CHAPTER 3

BACKGROUND KNOWLEDGE

3.1 Wavelet Analysis

The wavelet transform [9] was developed to solve the resolution problem, found in Fourier transform. The wavelet transform will gives a good time resolution and poor frequency resolution at high frequencies and good frequency resolution and poor time resolution at low frequencies. The wavelet analysis works in the sense that the signal is multiplied with a function and the transform is computed separately for different segments of the time-domain signal.

Let $x(t)$ be the signal and $\varphi(t)$ be the mother wavelet function. The continuous wavelet transform is defined by

$$\Psi_x^{\varphi}(a,b) = \frac{1}{\sqrt{|a|}} \int x(t) \varphi^{*}\left(\frac{t-b}{a}\right) dt \qquad (3\text{-}1)$$

The wavelet transform is the function of two parameters, a and b, which are scaling and translation parameter respectively. The mother wavelet $\varphi(t)$ is a transforming function. For the work in this thesis, we will use complex Morlet function [10].

The complex Morlet function is defined by

$$\varphi(f_c, f_b)(t) = \frac{1}{\sqrt{\pi f_b}} e^{2i\pi f_c t} e^{-t^2/f_b} \qquad (3\text{-}2)$$

The complex Morlet function is also the function of two parameters, f_b called bandwidth frequency and f_c called center frequency. Figure 3.1 shows the Morlet function between -10 and 10.

