CHAPTER 4

EXPERIMENTAL RESULTS

4.1 DNA Spectra

We start taking Fourier transform into any exon to see the peak exists at frequency $k=N/3$ in its spectrum plot. The length of the exon of *Drosophila melanogaster* (GENBANK, accession number NC206656) is 2558 base pair. Next, we converse this exon into 4 numerical sequences by using Anastassiou's mapping scheme. The spectrum plot by using Fourier transform of this exon is shown in Figure 4.1.

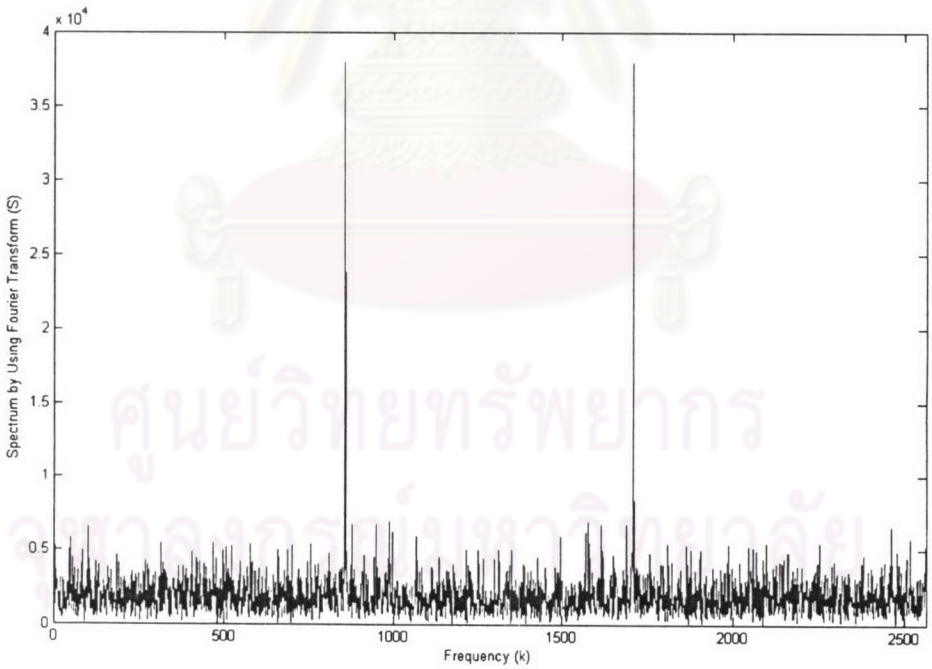

Figure 4.1 Fourier transform of exon of *Drosophila melanogaster*

It was shown that two peaks at $k=N/3$ is clearly noticed. This corresponds to a period three, refers to the length of each codon which composed of three nucleotides. Then we use wavelet transform with Morlet function instead of Fourier transform and plot its spectrum. The result is shown in Figure 4.2.

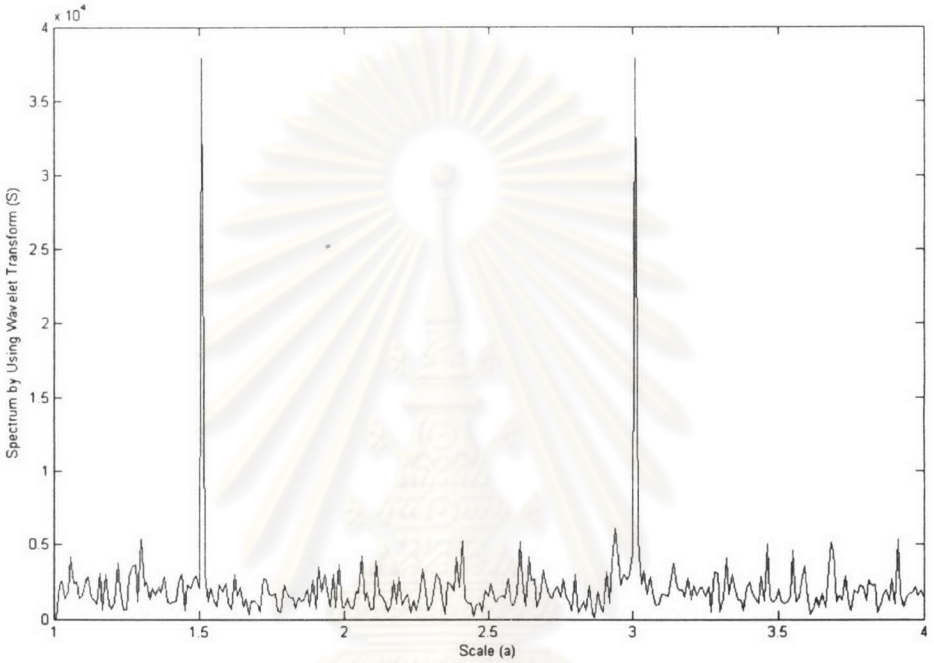

Figure 4.2 Wavelet transform of exon of *Drosophila melanogaster*

Noticeably, there are two peaks in the Figure 4.1 and 4.2. The result produced by using wavelet transform is resembled to the result obtained by Fourier transform. We zoom the Figure 4.2 at the second peak and show in Figure 4.3.

