Locating putative p53-binding sites in DNA sequences

A Project Report

submitted by

RAGHUNANDAN H K

(EE03B042)

in partial fulfillment of the requirements
for the award of the degree of

BACHELOR OF TECHNOLOGY
in
ELECTRICAL ENGINEERING

Under the guidance of

Prof.R David Koilpillai

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY, MADRAS.

May 2007

THESIS CERTIFICATE

This is to certify that the thesis titled **Locating putative p53-binding sites in DNA sequences**, submitted by **Raghunandan H K**, to the Indian Institute of Technology, Madras, in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Electrical Engineering**, is a bona fide record of the research work done by him under my supervision and guidance. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Prof. R David Koilpillai
Professor
Dept. of Electrical Engineering
IIT-Madras, 600 036

Place: Chennai
Date: 9th May 2007

ACKNOWLEDGEMENTS

I would sincerely like to thank Prof. David Koilpillai for his continued support and guidance throughout the project. This project would not have been possible but for his insightful guidance. I will forever cherish the time spent with him and he will continue to inspire me in my future endeavours.

I would also like to thank Dr. Sanjeev Kumar Gupta, Dept. of Biotechnology, IIT Madras for providing the necessary guidance and the material necessary to understand this inter-disciplinary subject. I would also like to take this opportunity to thank Dr. Shekhar C. Mande, for his valuable inputs.

I would like to thank my parents for their love and affection and for instilling in me a sense of discipline which is indispensable in any walk of life. It is their guidance that has made me what I am today. A note of thanks is also due to my sister for her inputs in the field of genetics.

My friends at IIT Madras have been very cooperative and helpful throughout my stay on campus. We have had numerous discussions and arguments on a wide spectrum of topics and these have given me a very broad perspective of life. I would like to thank all of them for it. I would like to specially thank my friend Kumar Appaiah for being a constant source of inspiration for us to work more diligently towards our goals.

Finally I would like to thank all my Professors at IIT Madras, not only for helping me gain an understanding of Engineering in general and Electrical Engineering in particular, but also for the inspiration that I draw from them which will be a motivating force throughout my student life.

ABSTRACT

Genomic signal processing deals with applying signal processing techniques to sequences occurring in life forms, particularly the DNA and RNA sequences. Such techniques become extremely important in obtaining useful information from the large sets of data we are presented with, in the form of the human genome. Applications of some signal processing methods like Fourier transformation and spectrograms to DNA strings are studied. We then study the problem of locating putative p53-binding regions on a DNA sequence. These regions are known to play an important role in tumor suppression and thus in the prevention of cancer. We validate the profile Hidden Markov Model based method for locating the p53-binding regions and extend the method with a filtering technique based on the concept of offending bases. Simulation results are presented highlighting the effectiveness of the proposed technique.

Table of Contents

List of Figures

Abbreviations

DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
DFT	Discrete Fourier Transform
HMM	Hidden Markov Model
pHMM	profile Hidden Markov Model
TP53	Tumor Protein 53
TAD	Transcription Activation Domain
DBD	DNA Binding core Domain
OD	Oligomerisation Domain

Chapter 1

Introduction

1.1 Genomic Signal Processing

Genomic signal processing [1] is primarily the processing of Deoxyribonucleic acid (DNA) sequences, Ribonucleic acid (RNA) sequences, and proteins. A DNA sequence is made up of an alphabet of four elements namely *A, T, C, and G*. These letters represent the bases Adenine, Thymine, Cytosine and Guanine respectively. A DNA string contains the genetic information of living organisms. More information on DNA, genes and the Genetic code can be found in Appendix A.

Another example of discrete element sequences in life forms is the protein. Proteins govern a number of functions in living organisms. A protein is a sequence of amino acids. There are twenty amino acids and hence a protein is a sequence defined on an alphabet of twenty. The twenty amino acids are denoted by letters from the English alphabet except B, J, O, U, X, and Z.

It has been shown that [1] [2] we can perform a number of signal processing operations like Fourier transformation, digital filtering, wavelet transformations, and Markov modelling on gene sequences. These DSP techniques have important practical applications and assist in obtaining useful information from these gene sequences.

In this thesis we study the problem of locating p53-binding regions on a DNA sequence. p53 is a protein that was first identified in 1979. Its role as a tumor suppressor was established in 1989 and since then, it has received considerable attention from the scientific community. Owing to its role as a transcription factor and a tumor suppressor, the sequences on the DNA that can bind to the p53 protein have a high likelihood of being in the vicinity of genes that regulate the cell cycle. Hence a method of finding such genes is to locate DNA sequences that are p53-binding.

With the completion of the Human Genome Project, vast data are available regarding the human genome and computational methods are being developed to analyze this data. One such method is the use of Hidden Markov Models (HMM) to detect patterns in sequences [10]. Anders Krogh and others have developed a HMM equivalent of profile analysis for investigating protein families. A similar method called profile Hidden Markov Models (pHMM) has been used for the identification of p53-binding regions in [11]. In this thesis we validate the pHMM method for locating putative p53-binding sequences and extend the method by filtering sequences with offending bases.

1.2 Contribution of the thesis

- We verify the Discrete Fourier Transform (DFT) technique and the Spectrogram technique for locating coding regions on the DNA sequence and demonstrate their equivalence
- We validate the pHMM method for locating putative p53-binding DNA sequences
- We extend the pHMM method by filtering sequences with offending bases

1.3 Thesis Outline

The outline of the thesis is as follows. Chapter 2 presents some of the general signal processing methods like Fourier transformation and Spectrograms that can be applied to the study of DNA sequences with some simulation results to illustrate the methods. In Chapter 3, we describe the p53 protein and its role in tumor suppression. Chapter 4 presents the problem of locating putative p53-binding regions in a DNA sequence. The profile Hidden Markov Model approach to this problem is studied in detail. An extension to this method based on filtering out sequences with offending bases is then presented. The methods studied are demonstrated using computer simulations. Chapter 5 presents some concluding remarks and possible directions of future research.

Information regarding DNA, RNA, Genes, and protein synthesis have been included in Appendix A.

Chapter 2

Signal Processing on DNA sequences

One of the applications of signal processing techniques to DNA sequences is in the identification of coding regions on the DNA. These are the regions that encode for proteins and hence constitute an important part of the human genome. We study two techniques, the Fourier transform technique [1] and the Spectrogram technique [2], for the identification of coding regions and demonstrate their equivalence.

2.1 Fourier transformation

Fourier transform techniques have been found to be useful in extracting information from DNA sequences [1]. First, let us define *indicator sequences* for the bases in DNA. There are four indicator sequences corresponding to the four bases A, T, G, and C. The indicator sequence corresponding to a particular base indicates the presence or absence of that base at the specified position by a 1 or a 0 respectively. For example, for the DNA sequence ATGTCGTAAGGTCCGT, $x_A[n]$ = 1000000110000000, $x_T[n]$ = 0101001000010001 ...

Let the Discrete Fourier Transform (DFT) of a length-N block of $x_A[n]$ as $X_A[n]$, that is

$$X_A[n] = \sum_{n=0}^{N-1} x_A[n] e^{-j2\pi kn/N} \qquad\qquad 0 \leq k \leq N-1$$

One of the many application of Fourier transform techniques to study DNA sequences is the identification of protein coding regions in genes, that is, the regions which can be decoded into proteins. It has been noticed [3] that the protein-coding regions have a *period-3 component* because of coding biases in the translation of codons into amino acids. The period-3 property is not present outside the coding regions and can be exploited to locate such regions. Thus if N is a multiple of 3, then

$$S[k] \triangleq |X_A[k]|^2 + |X_T[k]|^2 + |X_C[k]|^2 + |X_G[k]|^2$$

should peak at $k = N/3$. Thus, if we track this value for short windows along the DNA sequence, we can identify the coding regions depending on the peaks of $S[N/3]$. The window length should be long enough, so that the periodicity effect can be captured above the random variations. But a long window implies more computations and also poorer resolution.

2.1.1 Simulation Results

The gene F56F11.4 in the C.elegens chromosome III [5] was selected for implementing the above method. This gene is known to have five coding regions. The simulations were carried out with a shift length of 30 samples between adjacent windows. The window lengths were varied to see the effect of changing window length and a Hamming window was used before taking the DFT so as to remove high frequency noise. Figures 2.1, 2.2, 2.3 show the plots of $S[N/3]$ as a function of the position along the gene for the following three cases

i Window size 351 samples, with a Hamming window

ii Window size 351 samples, with a rectangular window

iii Window size 651 samples, with a Hamming window

As can be seen from the figures, use of a Hamming window reduces high frequency noise. Also, simulation with longer windows gives less noise, but the peaks are *spread out* leading to poorer resolution.

2.2 DNA Spectrograms

Spectrograms visually provide a significant amount of information about signals. For example, trained observers can figure out the words uttered in voice signals by simple visual inspection of their spectrograms. Similarly, spectrograms can be powerful visual tools for DNA sequence analysis [2].

