Gene Prediction System

by

Hazrina Yusof Hamdani
6498

Final Report Submitted In Partial Fulfillment Of
The Requirement For The
Bachelor of Technology (Hons)
(Information and Communication Technology)

JANUARY 2008

Universiti Teknologi PETRONAS
Bandar Sri Iskandar
31750 Tronoh
Perak Darul Ridzuan

CERTIFICATION OF APPROVAL

Gene Prediction System

by

Hazrina Yusof Hamdani

A project final report submitted to the
Information and Communication Technology Programme
Universiti Teknologi PETRONAS
in partial fulfillment of the requirement for the
Bachelor of Technology (Hons)
(Information and Communication Technology)

Approved by,

(Siti Rohkmah Mohd Shukri)

CERTIFICATION OF ORIGINALITY

This is to certify that I am responsible for the work submitted in this project, that the original work is my own except as specified in the references and acknowledgements, and that the original work contained herein have not been undertaken or done by unspecified sources or persons.

(HAZRINA YUSOF HAMDANI)

ABSTRACT

With the increasingly popularity of genome sequencing, transforming such raw sequence data into knowledge remains a hard task. This project will develop an application for gene prediction using development tools such as Perl and PHP. This project also will identify stretches of sequence for genomic DNA that is biologically functional including protein coding regions. There are 3 steps involve which are transcription, splicing and translation. Transcription is the process of copying DNA to RNA. Meanwhile, splicing is modification of genetic information after a few transcriptions. It will determine introns (useless information) and exons (useful information) in DNA sequence. Hidden Markov Model which is a mathematical functional will be used in this step in order to predict the DNA sequence. The introns will be removed and exons will be combined together to produce mRNA. The final step is translation which is the process by which mRNA is translated into proteins. The protein sequence is the final output of Gene Prediction System. More knowledge on gene prediction will be explored in order to get better understanding of how gene prediction works.

ACKNOWLEDGEMENT

Alhamdu Lillah, thank you to Almighty God, who made all things possible and for all the blessings. I would like to acknowledge and extend my heartfelt gratitude to the following persons who have made the completion of this final year project possible: my supervisor, Siti Rohkmah Mohd Shukri, for her vital encouragement, support, assistance and for always been there for me. Thank you for being my role model and mentor. Dr Mohd Azmuddin Bin Abdullah, my co-supervisor for support and assistance. I also would like to thank members and staffs of Computer and Information Sciences Department. Most especially thank you to my beloved family, my father, Yusof Hamdani Jalalludin, my mother, Hasmawati Hassan, my brothers and sister for unconditional support, encouragement to pursue my interests and also motivation to finish up the project. Thank you to all my friends for the constant reminders and much needed motivation. Thank you very much.

Hazrina Yusof Hamdani
April 11, 2008

TABLE OF CONTENTS

List of Figures

List of Tables

1 INTRODUCTION

1.1 Background of the Project

Due to the steadily growth of bioinformatics and related scientific discipline in bioinformatics, there have been some research and projects had and planned to be conducted based on the field. Furthermore, it would be interesting to do a project on gene prediction information system for this final year project. What is bioinformatics? Bioinformatics is combination of biology field and information technology field. Westhead says that bioinformatics is the marriage of biology and information technology [5]. The discipline that includes any computational tools and methods used to manage, analyze and manipulate large sets of biological data. Gene prediction is the area of computational biology that is concerned with algorithmically identifying stretches of sequence, usually genomic DNA, that are biologically functional. Gene prediction or gene finding is the most important step to understand the genomic species once it has been sequenced.

1.2 Problem Statement

In earliest day, gene prediction was based on experimentation on living cells and organisms. Statistical analysis of the rates of homologous recombination of several different genes could determine their order on a certain chromosome, and information from many such experiments could be combined to create a genetic map specifying the rough location of known genes relative to each other.

Nowadays in Malaysia, bioinformatics is not established yet. By developing this project it will provide newcomers especially students in secondary school and people who love the biological field and computer field to understand the principles of bioinformatics applications and to gain some practice for their own usage.

A traditional way to find genes was to do experiments in the laboratory. The experiment conducted by researchers will extract and sequence the RNA because RNA serves as the template for translation of genes into protein. Unfortunately by using this traditional way, it will cause some problems such as a few genes will dominate the sequence and it will be hard to prevent duplication. In order to prevent such problems, a computational system of gene finding should and will be introduced.

1.3 Objectives

- To develop application for gene finding.
- To identify stretches of sequence for genomic DNA that is biologically functional including protein coding regions.
- To explore gene prediction knowledge.

1.4 Scopes

- Gene finding for Eukaryotes family.
- Development tools using Perl/ PHP
- Use combine predictive and comparative algorithms for gene finding.

2 LITERATURE REVIEW/ THEORY

2.1 Eukaryote

Eukaryote defines as organisms whose cells are organized into complex structures by internal membranes and a cytoskeleton. Eukaryote group are animals, plants, fungi and protests. All living creatures including Eukaryotes have DNA. To find DNA sequence using the simplest method is to search for open reading frame (ORFs). An ORF is a length of DNA sequence that contains a contiguous set of codons, each of which specifies an amino acid. (David Mount, 2001).

2.2 Structure of Gene

To reproduce themselves, organisms' information been stored in DNA. The term structure of gene mean a unit of information and corresponds to discrete segment of DNA. It encodes the amino acid sequence of a polypeptide. In higher organism such as human, the genes are present on a series of long DNA molecules called chromosomes. In human, 30 000 genes arranged on 23 choromosomes [19]. Before the gene (DNA) been transcript to RNA, the gene structures contains of promoter, exon and intron.

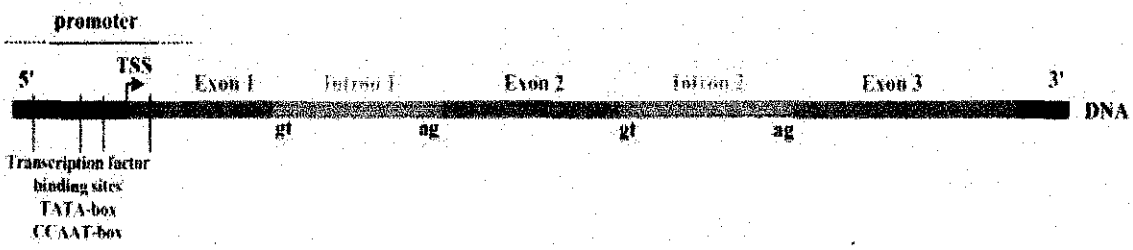


Figure 1: Structure of DNA.

The promoter contains a specific DNA sequences known as response elements such as TATA box and CCAAT box. TATA box is considered as core promoter sequence is frequently found in eukaryote. The sequence is TATAA.

There are differences between promoter in eukaryote and promoter in prokaryotes. Promoter in prokaryote is recognized by RNA polymerase, an enzyme that makes an RNA copy of a DNA or RNA template [20] and an associated sigma factor, a prokaryotic transcription initiation factor that enables specific binding of RNA polymerase to gene promoters [21]. Meanwhile in eukaryote the process is more complicated because it involves 7 different factors for the transcription.

Promoter in eukaryote is difficult to characterize because they typically lie upstream of the gene. The regulatory elements can up until several kilobases away from the transcriptional start site. As been mentioned before, many eukaryotic promoters contain a TATA box. TATA box lies very close to the transcriptional start site (within 50 bases) [22].

In eukaryote group such as animal and plant, gene are encoded in several pieces called exon and been separated by non- coding DNA segments called introns. Meanwhile in prokaryote group such as microbase, gene is encoded by a simple DNA segment. That is why the method used for predict gene and finding protein between eukaryote group and prokaryote group are different. The number and size of introns vary between genes and introns can be removed from RNA transcripts by a process call splicing [19]. The process of removing introns will be discussed later.

