An Evolutionary Studies of Deoxyribose Nucleic Acid (DNA) Sequences Towards Signal Processing

Masthan Sheik[1], Dr. Subhasis Bose[2]

[1]Research Scholar, Department of Electronics and Communication Engineering, Sri Satya Sai University of Technology & Medical Sciences, Sehore, M.P., India.
[2]Research Guide, Department of Electronics and Communication Engineering, Sri Satya Sai University of Technology & Medical Sciences, Sehore, M.P., India.

Abstract

Large quantities of hidden information are encoded in the DNA sequences, which are essential to the functioning of an organism In the DNA sequence, areas of importance include the following: exons; introns; genetic regions; intergenic; untranslated; translation sites; promoting regions; splicing locations; microsatellites; and minisatellites repeated many times. Included are several parts of the genome. Models that depend on and represent these features have been developed using a variety of computation methods. For model-dependent methods, prior knowledge of the data sets and higher precision are required. Model dependency means these methods can't be improved upon. Model-independent methods, on the other hand, don't need any prior knowledge or training. In many cases, the independent models rely on the fact that the DNA sequence has distinct intervals depending on where you are. Using signal processing methods, these periodicities may be identified across various genetic regions. It is called Genomic Signal Processing because it uses Signal Processing Tools for genetic data processing. This research has developed signal processing methods for locating short exons, receiver splice sites, microsatellites, and minisatellites in DNA sequences.

Keywords: *Deoxyribose Nucleic Acid, Signal Processing, Model dependency, DNA Sequence, Genomic Signal Processing.*

1. INTRODUCTION

Recent advances in bioinformatics have made it possible to analyse genomic data sets using signal processing methods. Genomics The application of signal processing methods for the extraction of genomic datasets' hidden information is known as signal processing (GSP). Genomic information is made up of DNA sequences found in cell nuclei. DNA sequences contain the four nucleotide bases adenine ('A,' thymine ('T,' guanine ('G,' and cytosine ('C'). The intergenic and genic portions of eukaryotic cell DNA sequences (nuclei cell) may be separated. Exonic and intronic regions were created by further dividing the genomic domains. The intronic regions lack a period-3 component. These areas behave like period-3 due to codon bias, and they are connected to the protein coding regions. Protein coding regions are in charge of synthesizing different proteins in living organisms. As part of a larger picture, splice sites help to properly identify protein coding regions. Accepters' splice sites (intron and exon border) and donors' splice sites may be distinguished (boundary of exon and intron). DNA sequences include repeating patterns from many eras, not only period-3. More than two contiguous DNA sequence repeating patterns are used in tandem repeats (TRs). For the study of tandem repetitions, factors such as pattern size, pattern structure, copy count, and pattern location are important factors. Depending on the size of the repeating pattern, TRs may be classified as microsatellites, minisatellites, or satellites. At this time, the human genome has more than 10,000 STR sequences. Tandem repeats are caused by neurodegenerative diseases including Huntington's disease, fragile X syndrome, and mythological dystrophy, as well as cancer. Individual genetic profiling begins with the use of short tandem repeats (STRs). In the medical, forensic, and phylogenic disciplines, studying genetic data is critical for the

well-being of humans and other living things.

2. LITERATURE REVIEW

Inbamalar, T. M. (2015) Computational methods are used in bioinformatics and genomic signal processing to address a wide range of biological issues. Specifically, they're interested in genetic materials like DNA, RNA, and proteins, as well as the information associated with them. One of the most difficult and time-consuming jobs in analysis is locating the DNA sequences that encode proteins. Existing DSP techniques offer a less precise and computationally expensive solution with a higher level of background noise. In order to identify protein coding regions in DNA sequences, it is critical to increase accuracy, computational complexity, and background noise. To find protein coding areas in DNA sequences, this study introduces a novel DSP-based approach that uses artificial intelligence. Here, DNA sequences are represented as electron ion interaction potential (EIIP) sequences, which are then transformed to numeric sequences. Once this is done, a discrete wavelet transform is applied. After determining the energy's absolute value, the next step is determining the appropriate threshold. The test is carried out with the help of the NCBI's databases. The comparison analysis has been completed, and it confirms the suggested system's efficacy.

