COMPARISON OF DNA SEQUENCES BASE COMPOSITION IN CODING AND NON-CODING SEQUENCES

A

Dissertation Report

Submitted in Partial Fulfilment of the Requirements

For the Award of Degree of

Masters of Science

In

Biotechnology

SUBMITTED BY

PRIYANKA DUBEY

REGISTRATION NO. 301701022

UNDER THE GUIDANCE OF

DR. VIKAS HANDA

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

DEPARTMENT OF BIOTECHNOLOGY

TIET, PATIALA

AUGUST, 2019

CANDIDATE DECLARATION

This is to certify that dissertation entitled "Comparison of DNA sequence base composition in in coding and non-coding sequences" submitted by "Priyanka Dubey(301701022)" in partial fulfilment of the requirement of award of Masters of sciences in Biotechnology is a record of original dissertation work done by me, under the guidance and supervision of Dr. Vikas Handa (Assistant professor, Department of Biotechnology TIET, Patiala) and it has not formed the basis for the award of any Degree or Diploma or other similar title to any candidate of any University.

Priyanka Dubey

M.S.C in Biotechnology

Roll NO 301701022

Date: 10 August 2019

Place: Patiala

CERTIFICATE

This is to certify that dissertation entitled "Comparison of DNA sequence base composition in in coding and non-coding sequences" submitted by "Priyanka Dubey (301701022)" in partial fulfilment of the requirement of award of Masters of sciences in Biotechnology is an authentic work carried out by her under my supervision and guidance.

To the best of my knowledge the matter embodied in this dissertation has not been submitted to award of any degree or certificate in any other university/ institute.

12.08.19

Dr.Vikas Handa

Assistant Professor

Department of Biotechnology

TIET, Patiala

Dr. Moushumi Ghosh

Head of Department

Department of Biotechnology

TIET, Patiala

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

A	Adenine
C	Cytosine
C V	Coefficient of variation
DNA	Deoxyribonucleic acid
G	Guanine
GC%	Guanine Adenosine percentage
NCBI	National Centre for Biotechnology Information
ORF	Open reading frame
RNA	Ribonucleic acid
R.W	Random walk
T	Thymine
U	Uracil

LIST OF FIGURES

ABSTRACT

DNA stores genetic information within the variety of sequence of 4 bases A, T, G and C. Just in case of RNA viruses, it's hold on within the variety of RNA base sequence comprising of G, A, U and C. The organization of genomes of prokaryotes and eukaryotes are simple and complex respectively. GC content typically calculated as a percentage value and sometimes known as G+C quantitative relation or rate-ratio. GC-content proportion is calculated as Count (G + C)/Count (A + T + G + C) * 100 percent. Here we calculated the GC% for both coding and non-coding sequences of prokaryotes and exons and introns of eukaryotes and observed that in both the cases the GC% of coding region and exons regions are always higher than non-coding or introns region. We analysed the coefficient of variation and displacement in 2D-random walk of DNA sequence for the data and analysed that in prokaryotes the coding is higher while in the case on eukaryotes introns have higher coefficient of variation and 2D-random walk displacement so no inference can be drawn because the coefficient of variation and random walk seems to be getting affected by the length of the sequence.

Keywords: Coding sequence, GC %, non-coding sequence, exon and intron.

CHAPTER - #1

INTRODUCTION

INTRODUCTION

DNA stores genetic information in the form of sequence of four bases A, T, G and C. In case of RNA viruses, it is stored in the form of RNA base sequence comprising of G, A, U and C. Erwin Chargaff found that in DNA the G is always equal to C and the A is always equal to T which he called the Chargaff rule (Chargaff *et al.*, 1950, 1952). His second rule was called Parity rule 2 (PR2) in which he stated that in each strand, the number of as is roughly equal to the number of T ratio and Gs are roughly equal to Cs. Further concept of sequence complementarity in the double helical DNA was discovered by Watson & Crick (Pauling, L., and Corey, R. B., *et al.*,1953). Subsequently genetic code, functional aspects of various elements and motifs were deciphered. The base sequence function deciphering revealed several interesting molecular phenomena associated with storage and expression of genetic information. The base composition and other DNA sequence attributes is relevant till the date and may be used as a characteristic of genomes or DNA sequence function.

Genome is the complete set of genes or genetic material present in a cell or an organism. The variation in the living world is largely owing to the organisms' genetic information which differs due to variation in genome size and its base sequence. The genome size (or "C-value") of an organism is defined as the total amount of DNA contained within a single (i.e., haploid) set of its chromosomes. Genome sizes of bacteriophages and viruses range from a few thousand bases to several hundred kilobases. Bacterial genomes range from 0.5 Mb to 10 Mb. Primitive multicellular organisms such as nematodes have genome size about four times larger. On the higher side the largest genome is found in an Amoeba dubia, a one-cell organism, with 670,000 Mb, 200 fold larger than the human genome and 20,000 fold larger than the one found in yeast and the other one is Paris japonica which has 149 million base pairs**.**

One of the most important component of a genome is genes which are a DNA sequence that encode for proteins and some functional RNAs. In cellular organism's genomes, thousands to tens of thousands of genes are present. The coding sequence of a gene determines the amino acid sequence of proteins encoded by it.

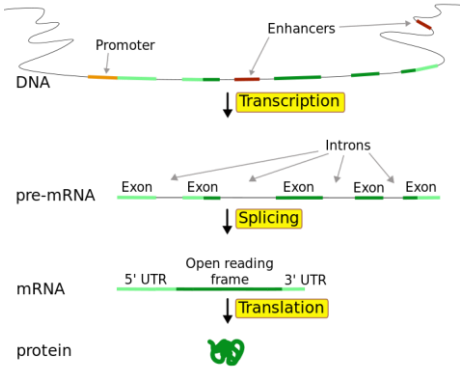


Figure 1: Translation

(https://pubs.niaaa.nih.gov/)

The coding sequence rarely found interrupted by the presence of non-coding sequences in prokaryotes while in eukaryotic genes, the coding sequences are often found interrupted by non-coding sequences, known as introns. The proportion of interrupted genes, average number of introns and the average size of introns increase as we move up the evolutionary complexity of organisms. The interrupted coding sequences of genes are known as exons.

Living organisms of varying evolutionary complexity not only differ in genome size and presence of different genes but also in genome complexity. One of the preliminary experiments that revealed genome complexity was based on reassociation kinetics of genomic DNA. When double stranded DNA in solution is heated it denatures releasing the complementary strand. If the solution is cooled quickly the DNA remains in a single stranded state. However, if the solution is cooled slowly reassociation will occur. The reassociation of a pair of complementary sequence results from their collision and the rate depends on their concentration. As the two strands are involved the process follows second order kinetics. The size of the DNA fragment affects the rate of reassociation and is conveniently controlled if the DNA is divided into small fragments. The rate of reassociation is dependent on genome size. This proportionality is only true in the absence of repeated sequence. When the studies were done on the reassociation of calf thymus DNA, the results indicated that the concentration of DNA sequence that reassociate rapidly is 100000 times the concentration of those sequence that reassociate slowly. So, the highly repetitive DNA renatures fast and while the moderately repetitive DNA renatures at an intermediate rate relative to the slow component, which consists of non-repetitive and renatures at a very slow rate.

In prokaryotes almost entire genome comprises of non-repetitive or unique sequences. Most of the coding sequences are found in non-repetitive DNA. However, in the case of higher eukaryotes large proportions of repetitive DNA are found. The repetitive DNA sequences can be classified into two forms. Tandem repeat sequences where multiple copies of repeat unit are present contiguously and they belong to the rapidly reassociating component of genome while moderately fast component consists of dispersed repeat sequences which have fewer repeat units scattered in the genome. Mini-satellite and micro-satellite DNA are the examples of tandem repeats while transposons are exemplifying dispersed repeats. Most of the genes are found to be present in non-repetitive DNA. Bacterial genomes entirely consist of genes whereas in higher eukaryotes genes can be small islands in the non-coding DNA.

In eukaryotes, DNA is complexed with histones and other chromatin proteins to form compact structures known as chromosomes which are found in nucleus. A similar type of nucleoid structure is found in prokaryotes. Nucleoids lack histones and are not stored in an organelle. Eukaryotic chromosomes when stained using Giemsa stain develop alternate dark and light bands) in which the G bands are positive staining regions and R bands are negative staining regions. In G banding the chromosome are subjected to controlled digestion with trypsin before Giemsa staining, which reveals alternatively positive G bands and negative R bands staining regions.

The AT rich DNA is mostly associated with condensed chromatin whereas GC sequence is located in the dispersed chromatin (Saccone *et al*. 2002; Bernardi *et al*.2015).

GC content is the percentage of nitrogenous bases on DNA or RNA molecules that are either guanine or cytosine. The genome of various organism differs in their overall GC content. GC content varies widely among different species, for example GC content of *Plasmodium falciparum* the GC% is 27.11% on the other hand in case of *Thermus aquaticus* the GC% is 67.1%. Thus GC content of a species is used as one of the genomic characteristics. GC content is a primary factor shaping amino acid compositions. GC content shapes amino acid composition to trade off the cost of amino acids with bases, which could be caused by the energy efficiency (Bolívar and Guéguen *et al*. 2019). Amino acid composition reflects the usage of twenty standard amino acids in proteins. Understanding the changes of amino acid composition among homologous proteins is key to the investigation of protein functioning, as the proteins can acquire new functions through amino acid substitutions (Misawa *et al*. 2008). Understanding how proteins evolve is important, and the order of amino acids being recruited into the genetic codons was found to be an important factor shaping the amino acid composition of proteins. The variety of available genomic sequences also allows us to compare the compositional distribution patterns of different genomes ranging from unicellular to multicellular eukaryotes. It has been reported that a high GC content of a genome increases the propensity of few signature amino acids like G, A, R, and P which are known to promote protein intrinsic disorders (Misawa *et al*. 2008).

The GC content not only varies among genomes of different species but also within a genome. An *isochore* is a large region of DNA with a high degree uniformity in guanine (G) and cytosine(C) GC and CG (collectively GC content) During the evolution of the gene rich and moderately gene rich isochores underwent a major GC increases. GC rich regions differs significantly from that of GC poor regions. For example, the average GC content in human genomes ranges from 35% to 60% across 100-Kb fragments, with a mean of 46.1%. The development of the isochore model greatly increased the appreciation of the complexity and compositional variability of eukaryotic genomes. (Bernardi et al. 1985; Filipski 1987; Sueoka 1988; Wolfe *et al*. 1989; Eyre-Walker 1992). The GC rich isochores are found to be more thermostable, bendable, ability to B-Z transition and curvature of the DNA helix. GC rich isochores are known to be gene rich and show a high level of transcription. Large scale DNA deformation is ubiquitous in transcriptional regulation in prokaryotes and eukaryotic alike. GC content is found to be variable with different organisms, the process of which is envisaged to be contributed to by variation in selection, mutational bias, and biased recombination-associated DNA repair (Saccone *et al*. 2002; Bernardi 2015).

There are certain GC rich regions found in higher eukaryotes that contain high content of CpG dinucleotide, known as CpG islands. The CpG dinucleotides are mutagenic and as a result found underrepresented in the entire genome of higher eukaryotes except in CpG islands. The

CpG islands are often present in 5' regions of mammalian genes and most of the CpG islands remain unmethylated (Takai and Jones *et.al* 2001).

The AT rich DNA is mostly associated with condensed chromatin whereas GC sequence is located in the dispersed chromatin (Saccone *et al*. 2002; Bernardi *et al*. 2015). The GC rich sequence is more liable to mutation. DNA sequence can contain information and some of the information is merely understood. The change of base composition even in synonymous sites affects mutation probability of nonsynonymous sites and thus of encoded proteins. There is a unique type of housekeeping genes, which are especially unsafe when prone to mutation. Natural selection which usually removes deleterious mutations, in the case of these genes only increases the hazard because it can descend to sub organismal (cellular) level.