2.3 Deoxyribonucleic Acid (DNA)

DNA, which stands for deoxyribonucleic acid have huge capacity to store genetic information. It is the substance that makes our genes in a form of a large macromolecule consisting of a chain of four constituents that called nucleotide. A nucleotide is made up of one phosphate group linked to a pentose sugar which is itself linked to one of four types of nitrogenous organic bases. It symbolized by 4 letters A, C, G and T. Forming a bond between the 5' (5 *prime*) and 3' (3 *prime*) positions of the constituent nucleotides that make the DNA molecule. Example of of the DNA:

TGACT = Thymine-Guanine-Adenine-Cytosine-Thymine

Kendrew and Perutz discovered that the DNA molecule consists of two complementary strands strand: thymine (T) facing adenine (A) and guanine (G) facing cytosine (C). Rosetta stone explains about DNA sequences for example, when living organisms reproduce, each of their genes must be duplicated. In order to do this, nature doesn't go about it like a photocopier but it will make an exact copy.

2.4 Ribonucleic Acid (RNA)

RNA which stands for ribonucleic acid is a nucleic acid polymer consisting of nucleotide monomers. Its roles are to translate genetic information from DNA into protein products; RNA will acts as a messenger between DNA and ribosome (the protein synthesis complexes). RNA serves as the template for translation of genes into proteins, transferring amino acids to the ribosome to form proteins. RNA contains ribose sugars and it uses predominantly uracil. It has four different bases which are adenine, guanine, cytosine and uracil. RNA is a single-stranded molecule and has shorter chain of nucleotides.

2.5 Steps in Gene Finding

Gene finding consists of 3 steps which are transcription, splicing and translation. These steps describe the process of translation of a gene to a protein. It is call as Central Dogma principle. Central dogma states that information is transfer from DNA -> RNA -> protein and such information cannot be transferred back from protein to either protein or nucleic acid [18].

2.5.1 Transcription

Transcription is the process of copying DNA to RNA. The transcript of the gene is a molecule that must be processed to remove introns (extra sequences). Introns are bordered by donor and acceptor which are GT and AG. Introns break up the amino acid coding sequence into segments called exons. The transcript of these genes is called primary transcript or pre-mRNA. Pre-mRNA is processed in the nucleus to remove the introns and join the exons together into mRNA.

2.5.2 Splicing

Splicing is a modification of genetic information after a few transcriptions, in which introns of precursor messenger RNA (pre-mRNA) are removed, exons of it are joined, and matured mRNA has been created. mRNA is a molecule of RNA encoding a chemical "blueprint" for a protein product [23]. In mRNA as in DNA, genetic information is encoded in the sequence of four nucleotides arranged into codons of three bases each [23]. Codon is a word of 3 nucleotides been used when translating a DNA sequence into a protein.

Matured mRNA can be recognize by it structure. The mRNA structure contains of 5' cap, 2 untranslated regions which are 5' UTR and 3' UTR, coding region or coding sequence (CDS) and Poly A tail.

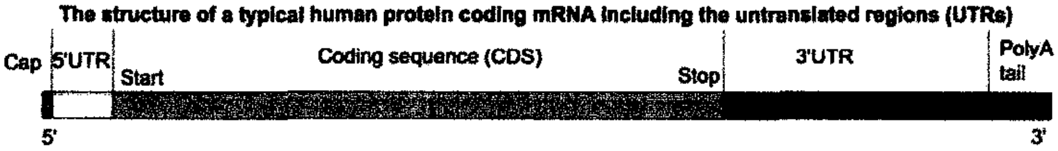

The structure of a typical human protein coding mRNA including the untranslated regions (UTRs)
Cap
5'UTR
Start
Coding sequence (CDS)
Stop
3'UTR
PolyA tail
5'
3'

Figure 2: Structures of mRNA

mRNA structure, start with 5' cap. 5' cap is a altered nucleotide end to the 5' end of mRNA (5' UTR). The 5' capping is important to create a mature mRNA which is then will be able to be put through the process of translation because the 5' cap ensure the mRNA's stability while it undergo translation in the process of protein synthesis [24].

As been mentioned before this, 5' UTR and 3' UTR are the untranslated region. UTR means untranslated region. 5' UTR and 3' UTR have it own roles in gene expression but the untranslated region also can be classified into several general roles which are the region will make sure the mRNA stability, mRNA localization and the region will make sure the translation work efficiently [23].

5' UTR (five prime untranslated region) known as the leader sequence. It starts at the +1 position which is the location where transcription begins and ends before the start codon of the coding region. The size of 5' UTR maybe a hundred or more nucleotide long [25]. Meanwhile, 3' UTR (three prime untranslated region follows the coding region [26].

The poly A tail is a long sequence of adenine nucleotides added to the “tail” of mRNA. The size of poly A tail often several hundred nucleotides long. The general role of poly A tail and also the 5' cap is to protect the mRNA [27]. In eukaryote, the poly A tail is added onto transcripts that contain a specific sequence, the AATAA signal [23]. The important of the poly A tail is to demonstrate by a mutation in the human alpha 2-globin gene that changes the original sequence AATAA into AATAAG, which can lead to hemoglobin deficiencies.

Finally the last structure in mRNA is coding region or coding sequences (CDS). Coding region is a composed of codons, which are decoded and translated into one (mostly eukaryote) and several (mostly prokaryote) proteins by the ribosome [23]. It will begin with the start codon (ATG) and end with the one of the three possible stop codons (TAA, TAG or TGA) [23]. The coding region also been called as open reading frame (ORF). Three possible sets of codon can be read from any sequence depending on which base is chosen as the start [19]. Each set of codons is known as a reading frame [19]. Theoretically, the coding region can be read in 6 reading frames in organisms with double- standed DNA (3 in th forward and 3 in the reverse direction) [28].

There are 3 types of mRNA molecule which are monocistronic, distronic and polycistronic. For eukaryotic, the mRNA molecule is monocistronic meanwhile for eukaryotic, the mRNA molecule is polycistronic. The mRNA molecul is said to be monocistronic when it is contains the genetic information to translate on a single protein [23]. Meanwhile for polycistronic mRNA carries the informations of several proteins, which are translated into several proteins [23]. It is mean that eukaryotic mRNA only contains a single open reading frame.

One of the methods to determine promoter, extron, intron, 5' cap, untranslated region, coding region (reading frame) and poly A tail is by using Hidden Markov Model (HMM). This method will find the coding and non-coding regions of unlabeled string of pre-mRNA.

2.5.3 Translation

Translation is the process by which messenger RNA is translated into proteins in eukaryotes.