Susana Vinga is the author (2014) The field of information theory (IT) has found widespread use in molecular biology because of its focus on communication systems analysis. Concepts from IT, such as entropy and mutual information, were particularly useful for alignment-free sequence analysis. From genome global analysis and comparison, including block-entropy estimation and resolution-free metrics based on iterative maps, to local analysis, including motif classification, transcription factor binding site prediction and sequence characterization using linguistic complexity and entropic profiles, this review examines several aspects of IT applications in great detail. DNA, RNA, and protein features have been combined with sequence-independent properties such as gene mapping and phenotypic analysis, and models based on communication systems theory have been used to describe information transmission channels at the cellular level and also during evolutionary processes. However, this review attempts to classify existing methods and indicate their relation to broader transversal topics such as genomic signatures, data compression and complexity, time series analysis as well as classification of living organisms, providing a resource for future developments in this promising field.

Theodora S. Mabrouk (2017) The discipline of bioinformatics has now firmly established itself as a molecular biology control and includes a wide range of information from structural biology, genomics, and gene expression research. When it comes to managing biological data, bioinformatics applies computers to the task. These GSP methods have been widely used in bioinformatics and will continue to play an important role in the study of biological problems. Genomic Signal Processing (GSP) refers to the use of DSP techniques for genomic data analysis (e.g., DNA sequences). Applications of GSP in bioinformatics, including as the identification of DNA protein coding regions, the identification of reading frames, and the detection of cancer, have recently received considerable attention. Malignant neoplasm, the medical name for cancer, is one of the world's most deadly illnesses and has seen a rise in the mortality rate in recent years. Early diagnosis may provide a promising method to determining and taking action to address this risk. Cancerous cells are often caused by genetic abnormalities, and GSP is a technique for detecting them. This review examines the use of GSP in bioinformatics in general. Recent findings and what has been achieved so far as a new area of study are collected using GSP methods, which are specifically utilised for cancer diagnosis.

Ning Yu (2018) Using data-driven machine learning, particularly deep learning, in bioinformatics is becoming more and more essential. DNA sequences are often transformed to numerical values in machine learning for data representation and feature learning. Genomic Signal Processing (GSP) uses a similar transformation to extract and recognise signals from genomic sequences that are numerical sequences. Encoding strategy is another term for this kind of conversion. GSP applications and machine learning models' performance may be significantly impacted by the various encoding methods. Genomic sequence coding schemes in GSP as well as other genome analysis applications are collected, discussed and summarized for use as a complete reference in machine learning for the representation and feature learning in genomic data.

Carl Sedlá, M. Sc (2018) Modern environmental microbiology research makes use of genomic data, particularly DNA sequencing, to characterize microbial populations. Metagenomics is the study of all genetic material found in a

sample of environmental material. Bioinformatics plays an irreplaceable role during data processing in this doctorate thesis on metagenomics. Theoretically, this thesis compares and contrasts two distinct metagenomics methods, outlining their key concepts and flaws. Targeted sequencing is a well-established area with a broad variety of bioinformatics methods for the first method. There are ways to enhance techniques for comparing samples from different settings. Here, data is transformed into a bipartite network, where one partition is made up of taxa and the other is composed of data from samples or habitats, using a brand-new method. A graph like this captures both the qualitative and quantitative aspects of microbial networks under study. It enables for significant data reduction while maintaining the ability to identify communities of comparable samples and their characteristic microorganisms automatically. The second strategy makes use of massively parallel shotgun sequencing of the whole metagenome. A fresh approach, using less established bioinformatics methods, is being used. The most difficult part of metagenomics is binning, the rapid grouping of sequences. A genomic signal processing technique is used in this thesis's method. A new method was devised after a careful examination of the genetic information contained in genomic signals and a comprehensive study of its redundancy. Character sequences are transformed into a variety of phase signals using this method. Additionally, it can handle nanopore sequencing data as a native current signal immediately without converting it beforehand.

Po-Yen L. Wu (2017) Facilitated by -omic data, precision medicine is a promising medical model that may revolutionize the quality of the current healthcare system. Currently, -omic data are being rapidly accumulated because of the advent of high-throughput -omic assays. Though challenging, abundant information embedded in these data is encouraging for the realization of precision medicine. Data analytics, including data pre-processing and data modeling techniques, has been successfully applied to many -omic applications, and biomarkers identified from -omic data are viewed as catalyzers for precision medicine.

