Design of 2D Genome Visualization Tool for DNA Sequence Analysis

Ankita Tripathi
Student, M.Tech
CSE, VNSIT Bhopal

Santosh Mishra
Computer Science and Engineering,
VNSIT Bhopal

ABSTRACT
In this paper sequencing of genomic data requires methods to allow this data to be visualized and analyzed. With the emergence of genomic signal processing, graphical representation techniques play a key role in applying digital signal processing techniques like Fourier transforms, and more recently, wavelet transform for visualizing DNA sequence.[2] The choice of the graphical representation technique for a DNA sequence affects how well its biological properties can be reflected in the graphical domain for visualization and analysis of the characteristics of special regions of interest within the DNA sequence. This paper presents a summary of various DNA graphical representation methods and their applications in envisaging and analyzing long DNA sequences. [1] [2]

Keywords
Genomes, Nucleotides, DNA, sequence, Biology

1. INTRODUCTION
In this paper rising field of bioinformatics guarantees to lead to advances in understanding basic biological processes and, in turn, advances in the diagnosis, DNA sequencing and issue functioning. Bioinformatics has transformed the discipline of biology from a strictly lab-based science to Associate in nursing knowledge science likewise. Increasingly, biological studies begin with a scientist conducting giant numbers of data and web-site searches to formulate specific hypotheses or to vogue large-scale experiments.[2] The implications behind this change, for both science and medication, are staggering. Equally exciting is the potential for uncovering evolutionary relationships and patterns between totally different styles of life. With the aid of nucleotide and macromolecule sequences, it should be potential to realize the ancestral ties between totally different organisms.[3] Thus far, experience has educated United States that closely connected organisms have similar sequences and that a lot of distantly connected organisms have a lot of dissimilar sequences. Proteins that show significant sequence conservation, indicating a clear evolutionary relationship, are a foresaid to be from the same macromolecule family. By studying macromolecule folds (distinct macromolecule building blocks) and families, scientists are in a position to reconstruct the biological process relationship between Two species and to estimate the time of divergence between Two organisms since they last shared a standard relative. [3]

1.1 Bioinformatics
In Bioinformatics is the field of science within which biology, computer science and knowledge technology merges to create one discipline. Biology in the Two1st century is being remodeled from a strictly lab-based science to associate science similarly. The ultimate goal of the sector is to alter the invention of latest biological insights similarly on produce a worldwide perspective from those unifying principles in biology is discerned. At the beginning of the genomic revolution a bioinformatics concern was the creation and maintenance of a database to store biological info. such as nucleotide and aminoalkanoic acid sequences.[2] Development of this type of information concerned not solely style problems however the event of complicated interfaces wherever by researchers might each access existing information similarly as submit new or revised information.[3]

1.2 Genes
A gene is a small piece of the genome. It is the genetic equivalent of the atom. As an atom is the fundamental unit of matter, a gene is the fundamental of heredity. [4]

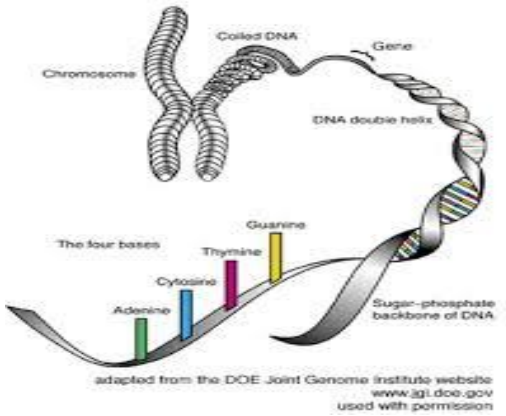

Figure 1.1 Gene in DNA

In Genes area unit found on chromosomes and are created of desoxyribonucleic acid. Different genes confirm the completely different characteristics, or traits, of an organism. In the simplest terms (which are literally too simple in several cases), one gene may confirm the color of a bird's feathers, while another factor would confirm the form of its beak.[5]
The number of genes within the ordination varies from species to species. More advanced organisms tend to have additional genes. Bacteria have many hundred to many thousand genes. Estimates of the number of human genes, by contrast, range from twenty five Thousand to Thirty Thousand.[3]

1.3 Genetics
Genetics is the study of genes, and tries to explain what they are and therefore the manner they work. Genes are but living organisms inherit choices from their ancestors; for example, children typically look like their parents as a results of they have transmissible their parents' genes. Genes are created from associate degree extended molecule far-famed as DNA, which is derived and transmissible across generations. DNA is created of simple units that line up in associate degree extremely specific order within this huge molecule. The order

of these units carries genetic information, similar to how the order of letters on a page carries knowledge. The language used by DNA is termed the ordination, which lets organisms browse the knowledge among the genes. This information is the directions for constructing and operating a living organism. [4]

1.4 Genetic Engineering

Genetic engineering, also known as genetic modification, is the direct human manipulation of an organism's ordering victimisation trendy DNA technology. It involves the introduction of foreign DNA or artificial genes into the organism of interest. The introduction of new DNA doesn't need the employment of classical genetic strategies. The most common style of biotechnology involves the insertion of latest genetic material at an unspecified location within the host ordering.[4] This is accomplished by isolating and repeating the genetic material of interest victimisation molecular biological research strategies to get a DNA sequence containing the specified genetic parts for expression, and then inserting this construct into the host organism. Genetic engineering techniques are applied in numerous fields together with analysis, biotechnology, and medicine. Medicines such as insulin and human endocrine are currently made in microorganism. [3]