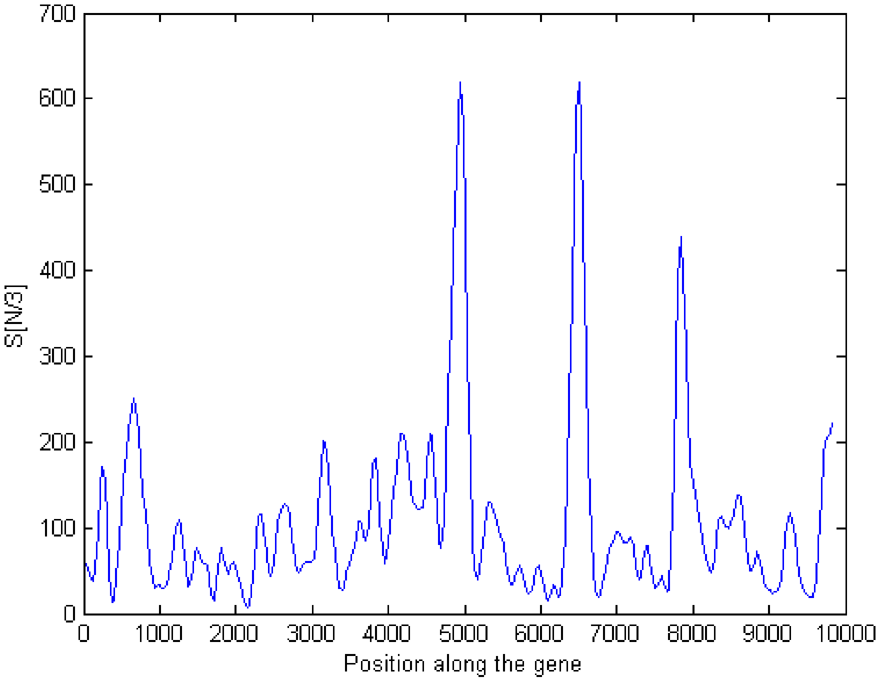


Figure 2.1: $S[N/3]$ for the gene F56F11.4 in the C.elegans chromosome III with Window length = 351, Window Type : Hamming Window

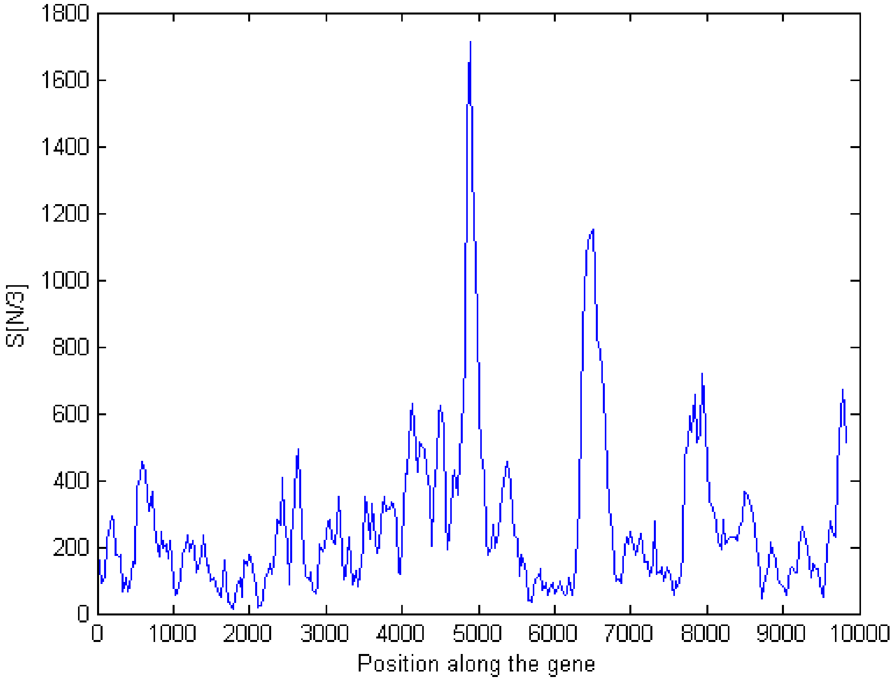


Figure 2.2: $S[N/3]$ for the gene F56F11.4 in the C.elegans chromosome III with Window length = 351, Window Type : Rectangular Window

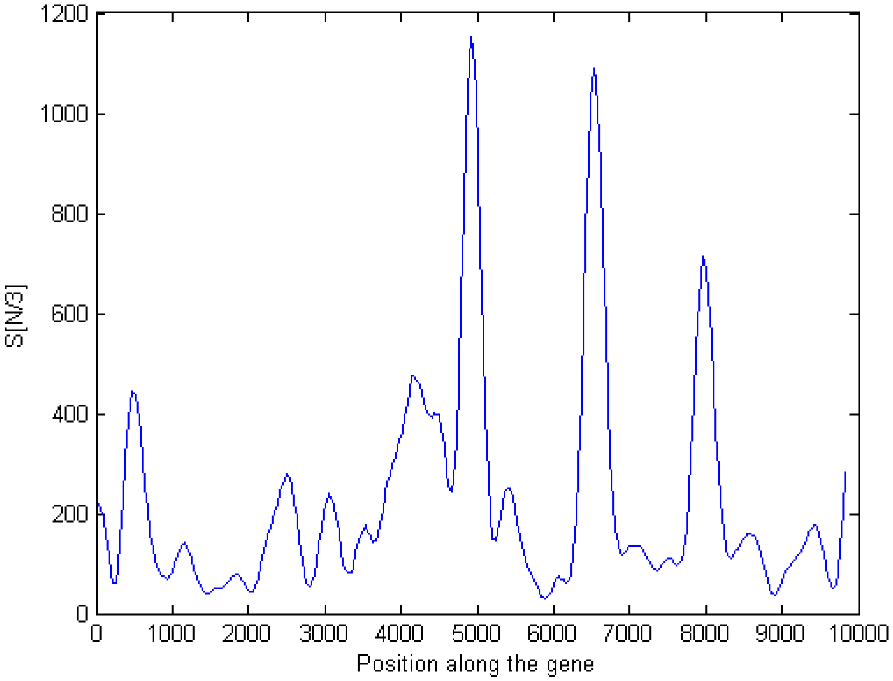


Figure 2.3: $S[N/3]$ for the gene F56F11.4 in the C.elegans chromosome III with Window length = 651, Window Type : Hamming Window

Here we define a spectrogram as a display of the magnitude of the DFT. Since we can use colour spectrograms, we have a dimensionality of three. Also note that any three of the four indicator sequences are enough to define the DNA string since the four indicator sequences should add up to 1 at all positions.

To reduce the dimensionality from four to three in a symmetric manner, we can adopt the method where each of the four bases is assigned to a vertex of a regular tetrahedron in space. Thus, the three numerical sequences x_r, x_g, and x_b can be defined by considering the four 3-D vectors having unit magnitude and pointing from the centre to the vertices of the tetrahedron. For example, as shown in Figure 2.4, we can choose A(0,0,1), T($2\sqrt{2}/3$,0,-1/3), C(-$\sqrt{2}/3$,$\sqrt{6}/3$,-1/3), and G(-$\sqrt{2}/3$,-$\sqrt{6}/3$,-1/3) as the vertices of the tetrahedron. Then

$$x_r[n] = \frac{\sqrt{2}}{3}(2x_T[n] - x_C[n] - x_G[n])$$

$$x_g[n] = \frac{\sqrt{6}}{3}(x_C[n] - x_G[n])$$

$$x_b[n] = \frac{1}{3}(3x_A[n] - x_T[n] - x_C[n] - x_G[n])$$

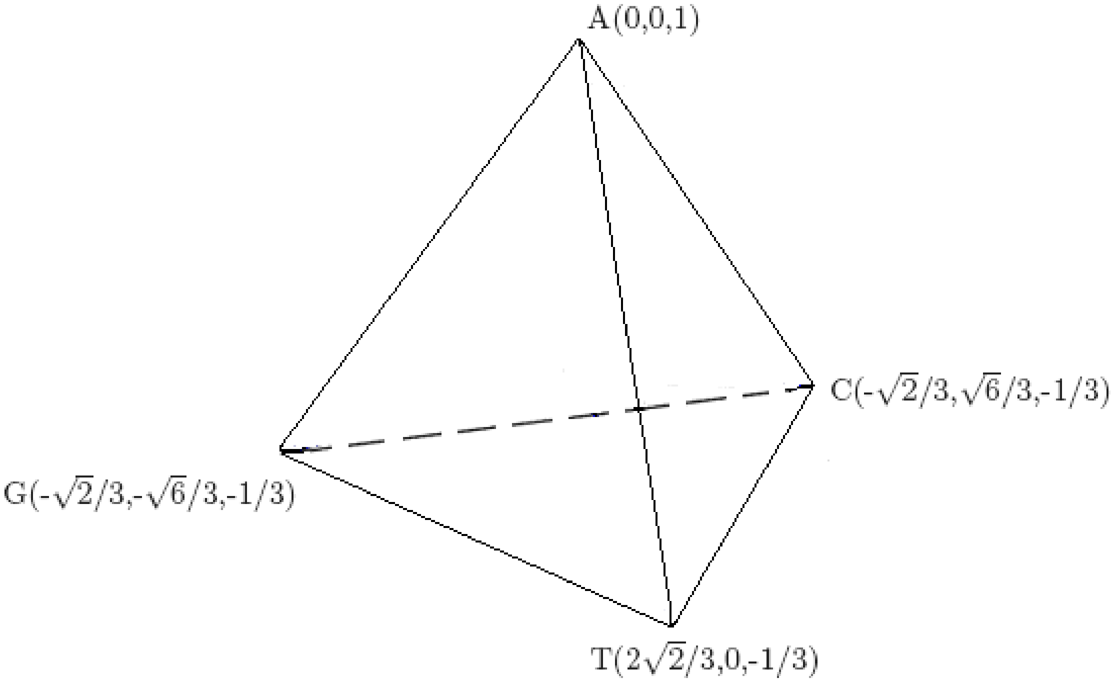


Figure 2.4: A schematic representation of the conversion scheme from the set $\{x_A[n], x_T[n], x_G[n], x_C[n]\}$ to the set $\{x_r[n], x_g[n], x_b[n]\}$

The DNA spectrogram is now obtained by a superposition of the three primary colours red, green and blue. The intensity of each of these colours is made proportional to the magnitude of the DFT of the corresponding signal.