The evolutionary pattern of base composition and its potential cause have not been well understood. Genome sequencing make it possible to examine the overall genome organization and phylogenetic inference from an evolutionary perspective, an efficient and effective DNA repair system is essential to the maintenance of genome integrity; mutations in DNA repair genes can lead to hyper mutated genomes, severe diseases, and cancers. Some nucleotide changes must inevitably escape this surveillance system to provide genetic variability to fuel the evolutionary process. The evolutionary rates of genes positively correlate with GC contents with P-value significantly lower than 0.05 for 94% homologous genes.

The isochore model of the mammalian genome describes patterns in base composition. Although, on an average the mammalian genome is 40%GC the distribution is far from uniform. The existence of these compositionally homogeneous stretches (meaning equal regions). The isochore classes differ in more than base composition: they are not uniform in their representation in the genome; they are not random in their distribution within or among chromosomes; and they are strikingly different in their gene densities (Saccone *et al*. 2002; Bernardi *et al*.2015).

Rates of recombination can vary among genomic regions in eukaryotic and this is believed to have major effects on their genome organization in terms of base composition, DNA repeats density, intron size, evolutionary rates and gene order. Rates of crossing over have been shown to correlate not only with the GC content of synonymous sites, where weak natural selection is expected to act on codon-usage bias, but also with the GC content of noncoding sites. This is unlikely to be because GC bases are recombinogenic, as the correlation is far stronger with silent DNA than with total DNA. This unexpected correlation may reflect the action of weak selection on noncoding GC, which would be less effective in regions of reduced recombination.

In silico analysis of DNA sequences is an important area of computational biology in the post-genomic era. Over the past two decades, computational approaches for *ab initio* prediction of gene structure from genome sequence alone have largely facilitated our understanding on a variety of biological questions. Although the computational prediction of protein-coding genes

has already been well-established, we are also facing challenges to robustly find the non-coding RNA genes, such as miRNA. Two main aspects of ab initio gene prediction include the computed values for describing sequence features and used algorithm for training the discriminant function, and by which different combinations are employed into various bioinformatics tools. There are many general properties of DNA sequence, such as GC content and base composition, having been well used for *in silico* analysis.

Coding sequences are not only more conserved but also contain ORF must have bearing on base composition or other sequence attributes. Introns are non-coding sections of a gene, transcribed into the precursor mRNA sequence, but ultimately removed by RNA splicing during the processing to mature messenger RNA. Many introns appear to be mobile genetic elements.

GC skew is when the nucleotides G and C are over or under abundant in a very specific region of deoxyribonucleic acid or ribonucleic acid. In equilibrium conditions (without modification or selective pressure and with nucleotides indiscriminately distributed within the genome) there's an equal frequency of the four deoxyribonucleic acid bases (Adenine, Guanine, Thymine, and Cytosine) on each single strands of a deoxyribonucleic acid molecule. GC skews are shaped by asymmetric accumulation of specific mutations which are determined at two levels, namely strand biased mutation and subsequent selection G (Marais *et al*., 2015). GC skews, as expressed by (G-C)/ (G+C) and AT skew, expressed by (A-T)/ (A+T) by bacterial chromosomes (Marais *et al*., 2015). The GC skew is a useful indicator of the DNA leading strand, lagging strand, replication origin, and replication terminal. Most bacteria and archaea contain only one DNA replication origin. The GC skew is positive and negative in the leading strand and in the lagging strand respectively; therefore, it is expected to see a switch in GC skew sign just at the point of DNA replication origin and terminus (Marais *et al*., 2015).

The coding sequences of genome interest most to the biologists since they are responsible for encoding proteins. However scientists are increasing getting interested in non-coding part of genomes also. The tandem repeat sequences are relatively easier to identify when compared to dispersed repeats. However it is difficult to differentiate between coding sequences and non-repetitive non-coding sequences such as regulatory regions introns unless DNA sequence is searched for open reading frames (ORFs). This is in particular a problems in eukaryotes where certain exons can be very short and thus do not have significantly long ORFs. It was interesting to investigate a difference between coding and non-coding sequences based on base composition and other related attributes rather than searching for ORFs. In the present study some of such attributes have been analysed to find if coding and non-coding sequences can be differentiated. DNA walk (random walk) representation can be used to extract useful non-trivial characteristics from sequence data, which are then applicable to evolutionary sequence comparisons or delineation of yet unknown genetic function and diversity (Berger et al. 2004).

CHAPTER - #2

REVIEW OF LITERATURE

DNA base composition is a fundamental genomic feature that impacts codon usage, DNA methylation, speciation, genome organization and phylogenetic inference. Genomic sequencing makes it possible to examine all genomic patterns as well as their potential mechanisms (Carter *et al.,* 2013). Significant research progress has been made in three areas: DNA base composition, mutation and DNA repair systems. Among these three research areas, the first is the individual-strand base composition (Carter *et al., 2013).* The Chargaff's first parity rule (PR1) (*i.e.* [A] = [T]and [G] =[C]) of nucleotide base composition in double stranded DNA was an integral pre-requisite to the Watson–Crick's double-helical model (G Marais *et al., 2004).* The less-known, second parity rule (PR2) [A] ≈ [T]and [G] ≈[C] summarize the observation that for each individual strand of a DNA duplex (Chargaff *et al.,* 1950, 1952). Although the validity of PR2 was previously demonstrated at the genome level and various theories have been proposed no further large-scale studies using a diverse set of species with sequenced whole genomes have been reported (Marais *et al., 2004).* This rule has been proved largely true except in some small DNA molecules such as the mitochondrial (MT) DNAs.

Beside the overall genomic nucleotide, it was found that DNA template strands have more pyrimidine nucleotides (i.e. RNAs are purine rich), which was lately named the Szybalski's rule Forsdyke (Dang *et al*., 1998; Lao and Forsdyke, 2000). More critically, no studies on individual-strand base composition across single nucleotide polymorphisms (SNPs) have been reported using a population of related individuals. Because these polymorphic sites constitute the dynamic part of the genome and are most abundant, it would be interesting to study whether there is a pattern, which may contribute to our understanding of the underlying mechanisms for individual-strand base equality. The second related research area is mutation bias. Mostly studies were performed in the pre-genomics era and it was sometimes based on incomplete genomic data. Base mutation was shown to have a bias in the direction of A: T and that newly emerged low-frequency SNP alleles are typically A: T rich. Human chromosomes comprise two subsets of bands, the GC-richest H3+ and the GC-poorest L1+ bands, accounting for about 17 and 26%, respectively, of all bands. The former are a subset of the R bands and the latter are a subset of the G bands. These band showed the highest and the lowest gene densities, respectively, as well as a number of other distinct features. Here we report that human and chicken interphase nuclei are characterized by the following features :(1) the gene-richest/GC-richest chromosomal regions are predominantly distributed in internal locations, whereas the gene-poorest/GC-poorest DNA regions are close to the nuclear envelope. (2) The interphase chromosomes seem to be characterized by a polar arrangement, because the gene-richest/GC-richest bands and the gene poorest bands are predominantly located in the distal and proximal regions, respectively, of chromosomes, and because interphase chromosomes are extremely long most of the chromosomes are only composed of GC rich regions.

Large-scale DNA deformation is ubiquitous in transcriptional regulation in prokaryotes and eukaryotes alike. Though much is known about how transcription factors and constellations of binding sites dictate where and how gene regulation will occur, less is known about the role played by the intervening DNA (Johnson *et al.* 2013). Non uniformity of nucleotide composition within genomic sequences from a variety of taxa ranging from phages to mammals was revealed several decades ago by thermal melting and gradient centrifugation experiments (Inman 1966; Filipski et al. 1973). The human genome appears to be highly compositionally heterogeneous both within and between individual chromosomes; the heterogeneity goes much beyond the predictions of the isochore model. The extent of the compositional heterogeneity in a genomic sequence strongly correlates with its GC content in all multicellular eukaryotes studied regardless of genome size. The degree of regional homogeneity in base composition in the human genome is a fundamental property of the genome sequence. Not only does it characterize the organization and evolution of the genome, but it also provides a context for many practical sequence analyses. Statistical quantities such as GC%, used in sequence analyses for purposes such as computational gene recognition, should be sampled from a homogeneous region of the sequence. The human genome consists of regions that differ in their GC content. Regions greater than 3 kb in length with a high degree of GC content uniformity are termed isochores. In general, genomes of higher eukaryotes contain regions of high GC content, absent from genomes of lower organisms (Bernardi, *et al*. 1993). During the evolution of homeotherms, the gene-rich and moderately GC-rich isochores. This increase was maintained during the evolution of mammalian and avian orders (Bernardi, 2000). The genomic organization of GC-rich regions differs significantly from that of GC-poor regions. For example, it has been found that chromosomal regions of high GC content exhibit higher gene densities (Bernardi, 2000; Lander et al., 2001) and, hence, higher CpG island densities (Cross et al., 2000).

The AT-rich DNA is mostly associated with condensed chromatin, whereas the GC-rich sequence is preferably located in the dispersed chromatin. The AT-rich genes are prone to be tissue-specific (silenced in most tissues), while the GC-rich genes tend to be housekeeping (expressed in many tissues). It shows another important property of DNA base composition, which can affect genes with high AT content. The GC-rich sequence is more liable to mutation. We observed that Spearman correlation between human gene GC content and mutation probability is above 0.9. The change of base composition even in synonymous sites affects mutation probability of nonsynonymous sites and thus of encoded proteins. There is a unique type of housekeeping genes, which are especially unsafe when prone to mutation.

DNA base composition is a fundamental genome feature. However, the evolutionary pattern of base composition and its potential causes have not been well understood. The findings from comparative analysis of base composition at the whole genome level across 2210 species, the polymorphic site level across eight population comparison sets, and the mutation-site level in 12 mutation-tracking experiments. Firstly, demonstrate that base composition follows the individual-strand base equality rule at the genome, chromosome and polymorphic site levels.

More intriguingly, clear separation of base-composition values calculated across polymorphic sites was consistently observed between basal and derived groups, suggesting common underlying mechanisms. Individuals in the derived groups show an A&T-increase/G&C-decrease pattern compared with the basal groups. Spontaneous and induced mutation experiments indicated these patterns of base composition change can emerge across mutation sites.

The isochore model of the mammalian genome describes patterns in base composition 4. Although on average the mammalian genome is 40% GC, the distribution is far from uniform. Rather, the genome is a mosaic of long stretches of DNA (>300kb in length) that are homogeneous in base composition but varying from less than 38% GC to more than 55% GC. The existence of these compositionally homogeneous stretches of DNA, which have been termed isochores (meaning 'equal regions' Evolutionary modelling research demonstrated the relative effect of mutation rate and fixation probability in shaping human base composition (Gaylord strew *et al., 2019*). For example, substantial biases towards lower overall GC content were discovered from sequence comparison between human and other primates at both paralogous repeat elements and orthologous genes. However, genetic relationships of human individuals were not examined in earlier studies where AT frequency was examined across human SNPs. It can be questioned this change in base frequency will emerge if we compare individual-strand base composition across polymorphic sites between populations separated by a bottleneck event. In the third research area, fine control of the mutation process in living organisms has been revealed by the elucidation of DNA repair mechanisms in model species and the discovery of hundreds of associated genes in humans.