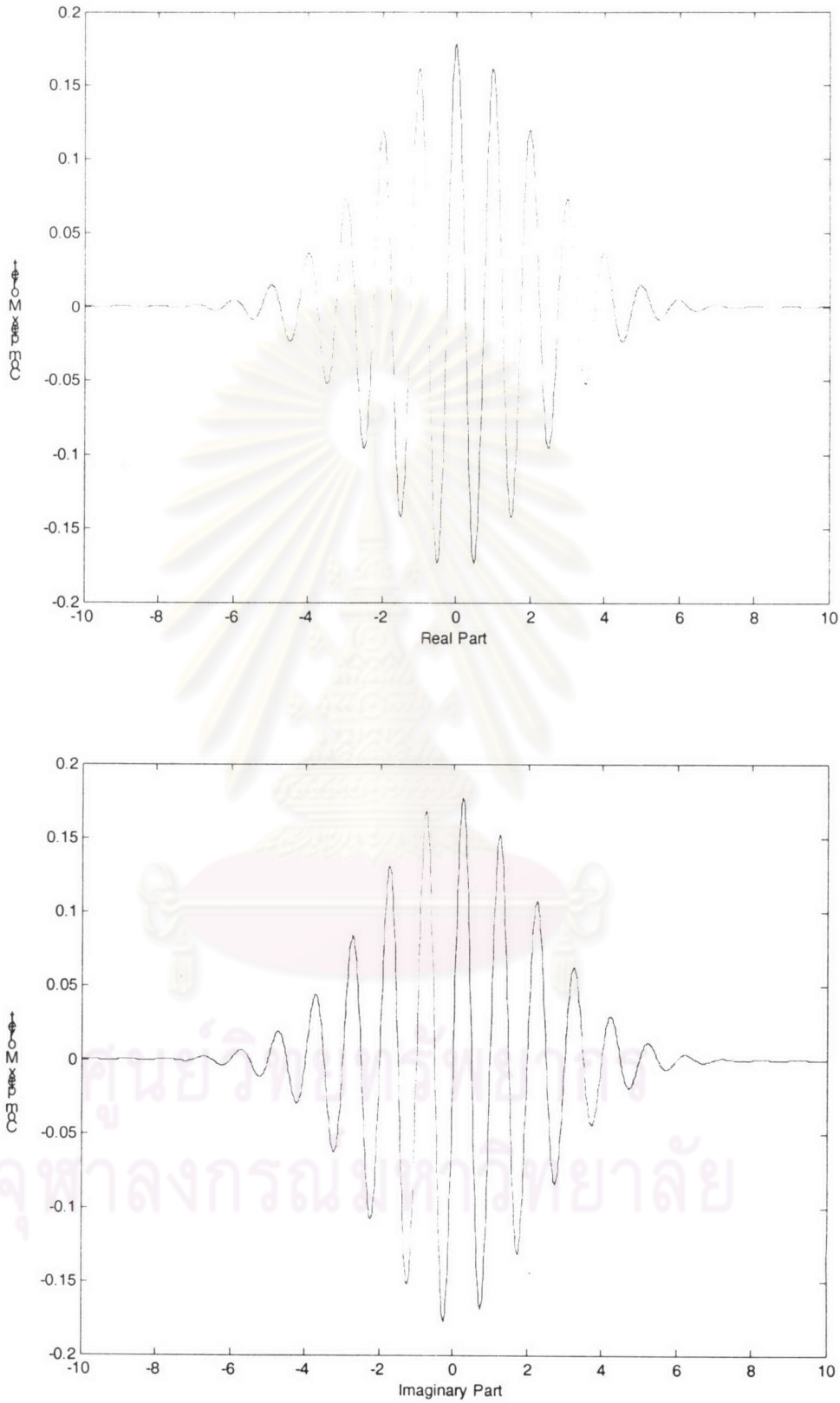

Figure 3.1 Morlet function with f_c=1 and f_b=10

Applying the complex Morlet function into wavelet transform, we obtain

$$\Psi_x^{\varphi}(a,b,f_c,f_b)=\frac{1}{\sqrt{a\pi f_b}}\int x(t)e^{2i\pi f_c\left(\frac{t-b}{a}\right)}e^{-\left(\frac{t-b}{a}\right)^2/f_b}\,dt \qquad (3\text{-}3)$$

There will be totally four varying parameters. The frequency *k* in Fourier transform will be related to with a scaling parameter *a* in wavelet transform. The benefit of this transformation is flexibility, which can perform like Fourier transform does in one step by varying these four parameters. The bandwidth parameter will determine bandwidth or window of calculation at one translation.

Consider the sinusoidal signal

$$s(t)=\begin{cases}\sin\left(\frac{\pi}{2}\times 0.05t\right) & 0\le t\le 300\\ \sin\left(\frac{\pi}{2}\times 0.1t\right) & 300<t<590,610<t\le 1000\\ 0 & 590\le t\le 610\end{cases} \qquad (3\text{-}4)$$

The signal *s(t)* composes of two frequencies between *t*=0 and *t*=1000 and a small discontinuity at *t*=590 to *t*=610. The joint of two frequencies occurs at *t*=300.

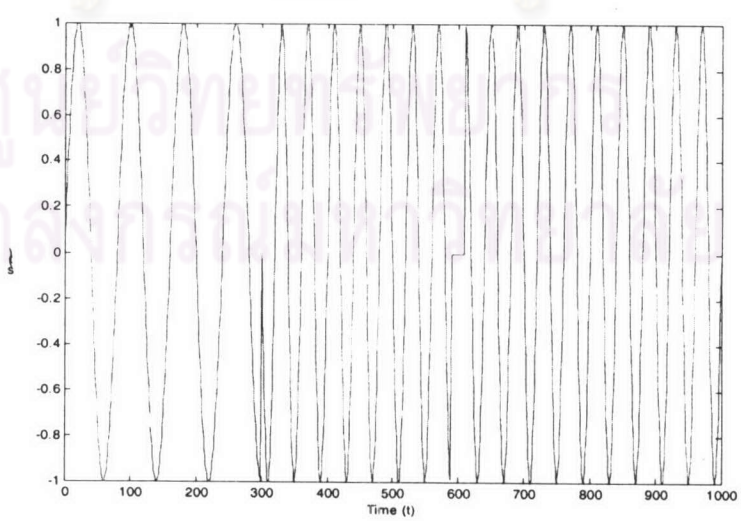

Figure 3.2 Plot of *s(t)*

Later, *s(t)* is applied with the wavelet transform using complex Morlet function with bandwidth frequency (f_b) of 1.5 and center frequency (f_c) of 1. The result is presented in Figure3.3.

