

การวิเคราะห์การประมวลผลสัญญาณดิจิทัลของลำดับดีเอ็นเอ



นางสาวอารยา วิวัฒน์วานิช

สถาบันวิทยบริการ

จุฬาลงกรณ์มหาวิทยาลัย

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาวิทยาการคอมพิวเตอร์ ภาควิชาคณิตศาสตร์

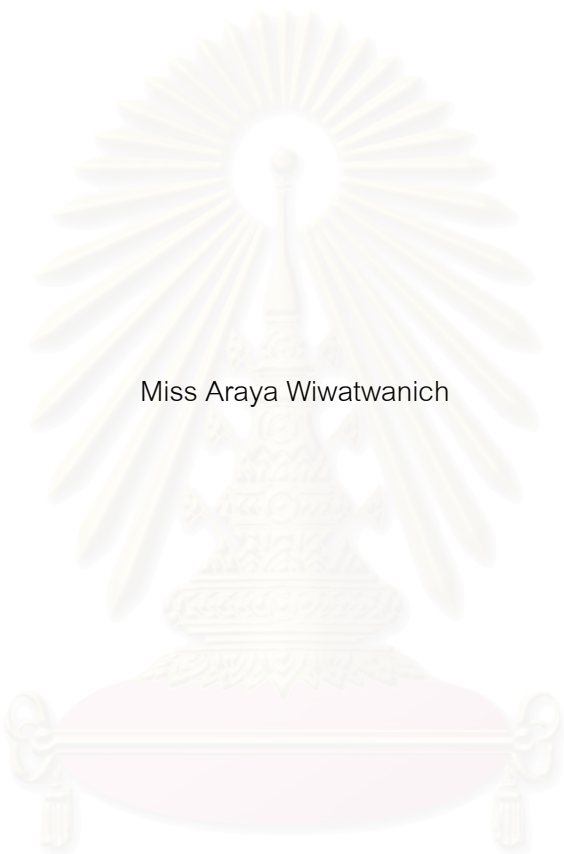
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2546

ISBN 974-17-5376-4

ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

DIGITAL SIGNAL PROCESSING ANALYSIS OF DNA SEQUENCES



Miss Araya Wiwatwanich

สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Computational Science

Department of Mathematics

Faculty of Science

Chulalongkorn University

Academic Year 2003

ISBN 974-17-5376-4

อารยา วิวัฒน์วานิช : การวิเคราะห์การประมวลผลสัญญาณดิจิทัลของลำดับดีเอ็นเอ (DIGITAL SIGNAL PROCESSING ANALYSIS OF DNA SEQUENCES) อ. ที่ปรึกษา: อ.ดร. ไพศาล นาคมหาขลาสินธุ์ อ. ที่ปรึกษาร่วม : อ.ดร. ปรีเปรม พัฒนมหกุล, อ.ดร. รัฐ พิษญาณกุล, 33 หน้า.
ISBN 974-17-5376-4

รูปพล็อตของสเปกตรัมของดีเอ็นเอสามารถใช้ทำนายตำแหน่งของเอ็กซอนได้ในระดับหนึ่ง จุดมุ่งหมายของงานวิจัยนี้คือการประยุกต์เทคนิคของการประมวลผลสัญญาณดิจิทัลเพื่อเพิ่มประสิทธิภาพในการทำนายโดยใช้ยีนจำนวนหนึ่งจาก *Caenorhabditis elegans* มาทดสอบวิธีการ

ในกระบวนการวิเคราะห์สเปกตรัม ดีเอ็นเอลำดับใดๆจะถูกแปลงไปเป็นลำดับของเลขฐานสองโดยใช้กฎการแปลง 4 แบบซึ่งขึ้นอยู่กับชนิดของนิวคลีโอไทด์, ชนิดของกรดอะมิโน, การชอบน้ำไม่ชอบน้ำของกรดอะมิโน, และ ตำแหน่งในโคดอน ตามลำดับ พบว่าการแปลงตามชนิดของนิวคลีโอไทด์เป็นการแปลงแบบเดียวที่บ่งชี้ว่าลำดับดีเอ็นเอมีค่าเป็น 3 ได้อย่างชัดเจน

สเปกตรัมที่มาจาก การแก้ปัญหาค่าสูงสุดได้ถูกปรับปรุง โดยมีการปรับเปลี่ยนตัวอย่างของเอ็กซอนและอินทรอนที่เหมาะสมเพื่อใช้ในปัญหาค่าสูงสุด พบว่าตัวอย่างดังกล่าวไม่จำเป็นต้องยาวจนเกินไป และการใช้ตัวอย่างของเอ็กซอนและอินทรอนที่มาจากยีนเดียวกันจะช่วยลดจำนวนยอดแหลมที่ไม่ต้องการในรูปพล็อตได้



สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

ภาควิชา คณิตศาสตร์
สาขาวิชา วิทยาการคอมพิวเตอร์
ปีการศึกษา 2546

ลายมือชื่อนิติกร.....
ลายมือชื่ออาจารย์ที่ปรึกษา.....
ลายมือชื่ออาจารย์ที่ปรึกษาร่วม.....
ลายมือชื่ออาจารย์ที่ปรึกษาร่วม.....

4472502023 : MAJOR COMPUTATIONAL SCIENCE
 KEYWORDS : DIGITAL SIGNAL PROCESSING/ SPECTRAL ANALYSIS /
 EXON IDENTIFICATION/

ARAYA WIWATWANICH: DIGITAL SIGNAL PROCESSING ANALYSIS OF DNA
 SEQUENCES THESIS ADVISOR: PAISAN NAKMAHACHALASINT, Ph.D.
 THESIS CO-ADVISORS: PREPRAME PATTANAMAHAKUL, Ph.D.,
 RATH PICHYANGKURA, Ph.D., 33 pp. ISBN 974-17-5376-4

DNA spectra plot can fairly indicate exon locations. The main goal of this work is to apply the Digital Signal Processing techniques to predict exon locations more accurately. A group of genes from *Caenorhabditis elegans* was used to test the methods.

In spectral analysis process, a DNA sequence was transformed to binary sequences by 4 mapping rules depending on nucleotides, amino acids, hydrophobicity, and codon bias, respectively. The triplet-periodicity discovered in exons stands out only when we used 4 binary indicator sequences corresponding to 4 nucleotides.

The optimized spectral content measure proposed by Anasstasiou was developed. Alternating samples used in the optimization problem shown that it is not necessary to use a very large sample. The attempt to discriminate between exons and introns of the same genes succeeded in reducing unwanted peaks.

สถาบันวิทยบริการ
 จุฬาลงกรณ์มหาวิทยาลัย

Department **Mathematics**

Field of study **Computational Science**

Academic year **2003**

Student's signature.....

Advisor's signature.....

Co-advisor's signature

Co-advisor's signature

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Dr. Paisan Naknahachalasint, my thesis advisor, for his invaluable guidance and encouragement in preparing and writing this thesis. I am also indebted to Dr. Preprame Pattanamahakul and Dr. Rath Pichyangkura, my thesis co-advisors, for their informations and discussions. Moreover, I wish to express my gratefulness to the other members of my committee, Professor Chidchanok Lursinsap and Dr. Chatchai Srinitiwara Wong.

My thankfulness goes to my mother for her patience and unconditional love, to Jakree Anantasirisombat for his exchange of view, to Amares Kocharat for taking care of me and my computer, and to my all dear friends for still being there for me.

Finally, I want to thank the Ministry of university affairs for the financial support given during my study.



สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

CONTENTS

	page
Abstract in Thai	iv
Abstract in English	v
Acknowledgements	vi
List of Figures	viii
I INTRODUCTION	1
1.1 Discrete Fourier Transforms	1
1.2 Digital Signal Processing	1
1.3 DSP Techniques in the Field of Biology	2
II BACKGROUND TO DNA AND PROTEIN SYNTHESIS	4
2.1 Bases	4
2.2 Nucleotide and Polynucleotide	6
2.3 The Construction of Double-Stranded Shape	7
2.4 Gene and RNA Transcription	7
2.5 Amino Acids and Translation	8
III LITERATURE REVIEW ON EXON IDENTIFICATION	10
IV EXPERIMENTAL METHODS	17
4.1 Spectral Analysis with Other Sequence Alphabets	17
4.2 Exon Prediction	25
V CONCLUSION	31
References	32
VITA	33

List of Figures

2.1	DNA double helix structure.....	4
2.2	Structure of purine and pyrimidine bases.....	5
2.3	Structure of Deoxyribose sugar.....	6
2.4	A single strand of DNA.....	6
2.5	Double strand DNA.....	7
2.6	Example of mRNA product.....	8
2.7	The genetic code.....	9
2.8	The protein synthesis process.....	9
3.1	Plot of the spectrum for exons from Y73B3A.1.....	12
3.2	Spectra of F56F11.4.....	13
3.3	Plot of $ aA + tT + cC + gG ^2$ for F56F11.4.....	15
3.4	Plots of $ W ^2$ using $a, t, c,$ and g in (3.9) for C05D9.1, D1005.1, and F13C5.2.....	16
4.1	Plot of $S_b[k]$ for exons from Y73B3A.1.....	21
4.2	Plot of $S_h[k]$ for exons from Y73B3A.1.....	21
4.3	Plot of $S_c[k]$ for exons from Y73B3A.1.....	21
4.4	Plot of $S_b[k]$ for introns from B0344.2.....	22
4.5	Plot of $S_h[k]$ for introns from B0344.2.....	22
4.6	Plot of $S_c[k]$ for introns from B0344.2.....	22
4.7	Plot of $S_d[k]$ for exons from Y73B3A.1.....	24
4.8	Plot of $S_d[k]$ for introns from B0344.2.....	24
4.9	Plots of $ W ^2$ with different sets of $a, t, c,$ and g for F56F11.4.....	26
4.10	Plots of $ W ^2$ using the best solutions for C05D9.1, D1005.1, and F13C5.2.....	28

4.11 Plots of $|W|^2$ for F59C12.2 using coefficients in (4.10)

4.12 Plots of $|W|^2$ for F59C12.2 using coefficients in (4.11)



สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

CHAPTER I

INTRODUCTION

1.1 Discrete Fourier Transforms (DFT)

The Fourier Transform is used to transform a function in a **time domain** to another function in a **frequency domain**. It describes the continuous spectrum of a non-periodic time signal [7]. In case the time variable is discrete, the input function for the transform is considered as a complex vector $[f_0, f_1, \dots, f_{N-1}]$, where the length N of the vector is a fixed parameter. The Discrete Fourier Transform of vector $[f_0, f_1, \dots, f_{N-1}]$ is also a sequence of complex numbers of the same length:

$$F_k = \sum_{n=0}^{N-1} e^{-\frac{2\pi i kn}{N}} f_n, \quad k = 0, 1, \dots, N - 1. \quad (1.1)$$

The sequence F_k provides a measure of the frequency content at frequency k .

1.2 Digital Signal Processing (DSP)

There are many types of signals in the real world, for example, human speech that is a result of the vocal cords vibration, voltages generated by the heart and brain, radars and sonar echoes, seismic vibrations and so on. These signals contain some information that needs to be extracted. However, it is almost useless when we try to analyze that information in the time domain. In contrast, the frequency domain helps us to directly examine the key information encoded in the form of frequency, phase, and amplitude of the component sinusoid [10]. Then, the Fourier Transforms become the important tools in signal processing. In particular, the

Discrete Fourier Transform plays very significant role in digital signal processing. (Note that digital signals refer to discrete time signals.)

The development of digital signal processing dates from the 1960's with the use of mainframe digital computers for number-crunching applications such as the Fast Fourier Transform (FFT), which allows the frequency spectrum of a signal to be computed rapidly. These techniques were not widely used at that time, because suitable computing equipment was available only in universities and other scientific research institutions.

DSP technology is nowadays commonplace in such devices as mobile phones, multimedia computers, video recorders, CD players, hard disc drive controllers and modems, and will soon replace analog circuitry in TV sets and telephones. An important application of DSP is in signal compression and decompression. In CD systems, for example, the music recorded on the CD is in a compressed form (to increase storage capacity) and must be decompressed for the recorded signal to be reproduced. Signal compression is used in digital cellular phones to allow a greater number of calls to be handled simultaneously within each local "cell". DSP signal compression technology allows people not only to talk to one another by telephone but also to see one another on the screens of their PCs, using small video cameras mounted on the computer monitors, with only a conventional telephone line linking them together.

1.3 DSP Techniques in the Field of Biology

Biomolecular sequences, DNA and proteins, have been represent by character strings, in which each character is in a set of the correspoding alphabets. In the case of DNA, there are 4 alphabets corresponding to 4 types of nucleotides A,T,C,and G; in the case of proteins, the size of alphabet set is 20 according to the

number of amino acid types. We can regard these genomic sequences as discrete time signals which the time domain is the position of characters on individual string. If we properly map a character string into one or more numerical sequences, then the DSP provides a set of useful tools to reveal some hidden information about life.



สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

CHAPTER II

BACKGROUND TO DNA AND PROTEIN SYNTHESIS

The structure of **DNA (deoxyribonucleic acid)** was first described by **Watson and Crick in 1953**. They found that the shape of DNA molecule is double helix, which looks like a spiral staircase (Figure 2.1). To clear up brief knowledge about DNA, it is important to inform the DNA compositions first, and then we will discuss about the process of protein synthesis.



Figure 2.1: DNA double helix structure

(<http://www.bartleby.com/61/indexillus10.html>)

2.1 Bases

There are four types of bases found in DNA: **Adenine (A)**, **Guanine (G)**, **Thymine (T)**, and **Cytosine (C)**. They can be separated into two groups by

their chemical structures. A and G represent a **purine** ring on their structures, so they are put in the purine group. And by having the **pyrimidine** ring structure, T and C lie in the pyrimidine group (Figure 2.2).

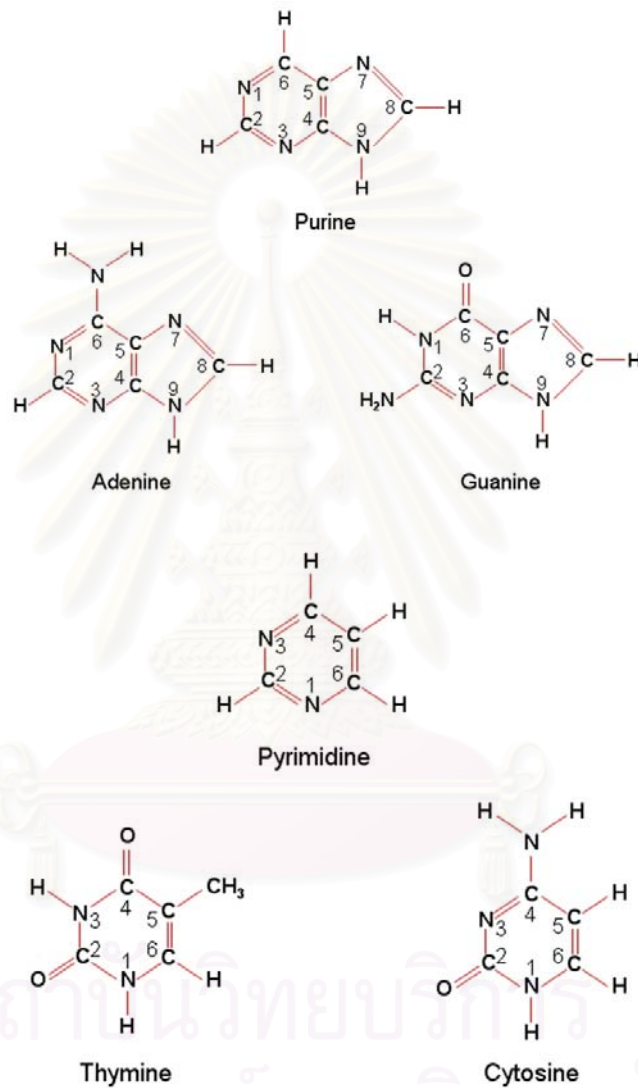


Figure 2.2: Structure of purine and pyrimidine bases

2.2 Nucleotide and Polynucleotide

Figure 2.3 displays the common structure of **deoxyribose** sugar with 5 carbon positions. When incorporated into a nucleotide, the sugar blocks further polymerization that is it contains a phosphate group on the 5' carbon and bound to a base on the 1' carbon. The next nucleotide is linked at the 3' carbon through the phosphate group (Figure 2.4).

