFT applied on the synthetic intron sequence does not produce any peak.

4.2.2 KLT Applied on Synthetic Intron Sequence

As is the case, the autocorrelation function matrix is determined and the plot of the eigen of the Toeplitz matrix is as shown below.

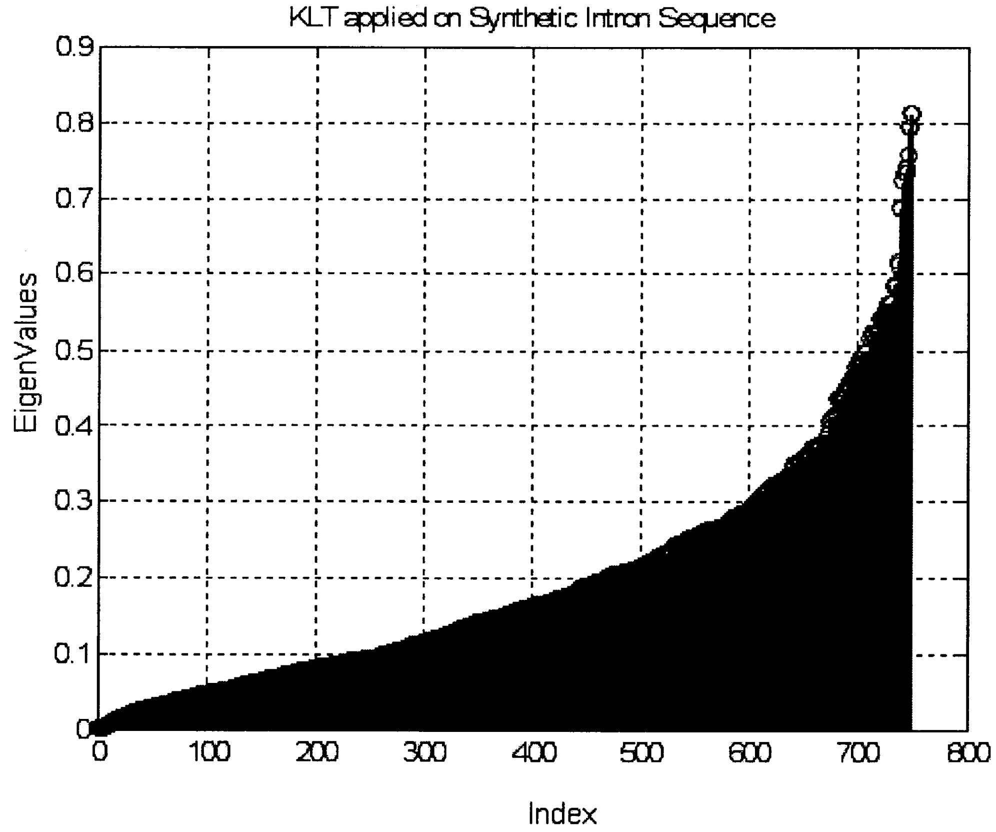


Figure 4.4 KLT applied on a synthetic intron sequence.

There are plenty of eigen values. The comparison of the two KLT plots shows the difference in the spectrum.

CHAPTER 5

APPLICATION OF TRANSFORMS ON GENOMES SEQUENCES

5.1 Logic of Application of Discrete Fourier Transform

As is done in many papers, Fourier Transform is used to distinguish a coding region from a non-coding region. The transform is applied to the DNA strand. Fourier Transform converts a signal from the time domain to the frequency domain. After the application of the transform, the power spectral density of the DNA strand is obtained. Coding and the non-coding regions are both made of the same codons. But their functions are different. The logic behind the application of Fourier Transform is to distinguish if the transform can bring out a difference between the two coding regions when transformed to another domain.

The indicator sequence of the entire DNA strand is taken. That is a vector that returns a ¾ for the presence of a nucleotide and a -1/4 for the absence of the nucleotide. The DFT of all the sequences is found. The individual DFTs are squared and added. The sum is plotted. It is observed that there is a peak at f = 1/3. This is in accordance to what is written in paper [14]. Similarly, if a non-coding region is taken, no peak can be observed.

5.1.1 Algorithm

- Consider a DNA sequence ACGTACGTACGTACGTACGT

- Form indicator sequences – They are strings of numbers that indicate the presence of the base. That is, an indicator sequence of A will have 3/4 to indicate the presence of A and -1/4 in the absence. This will eliminate the DC component in the spectral domain.

- The stand and the **indicator sequences** are as given –

Given DNA Strand		A	C	G	T	A	C	G	T	A	C	G	T
Indicator Sequence of A	X_A	$+\frac{3}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$+\frac{3}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$+\frac{3}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$
Indicator Sequence of C	X_C	$-\frac{1}{4}$	$+\frac{3}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$+\frac{3}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$+\frac{3}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$
Indicator Sequence of G	X_G	$-\frac{1}{4}$	$-\frac{1}{4}$	$+\frac{3}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$+\frac{3}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$+\frac{3}{4}$	$-\frac{1}{4}$
Indicator Sequence of T	X_T	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$+\frac{3}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$+\frac{3}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$-\frac{1}{4}$	$+\frac{3}{4}$

(5.1)

- Apply **DFT** to the indicator sequences, by using the formula

$$X_Y[k] = \sum_{n=0}^{N-1} x_Y(n) e^{-j2\pi kn/N} \tag{5.2}$$

where $0 \leq k \leq N\text{-}1$

$Y = A, C, G, T$

- To Obtain the **Power Spectral Density**

$$S[k] \triangleq |X_A[k]|^2 + |X_C[k]|^2 + |X_G[k]|^2 + |X_T[k]|^2 \tag{5.3}$$

- **Plot** the Sum and Observe

5.1.2 DFT applied on a Coding Sequence

5.1.2.1 Details of the Sequence.

Organism:	Schizosaccharomyces Pombe
Length of Sequence:	2784
Coding Region:	891:1982
Length of Coding Sequence:	1092

5.1.2.2 Spectrum of the Coding Sequence.

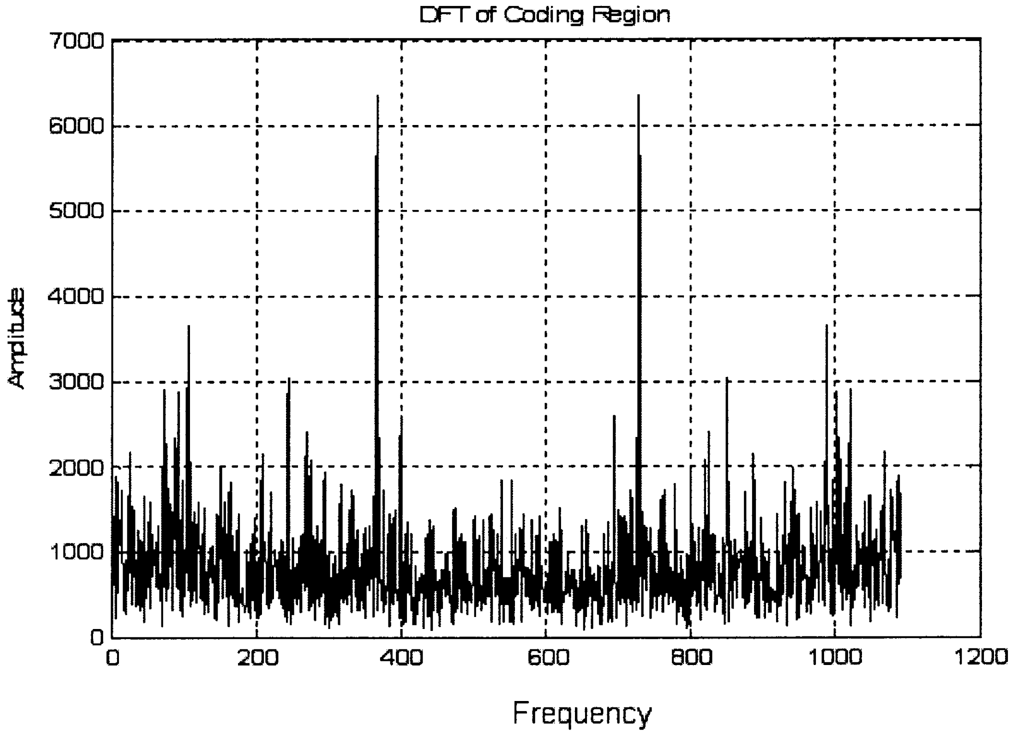


Figure 5.1 DFT of a coding sequence.

5.1.2.3 Observation.

- There are two distinctive peaks can be observed at frequency = 364 and 728
- Figure of Merit $= \dfrac{\text{Peak Value}}{\text{Mean of all Values}} = \dfrac{6350}{810.4157} = 7.8355$

$$\text{Figure of Merit For Coding Region when DFT if applied} = 7.8355 \qquad (5.4)$$

5.1.2.4 Inference.

- The spectrum shows a peak at 364 which is equal to one third of the entire sequence.
- The peak at 364 corresponds to the fact that Fourier Transform is symmetric. So if there is a peak at one third the frequency, there should be another at two third.

Power spectral density of the coding region of DNA shows a peak at sample value k = N/3 where N is the length of the sequence. It is sufficient to calculate the DFT at sample value k = N/3. [2]

It must be noted that that though the coding regions exhibit this peak at f = 1/3, the strength of the signal varies from one organism to another.

5.1.3 DFT applied on a Non-Coding Region

The previous section dedicated itself to the application of DFT on an entire coding region. The peak at sample frequency f= 1/3 could be observed. To have a similar basis for comparison between the properties of two sequences, it is but logical that the same transform is applied on the non-coding region too.

From the paper [14], it is predicted that the application of Fourier Transform on both the sequences will result in the distinctive difference. Previous papers have proved the fact that the coding region has signal and noise, whereas the non-coding region does not have any signal. It is purely noise. Thus the plot to be expected is the plot of DFT applied on random noise, which is once again going to be random.

5.1.3.1 Details of the sequence.

Organism:	Schizosaccharomyces Pombe
Length of Sequence:	2784
Non -Coding Region:	concatenation of 1:890 and 1983: 2784
Length of Non - Coding Sequence:	1692

5.1.3.2 Spectrum of the Non-Coding Sequence.

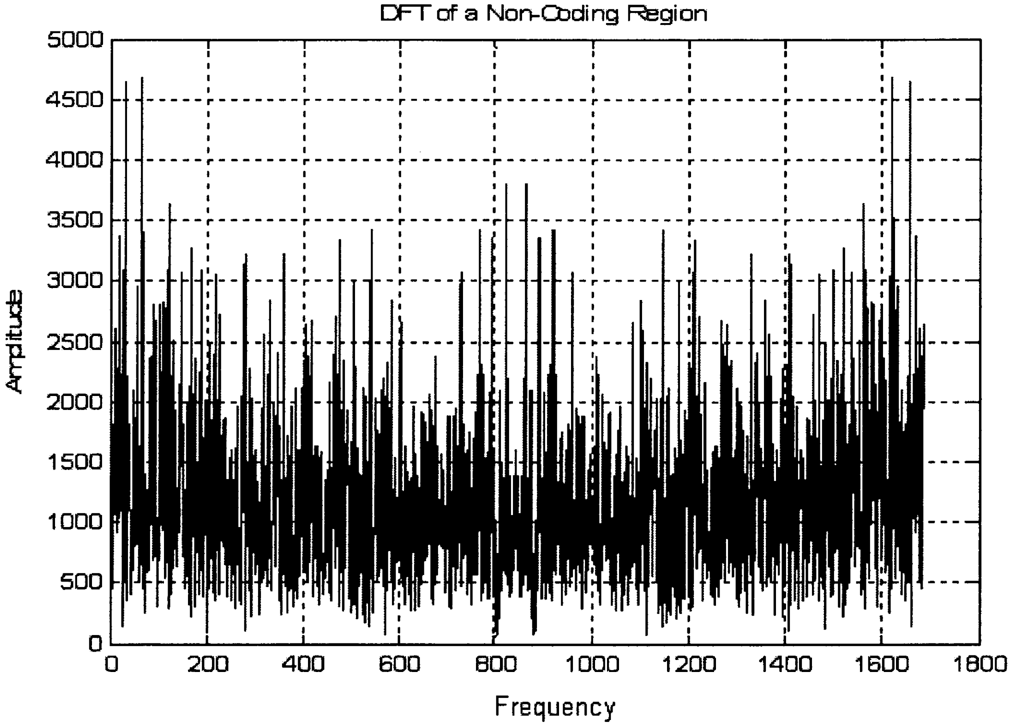


Figure 5.2 DFT of a Non-Coding Sequence

5.1.3.3 Observation.

- This plot is the spectrum of random noise. Thus there are no observable peaks
- Figure of merit $= \dfrac{\text{Maximum Value}}{\text{Mean of all Values}} = \dfrac{5.0205\text{e}+003}{1.1987\text{e}+003} = 4.1881$

$$\text{Figure of Merit for a Non-coding Region when DFT is applied} = 4.1881 \qquad (5.5)$$

5.1.3.4 Inference.

- Non-coding region does not give any peaks in its spectrum.
- The difference in the Figure of Merits of the coding and non-coding spectra , when DFT is applied to the signal is = 7.8355 - 4.1881 = 3.6474

Difference in the Figure of Merits in Coding and Non-Coding (5.6)
Region When DFT if applied = 3.6474

5.1.3 Conclusion

- The basic difference between the coding region and the non-coding region, in the frequency domain, is that the coding region has a peak at its 1/3 whereas the non-coding region does not have a peak.
- This property is referred to as the *period-3 property* exhibited exclusively by the coding region.