2.2.1 Simulation Results

The DNA spectrogram defined in Section 2.2 can also be used for the problem of identification of protein coding regions in a DNA sequence. If the DFT size N is an integral multiple of 3, then, owing to the period-3 property mentioned in Section 2.1, the spectrogram at $N/3$ should show bright spots corresponding to the protein coding regions.

We generate the spectrogram for the gene F56F11.4 in the C.elegans chromosome III which, as seen in Section 2.1.1, has five coding regions. As with the earlier method, the window size was 351 samples. This was zero padded with an equal number of samples before taking FFT so as to improve frequency resolution.

The spectrogram thus obtained is shown in Figure 2.5. The spectrogram of a random sequence has also been shown in Figure 2.6 for comparison.

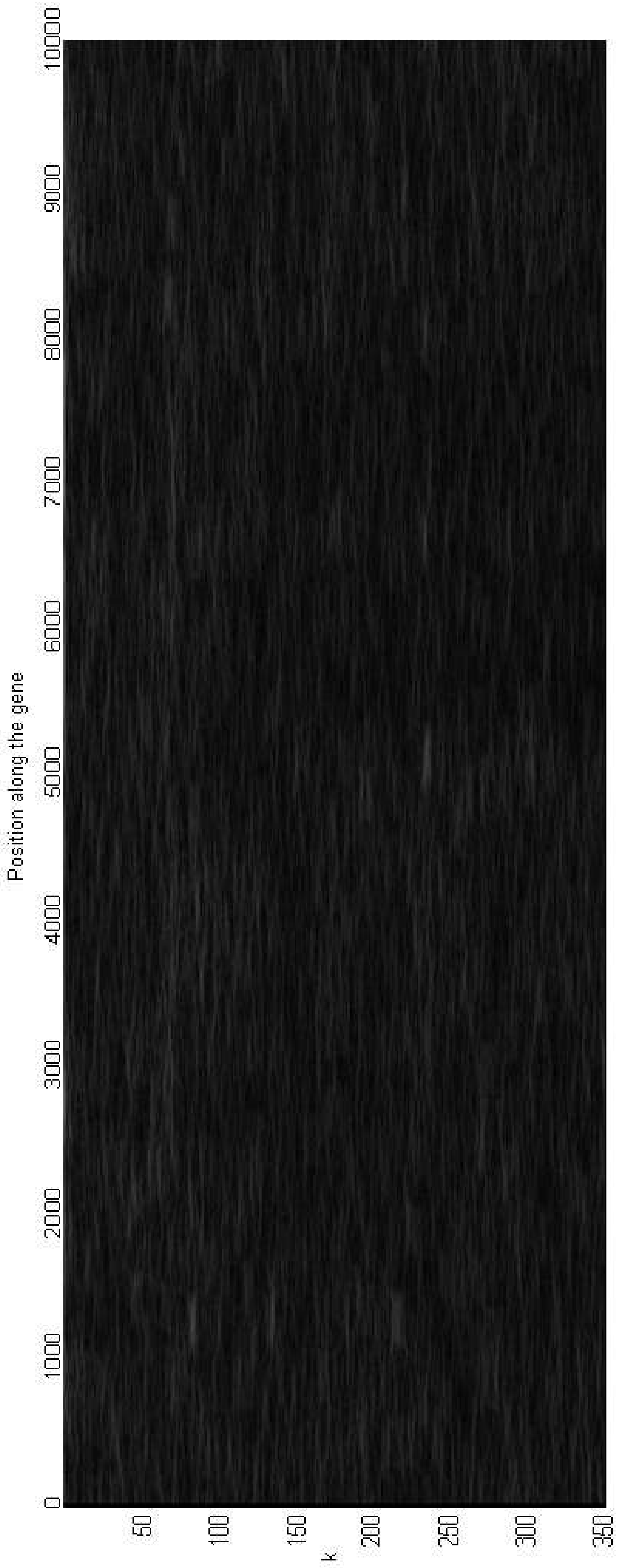


Figure 2.5: Spectrogram of the gene F56F11.4 in C.elegans chromosome III

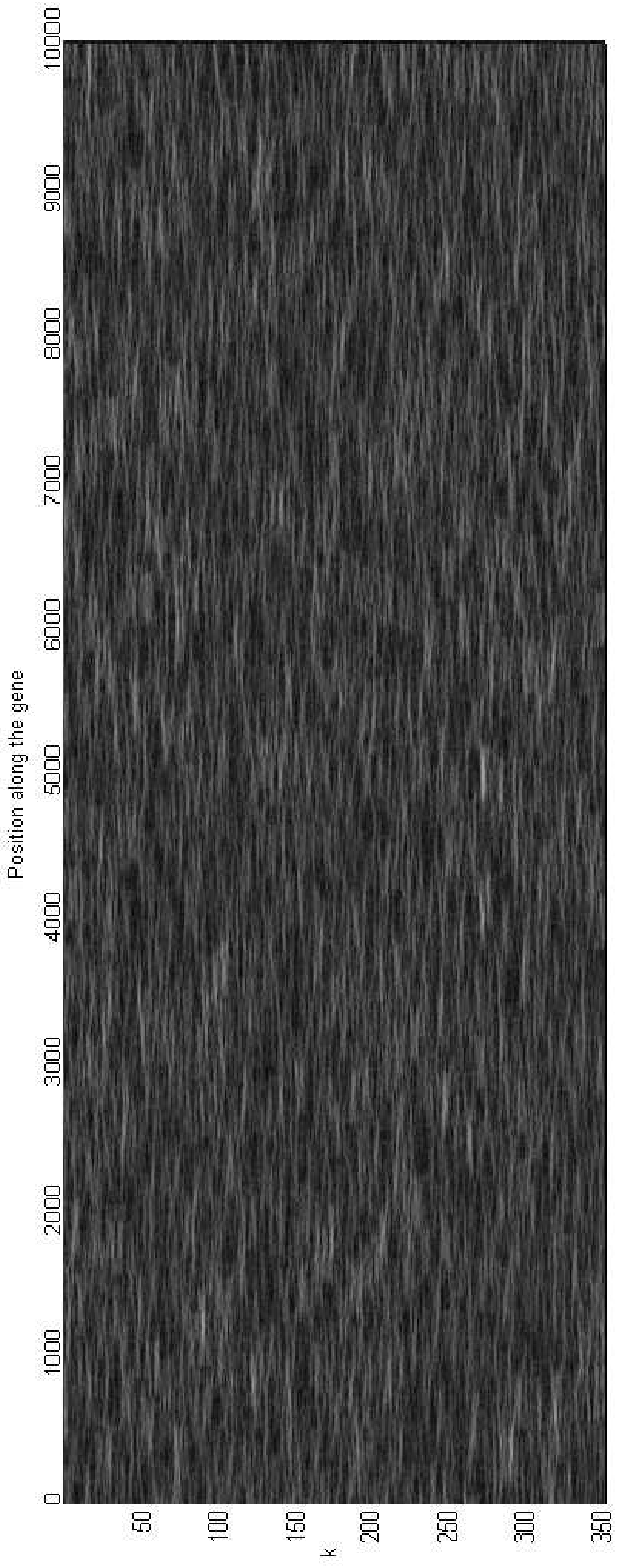


Figure 2.6: Spectrogram of a random sequence

As we can see in Figure 2.5, at $k = N/3$, we can see bright spots corresponding to the protein coding regions. Compare this with the peaks in Figure 2.1. It it clear that the three peaks in the central portion of Figure 2.1 have translated into the three bright spots. The other two peaks appear slightly less bright in the spectrogram but are still discernible.

Chapter 3

The p53 protein and its role in cancer suppression

The p53 protein is a transcription factor that regulates cell cycle and hence functions as a tumor suppressor. It prevents proliferation of genome mutation and is hence referred to as the *guardian of the genome*

3.1 The p53 protein - Structure

The human p53 protein is 393 amino acids long and consists of three principle domains as shown in Figure 3.1:

- An N-terminal transcription-activation domain (TAD), which activates transcription factors
- A central DNA-binding core domain (DBD). Contains zinc molecules and Arginine Amino Acid Residues
- A C-terminal homo-oligomerisation domain (OD). Tetramerization greatly increases the activity of p53 in vivo

Of these the most important region would be the central DNA-binding core domain. This is the region that binds to the DNA and activates the synthesis of proteins necessary for suppressing cancer. Hence mutations that deactivate p53 occur mostly in this region and destroy its ability to bind to target DNA sequences, thus preventing activation of those genes.