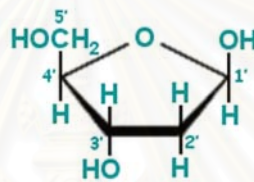


Figure 2.3: Structure of deoxyribose sugar

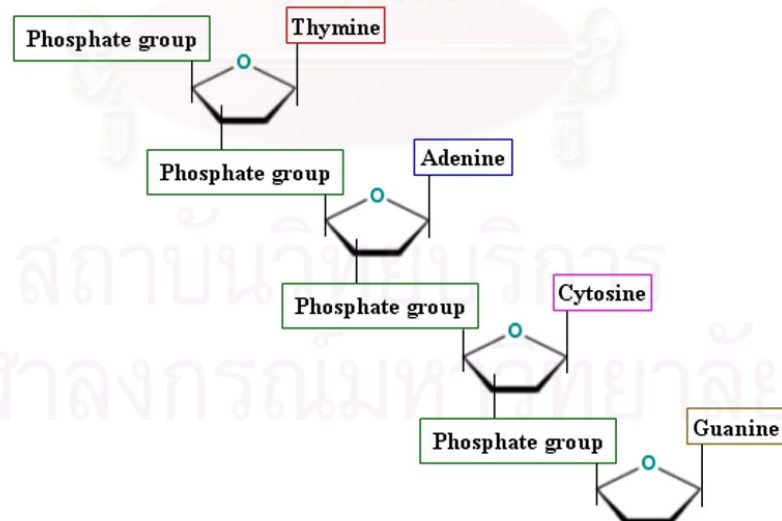


Figure 2.4: A single strand of DNA

A sequence of nucleotides is called **polynucleotide**. The polynucleotide shown

in Figure 2.4 would be written TACG, since sequences are read in the $5' \rightarrow 3'$ direction by convention.

2.3 The Construction of Double-Stranded Shape

When two polynucleotide strands of equal size are ready to form a helix, they will align in an **antiparallel** fashion. This means that one strand is oriented in the $5' \rightarrow 3'$ direction and the other in the $3' \rightarrow 5'$ direction. Later, they are held together by hydrogen bonds between individual bases. Specifically, A only pairs with T, and G only pairs with C. Then, the length of DNA is measured by the number of basepairs. Figure 2.5 exhibits an example model of 10-basepaired DNA.

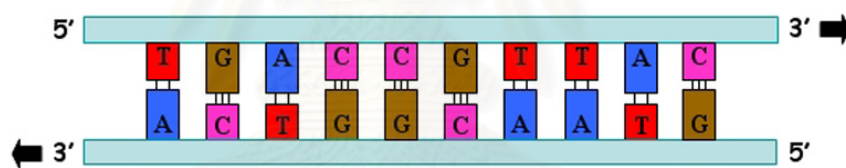


Figure 2.5: Double-stranded DNA

2.4 Gene and RNA Transcription

A **gene** is an ordered sequence of nucleotides located at a particular position on a DNA strand. Before the synthesis of a protein begins; the 2 DNA strands in a gene that codes for a protein unzip from each other. One strand of the DNA double helix is used as a template by the **RNA polymerase** to synthesize a **messenger RNA (mRNA)**. Note that RNA (Ribonucleic Acid) is a single strand of polynucleotide which contains base U (Uracil) instead of base T. Figure

2.6 shows 9 bases in a DNA sequence and their complementary mRNA bases. This process is called **transcription**.

Actually, there is a further complication to most **eukaryote** genes and a few **prokaryote** genes. Within a gene there may be several sub regions which are responsible for protein coding called **exons**. The regions between two successive exons are called **introns**. Introns are eliminated after they have directed the transcription and the remaining exons are spliced to make final mRNA (Figure 2.8).

DNA	T	C	G	A	T	A	G	C	T
mRNA	A	G	C	U	A	U	C	G	A

Figure 2.6: Example of mRNA product

2.5 Amino Acids and Translation

A sequence of bases on mRNA will be translated by **tRNA (transfer RNA)** to a sequence of **amino acids**. In the translation process, each set of 3 mRNA bases (the mRNA base triplet is called a **codon**) is a genetic code for an amino acid. Thus the number of bases on mRNA is a multiple of three. The mRNA sequence always begins with **AUG (Methionine)** and ends with a **STOP codon**. Although there are 64 possible codons, there are only 20 types of amino acid in living beings. Figure 2.7 shows the genetic code for the 20 amino acids and 3 stop codons. The protein synthesis process is visualized in Figure 2.8.

Since the protein synthesis process is performed tripletwise, there exist 3 possible reading frames per mRNA strand [1].

Example Given mRNA sequence UGGUUUGGCUCA, we can translate it to

amino acid sequence in 3 manners as

U-G-G-U-U-U-G-G-C-U-C-A-C-A

FRAME1 Trp - Phe - Gly - Ser

FRAME2 Gly - Leu - Ala - His

FRAME3 Val - Trp - Leu - Thr

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

Figure 2.7: The genetic code

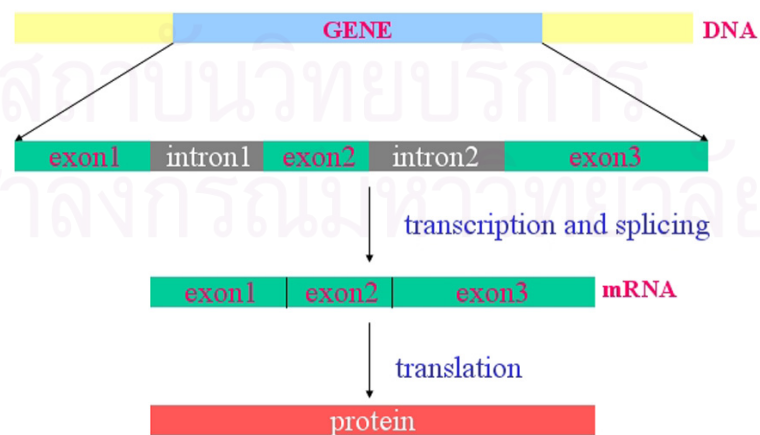


Figure 2.8: The protein synthesis process

CHAPTER III

LITERATURE REVIEW ON EXON IDENTIFICATION

For the time being, genome sequences of more than 800 organisms are either complete or being determined. Driven by this explosion of genome data, gene finding programs have also proliferated. Those programs are designed to determine where the gene starts and ends together with a specification of the alternating exons and introns (in eukaryota). Nevertheless, none of them can satisfactorily solve the problem.

Today, most of gene predictors are based on Hidden Markov Model (HMM), include GENIE (Kulp et al., 1996), GENSCAN (Burge, 1997), VEIL (Henderson et al., 1997), etc. An HMM is a stochastic model whose parameters depend on a training data. The more complicated models need the greater deal of training data. This leads to a negative aspect that gene prediction may bias toward genes with similar features to those used as training set.

In recent years, there has been another approach to the problem of gene prediction borrowed from the field of digital signal processing. Valuable information in DNA is expected to be found in the frequency domain. Thus the spectral analysis has been performed on DNA sequences to educe frequency components that may be present. Since a nucleotide sequence depicts a sequence of 4 alphabets (A, T, C, and G), applying a mathematical framework to it requires a suitable numerical mapping. For instance, the binary indicator sequences method is implemented to convert a nucleotide sequence into 4 binary sequences [13].