From an evolutionary perspective, an efficient and effective DNA repair system is essential to the maintenance of genome integrity; mutations in DNA repair genes can lead to hyper mutated genomes, severe diseases, and cancers (Ruobing & Guan*, et al. 2019*). On the other hand, some nucleotide changes must inevitably escape this surveillance system to provide genetic variability to fuel the evolutionary process. However, evolutionary divergence of these DNA repair genes themselves in different human groups has not been extensively studied. One recent study suggested that low fidelity DNA replication by polymerase is partly responsible for the observed multi nucleotide mutation in the human population. Contrasting genome-wide sequence polymorphisms between populations separated by a bottleneck would allow us to probe potential roles of DNA repair systems in long-term genome evolution. During a DNA sequencing task, the nucleotides of the reads must be placed in the correct order to reconstruct the original sequence. This approach is expected to help improve assembly projects by reducing search spaces when grouping related sequence fragments. Massively parallel next-generation sequencing technologies (Sanger's method of the 1980's) provide high throughput results at a low cost but the reads are often too short to be able to determine their adjacency (Bolívar *et al., 2019).* Some of the limitations encountered in the assembly process include read coverage and size. The absence of placement information such as read coverage forms a bottleneck in the reassembly process. When the read sequences are very short, then special procedures must be

taken to maximize their informational content to achieve placement evidence. It may be necessary to form reads by de novo assembly methods as in. Despite these limitations, technologies such as Velvet and Oases have been used for many genome assembly projects and assembling reads using approaches from probability theory, or from the memory-based, are gaining popularity (Bolívar *et al., 2019).*

RNA interference (RNAi) technology is widely used in scientific research as a genetic tool *(Boettcher and* McManus,*et al. 2015; Blake et al., 2017).* It is more likely to be used as a new approach in agricultural pest control (Burand and Hunter,*et al*. 2013; Kimetal *et al*.,2015;Jogaetal *et al*.,2016).RNAi can be triggered by introducing double-stranded RNA(dsRNA),which is processed into effective small interfering RNAs (siRNAs) by the Dicer enzyme. Then, the generated siRNAs are incorporated into the RISC complex with other proteins, enter into the subsequent RNAi pathway, and then cause the gene silencing effect (Fireetal *et al* 1998;Tijsterman and Plasterk *et al*. 2004; Winteretal *et al.* 2009). Therefore, the Dicer enzyme processing of the dsRNA into siRNAs is the key step in the RNAi pathway; however, it is not clear how dsRNA is recognized and cleavage by the Dicer enzyme, or what kinds of siRNAs will be produced in vivo.

The isochore model of the mammalian genome describes patterns in base composition 4. Although on average the mammalian genome is 40% GC, the distribution is far from uniform. Rather, the genome is a mosaic of long stretches of DNA (>300kb in length) that are homogeneous in base composition but varying from less than 38% GC to more than 55% GC. The existence of these compositionally homogeneous stretches of DNA, which have been termed isochores (meaning 'equal regions') has been demonstrated by fractionating (Xianran Li1 *et al., 2015).* Exogenous dsRNA enters the insect body and can induce the RNAi effect only when it is cleaved into siRNA. However, what kinds of base composition are easier to cut and what kinds of siRNA will be produced in vivo is largely unknown. In this study, we found that dsRNA processing into siRNA has sequence preference and regularity in insects. They injected 0.04 mg/g dsRNA into Asian corn borers or cotton bollworms according to their body weight, and then the siRNAs produced in vivo were analysed by RNA-Sequence.

Discovery was done that has large number of siRNAs were produced with GGU nucleotide residues at the 50- and 30-ends and produced a siRNA peak on the sequence. Once the GGU site is mutated, the number of siRNAs will decrease significantly and the siRNA peak will also be lost. However, in the red flour beetle, a member of Coleoptera, dsRNA was cut at more diverse sites, such as AAG ,GUG, and GUU; more importantly, these enzyme restriction sites have a high conservation base of A/U. Our discovery regarding dsRNA in vivo cleavage preference and regularity will help us understand the RNAi mechanism and its application (Ruobing Guan *et al.,2019).* With the proposal of the neutral theory of molecular evolution in 1968, Kimura introduced a revolutionizing concept to evolutionary biology, which at that time was

strongly influenced by the view that evolution is driven by positive Darwinian selection. Instead, Kimura proposed that at the molecular level deleterious mutations are common, while advantageous mutations are rare, and that in a finite population most evolutionary changes are a consequence of the fixation of neutral mutations due to genetic drift. Kimura thus put a new emphasis on the stochasticity of population genetics, and further established the relationship between sequence conservation and functional importance, which is key to many bioinformatics software for the identification of conserved coding as well as non-coding elements *(Bolívar1 et al., 2013).* Sequence reads and contigs often exhibit the same kinds of base usage that is also observed in the sequences from which they are derived, we offer a base composition analysis tool. The tool uses these natural patterns to determine relatedness across sequence data. They have introduced spectrum sets (sets of motifs) which are permutations of bacterial restriction sites and the base composition analysis framework to measure their proportional content in sequence data (Bolívar1 *et al.,* 2013*).*

CHAPTER - #3

OBJECTIVES

OBJECTIVES

1. To download and segregate DNA sequences as coding and non-coding regions (exons & introns respectively in the case of human genes)
2. To determine frequency of the four bases, dinucleotides, trinucleotides and tetranucleotides for all the sequences
3. Analysis of the frequencies of bases and their permutation to study effect of base composition on coding and non-coding sequences

CHAPTER - #4

MATERIALS AND METHODS

DATA SOURCE: The DNA sequence data was obtained from NCBI mentioned in the table below:

PROKARYOTES:

S.NO	NAME OF THE ORGANISM	ACCESSION NUMBER
1.	*Escherichia coli*	NC_000913.3
2.	*Haemophilus influenzae*	NC_000907.1
3.	*Mycobacterium tuberculosis*	NC_000962.3
4.	*Bacillus subtilis*	NC_000964.3
5.	*Thermus aquaticus*	NZ_CP010822.1
6.	*Deinococcus radiodurans*	NC_001263.1
7.	*Agrobacterium tumefaciens*	NZ_005042.1
8.	*Prochlorococcus marinus*	NC_005042.1
9.	*Borrelia burgdoferi*	NC_001318.1
10.	*Chlamydia trachomatis*	NC_075110.1
11.	*Acidobacterium capsulatum*	NC_012483.1
12.	*Nitrospira moscoviensis*	NZ_CP000021.9

Human genes:

S.NO	NAME OF THE GENE	ACCESSION NUMBER
1.	Hsa_Dnmt 1	NC_000019.10
2.	Hsa_TRDnmt 1	NC_000010.11
3.	Hsa_Dnmt3a	NC_000002.12
4.	Has_Dnmt3b	NC_000020.11

5.	Nanog_chr6	NC_000072.6
6.	Plg_chr6	NC_000002.12
7.	Pole4_chr11	NC_007122.7
8.	GADD45a_chr1	NC_000001.11
9.	Hsa_Tet1	NC_000076.6
10.	Hsa_Dkk1	NC_000010.1
11.	Hsa_Ercc5	NC_000013.11
12.	Has_Smarca4	NC_000019.10

SEQUENCE ANALYSIS TOOLS:

1. **FCGR**: It was used for the counting of bases in di, tri, and tetra nucleotides for statistical analysis of the frequencies.
2. **MS EXCEL**: M.S. Excel tool was used for the statistical and computational analysis of the frequencies.

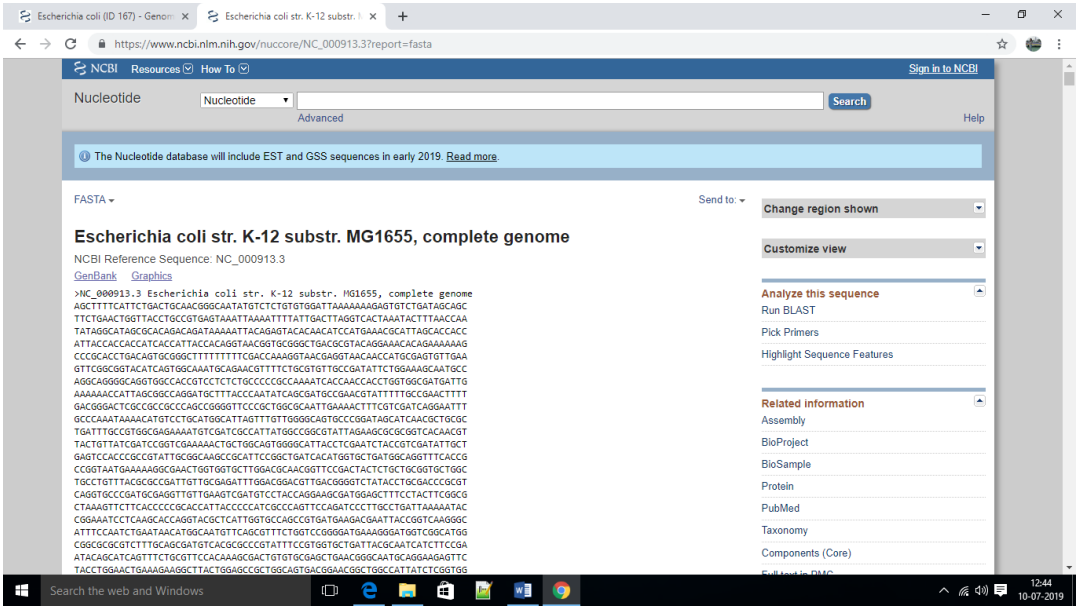
Following tools were used:

S.NO	TOOLS	CLASS	FUNCTION
1	LEN	Text	It is use to find the number of characteristics between the coding and non-coding regions.
2	SUM	Math and Trigonometry	It is use to find the sum of coding and non-coding regions.
3	STDEVP	Statistical	It is use to find the deviation
4	AVERAGE	Statistical	It is use to find the arithmetic of data.
5	COUNTIF, COUNT, IF,	Statistical	It is used to count the no of sequences

3. **Notepad++:** It was used for recording and then running Macros for DNA sequence manipulation and analysis. For various sequence manipulations Macros were recorded and run multiple times to carry out the desired process.

METHODS:

1. The nucleotide sequences of bacterial and human chromosomes were obtained from NCBI with the particular given accession number in the GenBank format.

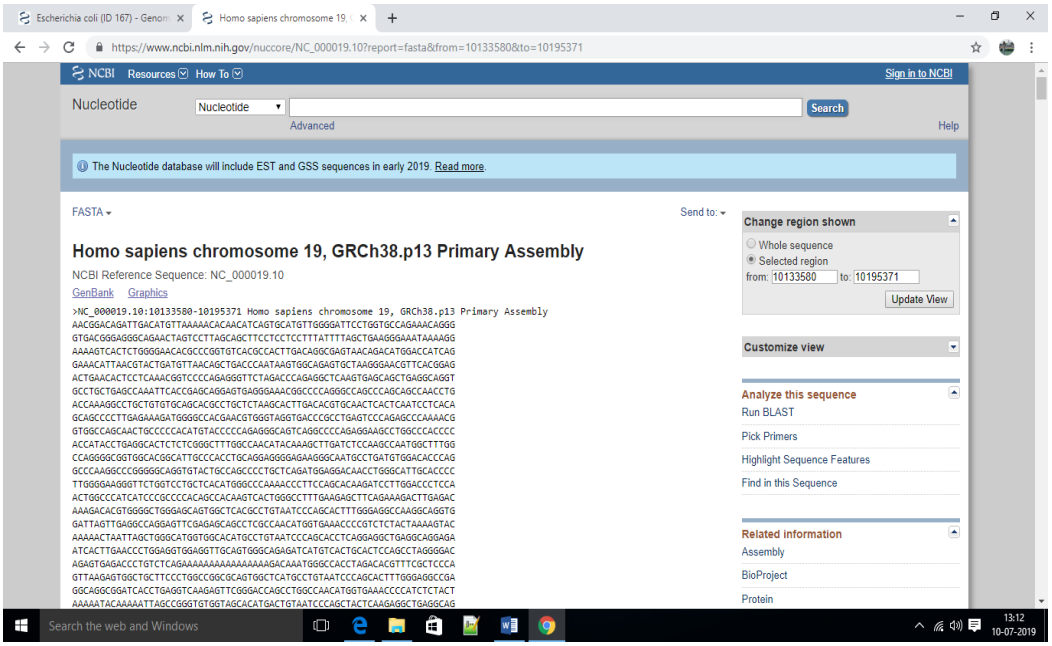


Screenshot 1

Annotations of the sequence were analysed using MS Excel to determine the start and end positions of all the genes present in the sequence. The start and end positions of the coding sequences were used to find the start and end positions of the non-coding regions (comprising of promoter, regulatory and intergenic sequences). Using the positions of all the coding and non-coding regions, the sequences were downloaded from NCBI. All the coding sequences were joined by putting a 'J' character as separator between adjacent sequences. 'J' character was selected as it does not represent any base. Similarly, non-coding sequence were also down loaded and joined into a single string of characters.