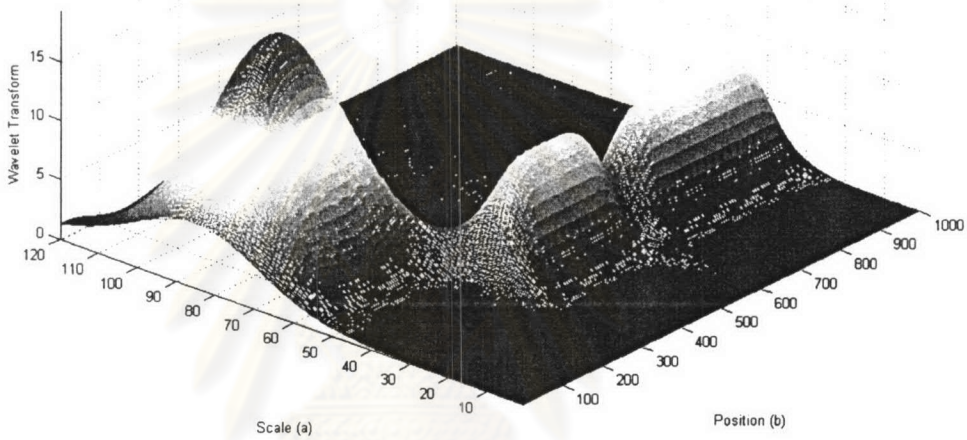

Figure 3.3 Wavelet transform of *s(t)*

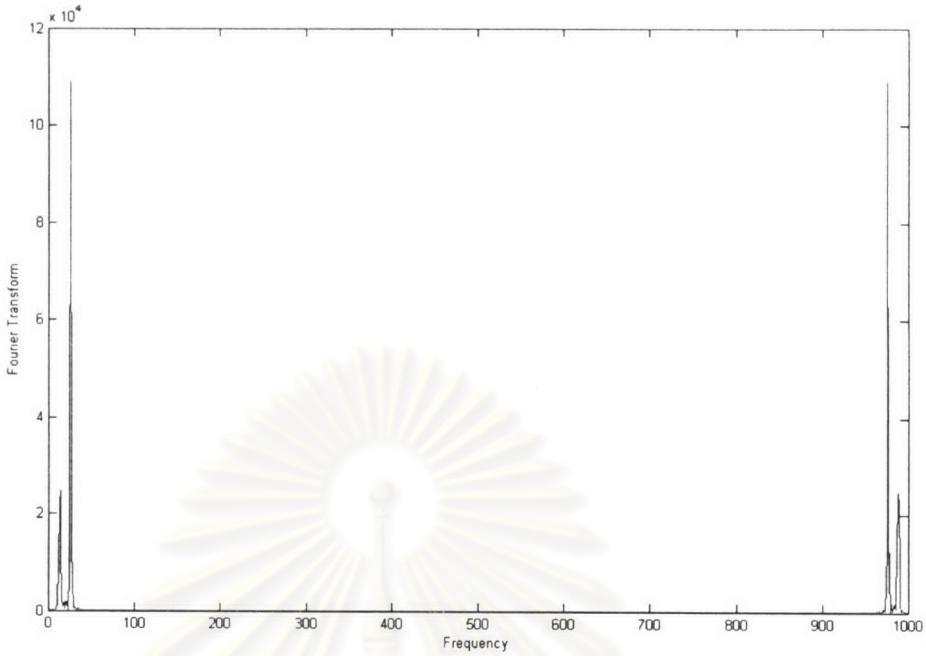

Figure 3.4 Fourier transform of *s(t)*

From the Figure 3.3, the wavelet can identify the location of frequencies change at position b=300 (corresponds to t=300) and the location of discontinuity at position b=600 (corresponds to t=600). The wavelet also indicates what frequencies the signal contains. Figure 3.5 represents the same result of wavelet transform of *s(t)* as Figure 3.3 represents, but this figure shows another viewpoint. Consider the first exponent term under integration in Equation 3-3, the value of f_c/a represents what frequencies the signal *s(t)* contains. As we cam see in the Figure 3.5, the two peaks locate at a=40 and 80. The frequencies can be calculated by using f_c/a which are 0.0125 hertz and 0.025 hertz respectively.

In comparison to spectrum by using Fourier transform, shown in Figure 3.4, the Fourier transform cannot specific the location of the sudden change of frequencies and the location of discontinuity.

In addition, the Fourier transform in Figure 3.4 gives information about the frequencies that contained in signal *s(t)* contains. This signal contains two frequencies, 0.012 hertz and 0.025 hertz. Figure 3.5 Wavelet transform of *s(t)*.