Given a genomic sequence $x(n)$ of length N (with position n running from 0

to $N - 1$), we form an indicator sequence $u_A(n)$ for base A such that

$$u_A(n) = \begin{cases} 1, & \text{if } x(n) = A; \\ 0, & \text{otherwise, } \quad n = 0, 1, \dots, N - 1. \end{cases} \quad (3.1)$$

This assignment is also applied to the rest bases (T, C, and G). The sequences $u_A(n)$, $u_T(n)$, $u_C(n)$, and $u_G(n)$ are called the binary indicator sequences. Let $U_A(k)$ be the **normalized** discrete Fourier transform of $u_A(n)$:

$$U_A(k) = \frac{1}{N} \sum_{n=0}^{N-1} e^{-\frac{2\pi i k n}{N}} u_A(n), \quad k = 0, 1, \dots, N - 1. \quad (3.2)$$

Thus, the sequence $U_A(k)$, $U_T(k)$, $U_C(k)$, and $U_G(k)$ provide a four-dimensional representation of the frequency spectrum of the nucleotide sequence.

From equation (3.2), it follows that

$$U_A + U_T + U_C + U_G = \begin{cases} 0, & \text{if } k \neq 0; \\ N, & \text{if } k = 0. \end{cases} \quad (3.3)$$

The total DFT power spectrum of the nucleotide sequences is the sum of the four individual spectra, namely:

$$S[k] = |U_A[k]|^2 + |U_T[k]|^2 + |U_C[k]|^2 + |U_G[k]|^2. \quad (3.4)$$

To illustrate the DFT power spectrum, we plot $S[k]$ for any interested sequences. It is known [2] that the spectrum of protein coding regions (exons) typically has a peak at the frequency $k = N/3$ corresponds to a period of three samples equals to the length of each codon. We choose coding region of Y73B3A.1 (standard name), from *Caenorhabditis elegans*, as a candidate to exhibit the period-3 characteristic. The plot is shown in Figure 3.1. The peak at the frequency $k = N/3$ is not observed in non-coding regions (introns) nor in random-synthetic sequences. This property has been used [12] to design a gene prediction algorithm.

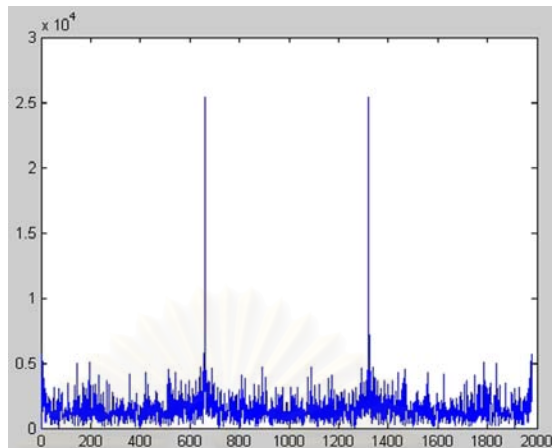


Figure 3.1: Plot of the spectrum for exons from Y73B3A.1

Given a base sequence, the regions which are likely to be exons are expected to show the period-3 peculiarity. We need the spectrogram that provide a localized measure of the frequency content. Thus, we apply the **short-time Fourier transform (STFT)** to our sequence by evaluating the DFTs on a sliding window of small length. Once we have slid the window along the full-length sequence, we obtain an array of DFT spectra whose index can be interpreted as base position.

As an example spectra, we evaluated a 351-point STFT (window size = 351) slided by 3 bases for a part of the gene F56F11.4 in *Caenorhabditis elegans*, and plot the spectra $S_n[N/3]$.

Compare the plot result in Figure 3.2 with real locations of the five exons in table 3.1, the peak due to the first exon is not dominant. Luckily, the four visible peaks can excellently indicate the rest exons.

Being based on a universal property of coding sequences, this gene-prediction methodology is independent of a training set of genes from which priors can be estimated.

In 2000, D. Anasstasiou has proposed a predictor based on DSP. He defined

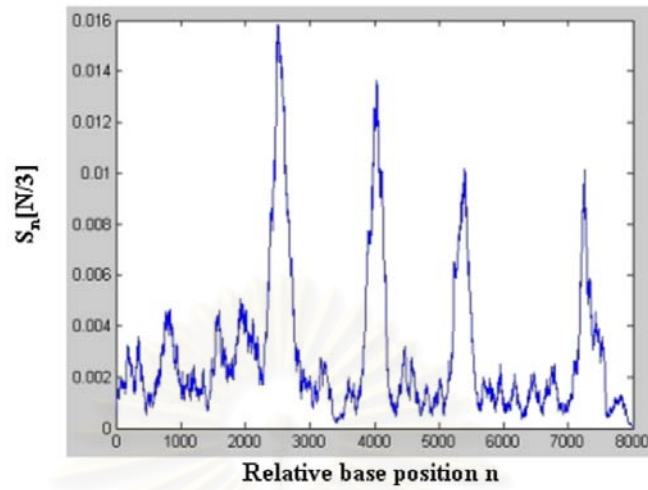


Figure 3.2: Spectra of F56F11.4

Relative position	Exon length
929-1135	207
2528-2857	330
4114-4377	264
5465-5644	180
7255-7605	351

Table 3.1: Locations of the five exons

the following normalized DFT coefficients at frequency $k = N/3$:

$$\begin{aligned}
 A &= U_A[N/3], & T &= U_T[N/3], \\
 C &= U_C[N/3], & G &= U_G[N/3],
 \end{aligned} \tag{3.5}$$

$$W = a.A + t.T + c.C + g.G.$$

Anastasiou claimed that, for properly chosen values of a , t , c , and g , the magnitude of W is a powerful predictor. Then he set up an optimization problem to find the values of a , t , c , and g that maximize the discriminatory capability be-

tween exons and random base sequence.

At the beginning of his proof, he collected all genes from chromosome XVI of *S. cerevisiae* (GENBANK accession number NC_001148), for which there were no introns and for which the evidence was labeled ‘experimental’. For each of the chosen genes, he evaluated the corresponding numbers A, T, C , and G , thus creating a set of statistical samples. Next, he generated a random base sequence, with the same number and length as the protein coding sample, thus creating a different set of random variables A_R, T_R, C_R , and G_R .

Since $A + T + C + G = 0$ (because of equation 3.3), each of the four variables is a linear combination of the other, for example, $C = -A + T + G$ which makes $W = (a - c).A + (t - c).G + (g - c).G$. Thus we can eliminate one parameter by setting its coefficients to a constant in order to define an optimization problem with a unique solution.

According to Anastasiou’s paper, he set $c = 0$ so that $W = a.A + t.T + g.G$. This reduction of dimensionality would not have enhanced predictive power. He formulated the optimization problem for finding the complex numbers a, t , and g maximizing the quantity:

$$p(a, t, g) = \frac{E \{|a.A + t.T + g.G|\} - E \{|a.A_R + t.T_R + g.G_R|\}}{std(|a.A + t.T + g.G|) + std(|a.A_R + t.T_R + g.G_R|)} \quad (3.6)$$

under the constraints (because W is also invariant to rotation and scaling):

$$E \{arg \{a.A + t.T + g.G\}\} = 0, \quad (3.7)$$

$$|a| + |t| + |g| = 1. \quad (3.8)$$

Note that E and std stand for mean and standard deviation, respectively.

Using optimization techniques based on iterated random perturbations start-

ing from an initial guess yields the solution:

$$\begin{aligned} a &= 0.10 + 0.12i, & t &= -0.30 - 0.20i, \\ c &= 0, & g &= 0.45 - 0.19i, \end{aligned} \quad (3.9)$$

corresponding to a value of $p(a, t, g) = 2.18$.