In case of Eukaryotes, following steps were followed:

1. The exons and introns of 12 different genes were made into a single file and each sequence was separated by exon and introns numbering respectively



Screenshot 2

2. Then, all the exons are compiled in a separate file which is again separated by J
3. The process was repeated for introns
4. The above given two steps was repeated for all the 12 genes of human chromosome 21
5. The conversion was made using Notepad++ (it was converted into one-word sequence file)

CHAPTER - #5

RESULTS

The coding sequences of genomes are studied with much attention as they encode for the proteins. However, many non-coding sequences are also equally or comparably important such as promoters and other regulatory sequences. In this study, it has been attempted to compare coding and non-coding sequences based on analysis of the DNA base compositions and frequency of their permutations that is di-, tri- and tetra-nucleotide motifs. In prokaryotes, most of the DNA sequence is coding as prokaryotic DNA rarely contains any kind of repetitive sequences. As a result, in case of prokaryotes comparison was performed between coding sequences and any other sequences present in the genome. However eukaryotic genomes, especially higher eukaryotes only a small fraction of genome comprises of coding sequences. For example, human genome has only ~2% of genome coding for proteins and rest of the genome consists of promoters and transcription regulatory sequences, introns and tandem and dispersed repeat sequences. Keeping this fact in consideration, comparison has been performed between exons (coding) and introns (non-coding sequences) of randomly selected human genes.

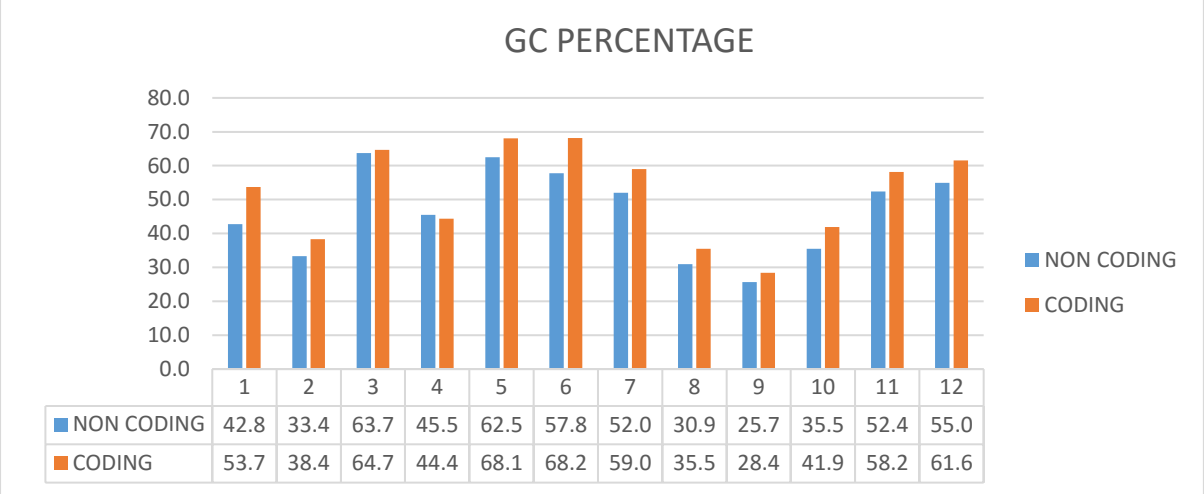
A sample of twelve randomly selected genomic sequences were downloaded in FASTA format from twelve bacteria of diverse evolutionary lineages. The sequences were fragmented into coding and non-coding sequences each class was subjected to FCGR tool to determine all the four bases, 16 dinucleotides, 64 trinucleotides and 256 tetra nucleotides frequencies. Further twelve genes were selected from human genome in a random fashion and fragmented into exon and intron sequences. Similarly, base composition and di-, tri- and tetra-nucleotide frequencies were determined for exons and introns of each gene for further analysis.

Measurement of GC% of coding and non-coding sequences:

One of the important characteristic of a genome or other DNA sequence is its GC% which is related to several properties of the DNA such as bendability, thermosability and resistibility. Additionally, it has also been found that GC rich sequences are associated with gene regions. In order to investigate it GC% was determined for all the coding and non-coding sequences of bacterial as well as human origin.

A similar observation was made when GC% or exons was compared with intron. The GC% of exons was found to be greater than in introns in most of the cases. A paired sample *t*-test showed that the difference was statistically significant with a p-value of 0.000105842. Thus it was inferred that irrespective of the sources i.e. bacterial or eukaryotic genomes the coding sequences have higher GC% when compared to their non-coding counterparts.

GC% was calculated of all twelve bacterial genomes:

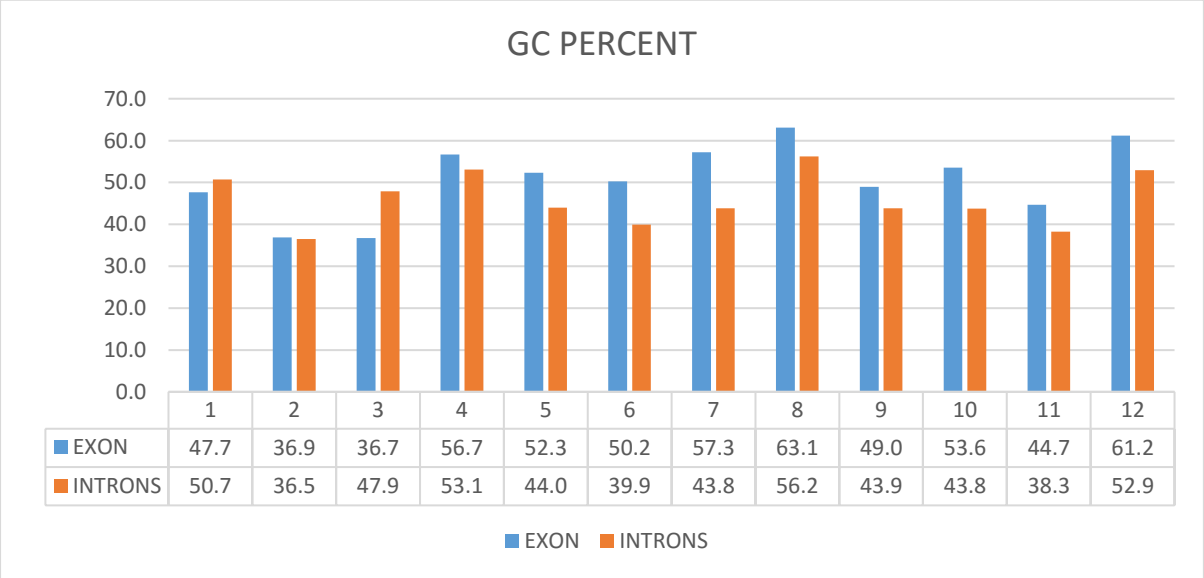


	1	2	3	4	5	6	7	8	9	10	11	12
NON CODING	42.8	33.4	63.7	45.5	62.5	57.8	52.0	30.9	25.7	35.5	52.4	55.0
CODING	53.7	38.4	64.7	44.4	68.1	68.2	59.0	35.5	28.4	41.9	58.2	61.6

Figure 2: Graphical representation of G C content in prokaryotes (Non Coding vs Coding)

It was observed that in most of the bacteria, the coding sequences had higher GC% when compared to the respective non-coding sequence. The difference was analysed by paired sample *t*-test and it was found to be statistically significant with a p-value of 0.014867.

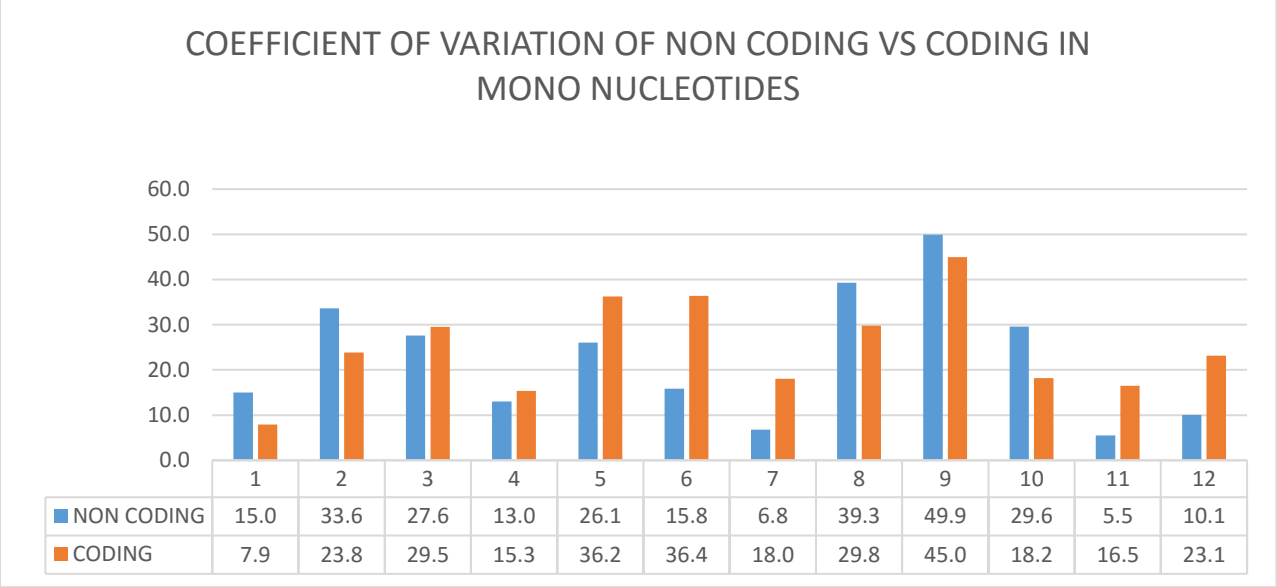
In case of eukaryotes:



	1	2	3	4	5	6	7	8	9	10	11	12
EXON	47.7	36.9	36.7	56.7	52.3	50.2	57.3	63.1	49.0	53.6	44.7	61.2
INTRONS	50.7	36.5	47.9	53.1	44.0	39.9	43.8	56.2	43.9	43.8	38.3	52.9