5.2 Logic and Purpose of Application of Karhunen-Loeve Transform

Karhunen-Loeve Transform, or the KLT adapts itself to the signal. i.e. it draws the basis vectors from the signal and thus help define the signal better than any fixed transforms.

Since KLT can be obtained by the eigen decomposition of the Correlation or the Covariance matrix, an autocorrelation function matrix of the sequence has to be determined. Autocorrelation function matrix should be a toeplitz matrix. It is usually determined by the calculating the correlation with time lag 0 till time lag of 350. A 351x351 toeplitz matrix is determined and its eigen vectors are obtained and plotted. Since the eigen of white noise is a straight line, the eigen plot of the non-coding region should be a near straight line, whereas the eigen of the coding region must behave more like the eigen of a signal. Fourier Transform shows the peak for the coding region, implying that there exists a signal. Since it is not a clear spectrum, coding sequence is a signal and noise, whereas non-coding sequence is a noisy sequence. Though the eigen values of the coding region are not going to be without eigen values, the figure of merit of the coding region will be higher than that of a non-coding region.

5.2.1 Algorithm

- Consider a DNA coding strand. Calculate the indicator sequence – a series of $\frac{3}{4}$ and $-\frac{1}{4}$ to indicate the presence or absence of the base.
- Take each strand into consideration.
- Calculate the autocorrelation lags from r(0) to $r\left(\frac{N}{2} - 1\right)$
- Form the Toeplitz Autocorrelation Function Matrix for each base. A typical Toeplitz matrix looks like this

$$\text{Toeplitz ACF matrix} = \begin{bmatrix} r(0) & r(1) & r(2) & \vdots & r(\frac{N}{2} - 1) \\ r(1) & r(0) & r(1) & \vdots & r(\frac{N}{2} - 2) \\ r(2) & r(1) & r(0) & \vdots & r(\frac{N}{2} - 3) \\ \cdots & \cdots & \cdots & \vdots & \cdots \\ r(\frac{N}{2} - 1) & r(\frac{N}{2} - 2) & r(\frac{N}{2} - 3) & \vdots & r(0) \end{bmatrix}$$

- Calculate the eigenvalues for each Autocorrelation Function Matrix.
- Add the individual strings of eigenvalues.
- Plot the sequence.

5.2.2 KLT applied on a Coding Region

5.2.2.1 Details of the Coding Sequence.

Organism:	Schizzosaccharomyces Pombe
Length of Sequence:	2784
Coding Region:	891:1982
Length of Coding Sequence:	1092
ACF matrix dimension:	351 x 351

5.2.2.2 Eigen Plot of the Coding Sequence.

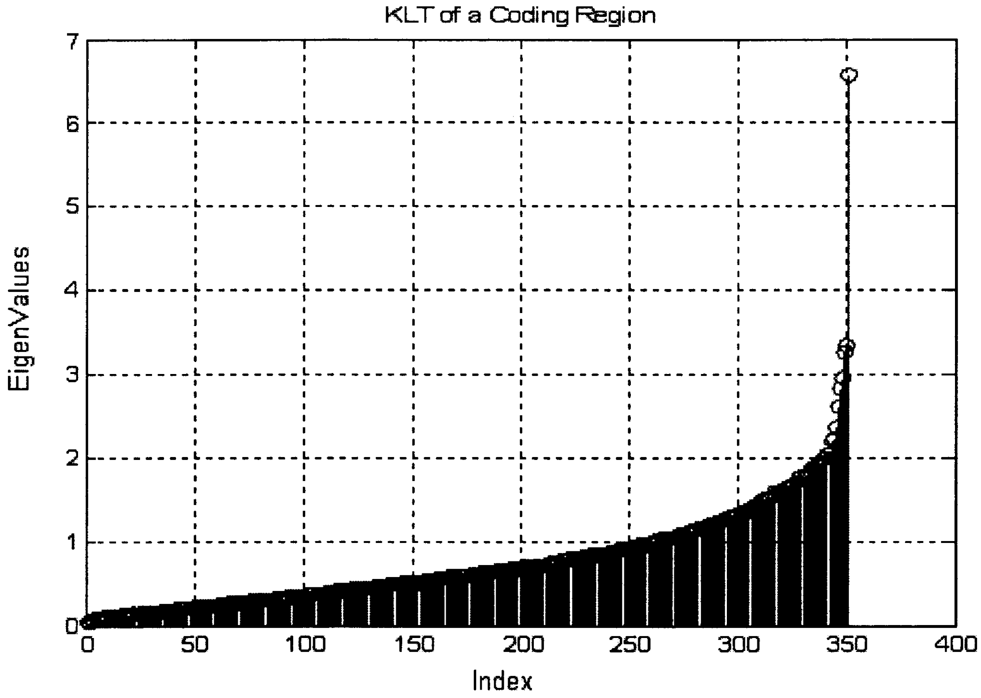


Figure 5.3 KLT of a coding sequence.

5.2.2.3 Observation.

- $$\text{Figure of Merit} = \frac{\text{Maximum Eigen Value}}{\text{Mean of all Eigen Values}} = \frac{6.5787}{0.7500} = 8.7716$$

5.2.2.4 Inference.

- The figure of merit calculated for the same coding DNA strand differs from that when calculated by using the DFT.

5.2.3 KLT applied on a Non-Coding Region

5.2.3.1 Details of the Sequence.

Organism:	Schizzosaccharomyces Pombe
Length of Sequence:	2784
Non -Coding Region:	concatenation of 1:890 and 1983:2784
Length of Non - Coding Sequence:	1692
ACF matrix dimension:	351 x 351

5.2.3.2 Eigen Plot of Non-Coding Sequence.

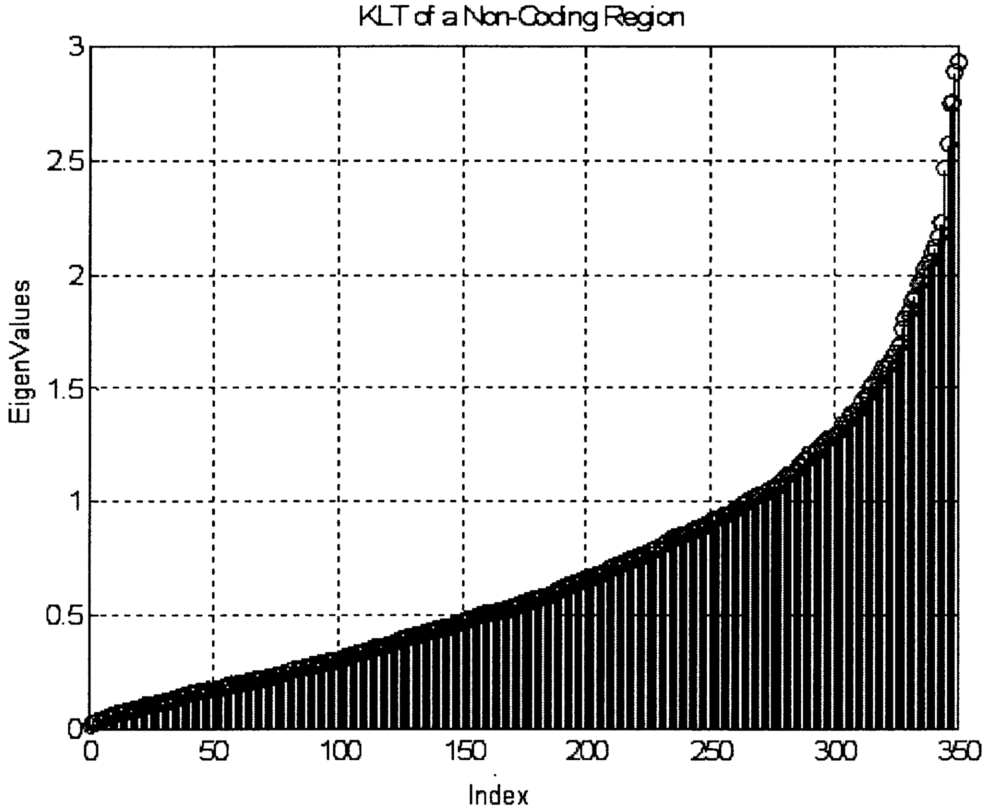


Figure 5.4 KLT of a non-coding region.

5.2.3.3 Observation.

- Figure of Merit = $\dfrac{\text{Maximum Eigen Value}}{\text{Mean of all Eigen Values}} = \dfrac{2.9229}{0.7054} = 4.1438$
- The difference between the Figure of Merits of the coding and non-coding region using the KLT = 8.7716 – 4.1438 = 4.6278.

5.2.3.4 Inference.

- The Figure of Merit of coding region is almost double of the figure of merit of a non-coding region, when KLT is applied to the sequence.

5.3 Conclusion

KLT could be a better way to demarcate the coding and the non-coding region of the DNA, than Fourier Transform. The Figure of Merit stands out much better when the KLT is applied to the signal than when DFT is applied. In KLT, the Ratio of the Figure of Merits of the coding and the non-coding region comes out to be approximately more than the double. But when DFT is applied to the signal, the ratio of Figure of Merit of the coding and the non-coding region is less than the double.

This difference will prove its efficacy when the sliding window is applied to the whole sequence. It shall be discussed in the subsequent chapter.

CHAPTER 6

SLIDING WINDOW TECHNIQUE

6.1 Logic and Purpose

In a given DNA strand, the coding regions are the most important. But it does not mean that the introns can be completely excluded, for the combination of exons and introns form the DNA strand. It is then necessary for the existence of some effective mechanism for the easy demarcation of the approximate regions where the exons and introns exist.

One method to do so is putting into use the sliding window technique, and apply the two different transforms. If the method were all right, the figure of merit should be very high for the coding regions and low for the non-coding regions.

Another way to detect the signal is by the use of Neyman Pearson Detection. After the vector of figure of merits is obtained, two vectors are extracted, one that belongs to the exons and the other to the introns. A combined plot of the cumulative distribution function of exons and 1-CDF of introns is obtained. The point on the x-axis corresponding to the meeting point will be the threshold value. Since the plot is the CDF, the meeting point will at the same time depict the Probability of Detection and the Probability of False Alarm. Since it is desirable that the threshold be as low as possible, a new threshold is considered is considered. It is shifted to the left of the original threshold value. The point where the new threshold meets the CDF curve of exons is the Probability of Misdetection. The point where it meets the 1-CDF curve of introns will be the Probability of False Alarm. The Probability of Detection is the 1- Probability of Misdetection.

6.2 DFT in Sliding Window

As was discussed in the previous chapter, [14] applies Fourier Transform in a sliding window to demarcate the coding and the non-coding regions. The algorithm is very simple. The parameters defining a sliding window are the block size and the hop size. The power spectral density is calculated at each window of length 351. The Figure of Merit for each window is obtained. The plot of the vector of Figure of Merits is made. Another figure depicts the combined CDFs of exons and introns.

6.2.1 Algorithm

6.2.1.1 Algorithm to find the Figure of Merit of the DFT based Technique.

- The two parameters of a sliding window are the window size or block size or step size and the hop size.
- Take the whole DNA strand into consideration.
- Take a window of length 351.
- Calculate the Power Spectral Density at this window.
- Calculate the Figure of Merit, by the equation

$$\text{Figure of Merit} = \frac{\text{Maximum of Values}}{\text{Mean of all Values}}$$

- Hop the window by 1
- Repeat the whole procedure.
- Plot the graph of figure of merits.

6.2.1.2 Algorithm of Detection of Signal by using Neyman - Pearson Detection.

- Let the vector of Figure of Merits be denoted by f.
- Extract the portion that has the exons in it.
- Denote it as another vector , say E
- Calculate the Cumulative Distribution Function for E

- Extract the portion that has introns.
- Denote it as another vector , say *I*
- Calculate the Cumulative Distribution Function for I
- Plot the CDF of E and 1-CDF of I in the same plot.

6.2.3 Details of the Sequence

Organism:	Schizosaccharomyces Pombe
Length of Sequence:	2784
Coding Region:	891:1982
Length of Coding Sequence:	1091
Non -Coding Region:	concatenation of 1:890 and 1983:2784
Length of Non - Coding Sequence:	1692

6.2.4 Expected Output

As is proven in the previous chapters, the figure of merit for a coding region is higher than that of a non-coding region. Even a part of the coding region will also exhibit the period-3 property. Thus the typical plot should look like a low flat line extending from 1 to 890, a high plateau (of approximately double the amplitude) ranging from 891 to 1982, and then a low flat line from 1983 to 2784.

Though it is not possible to obtain the typical flat spectrum, there should at least be an observable difference in the coding and the non-coding regions.