He evaluated 351-point STFT for a part of F56F11.4 (the same DNA sequence

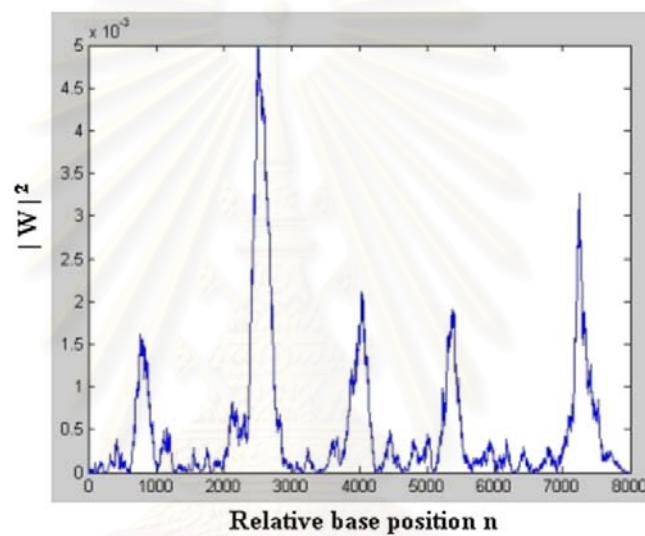


Figure 3.3: Plot of $|aA + tT + cC + gG|^2$ for F56F11.4

as in Figure 3.2), and plot the squared magnitude of W using the coefficients in (3.9). Now, in Figure 3.3 the first peak emerges at the position where the first exon located. However, the example presented in his paper seems to be the most illustrative that he was able to find. We use this coefficients to evaluate W for other genes from chromosome X of *Caenorhabditis elegans*, but the plots look unpleasant. A few examples are shown in Figure 3.4.

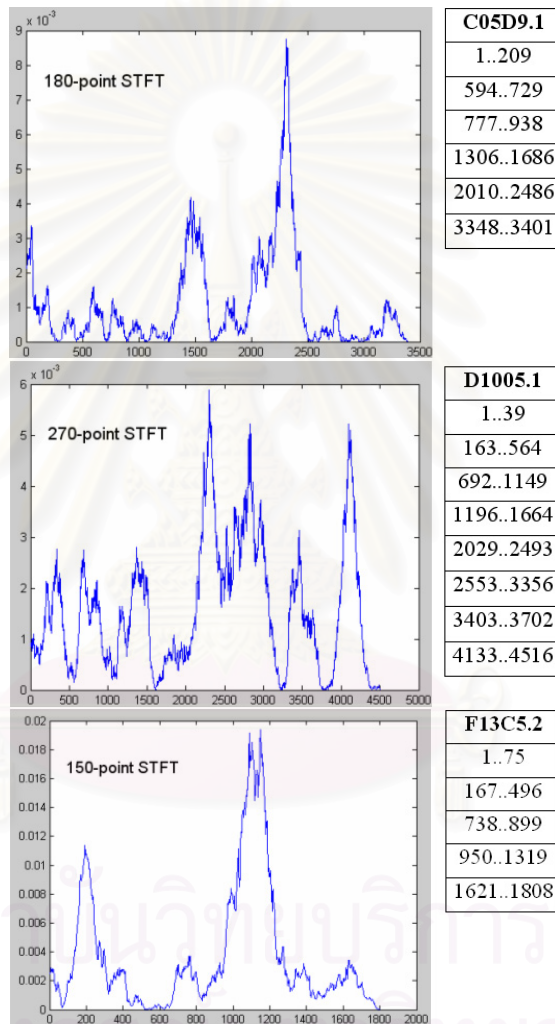


Figure 3.4: Plots of $|W|^2$ using $a, t, c,$ and g in (3.9) for C05D9.1, D1005.1, F13C5.2

CHAPTER IV

EXPERIMENTAL METHODS

4.1 Spectral Analysis with Other Sequence Alphabets

In previous study, a DNA sequence had been thought of as an ordered sequence of the 4 alphabets A, T, C, and G. Before the DSP tools can be applied, we have to map the genomic sequence to some numerical sequences. A good example is the 4 binary sequences in (3.1).

Along a DNA sequence, we can map three nucleotides at a time (like the way the protein synthesis process perform). Consequently, it becomes a sequence being comprised of 21 different residues, which come from the amino acid set {A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V, Stop}. These letters are the abbreviations for all amino acids as shown in Table 4.1.

Given a nucleotide sequence $x(n)$ of length N , we transform it into an amino acid sequence $z(m)$ of length $N/3$. We define a binary indicator sequence for amino acid A as

$$b_A(m) = \begin{cases} 1, & \text{if } z(m) = A; \\ 0, & \text{otherwise,} \end{cases} \quad m = 0, 1, \dots, N/3 - 1. \quad (4.1)$$

Binary indicator sequences for other amino acids are defined in the same manner. We called the sequences $b_A, b_R, b_N, b_D, b_C, b_Q, b_E, b_G, b_H, b_I, b_L, b_K, b_M, b_F, b_P, b_S, b_T, b_W, b_Y, b_V, b_{stop}$ the amino acid binary indicator sequences. In this work, the stop codons were neglected.

Amino acid	3-letter code	1-letter code	DNA codons
Alanine	Ala	A	GCT, GCC, GCA, GCG
Arginine	Arg	R	CGT, CGC, CGA, CGG, AGA, AGG
Asparagine	Asn	N	AAT, AAC
Aspartic acid	Asp	D	GAT, GAC
Cysteine	Cys	C	TGT, TGC
Glutamine	Gln	Q	CAA, CAG
Glutamic acid	Glu	E	GAA, GAG
Glycine	Gly	G	GGT, GGC, GGA, GGG
Histidine	His	H	CAT, CAC
Isoleucine	Ile	I	ATT, ATC, ATA
Leucine	Leu	L	CTT, CTC, CTA, TTA, TTG
Lysine	Lys	K	AAA, AAG
Methionine	Met	M	ATG
Phenylalanine	Phe	F	TTT, TTC
Proline	Pro	P	CCT, CCC, CCA, CCG
Serine	Ser	S	TCT, TCC, TCA, TCG, AGT, AGC
Threonine	Thr	T	ACT, ACC, ACA, ACG
Tryptophan	Trp	W	TGG
Tyrosine	Tyr	Y	TAT, TAC
Valine	Val	V	GTT, GTC, GTA, GTG
Stop codons	Stop	Stop	TAA, TAG, TGA

Table 4.1: Amino acid / codon mapping table

Spectral analysis is performed by taking the DFT to each of the indicator sequences. Let $\Pi = \{A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V, \text{Stop}\}$ and B_α , be the DFT of b_α , $\alpha \in \Pi$. The total DFT power spectrum of the amino acid binary indicator sequences is the sum of the 20 individual spectra, namely:

$$S_b[k] = \sum_{\alpha \in \Pi} |B_\alpha(k)|^2, \quad k = 0, 1, \dots, N/3 - 1. \quad (4.2)$$

Besides, amino acids can be classified along their hydrophobicities into two groups: hydrophobic and hydrophilic. There are 10 amino acids types classified in the hydrophobic group: F, L, I, Y, W, V, M, P, C, and A. In this work, we arranged the Stop codons in the hydrophilic group. Thus we can define another set of binary indicator sequences for an amino acid sequence by

$$h_X(m) = \begin{cases} 1, & \text{if } z(m) \text{ is an amino acid in the hydrophobic group;} \\ 0, & \text{otherwise,} \end{cases} \quad (4.3)$$

$$\text{and } h_Y(m) = 1 - h_X(m), \quad m = 0, 1, \dots, N/3 - 1.$$

Let H_X and H_Y be the DFT of h_X and h_Y , respectively. Accordingly, the total spectrum of these 2 binary indicator sequences is

$$S_h[k] = |H_X[k]|^2 + |H_Y[k]|^2, \quad k = 0, 1, \dots, N/3 - 1. \quad (4.4)$$

Other instinctive alphabets may be used to take advantage of the statistical properties for certain genome regions. For instance, coding regions may be described more completely with a 12 alphabets due to the inherent codon bias in exons. Conversely, this codon bias is not common in introns. We define the phase of position n for a nucleotide to be the number $p \equiv n \pmod{3}$, where $p \in \{0, 1, 2\}$. The alphabet taking into account phase information is thus $\Gamma = \{A_0, A_1, A_2, T_0, T_1, T_2, C_0, C_1, C_2, G_0, G_1, G_2\}$.

For example, if the original DNA sequence is C T A T G A G C C T G A G T, then the 12-letter sequence must be $C_0T_1A_2T_0G_1A_2G_0C_1C_2T_0G_1A_2G_2T_2$.