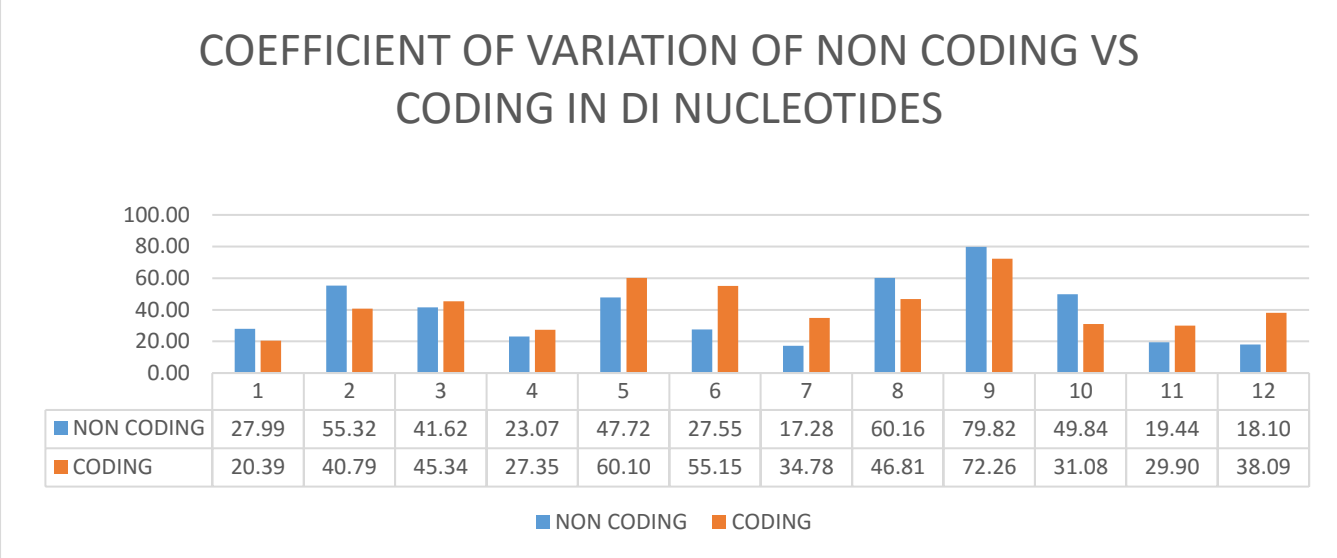
Figure 3: Graphical representation for GC% of exons and introns

Measurement of coefficient of variation:



	1	2	3	4	5	6	7	8	9	10	11	12
NON CODING	15.0	33.6	27.6	13.0	26.1	15.8	6.8	39.3	49.9	29.6	5.5	10.1
CODING	7.9	23.8	29.5	15.3	36.2	36.4	18.0	29.8	45.0	18.2	16.5	23.1

Figure 4: Graphical representation of coefficient of variation of non-coding vs coding in mono nucleotides



	1	2	3	4	5	6	7	8	9	10	11	12
NON CODING	27.99	55.32	41.62	23.07	47.72	27.55	17.28	60.16	79.82	49.84	19.44	18.10
CODING	20.39	40.79	45.34	27.35	60.10	55.15	34.78	46.81	72.26	31.08	29.90	38.09

Figure 5: Graphical representation of coefficient of variation of non-coding vs coding in di nucleotides

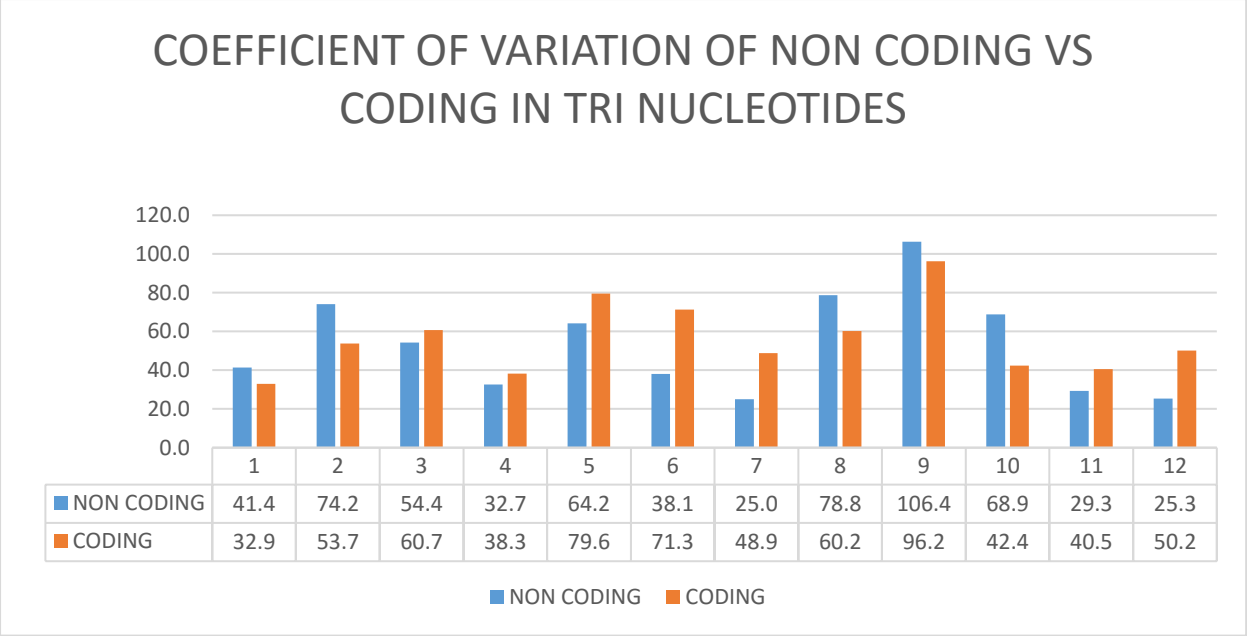


Figure 6: Graphical representation coefficient of variation of non-coding vs coding in tri nucleotides

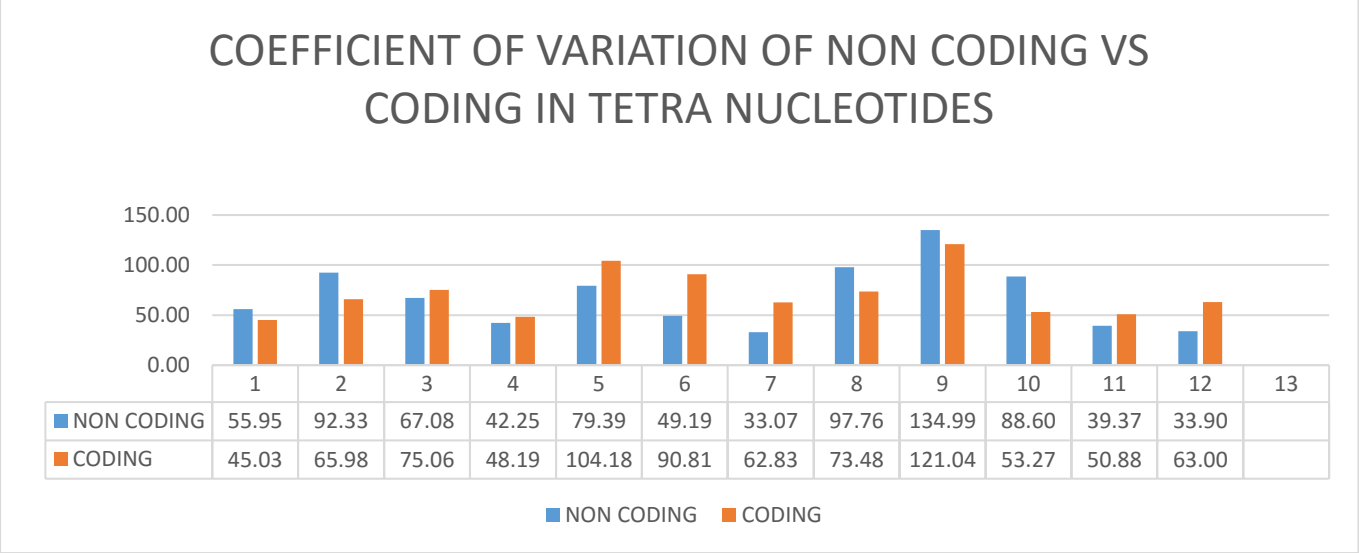
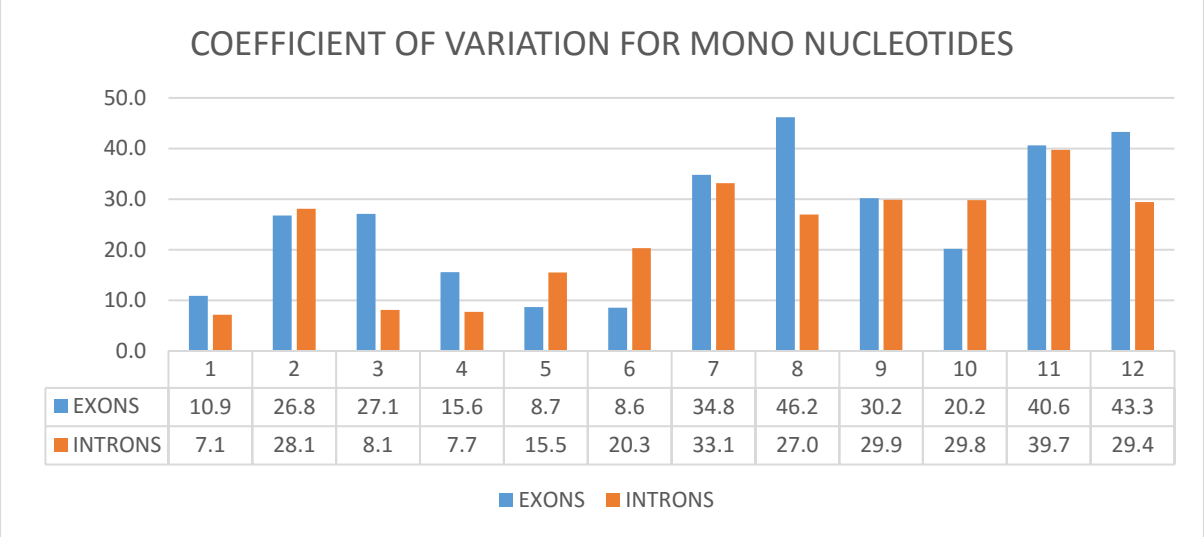


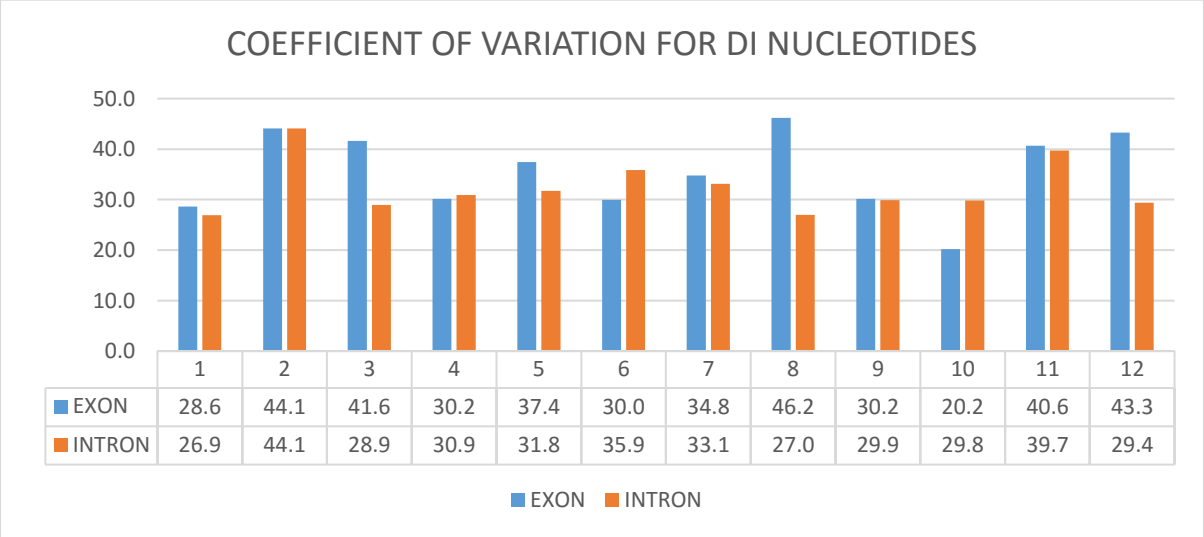
Figure 7: Graphical representation coefficient of variation of non-coding vs coding in tetra nucleotides

