Indian Institute of Information Technology

Allahabad

Project Report

Recognition of Exons and Introns using the principle borrowed from DNA Walk and Queuing Simulation

Under the guidance of Dr. T. Lahiri

Submitted By:

Sourabh Gupta (IIT2009066)

Deepak Bhoria (IIT2009112)

Candidate's Declaration

We hereby declare that the work presented in this project report entitled "Recognition of Exons and Introns using the principle borrowed from DNA Walk and Queuing Simulation", submitted as a 7th semester B.Tech (IT) Mini project is an authenticated record of our original work carried out from July 2012 to September 2012 under the guidance of Dr. T. Lahiri. Due acknowledgements have been made in the text to all other material used. This project work was done in full compliance with the requirements and constraints of the prescribed curriculum.

Place: IIIT Allahabad	Sourabh Gupta (IIT2009066)
Date: 25rd September, 2012	Deepak Bhoria (IIT2009112)

Certificate

(Deemed University)

(A center of excellence in IT, established by Ministry of HRD, Govt. of India)

Date: _______________

CERTIFICATE

This is to certify that this project work entitled "**Recognition of Exons and Introns using the principle borrowed from DNA Walk and Queuing Simulation** " is submitted by **IIT2009066** and **IIT2009112** as mid semester report.

COUNTERSIGNED

____

Dr. T. Lahiri

(Project Supervisor)

Acknowledgement

As understanding of the study like this is never the outcome of the efforts of a single person, rather it bears the imprint of a number of persons who directly or indirectly helped us in completing the present study. We would be failing in our duty if we don't say a word of thanks to all those whose sincere advise make our this documentation of topic a real educative, effective and pleasurable one.

It is our privilege to study at Indian Institute of Information Technology, Allahabad where students and professors are always eager to learn new things and to make continuous improvements by providing innovative solutions. We are highly grateful to the honorable **Dr. M. D. Tiwari, Director IIIT-Allahabad,** for his ever helping attitude and encouraging us to excel in studies.

Regarding this thesis work, first and foremost, we would like to heartily thank our supervisor **Dr. T. Lahiri** for his able guidance. His fruitful suggestions, valuable comments and support were an immense help for me. In spite of his hectic schedule he took pains, with smile, in various discussions which enriched us with new enthusiasm and vigor.

Table of Contents

Goal

The main aim of our project is to use visual and computational methods to analyze pattern of bases present in Introns and Exons using the principle borrowed from DNA walk and queue simulation.

Pre mature messenger-RNA contains exons and introns. Introns are removed and exons get combined through RNA Splicing to form proteins using Translation process [1]. To understand or simulate transcription and translation process of central dogma of molecular biology [2], we need to analyse the nucleotide sequence in introns and exons. We will be using stochastic methods for their analysis. By making the histogram according to frequency of arrival of individual nucleotides we can find the probability of occurrences of nucleotides in given pattern of exon and introns.

So we need a method to graphically represent the genome sequences in a way suitable for human perception to take advantage of the unique capability of the human visual pattern recognition system as well as a computational method for finding unknown patterns that occur imperfectly in a set of many sequences.

Motivation

The motivation behind our project is to use the computing power in analytical study of a genome represents how the frequency of each nucleotide of a pairing nucleotide couple changes locally.

DNA Walk [3], [4] is a simple algorithm used to draw a DNA sequence by simply assigning a direction to each nucleotide which provides a visual image how nucleotides are arranged. We can assign T, C, A, and G correspond the East, South, West, and North directions, respectively. Reading the nucleotide sequence nucleotide by nucleotide, we can create a graph.

But it provides only the 2Dimensional Visual image about how direction is changed according to current nucleotide. If we modify the DNA walk; we can draw histogram according to arriving individual nucleotides in introns and exons and hence find what probability distribution it follows.

Scope

- Making Sense of DNA and Protein Sequences.
- Unmasking Genes in the Human Genome.