Given a length- N DNA sequence $x(n)$, define 12 different length- N indicator sequences $c_\alpha(n)$, $\alpha \in \Gamma$ by

$$c_\alpha(n) = \begin{cases} 1, & \text{if } x(n) = \alpha; \\ 0, & \text{otherwise,} \end{cases} \quad n = 0, 1, \dots, N - 1. \quad (4.5)$$

And the total spectrum can be calculated as

$$S_c[k] = \sum_{\alpha \in \Gamma} |C_\alpha(k)|^2, \quad k = 0, 1, \dots, N - 1, \quad (4.6)$$

where C_α is the DFT of c_α for all $\alpha \in \Gamma$.

We performed the spectral analysis on coding regions. In this task, we again used joined coding regions of Y73B3A.1 as the sequence of study, after we used it for displaying period-3 characteristic in Chapter 3. The plots of $S_b[k]$, $S_h[k]$, and $S_c[k]$ for it are shown in Figures 4.1, 4.2, and 4.3, respectively. The DFT spectrum plot using codon bias binary indicator sequences set produces a very strong spectral content at frequency $k = N/3$.

Next, we needed the DFT power spectra of introns to compare with those of exons. We chose noncoding regions from B0344.2 with 12596-basepair length as our example. Since we can not talk about reading frame on introns, we proceeded the work with all 3 reading frames. The first, second, and third charts in Figures 4.4, 4.5, and 4.6 are spectra for introns from B0344.2 on reading frames 1, 2, and 3, respectively.

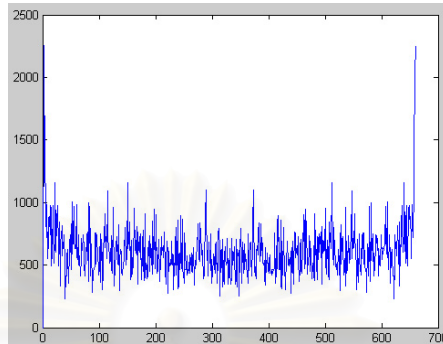


Figure 4.1: Plot of $S_b[k]$ for exons from Y73B3A.1

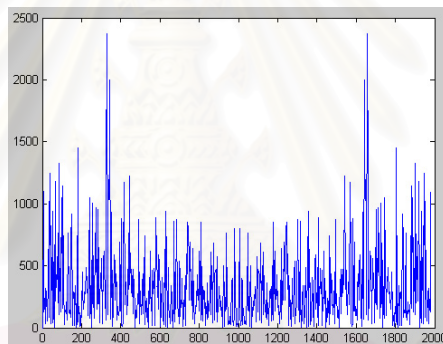


Figure 4.2: Plot of $S_h[k]$ for exons from Y73B3A.1

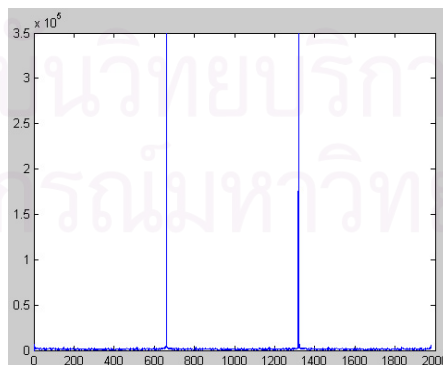


Figure 4.3: Plot of $S_c[k]$ for exons from Y73B3A.1

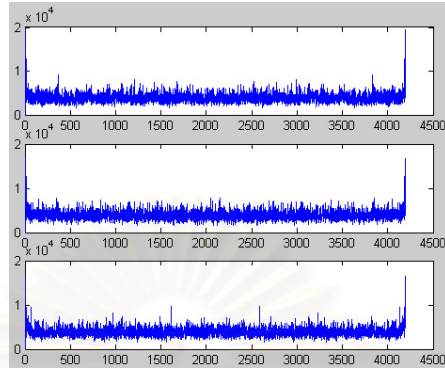


Figure 4.4: Plot of $S_b[k]$ for introns from B0344.2

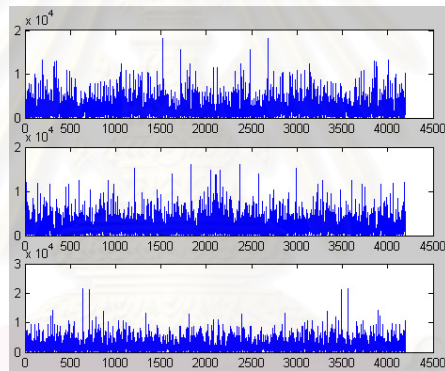


Figure 4.5: Plot of $S_h[k]$ for introns from B0344.2

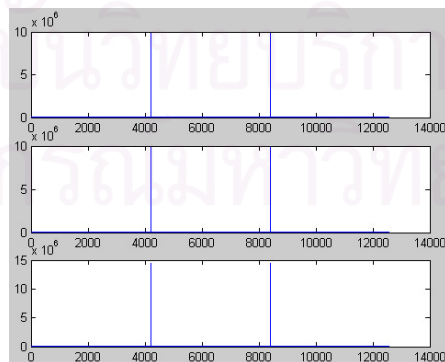


Figure 4.6: Plot of $S_c[k]$ for introns from B0344.2

Observe that the spectra for introns on 3 reading frames are not far from each other. This can be easily explained by looking at the DFT formula:

$$F_r[k] = \sum_{n=0}^{N-1} e^{-\frac{2\pi i k n}{N}} f(n), \quad k = 0, 1, \dots, N-1. \quad (4.7)$$

The subscript r of F refers to the reading frame $r \in \{1, 2, 3\}$. We can see that

$$F_2[k] = e^{\frac{2\pi i k}{N}} F_1[k], \text{ and } F_3[k] = e^{\frac{4\pi i k}{N}} F_1[k].$$

Thus changing reading frame will not alter the features of the plot.

Besides these 2 examples, we also extended this analysis on many other sequences from *C. elegans*. From those results, we found that there is no distinct characteristic to discriminate exons and introns. Notice that the spectra of both introns and exons using **codon bias** binary indicator sequences set show peaks at frequency $k = N/3$. We also found this characteristic in random sequences. This issue is an effect of the assignment of binary values to 12 alphabets in regard to the codon bias. Equation (4.5) forces all 12 binary indicator sequences to have the period of 3. So, we regarded a given DNA sequence of length N as a codon sequence of length $N/3$, and redefined the binary indicator sequence for $\alpha_i \in \Gamma$ by

$$d_{\alpha_i}(m) = \begin{cases} 1, & \text{if base in phase } i \text{ of codon } m \text{ is } \alpha; \\ 0, & \text{otherwise,} \end{cases} \quad (4.8)$$

$$m = 0, 1, \dots, N/3 - 1.$$

Example Given sequence C T A T G A G C C T G A G T C, we have

$$d_{T_0} = \{0, 1, 0, 1, 0\}$$

$$d_{T_1} = \{1, 0, 0, 0, 1\}$$

$$d_{T_2} = \{0, 0, 0, 0, 0\}.$$

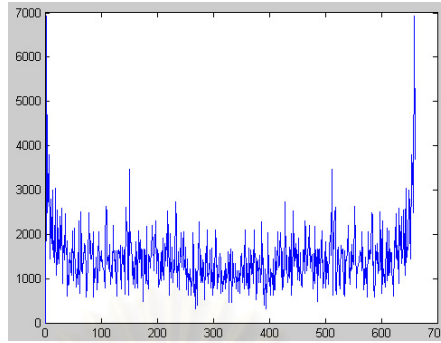


Figure 4.7: Plot of $S_d[k]$ for exons from Y73B3A.1

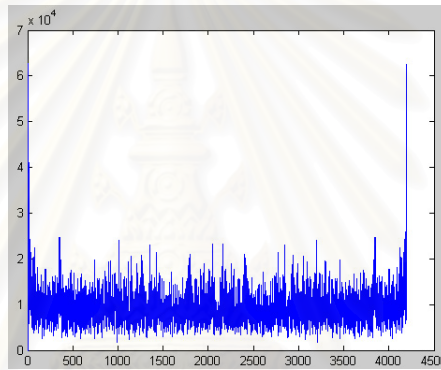


Figure 4.8: Plot of $S_d[k]$ for introns from B0344.2

After we took the DFT to each of the indicator sequences, we received 12 complex sequences: $\{D_{\alpha_i} \mid \alpha_i \in \Gamma\}$. The total spectral content of the 12-letter signal is

$$S_d[k] = \sum_{\alpha_i \in \Gamma} |D_{\alpha_i}(k)|^2, \quad k = 0, 1, \dots, N/3 - 1. \quad (4.9)$$

We calculated DFT power spectra for many exons and many introns for comparison. Still we cannot find any attribute to separate exons from introns. We put the plots for exons from Y73B3A.1 and introns from B0344.2 (reading frame 1) on display (Figures 4.7 and 4.8).