Measurement of coefficient of variation of exons and introns:



	1	2	3	4	5	6	7	8	9	10	11	12
EXONS	10.9	26.8	27.1	15.6	8.7	8.6	34.8	46.2	30.2	20.2	40.6	43.3
INTRONS	7.1	28.1	8.1	7.7	15.5	20.3	33.1	27.0	29.9	29.8	39.7	29.4

Figure 8: Graphical representation for coefficient of variation of exons and introns exons and introns in mono nucleotides



	1	2	3	4	5	6	7	8	9	10	11	12
EXON	28.6	44.1	41.6	30.2	37.4	30.0	34.8	46.2	30.2	20.2	40.6	43.3
INTRON	26.9	44.1	28.9	30.9	31.8	35.9	33.1	27.0	29.9	29.8	39.7	29.4

Figure 9: Graphical representation for coefficient of variation of exons and introns exons and introns in Di nucleotides

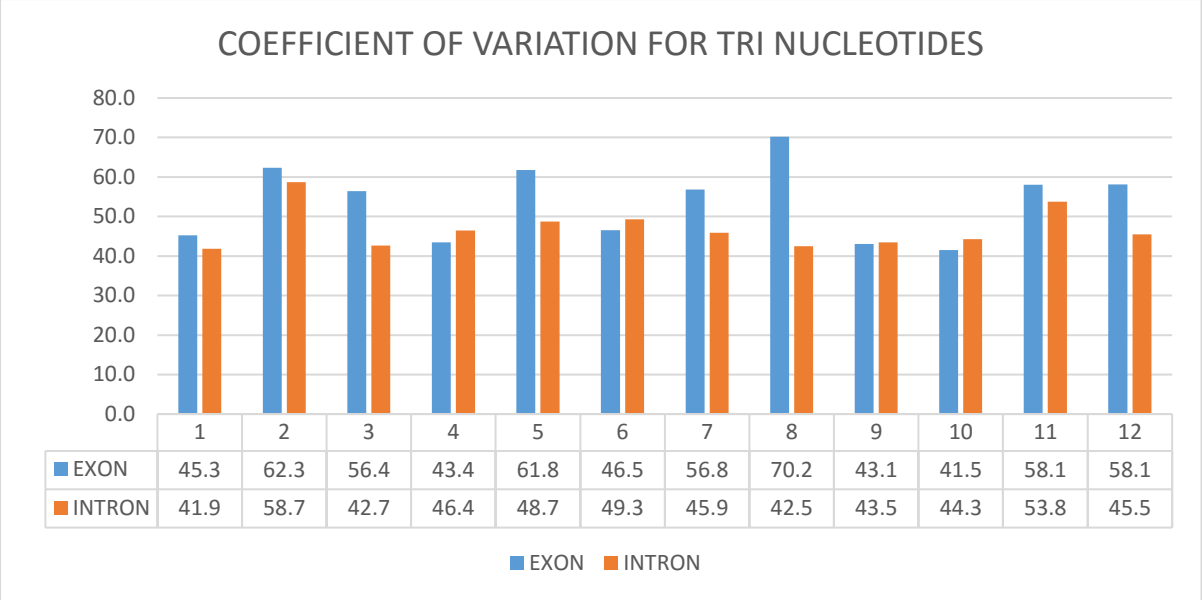


Figure 10: Graphical representation for coefficient of variation of exons and introns exons and introns in tri nucleotides

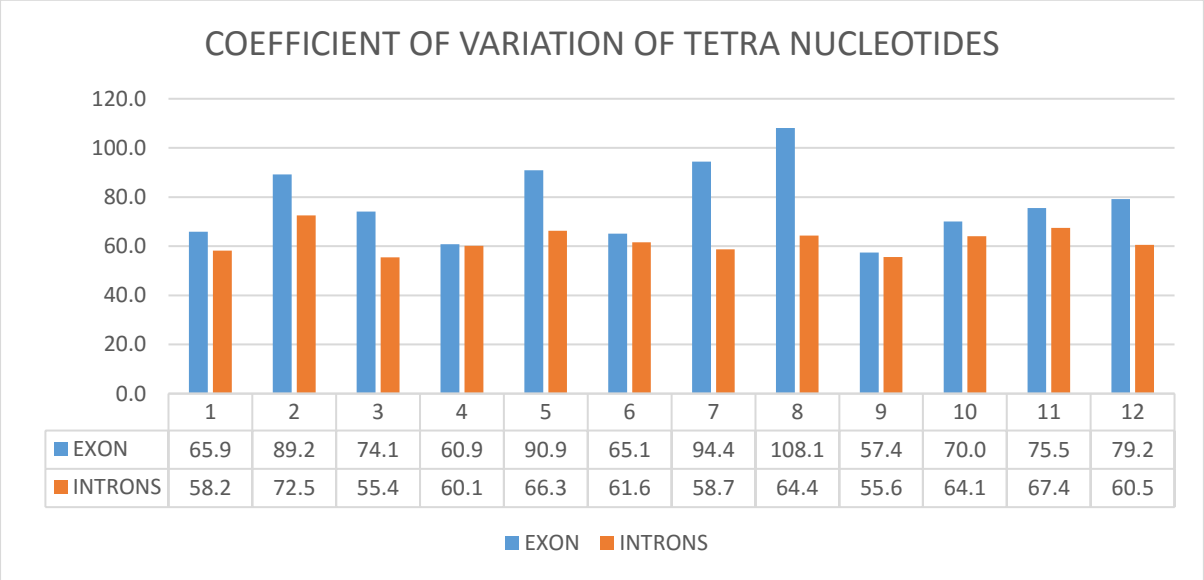


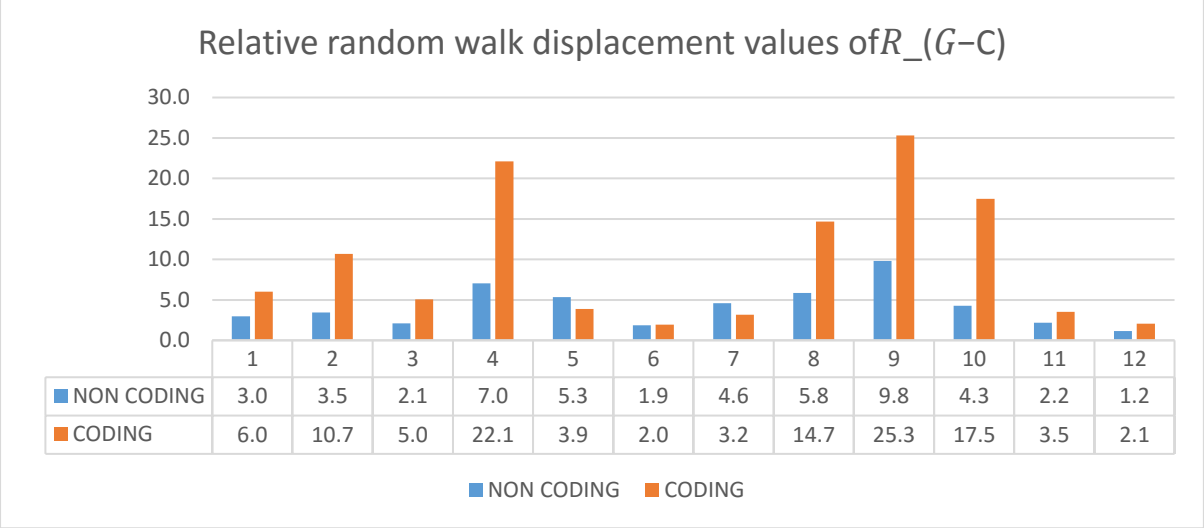
Figure 11: Graphical representation for coefficient of variation of exons and introns exons and introns in tetra nucleotides

Coefficient of variation was used as a measurement of unevenness of the base composition. Coefficient of variation was calculated for di, tri, tetra, frequencies. However, no distinct pattern could be observed between coding and non-coding sequences. Coding and non-coding sequences do not have significant difference in the unevenness of base composition.

Measurement of Random walk:

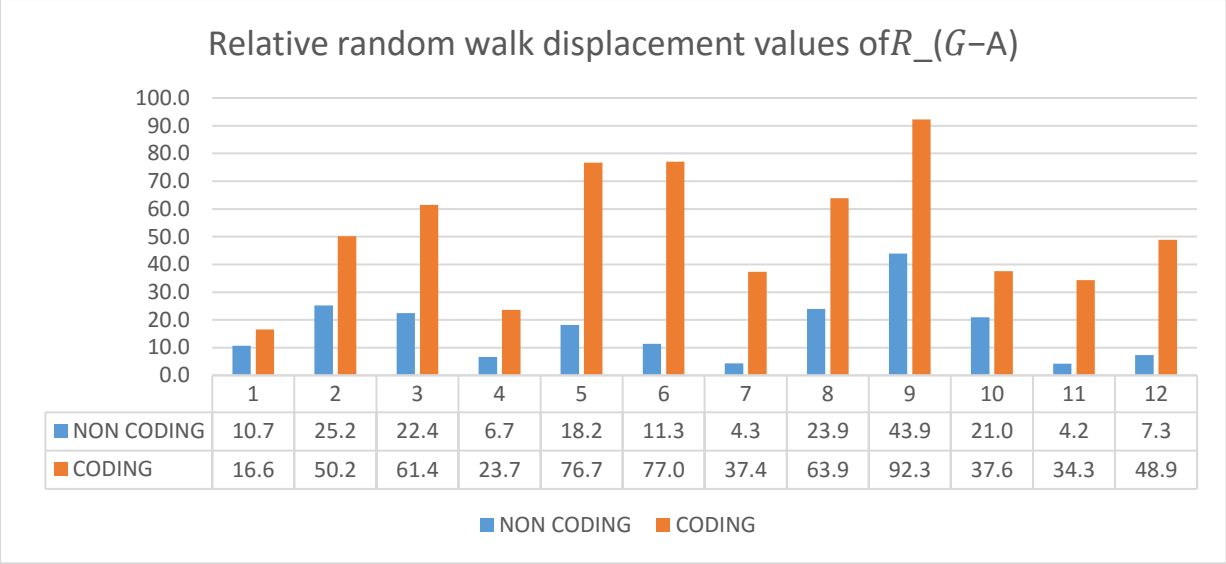
DNA sequence can be graphically represented in a manner similar to a 2D random walk. For example for each G, the graph moves one step up towards higher value on Y axis, for A, one step down towards lower value on Y-axis. Similarly, for a T, the graph moves one step right along with higher values on X-axis and for each C, one step left towards lower values on X-axis. The final displacement 'R' covered by the sequence will be$R_{G-A} = \sqrt{(G-A)^2 + (T-C)^2}$. However the sequence can be represented on two more distinct manners as: $R_{G-T} = \sqrt{(G-T)^2 + (A-C)^2}$ &$R_{G-C} = \sqrt{(G-C)^2 + (A-T)^2}$. The 2Df random walk displacement values were divided by the square root of the length of the sequence in order to normalize the displacement values and represented as R' (relative random walk displacement values). The coding and non-coding sequences of bacteria human gene were compared by calculating the above mentioned three DNA 2D random walk displacement values.