Anjuli Meiser is the person in question (2017) Ecological communities are increasingly being characterized via metagenome skimming, or low-coverage shotgun sequencing of several species assemblages, followed by reconstruction of individual genomes. This is a potential method for reconstructing the genomes of facultative symbionts such as lichen-forming fungus using metagenomic reads, such as metagenomic sequence data. The accuracy and completeness of metagenomic-based assemblies have not been compared to those derived from pure culture strains of lichenized fungus in any studies to far. We used metagenomic sequences from lichen thalli to reconstruct the genomes of Evernia prunastri and Pseudevernia furfuracea. Two distinct taxonomic binning techniques were used to extract fungal contigs, and gene prediction was conducted on the fungal contig subsets. We next used genome assemblies based on pure culture strains of the two fungal species as a reference to evaluate the quality and completeness of the metagenome-based assemblies.

3. ARTIFICIAL NEURAL NETWORKS

The promise of Artificial Neural Networks (ANNs) has attracted scholars from a wide range of disciplines. The NN approach has its roots in efforts to develop a computer model of the nervous system's information processing capabilities. The information learned is stored in a neural network, which may be thought of as a large collection of connections with different strengths (weights).

ANNs seem to be an excellent alternative technology. Synapses in the brain are mathematically approximated by artificial neural networks (ANNs), which were initially developed to investigate how perception works in the brain. Instead of understanding the brain's processes, neural networks have been used to mimic non-linear mapping and offer ANNs with a high internal representation capability. A new theory of neural networks, along with technical advances, has transformed them into an effective tool for detecting complex processes that are missing or damaged in some way, including information, collinearity, and temporal delays. Without basing models or postulated formulas on insufficient data, it is also possible to make use of this method. Numerous properties of the neural network favoured its use to the analysis of protein and nucleic acid sequences. To transmit knowledge to the network, ANNs utilise both positive and negative information. This may be done in both sequences with and without the feature of interest. As a result, they are more useful than conventional methods for identifying complex relationships that are based only on the frequency of particular residues occurring. It automatically detains which residues and which places are important using a knowledge-based ANN during training. In biological macromolecules research, neural networks are widely utilised because of their superior performance for parallel sequence processing Nucleic acid and protein sequences are being used to create maps of spatial structure and

characteristics.

4. GENOMICS

Genomics is the study of an organism's genome to uncover previously unknown traits. The genome refers to the whole collection of a living organism's DNA sequences. A single human cell has approximately 3 billion DNA base pairs, which is about the average for all living things. The DNA sequence contains the nucleotides Adenine, Thymine, Guanine, and Cytosine. DNA sequences are like a blueprint for building a human body. Chromosomes connect together DNA sequences. Which may be found in the cell's nucleus? Enzymes and chemical messengers help these cells make the specific proteins they need. DNA gene information is translated into messenger RNA molecules by enzymes (mRNA). In the cytoplasm, ribosomes read mRNA that has travelled from the nucleus. In order to produce a specific protein, the mRNA must be linked to an amino acid sequence, and this is done by the ribosome. To help with the development of organs and tissues, the protein acts as a chemical reaction monitor and signal transducer. Creating an abnormal protein is conceivable if the DNA sequence has changed. These mutant proteins may disrupt the body's normal functioning, increasing the risk of disease like cancer.