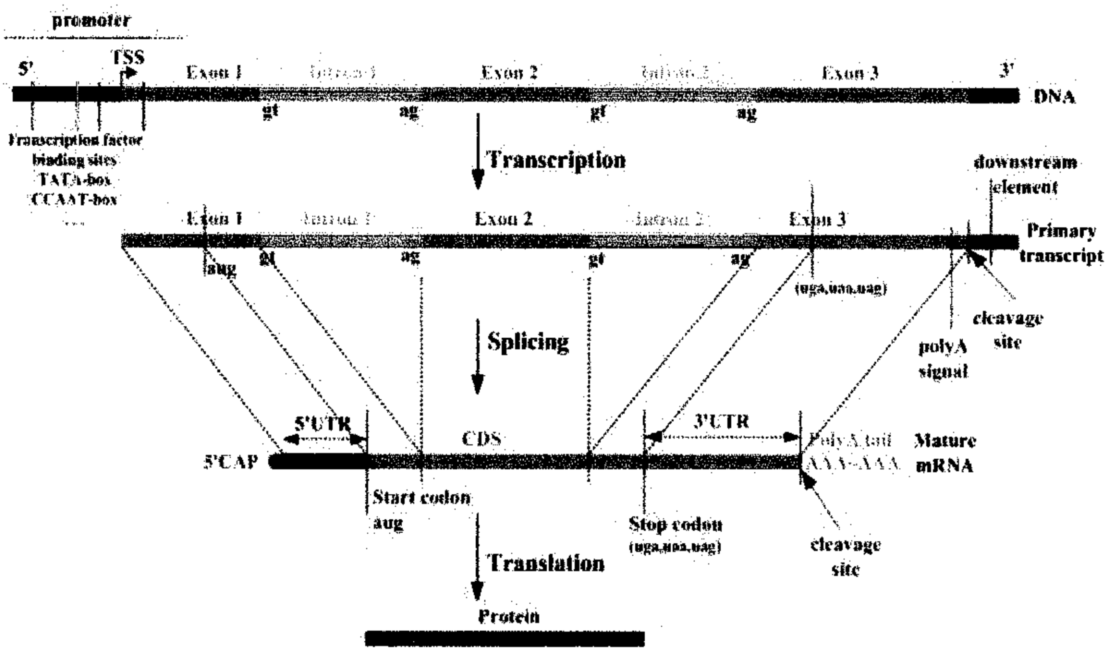
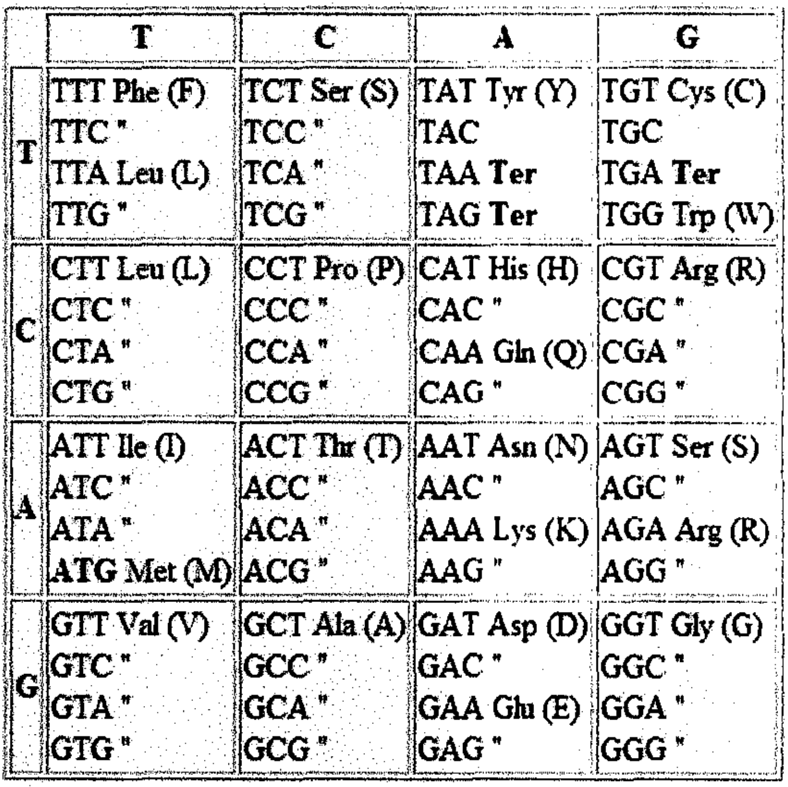


Figure 3: Transcription, Splicing and Translation Process

2.6 Amino Acid

To turn proteins into amino acid, we can use genetic code. The standard genetic code is universal and it uniquely relates A, T, G, and C to a suite of 20 amino-acid symbols.



	T	C	A	G
T	TTT Phe (F) TTC " TTA Leu (L) TTG "	TCT Ser (S) TCC " TCA " TCG "	TAT Tyr (Y) TAC TAA **Ter** TAG **Ter**	TGT Cys (C) TGC TGA **Ter** TGG Trp (W)
C	CTT Leu (L) CTC " CTA " CTG "	CCT Pro (P) CCC " CCA " CCG "	CAT His (H) CAC " CAA Gln (Q) CAG "	CGT Arg (R) CGC " CGA " CGG "
A	ATT Ile (I) ATC " ATA " ATG Met (M)	ACT Thr (T) ACC " ACA " ACG "	AAT Asn (N) AAC " AAA Lys (K) AAG "	AGT Ser (S) AGC " AGA Arg (R) AGG "
G	GTT Val (V) GTC " GTA " GTG "	GCT Ala (A) GCC " GCA " GCG "	GAT Asp (D) GAC " GAA Glu (E) GAG "	GGT Gly (G) GGC " GGA " GGG "

Table 1: Table of Standard Genetic Code

Genome sequencing centers will search newly sequences with gene prediction programs. After that, it will interpret the sequence database entry with this information. This annotation includes gene location, gene structure, which refers to positions of predicted exons/ introns and regulatitory sites, and any matches of the translated exons with the protein sequence database.

The amino acid sequence of the predicted gene may also be entered in the protein sequence databases. It is a good practice to reconfirm any gene prediction of interest and perform alignments of the predicted sequence with matching database sequence. [11]

2.7 Other Programs

Few systems that been developed by scientists in whole around the world just to predict the gene. Examples of gene prediction program that use several method such as ab initio and others are GENSCAN (Burger, 1997), GENIE (Kulp *et al.*, 1996), HMMGene (Krogh, 1997), GENEID (Parra *et al.*, 2000), GENEWISE (Birney and Durbin, 1997), PROCRUSTES (Gelfand *et al.*, 1996), GENOMESCAN (Yeh *et al.*,2001), AUGUSTUS (Stanke and Waack, 2003) and many more [34].

2.8 Hidden Markov Model

Hidden Markov Model or HMM is a machine learning approach that derives 'rules' from training data and applies them to new, uncharacterized test data to predict features similar to the ones learned. HMM is a mathematical formulation of a succession of hidden, mutually exclusive properties associated with one sequence or a multiple sequence alignment [1].

Hidden Markov Model has been used in several applications such as speech recognition, handwriting recognition gesture recognition and bioinformatics [29]. In bioinformatics especially in the context of DNA, HMM are used to predict the location and structure of genes.

For better understanding, the famous "Fair Bet Casino" will be used as example for basic HMM. [32]

Dealer flips a coin and player bets on the outcome: H (heads) or T (tails). Dealer may use a <u>fair</u> coin or a <u>biased</u> coin. The probability of using <u>Fair coin</u> but get heads or tails are:

(F) Fair: $p_F(H) = p_F(T) = ½$

(B) Biased; $p_b(H) = ¾$, $p_B(T) = ¼$

For security, the dealer switches between coins very rarely, only once every 5 times (example, probability of a switch is 0.2).

We can model this problem with a "machine". There are 2 possible states which are F-fair coin are used and B-biased coin is used. At the beginning of each step the m/c is in a *hidden* state. Why it is hidden? We can say it is hidden because it is unknown to us as observers. At each step, the m/c decides what *step* to take such as whether to switch between the coins or not and what *symbol* to emit either heads or tails.

There is a probability distribution associated with each of what step is chosen as the next step and what symbol to emit upon changing form one state to another. It can be represented with a 'machine' M= (Q, Σ a, $\in$) where Σ is alphabet of symbols ({H, T}). Q is set of states, each of which will emit symbols from Σ ({B, F}). a describing the probability of a transition from state s_k to state s_l . For example, a = (a_{BB},a_{BF},a_{FB},a_{FF}) with $a_{BB} = a_{FF} = 0.8$, $a_{BF} = a_{FB} = 0.2$. It can also be represented using the diagram below:

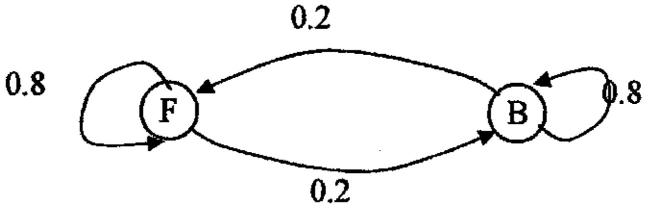


Figure 4: Representation of a

$\in$ describing the probability to emit symbol b in state s_k. $\in$ ($e_k(b)$ – a(N x $|\Sigma|$)matrix which are $e_F(H)$, $e_F(T)$, $e_B(H)$,$e_B(T)$. It also can be represented in a matrix form.

$e_F(H) = e_F(T) = ½$, $e_B(H) = ¾$,$e_B(T) = ¼$

$$e = \begin{array}{c|cc} & F & F \\ F & 0.8 & 0.2 \\ B & 0.2 & 0.8 \end{array}$$

A path $\Pi = \Pi_1 \Pi_2 .. \Pi_M \in \{F, B\}$ in the HMM is a sequence of states. For example, let x = THTHHTH and Π = FFBBBBB (M = 7). Then:

x	: T	H	T	H	H	T	H
Π	: F	F	B	B	B	B	B
$P(x_i \mid \Pi_i)$	: ½	½	¼	¾	¾	¼	¾
$P(\Pi_{i-1} \text{->} \Pi_i)$	: 0.8	0.8	0.2	0.8	0.8	0.8	0.8

The probabilities of path that can be used in HMMs are the probability of sequence x given path Π the probability of picking path Π the joint probability of Π and sequence x, the probability that sequence x is generated by the HMM and the probability that path Π is taken, given that x is observed.