Measurement of Random walk of coding and non-coding sequences:



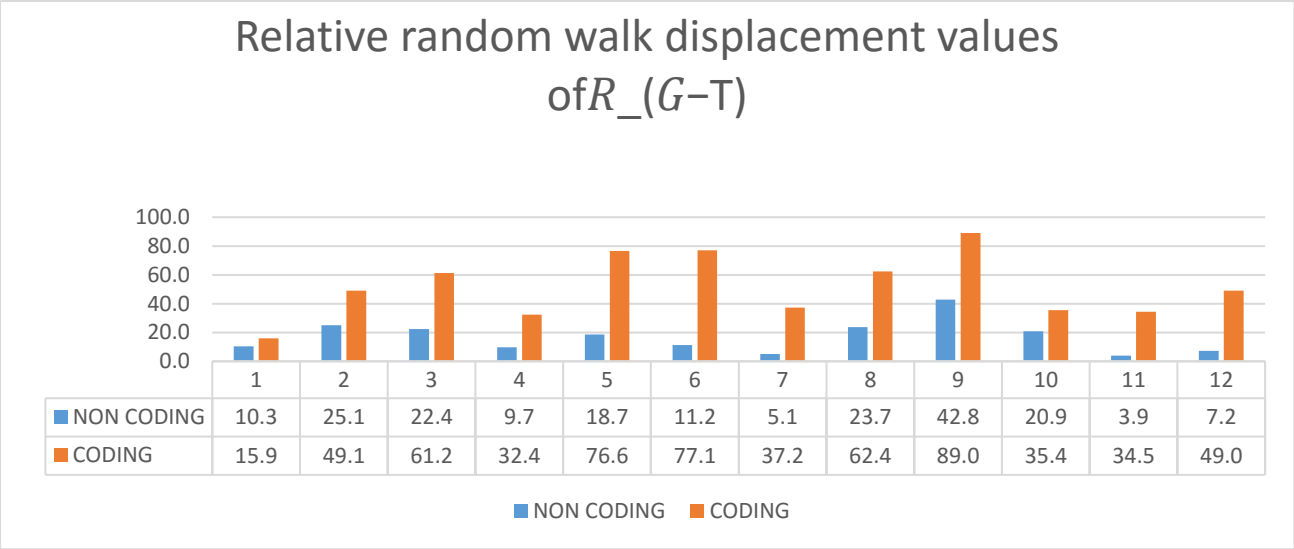
	1	2	3	4	5	6	7	8	9	10	11	12
NON CODING	3.0	3.5	2.1	7.0	5.3	1.9	4.6	5.8	9.8	4.3	2.2	1.2
CODING	6.0	10.7	5.0	22.1	3.9	2.0	3.2	14.7	25.3	17.5	3.5	2.1

Figure 12: Graphical representation for random walk in non-coding vs coding



	1	2	3	4	5	6	7	8	9	10	11	12
NON CODING	10.7	25.2	22.4	6.7	18.2	11.3	4.3	23.9	43.9	21.0	4.2	7.3
CODING	16.6	50.2	61.4	23.7	76.7	77.0	37.4	63.9	92.3	37.6	34.3	48.9

Graphical representation for random walk in non-coding vs coding



	1	2	3	4	5	6	7	8	9	10	11	12
NON CODING	10.3	25.1	22.4	9.7	18.7	11.2	5.1	23.7	42.8	20.9	3.9	7.2
CODING	15.9	49.1	61.2	32.4	76.6	77.1	37.2	62.4	89.0	35.4	34.5	49.0

Figure 13: Graphical representation for random walk in non-coding vs coding

Measurement of random walk in eukaryotes:

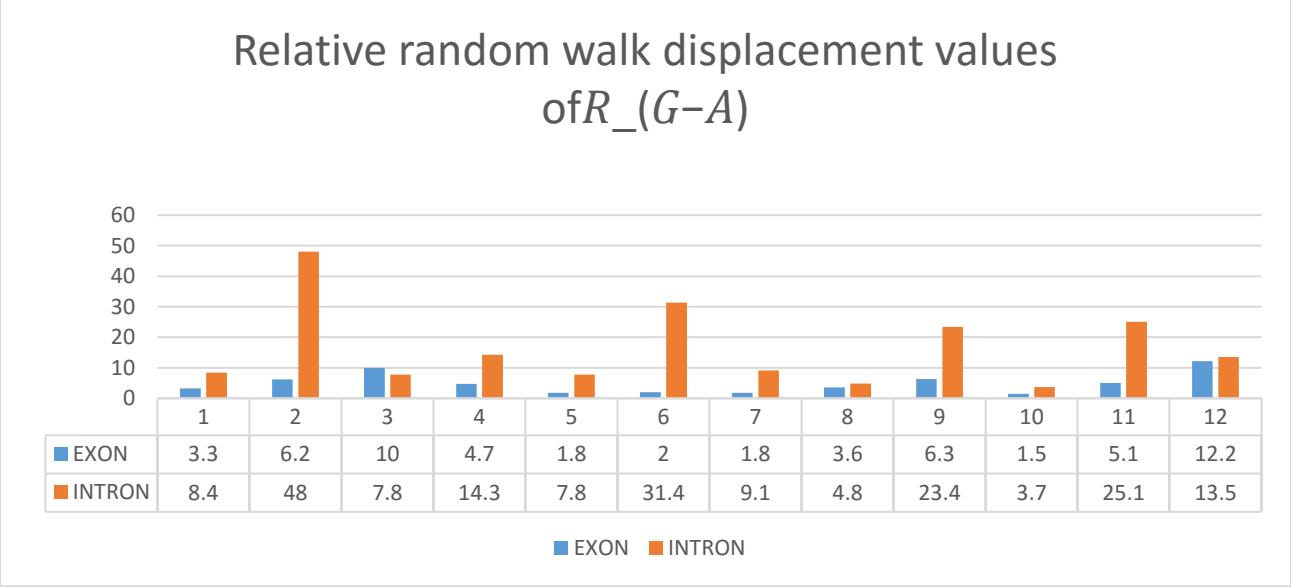


Figure 14: Graphical representation of random walk for exons and introns

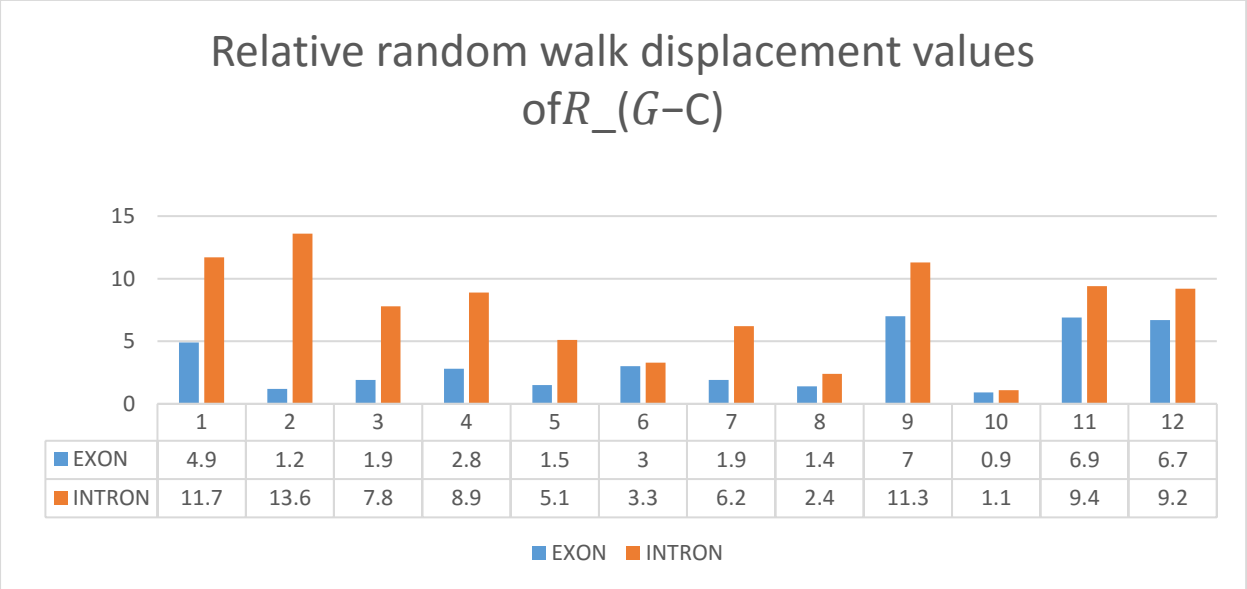


Figure 15: Graphical representation of random walk for exons and introns

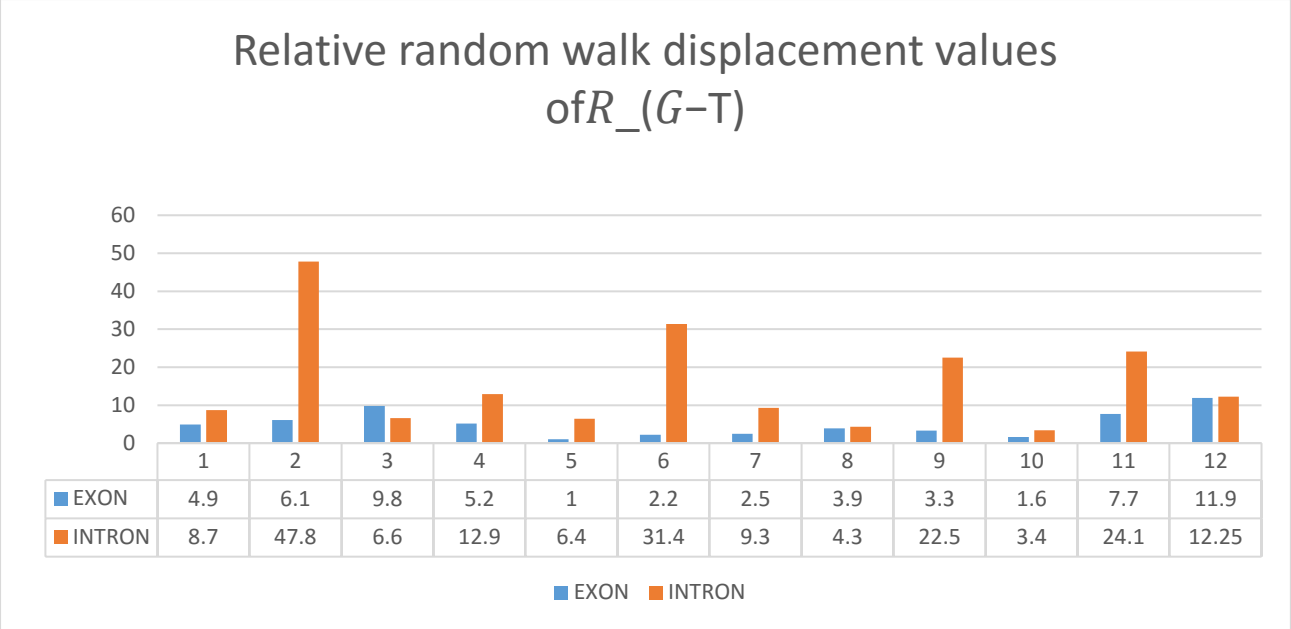


Figure 16: Graphical representation of random walk for exons and introns

In the measurement of random walk in case of prokaryotes the values pertaining to coding sequences are greater than those of non-coding sequences in all three cases *i.e.* $R'_{G\text{-}A}$, $R'_{G\text{-}T}$ and $R'_{G\text{-}C}$. While introns (non-coding sequences) show greater displacement than in exons (coding) in all three cases *i.e.* $R'_{G\text{-}A}$, $R'_{G\text{-}T}$ and $R'_{G\text{-}C}$. The opposite trend observed in bacteria and human genes either indicates that introns exhibit higher 2D random walk displacement than exons or may be it is influenced by the total length of the sequences for introns are much longer than exons just as coding sequences are much longer than non-coding sequences in bacteria.

No inference can be drawn because the coefficient of variation is getting affected by the length of the sequence.

The unevenness of the base composition is expected to have bearing on sequence complexity and thus it may be related to the information content of the DNA sequence.