Literature survey

DNA (Deoxyribonucleic acid**)** is often called the Blueprint of Life. DNA is a nucleic acid containing the genetic instructions used in the development and functioning of all known living organisms (with the exception of RNA viruses). The DNA segments carrying this genetic information are called genes. Likewise, other DNA sequences have structural purposes, or are involved in regulating the use of this genetic information. Along with RNA and proteins, DNA is one of the three major macromolecules that are essential for all known forms of life.

DNA consists of two long polymers of simple units called nucleotides, with backbones made of sugars and phosphate groups joined by ester bonds. These two strands run in opposite directions to each other and are therefore anti-parallel, one backbone being 3' (three prime) and the other 5' (five prime). This refers to the direction the 3rd and 5th carbon on the sugar molecule is facing. Attached to each sugar is one of four types of molecules called nucleobases (informally, *bases*).

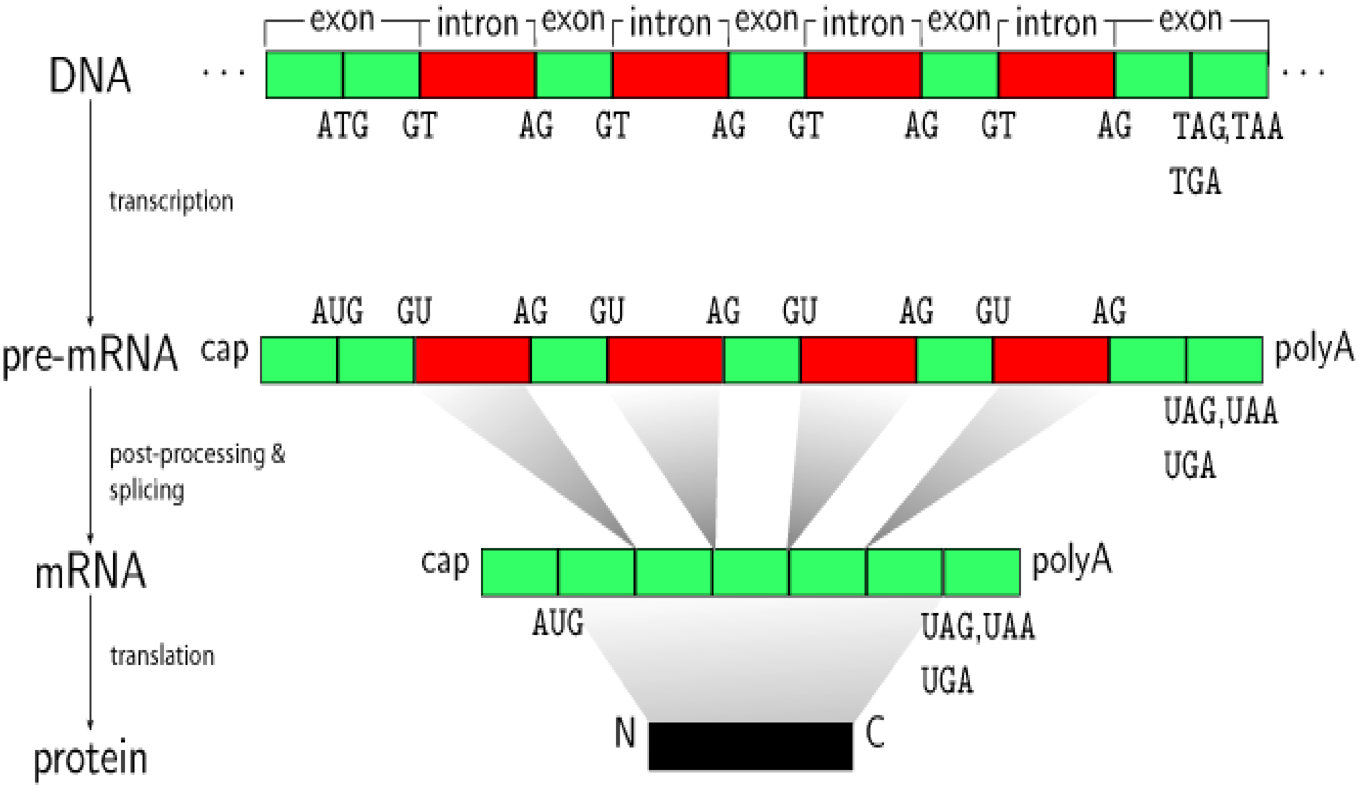
In simple terms DNA contains instruction for making proteins in a cell. It happens in two steps:

- **Transcription:**
- **Translation:**

Transcription: The synthesis of new proteins begins with the transcription. DNA is transcribed to RNA to produce mRNA (messenger Ribo Nucleic Acid), rRNA (ribosomal RNA) or tRNA (transport RNA).