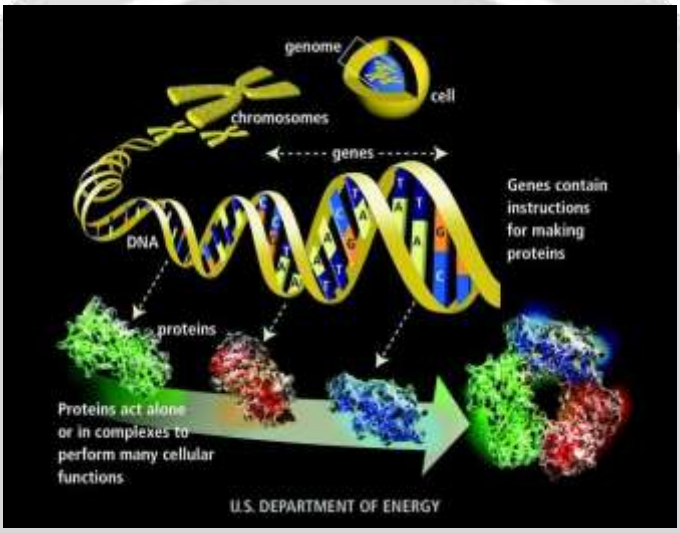


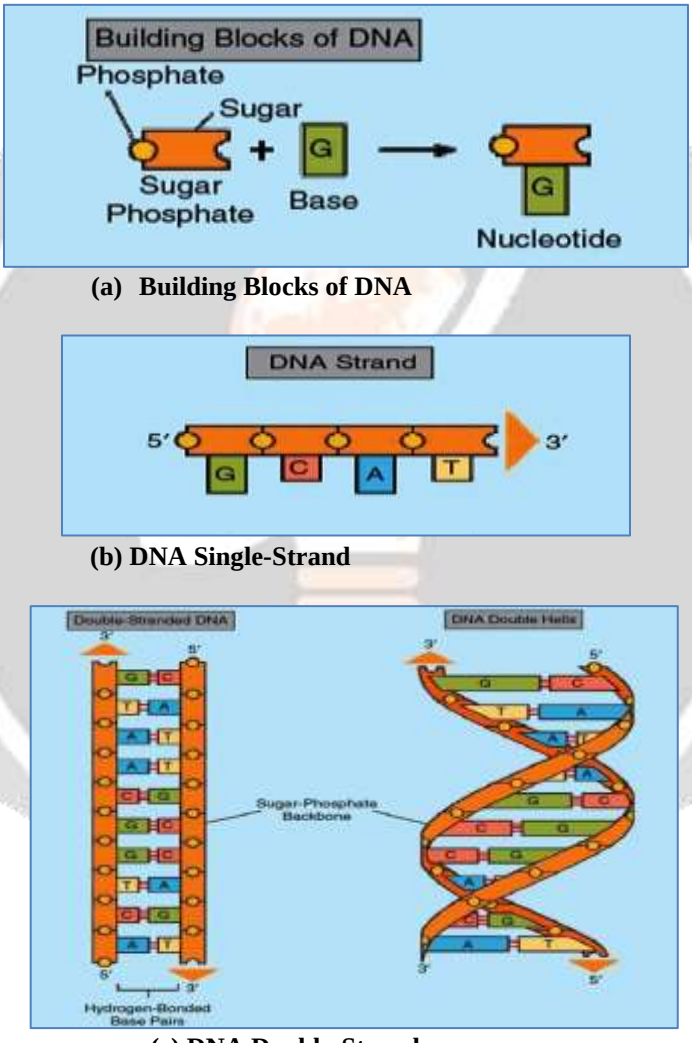
Figure 1: The Genome (Inside the Cell) Contains All of an Organism's Genetic Instructions

5. DEOXYRIBOSE NUCLEIC ACID (DNA)

To represent the genetic blueprint used by all known living organisms, DNA molecules contain digital information. Nucleotides make up the double stranded helix seen in the eukaryotic cell nucleus. Adenine (A), thymine (T), guanine (G), and cytosine (C) are among the chemical components that make up a nucleotide or base, together with deoxyribose sugar and phosphate. Fig. 2 shows a close-up of him (a). There are about three billion nucleotides in a single human cell. A zipper or high temperature may break a DNA double helix, which has two strands, in half. The complementary nature of separate strands is shown by the fact that learning about one immediately makes you aware of the other. As illustrated in Fig.1, the nucleotides that are covalently linked together form a long backbone that alternates between sugar and phosphate (b). In the one strand of DNA, each nucleotide was connected to another nucleotide through chemical contact (hydrogen bond). In other words, if A is connected to T, then T is related to B, and so on. A hydrogen bond between nucleotides and the stacking relationship between nucleotide bases have maintained the two-strand helix DNA structure. If you combine all these connections into one strong helical structure (like rope), you get something that is considerably stronger. When comparing different nucleotides, just the bases differ, thus the bases' names are given. DNA has chemical polarity because of the way nucleotides are linked together. The phosphate end is shown as 5', whereas the sugar end is shown as 3'. (Left to right). When cell machinery utilizes bases as a signal for amino acid synthesis, the DNA sequences create a lengthy, one-dimensional string that is typically listed from the 5'end to the 3'end. Figure 2 shows the covalent connection between nucleotide subunits as the basis for this standard (c). To date, no other discovery in molecular biology has come close to