6.2.5 Plot of DFT in a sliding window

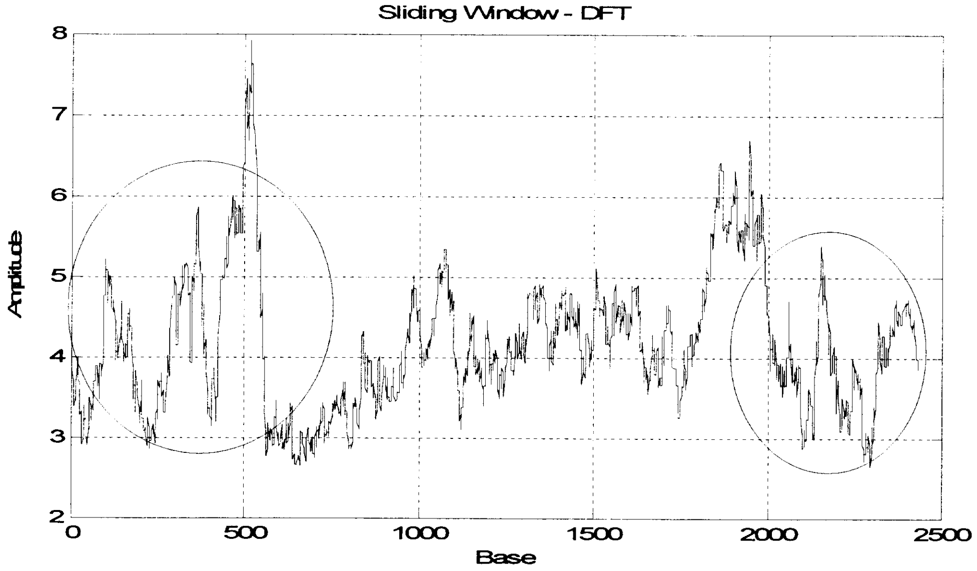


Figure 6.1 DFT as applied in a sliding window. The circled regions depict the introns.

6.2.5.1 Observation.

- The plot goes contrary to the assumption that the introns will have a lower Figure of Merit than the Exons.
- Even if it were the other way round, there is no clear cut demarcation between the intronic and the exonic regions.

This could only mean that the DFT sliding window technique is not exactly a suitable option. It remains to be seen as to what the plot of Neyman Pearson has to say about the detection of Exons and Introns.

6.2.6 Plot of the CDF of the Introns and Exons

The following is the plot of the CDF of Introns and Exons. The prediction is to be made by the Neyman Pearson.

In this case, Neyman Pearson detection can be made only *a posteiori*. Taking a peek into how the detection is actually done, it goes back to say that after the window is slid and a perfect demarcation is met, one can distinguish between the large and the small peaks, and plot the CDFs of the exons and introns. In this case, when the organism was taken from the NCBI database, the intron and the exon location were known. So it was easy to make the exact clear cut demarcation. But given a sequence and asked to find the exon and intron location *a priori*, this method is not very convenient.

This is the plot of the Neyman Pearson detection.

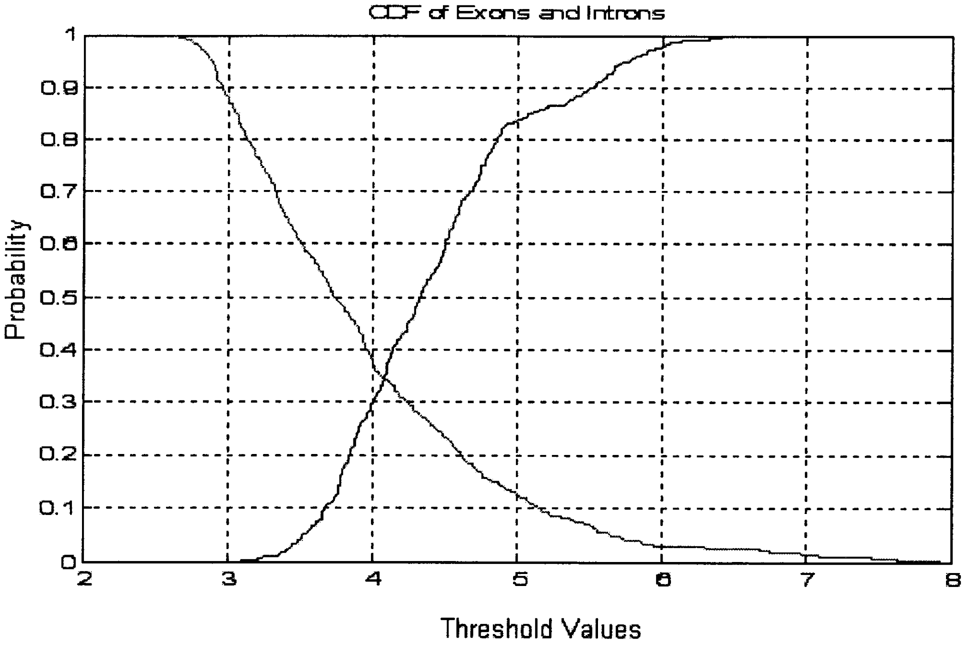


Figure 6.2 Plot of CDF of exons and introns.

6.2.6.1 Observation.

- Approximately around 4.1 is the discriminating factor.

- It means to say that the regions above the 4.1 will depict the exonic and the regions below 4.1 will depict the intronic regions.

- The point at the Y axis corresponding to the threshold value is approximately 0.35. This means that the Probability of False Alarm and the Probability of Miss Detection are both 0.35 (Once again bringing to notice that this is the CDF plot and not the PDF plot. So there is no need to integrate the areas to get the Probability of False Alarm and the Probability of Misdetection.)

- The idea would be shift the threshold towards the left, so that it is made as less as possible. So in case the Threshold is made at 3.8, the probability of misdetection will be 0.1 and the probability of false alarm is 0.5.

- Though the above is a good measure, the plot does not support this.

6.2.6 Conclusion

- DFT may not be the best way to differentiate the coding and the non-coding regions.

6.3. Application of Karhunen Loeve Transform in Sliding Window

6.3.1 Algorithm

6.3.1.1 Algorithm to find the Figure of Merit of the KLT based sliding window technique.

- Take one base indicator sequence into consideration.
- Take a window of 351.
- For that window, calculate 175 x 175 Toeplitz Autocorrelation Function Matrix.
- Find the eigenvalues of the matrix.
- Do the same for each indicator sequence.
- Add the eigen values for all the four different bases.
- Calculate the Figure of Merit given by the formula

$$\text{Figure of Merit} = \frac{\text{Maximum of Eigenvalues}}{\text{Mean of all Eigenvalues}}$$

- Hop the window by one.
- Repeat the above procedure till then end of the sequence
- Plot

6.3.2 Algorithm of Detection of Signal by using a Neyman - Pearson Detector

- Let the vector of Figure of Merits be denoted by *f*.
- Extract the portion that has the exons in it.
- Denote it as another vector , say *E*
- Calculate the Cumulative Distribution Function for E
- Extract the portion that has introns.
- Denote it as another vector , say *I*
- Calculate the Cumulative Distribution Function for *I*
- Plot the CDF of *E* and 1-CDF of *I* in the same plot.

6.3.3 Details of the Sequence

Organism:	Schizosaccharomyces Pombe
Length of Sequence:	2784
Coding Region:	891:1982
Length of Coding Sequence:	1091
Non -Coding Region:	concatenation of 1:890 and 1983:2784
Length of Non - Coding Sequence:	1692

6.3.4 Speculated Output and the Reason

KLT adapts itself to the signal. So it is enough to observe a clear demarcation between the two regions. Theoretically speaking, the exons should have a high peak and the introns should be very low. But it is observed that in a few organisms, it is not the case. That is, though there is a clear demarcation between the two regions, the introns show a higher peak than the exons. The reason for this is still under research. For the organism used in all the discussions till now, the plot goes as follows.

6.3.5 Plot of Eigen Values in Sliding Window

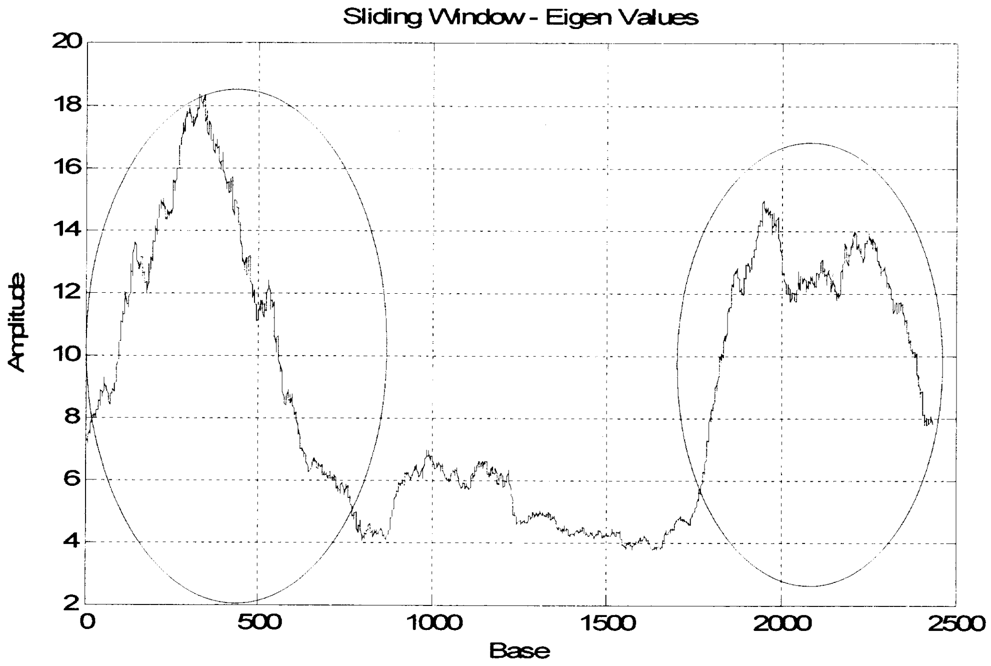


Figure 6.3 KLT in a sliding window.

6.3.5.1 Observation.

- There is a clear distinction between the two regions.
- The graph is a clear contradiction to the fact that the exonic regions have a greater Figure of Merit than when compared to the intronic regions.

6.3.6 Plot of the CDF

The Neyman Pearson based detector clearly predicts the output of the signal detection. The graph shows the plot of CDF of exons and 1 – CDF of introns.

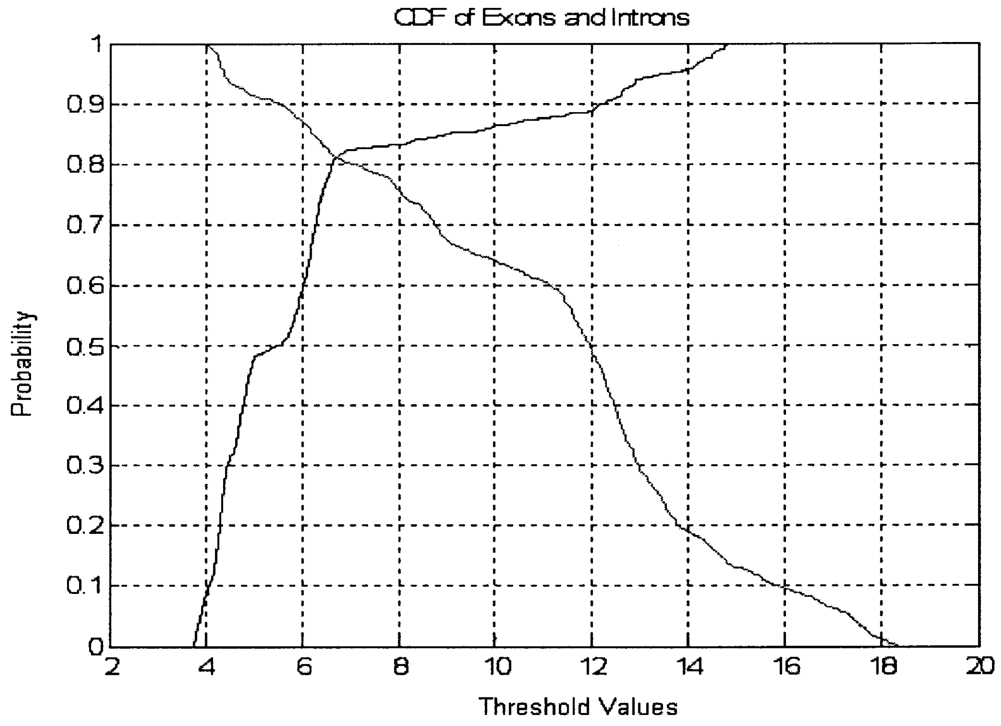


Figure 6.4 Plot of CDF of exons and introns.

6.3.6.1 Observation.

- Approximately around 7 is the discriminating factor.
- The actual answer should have been that the regions above 7 are the exonic and the regions below 7 are the intronic regions. But it is known *a priori* that it is the other way round. One possible explanation that can be given is that the KLT adapts itself to the signal. There must be something in the signal by itself that makes the introns have a higher peak than the exons. The current state of research can just state the fact that the KLT is able to clearly distinguish between the two regions. There is a very clear cut demarcation between the intronic and the exonic regions. The demarcation is much better than what DFT applied in sliding window has to provide.

6.3 Conclusion

KLT is successful in clearly demarcating the exonic and the intronic regions. Current research of the thesis can help state the fact that the demarcation criteria once met, the sequence should be tested for the period -3 property. If they satisfy that condition, then that high peak belongs to an exon, or else it belongs to an intron.

CHAPTER 7

COMPARISON OF THE TWO TRANSFORM FOR A LARGE NUMBER OF ORGANISMS

The thesis now reaches the culmination of what could be a better option for a large number of organisms. Organisms are now considered as random, without a choice at the order or the kingdom of the same. DNA sequences of many organisms were taken. The exons and introns were taken out separately and sliding window for DFT and KLT was applied on them. The ultimate sequence of exons was the addition of all the individually transformed exon sequences. Similarly, the ultimate sequence of introns was the addition od all the individually transformed intron sequences. The combined plot of CDF of exons and 1-CDF of introns were plotted for DFT and KLT. This is basically to give a big picture as to what transform works much better than the other.