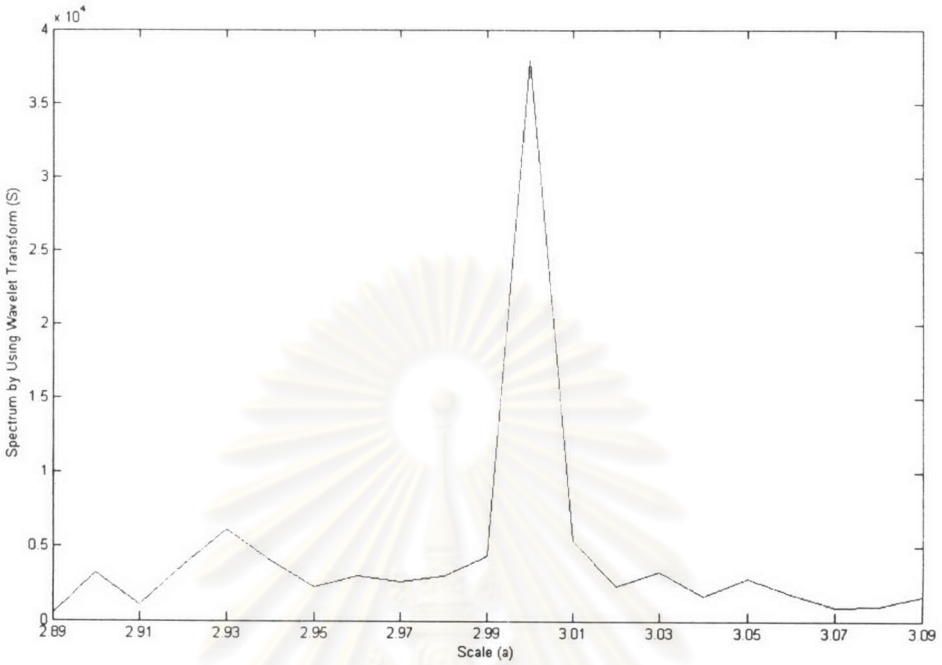

Figure 4.3 A peak of wavelet transform of exon of *Drosophila melanogaster*

Then, we consider the non-coding regions, called introns. All nucleotides in non-coding regions do not produce any amino acid. Therefore, we assume there are periodicities of three in non-coding regions in comparison to the periodicities in coding regions. To prove this fact, both Fourier transform and continuous wavelet transform are applied to any intron of *D. melanogaster* (GENBANK, accession number NC206698). The result of the Fourier transform is shown in Figure 4.4, and the result of the continuous wavelet analysis is shown in Figure 4.5. We can see from both figures that there is no peak at $k=N/3$ for Fourier transform and $a=3$ for continuous wavelet transform. This fact matches to the laboratory result and analysis in biology, which states that there is no correlation in any intron, while the long-range correlation is found in the exon [13] [14] [15]. For the work in this thesis, we will use $a=3$ for wavelet analysis to distinguish between exon and intron.

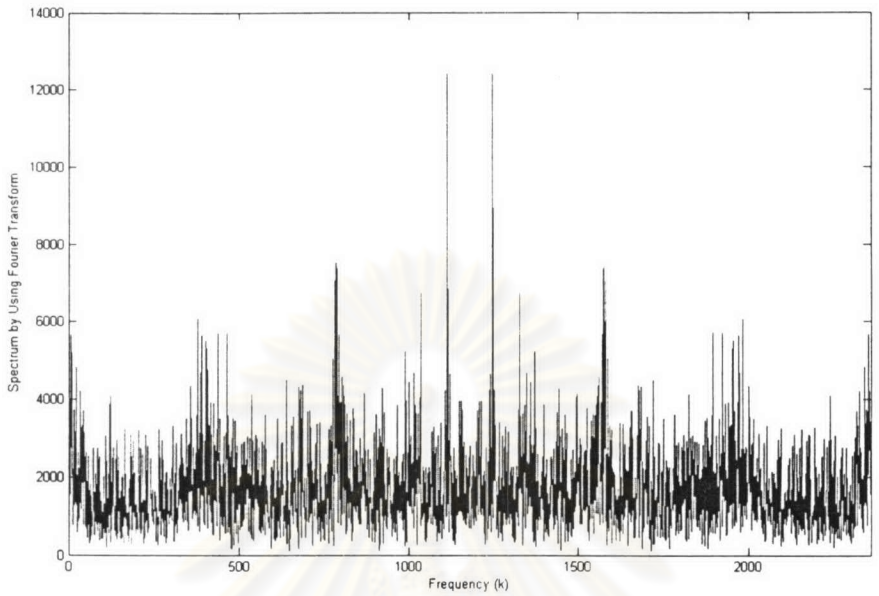

Figure 4.4 The Fourier transform of intron of *Drosophila melanogaster*

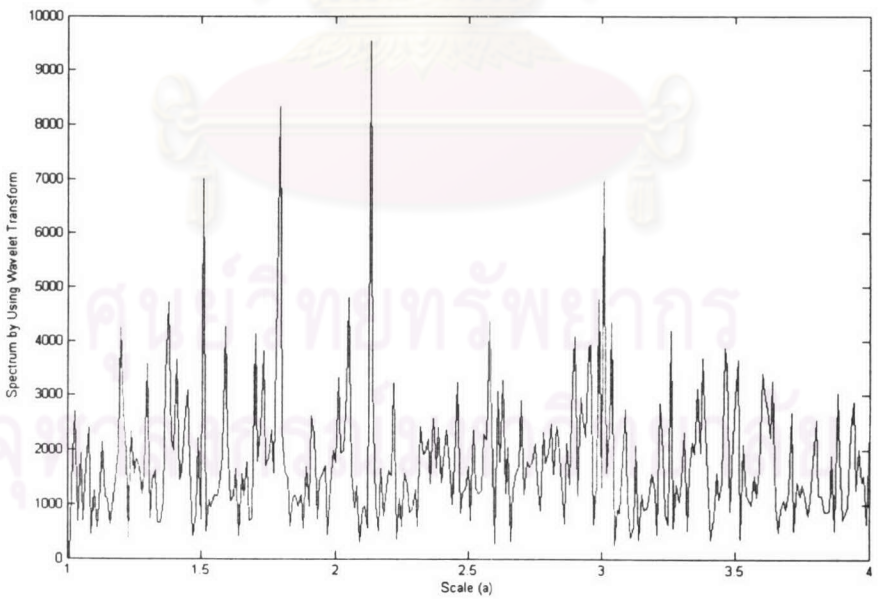

Figure 4.5 The wavelet transform of intron of *Drosophila melanogaster*

4.2 Wavelet Analysis

This section is explained the results of the analysis. The analysis is mainly applied into two DNA sequences.