matching the discovery of the DNA double helix. This molecule's legendary status has never been achieved by another. The basic signal of life, genetic information, happens as a discrete signal, in contrast to most natural signals like heat, sound, and electromagnetic waves. Different repeating DNA structures may be found in a DNA sequence. There are periodicities associated with these recurring structures (patterns). Two kinds of periodicities exist: periodicity owing to codon bias (three base periodicity) and periodicity associated with a repeat motif of a certain size. Only in the DNA genetic domain can codon biais occur at such a frequency. Information about protein synthesis may be found in the DNA genic region. All cells in an organism contain the same genes, but only a subset of those genes are active in the particular cell family. As a result, each gene is in charge of generating a unique protein.

For example, active genes are different in nerve cells from genes active in blood cells.



(a) Building Blocks of DNA

(b) DNA Single-Strand

(c) DNA Double-Strand

Figure 2: Basic Building Blocks of The DNA Structure

6. DNA SEQUENCING

A genetic substance called DNA (deoxyribonucleic acid) was demonstrated in 1994 by Oswald Theodore. DNA's double helix shape and four nucleotide bases were discovered by James D. Watson and Francis Crick in 1953, setting the stage for molecular biology as we know it today. As the genetic code that controls the structure, cell function, and many other aspects of life, DNA is critical to the survival of both species and people. In order to provide biologists and medical organisations with a broad range of applications, DNA sequencing technologies have

been created, including breeding, the detection of dangerous genes, molecular cloning and evolutionary study. Ideally, these tools will be precise, reasonably-priced, easy to use, and relatively fast to implement. Over the past 30 years, DNA sequencing technologies have advanced significantly, and they are now the engine of the genome-era, which is characterised by an enormous amount of genetic data. Frederick Sanger developed the Sanger sequencing technique in 1977, which uses the chain termination method to sequence DNA. Walter Gilbert developed a new method for sequencing DNA based on chemically altering DNA and then cleaving it at certain points along the way. In the first wave of commercial and laboratory sequencing applications, the Sanger sequencing method was the primary technology. AB370, which had been in development for many years, was first made available to the public in 1987 thanks to Applied Biosystems. AB370 uses a capillary electrophoresis method that improves accuracy and speed, allowing for the detection of 500 bases per day with a reading length of 600 bases and 96 bases per day. As of this writing, the AB3730xl model can read for up to 900 times per second and produce 2.88 M bases every day. The automated sequencing equipment and related software developed in 1998 using the capillary sequencing machine and the Sanger sequencing method played a major role in the completion of the human genome project in 2001. This study sparked the creation of low-cost, high-speed sequencing equipment for the future generation (NGS). The NGS technologies, as opposed to the Sanger method, provide superior results at a lower cost, as well as the ability to do several analyses in simultaneously. Biological explanatory analysis of hidden genomic sequence information is becoming an increasingly important problem in genomic data processing.

7. TANDEM REPEATS IN DNA

Genome analysis is a critical component of current research into how organisms operate biologically. In eukaryotes involved in gene variations and regulatory functions of gene expressions, finding DNA repeats utilizing abinitio methods is critical. In DNA sequences, the repeats may be classified as either tandem or dispersed. An arbitrary sequence of DNA symbols is found in tandem repeats (TRs) where two or more adjacent copies may be discovered. In contrast, the scattered repeats are made up of two or more copies that are not contiguous. Repeats in Fig.3 are shown.