But how likely is that sequence x was generated by M and how to determine the most likely path, given the data? The answers are using the Forward (sum) algorithm by calculating the P(x) and the Veterbi (max) algorithm to calculates Π_{max} that maximizes $P(x, \Pi)$[33].

The Forward algorithm goal is to determine the probability P(x) of generating sequence $x = x_1 \ldots x_M$ with M. The method used one dynamic programming. For example:

Let s_0 = designated start state, Q = $\{s_0, \ldots s_{N-1}\}$ and let $f_j(i)$ = probability of generating the prefix $x_1 \ldots. x_i$ and ending in states s_j. Below is pseudocode for the example and the final result that will get is $P(x) <- \sum_{k=0,|Q|-1} f_j(M)$

```
$f_j(i) = \sum_{p:s0..>sj} P(x_1 \ldots x_j, p)$ over all paths p form $s_0$ to $s_j$j
$f_0(0) <- 1; f_j(0) <- 0$ for j=1,...|Q|-1 ($s_0$=start state)

for i=1 to M do
 for j=1 to |Q| do
  $f_j(i) <- e_j(x_i) \sum_{k=0,|Q|-1} a_{k,j} f_k(i-1)$
```

So, $P(x) <- \sum_{k=0,|Q|-1} f_j(M)$

The Veterbi algorithm goal is to determine the probability P(x) of generating sequence $x = x_1 \ldots x_M$ with M. The method is by using dynamic programming. For example:

Let s_0 = designated start state, Q = $\{s_0, \ldots s_{N-1}\}$ and let $v_j(i)$ = max probability of generating the prefix $x_1 \ldots. x_i$ and ending in states s_j over all paths p: s_0 to s_j. Below is pseudocode for the example and the final result that will get is $\max_\Pi P(x, \Pi)$.

```
$v_j(i) = \max_{p:s0..>sj} P(x_1 \ldots x_j, p)$ over all paths p form $s_0$ to $s_j$j
$f_0(0) <- 1; f_j(0) <- 0$ for j=1,...|Q|-1 ($s_0$=start state)

for i=1 to M do
 for j=1 to |Q| do
  $v_j(i) <- e_j(x_i) \max_{k=0,|Q|-1} a_{k,j} v_k(i-1)$
```

So, $\max_\Pi P(x, \Pi)$ is max $_{j=0,|Q|-1}$ $v_j(M)$ and the Π_{max} will be retrieve using trace back pointers.

Forward algorithm will calculate the probability for each path that will be taken to reach the final data. Meanwhile, Viterbi algorithm will determine the path that will be taken. For better understanding, let see figure 5. Figure 5 will represent the Forward and Viterbi algorithms. The blue arrow will represent the probability for each path and. Meanwhile for figure 6, the yellow circle and yellow arrow will occur after the path has been determined using Viterbi algorithm. The final result for this example is:

X: T H H H H
Result: F B B F F

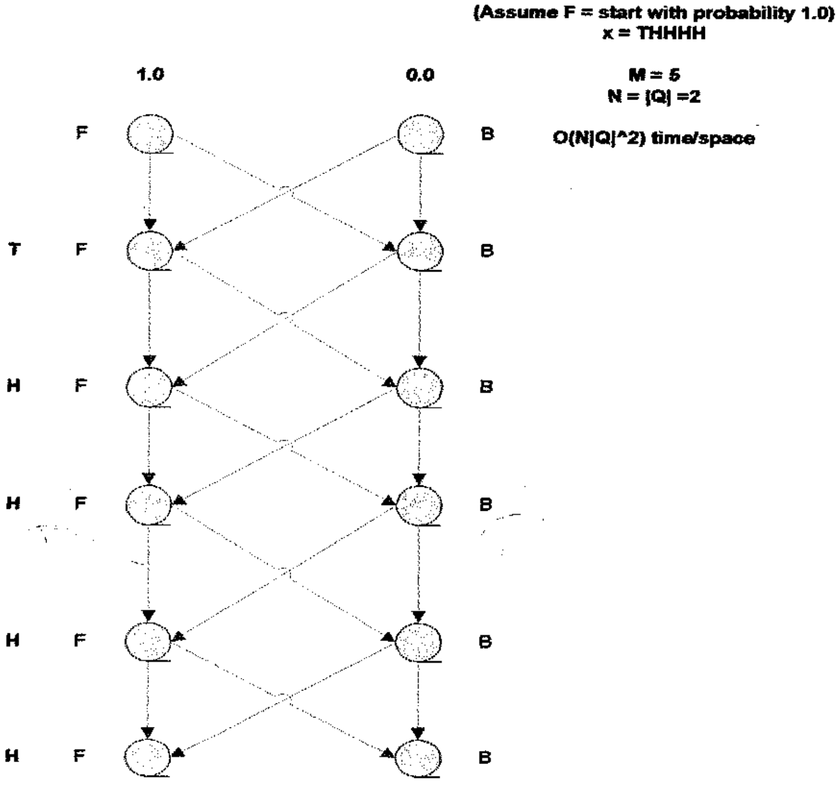


Figure 5: The network for the Forward algorithms

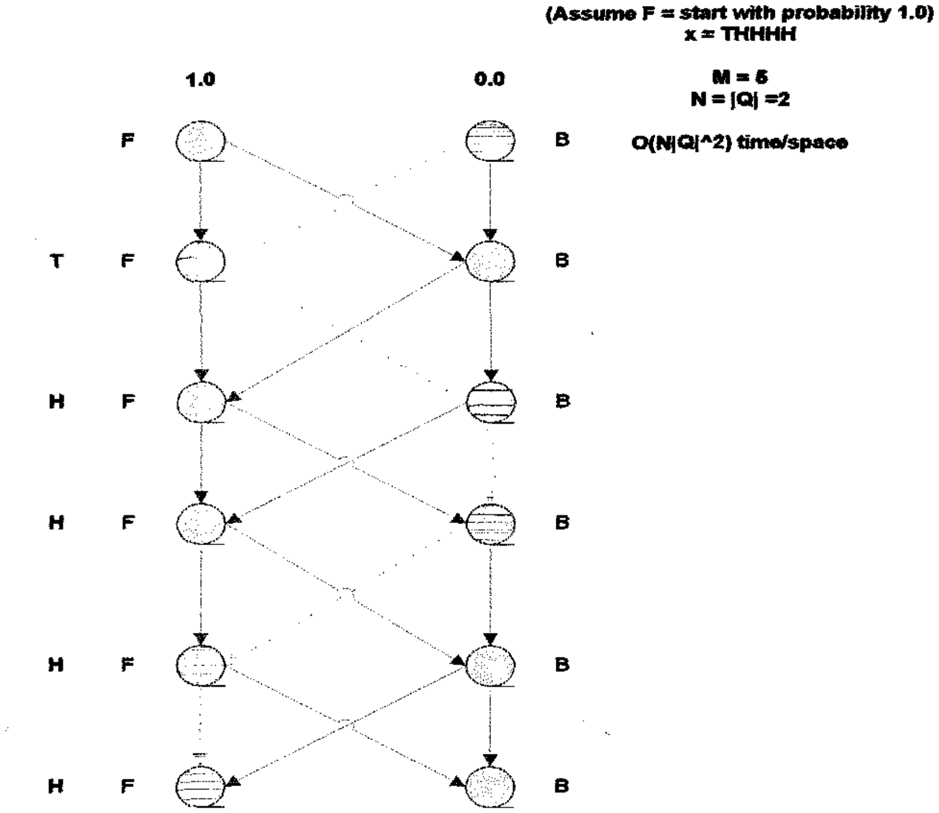


Figure 6: The network for Viterbi algorithms

Where should the Hidden Markov Model (HMM) been implement in gene prediction? It has been implemented in the second step of gene finding which is the "splicing".

2.9 The Measurement of System Accuracy

The system accuracy can be measured using several ways, by classifying all prediction calls on the test sequence as:

- True positive (TP): Gene evaluated as genes
- False positive (FP): Non- genes evaluated as genes
- True negative (TN): Non- genes evaluated as non- genes
- False negative (FN): Genes evaluated as non- genes

Below are formula to find sensitivity of a program and specificity of a program. Sensitivity of a program is the ability of a program to identify as many correct genes as possible. Meanwhile, the specificity of a program is the measurement of the proportion of the correct genes out of the total genes identified.