7.1 DFT Applied on Large Number of Organisms

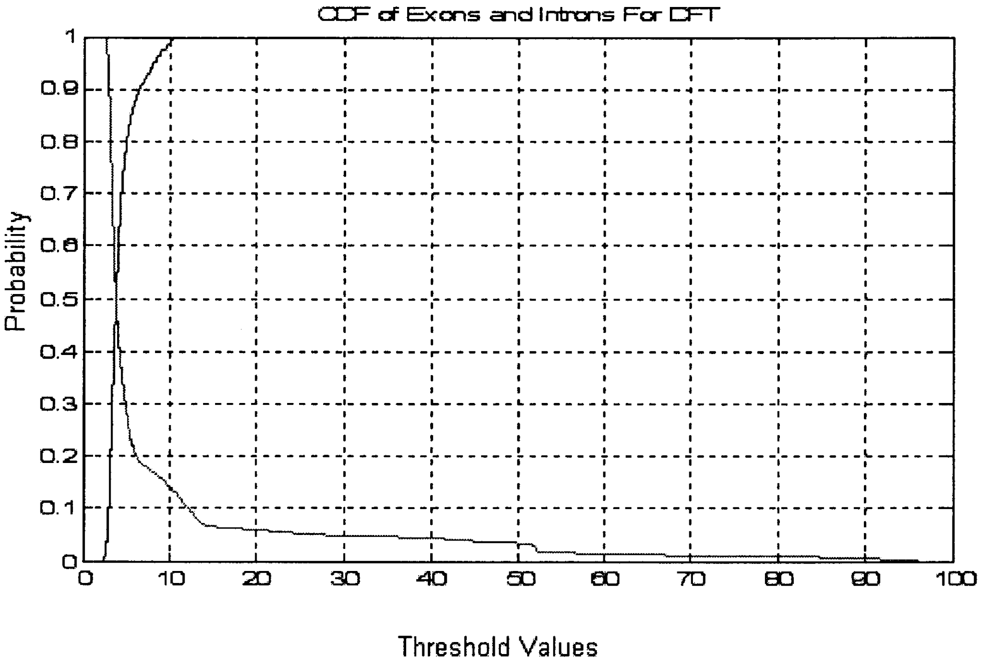


Figure 7.1 CDF plot of DFT applied on large number of organisms.

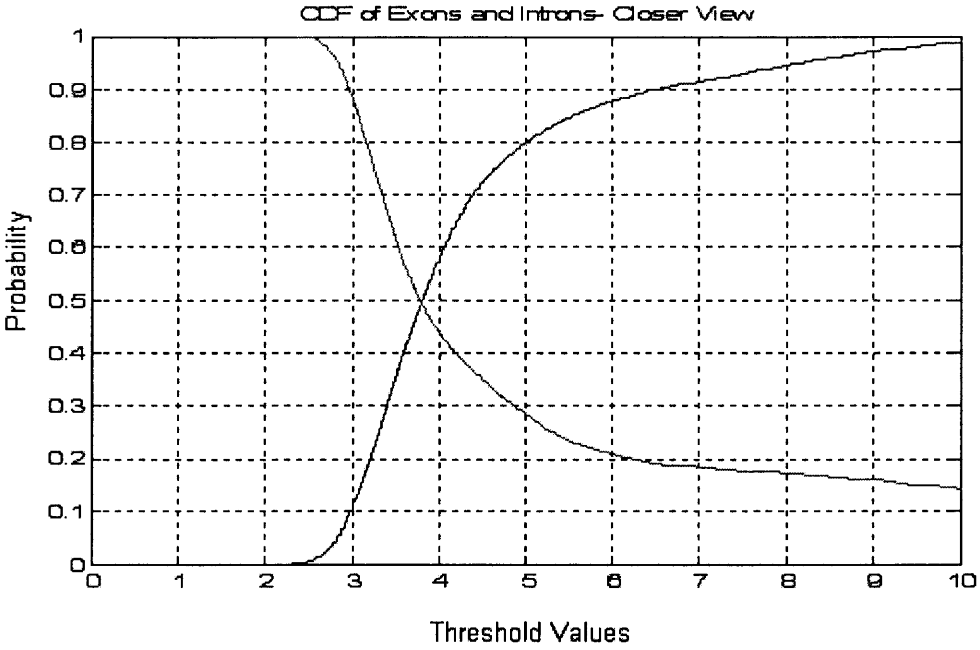


Figure 7.2 A closer view of Fig. 7.1.

7.2 KLT Applied on Large Number of Organisms

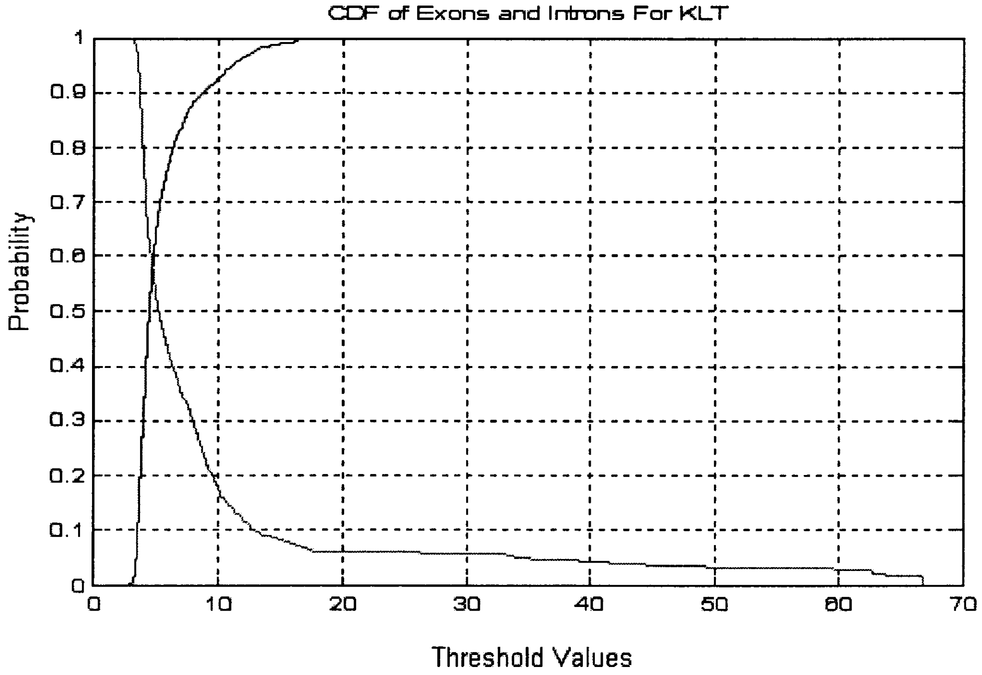


Figure 7.3 CDF plot of KLT applied on large number of organisms.

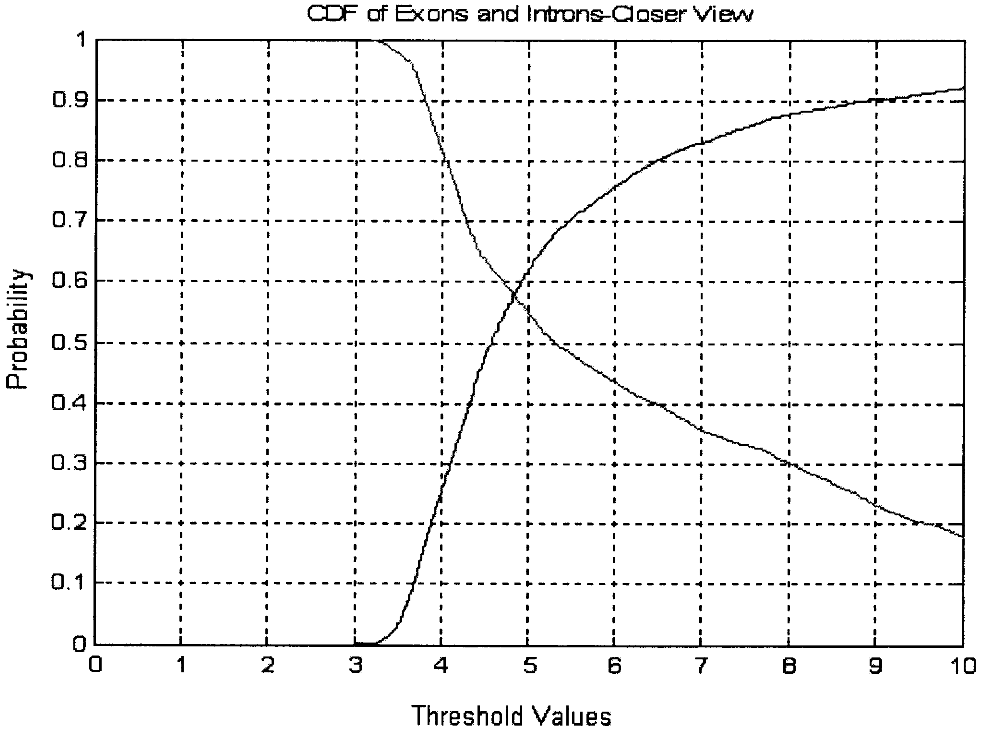


Figure 7.4 A closer View of Fig.7.3.

7.3 Observation

The observation can be clubbed into one line – DFT is not very effective! The plots clearly indicate the fact that as far as DFT is concerned, it is an equally likely probability that the detection is made perfectly. That means that the region has an equal probability that it is an exon as well as an intron. Whereas, the demarcation is better set off in the application of KLT. Though KLT inverses the region, an observer can at least observe and now be sure of the fact that where it is a peak, it is more likely that it is an intron than an exon. It is approximately 60% accurate that the region is an intron.

CHAPTER 8

SUMMARY

The thesis started off with the description about basic concepts of DNA, its structure, and the importance of codons. The nature of codons results in the period-3 property for a coding region, which is a good discriminating factor between the exons and the introns. Reference paper [14] indicates the detection of the coding and the non-coding region with the detection based on the difference in the figure of merits. But the Neyman Pearson detection of this sliding window technique does not support the efficacy of this method of coding region prediction.

KLT, on the other hand draws the basis vectors from the signal. The eigen vectors of the autocorrelation function matrix of the signal helps in the determination of a unique set of basis vectors of the sequence. The figure of merit seems to be higher for the coding region when KLT is applied to it than when DFT is applied. Moreover, the KLT helps in the better demarcation of the signal when applied in a sliding window.

Both the transforms, DFT and KLT were applied to a large number of organisms, irrespective of the fact as to whether or not the organisms belonged to the same family. Neyamn Pearson detection for the signal clearly indicated the fact that KLT transform worked better for a large number of organisms.

DFT was one transform that was used and observed by many papers in the yester years of research. Even currently, a few pioneers are working on the properties of codon region based on the Fourier Transform result [12]. But it is a fact that DFT is a fixed transform, whereas, KLT draws its basis vectors from the signal. So the application of KLT was considered. But when computation is considered, DFT is much easier. There are not many matrix calculations. When the sequence is exceptionally large, KLT takes a very long time to compute. Not to mention the fact that KLT works exactly the other way round in demarcating the sequence. The prediction is that the peaks ought to appear for an exonic region and a low for intron. And the KLT seems to contradict this statement to perfection.

Overall, if computational efficiency is to be considered, then DFT is the correct transform. But more than computation, it is the efficiency of demarcating the signal that is more important. In this case, KLT triumphs over DFT.

As a future work, it would be very interesting to look at the big picture for the organisms of the same family. For example, the application of DFT and KLT to the exons and introns of many different types of bacteria could be considered. This may help to enable future researchers determine the fact as to what transform is applicable to what family.

APPENDIX A

DNA

This appendix describes the chemical structure of the nucleotides and the bonds between the different nucleotide molecules.

Deoxyribonucleic acid or DNA is the basic hereditary unit of life. A double helical structure, it is made up of nucleotides attached to the sugar phosphate bone. Nucleic acids are linear, unbranched polymers of nucleotides.

Nucleotides are made of five carbon sugar. Deoxyribose is a pentose that has a hydrogen atom attached to its 2' carbon atom.