The spectral analysis is not successful in observing any latent periodicity in exons by the methods developed. The triplet-periodicity discovered in exons

stands out only when we view a DNA sequence as a 4-nucleotide string.

4.2 Exon Prediction

As we see in Chapter 3, W is still not a satisfactory predictor. The goal of this work is to find some reasons for its weakness and try to improve its efficacy. We started by following Anastassiou's method.

Our first duty was to collect all single-exon genes with 'experimental evidence' from chromosome XVI of *S. cerevisiae* (NC_001148), the same sample data as his. The genome of *S. cerevisiae* is now complete in GENBANK. Most of genes in chromosome XVI are intronless and convenient to collect. This may be a reason for Anastassiou to choose this data as a statistical sample.

The method was slightly changed and there was the slightest effect to the statistical properties. We joined the chosen genes together getting the sequence of about 480,000 basepairs. Then we evaluated the 1500-point STFT with no overlap for this protein coding sample, creating the vectors A, T, C , and G . We generated a random base sequence with the same length as the sample, and computed another set of random variables: A_R, T_R, C_R , and G_R .

We implemented the method by MATLAB, a useful mathematical tool. The optimization problem can be solved by an iterative technique available in MATLAB toolbox. Certainly, the best choice of the initial values of a, t, c , and g must be equations (3.9). The solution thus depends on the seed of the random number generator. For different seeds, the obtained random sequences will not repeat. We solved the problem many times with various random sequences, the solutions varied a little. Then we evaluated the 351-point STFT sliding the window by 3 bases for the same DNA sequence as in Figure 3.2 (F56F11.4). Using those solutions as the coefficients sets, we plot the squared magnitudes of $aA + tT + cC + gG$. Some

examples are shown in Figure 4.9.

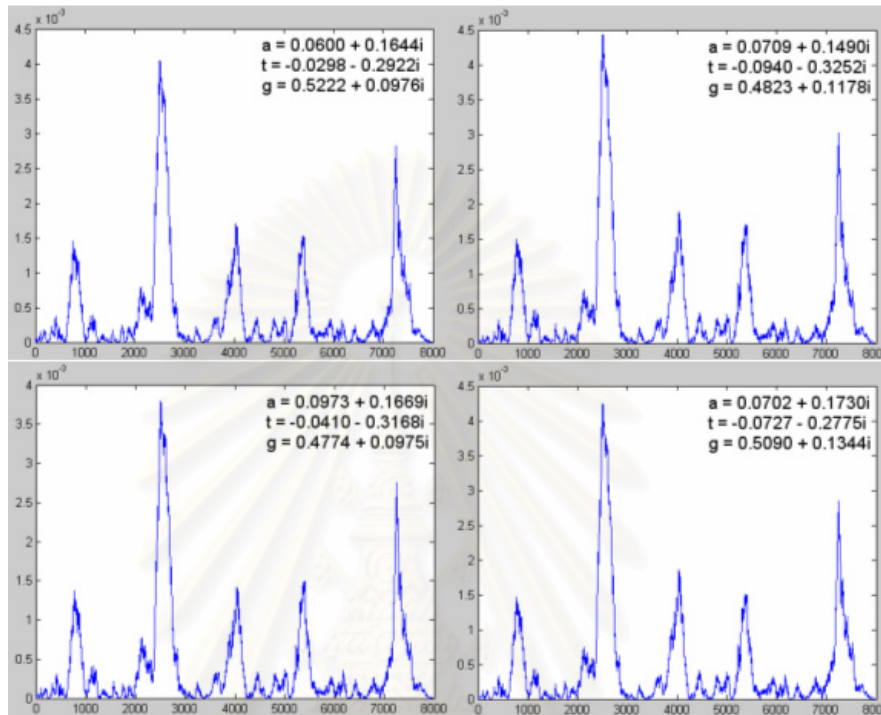


Figure 4.9: Plots of $|W|^2$ with different sets of $a, t, c,$ and g for F56F11.4

As we see in Figure 4.9, the difference of the seeds does not affect the feature of the plots. Recall that this example seems to be the most illustrative that we was able to find, it is not strange if there are many sets of coefficients that can apply to this genome.

Further assumption is that the values of the coefficients may change for different genomes. The statistical properties of coding regions from chromosome XVI of *S.cerevisiae* may not be able to explain the statistical properties of genomes from other organism.

In this work, we decided to choose some genes from chromosome X of *C. elegans* to test our methods. Hence we ought to use coding regions from this chromosome as our sample. Although we could not find any coding region for which the evi-

dence was labeled ‘experimental’ and intronless, we first tried to select exons from F57C12.4. We joined them together and got a sequence of only 4578 basepairs long.

We adjusted the method by sliding the window of STFTs by 3 bases to get the random variables A , T , C , and G . By the same method, we received A_R , T_R , C_R , and G_R for a synthesized random sequence. We had an idea that the size of the window may cause different results. So, we varied the window size from 150 to 360 basepairs with the increment of 30. Next, we solved the optimization problem yielding a set of solutions. The best features of our examples: C05D9.1, D1005.1, and F13C5.2 are shown in Figure 4.10.

This results look scarcely better compared to Figure 3.4. We changed the sample from coding regions of F57C12.4 to those of ZK1193.2 (3,753 basepairs), C33E10.6 (3,114 basepairs), and T27B1.1 (5,241 basepairs), respectively, but the performances look alike. However, this task lets us know that using a very large sample is not necessary.

The optimization problem should be set up to find the values of a , t , c , and g that maximize the discriminatory capability between exons and introns. In most genes, their joined introns are longer than their joined exons. We needed intron samples and exon samples to be the same length. Thus it is doubtful which part of joined introns should be chosen to compare with joined exons. This might become the reason for using random sequence instead of introns. But the proof that the order of bases in introns are random is still unclear. We should try to find the proper noncoding regions for the optimization problem.

We started by choosing many genes from *C. elegans* which the total lengths of their exons and introns are nearly equal. For individual genes, their exons and introns were splitted from each other. We cut the rear part of the longer sequence

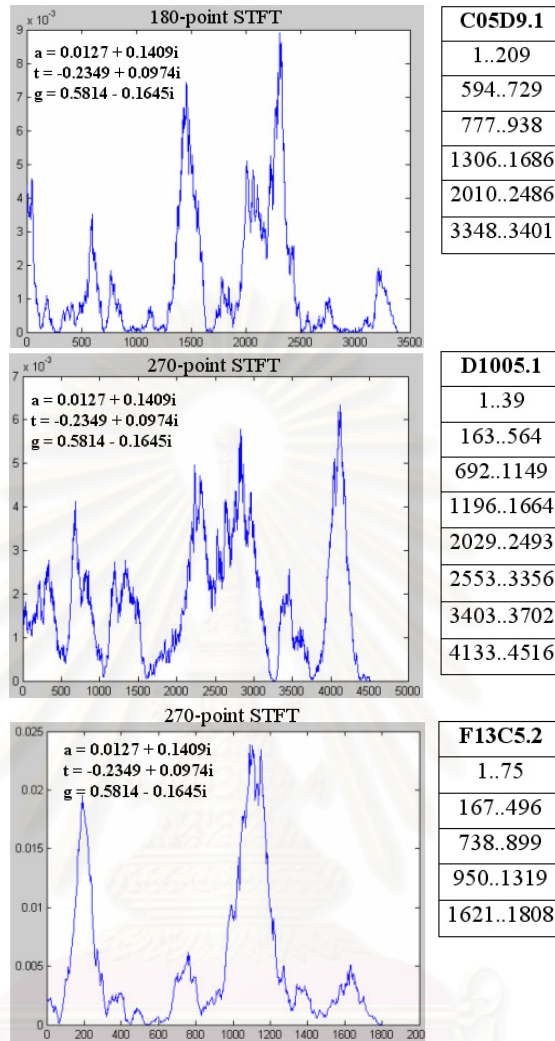


Figure 4.10: Plots of $|W|^2$ using the best solutions for C05D9.1, D1005.1, and F13C5.2

off. Then we used all joined exons to create A, T, C, G , and all joined introns to create A_R, T_R, C_R, G_R for the optimization problem.