- Actual Positive (AP) = TP + FN
- Actual Negative (AN) = FP + TN
- Predicted Number of Positive (PP) = TP + FP
- Predicted Number of Negative (PN) = TN + FN
- Sensitivity (SS) = TP / AP
- Specificity (SP) = TP / PP
- Correlative- Coefficient (CC) = $(TP \times TN - FP \times FN)/\sqrt{AN \times PP \times AP \times PN}$.

Before the HMM method been found out, gene prediction system were based on ad hoc features recognition, such as Grail. Using the formula given above, Grail achieves the sensitivity of 0.72 and specificity of 0.84. Meanwhile, HMM achieves the sensitivity of 0.93 and specificity of 0.93. The resource of the result is from Steve Skiena's website, 2002. [8]

cannot do rather than user understand a system on paper which means user can simulate and view the system flow.[6]

The methodology can display the suitable human computer interface for the system based on user perspective. Users can try the system prototype and provide suggestion to improve the interface of the system.

Below is prototyping- based methodology's figure.

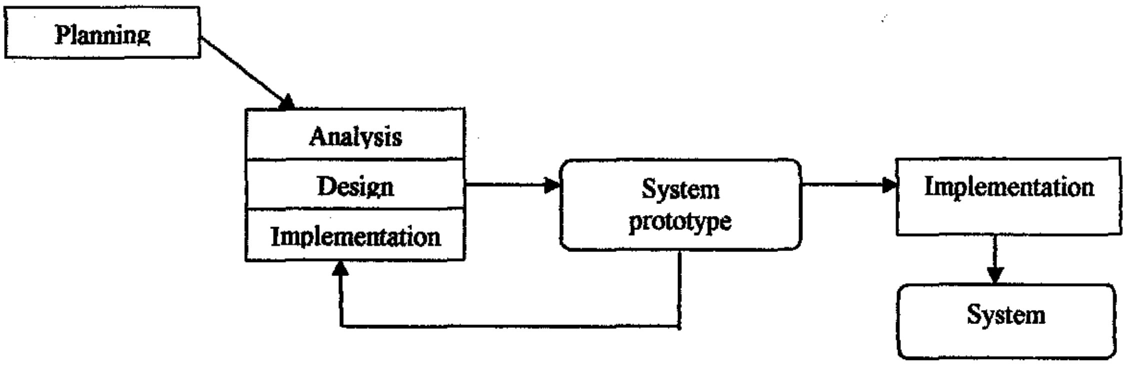


Figure 7: A prototyping- based Methodology

3.2 Tools

All systems in this world have certain tools that are used in order to develop it based on the requirements. Below are the tools that are planned to be used for this project:

3.2.1 HTTP Server

The system has been developed using web base. To turn a computer to be a server, an Apache has been installed to the target computer. Apache is a web server notable for playing a key role in the initial growth of the World Wide Web [13]. Apache support variety of features such as PHP (Hypertext Preprocessor) and Perl which are languages that have been used to develop the system [12}.

3.2.2 Languages

The basic language in developing a system for web base is HTML. HTML, which stands for Hypertext Markup Language is the main markup language for web pages [13]. The system web pages structure basically will be develop using HTML.

PHP, which stand for Hypertext Preprocessor is a reflective programming language originally designed for producing dynamic web pages [15]. PHP is a widely-used general-purpose scripting language that is especially suited for Web development and can be embedded into HTML. PHP generally runs on a web server, taking PHP code as its input and creating Web pages as output [15]. PHP also has many and varied applications, compounded by the availability of many standard and third-party modules [16]. PHP is a powerful language in text processing. It cans manipulate a string that will be used as input in the system within a split second. It also been supported in Apache

3.3 Flowchart

Below is the flowchart for Gene Prediction System. At the main page of the gene prediction system, user will need to enter input sequence by using two methods either entered the input sequence in given or uploaded the sequence file. Refer to figure 7 and figure 8. User is given these choices because sometimes the input sequence that user wants to use just a small size. For example, the size only 1822 base pairs (bp). There are the sizes of input sequence which the length until 191075 bp [17] and it is appropriate if just to upload it.

The output will be represent in graphical and text output. Refer to figure 9.

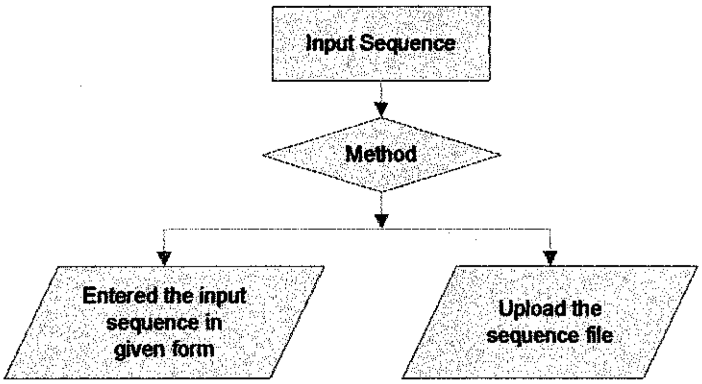


Figure 9: The flowchart for input selection

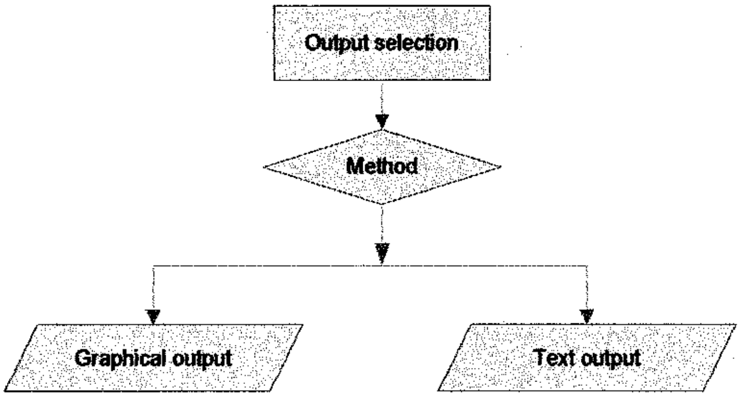


Figure 10: The flowchart for output selection

3.4 Input Format

There are several formats that have been used by bioinfomatics scientists in all over the world to represent either nucleic acid sequence or protein sequence. The example of the format which are RAW format, FASTA format, PIR format, MSF format, CLUSTAL format, TXT format and graphic format.

RAW format which is a sequence format that doesn't contain any header. The numbers and spaces are usually tolerated [1]. Meanwhile, PIR format is much less similar to FASTA format but less common. The MSF format and CLUSTAL format are multiple sequence alignment but the different between them is CLUSTAL format works with T-Coffee meanwhile MSF doesn't work with T-Coffee [1]. TXT format is text format and graphic formats such as GIF, JPEG, PNG and PDF.[1]

In the system, FASTA format will be used as a default format. FASTA format contains a header line, follow by a sequence. The FASTA format is better because it is easy to manipulate and parse sequences using text-processing tools and scripting language like Python, PHP (for website) and Perl. Almost all the gene prediction system recognizes the FASTA format.

Other formats that will be used are the graphical formats. It is important for the user to view some parts of the DNA sequence that has been predicted earlier. Graphical format can represent information of DNA more clearly.