4.2.1 Result of *Caenorthabditis elegans*

Afterward, we test a sample of nucleotide sequence, *Caenorthabditis eegans* (GENBANK, accession number AF099922). This sequence is length of 8000 base pairs. The mapping of this sequence into numerical sequences is applied by using the same method mentioned in Chapter 2 before applying continuous wavelet transform.

Consider the continuous wavelet transform with Morlet function described in the chapter 3. Under the integration, the Morlet function contains two exponential terms. The first term $e^{-2\pi f_c\left(\frac{t-b}{a}\right)}$ is an oscillator while the second term $e^{-\frac{1}{f_b}\left(\frac{t-b}{a}\right)^2}$ is an envelope over the oscillator. We vary the value of f_b in the second term to limit the range of computation each time. It functions as a window that moves along the time axis with parameter b whereas the value of a is fixed at a=3. By changing the value of b, this term can move and let the continuous wavelet transform performs only in the range of this window function.

The DNA of *C. elegans* will be transformed. After transformation, the result is optimized by means of Anastassiou's method to maximize the discrimination between exon and intron. The 3D plot is shown in Figure 4.6. The color plot is shown in Figure 4.7. In Figure 4.7, level of color shows the magnitude of spectrum. The light color represents the high magnitude while the dark color represents the low magnitude.

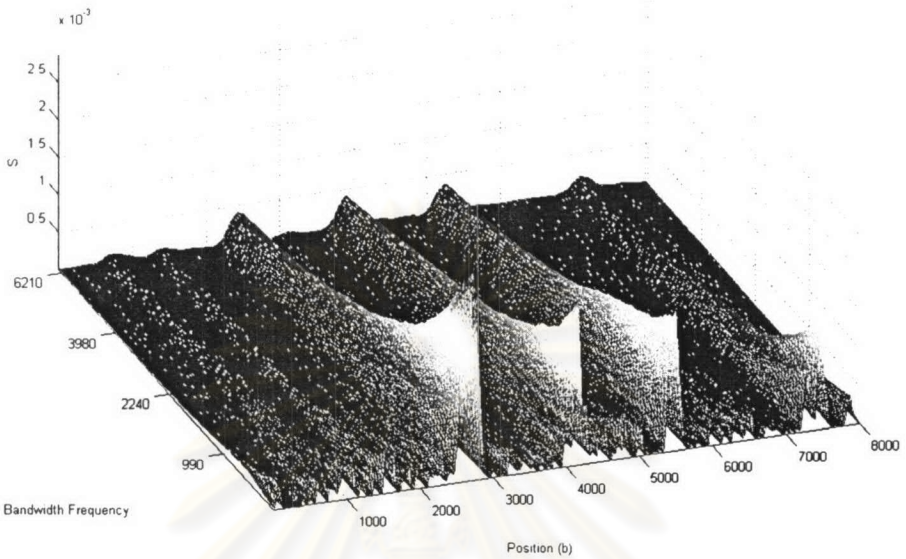

Figure 4.6 Spectrum 3D plot by using wavelet transform

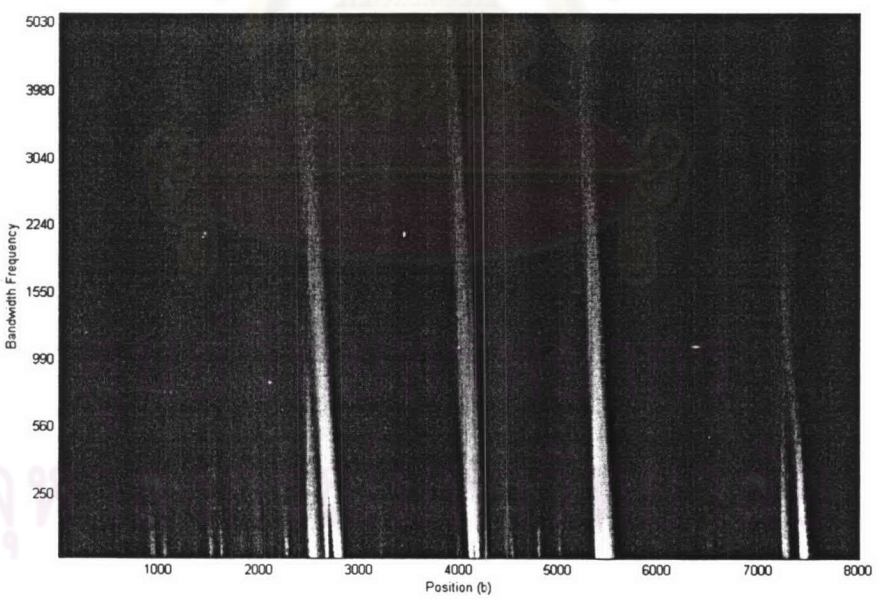

Figure 4.7 Spectrum 3D plot by using wavelet transform

As we can see from the two figures above, the spectrum plot shows the location of the exon in DNA of *C. elegans*. The actual locations of the exons are displayed in Table 4.1.