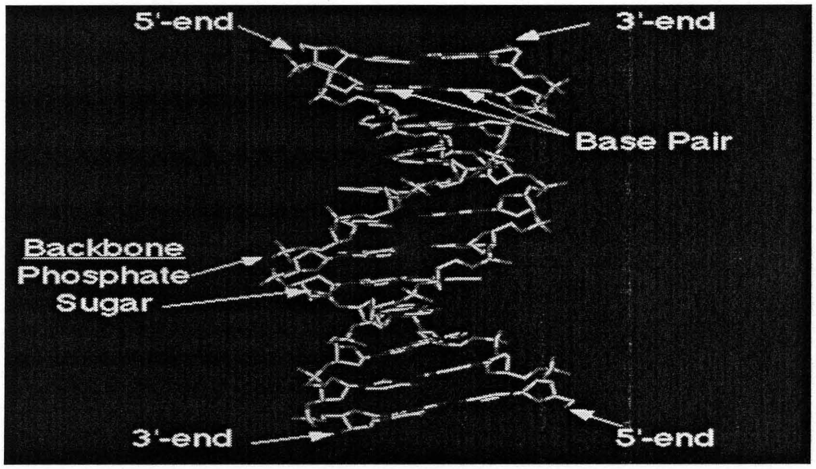


Figure A.1. Structure of DNA

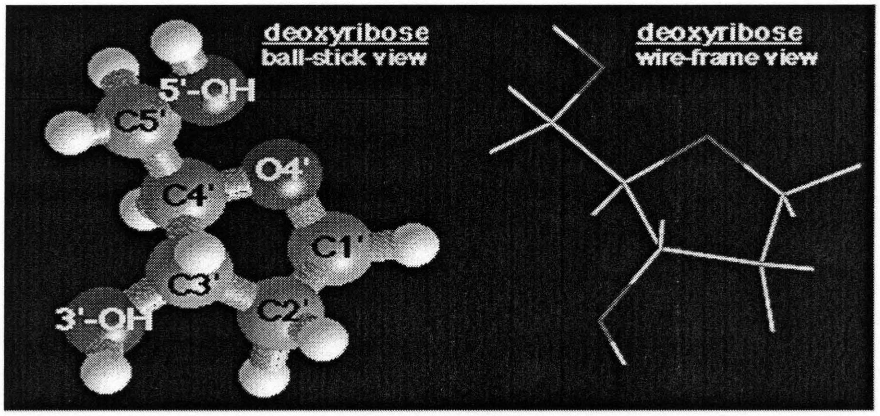


Figure A.2. Structure of Deoxyribose

Ribose is a pentose that has a hydroxyl group atom there. The monomers of DNA are the deoxyribonucleotides or the deoxyribose-containing nucleotides.

A nitrogen-containing ring structure is called a base. The base is attached to the 1' carbon atom of the pentose. In DNA, four different bases are found.

1. Two purines, called Adenine (A) and Guanine(G)

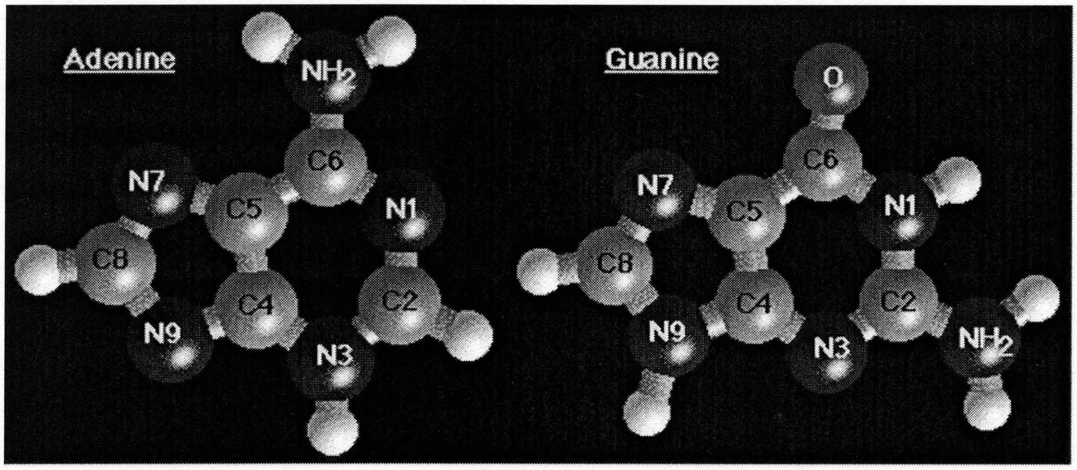


Figure A.3. Structure of Adenine and Guanine

2. Two pyrimidines called Thyamine (T) and Cytosine (C)

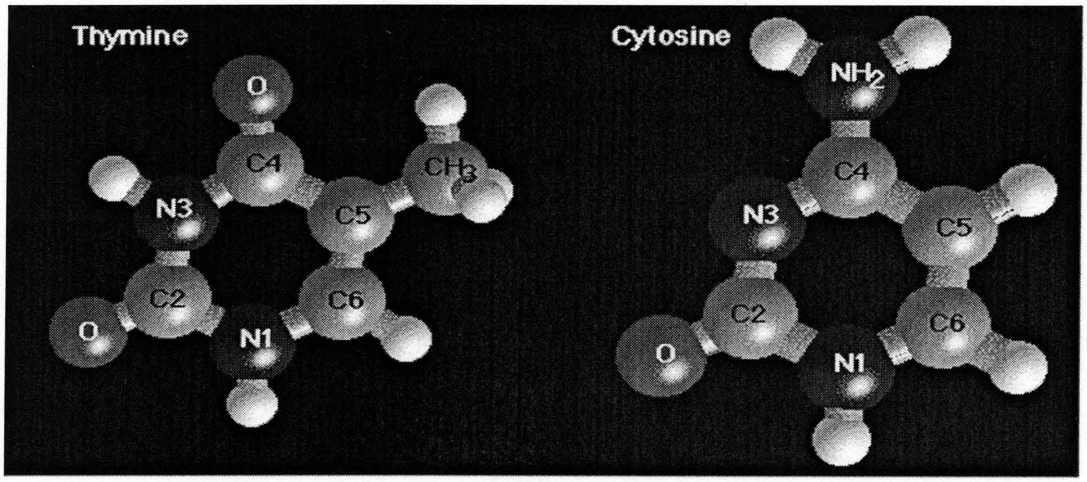


Figure A.4. Structure of Cytosine and Thymine

The combination of a base and a pentose is called a nucleoside.

The two polynucleotide strands are wound around each other. The phosphate group bonded to the 5'carobon atom of one deoxyribose is covalently bonded to the 3'carbon of the next. Each base forms hydrogen bonds with the one directly opposite to it, forming base pairs, also called as the nucleotide pairs. The hydrogen bonds between the strands of the double helix are weak enough that they can be easily separated by enzymes.

Adenosine (A) always binds with Thymine (T) with a double bond

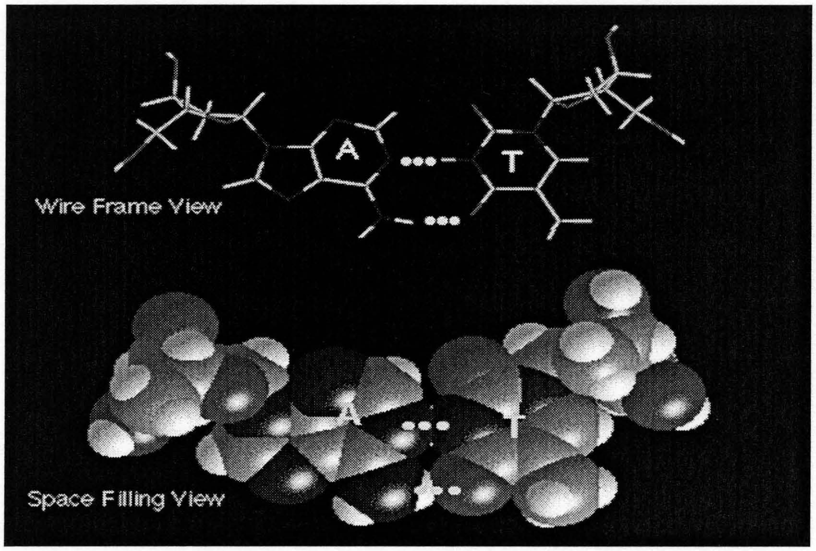


Figure A.5. Example of dA-dT base pair as found within DNA double helix

Cytosine (C) pairs up with Guanine (G) with a triple bond.

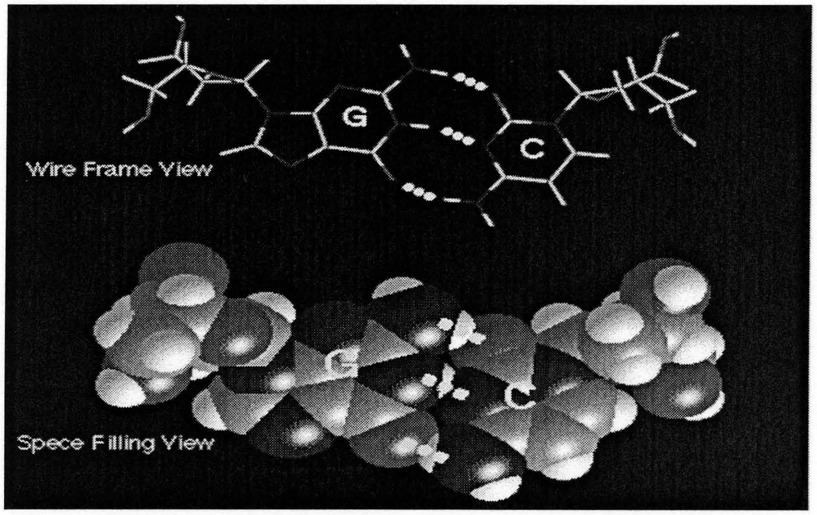


Figure A.6. Example of dG-dC base pair as found within DNA double helix

APPENDIX B

APPLICATION OF A FEW TRANSFORMS THAT DID NOT WORK

The thesis now reached a stage where it is known what to predict when Fourier Transform was applied to the DNA sequence. It would not be interesting to note what happens when a few other fixed transforms are applied to the DNA sequence. The fixed transforms are chosen in such a way that they had some relation to Fourier Transform. A common DNA sequence is taken, to predict the application of all the fixed transforms. The details of the organism are given as follows:

Details of the sequence

Organism:	Schizosaccharomyces Pombe
Length of Sequence:	2784
Coding Region:	891:1982
Length of Coding Sequence:	1091
Non -Coding Region:	concatenation of 1:890 and 1983:2784
Length of Non - Coding Sequence:	1692

B. 1 Application of DCT on DNA sequences

DCT is defined by the equation

$$\Phi(r,n) = \Phi_r(n) = \left(\frac{1}{c_r}\right)\cos\frac{(2n+1)r\pi}{2N}$$

where

$$c_r = \left\{\begin{matrix} \sqrt{N} & r = 0 \\ \sqrt{N/2} & r \neq 0 \end{matrix}\right\}$$

where $o < n$, $r < N-1$

To prove that DCT is a 2N point DFT

Let {x(n)} , n = 0,1,.. N-1 be a given sequence. Then the extended sequence {y(n)} symmetric about (2N-1)/2 point can be constructed so that

$$\begin{aligned} y(n) &= x(n) && n = 0,1,\ldots,N\text{-}1 \\ &= x(2N - n - 1) && n = N, N+1, \ldots, 2N\text{-}1 \end{aligned}$$

If W_{2N} is used to denote $\exp(-j2\pi/2N)$, then it can be seen that the DFT of {y(n)} is given by

$$Y(m) = \sum_{n=0}^{2N-1} y(n) W_{2N}^{nm}$$

which implies

$$\begin{aligned} Y(m) &= \sum_{n=0}^{N-1} x(n) W_{2N}^{nm} + \sum_{n=N}^{2N-1} y(n) W_{2N}^{nm} \\ &= \sum_{n=0}^{N-1} x(n) W_{2N}^{nm} + \sum_{n=N}^{2N-1} x(2N - n - 1) W_{2N}^{nm} \\ &= \sum_{n=0}^{N-1} x(n) W_{2N}^{nm} + \sum_{n=0}^{N-1} x(n) W_{2N}^{(2N-n-1)m} \\ &= \sum_{n=0}^{N-1} x(n) [\, W_{2N}^{nm} + W_{2N}^{-(n+1)m}] \end{aligned}$$

where m = 0 , 1 , … , 2N-1

Multiplying both sides of the equation by the factor $\frac{1}{2} W_{2N}^{m/2}$ equation becomes,

$$\frac{1}{2} W_{2N}^{m/2} Y(m) = \sum_{n=0}^{N-1} x(n) \cos\left[(2N+1) \frac{m\pi}{2N} \right]$$

where m = 0 , 1 , … , N-1

It is easy to see that except for the required scale factors, DCT II and DFT are the same.

{Y(m)} is the 2N-point DFT of {y(n)}

Equation also shows that for m = 0 , 1 , … , N – 1 , the transformed sequence {Y(m)} properly scaled is the DCT – II of the N-point sequence {x(n)}.