3.5 Milestone

Below is the milestone for the first semester of 2 semester final project

No.	Detail/ Week	1	2	3	4	5	6	7	8	9	10	11		12	13	14
1	Selection of Project Topic	■	■										B			
2	Preliminary Research Work		■	■												
3	Submission of Preliminary Report				■								R			
4	Seminar 1						■									
5	Project Work					■	■	■					E			
6	Submission of Progress Report								■							
7	Seminar 2										■		A			
8	Project Work Continues									■	■	■		■		
9	Submission of Interim Report Final Draft												K		■	
10	Oral Presentation															■

Table 2: Milestone for the first semester of 2 semester final project

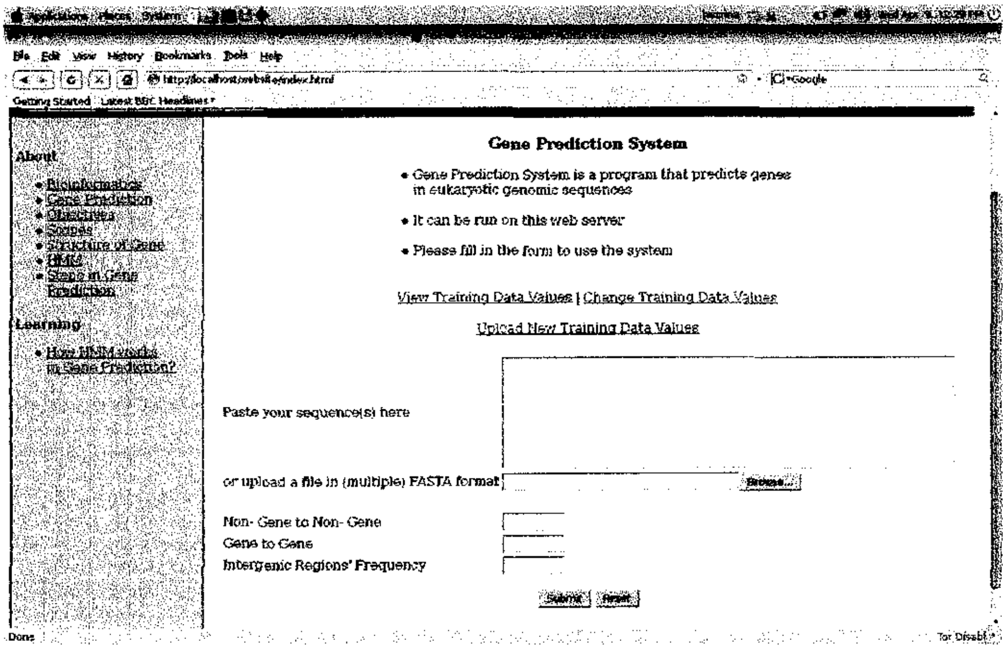
Below is the milestone for the second semester of 2 semester final project

No.	Detail/ Week	1	2	3	4	5	6	7		8	9	10	11	12	13	14
1	Project Work Continue	■	■	■					B							
2	Submission of Progress Report 1				■											
3	Project Work Continue				■	■	■	■	R							
4	Submission of Progress Report 2									■						
5	Seminar								E			■				
6	Project Work Continue									■	■	■				
7	Poster Exhibition								A				■			
8	Submission of Dissertation (soft bound)												■			
9	Oral Presentation								K						■	
10	Submission of Project Dissertation (Hard Bound)															■

Table 3: Milestone second semester of 2 semester final project

4 RESULT AND DISCUSSION

Below are the figures of gene prediction prototype. Figure 11 is the prototype form to be entered by user. User will entered gene sequence which contains A, C, G and T. If user entered the sequence either in capital letter or small letter, the system still will be working. User need to enter probability for non-gene to non-gene, probability for gene to gene and intergenic regions' frequency. These four parameters will be used in calculation of Hidden Markov Model. Another parameter that need in the calculation is the training data. Training data is data of DNA sequence that has been trained a number of times to determine the probability of A-T-C-G position in an organism. Training data for one organism is different compare to the other.


File Edit View History Bookmarks Tools Help
http://localhost/website/index.html
Google
Getting Started Latest BBC Headlines
About
Bioinformatics
Gene Prediction
Objectives
Scopes
Structure of Gene
HMM
Steps in Gene Prediction
Learning
How HMM works in Gene Prediction?
Browse...
Submit
Reset
Done

Gene Prediction System

- Gene Prediction System is a program that predicts genes in eukaryotic genomic sequences
- It can be run on this web server
- Please fill in the form to use the system

View Training Data Values | Change Training Data Values

Upload New Training Data Values

Paste your sequence(s) here

or upload a file in (multiple) FASTA format

Non- Gene to Non- Gene

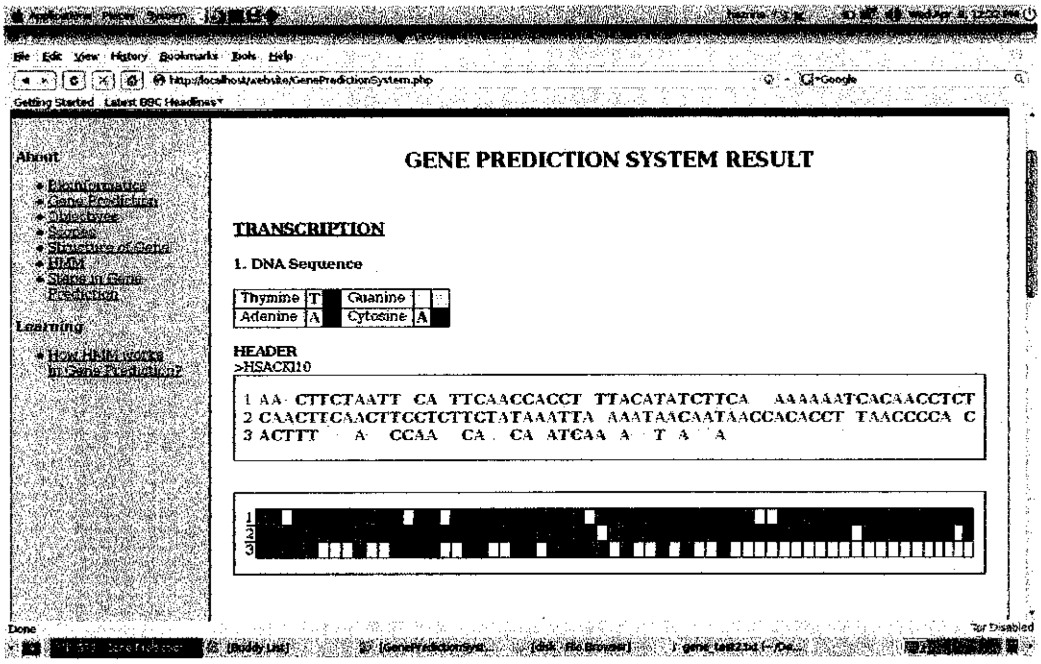
Gene to Gene

Intergenic Regions' Frequency

Figure 11: Gene prediction form

After user entered the sequence and fill up all the forms and clicked the submit button, the gene prediction engine will process the data. There are 3 steps involved in the process which; are transcription, splicing and translation.

In the transcription, the system will capture the DNA sequences and convert the sequences into colored sequences. The colors that have been used is base on the standard color of DNA sequences [31]. For Thymine and Guanine, they have been represented using red and yellow. Meanwhile for Adenine and Cytosine, they have been represented using green and blue. The colors representation is to make the human eyes capture the differentiation between the letters.



GENE PREDICTION SYSTEM RESULT

TRANSCRIPTION

1. DNA Sequence

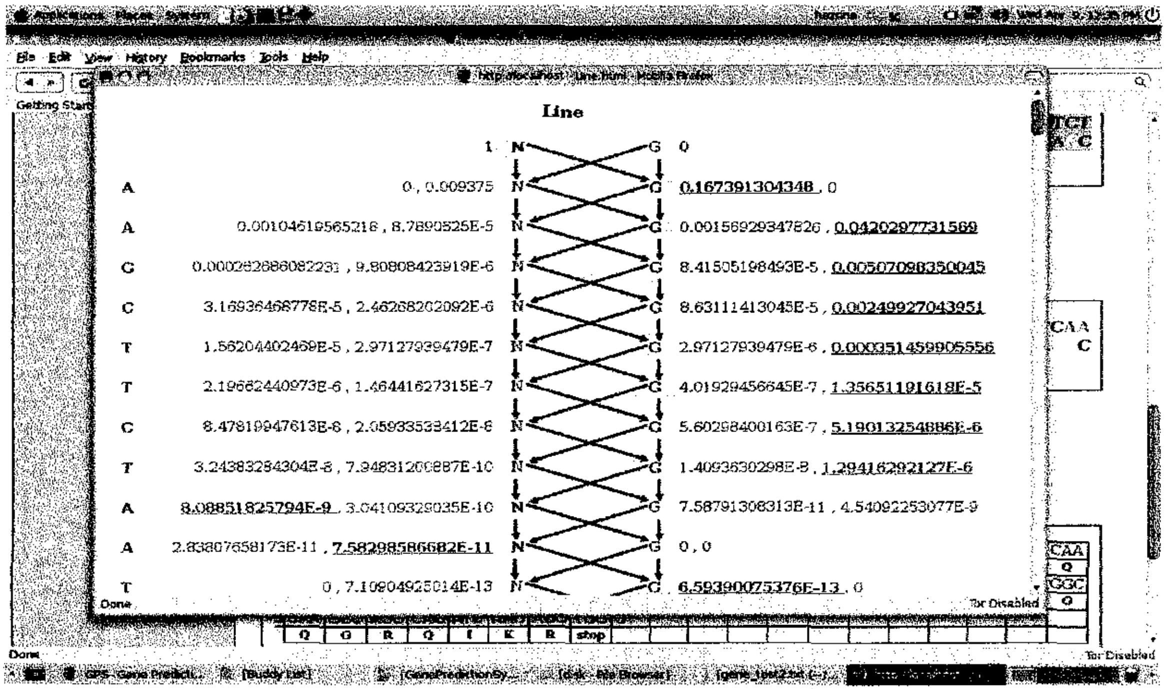
Thymine	T		Guanine		
Adenine	A		Cytosine	A	