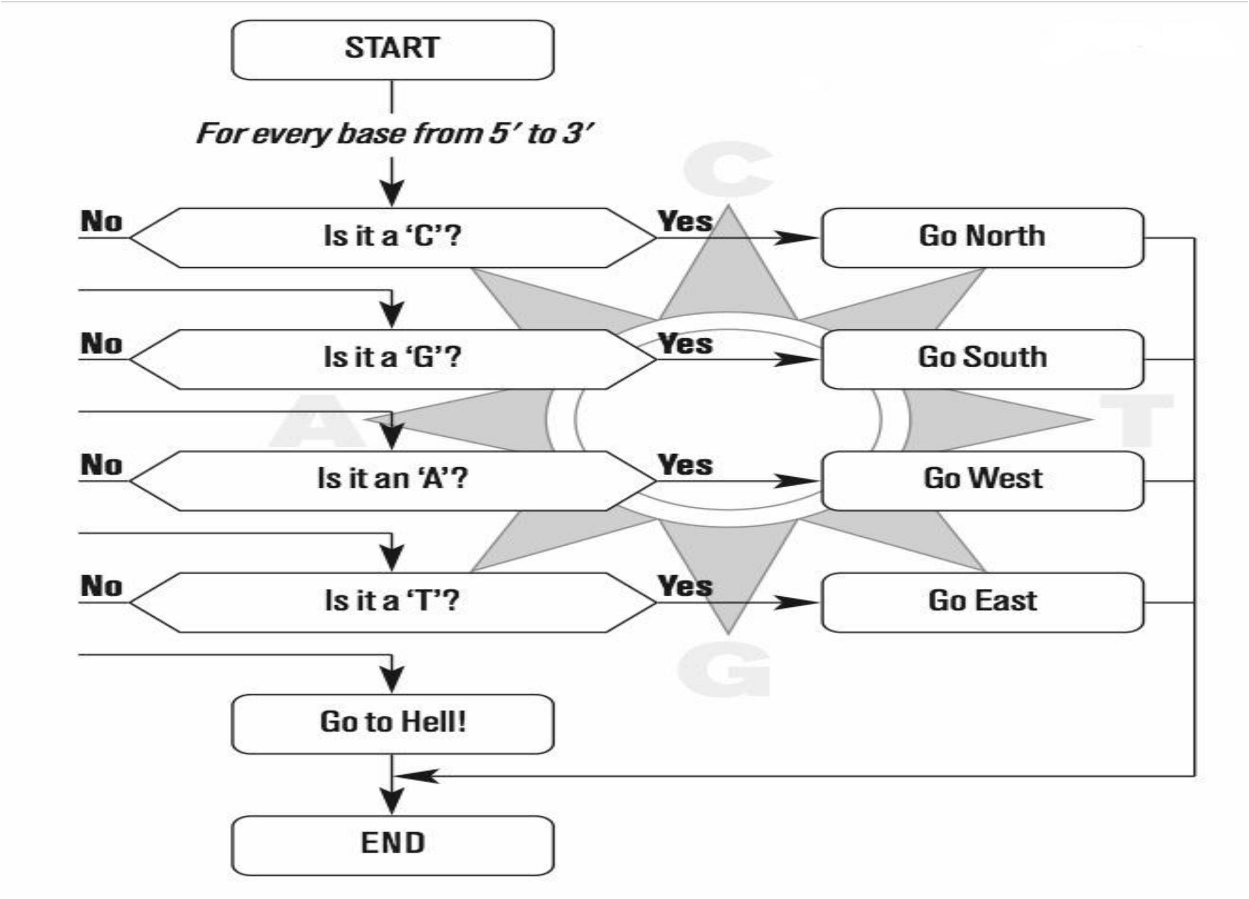
Translation: In translation, messenger RNA (mRNA) produced by transcription is decoded by the ribosome to produce a specific amino acid chain, or polypeptide that will later fold into an active protein.