From the above calculations, it is clear that if the peak of the spectrum is at N/3 for DFT, it is at 2N/3 for a DCT. The plot is shown as follows:

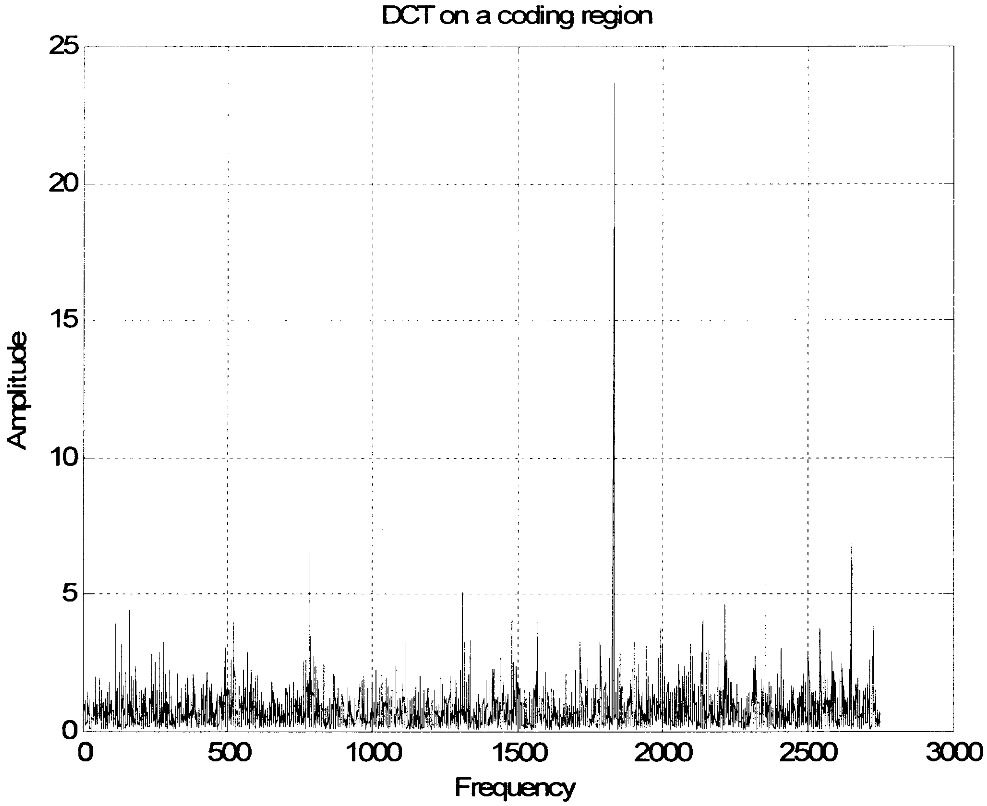


Figure B.1 DCT on a coding region

B.2 Relation between DHT and DFT

Differences between Fourier Transform and Hartley Transform are:

- DFT is complex, DHT is real.
- Direct and Inverse Transformation have the same integral operation in DHT, whereas it is different in DFT.
- DFT is symmetric transform, whereas DHT has neither odd nor even symmetry
- Hartley Transform is given by $H(f) = \int_{-\infty}^{\infty} V(f)\mathrm{cas}(2\pi ft)dt$
- Inverse Hartley Transform is $V(t) = \int_{-\infty}^{\infty} H(t)\mathrm{cas}(2\pi ft)dt$

To determine the relation between DFT and DHT, H(f) is made of odd O(f) and even E(f) components

$$O(f) = \frac{H(f) - H(-f)}{2} = \int_{-\infty}^{\infty} V(t)\sin(2\pi ft)dt \qquad E(f) = \frac{H(f) + H(-f)}{2} = \int_{-\infty}^{\infty} V(t)\cos(2\pi ft)dt$$

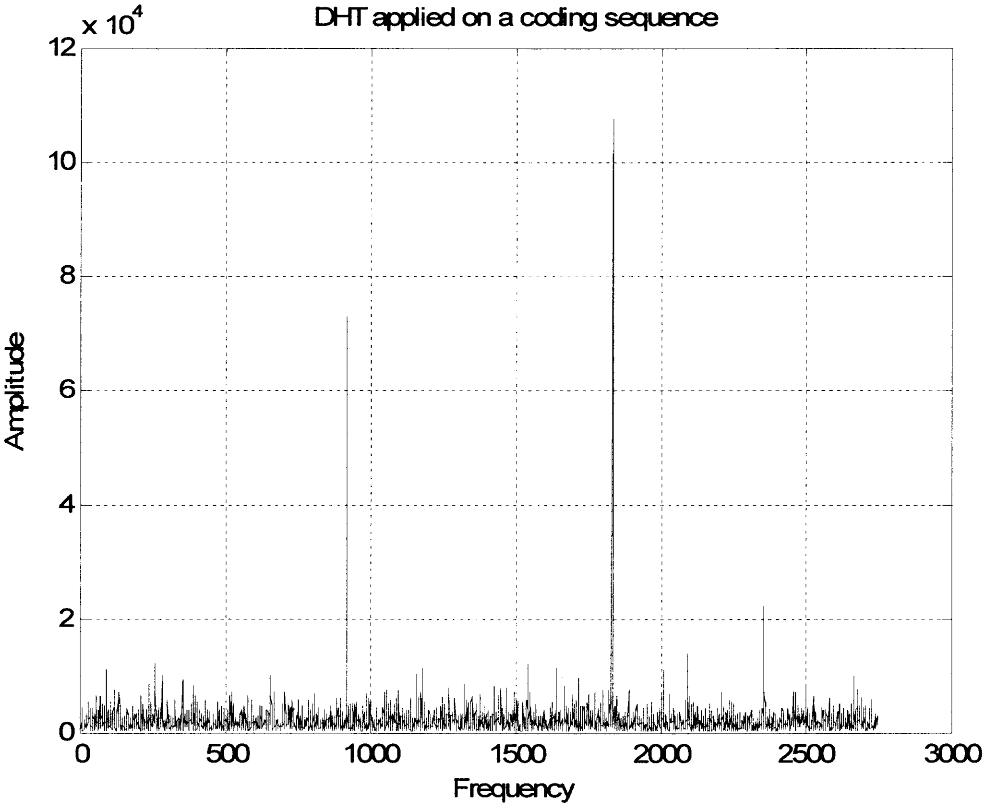


Figure B.2 Application of DHT on a coding sequence

B.3 Comparison between DFT and Walsh Hadamard Matrix

- The Walsh-Hadamard transform is similar to Fourier series analysis, but uses square waves instead of sinusoidal waves. It is used predominantly in communication.
- DFT consists of a projection onto a set of orthogonal sinusoidal waveforms. WHT consists of a projection onto a set of square waves called Walsh Functions

- FT coefficients are called Frequency Components. HT coefficients are called Sequence Components.
- In FT, waveforms are ordered by frequency. In HT, waveforms are ordered by their zero-crossings.
- FT is complex. HT is real.
- The Hadamard matrix of order 2j is given by

$$H_{2J2J} = \begin{vmatrix} H_{JJ} & H_{JJ} \\ H_{JJ} & -H_{JJ} \end{vmatrix} \qquad \text{where } H_{22} = \begin{vmatrix} 1 & 1 \\ 1 & -1 \end{vmatrix}$$

- Disadvantage of this transform is that only matrix dimension will be a power of two.

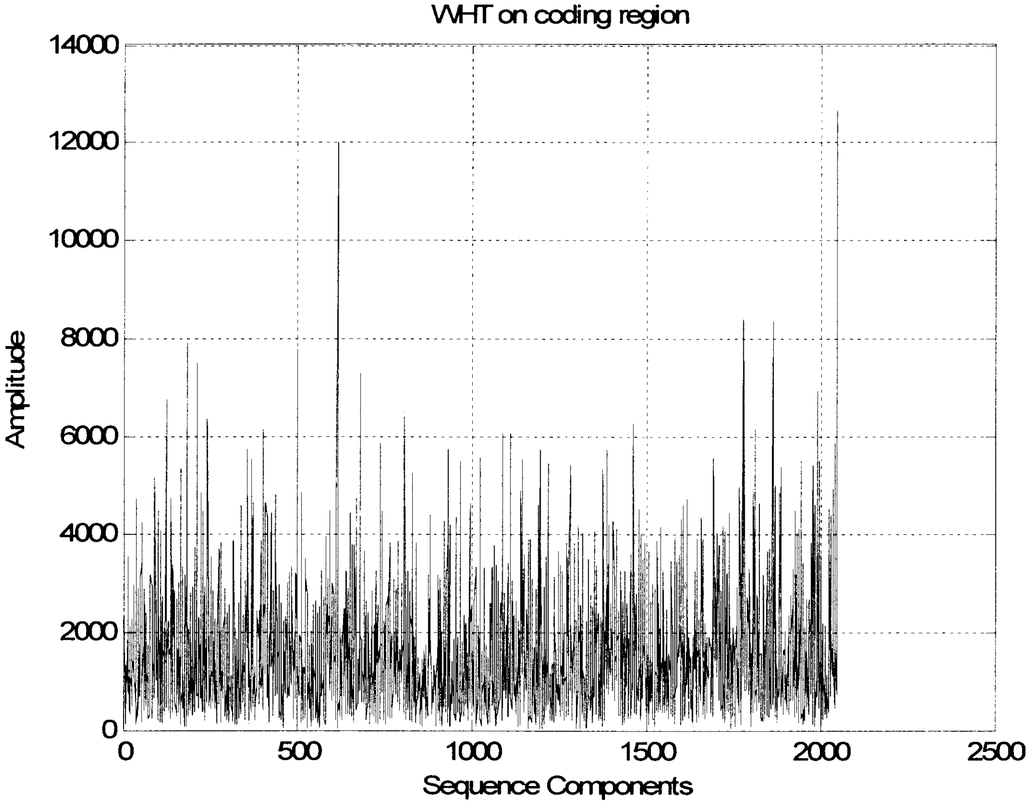


Figure B.3 WHT applied on coding sequence

APPENDIX C

CHARACTERIZATION OF A DISCRETE TIME STOCHASTIC PROCESS [32]

It is usually not easy to fine the joint probability density function for an arbitrary set of observations made on a stochastic process. Therefore we must specify the first and the second moments.

Consider a discrete time stochastic process represented in time series as

$$u(n)\ ,u(n-1)\ ,u(n-2),\ldots,u(n-M)$$

u(n) is used to define such a process.

Mean- Value Function of the process is

$$\mu(n) = E[u(n)]$$

where E denotes the statistical expectation operator.

Autocorrelation Function is defined by

$$r(n,n-k) = E[u(n)\,u^*(n-k)]$$

where k = ... , -2 , -1 , 0 , 1 , 2 , ...

* denotes complex conjugate.

Autocovariance Function is defined by

$$c(n,n-k) = E[(u(n)-\mu(n))\,(u(n-k)-\mu(n-k))^*]$$

where k = ... , -2 , -1 , 0 , 1 , 2 , ...

Mean-value , autocorrelation function and the autocovariance functions are related by this formula

$$c(n,n-k) = r(n,n-k) - \mu(n)\mu^*(n-k)$$

For strictly stationary processes, the mean, autocorrelation and autocovariance assume simpler forms

Mean	$\mu(n)$	$=$ constant
Autocorrelation Function	$r(n\ ,n - k)$	$= r(k)$
Autocovariance Function	$c(n\ ,n - k)$	$= c(k)$
Mean Square Value	$r(0)$	$= E[\lvert u(n)\rvert^2]$
Variance	$c(0)$	$= \sigma_u^2$

Mean Ergodic Theorem

The expectations or ensemble averages of a stochastic process are averages across the process. Time averages are used t build a stochastic model of a physical process by estimating unknown parameters of the model.

Consider a discrete time stochastic process u(n) that is wide sense stationary. Let μ be the mean and c(k) be the autocovariance function for lag k. For an estimate of the mean μ , the time average used is

$$\mu(N) = \frac{1}{N}\sum_{n=0}^{N-1} u(n)*u'(n)$$

where N is the total number of samples used in the estimation.

The estimate of $\hat{\mu}(N)$ is a random variable with a mean and variance of its own. $\hat{\mu}(N)$ is called the unbiased estimator of the ensemble mean of the process.

Process u(n) is mean ergodic in the mean – square error sense , if the mean –square value of the error between the ensemble average μ and the time average $\hat{\mu}(N)$ approaches zero as the number of samples , N, approaches infinity;, that is

$$\underset{N\to\infty}{\mathrm{Lt}}\ E[\lvert \mu - \hat{\mu}(N)\rvert^2] = 0$$

Correlation Matrix

Let the Mx1 observation vector $\underline{u}(n)$ represent the elements of the zero-mean time series $u(n)\ , u(n-1)\ , \ldots\ , u(n-M+1)$.

$$\underline{u}(n) = [\,u(n)\,, u(n-1)\,, \ldots\,, u(n-M+1)]^{T}$$

where T denotes transposition . Correlation matrix of a stationary discrete- time stochastic process is represented by the time series as the expectation of the outer product of the observation vector $\underline{u}(n)$ with itself.

Let $\underline{R}$ denote the MxM correlation matrix. It is represented as

$$R = E[\underline{u}(n)\underline{u}^H(n)]$$

where H denotes Hermitian Transposition (which is the complex conjugate and transpose of a matrix).

Correlation Matrix in the expanded form will be

$$R = \begin{bmatrix} r(0) & r(1) & \cdots & r(M-1) \\ r(-1) & r(0) & \cdots & r(M-2) \\ \vdots & \vdots & \ddots & \vdots \\ r(-M+1) & r(-M+2) & \cdots & r(0) \end{bmatrix}$$

The element r(0) is always real.

Properties of Correlation Matrix

1. The correlation matrix of a stationary discrete-time stochastic process is Hermitian.

   When a matrix is Hermitian , $R^H = R$

   This means that the correlation matrix can be written as

$$R = \begin{bmatrix} r(0) & r(1) & \cdots & r(M-1) \\ r^*(1) & r(0) & \cdots & r(M-2) \\ \vdots & \vdots & \ddots & \vdots \\ r^*(M-1) & r^*(M-2) & \cdots & r(0) \end{bmatrix}$$

2. The correlation matrix of a stationary discrete-time stochastic process is Toeplitz This means that if all the elements on the main diagonal are equal and if the elements on any other diagonal parallel to the main diagonal are also equal, then the matrix is Toeplitz.

It may be stated that if the discrete time stochastic process is wide sense stationary, then its correlation matrix R is Toeplitz.

3. Correlation matrix of a discrete- time stochastic process is always nonnegative definite and almost always positive definite.