Table 4.1 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

The value of f_b which best matches the actual location of the exons for *C. elegans* is about 2000 to 2250.

Bandwidth of Morlet function controlled by the second exponent term in Morlet function, which corresponds to this range of f_b, ranges about 340-360 base pairs.

As we can see from the figures above, f_b directly affects the frequency analysis. The bigger f_b will produce the bigger bandwidth. Figure 4.8, 4.9, 4.10 and 4.11 show the frequency analysis in two dimensions at f_b =90, 560, 2130 and 5030 respectively.

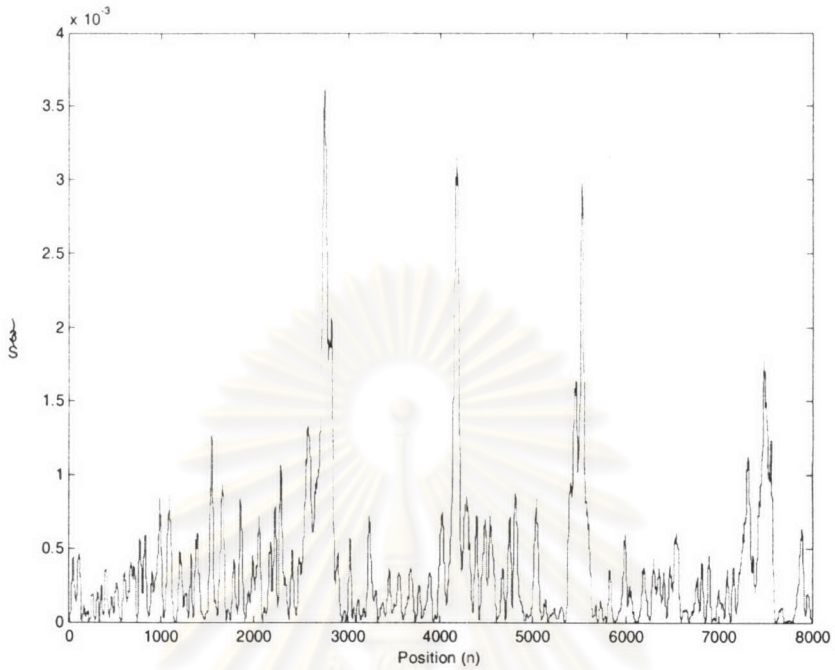

Figure 4.8 Frequency analysis of *Caenorthabditis elegans* at f_b =90

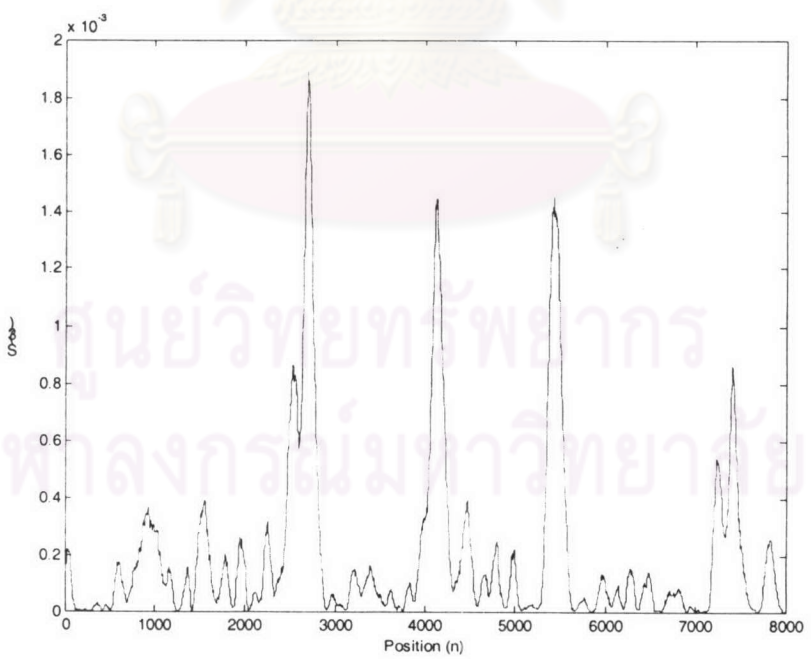

Figure 4.9 Frequency analysis of *Caenorthabditis el*egans at h f_b =560

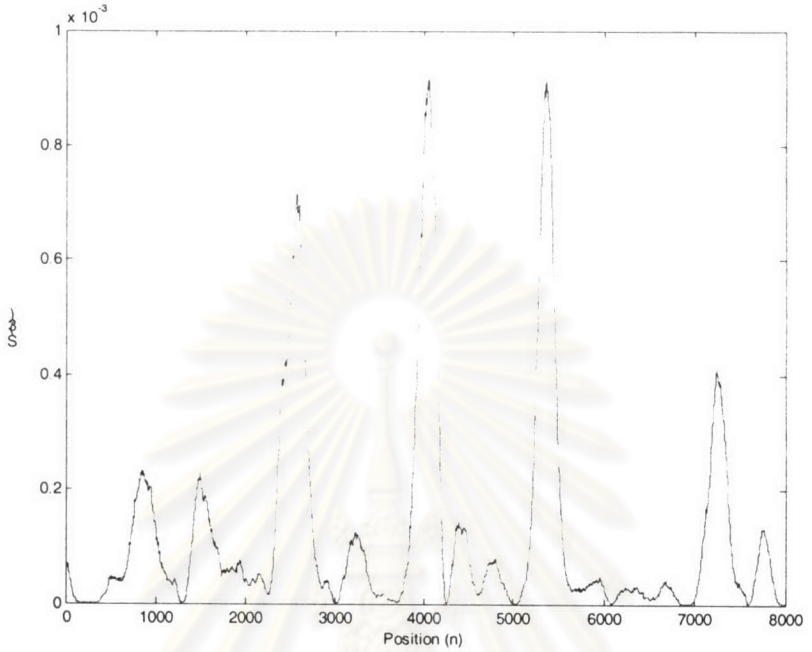

Figure 4.10 Frequency analysis of *Caenorthabditis elegans* at f_b =2130

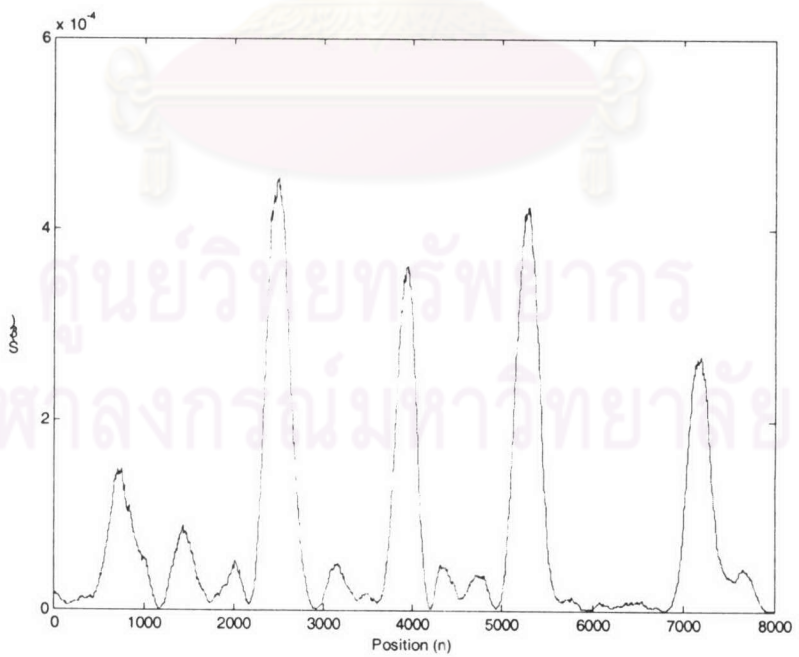

Figure 4.11 Frequency analysis of *Caenorthabditis elegans* at f_b =5032

Bandwidth in Figure 4.8 is roundly at 72 base pairs. The result has lost some information in the exons, which length greater than 72 base pairs. The Figure 4.9 has the bandwidth about 180 base pairs. The result becomes clearer because the first and forth exons have their length about 180 base pairs, actually 207 and 180 base pairs respectively. Figure 4.10 shows the best matched result with f_b =2130, corresponding to the bandwidth of 351 base pairs, which