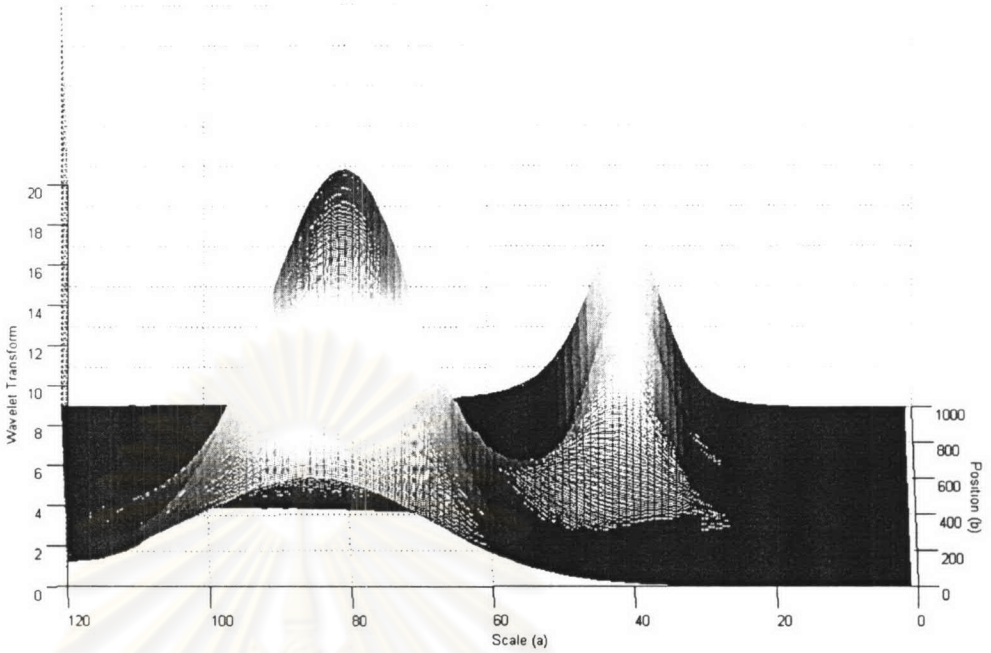

Figure 3.5 Wavelet transform of *s(t)*

Beyond, the wavelet analysis can identify the position of frequency in the time domain when each frequency exists, while the Fourier analysis cannot do. From the figure above, the first peak (a=80) exists from b=0 to b=300 while the second peak (a=40) exists from b=300 to b=1000. In Figure 3.3, the second peak is not continue around b=600. This shows the discontinuity in time domain. This value of b in frequency domain corresponds to value of t in time domain. This is outstanding advantage of wavelet transform versus Fourier transform.

3.2 Basic Knowledge about DNA

All matters including living things are made of the same common structural unit, which is atom. Atom is not a living unit, but the interconnection between being and non-being is biomolecule. Thousand atoms combine altogether to form a biomolecule [11].

Living organisms are classified into two groups: Prokaryotes and Eukaryotes. Prokaryotes (Figure 3.6) are organisms having cells without nucleus membrane or nucleus envelope, whereas eukaryotes are those having cells with nucleus membrane or nucleus envelope. In prokaryotes, organs in cells have no organelle. Thus, genetic material or nucleic acid is amalgamated altogether. Blue-green algae and bacteria are the good examples of prokaryotes. Eukaryotes have nucleus membrane. Thus, genetic material is separated from other materials.

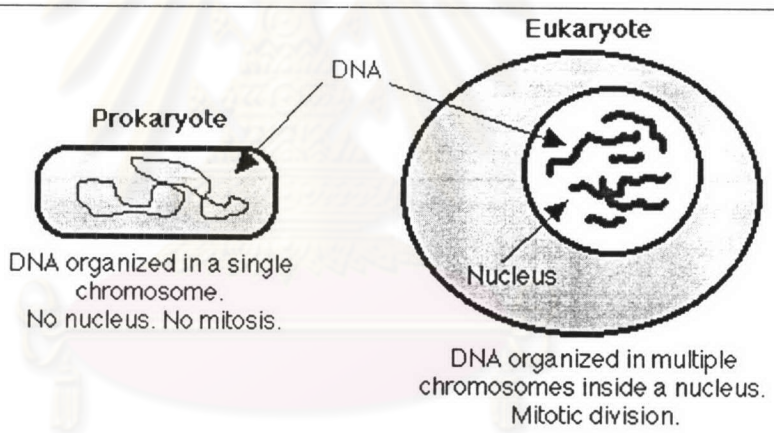

Figure 3.6 Prokaryote and eukaryote
(http://www.kubrussel.ac.be/onderwijs/etew/tew/vakken/tweedekan/ecsector/biotech/prokaryote_eukaryote.gf)