   This means that if $\underline{a}$ is an arbitrary (non-zero) Mx1 complex –valued vector, y is the inner product $\underline{a}$ and the observation vector $\underline{u}(n)$, defined by $y = \underline{a}^H \underline{u}(n)$, and R is the correlation matrix of u(n) , then

$$\begin{aligned} E\,[|y|^2] &= E[yy^*] \\ &= E[\underline{a}^H \underline{u}(n) \underline{u}^H (n) \underline{a}] \\ &= \underline{a}^H E[\underline{u}(n) \underline{u}^H (n)] \underline{a} \\ &= \underline{a}^H R \underline{a} \end{aligned}$$

   Since

$$E[|y|^2] \geq 0$$

   Therefore

$$\underline{a}^H R \underline{a} \geq 0$$

   A Hermitian form that satisfies this condition for every non-zero a is said to be non-negative or positive semidefinite.

4. Correlation Matrix of a wide- sense stationary process is non-singular due to the unavoidable presence of noise.

   The matrix R is said to be non-singular if its determinant, denoted by det(R) is nonzero.

APPENDIX D

TRANSFORMS AND THEIR PROPERTIES [28]

Let us consider a sequence of sampled signals {f(k)} as the weighted sum of component sequences. This is to highlight certain features of the signal.

$$f(n) = \sum_{k=-\infty}^{\infty} f(k)\delta(n\text{-}k)$$

where the component sequence $\delta(n\text{-}k)$ is the Kronecker delta sequence:

$$\delta(n - k) = \begin{Bmatrix} 1, & n - k = 0, \\ 0, & \text{otherwise} \end{Bmatrix}$$

To represent the signal as a superposition of component sequences, we consider the signal to be defined from the interval $0 \leq k \leq N\text{-}1$. Signal {f(k)}can then be defined as an N dimensional vector f

$$\underline{f} = \begin{bmatrix} f_0 \\ f_1 \\ f_2 \\ \vdots \\ f_{N-1} \end{bmatrix} = f_0 \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + f_1 \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \cdots + f_{N-1} \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

$$= f_0\underline{e}_0 + f_1\underline{e}_1 + \ldots + f_{N-1}\underline{e}_{N-1}$$

The sequence {f(k)} is now represented as a point in the N dimensional Euclidean space spanned by the basis vectors $\{e_0 , e_1 , \ldots e_{N-1}\}$ .These basis vectors are linearly independent , since the linear combination $c_0\underline{e}_0 + c_1\underline{e}_1 + \ldots + c_{N-1}\underline{e}_{N-1}$ can vanish only when $c_0 = c_1 = \ldots = c_{N-1} = 0$

D.1 Block Transforms

Let the signal and the spectral vectors be defined as

$$\underline{f}^T = [f_0 , f_1 , \dots , f_{N-1}]$$

$$\underline{\theta}^T = [\theta_0 , \theta_1 , \dots , \theta_{N-1}]$$

Let the real orthonormal sequences $\phi_r(k)$ be the rows of a transformation matrix , $\phi(r , k)$

$$\Phi = [\phi(r , k)]$$

D.1.1 Properties of Block Transforms

D.1.1.1 Orthogonality

It is evident that

$$\underline{\theta} = \Phi \, \underline{f}$$

And

$$\underline{f} = \Phi^{-1}\underline{\theta} = \Phi^T\underline{\theta}$$

which implies that

$$\Phi^{-1} = \Phi^T$$

This is the principle of orthogonality.

D.1.1.2 Orthonormality

Let Φ_r be a column vector representing the basis sequence$\{ \phi_r(k) \}$

$$\underline{\Phi}_r^T = [\phi_r(0), \ \phi_r(1), \dots , \phi_r(N-1)]$$

If f were written as the sum of the basis vectors,

$$\underline{f} = [\underline{\Phi}_0 \underline{\Phi}_1 \ldots \underline{\Phi}_{N-1}] \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_{N-1} \end{bmatrix} = \theta_0 \underline{\Phi}_0 + \theta_1 \underline{\Phi}_1 + \ldots + \theta_{N-1} \underline{\Phi}_{N-1}$$

The orthonormality condition is

$$\underline{\Phi}_n^T \underline{\Phi}_s = \delta_{n-s}$$

D.1.1.3 Unitary Matrix

For complex valued signals and bases, the transformation becomes

$$\underline{\theta} = \Phi^{*} \underline{f} \leftrightarrow \underline{f} = \Phi^{T} \underline{\theta}$$

with the property

$$\Phi^{-1} = \Phi^{*T}$$

D.2 Discrete Fourier Transform

It is the most important orthogonal transform. It is the set of orthogonal complex sinusoids.

$$x_r(n) = e^{j2\pi rn/N} = W^{rn}$$
$$W \triangleq e^{j2\pi/N}$$

where r = 0 , 1 , ... , N-1

Forward and Inverse DFT are given by

$$X(k) = \text{DFT}\{x(n)\} = \sum_{n=0}^{N-1} x(n)\, e^{-j2\pi nk/N}$$
$$x(n) = \text{IDFT}\{X(k)\} = \frac{1}{N} \sum_{k=0}^{N-1} X(k)\, e^{j2\pi nk/N}$$

The transform kernels are given by

$$\text{DFT Kernel} = \Phi = [W^{-nk}]$$

$$\text{IDFT Kernel} = \Phi^{-1} = \frac{1}{N}[W^{nk}] = \frac{1}{N}\Phi^{*}$$

A few points that are to be noted along with the information about this transform are:

- The set of coefficients {X(k)} constitute the frequency spectrum of the sample.
- X(k) and x(n) are periodic in their arguments with period N.
- The first coefficient X(0)/N is the 'DC' value of the signal .
- The fundamental $\{x_1(n) = e^{j2\pi n/N}\}$ is a unit vector in the complex plane that rotates with the time index n.

The unitary DFT is a normalized DFT wherein the scaling factor N changes. The equations thus become:

$$X'(k) = \frac{1}{\sqrt{N}} X(k) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x(n) W^{nk}$$

This makes

$$x(n) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} X'(k)\, W^{-nk}$$

The unitary transformation is

$$\Phi = \frac{1}{\sqrt{N}}[W^{-nk}]$$

$$\Phi^{-1} = \frac{1}{\sqrt{N}}[W^{nk}] = \Phi^{*T} = \Phi^{*}$$

The basis vectors of the unitary DFT (the columns of Φ^{*}) are the eigenvectors of a circulant matrix. i.e. if the kth column of Φ^{*} is denoted by

$$(\Phi_k^{*})^T = [W^0 , W^{lk} , \ldots , W^{(N-1)k}]$$

then Φ_k^{*} are the eigenvectors in

$$\Re\Phi_k^{*} = \lambda_k \Phi_k^{*}$$

where is the circulant matrix given by

$$\Re = \begin{bmatrix} h_0 & h_{N-1} & \cdots & h_1 \\ h_1 & h_0 & \cdots & h_2 \\ \vdots & \vdots & \vdots & \vdots \\ h_{N-1} & h_{N-2} & \cdots & h_0 \end{bmatrix}$$

D.3 Karhunen Loeve Transform

It is a unitary transform that diagonalizes the covariance or correlation matrix. It is efficient because the processing of any one coefficient has no direct bearing on the others.

Theory

Let R be the (N x N) correlation matrix of a random sequence. $x = (x_1, x_2, \ldots, x_n)^T$

$$R = E\ [xx^H]$$

$$R = E \begin{bmatrix} x_1x_1^* & x_1x_2^* & x_1x_3^* & \cdots & x_1x_N^* \\ x_2x_1^* & x_2x_2^* & x_2x_3^* & \cdots & x_2x_N^* \\ x_3x_1^* & x_3x_2^* & x_3x_3^* & \cdots & x_3x_N^* \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_Nx_1^* & x_Nx_2^* & x_Nx_3^* & \cdots & x_Nx_N^* \end{bmatrix}$$

Unitary matrix diagonalizes R be defined as ϕ

$$\phi^{-1} = \phi^H$$

$$\Rightarrow \phi\phi^H = I$$

$$\Rightarrow \phi^H R\phi = \phi^{-1}R\phi = \Lambda = \text{diag}[\lambda_1, \lambda_2, \ldots, \lambda_k, \ldots, \lambda_N]$$

where $\lambda_{1=1,2,3,\ldots,N}$ are the eigenvectors of R

ϕ is KLT matrix

Let y be the forward transform of x.

$$y = \phi^{-1}x = \phi^{H}x$$

Let the inverse transform of y be

$$x = \phi y$$

where $y = (y_1 , y_2 , \dots , y_N)^T$ represents the random sequence in the transform domain. Correlation matrix for y is the given by

$$\begin{aligned} E[yy^H] &= E[\phi^H xx^H \phi] \\ &= \phi^H E[xx^H]\phi \\ &= \phi^H R\phi \\ &= \Lambda \end{aligned}$$

y has no cross correlation. X has been decorrelated by KLT matrix ϕ.

Note: The KLT is a signal dependent transformation, the implementation of which requires the estimation of the correlation matrix of the input vector, the diagonalization of this matrix, and the construction of the required basis vectors, these computation make KLT impractical for real-time applications. DCT provides the pre-determined set of basis vectors that are a good approximation to KLY. For a stationary zero mean, first order Markov process, DCT is asymptotically equivalent to KLT, both as the sequence length increases and also as the adjacent correlation coefficient tends towards unity; the adjacent correlation coefficient of a stochastic process is defined as the autocorrelation function of unit time lag, divided by the autocorrelation function of the process for zero lag. Whereas KLT is signal dependent, the DCT is signal independent and can therefore be implemented in a computationally efficient manner. The DCT basis vectors are a good approximation to the KLT for some signals.

APPENDIX E

SIGNAL DETECTION USING NEYMAN PEARSON THEOREM [27]

Neyman Pearson Approach

Take into consideration a Neyman Pearson approach to signal detection with hypothesis testing. Assume a realization of a random variable whose PDF is either N(0,1) or N(1,1) This notation of $N(\mu, \sigma^2)$ means that the it is a Gaussian PDF with mean μ and variance σ^2. It is now necessary to determine if the $\mu = 0$ or $\mu = 1$based on a single observation x[0]. The two hypotheses are summarized as follows:

$$H_0 : \mu = 0$$

$$H_1 : \mu = 1$$

where H_0 is referred to as the Null Hypothesis and H_1 as the alternative hypothesis. The point where they meet is called as the threshold. If we decide H_1 but H_0 is true, then we make Type I error. This is also called as the False Alarm. Type 1 error probability is denoted by $(P(H_1 = H_0))$ and is called as the Probability of False Alarm, denoted by α or P_{FA}.

If we decide on H_0 but H_1 is true then we make the Type II error. This is called the Miss. Type II error probability is denoted by $(P(H_0 = H_1))$ and is called as the Probability of Miss , denoted by β or P_M .

To design an optimal detector, we seek to minimize the error $P(H_0;H_1)$, or equivalently to maximize 1 - $P(H_0;H_1)$. The latter is just $P(H_1;H_1)$ and is called as the Probability of Detection. It is denoted by 1- β or P_D.

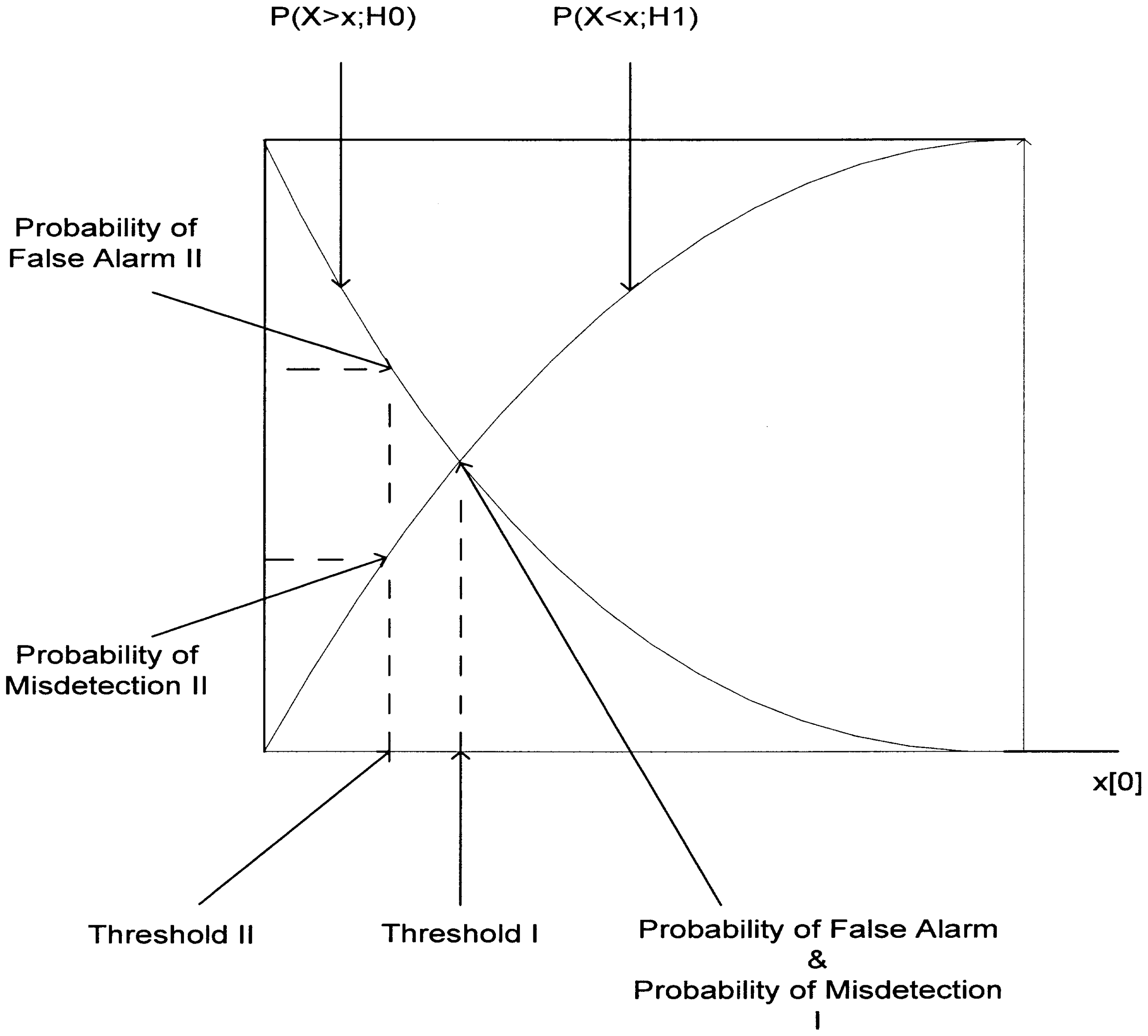

P(X>x;H0)
P(X<x;H1)
Probability of False Alarm II
Probability of Misdetection II
x[0]
Threshold II
Threshold I
Probability of False Alarm & Probability of Misdetection I

Figure E.1 Neyman Pearson Detector

APPENDIX F

MATLAB SOURCE CODE

% to obtain the indicator sequence without canceling the DC component.

```
function y= insert(x)
l=length(x);
for i=1:1:l
y(2*i-1)=x(i);
end

function y = indicatorA(x);
x=insert(x);
x1=strrep(x,'A','1');
x2=strrep(x1,'T','0');
x3=strrep(x2,'C','0');
x4=strrep(x3,'G','0');
y = str2num(x4);

function y = indicatorC(x);
x=insert(x);
x1=strrep(x,'C','1');
x2=strrep(x1,'A','0');
x3=strrep(x2,'T','0');
x4=strrep(x3,'G','0');
y = str2num(x4);

function y = indicatorG(x);
x=insert(x);
x1=strrep(x,'G','1');
x2=strrep(x1,'A','0');
x3=strrep(x2,'T','0');
x4=strrep(x3,'C','0');
y = str2num(x4);

function y = indicatorT(x);
x=insert(x);
x1=strrep(x,'T','1');
x2=strrep(x1,'A','0');
x3=strrep(x2,'C','0');
x4=strrep(x3,'G','0');
y = str2num(x4);
 function [ya yt yc yg]=indicator(x);
ya=indicatorA(x);
```

```
yt=indicatorT(x);
yc=indicatorC(x);
yg=indicatorG(x);

% to calculate the indicator sequence by canceling out the DC component

function y = indicatorA34(x);
x=insert(x);
x1=strrep(x,'A','3/4');
x2=strrep(x1,'T','-1/4');
x3=strrep(x2,'C','-1/4');
x4=strrep(x3,'G','-1/4');
y = str2num(x4);

function y = indicatorC34(x);
x=insert(x);
x1=strrep(x,'C','3/4');
x2=strrep(x1,'A','-1/4');
x3=strrep(x2,'T','-1/4');
x4=strrep(x3,'G','-1/4');
y = str2num(x4);

function y = indicatorG34(x);
x=insert(x);
x1=strrep(x,'G','3/4');
x2=strrep(x1,'A','-1/4');
x3=strrep(x2,'T','-1/4');
x4=strrep(x3,'C','-1/4');
y = str2num(x4);

function y = indicatorT34(x);
x=insert(x);
x1=strrep(x,'T','3/4');
x2=strrep(x1,'A','-1/4');
x3=strrep(x2,'C','-1/4');
x4=strrep(x3,'G','-1/4');
y = str2num(x4);

function [ya yt yc yg]=indicator34(x);
ya=indicatorA34(x);
yt=indicatorT34(x);
yc=indicatorC34(x);
yg=indicatorG34(x);
```

% to obtain the power spectral density of the entire sequence

```
function S = indfft(q,N)
[a t c g] = indicator(q);
A = abs(fft(a));
T = abs(fft(t));
C = abs(fft(c));
G = abs(fft(g));
S = A.^2 + T.^2 + C.^2 + G.^2;
S(1) = S(2);
N = length(S);
```

% application of KLT over the entire sequence

```
Function y = karhu(q);
[a t c g] = indicator34(q);

N = length(a);

for i = 0:1:350
   wa = a(i+1:N)*a(1:N-i)'/(N - i);
   Wa(i+1) = wa;

   wt = t(i+1:N)*t(1:N-i)'/(N - i);
   Wt(i+1) = wt;

   wc = c(i+1:N)*c(1:N-i)'/(N - i);
   Wc(i+1) = wc;

   wg = g(i+1:N)*g(1:N-i)'/(N - i);
   Wg(i+1) = wg;
end
la = sort(abs(eig(toeplitz(Wa))));

lt = sort(abs(eig(toeplitz(Wt))));

lc = sort(abs(eig(toeplitz(Wc))));

lg = sort(abs(eig(toeplitz(Wg))));

S = la+lt+lc+lg;

plot(S)
```

% application of DFT in a sliding window

```
function y = sldfft(q,l,h);
n = length(q);
for i = 1:h:n-l
   q1 = q(i:i+l-1);
   S = indfft(q1);
   s = max(S)/mean(S);
   y(ceil(i/h)) = s;
end
```

% application of KLT as a sliding window

```
function  y = karhunen(q);
[a t c g]  = indicator34(q);

N = length(a);
for j = 1:1:N-rem(N,351)
   aa = a(j:j+350);
   cc = c(j:j+350);
   gg = g(j:j+350);
   tt = t(j:j+350);
   n = length(aa);
     for i = 0:1:174
        wa = (aa(i+1:n)*aa(1:n-i)')/(n-i);
        oa(i+1)= wa;
        wc = (cc(i+1:n)*cc(1:n-i)')/(n-i);
        oc(i+1)= wc;
        wg = (gg(i+1:n)*gg(1:n-i)')/(n-i);
        og(i+1)= wg;
        wt = (tt(i+1:n)*tt(1:n-i)')/(n-i);
        ot(i+1)= wt;
     end
    la = eig(toeplitz(oa));% la = la(1:end-1);
    lc = eig(toeplitz(oc));% lc = lc(1:end-1);
    lg = eig(toeplitz(og));% lg = lg(1:end-1);
    lt = eig(toeplitz(ot));% lt = lt(1:end-1);

    l = la + lc+ lg + lt;

    ml = max(l)/mean(l);
    L(j) = ml;

end
```

% Obtain the CDF plots for Neyman Pearson Detection

```
function [xCDF yCDF] = exoncdf(a);
[yy xx n emsg eid] = cdfcalc(a);
k = length(xx);
n = reshape(repmat(1:k,2,1),(2*k),1);
xCDF = [-Inf ; xx(n) ; Inf];
yCDF = [0 ; 0 ; yy(1+n)];


function [xCDF yCDF ] = introncdf(a);
[yy xx n emsg eid] = cdfcalc(a);
k = length(xx);
n = reshape(repmat(1:k,2,1),(2*k),1);
xCDF = [-Inf ; xx(n) ; Inf];
y1CDF = [0 ; 0 ; yy(1+n)];
yCDF = 1 - y1CDF;


function exintplot(a1,a2)
[x1 y1] = exoncdf(a1);
[x2 y2] = introncdf(a2);
figure, plot(x1,y1,x2,y2);
title('CDF of Exons and Introns')
xlabel('Base');
ylabel('Percentile');
grid on;
```

REFERENCES

[1] J.D.Watson and F.H.C Crick, "Structure of DNA", Contribution to the Discussion of Provirus, 1953.

[2] P.P. Vaidhyanathan, "Genomics and Proteomics: A Signal Processors's Tour" , IEEE Circuits and Systems Magazine, Fourth quarter, 2004.

[3] Pierre Francois Baignée, Pierre Baldi, "Flexibility of genetic code with respect to DNA structure".

[4] Paul Cristea, "Genetic Signal Analysis" (http:// www.dsp.pub.ro/articles).

[5] Ramiro Pablo Costa, "Gene Prediction Algorithms", Computational Biology Review, May 2003.

[6] M.Q. Zhang, "Discriminal analysis and its applications in DNA sequence motif recognition", Henry Stewart Publication, Briefings in Bioinformatics Vol.1, No:4, Nov 2000.

[7] J.M. Claverie, "Computational Methods for the identification of genes in vertebrate genomic sequences", Human Molecular Genetics, Vol.6, No:10, Review, pp 1735 – 1744, 1997.

[8] C. Burge and S. Karlin, " Prediction of complete gene structure in Human Genomic DNA", Journal of Molecular Biology, pp. 78- 94, Apr 1997.

[9] S. Salzberg, "Locating Protein Region in Human DNA using a decision tree algorithm", Journal of Computational Biology, Vol.2, pp 473- 485,1995.

[10] C.H. Wu, "Artificial Neural Networks for Molecular Sequence Analysis", Computer and Chemistry, Vol. 21, No:4, pp 237 – 256.

[11] Liaofu Luo , Weijiang Lee, Lijun Jia, Fengmin Ji, Lu Tsai, "Statistical Correlation of Nucleotides in a DNA sequence", Physical Review E, Vol.58, No:1, July 1998.

[12] Suryaprakash Datta , Amir Asif, " DFT based DNA splicing algorithms for prediction of protein coding regions", IEEE 2004.

[13] Gail L. Rosen and Jeffrey D. Moore, "Investigation of Coding Structure in DNA", IEEE 2003.

[14] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, "Prediction of probable genes by Fourier Analysis of genomic sequences", CABIOS, Vol.13, No:3, pp. 263 – 270, 1997.

[15] James W. Fickett, "Recognition of Protein Coding Regions in DNA sequences", Nucleic Acids Research, Vol.10, No:17, 1982.

[16] P.P Vaidyanathan, and B-J. Yoon, "Gene and exon prediction using all-pass based filters." Workshop on Genomic Sig. Proc. and Stat., Raleigh, NC, Oct.2002.

[17] C.K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciotino, M.Simons and H.E. Stanley, "Long-range correlations in nucleotide sequences", Nature, Vol.356, pp.168 – 170, March 1992.

[18] V.R. Chechetkin and A.Y.Turygin, "Size Dependence of three-periodicity and long range correlations in DNA sequences", Physical Letters A, Vol. 199, pp. 75-80,1995.

[19] B.D.Silverman, R.Linsker, "A measure of DNA periodicity", Journal of Theoretical Biology, Vol.118, pp.295 – 300, 1986.

[20] R.F. Voss, "Evolution of long range fractal correlations and 1/f noise in DNA sequences", Physical Letters,Vol.68, pp.3805- 3808.

[21] Niranjan Chakravarthy , A. Spanis, L.D. Iasemidis, K. Tsakalis, "Autoregressive Modeling and feature Analysis of DNA sequences", EURASIP Journal on Applied Signal Processing, Vol.1, pp. 13-28, 2004.

[22] Ivo Grosse, Hanspeter Herzel, Sergey V.Buldyrev, H. Eugene Stanley, "Species independence of mutual information in coding and non-coding DNA", Physical Review E, Vol.61, No:5, May 2000.

[23] Serban Mereuta, "Spectral analysis of DNA sequences – a brief review", manuscript received April 12, 2005. Work supported in part by CNCSIS under Grant 17/8/2005.

[24] D. Anastassiou, "Genomic Signal Processing", IEEE Signal Processing Magazine, pp.8-20, July 2001.

[25] A.V. Oppenheim and R.W.Schafer, *Discrete –Time Signal Processing,* Prentice Hall, Inc, NJ, 1999.

[26] A.V. Oppenheim, A.S.Willisky, and S.H.Nawab, *Signals and Systems,* Prentice Hall, Inc., 1997.

[27] Steven M.Kay, *Fundamentals of Statistical Signal Processing Volume II Detection Theory,* Prentice Hall, NJ, 1993.

[28] Ali N.Akansu, Richard A.Haddad, *Multiresolution Signal Decomposition, Transforms, Subbands and Wavelets,* Academic Press, CA, 2001.

[29] Carl W. Helstrom, *Elements of Signal Detection and Estimation,* Prentice Hall, NJ, 1995.

[30] K.R. Rao, P.Yip, *Discrete Cosine Transform Algorithms, Advantages, Applications*, Academic Press, Inc. 1990.

[31] Alberto Leon-Garcia, *Probability and Random Processes for Electrical Engineering,* Taj Press, 2004.

[32] Simon Haykin, *Adaptive Filter Theory*, Pearson Education, Inc. 2002

[33] http://genealogy.about.com/library/authors/ucroderick/a.htm *

[34] http://animal.genome.org/edu/gene/genetic-code.html *

[36] http://www.genelex.com/paternitytesting/paternityslide1.html *

[37] http://ncbi.nlm.nih.gov *

[38] http://ltspc89.epfl.ch/~vandergh/RESEARCH/DNA_analysis/dna.html *

[39] http://www-ee.uta.edu/dip/Courses/EE5356/KLT(new).pdf *

[40] http://www.emc.maricopa.edu/faculty/farabee/BioBookDNAMOLGEN.html *

[41] http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/N/Nucleotides.html *

[42] http://www.rae.org/introns.html *

[43] http://www.uky.edu/Classes/BIO/520/BIO520WWW/StudentPresent/grieser.htm *

[44] http://post.queensu.ca/~forsdyke/introns.htm *

[45] http://mathworld.wolfram.com/FredholmIntegralEquationoftheFirstKind.html *

* All web references were last accessed on 22 November, 2005.