equals the largest exon on *C. elegans*. Thus, exon predictors can work well if the bandwidth is longer than any exon. The difference between Figure 4.10 and 4.11 is amplitude of the spectrum. Noticeably, the amplitude of the spectrum in Figure 4.10 is bigger than that in Figure 4.11 since the bandwidth of Figure 4.10 is about 351, which can exactly cover the longest exon in the DNA. Therefore the result becomes improved. The Figure 4.11 uses bandwidth about 540 base pairs. As the result, the outcome becomes coarse.

At this point, we notice that spectrum plots of wavelet analysis of the *C. elegans* look better than that produced by Fourier transform. The curve obtained by wavelet transform is smoother and noise is suppressed. By wavelet analysis, the peaks of locations of exons are higher than those of introns. This better displays location of exons.

4.2.2 Result of *Saccharomyces cerevisiae*

We try to apply this method to another nucleotide sequence, *Saccharomyces cerevisiae* (GENBANK, accession number NC 001135). This nucleotide sequence has its length of 12000 base pairs, composed of six exons. The actual locations of exons on that sequence are show in Table 4.2.

Table 4.2 Locations of the six exons on the *Saccharomyces cerevisiae*

Position	Exon Length
761-1429	669
1687-3135	1449
3387-4931	1545
5066-6757	1692
7147-9918	2772
10143-10919	777

We apply the same technique with the new nucleotide sequence with various f_b. Three dimensional spectrum plot is displayed in Figure 4.12.