The study of nucleic acid is swiftly widened in recent years due to development of technology. The nucleic acid is made up of three chemical entities, which are

1. five-carbon sugar
2. ring-shape nitrogen base
3. phosphate group

The nucleic acid can be divided into 2 groups according to the five-carbon sugar: Deoxyribonucleic acid (DNA) following deoxyribose sugar and Ribonucleic acid (RNA) following Ribose sugar (Figure 3.7). There are four types of the ring-shape nitrogen base found on DNA: Adenine (A), Thymine (T), Cytosine (C) and Guanine (G). While, RNA also contain four types of bases (T on DNA is substituted by Uracil (U) on RNA)

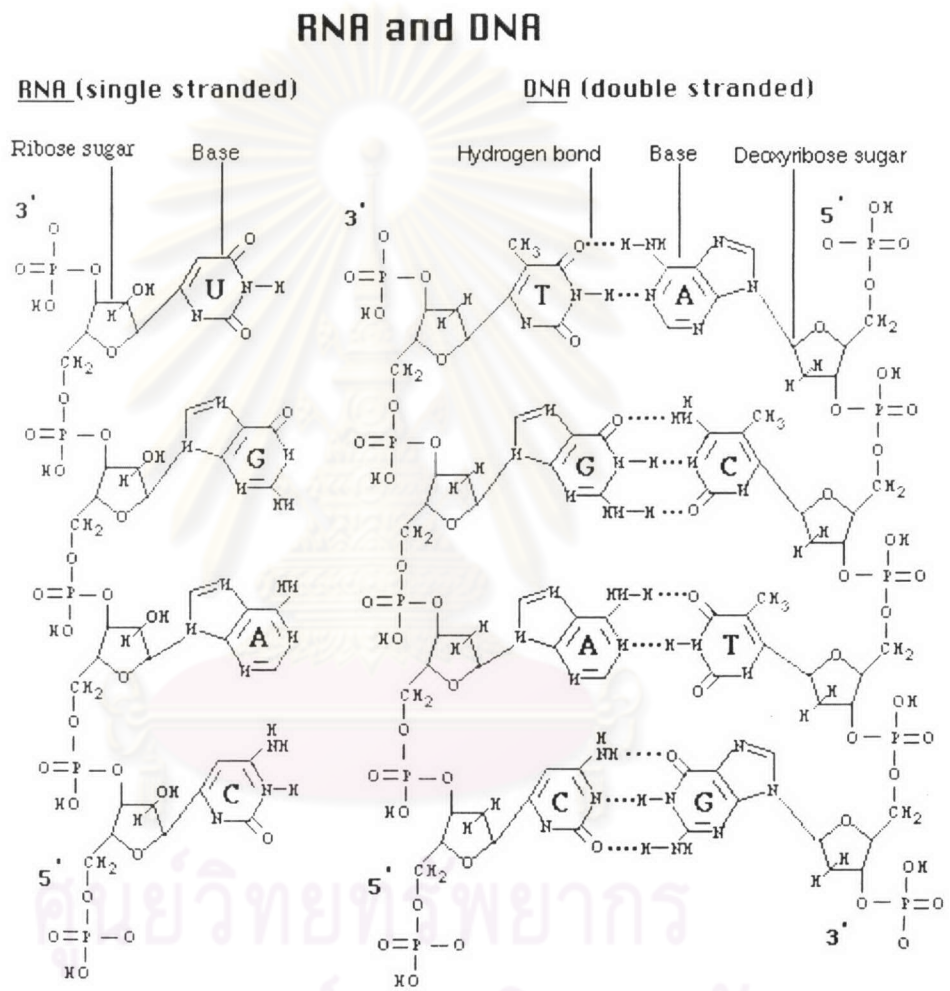

Figure 3.7 Nucleic acids: RNA and DNA

(http://ntri.tamuk.edu/cell/chapter7/dna-rna.gif)

These bases can be grouped into two main groups by their molecular structures. They are purines and pyrimidines. Purines are two-ring molecular structure whereas pyrimidines are single-ring molecular structure. Adenine and guanine are purines. Cytosine and thymine are pyrimidines. When a base binds to phosphate groups, this is called nucleotide [12].