3.2 The p53 protein - Functions

The main function of the p53 protein is cancer suppression. As seen in Section 3.1, it binds to specific sites on the DNA sequences and activates the correspond-

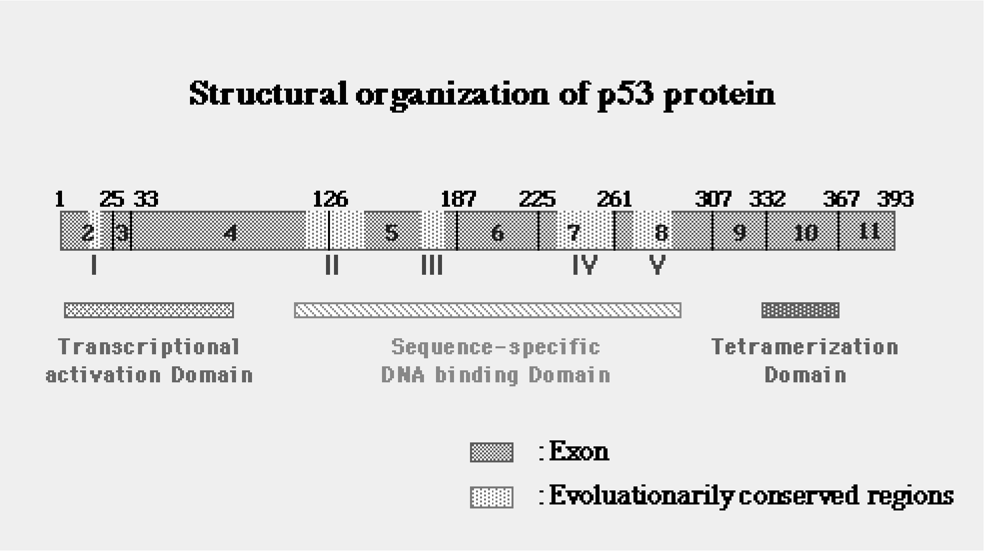

Structural organization of p53 protein
1
25
33
126
187
225
261
307
332
367
393
2
3
4
5
6
7
8
9
10
11
I
II
III
IV
V
Transcriptional activation Domain
Sequence-specific DNA binding Domain
Tetramerization Domain
: Exon
: Evoluationarily conserved regions

Figure 3.1: A schematic showing the different domains of the p53 protein [18]

ing genes for the synthesis of proteins. These proteins carry out the necessary functions. Hence p53 controls the cell cycle through these proteins.

p53 maintains cell stability through multiple pathways like :

- Activating DNA repair proteins when DNA has sustained damage
- Holding the cell cycle at the regulation point on DNA damage recognition so that the DNA repair proteins will have time to fix the damage and the cell can continue cell cycle
- Initiating apoptosis, the programmed cell death, if DNA damage proves to be irreparable

In a normal cell, p53 is bound to MDM2, and hence is inactive. DNA damage occurs due to a various reasons like exposure to UV radiation, carcinogenic chemicals through smoking or DNA damaging drugs. DNA damage is sensed and causes certain proteins to inhibit MDM2 and this activates the p53 protein. Once activated, p53 will either induce a cell cycle arrest to allow repair and survival of the cell or apoptosis to discard the damaged cell. The functioning of p53 is shown schematically in Figure 3.2

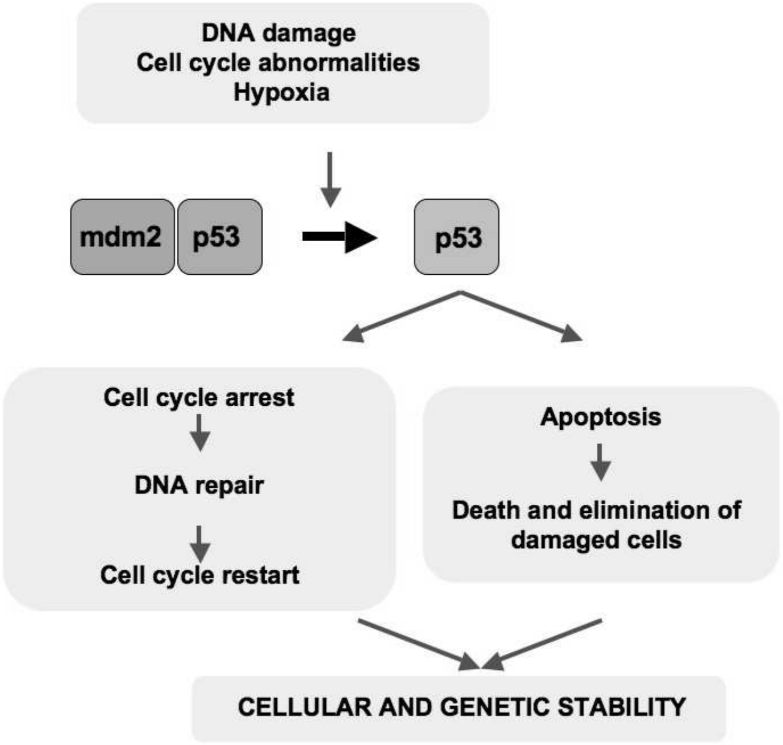


Figure 3.2: A schematic showing the function of the p53 protein

3.3 p53 binding sites

As we saw in Section 3.2, the p53 protein binds to certain regions on the DNA sequence and activates the corresponding genes to synthesize proteins. Such sequences of the DNA to which the p53 protein can bind are called *p53 binding sequences* or *p53 binding sites*. These sequences are usually found in and ”around” genes which produce proteins which regulate the cell cycle.

The genes containing p53-binding sites are those which play an important role in cancer suppression. Cancer is caused not only by mutations in the p53 protein. Cancerous mutations in genes with p53 binding sites can also cause cancer since the proteins they produce will not be able to regulate the cell cycle.

The p53-binding sequences are known to have lengths varying from 20 base pairs to 34 base pairs. The first 10 and the last 10 base pairs together form the *p53-binding site template*. Insertions up to 14 base pairs are made between these two regions. The combined physical structure of the binding sequence should be such that it can bind only to the p53 protein. This imposes certain constraints on the bases that can constitute the sequence. But the exact form of such constraints is not yet known.

Research in the area of p53 binding sites aims to develop a way of classifying any given DNA sequence as p53-binding or non p53-binding. The method studied and presented in this thesis will provide a way by which one can find all the genes that play an important role in suppressing cancer. This knowledge is very useful for research in gene therapy which seeks to treat diseases by correcting defective genes responsible for disease development.

Chapter 4

Locating putative p53-binding sites in DNA sequences

4.1 Introduction

As described in Section A.3 proteins are synthesized from genes only when they are expressed/activated. In the case of genes that regulate cell cycle, this is controlled by the p53 protein. This protein binds to the DNA sequence and activates the gene for protein synthesis. The sites on the DNA that the p53 can bind to are called p53-binding sites.

In this Chapter, we study a method of scoring any given DNA sequence based on it's probability of being p53-binding. This will provide a method by which one can find the genes that play an important role in suppressing cancer. This knowledge is very useful for research in gene therapy.

4.2 p53-binding DNA sequences

The basic structure of the p53-binding sequences was described in Section 3.3. The statistical properties of the first ten bases are the same as those of the last ten bases. Every position in the template has a probability distribution for the base that can occur at that position. Certain positions are very specific with respect to the bases that can occur at those positions. For example the fourth (and the fourteenth) position has a very high probability for the occurrence of the base C. Certain positions are less specific regarding the bases that can occur.

Hence the problem is to learn the probability distributions for the different positions using sequences known to be p53-binding and use them to score any given DNA sequence based on the probability that it is p53-binding. This suggests the use one of the many machine learning methods available in the literature [10]. We use a Hidden Markov Model based approach, called the profile Hidden Markov Model, because of its closeness to the binding sequence model and its computational efficiency, which is an important criterion given the large amounts of data that we intend to use it on.

4.3 The Hidden Markov Model - Description

4.3.1 The Markov Model

A Markov Model consists of a set a states each of which is associated with a particular symbol. The system begins in a particular state depending on a probability distribution called the *initial state distribution*. Transitions from one state to another are controlled by a *state transition matrix*. When the system is in a state, it emits the symbol corresponding to that state. The sequence of symbols thus emitted will be a Markov sequence, i.e

$$P(O_{n+1}|O_n, O_{n-1}, \ldots, O_1) = P(O_{n+1}|O_n)$$

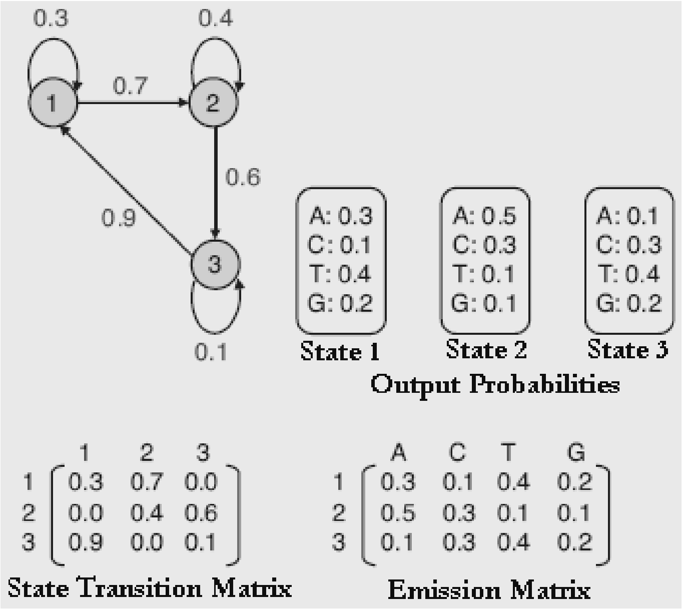
where O_i denotes the observation symbol at time instant i.