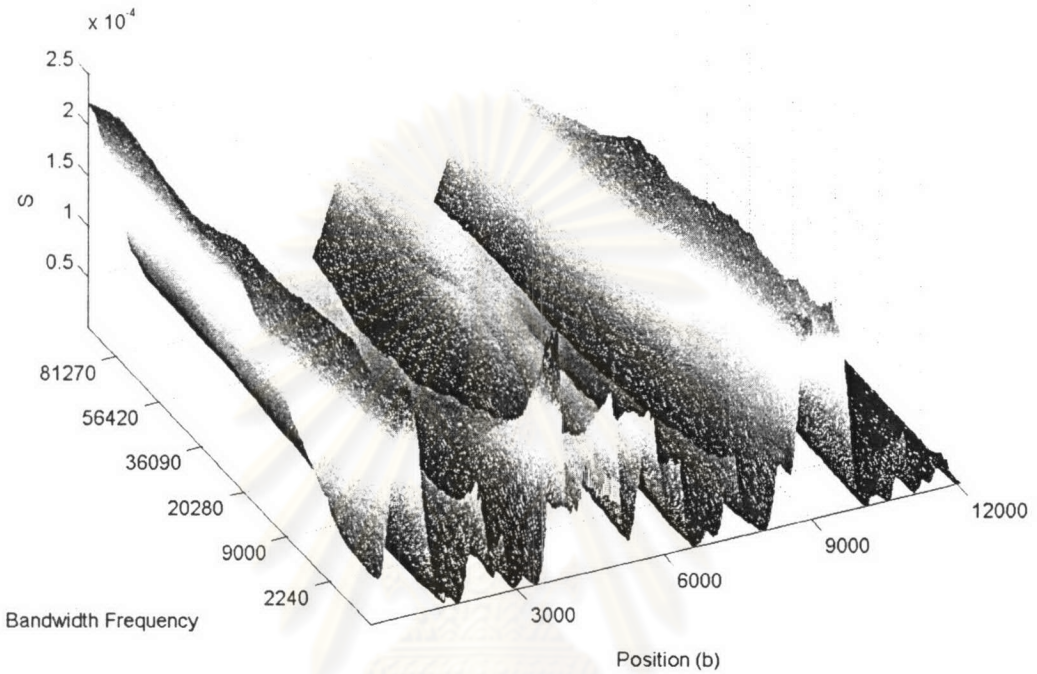

Figure 4.12 Spectrum 3D plot by using wavelet transform of *Saccharomyces cerevisiae*

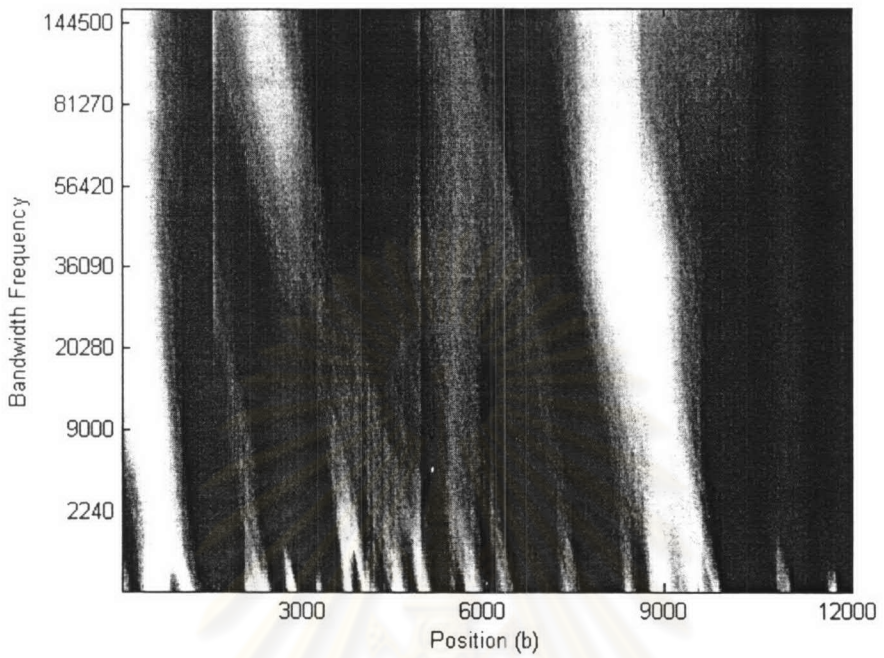

Figure 4.13 Spectrum color plot by using wavelet transform of *Saccharomyces cerevisiae*

Figure 4.12 also shows color map of the spectrum plot. The light color represents the high magnitude of *S*. Two dimensional plots of the spectrum are shown Figure 4.14 for f_b=44180 and Figure 5.15 for f_b=135310.