HEADER
>HSACKI10

```
1 AA CTTCTAATT CA TTCAACCACCT TTACATATCTTCA  AAAAAATCACAACCTCT
2 CAACTTCAACTTCCTCTTCTATAAATTA AAATAACAATAACCACACCT TAACCCCA C
3 ACTTT   A  CCAA  CA . CA ATCAA A  T A  A
```

Figure 12: The transcription

In the splicing, the system will predict the introns (non-gene/useless information) and exons (gene/useful information) of DNA sequences. How does the system predict the introns and exons? To do the prediction, Hidden Markov Model will be used as the algorithms of the prediction.



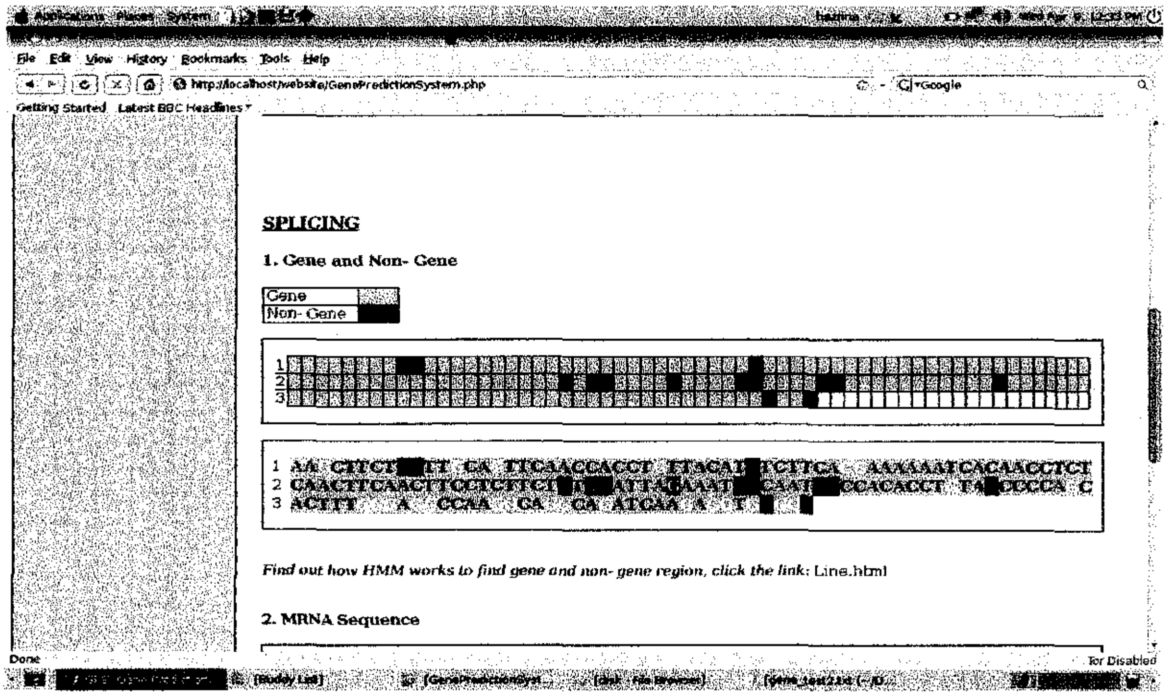
File Edit View History Bookmarks Tools Help
http://localhost/Line.html - Mozilla Firefox
Done
Tor Disabled
Q G R Q I K R stop

Line

```
1                                            N        G  0
A                              0 , 0.009375  N        G  0.167391304348 , 0
A      0.00104619565216 , 8.7890525E-5       N        G  0.00156929347826 , 0.0420297731569
G      0.000262686082231 , 9.80808423919E-6  N        G  8.41505198493E-5 , 0.00507098350045
C      3.16936468778E-5 , 2.46268202092E-6   N        G  8.63111413045E-5 , 0.00249927043951
T      1.56204402469E-5 , 2.97127939479E-7   N        G  2.97127939479E-6 , 0.000351459905556
T      2.19662440973E-6 , 1.46441627315E-7   N        G  4.01929456645E-7 , 1.35651191618E-5
C      8.47819947613E-8 , 2.05933533412E-8   N        G  5.60298400163E-7 , 5.19013254886E-6
T      3.24383284304E-8 , 7.94831200887E-10  N        G  1.40936302983E-8 , 1.29416292127E-6
A      8.08851825794E-9 , 3.04109329035E-10  N        G  7.58791308313E-11 , 4.54092253077E-9
A      2.83807658173E-11 , 7.58298586682E-11 N        G  0 , 0
T                      0 , 7.10904925014E-13 N        G  6.59390075376E-13 , 0
```

Figure 13: Hidden Markov Model

After the DNA sequences have been determined by highlighted either N (non-gene) or G (gene), the sequence will be represented using in the form shown below. (Refer figure 14).



File Edit View History Bookmarks Tools Help
http://localhost/website/GenePredictionSystem.php
Getting Started Latest BBC Headlines
Done
Tor Disabled

SPLICING

1. Gene and Non- Gene

Gene	▨
Non- Gene	■

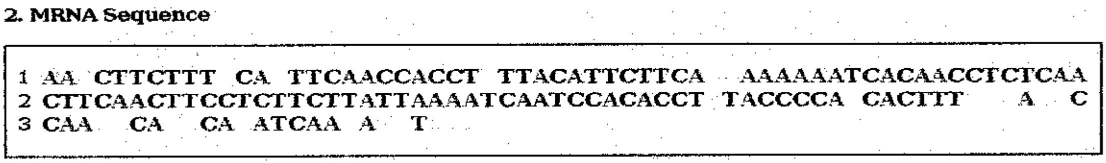
```
1 AA CTTCT  TT  CA TTCAACCACCT TTACAT TCTTCA  AAAAAATCACAACCTCT
2 CAACTTCAACTTCCTCTTCT T  ATTTAGAAAT  CAAT  CCACACCT TA CCCCA C
3 ACTTT   A  CCAA  CA  CA ATCAA A  T
```

Find out how HMM works to find gene and non- gene region, click the link: Line.html

2. MRNA Sequence

Figure 14: Sequence of gene and non gene in DNA sequence

After the introns(non gene) been removed, the exons (gene) will be combined and become the MRNA sequences.