4.3.2 The Hidden Markov Model

The model described in Section 4.3.1 is constrained because the state sequence and the sequence of symbols should both be Markov since the difference between the state and the emitted symbol is only one of notation. Hence this model is extended to form the Hidden Markov Model (HMM) [4].

A schematic of a simple HMM is shown in Figure 4.1. In addition to the initial probability vector (denoted by π) and the state transition matrix (denoted by A) defined above, a HMM consists of an emission probability matrix (denoted by B) which gives the probability of a particular symbol being emitted in a particular state. Hence the HMM, denoted by λ, is characterized by the three entities, π, A

and, B.



$$\begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline 1 & 0.3 & 0.7 & 0.0 \\ 2 & 0.0 & 0.4 & 0.6 \\ 3 & 0.9 & 0.0 & 0.1 \end{array} \qquad \begin{array}{c|cccc} & A & C & T & G \\ \hline 1 & 0.3 & 0.1 & 0.4 & 0.2 \\ 2 & 0.5 & 0.3 & 0.1 & 0.1 \\ 3 & 0.1 & 0.3 & 0.4 & 0.2 \end{array}$$

Figure 4.1: Schematic of a simple Hidden Markov Model

To use HMMs, three basic problems need to be solved. They are :

1. Given an observation sequence, how do we adjust the model parameters (λ) such that the probability of the sequence being emitted by the HMM is maximized ?

2. Given an observation sequence and the model, how do we choose a corresponding state sequence which is optimal ?

3. Given an observation sequence, and a HMM, how do we efficiently compute the probability that the sequence was emitted by the given HMM ?

The first problem is that of training the HMM. We iteratively adjust the parameters λ using the *Baum Welch* or the *Expectation Maximization (EM)* algorithm [8]. The second problem attempts to learn the hidden part of the model. If the states have some physical meaning, then knowing the state sequence can provide information about the workings of the system. This problem is solved using the *Viterbi's algorithm* [7]. The third problem is that of scoring a given sequence with respect to the model. In the present case, since we are using the HMM to model p53-binding sequences, the score should indicate the probability of the sequence being p53-binding. This problem is solved using the *forward-backward algorithm* [6].

4.4 The Hidden Markov Model - Algorithms

This section describes the algorithms [4] for solving the problems related to HMMs that were described in Section 4.3.2. The notations used in this section are as follows.

N	Number of states in the model
q_t	The state at time instant t
S_i	The ith element of the set of all states S
O_t	The observation symbol at time instant t
v_i	The ith element of the set of all symbols V
B	Observation symbol probability distribution
A	Transition probability distribution
π	Initial state distribution
λ	The HMM, $\lambda = (A, B, \pi)$

$$b_j(k) = P(O_t = v_k | q_t = S_j)$$

$$\pi_i = P(q_1 = S_i)$$

$$a_{ij} = P(q_{t+1} = S_j | q_t = S_i)$$

4.4.1 The forward-backward algorithm

The forward-backward algorithm [6] is used to calculate the probability of the observation sequence $O = O_1 O_2 \ldots O_T$ given the model λ, i.e, $P(O|\lambda)$. The straight forward procedure is computationally expensive and hence we used the forward-backward approach. Consider the forward variable $\alpha_t(i)$ defined as

$$\alpha_t(i) = P(O_1 O_2 \ldots O_t, q_t = S_i | \lambda)$$

We can solve for $\alpha_t(i)$ inductively, as follows

1. Initialization:
$$\alpha_1(i) = \pi_i b_i(O_1) \qquad 1 \leq i \leq N$$

2. Induction:
$$\alpha_{t+1}(j) = \{\sum_{i=1}^{N} \alpha_t(i) a_{ij}\} b_j(O_{t+1}) \qquad {\scriptstyle 1\leq t\leq T-1 \atop 1\leq j\leq N}$$

3. Termination:
$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

Step 1 is clear from the definition of $\alpha_t(j)$. The induction step is the most important step in the algorithm. The i^{th} term in the summation is $P(O_1O_2\ldots O_t, q_{t+1} = S_j, q_t = S_i|\lambda)$. Hence the result of the summation is $P(O_1O_2\ldots O_t, q_{t+1} = S_j|\lambda)$ which when multiplied with $b_j(O_{t+1}$ given $\alpha_{t+1}(j)$. In the termination step, each $\alpha_T(i)$ is equal to $P(O, q_T = S_i|\lambda)$. Hence the result.

4.5 The Viterbi Algorithm

The Viterbi algorithm [7] is used to find the optimal state sequence given the model and the observation sequence, i.e $Q = q_1q_2\ldots q_T$ such that $P(Q|O,\lambda)$ is maximized. This is equivalent to maximizing $P(Q, O|\lambda)$. Define the quantity,

$$\delta_t(i) = \max_{q_1,q_2,\ldots,q_{t-1}} P(q_1q_2\ldots q_t = i, O_1O_2\ldots O_t|\lambda)$$

i.e, $\delta_t(i)$ is the highest score along a single path, at time t, which accounts for the first t observations and ends in state S_i. By induction we have

$$\delta_{t+1}(j) = [\max_i \delta_t(i) a_{ij}] \centerdot b_j(O_{t+1})$$

To retrieve the state sequence, we need to keep track of the argument i that maximizes the above expression. Hence the algorithm for finding the best state sequence would be as follows

1. Initialization:

$$\delta_1(i) = \pi_i b_i(O_1) \qquad 1 \leq i \leq N \psi_1(i) = 0$$

2. Induction:

$$\delta_{t+1}(j) = \max_{1\leq i\leq N} [\delta_t(i) a_{ij}] b_j(O_{t+1}) \qquad {}^{1\leq t\leq T-1}_{1\leq j\leq N}$$

$$\psi_{t+1}(j) = \underset{1\leq i\leq N}{\operatorname{argmax}} [\delta_t(i) a_{ij}] \qquad {}^{1\leq t\leq T-1}_{1\leq j\leq N}$$

3. Termination:

$$P^* = \max_{1\leq i\leq N} [\delta_T(i)]$$

$$q_T^* = \underset{1\leq i\leq N}{\operatorname{argmax}} [\delta_T(i)]$$

4. Path trace-back:

$$q_t^* = \psi_{t+1}(q_{t+1}^*) \qquad t = T-1, T-2, \ldots, 1$$

The steps above constitute the Viterbi algorithm which has been applied to a variety of problems. As can be inferred, ψ is the array that stores the index of the state that maximizes the probability at each step for all paths. Finally in the back tracking step, the indices for the optimal path are obtained from ψ.

4.6 The Baum Welch Algorithm

The Baum Welch or the Expectation Maximization [8] [13] algorithm solves the problem of training the HMM. Here we need to adjust the parameters of $\lambda = (A, B, \pi)$ to maximize the probability of the observation sequence given the model. Since finding λ, to maximize this probability globally is not possible, we try to choose λ such that $P(O|\lambda)$ is locally maximized. We need to define two quantities, β and ξ :

$$\beta_t(i) = P(O_{t+1} O_{t+2} \ldots O_T | q_t = S_i, \lambda)$$

$\beta_t(i)$ is called the backward variable. Again $\beta_t(i)$ can be found inductively as follows:

1. Initialization:

$$\beta_T(i) = 1 \qquad 1 \leq i \leq N$$

2. Induction:

$$\beta_{t-1}(i) = \sum_{j=1}^{N} a_{ij} b_j(O_t) \beta_t(j) \qquad t = T, T-1, \ldots, 2, 1 \leq i \leq N$$

Also, $\xi_t(i,j)$ is the probability of being is state S_i at time t and state S_j at time $t+1$ given the model and the observation sequence, i.e,

$$\xi_t(i,j) = P(q_t = S_i, q_{t+1} = S_j | O, \lambda)$$

We can write $\xi_t(i,j)$ in terms of the forward and the backward variables as:

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{P(O|\lambda)}$$

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}$$

Also, define $\gamma_t(i)$ as the probability of being in state S_i at time t, given the observation sequence and the model. Hence:

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j)$$

Note that $\sum_{t=1}^{T-1} \gamma_t(i)$ is the expected number of transitions from S_i and $\sum_{t=1}^{T-1} \xi_t(i,j)$ is the expected number of transitions from state S_i to state S_j, where the expectation is over time.

Using the above formulae, we can give a method for re-estimation of the parameters of the HMM. The set of re-estimation formulas are:

$$\hat{\pi}_i = \gamma_1(i)$$

$$\hat{a_{ij}} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

$$b_j\hat{(k)} = \frac{\sum_{\substack{t=1\\O_j=v_k}}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(i)}$$

It has been shown that [13] [12] that either:

1. The initial model λ defines a critical point of the likelihood function, in which case $\hat{\lambda} = \lambda$, or

2. The model $\hat{\lambda}$ is more likely than the model λ in the sense that $P(O|\hat{\lambda}) > P(O|\lambda)$

If we iteratively update λ with $\hat{\lambda}$, we can improve the probability of O being observed given the model until some limiting point is reached. It should be noted that this procedure yields only the local maxima, and in most problems of interest, the optimization surface is very complex and will have many local maxima [4].