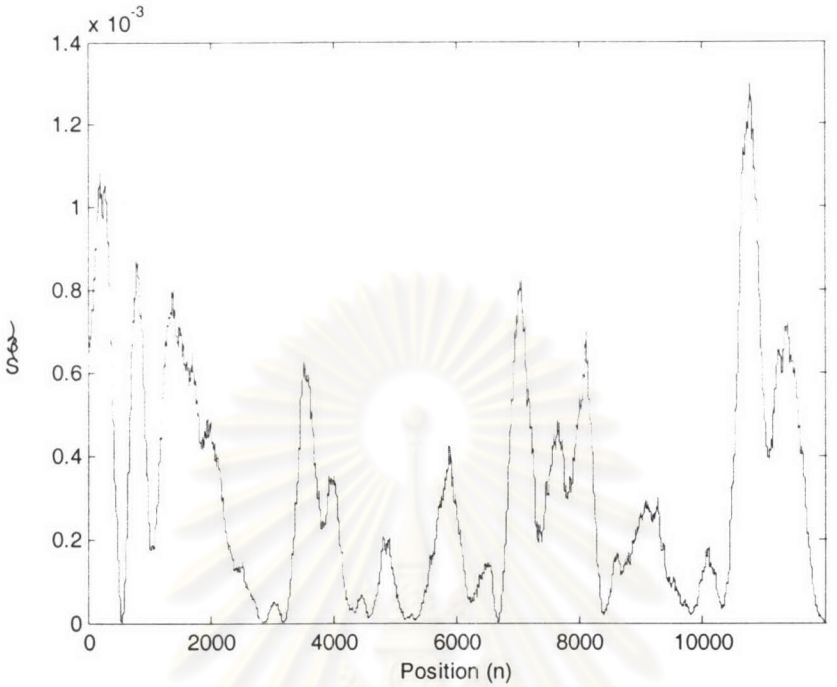

Figure 4.14 2D plot of spectrum of *Saccharomyces cerevisiae* with f_b=44180

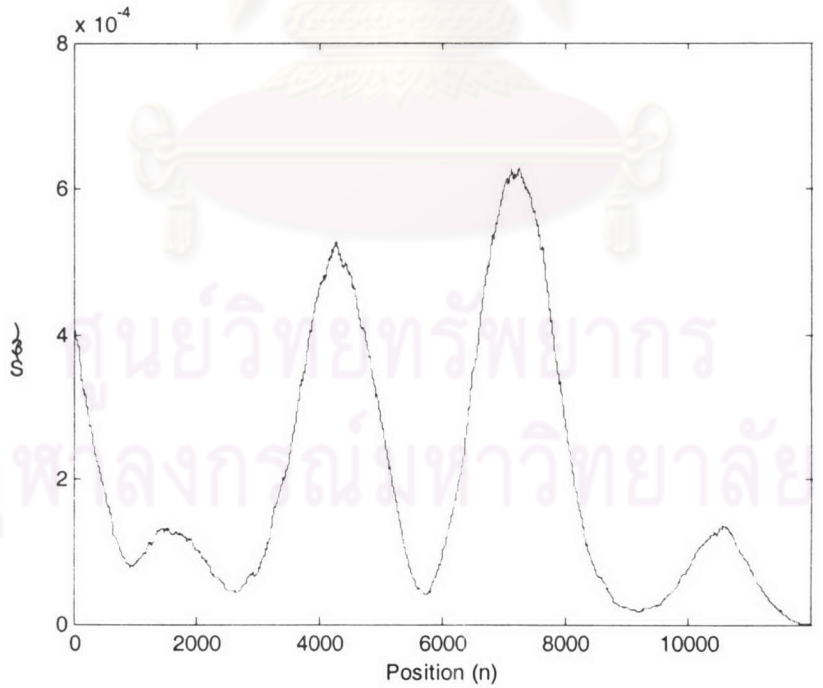

Figure 4.15 2D plot of spectrum of *Saccharomyces cerevisiae* with f_b=135310

Noticeably from the two figures above, the Figure 4.14 better matches the real locations of the exons on *S. cereviside*, while the Figure 4.15 does not match well although the Figure 4.15 uses bigger window which can cover all exons. We found that in Figure 4.14 the first three peaks locate at the positions of the exons on the sequence, but the last three peaks are not exactly matched. This fact can be explained. The reason is that the two DNA strands are complements of each other because the Saccharomyces Cereviside is a very undeveloped and simple organism. In most case of very simple organism, one gene is exactly one exon.

. The first three exons of Saccharomyces Cereviside are translated from 5' to 3' while the last three exons of the gene are translated on the complementary strand from 3' to 5' relatively to the first three exons (actually 5' to 3' relative to direction on the complementary strand). This solution will solve the disadvantage of this method.

Then we try to translate with the complementary base pair and process the DNA sequence from 3' to 5'. The frequency responses of that process are shown in Figure 4.16 and 4.17.

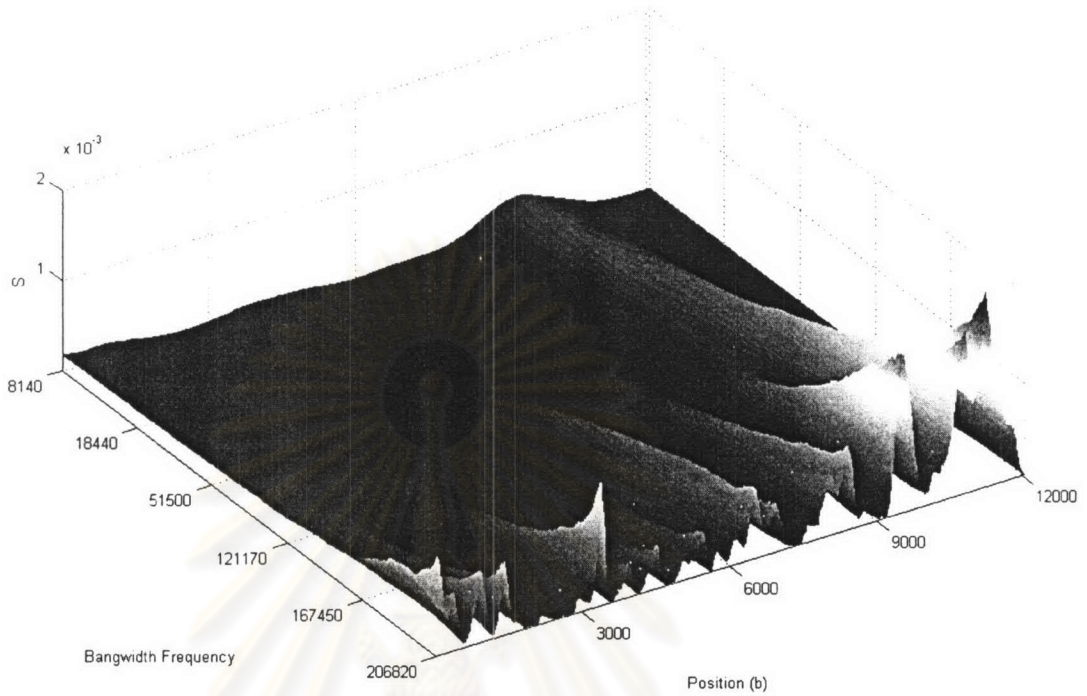

Figure 4.16 3D plot of frequency response of complement strand of *Saccharomyces cerevisiae*