Nucleotides are covalently linked together to form the polynucleotide chain. In this reaction, the oxygen of the 3' hydroxyl group on the chain attacks the phosphate of another nucleotide at 5' hydroxyl group. Notice that the protein chain is synthesized from DNA in a 5' to 3' direction.

DNA is double-stranded in shape of double helix. DNA is made up of two polynucleotide chains. These two strands are held together by hydrogen bonds forming between two strands. Only certain bases can pair well with other bases. A always pairs with T with double-hydrogen bond. C always pairs with G with triple-hydrogen bond (Figure 3.8).

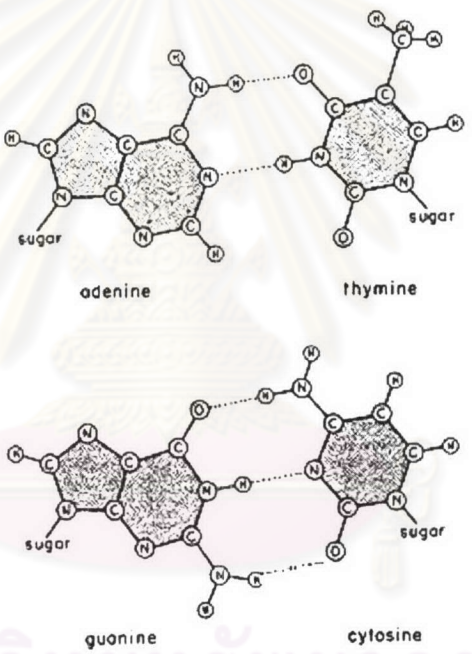

Figure 3.8 Bases and their bonding

(http://cecelia.physics.indiana.edu/life/dna/bases.jpg)

When a cell divides itself into new two cells, nucleus will be separated in order to conserve genetic properties for growing cells. DNA in nucleus will be replicated. Double helix DNA will be denaturized, the process of disruption the hydrogen bonds between bases on different strands by a certain enzymes. Then, two new DNAs will be replicated by means of the two templates. Each cell will be made up of double-stranded DNA, one of the old DNA and one of the new generated DNA.

Genome presents the total information content in DNA in each cell of an organism. Genomes are made up of very long DNA molecules housed on chromosomes. For example, the genome of the bacteria *Escherichia coli* consists of a single chromosome which is a single molecule of $4x10^6$ base pairs. The human genome is made up of $3.5x10^9$ base pairs of DNA distributed over 23 chromosomes. Chromosomes included many genes. A gene is a section of chromosomal DNA that contains the information to make a specific polypeptide through the production of a specific RNA, which is usually single-stranded. In order to be decoded by the cell, DNA is firstly duplicated into messager RNA (mRNA). The process of synthesizing a specific mRNA is called transcription (Figure 3.9).