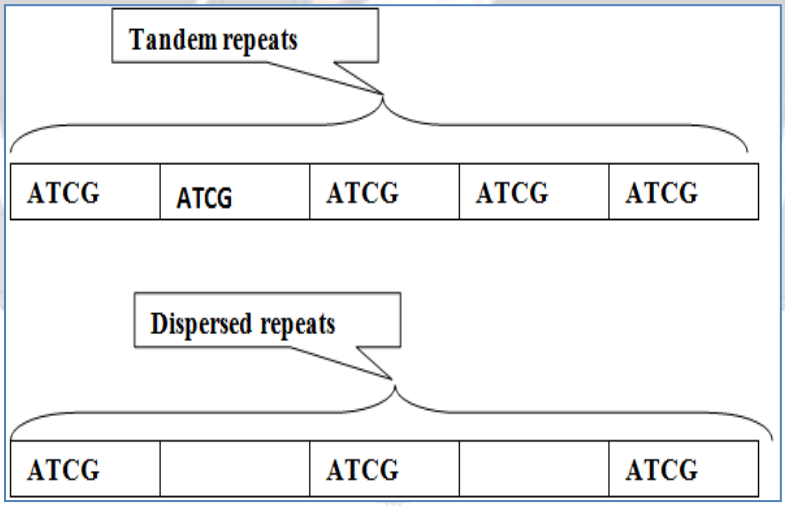


Figure 3: Repeats in DNA

TRs may be further categorized as satellites, minisatellites and microsatellites (MSs), as illustrated in Fig.4 based on the repeated motive (pattern) size. The satellites vary from 100 K base pairs (bps) to 1 Mbps with a pattern of more than 100 bps. The minisatellite duration varies from 1 to 20 kbps and its pattern size ranges from 9 to 80 bps. Microsatellite repeats (MSs) classified as short tandem repeats (STRs) are fewer than 150 bps long with 2 to 6 bps in pattern size.

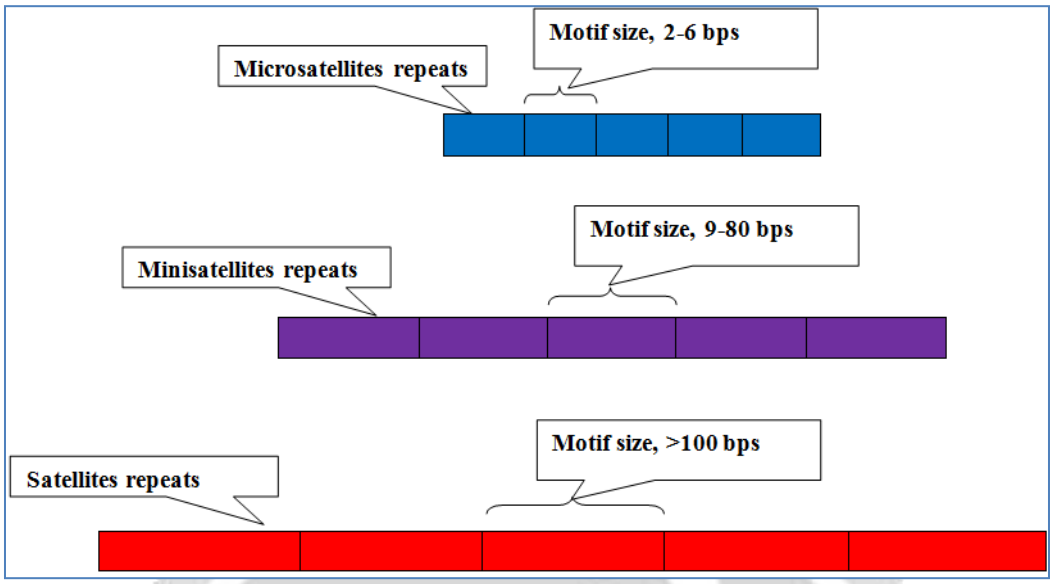


Figure 4: Tandem Repeats in DNA

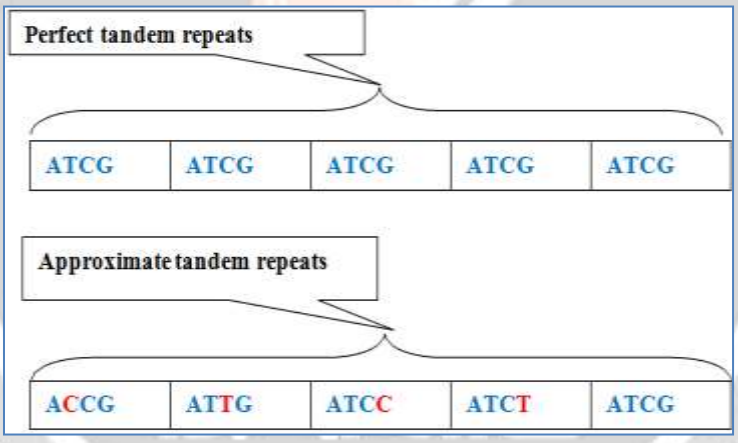


Figure 5: Perfect and Imperfect Repeats in DNA

Perfect tandem repeats (PTRs) are another name for tandem repetitions (ATRs). Repeater patterns are accurately copied in PTRs, whereas repeating patterns are inaccurately copied due to changes in faulty or approximate tandem repeats. PTR. The PTRs and ATRs are shown in Fig. 5.