The solution is

$$\begin{aligned}
 a &= 0.4587 - 0.0159i, & t &= 0.1433 + 0.0798i, \\
 c &= 0, & g &= 0.3674 - 0.1749i,
 \end{aligned} \tag{4.10}$$

which makes the squared magnitude plot of W for F59C12.2 looks preferable (Figure 4.11).

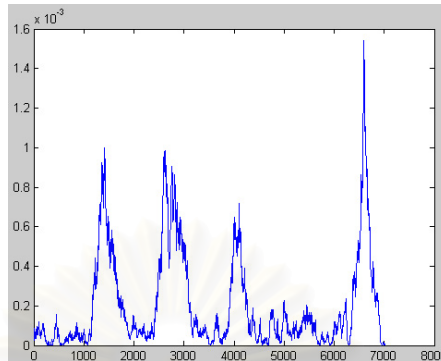


Figure 4.11: Plots of $|W|^2$ for F59C12.2 using coefficients in (4.10)

The variables received from exons of F57C12.4 and a random sequence give a solution

$$\begin{aligned} a &= -0.1317 + 0.1102i, & t &= -0.5012 + 0.1030i, \\ c &= 0, & g &= 0.2960 - 0.1125i, \end{aligned} \quad (4.11)$$

which makes the inferior plot. (Figure 4.12)

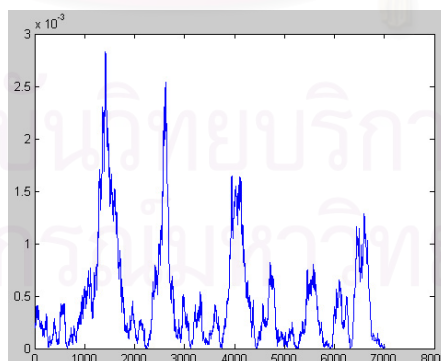


Figure 4.12: Plots of $|W|^2$ for F59C12.2 using coefficients in (4.11)

It is obvious that some unwanted peaks in Figure 4.12 are invisible in Figure 4.11. This advantage occurs when we use only 4,697 basepairs length of coding

and noncoding regions. With the solution in (4.10), we can reduce some unwanted peaks in several genes of our sample such as B0310.2, ZK1193.4, and F19G12.5. We believed that there are many genes outside our sample which give superior solution. That solution is expected to decrease the percentage of mismatched peaks.



สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

CHAPTER V

CONCLUSION

After we labored with many experiments, we realized that there is no measure that can indicate locations of exons exactly even in genomes of our case study organism, *C.elegans*. The prediction will become further complicate in advanced organism. However, the methods presented here shows fairly good performance to roughly forecast where the exons are. The exact locations must be obtained by biological experiment. It has been emphasized that the gene sample available in current databases is perhaps atypical, and this can affect the performance of gene predictors.

What we need further is to develop an algorithm that can identify short exons, and to predict very long exons, more accurately. Mention here that the prediction of exons locations is not a simple task, even the most elaborated methods sometimes fail to detect coding regions. In general a variety of methods are used in conjunction in a way that they complement to one another to obtain better results.

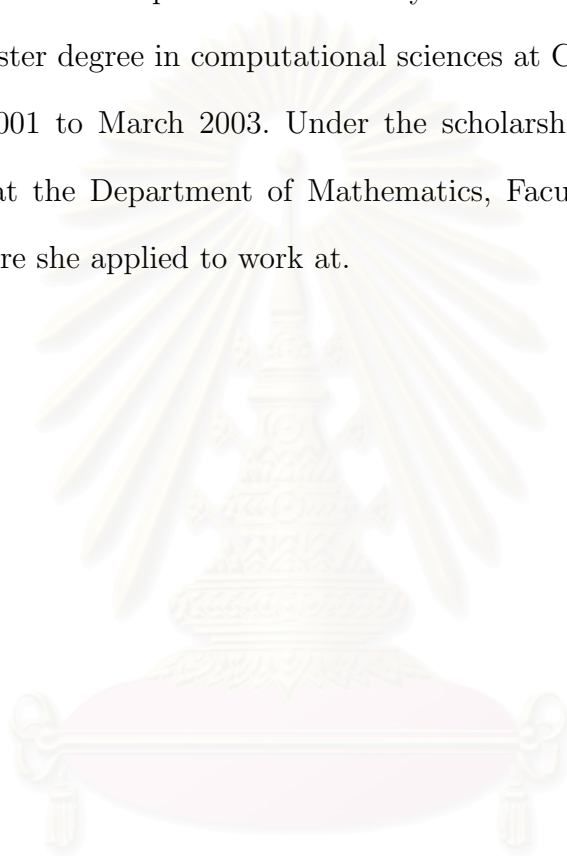
There are many future challenging problems remain to be solved in the gene-prediction field, such as to predict gens that encode non-coding RNAs, to predict polyadenylation sites, to predict replication origins, and so on.

Bibliography

- [1] Anastassiou,D. Frequency-domain analysis of biomolecular sequences. *Bioinformatics* **16(12)** (2000): 1073-1082.
- [2] Chechetkin,V.R. and Turygin,A.Y. Size-dependence of three-periodicity and long-range correlations in DNA sequences. *Physics Letters A* **199** (1995): 75-80.
- [3] Claverie,J.M. Computational methods for the identification of genes in vertebrate genomic sequences. *Oxford University Press* **6(10)** (1997): 1735-1744.
- [4] Finney,R.L., Weir,M.D., and Giordano,F.R. *Thomas' Calculus*. 10th ed. New York: Addison-Wesley Publishing, 2001.
- [5] Hanson,R.W. *Fast Fourier transform analysis of DNA sequences*. Bachelor of Arts, Reed College, Oregon, 2003.
- [6] Henderson,J., Salzberg,S., and Fasman,K.H. Finding genes in DNA with a Hidden Markov Model. *Journal of Computational Biology* **4(2)** (1997): 127-142.
- [7] Press,W.H., Teukolsky,S.A., Vetterling,W.T., and Flannery,B.P. *Numerical recipes in Fortran 77: the art of scientific computing*. 2nd ed. New York: Cambridge University Press 1986, 2001.
- [8] Salzberg,S., Delcher,A.L., Fasman,K.H., and Henderson,J. Finding genes in DNA with a Hidden Markov Model. *Journal of Computational Biology* **5(4)** (1998): 667-680.
- [9] Sinden,R.R. *DNA structure and function*. San Diego: Academic Press, 1994.
- [10] Smith,S.W. *Spectral Analysis*. <http://www.dspguide.com/specanal.htm>: California Technical Publishing, 2003.
- [11] Sneddon,I.N. *The use of integral transforms*. New York: McGraw-Hill, 1972.
- [12] Tiwari,S., Ramachandran,S., Bhattacharya,A., Bhattacharya,S., and Ramawamy,R. Prediction of probable genes by Fourier analysis of genomic sequences. *CABIOS* **13(3)** (1997): 263-270.
- [13] Voss,R. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Physical Review Letters* **68** (1992): 3805-3808.

VITA

Miss. Araya Wiwatwanich was born on April 24, 1979 in Chonburi. She graduated with a Bachelor degree in mathematics from Burapha University in 2001. She got a scholarship from the Ministry of university affairs to her further study for a master degree in computational sciences at Chulalongkorn University during June 2001 to March 2003. Under the scholarship requirement, she will be a lecturer at the Department of Mathematics, Faculty of Science, Burapha University where she applied to work at.



สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย