CHAPTER 2

LITERATURE REVIEW

There are many attempts to study the hidden information in biological sequences. One field of the study is to know if there exists some structure or construction rule controlling the arrangement of nucleotides in DNA sequences.

In early 1970s, A. Tsonis, P. Kumar, J.B. Elsner and P.A. Tsonis [3] found the periodicities of three in coding sequences of DNA and periodicities of variable length in non-coding sequences by using Fourier analysis and dot matrix analysis. When they applied wavelet transform with Morlet mother wavelet to the DNA sequences, they found that there are two construction patterns in DNA, indicating coding sequences and non-coding sequences. These two patterns in some organism cannot be found by using the Fourier analysis. However, by using dot matrix analysis, this evidence sometimes can be found. These periodicities are later found in DNA sequences of *Escherichia coli* by using the Fourier transform, by A. Fukushima, M. Kinouchi, Y. Kudo, S. Kanaya, H. Mori and T. Ikemura [4] in 2001.

In 1998, S.A. Buldyrev, N.V. Dokholyan, A.L. Goldberger, S. Havlen, C.K. Peng, H.E. Stanley and G.M. Viswanathan [5] supported the idea that non-coding sequences in gene are correlated. They mapped DNA string into numerical string by helps of random walk. They found that the coding sequences are less correlated than non-coding sequences and the correlation is remarkably long-range.

In 1998, A. Arnedo, Y. D'Aubenton-Carafa, B. Audit, E. Bacry, J.F. Muzy and C. Thermes [6] clearly identified the long-range correlation in non-coding sequences by using wavelet transform. Complex gaussian function was chosen as a mother wavelet in order to blind linear trends of the sequences. Their results are strongly visualized corresponding to the slopes of the trends.

In 2001 based on their previous work [8], R.P. Costa and D. Anastassiou [7] applied digital signal processing technique to the field of biology in order to implement a set of functions, to analyze interesting frequency properties of DNA. His goal is to predict location of coding regions within a portion of DNA sequences.

Their work can be separated into two main procedures. They first mapped nucleotide sequences into numerical sequences, and then performed the digital signal processing with the numerical sequences.

Mapping Bio-molecular Sequence into Numerical Sequence

Because the method of digital signal processing on the bio-molecular sequence, the nucleotide sequence has to be mapped into numerical sequence in order to mathematically process that information. There are several possible mapping methods available, but an easy mapping rule is presented here.

The number *a*, *t*, *g* and *c* were assigned for mapping the character A, T, G, and C respectively. Then, a DNA sequence of length N can be presented as

$$x(n) = au_A(n) + tu_T(n) + gu_G(n) + cu_C(n), \tag{2-1}$$

where

n = 0, 1, 2, 3, ..., N-1.

$\mathcal{U}_X(n)$ represents the binary (0 and 1) indicator function for the corresponding nucleotide. It takes the value 1 at index n if the corresponding nucleotide is present at that position, and takes the value 0 if the corresponding nucleotide is absent at that position.

$$\mathcal{U}_X(n) = \begin{cases} 1 & \text{if the nucleotide at position } n \text{ is X} \\ 0 & \text{if the nucleotide at position } n \text{ is not X} \end{cases} \tag{2-2}$$

For the reason that at one position, there is only one type of nucleotide presents. The expression becomes

$$\mathcal{U}_A(n) + \mathcal{U}_T(n) + \mathcal{U}_G(n) + \mathcal{U}_C(n) = 1, \tag{2-3}$$

where

n = 0, 1, 2, 3, ..., N-1.

After mapping these bio-molecular sequence into numerical sequence, digital signal processing can be applied with that sequence.

Digital Signal Processing of Numerical Sequence

Taking the Fourier transform to Equation 2-1. *X(k)* is defined as

$$X(k) = aU_A(k) + tU_T(k) + gU_G(k) + cU_C(k), \quad (2\text{-}4)$$

where

k = 0, 1, 2, 3, ..., N-1.

The sequences $U_X(k)$ represent the frequency information of each nucleotide. By combining all four terms of nucleotides and taking its spectrum properties, the spectrum is defined.

$$S(k) = \left|U_A(k)\right|^2 + \left|U_T(k)\right|^2 + \left|U_G(k)\right|^2 + \left|U_C(k)\right|^2 \quad (2\text{-}5)$$

The Equation 2-5 is the measurement of the total spectral information of the DNA sequence at frequency *k*. Then, this frequency analysis method was applied into an exon. He set *a*=1, *t*=1, *g*=1 and *c*=1 for the mapping. The result is shown in Figure 2.1.

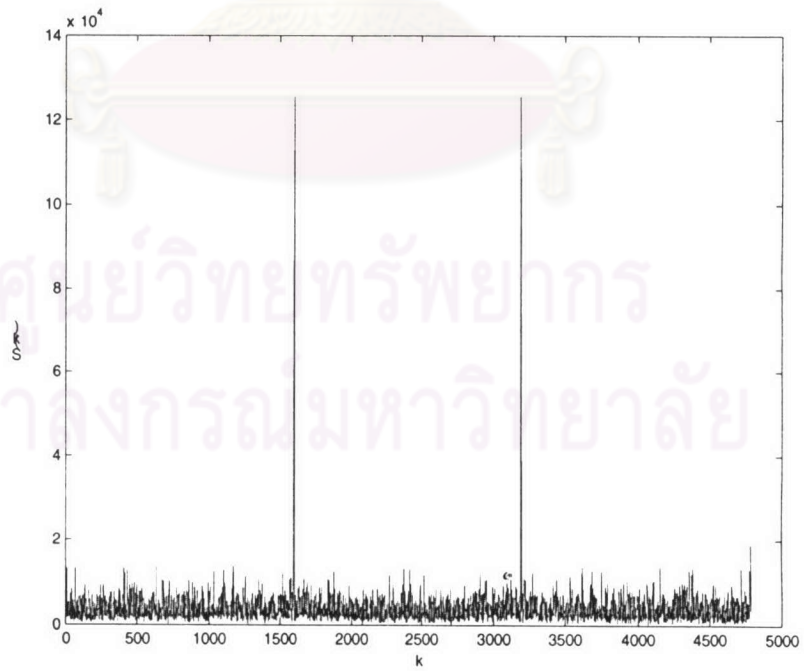

Figure 2.1 Fourier transform of an exon

Because of its symmetrical structure, the first half of Figure 2.1 is shown in Figure 2.2.

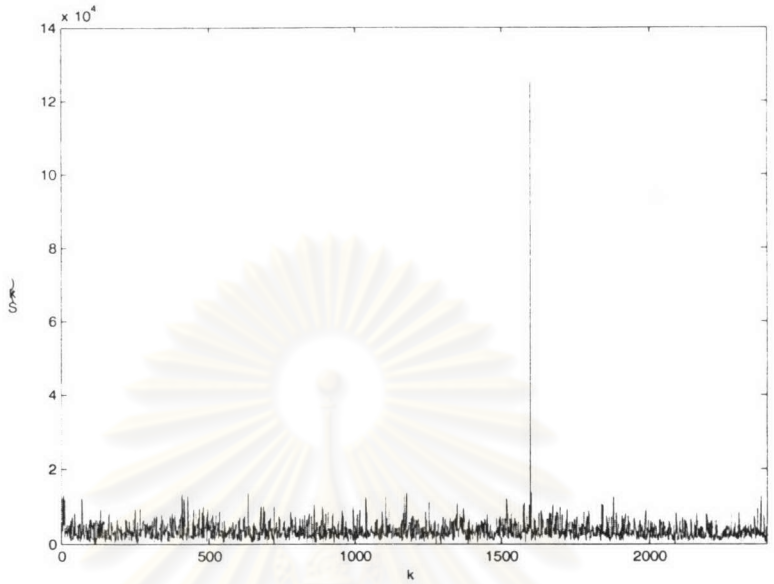

Figure 2.2 The first half figure of Figure 2.1 of an exon

He found that the value of *S(k)* is maximum at the particular frequency $k=N/3$. This corresponds to a period of three, refers to the length of each codon which composed of three nucleotides. It also has been shown that a protein coding region in DNA typically has a peak at the frequency of $N/3$. The specific frequency of exon shown in Figure 2.1 and 3.14 is enlarged and presented in Figure 2.3.

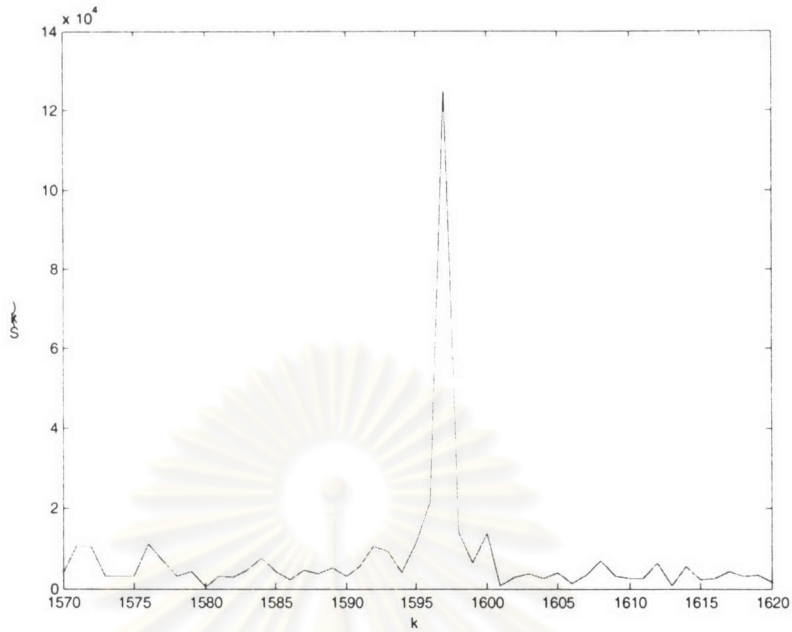

Figure 2.3 A peak of Figure 2.2

This result indicates that the hidden information located at the frequency $k=N/3$ corresponds to the triplet which leads to the amino acid synthesis. Therefore, the frequency response at the frequency $k=N/3$ is used for the following steps.

Then, the frequency response of that exon will be normalized by $1/N$. He substitutes the normalized frequency response of nucleotide A, T, G and C by alphabets A, T, G and C respectively.

$$A = \frac{1}{N} U_A\left(\frac{N}{3}\right)$$

$$T = \frac{1}{N} U_T\left(\frac{N}{3}\right)$$

$$G = \frac{1}{N} U_G\left(\frac{N}{3}\right) \qquad (2\text{-}6)$$

$$C = \frac{1}{N} U_C\left(\frac{N}{3}\right)$$

And from Equation 2-4, he got

$$X = aA + tT + gG + cC \qquad (2\text{-}7)$$

By means of random number generator, a random sequence with uniform distribution of each nucleotide was created.

On this assumption, the pure exon is resembled to the random sequence. In other words, if DNA sequence can give the zero, or in fact the minimum of discriminatory compatibility between pure random sequence and DNA itself, that DNA sequence is considered as an exon. However if DNA sequence gives the maximum difference between pure random sequence and DNA itself, that DNA sequence is considered as an intron.

The next step is to maximize the discrimination between protein coding regions with respective random variable sequence of each nucleotide. He synthesized a random DNA sequence of the same length of the first sample. Afterward the generated sequence will be applied with the same process as the exon.

From Equation 2-3, he got

$$A+T+G+C=\begin{cases}0, & k\neq 0\\ N, & k=0\end{cases} \tag{2-8}$$

If any constant value is added to *a*, *t*, *g* and *c* then the value of *X* in Equation 2-4 and *S* in Equation 2-5 does not change. The sequences of each nucleotide represent a redundant set; therefore one of the four coefficients can be set to be zero. In this case, he set the coefficient c=0.

$$\begin{aligned} c &= 0 \\ X &= aA+tT+gG \end{aligned} \tag{2-9}$$

The maximization problem is to find the complex numbers for the coefficients *a*, *t* and g in order to make the Equation 2-10 maximum i.e.

$$f(a,t,g)=\frac{E\{|aA+tT+gG|\}-E\{|a\,A_R+t\,T_R+g\,G_R|\}}{std\{|aA+tT+gG|\}+std\{|a\,A_R+t\,T_R+g\,G_R|\}} \tag{2-10}$$

Where $E\{x\}$ represents the statistical mean of $\{x\}$ and $std\{x\}$ represents the statistical standard deviation of $\{x\}$. A_R, T_R and G_R are transformed sequences of random sequences.

Because *X* is also invariant to rotation and scaling, the maximization undergoes with two constraints in Equation 2-11 and 2-12.

$$E\{\arg(aA + tT + gG)\} = 0 \tag{2-11}$$

$$|a| + |t| + |g| = 1 \tag{2-12}$$

The problem of maximization is readily solved by helps of Matlab with optimization toolbox. The result of these coefficients (*a*, *t* and *g*) affects the comparison of random sequence and the DNA sequence. The discrimination will be magnified with these coefficients. The difference between location of exon and intron will be overstated to distinguish by far. The magnitude of *S(k)* at the particular frequency *k=N/3*, defined in Equation 2-5, guides where exons and introns located in nucleotide sequence.

A gene of *Caenorhabditis elegans* of length 8000 base pairs is performed by this method. The calculation of Fourier transform is windowed. The window of calculation is 351 base pairs in order to match the maximum length of exon on that DNA. The result is shown in Figure 2.4.

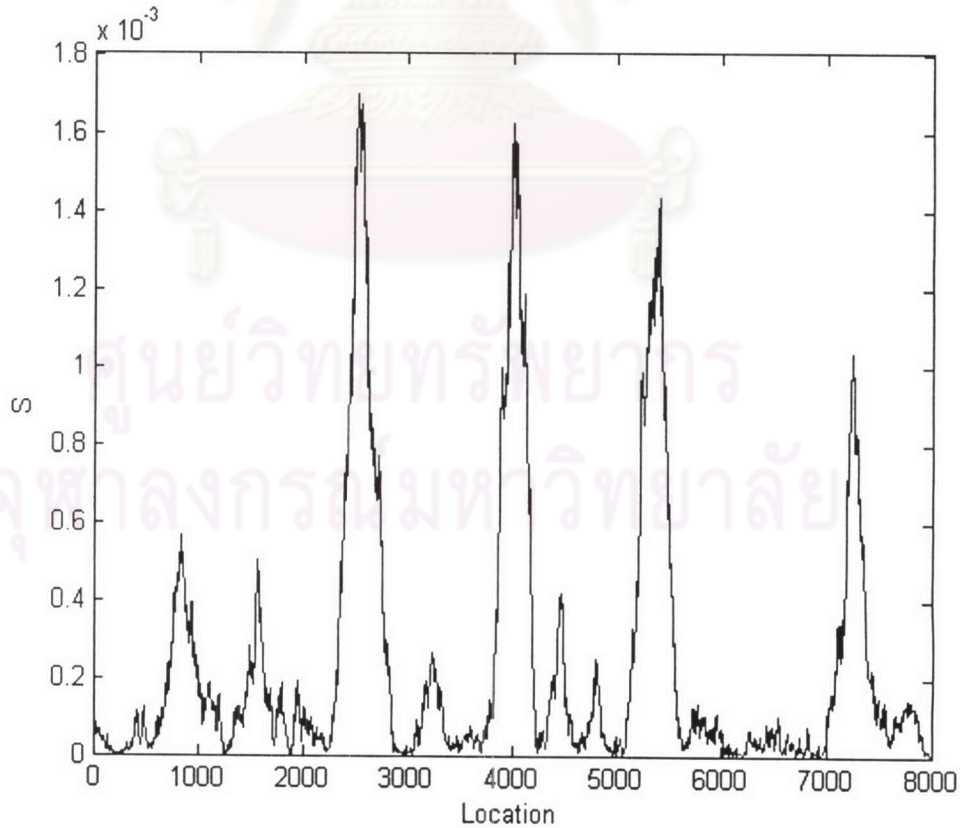

Figure 2.4 Spectrum plot of *Caenorhabditis elegans* by Fourier transform

The actual location of five exons on gene *C. elegans* is shown in Table 2.1.

Table 2.1 Locations of the five exons on the *Caenorthabditis elegans*

Position	Exon Length
929-1135	207
2528-2857	330
4114-4377	264
5465-5644	180
7255-7605	351

Locations of the five peaks of spectrum shown in Figure 2.4 match locations on exons in Table 2.1. The prediction of locations of exons and introns on gene *C. elegans* by this method is precisely correct.