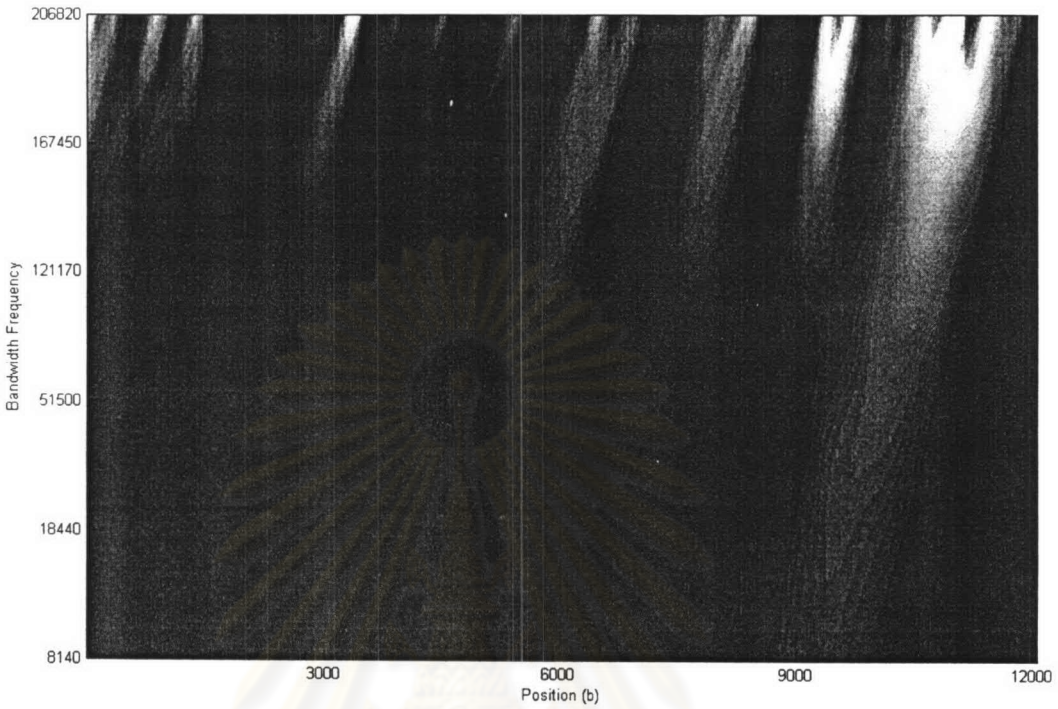

Figure 4.17 Color plot of frequency response of complement strand of *Saccharomyces cerevisiae*

Two dimensional representations of the frequency response are displayed in Figure 4.18 with f_b=44180 and Figure 4.19 with f_b=135310.

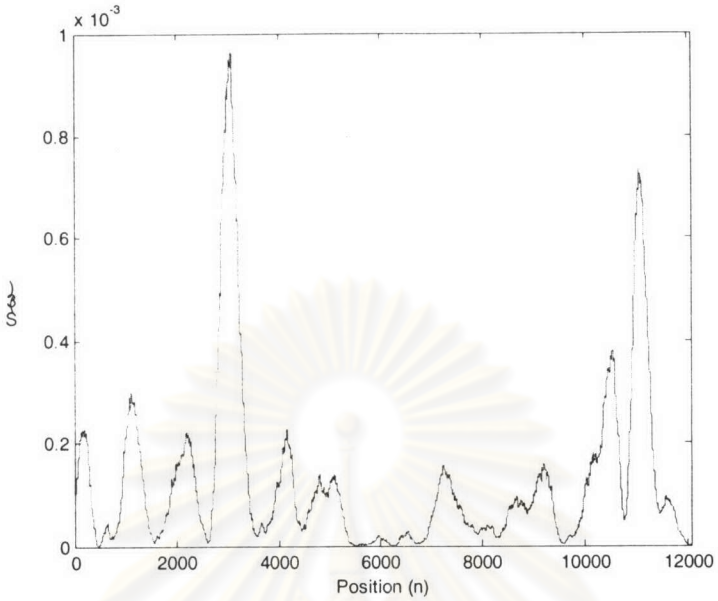

Figure 4.18 Frequency response of complement strand of *Saccharomyces cerevisiae* with f_b=44180

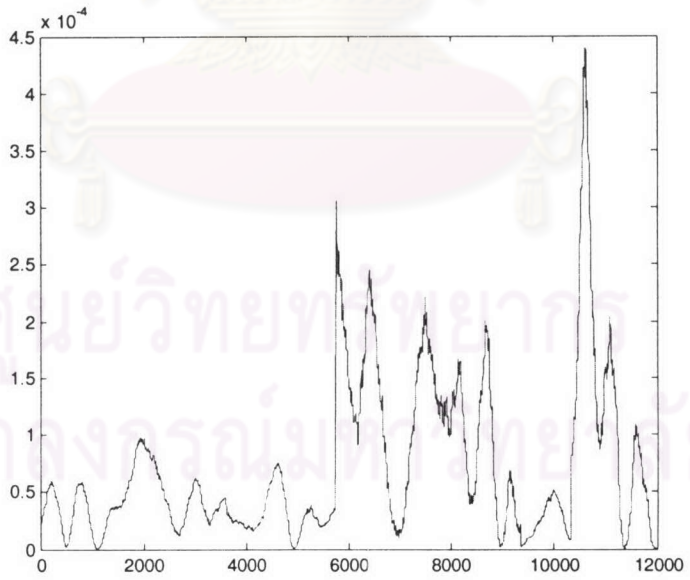

Figure 4.19 Frequency response of complement strand of *Saccharomyces cerevisiae* with f_b=135310

From the two figures above, reverse mapping is useful if the exon is decoded on a complementary strand. This technique can successfully represent the location of the last three exons of *S. cereviside*.

However, the result is possible to predict location of exons on the DNA, but it cannot precisely specify that location. By this method, the peak of frequency response foretells occurrence of periodicity of three, which is found in exon.