Among the three types of TR, MSs are important due to their well-established functions and links to cancer and other diseases. In addition to conditions like fragile X syndrome and Huntington's disease, MSs are also responsible for 40 additional neurological disorders including Frederick's ataxia and spinocelebellar ataxia type 31. There are several instances where altering the protein products is caused by increasing or deleting repeating patterns in the tiny repeats of tandems (microsatellites). The genes are discovered to be controlled by changes in the frequency of tiny tandem repeats (microsatellites) in non-coding regions. A variety of genetic disorders and cancers are caused by microsatellite mutations near particular genes. Microsatellites are more susceptible to mutation than other parts of the genome, therefore they are often studied in order to get a better understanding of evolution. Microsatellites are useful in a wide range of fields, including forensics, DNA fingerprinting, demography research, linkage analysis, evolutionary studies, and the behaviour of living creatures. The social behaviour of male voles is influenced by random changes in the microsatellite DNA length near the vasopressin receptor gene. A larger microsatellite area was achieved via increased bonding and care.

8. CONCLUSION

Developing accurate and efficient genomic annotation methods is the primary goal of this research. It's critical for the newly discovered genomic sequence that computational methods be model-independent. The sequence that helps characterize the structural and functional characteristics of DNA sequences is the first step following genome annotation. Genome annotation necessitates the identification of coding sequences (CDS). Identification of exonic eukaryotic regions is a prerequisite for determining the CDS. It's important to note that exonic regions correspond to protein-coding regions. determinants of protein shape include amino acid sequences in exonic regions. Computational genome annotation methods based on signal processing may provide results that are independent of any particular model. A certain region of the genome is defined by the frequency of certain events. By using signal processing-based CDS methods, we can detect the existence of the period-3 component in the exonic region. Nevertheless, the problem is complicated when there is a weak period-3 component in short exons or short exons that are separated by short introns. This problem was addressed in this research by combining the results of PCA mappings. Exon boundaries, also known as splice sites, must be identified in order to accurately determine CDS. The technique for identifying short exons was improved to show where the receiver splices start matching the exonic regions' start.

9. REFERENCES

1) T. M. Inbamalar, R. Sivakumar, "Improved Algorithm for Analysis of DNA Sequences Using Multiresolution Transformation", *The Scientific World Journal*, vol. 2015, Article ID 786497, 9 pages, 2015
2) Susana Vinga, Information theory applications for biological sequence analysis, *Briefings in Bioinformatics*, Volume 15, Issue 3, May 2014, Pages 376–389
3) Mai S. Mabrouk, Different Genomic Signal Processing Methods For Eukaryotic Gene Prediction: A Systematic Review. Biomedical Engineering: Applications, Basis and Communications Vol. 29, No. 01, 1730001 (2017)
4) Mgr. Ing. Karel Sedlář, (2018) Methods For Comparative Analysis Of Metagenomic Data, Brno University Of Technology
5) Po-Yen L. Wu, Advancing Precision Medicine Through Integrative Bioinformatics Approaches For Robust Biological Knowledge Discovery, Georgia Institute of Technology May 2017
6) Meiser, A., Otte, J., Schmitt, I. *et al.* Sequencing genomes from mixed DNA samples - evaluating the metagenome skimming approach in lichenized fungi. *Sci Rep* **7,** 14881 (2017).
7) Choong M. K. and Yan H., (2008), "multi-scale parametric spectral analysis for exon detection in DNA sequences based on forward-backward linear prediction and singular value decomposition of the double-base curves", *Bioinformation*, vol. 2, no. 7, pp. 273– 278.
8) Jiang R. and Yan H., (2008) "Studies of spectral properties of short genes using the wavelet subspace Hilbert-Huang transform (WSHHT)," *Physica A*, vol. 387, no. 16-17, pp. 4223–4247.