4.7 The profile Hidden Markov Model

4.7.1 Introduction

It was noted in Section 4.3 that HMMs can be used to model sequences whose internal states are hidden. This is used in modelling various kinds of sequences like in speech recognition, protein modelling etc. This can be used for modelling p53-binding sequences as well. However a modification of this called the profile Hidden Markov Model (pHMM) [11] proves to be beneficial in terms of computational efficiency and also fits the p53-binding sequence model better.

4.7.2 The pHMM structure

In certain problems like p53-binding sequence modelling, the profile/template of the sequence is fairly well known. In such cases, one can model the sequences better by constructing a HMM which resembles the template. For example, based on the template of the p53-binding sequences as explained in Section 4.2, a pHMM to model such sequences is given in Figure 4.2. The system begins at the ’start’ state and terminates at the ’end’ state.

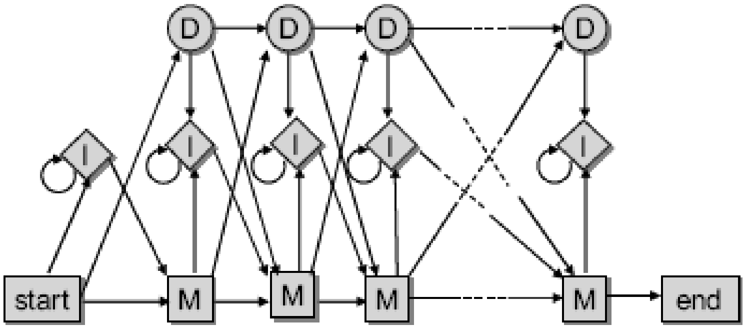


Figure 4.2: A state diagram of a profile Hidden Markov Model [11]

The model has three kinds of states denoted by M, I and D. The states M are called the *match states* and corresponding to the 20 base pairs in the template. Hence there are 20 M states. The states I are called *insertion states*. They are used to model insertions between the template symbols. Although it was mentioned in Section 4.2 that insertions take place only after the 10^{th} template symbol and before the 11^{th} template symbol, I states are included between each pair of adjacent template symbols to make the model more general. The D states are called *deletion states* and are used to model deletions of the template symbols.

4.7.3 Extending the pHMM Model

The pHMM of Section 4.7.2 can be initialized to a uniform distribution, i.e, the transition and emission probability distributions are uniform. It is then trained with sequences known to be p53-binding to update the parameters of the model. This way we are incorporating statistical information about p53-binding sequences into the model. But we can improve the results obtained if we make use of other pieces of information about p53-binding sequences. One such method is to filter the results obtained based on concept of *offending bases*.

It has been found that [14] certain bases at certain positions are incompatible with p53-binding. For example, base A cannot occur at the 7^{th} position on the template. Such bases are called offending bases. This information can be used to filter out all sequences that have been detected as p53-binding and have offending bases. Note that since we are starting with a uniform model, even if none of the training sequences has such bases, the emission probabilities will still be finite for offending bases. Hence the sequences should be filtered after they are obtained from the pHMM.

4.8 Simulations and Discussions

4.8.1 Training the model

The pHMM model described in Section 4.7.2 was trained with sequences obtained from the database of the *Laboratory of Statistical Genetics at the Rockefeller University* [15]. The set of training sequences was divided into 10 sets. The distance measure (ratio of difference between the lowest positive detection score and the highest negative detection score to the highest negative detection score) was calculated after every stage of training. Here the lowest positive detection score is the lowest score that was obtained for a sequence known to be p53-binding and the highest negative detection score is the highest score obtained for a sequence known to be non-p53-binding. The distance measure as a function of the number of training sets used is plotted in Figure 4.3.

4.8.2 Thresholding

The scores (obtained as the ratio of the probability of the sequence being emitted by this model to the probability of the sequence being emitted by a random model) for the human gene *acinus* is shown in Figure 4.4.

As we can see, there are a few peaks in the plot which correspond to p53-binding sequences. But for classifying any given DNA sequence as p53-binding or non-p53-binding, we need to define a suitable threshold for the scores obtained, above which the sequence is declared p53-binding. It should be noted that true positives and false positives both increase with decreasing threshold. Figure 4.5, shows the variation of true positives with the threshold. It is clear the the threshold should appear before the steep part of the curve, i.e, before 5.5. Also, the tolerance

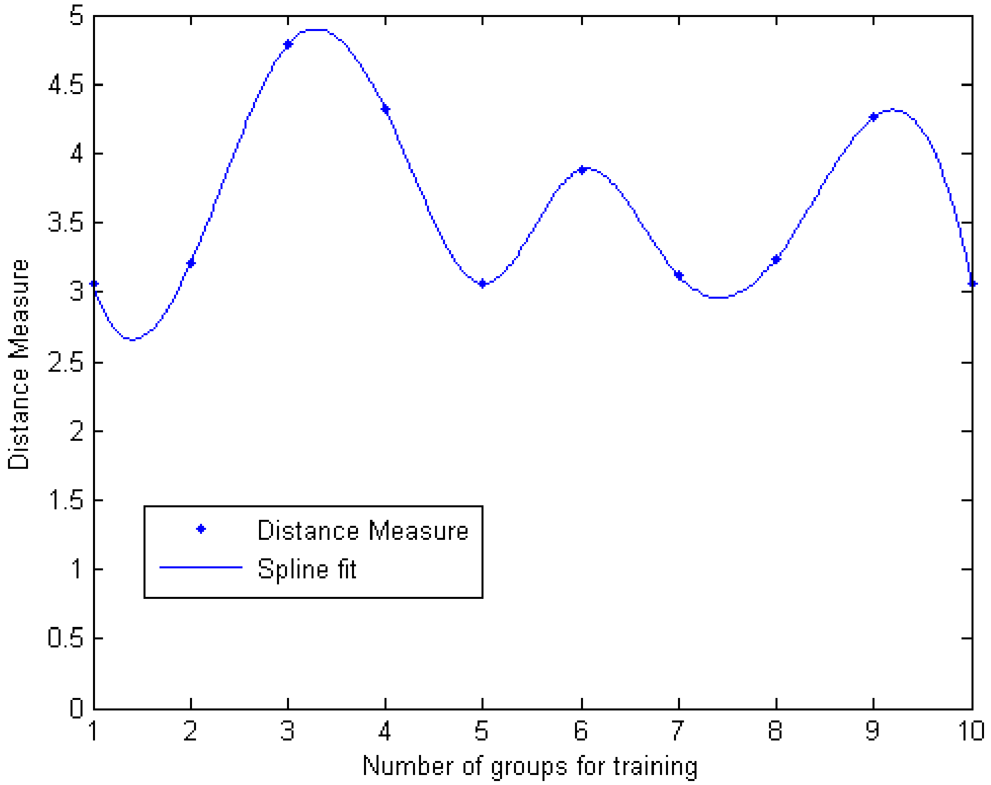


Figure 4.3: Distance measure as a function of the number of training sets

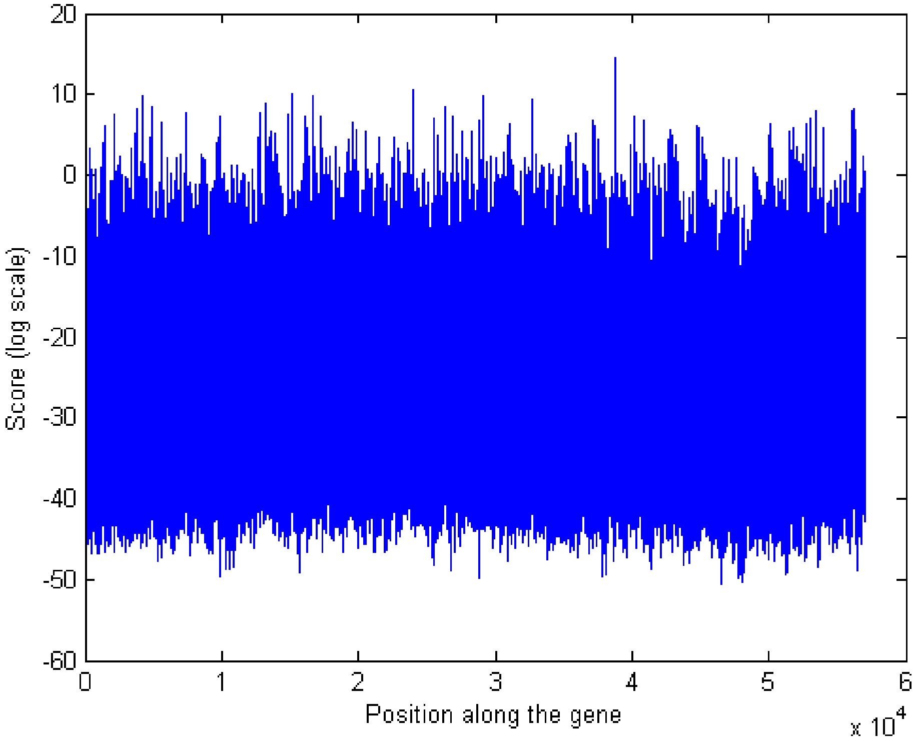


Figure 4.4: Scores obtained for the gene acinus

for true negative is much smaller than that for false positive since we would not want to lose any p53-binding sequences in the first step.