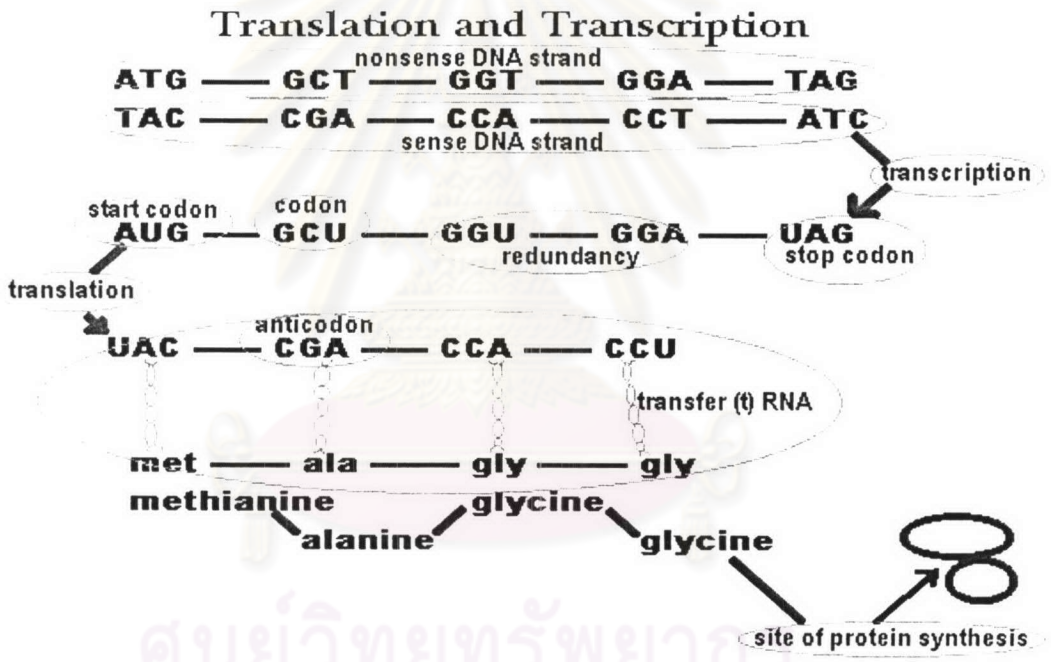

Figure 3.9 Transcription and translation
(http://www.agen.ufl.edu/~chyn/age2062/lect/lect_07/of7_1a.GIF)

A gene consists of many sub-regions. These sub-regions are distinguished into two types: exon and intron. Exons will be spliced together and then translated to amino acids. Introns will be left in a process of splicing (Figure 3.10).

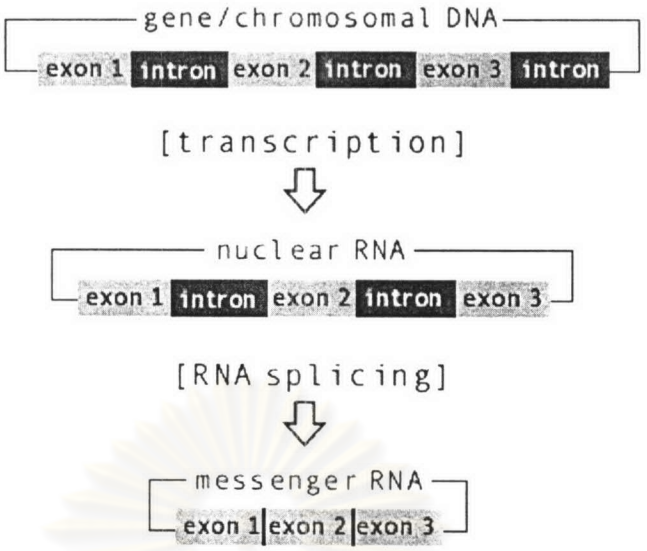

Figure 3.10 Splicing

(http://www.life.uiuc.edu/bio100/cindyk/proteinsyn/splicing.gif)

The three ordered arrangement of nucleotides is called a codon. A sequence of codon on the mRNA specifies the addition of an amino acid to a growing polypeptide chain. The chain is terminated when a stop codon is encountered.

A codon table is shown below. This table can be used to predict the amino acid sequence of a protein if the DNA sequence is known.

Table 3.1 Codon Table

First Position	Second Position				Third Position
	U	C	A	G	
U	PHE	SER	TYR	CYS	U
	PHE	SER	TYR	CYS	C
	LEU	SER	STOP	STOP	A
	LEU	SER	STOP	TRP	G
C	LEU	PRO	HIS	ARG	U
	LEU	PRO	HIS	ARG	C
	LEU	PRO	GLN	ARG	A
	LEU	PRO	GLN	ARG	G
A	ILE	THR	ASN	SER	U
	ILE	THR	ASN	SER	C
	ILE	THR	LYS	ARG	A
	MET	THR	LYS	ARG	G
G	VAL	ALA	ASP	GLY	U
	VAL	ALA	ASP	GLY	C
	VAL	ALA	GLU	GLY	A
	VAL	ALA	GLU	GLY	G

From the Table 3.1, there are 64 possible codons because one codon comprises of 3 bases whereas each base is possible 4 kinds. In fact, there are merely 20 kinds of amino acids. This shows that one amino acid is decoded from more than one pattern of codon.

From the codon table, AUG is always used to begin translation of an mRNA. The AUG, used to start protein synthesis, is called an initiation codon. On the other hands, certain codon results in the termination of a growing chain. These stop or termination codons are UAA, UAG and UGA.

Sequence of amino acids translated from a strand of mRNA can be different depends on where the reading start. There are 3 possible starting positions, which are called reading frames. Different reading frames give different result of amino acid translation, as shown in the figure below.

	A A U U C G A G U U U G U G
Frame 1	ASN TRP SER LEU
Frame 2	ILE ALA VAL CYS
Frame 3	PHE GLU PHE TRP

Figure 3.11 Reading Frames