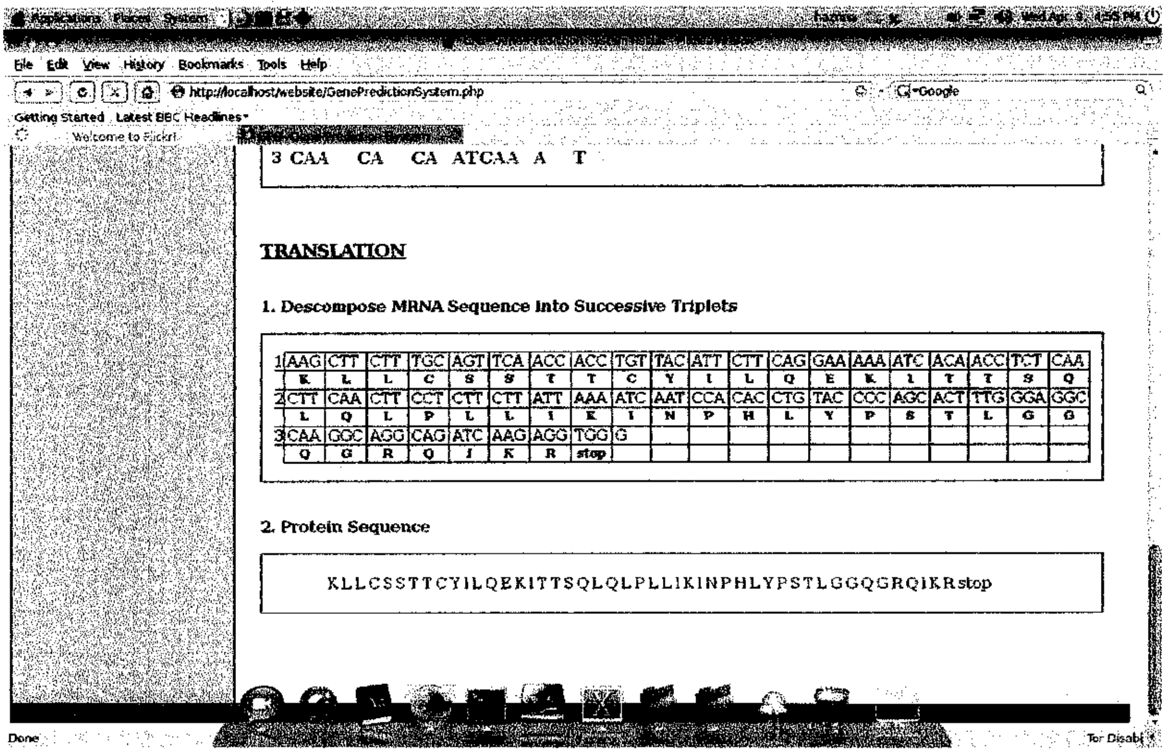


2. MRNA Sequence

```
1 AA CTTCTTT CA TTCAACCACCT TTACATTCTTCA AAAAAATCACAACCTCTCAA
2 CTTCAACTTCCTCTTCTTATTAAAATCAATCCACACCT TACCCCA CACTTT A C
3 CAA CA CA ATCAA A T
```

Figure 15: MRNA sequence

The final step of gene prediction is by translating the MRNA sequence to protein sequence. Tri-nucleotides will represent one element of protein. The final output will be protein sequence.



File Edit View History Bookmarks Tools Help
http://localhost/website/GenePredictionSystem.php
Getting Started Latest BBC Headlines
Welcome to Flickr!
Done

```
3 CAA CA CA ATCAA A T
```

TRANSLATION

1. Descompose MRNA Sequence Into Successive Triplets

1	AAG	CTT	CTT	TGC	AGT	TCA	ACC	ACC	TGT	TAC	ATT	CTT	CAG	GAA	AAA	ATC	ACA	ACC	TCT	CAA
	K	L	L	C	S	S	T	T	C	Y	I	L	Q	E	K	I	T	T	S	Q
2	CTT	CAA	CTT	CCT	CTT	CTT	ATT	AAA	ATC	AAT	CCA	CAC	CTG	TAC	CCC	AGC	ACT	TTG	GGA	GGC
	L	Q	L	P	L	L	I	K	I	N	P	H	L	Y	P	S	T	L	G	G
3	CAA	GGC	AGG	CAG	ATC	AAG	AGG	TGG	G											
	Q	G	R	Q	I	K	R	stop												

2. Protein Sequence

KLLCSSTTCYILQEKITTSQLQLPLLIKINPHLYPSTLGGQGRQIKRstop

Figure 16: Protein sequence

Even though there are similar systems that have been developed in whole around the world, Gene Prediction System is different from the other systems in term of the representation of output and the input that need to be used for calculation.

In term of output representation, Gene Prediction System is more presentable because compare to other program the output are been represent using colors. The colors usage in this system can helps scientist to find out which nucleotide is more, compare to other nucleotides. Using colors, users either scientist or non-scientist can see the nucleotides reperesentation clearly. They will attract to use the system. Other program, the output representation just in a plain text. No color been used.

In term of input that need to be used for calculation, other program has set the value for each organism but in Gene Prediction System the value need to be entered by scientist. It will make the system more generic.

4 CONCLUSION AND RECOMMENDATION

For future work, the system output should be displayed in 3-D to make it looks more interesting for newbie in bioinformatics. This can also attract the researcher in using the system.

My study indicates that the prediction of gene can be made by several methods such as HMM model, Homology, Ab initio and many more. By developing the gene prediction engine, it helps the scientist in microbiological fields to predict the gene with computers help. They can identify stretches of sequence for genomic DNA that is biologically functional including protein coding regions. Using the information gathered, people will get the benefits after the analysis if the gene helps to reduce or prevent diseases.

The research area which is bioinformatics is a new field in Malaysia. By doing some research and exploration in this field, it will give us some sense awareness that there are another research area that can be explored further by scientists in Universiti Teknologi PETRONAS and also in Malaysia.

5 REFERENCES

[1] Jean-Michel Claverie and Cedric Notredame, 2003, Bioinformatics for dummies, Wiley Publishing, Inc.

[2] http://en.wikipedia.org/wiki/genetic_code

[3] http://en.wikipedia.org/wiki/RNA

[4] http://www.nslij-genetics.org/gene/

[5] D.R Westhead, J.H Parish and R.M Twyman, 2002, Bioinformatics, BIOS Scientific Publishers Limited.

[6] Selecting Development Approach, February 2005, Centers for Medicare and Medicaid Services (CMS).

[7] Alan Dennis, Barbara Haley Wixom, David Tegarden, 2005, System Analysis and Design with UML Version 2.0: An Object Oriented Approach, John Wiley & Sons, Inc.

[8] http://www.cs.sunysb.edu/~skiena/549/lectures/geneprediction/

[9] http://www.scfbio-iitd.res.in/research/genomeanalysis.htm

[10] Jong-won Chang, Chungoo Park, Dang Soo Jung, Mi-hwa Kim, Jae-woo Kim, Seung-sik Yoo, Hong Gil Nam, Space-Gene,

[11] David W.Mount, 2001, Bioinformatics Sequences and Genome Analysis, Cold Spring Harbor Laboratory Press.

[12] http://httpd.apache.org/

[13] http://en.wikipedia.org/wiki/Apache_HTTP_Server

[14] http://en.wikipedia.org/wiki/HTML

[15] http://en.wikipedia.org/wiki/PHP

[16] http://en.wikipedia.org/wiki/Perl

[17] Youlian Pan, Christoph W. Sensen, Modular Neural Networks and Their Application in Exon Prediction.

[18] http://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology

[19] P.C Winter, G.I Hickey, H.L Fletcher, 2003 , Instant Notes Genetics, Viva Books Private Limited.

[20] http://en.wikipedia.org/wiki/RNA_polymerase

[21] http://en.wikipedia.org/wiki/Sigma_factor

[22] http://en.wikipedia.org/wiki/Promoter

[23] http://en.wikipedia.org/wiki/MRNA

[24] http://en.wikipedia.org/wiki/5%27_cap

[25] http://en.wikipedia.org/wiki/5%27_UTR

[26] http://en.wikipedia.org/wiki/3%27_UTR

[27] http://www.sumanasinc.com/webcontent/animations/content/lifecyclemrna.html

[28] http:// en.wikipedia.org/wiki/Open_reading_frame

[29] http://en.wikipedia.org/wiki/Hidden_Markov_model

[30] What is a hidden Markov Model? Sean R Eddy 2004 Nature Publishing Group

[31] http://seqcore.brcf.med.umich.edu/doc/educ/dnapr/sequencing.html

[32] Liliana Florea, 2006, Gene Finding, the George Washington University

[33] Alexander Isaev, Introduction to Mathematical Methods in Bioinformatics, Springer.

[34] Mario Stanke and Stephan Waack, Gene Prediction with a hidden Markov model and a new intron submodel, 2003