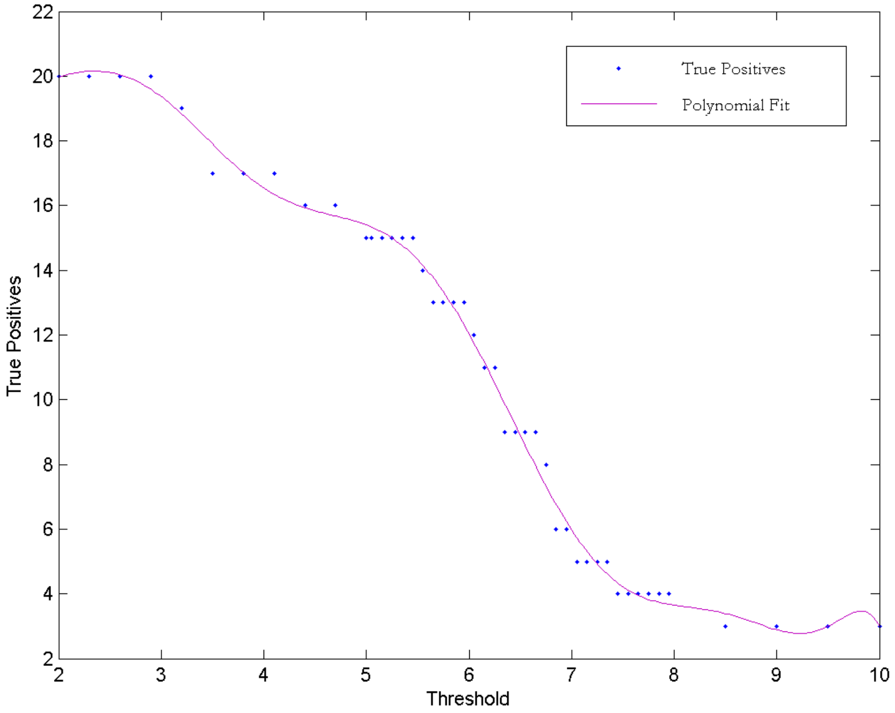


Figure 4.5: The variation of the number of true positives with the threshold

4.8.3 Filtering

As mentioned in Section 4.7.3, filtering of the sequences obtained from the model can be used to improve results. This is due to a decrease in the number of false positives after filtering. This is illustrated in Figure 4.6 where the false positives are plotted as a function of threshold with and without filtering.

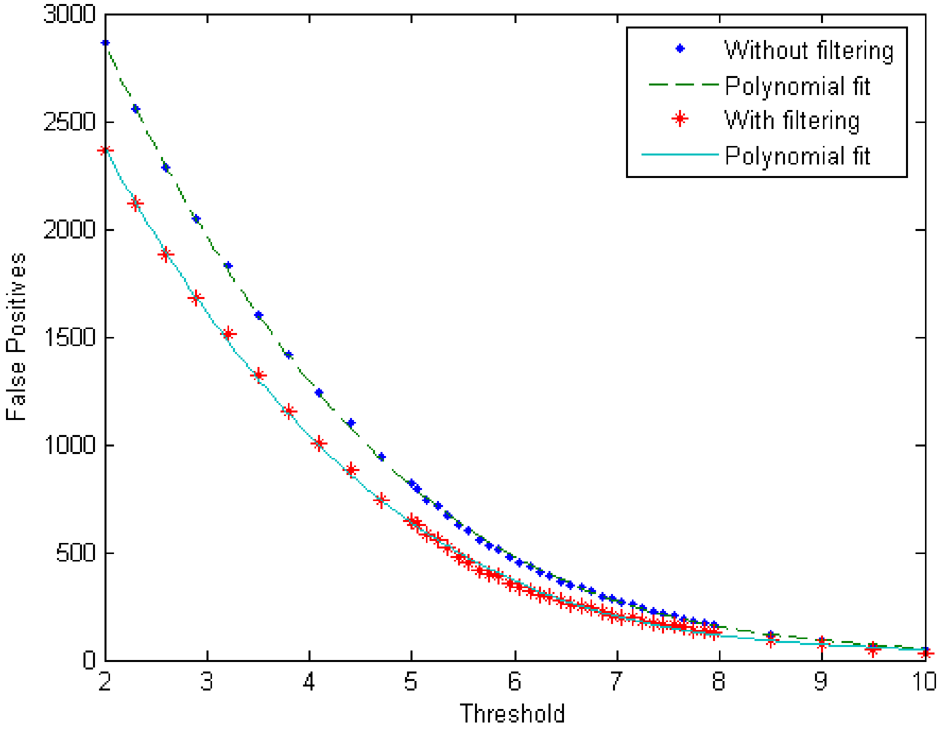


Figure 4.6: The number of false positives before and after filtering as a function of the threshold

Chapter 5

Conclusion and Future Work

In this project, problems in Genomics were considered. We studied the various ways of applying signal processing techniques to obtain useful information from biological sequences, particularly DNA sequences. The fourier transform technique and the Spectrogram technique for finding coding regions in a DNA were demonstrated using computer simulations. We then considered the problem of locating p53-binding sequences in a gene. The pHMM approach was used to solve this problem. Computer simulations were used to conduct experiments for training the model, finding optimum threshold values and to demonstrate the filtering technique used.

The p53 model discussed can be extended/modified to yield better results. For example, more filtering techniques can be explored but this would require a study of the biological literature regarding p53-binding sequences. In the model that was used, the deletion states were not used since all the sequences were of length greater than the template length. If the template length is increased, the deletion states will also be used.

The pHMM model discussed is quite general and can be used for various other applications one of which is protein modelling. Here pHMMs can be used to classify proteins based on their *family*, *functions* etc [16] [17]

Appendix A

DNA, Genes and the Genetic Code

A.1 Deoxyribonucleic acid (DNA)

DNA is a double stranded molecule as shown in Figure A.1 which is in the form of a double helix. Between the two strands of the backbone made of sugar-phosphate, there are pairs of bases Adenine, Guanine, Cytosine and Thymine. There are about three billion of these bases in a single human cell.

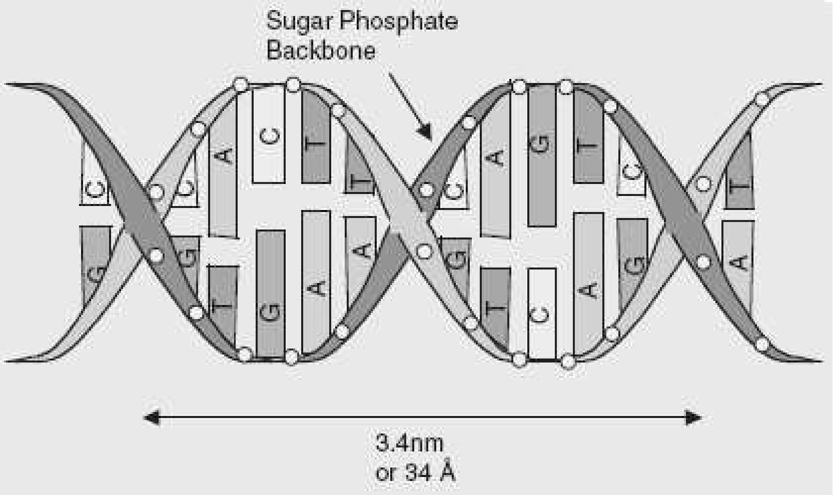


Figure A.1: The DNA double helix structure [1]

An opened up section of a DNA is shown in Figure A.2. The genome sequence corresponding to this example is $AGACTGAA$. The ordering is from the $5'$ end to the $3'$ end since they are scanned in that direction during the synthesis of amino acids. In the double stranded DNA, the base A always pairs with the T and the base C pairs with G. Thus the bottom strand contains no extra information and mostly only one of the two strands are active in gene expression.

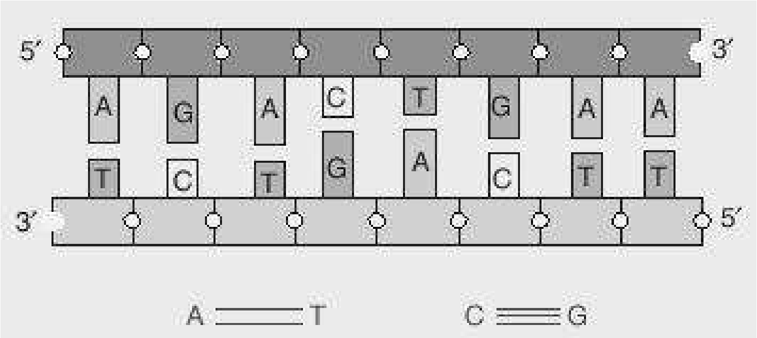


Figure A.2: An opened up section of a DNA sequence [1]

A.2 Genes

A DNA sequence can be separated into two regions namely genes and intergenic regions. Genes contain information for the generation of proteins. Each gene is responsible for the synthesis of a particular protein. Even though all the cells in a particular organism have the same DNA sequence and hence the same set of genes, only some selected genes are active in a particular family of cells. For example, the set of genes active in blood cells are different from those that are active in nerve cells which is reason they perform different functions. This is called gene expression and is a major topic of research.

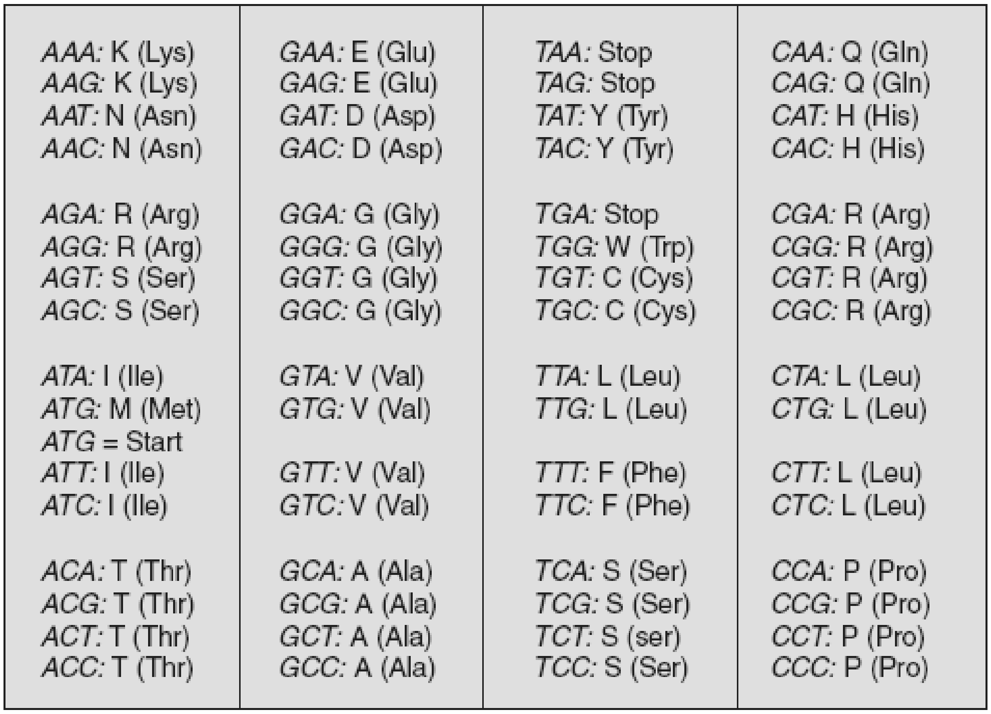
Genetic information flows from the DNA to the protein through another set of molecules called Ribonucleic acids (RNA). These are sequences similar to the DNA but are single stranded and are made up of bases Adenine, Guanine, Cytosine and uracil.

The gene is first copied on to a type of RNA called mRNA (messenger-RNA) and this moves out of the nucleus into the cytoplasm. This is used by a large molecule called ribosome for the synthesis of the appropriate protein. The translation from mRNA to protein is aided by molecules called tRNA (transfer-RNA) which store the genetic code.

A.3 The Genetic Code

The synthesis of amino acids based on the gene sequence is governed by a universal code, common to all life. This is called the Genetic Code, identification of which and associated work was awarded the 1968 Nobel Prize in Physiology or Medicine.

We saw in section A.2 that the DNA sequence is copied on to an mRNA and transferred out of the cytoplasm. This is called *transcription*. This is then used by the ribosomes to synthesize a string of amino acids which constitutes a protein. The conversion of mRNA into a sequence of amino acids is called *translation* and is governed by the Genetic Code. It is a three letter code, meaning that the bases in the mRNA (and hence the DNA) are divided into groups of 3 adjacent bases called a codon, each of which stands for a particular amino acid. Since the alphabet size is 4, there are 64 possible combinations but there are only 20 amino acids. Hence there exists a lot of redundancy in the system. The 64 codes with the corresponding amino acids is shown in Figure A.3.



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)	*GTG:* V (Val)	*TTG:* L (Leu)	*CTG:* L (Leu)
ATG = Start			
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)

Figure A.3: The Genetic Code [1]

The translation of the codon into an amino acid is physically made possible by the tRNA molecules. One end of the tRNA molecule matches the specific codon and the other end attaches to the corresponding amino acid as shown in

Figure A.4. The ribosomes work in conjunction with the mRNA and the tRNA to produce the protein. Hence it can be said that the tRNA molecules store the genetic code.

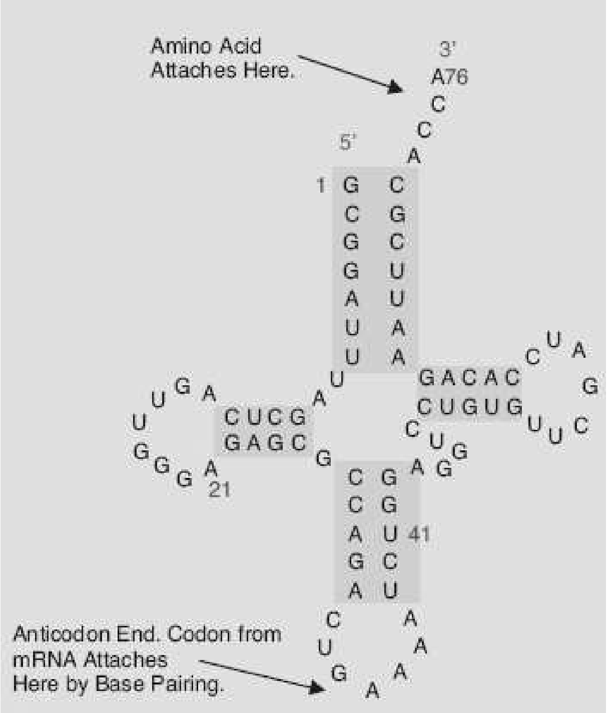


Figure A.4: A tRNA Molecule [1]

References

[1] Vaidyanathan, P.P., "Genomics and Proteomics: A Signal Processor's Tour", *Circuits and Systems Magazine, IEEE*, vol.4, issue 4, pp.6-29, Fourth Quarter 2004

[2] Anastassiou, D., "Genomic Signal Processing", *Signal Processing Magazine, IEEE*, vol.18, no.4, pp.8-20, July 2001

[3] Trifonov, E. N.,Sussman, J. L., "The pitch of chromatin DNA is reflected in its nucleotide sequence", *Proc. of the Nat. Acad. Sci., USA*, vol. 77, pp. 38163820, 1980

[4] Rabiner, Lawrence R., "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proceedings of the IEEE*, vol.77, no.2, pp.257-286, Feb 1989

[5] Riddle, D. L., Blumenthal, T., Meyer, B. J., Priess, J. R., "Introduction to C.elegans", *Cold Spring Harbor Laboratory Press, 1997*

[6] Baum, L. E., Egon, J. A., "An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology", *Bull. Amer. Meteorol. Soc.*, vol. 73, pp. 360-363, 1967

[7] Viterbi, A. J., "Error bounds for convolutional codes and an asymptotically optimal decoding algorithm", *IEEE. Trans. Informat. Theory*, vol. IT-13, pp. 260-269, Apr 1967

[8] Dempster, A.P., Laird, N.M., Rubin, D.B, "Maximum likelihood from incomplete data via the EM algorithm", *J. Roy. Stat. Soc.*, vol.39, no.1, pp.1-38, 1977

[9] Haussler, D., Krogh, A., Mian, I.S., Sjolander, K., " Protein modeling using hidden Markov models: analysis of globins", *System Sciences, 1993, Proceeding of the Twenty-Sixth Hawaii International Conference on*, vol.1, pp.792-802, 5-8 Jan 1993

[10] Krogh, A., Brown, M., Mian, I. S., Sjlander, K., Haussler, D., "Hidden Markov models in computational biology: Applications to protein modeling", *Journal of Molecular Biology*, 235, pp.1501-1531, 1994

[11] Huang, J., Shijun Li, "Mining p53 binding sites using profile hidden Markov model", *Information Technology: Coding and Computing, 2005. ITCC 2005. International Conference on*, vol.1, pp.146-151, 4-6 April 2005

[12] Baker, J., "The DRAGON system–An overview", *Acoustics, Speech, and Signal Processing, IEEE Transactions on*, vol.23, no.1, pp.24-29, Feb 1975

[13] Baum, L.E., Sell, G.R., “Growth functions for transformations on manifolds”, *Pac.J. Math.*, vol.27, no.2, pp.211-227, 1968

[14] Hoh, J., Jin, S., Parrado, T., Edington, J., Levine, A.J., Ott, J., “The p53MH algorithm and its application in detecting p53-responsive genes”, *Proc Natl Acad Sci U S A*, 99(13), pp.8467-8472, June 2002

[15] “Directory of p53 consensus DNA binding sites”, *http://linkage.rockefeller.edu/p53/*

[16] Friedrich, T., Pils, B., Dandekar, T., Schultz, J., Muller. T., “Modelling interaction sites in protein domains with interaction profile hidden Markov models”, *Bioinformatics*, vol.22, no.23 pp.2851-2857, Dec 2006

[17] Katsiapis, A. “Protein Modeling with Profile Hidden Markov Models”

[18] Linda, J. Ko, Prives, Carol, “p53 : puzzle and paradime”, *Genes and Development* no.10, pp.1054-1072, 1996