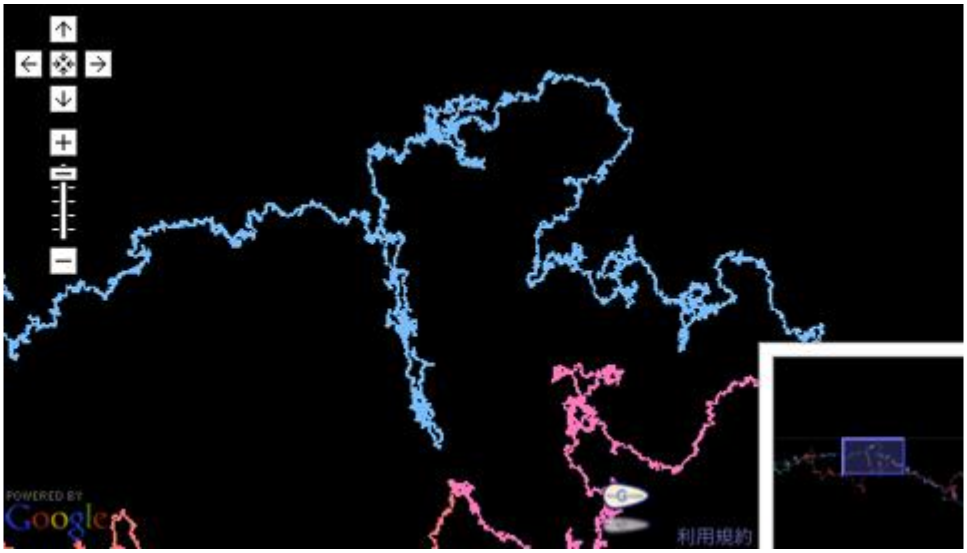
This process is the part of Central dogma of molecular biology.



An **intron** is any nucleotide sequence within a gene that is removed by RNA splicing while the final mature RNA product of a gene is being generated while an **exon** is a sequence of DNA that is translated into RNA and then protein. Introns need to be sliced out and exons are combined to form proteins. If we want to study the formation of proteins from DNA these introns and exons need to distinguished.

DNA Walk is a simple algorithm [3],[4] used to draw a DNA sequence by simply assigning a direction to each nucleotide which provides a visual image how nucleotides are arranged. We can assign T, C, A, and G correspond the East, South, West, and North directions, respectively. Reading the nucleotide sequence nucleotide by nucleotide, we can create a graph.


START
For every base from 5′ to 3′
No
Is it a ‘C’?
Yes
Go North
No
Is it a ‘G’?
Yes
Go South
No
Is it an ‘A’?
Yes
Go West
No
Is it a ‘T’?
Yes
Go East
Go to Hell!
END
C
A
T
G