CHAPTER - #6

DISCUSSION

DISCUSSION

All the genetic information is stored in the DNA. Although, there are certain changes which are heritable and which are not accurately encoded in the DNA sequences. DNA stores genetic information in the form of sequence of four bases C, A, T or U (in RNA), G and C in nucleic acids. Replication takes place in it. In prokaryotes, most of the DNA sequences is coding because we know prokaryotic DNA rarely contains any kind of non-coding or intergenic sequences. In case of lower eukaryotic DNA contains very few genes i.e., number of genes are always less. In higher eukaryotes mostly genes consist of introns and the number of introns are always more in number. Split genes are those which contains introns. Repetitive sequence is mainly found in intergenic regions that is the region between the adjacent genes.

Usually DNA sequence analysis is limited to sequence alignments, determining ORFs and GC% (isochores). This is an attempt is to analyse DNA sequence on the basis of the bases components and their attributes.

Except tandom repeats all the sequence appears to be very similar. This analysis was done to catch subtle difference between coding and non-coding sequences.

The coding sequence genes encode for proteins. The human genome appears to be highly multicellular eukaryotes regardless of genome size. The extent of the compositional heterogeneity in all multicellular. GC content is found to be variable with different organisms, the process of which is envisaged to be contributed to by variation in selection, mutational bias, and biased recombination-associated DNA repair.

G-C rich regions differs significantly from that of G-C poor regions. The average GC-content in human genomes ranges from 35% to 60% across 100-Kb fragments, with a mean of 46.1%.

In simpler organisms the whole DNA has unique sequences. Unique sequence are those which consist of coding sequences and does not have any repetitive sequence but in case of higher organism there can be large amount of repetitive DNA. Two types of repetitive DNA: 1. Tandem repeats 2. .Dispersed repeats.

The length of the non-repetitive DNA tend to increase as the organisms become more complex. The presence of large amount of DNA indicates the presence of repetitive DNA. Mostly genes are present in non-repetitive form of DNA.

GC content is a primary factor shaping amino acid compositions. GC content shapes amino acid composition to trade off the cost of amino acids with bases which could be caused by the energy efficiency. GC% of coding is higher in both prokaryotes and eukaryotes. Coefficient of variation is a measure of unevenness of base composition which in prokaryotes the coding is greater than non-coding while in case of eukaryotes the introns are greater than exons. In the

measurement of random walk in case of prokaryotes the coding is greater than non-coding in all three cases i.e. G-A, G-T and G-C. While intron is greater than exon in all three cases i.e. G-A, G-T and G-C.

In random walk the final displacement 'R' covered by the sequence will be $R_{G-A} = \sqrt{(G-A)^2 + (T-C)^2}$. However the sequence can be represented on two more distinct manners as: $R_{G-T} = \sqrt{(G-T)^2 + (A-C)^2}$ &$R_{G-C} = \sqrt{(G-C)^2 + (A-T)^2}$. The 2Df random walk displacement values were divided by the square root of the length of the sequence in order to normalize the displacement values and represented as R' (relative random walk displacement values). The coding and non-coding sequences of bacteria human gene were compared by calculating the above mentioned three DNA 2D random walk displacement values, further work required to be done.

The unevenness of the base composition is expected to have bearing on sequence complexity and thus it may be related to the information content of the DNA sequence. GC% was obtained from both the sequences i.e. bacterial and human genome sequences, but in case of coefficient of variation and random walk further work is required to be done

CHAPTER - #7

CONCLUSION

CONCLUSION

Through the present work we have analysed the information complexity. Here we have analysed the base sequence using different parameters like GC%, C.V (Coefficient of variation) and random walk. From all three parameters we have drawn the conclusion that GC% of coding is higher in both prokaryotes and eukaryotes. Coefficient of variation is a measure of unevenness of base composition which in prokaryotes the coding is greater than non-coding while in case of eukaryotes the introns are greater than exons. In the measurement of random walk in case of prokaryotes the coding is greater than non-coding in all three cases i.e. G-A, G-T and G-C. While intron is greater than exon in all three cases i.e. G-A, G-T and G-C. The unevenness of the base composition is expected to have bearing on sequence complexity and thus it may be related to the information content of the DNA sequence. GC% was obtained from both the sequences i.e. bacterial and human genome sequences, but in case of coefficient of variation and random walk further work is required to be done.

CHAPTER - #8

REFERENCE

REFERENCES

1. Akashi, H., & Gojobori, T. (2002). Metabolic efficiency and amino acid composition in the proteomes of Escherichia coliandBacillus subtilis. *Proceedings of the National Academy of Sciences, 99*(6), 3695-3700. doi:10.1073/pnas.062526999
2. Belozersky, A. N., & Spirin, A. S. (1958). A Correlation between the Compositions of Deoxyribonucleic and Ribonucleic Acids. *Nature, 182*(4628), 111-112. doi:10.1038/182111a0
3. Bolívar, P., Guéguen, L., Duret, L., Ellegren, H., & Mugal, C. F. (2019). GC-biased gene conversion conceals the prediction of the nearly neutral theory in avian genomes. *Genome Biology, 20*(1). doi:10.1186/s13059-018-1613-z
4. Bonham-Carter, O., Ali, H., & Bastola, D. (2013). A base composition analysis of natural patterns for the preprocessing of metagenome sequences. *BMC Bioinformatics, 14*(S11). doi:10.1186/1471-2105-14-s11-s5
5. Federico, C., Scavo, C., Cantarella, C. D., Motta, S., Saccone, S., & Bernardi, G. (2006). Gene-rich and gene-poor chromosomal regions have different locations in the interphase nuclei of cold-blooded vertebrates. *Chromosoma, 115*(2), 123-128. doi:10.1007/s00412-005-0039-z
6. Galtier, N. (2004). Faculty of 1000 evaluation for Recombination and base composition: The case of the highly self-fertilizing plant Arabidopsis thaliana. *F1000 - Post-publication Peer Review of the Biomedical Literature*. doi:10.3410/f.1020148.230461
7. Gardiner, K. (1996). Base composition and gene distribution: Critical patterns in mammalian genome organization. *Trends in Genetics, 12*(12), 519-524. doi:10.1016/s0168-9525(97)81400-x
8. Guan, R., Hu, S., Li, H., Shi, Z., & Miao, X. (2018). The in vivo dsRNA Cleavage Has Sequence Preference in Insects. *Frontiers in Physiology, 9*. doi:10.3389/fphys.2018.01768
9. Guo, Y. (2017). An overview on the DNA nucleotide compositions across kingdoms. *Gene Reports, 8*, 45-48. doi:10.1016/j.genrep.2017.05.003
10. Huang, Y., Chen, S., & Deng, F. (2016). Well-characterized sequence features of eukaryote genomes and implications for ab initio gene prediction. *Computational and Structural Biotechnology Journal, 14*, 298-303. doi:10.1016/j.csbj.2016.07.002
11. Johnson, S., Chen, Y., & Phillips, R. (2013). Poly(dA:dT)-Rich DNAs Are Highly Flexible in the Context of DNA Looping. *PLoS ONE, 8*(10). doi:10.1371/journal.pone.0075799
12. Li, X., Scanlon, M. J., & Yu, J. (2015). Evolutionary patterns of DNA base composition and correlation to polymorphisms in DNA repair systems. *Nucleic Acids Research, 43*(7), 3614-3625. doi:10.1093/nar/gkv197
13. Panda, A., Podder, S., Chakraborty, S., & Ghosh, T. C. (2014). GC-made protein disorder sheds new light on vertebrate evolution. *Genomics, 104*(6), 530-537. doi:10.1016/j.ygeno.2014.09.003

14. Paul, P., Malakar, A. K., & Chakraborty, S. (2017). Codon usage and amino acid usage influence genes expression level. *Genetica, 146*(1), 53-63. doi:10.1007/s10709-017-9996-4
15. Salvo, M. D., Puccio, S., Peano, C., Lacour, S., & Alifano, P. (2019). RhoTermPredict: An algorithm for predicting Rho-dependent transcription terminators based on Escherichia coli, Bacillus subtilis and Salmonella enterica databases.
16. Ulmschneider, M. B., & Sansom, M. S. (2001). Amino acid distributions in integral membrane protein structures. *Biochimica Et Biophysica Acta (BBA) - Biomembranes, 1512*(1), 1-14. doi:10.1016/s0005-2736(01)00299-1
17. Whittle, C. A., & Extavour, C. G. (2015). Codon and Amino Acid Usage Are Shaped by Selection Across Divergent Model Organisms of the Pancrustacea. *G3& #58; Genes|Genomes|Genetics, 5*(11), 2307-2321. doi:10.1534/g3.115.021402
18. Yin, S., Heckman, J., & Rajbhandary, U. L. (1981). Highly conserved GC-rich palindromic DNA sequences flank tRNA genes in Neurospora crassa mitochondria. *Cell, 26*(3), 325-332. doi:10.1016/0092-8674(81)90201-4
19. Zhang, R., & Zhang, C. (2005). Identification of replication origins in archaeal genomes based on theZ-curve method. *Archaea, 1*(5), 335-346. doi:10.1155/2005/509646
20. Takai and Jones (2001). Comprehensive analysis of CpG ISLANDS IN HUMAN CHROM0SOME 21 AND 22.

THESIS

ORIGINALITY REPORT

10%	6%	5%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Burridge, James, Celestina N. Jochua, Alexander Bucksch, and Jonathan P. Lynch. "Legume shovelomics: High—Throughput phenotyping of common bean (Phaseolus vulgaris L.) and cowpea (Vigna unguiculata subsp, unguiculata) root architecture in the field", Field Crops Research, 2016. Publication	1%
2	ethesis.nitrkl.ac.in Internet Source	1%
3	www.frontiersin.org Internet Source	1%

4	www.indjst.org Internet Source	1%
5	P. Khatri, V. Desai, A. L. Tarca, S. Sellamuthu, D. E. Wildman, R. Romero, S. Draghici. "New Onto-Tools: Promoter-Express, nsSNPCounter and Onto-Translate", Nucleic Acids Research, 2006 Publication	1%
6	www.biologicscorp.com Internet Source	1%
7	fs.wirtschaft.fh-trier.de Internet Source	1%
8	dspace.thapar.edu:8080 Internet Source	1%
9	Paola Astolfi, Dina Bellizzi, Vittorio Sgaramella. "Frequency and coverage of trinucleotide repeats in eukaryotes", Gene, 2003 Publication	<1%

10	C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, M. Simons, H.E. Stanley. "Statistical properties of DNA sequences", Physica A: Statistical Mechanics and its Applications, 1995 Publication	<1%
11	wiki2.org Internet Source	<1%
12	www.science.gov Internet Source	<1%
13	krishikosh.egranth.ac.in Internet Source	<1%
14	Judit Morla-Folch, Ramon A. Alvarez-Puebla, Luca Guerrini. "Direct Quantification of DNA Base Composition by Surface-Enhanced Raman Scattering Spectroscopy", The Journal of Physical Chemistry Letters, 2016 Publication	<1%

15	www.jbc.org Internet Source	<1%
16	epdf.tips Internet Source	<1%
17	Béatrice Heurtault, Patrick Saulnier, Brigitte Pech, Marie-Claire Venier-Julienne et al. "The influence of lipid nanocapsule composition on their size distribution", European Journal of Pharmaceutical Sciences, 2003 Publication	<1%
18	Ying Huang, Shi-Yi Chen, Feilong Deng. "Well-characterized sequence features of eukaryote genomes and implications for ab initio gene prediction", Computational and Structural Biotechnology Journal, 2016 Publication	<1%
19	link.springer.com Internet Source	<1%

	D. Rajyalakshmi, K. Kishore Raju, G.P. Saradhi 20 Varma. "Taxonomy of Satellite Image and Validation Using Statistical Inference", 2016 IEEE 6th International Conference on Advanced Computing (IACC), 2016 Publication	<1%
21	bmcgenomics.biomedcentral.com Internet Source	<1%
22	wikileaks.org Internet Source	<1%

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		