POWERED BY
Google
利用規約

DNA Walk

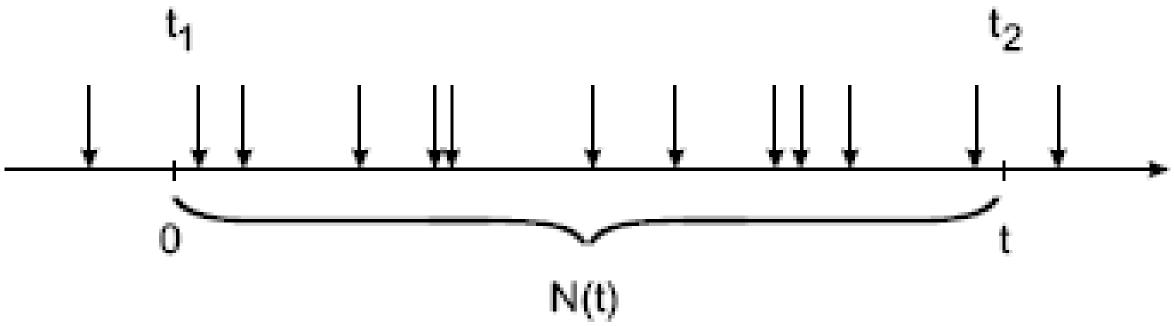
Queue simulation

The term queuing system[5] is used to indicate a collection of one or more waiting lines along with a server or collection of servers that provide service to these waiting lines. Customers requiring service are generated over time by an input source. The required service is then performed for the customers by the service mechanism, after which the customer leaves the queuing system.

BASIC COMPONENTS OF A QUEUING SYSTEM

- Input Process
- Service mechanism
- Queue discipline
- Output of the queue

Poisson process is one of the most important models used in queuing theory. Often the arrival process of customers can be described by a Poisson process.



The counter tells the number of arrivals that have occurred in the interval (0, t) or, more generally, in the interval (t1, t2).

The interarrival times are independent and obey the Exp(λ) distribution

$$\mathrm{P}\{N(t)=n\}=\frac{(\lambda t)^n}{n!}e^{-\lambda t}$$

Our Approach

If we apply DNA walk on these sequences of introns and exons we get a visual picture of how nucleotides arrive in the sequence but it doesn't tell us about probability of occurrence of nucleotide.

We can modify DNA Walk by constructing four arrays Garr, Carr, Aarr, Tarr. When nucleotide Guanine is encountered it puts its index in Garr, similarly for cytosine, thymine and adenine.

Data for intron and exons can be found from NCBI [6]. The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

Now we can plot graphs for these nucleotides. Although their sequence look random but they may follow some probability distribution. We will draw a histogram for occurrence of each nucleotide. This graphical representation shows a visual impression of the distribution of data. It is an estimate of the probability distribution of a continuous variable. A histogram consists of tabular frequencies, shown as adjacent rectangles, erected over discrete intervals (bins), with an area equal to the frequency of the observations in the interval. The height of a rectangle is also equal to the frequency density of the interval, i.e., the frequency divided by the width of the interval. The total area of the histogram is equal to the number of data. A histogram may also be normalized displaying relative frequencies. It then shows the proportion of cases that fall into each of several categories, with the total area equaling 1. The categories are usually specified as consecutive, non-overlapping intervals of a variable. The categories (intervals) must be adjacent, and often are chosen to be of the same size.

It can then be used to plot density of data, and often for density estimation: estimating the probability density function of the underlying variable. The total area of a histogram used for probability density is always normalized to 1. If the length of the intervals on the *x*-axis are all 1, then a histogram is identical to a relative frequency plot.

Clustering of repeats, palindromes, horizontally transferred genes, telomeres, and GC skew can be easily spotted using this visualization approach. Following is the DNA Walk of Escherichia coli, which is highly skewed in GC vector, and leading/lagging strands can be quickly identified from this diagram.

Software Used

We will be using Matlab for our project as it is a high-level language and interactive environment for numerical computation, visualization, and programming. The language, tools, and built-in math functions enable you to explore multiple approaches.

REFERENCES

[1] John W. Kimball (2012): Gene Translation, Kimball's Biology Pages

[2] Crick, F.H.C. (1958): On Protein Synthesis. Symp. Soc. Exp. Biol. XII, 139-163.

[3] Peng Zhou and Zhicai Shang (2009): 2D molecular graphics: a flattened world of chemistry and biology. *Journal of* Briefings in Bioinformatics Advance Acces .6-8.

[4] Zahhad , Ahmed, Elrahman (2012): Genomic Analysis and Classification of Exon and Intron Sequences Using DNA Numerical Mapping Techniques Published in MECS DOI: 10.5815/ijitcs.2012.08.03 29-30

[5] Azmat Nafees (2007): Queuing Theory and its Application Chapter 3 Methodogy 5 -12

Remarks by Board members