1.5 Genome

In modern molecular biology and genetics science, the genome is the entireness of associate degree organism's hereditary info. It is encoded either in DNA for many styles of virus, in RNA. The genome includes each the genes and the non-coding sequences of the DNA/RNA. The genome is distributed on chromosomes, which are created of compressed and entwined DNA. A gene is a phase of body DNA that directs the synthesis of a super molecule. [6]

1.6 DNA

Deoxyribonucleic acid (DNA) could be a supermolecule containing the genetic instructions employed in the event and functioning of all familiar living organisms (with the exception of RNA viruses). The DNA segments carrying this genetic data square measure known as genes. Likewise, other deoxyribonucleic acid sequences have structural functions, or are concerned in control the use of this genetic data. Along with RNA and proteins, DNA is one of the 3 major macromolecules that square measure essential for all familiar kinds of life. [5]

The information in deoxyribonucleic acid is hold on as a code created of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). In Human DNA consists of regarding three billion bases and a lot of than ninety nine % of these bases square measure identical altogether individuals. The order or sequence of these bases determines the knowledge available for building associate degreed maintaining an organism, similar to the way during which letters of the alphabet seem during a bound order to create words and sentences. [2] [3]

An necessary property of deoxyribonucleic acid is that it will replicate, or make copies of itself. Each strand of deoxyribonucleic acid in the helix will function a pattern for duplicating the sequence of bases. This is critical once cells divide as a result of every new cell has to have a definite copy of the deoxyribonucleic acid gift within the recent cell.
DNA consists of long polymers of straightforward units known as nucleotides, with backbones made of sugars and phosphate teams joined by organic compound bonds. These two strands run in opposite directions to every different and are so anti-parallel. Attached to every sugar is one among four varieties of molecules known as nucleobases. [4]

1.7 DNA Structure

The structure of DNA of all species comprises two helical chains each coiled round the same axis. Although each individual repeating unit is very small, DNA polymers can be very large molecules containing millions of nucleotides. For instance, the largest human chromosome, chromosome number 1, is approximately TwoTwo0 million base pairs long. In living organisms DNA will not typically exist as one molecule, but instead as a combine of molecules that are command tightly along. These two long strands entwine like vines, in the shape of a helix. The nucleotide repeats contain each the section of the backbone of the molecule, which holds the chain along, and a nucleobase, which interacts with the different deoxyribonucleic acid strand within the helix. A nucleobase connected to a sugar known as is a glycoside and a base linked to a sugar and one or a lot of phosphate teams is called a ester. Polymers comprising multiple linked nucleotides are known as a polynucleotide. [1] [2] [3]

The backbone of the DNA strand is created from alternating phosphate and sugar residues. The sugar in DNA is Two-deoxyribose, which is a monosaccharide (five-carbon) sugar. The sugars are joined along by phosphate teams that type phosphodiester bonds between the third and fifth carbon atoms of adjacent sugar rings. These asymmetric bonds mean a strand of deoxyribonucleic acid has a direction. In a helix the direction of the nucleotides in one strand is opposite to their direction within the other strand: the strands are parallel. The asymmetric ends of deoxyribonucleic acid strands are known as the (five prime) and (three prime) ends, with the 5 finish having a terminal phosphate cluster and the three finish a terminal radical. One major difference between deoxyribonucleic acid and RNA is the sugar, with the Two-deoxyribose in DNA being replaced by the various monosaccharide sugar carbohydrate in RNA. [6]

The bases lie horizontally between the two spiraling strands. The DNA double helix is stabilized primarily by two forces: hydrogen bonds between nucleotides.

In the aqueous environment of the cell, the conjugated π bonds of nucleotide bases align perpendicular to the axis of the DNA molecule, minimizing their interaction with the solvation shell and therefore, the Gibbs free energy. [3]

The four bases found in DNA are adenine (abbreviated A), cytosine (C), guanine (G) and thymine (T). These four bases are attached to the sugar/phosphate to form the complete nucleotide, as shown for adenosine monophosphate.

1.8 Grooves

Twin helical strands kind the DNA backbone. Another double helix is also found by tracing the spaces, or grooves, between the strands. These voids are adjacent to the base pairs and should give a binding web site. As the strands aren't directly opposite one another, the grooves are unevenly sized. One groove, the major groove, is Two 2Å wide and the different, the minor groove, is 12 Å wide. The narrowness of the minor groove means that the sides of the bases area unit a lot of accessible within the major groove.

As a result, proteins like transcription factors that can bind to specific sequences in double-stranded DNA typically build contacts to the perimeters of the bases exposed within the major groove.

1.8.1 Base pairing

In a DNA double helix, each type of nucleobase on one strand normally interacts with just one type of nucleobase on the other strand. This is called complementary base pairing. Here, purines form hydrogen bonds to pyrimidines, with A bonding only to T, and C bonding only to G. This arrangement of two nucleotides binding together across the double helix is called a base pair. As hydrogen bonds are not covalent, they can be broken and rejoined relatively easily. The two strands of DNA in a double helix can therefore be pulled apart like a zipper, either by a mechanical force or high temperature. [10]

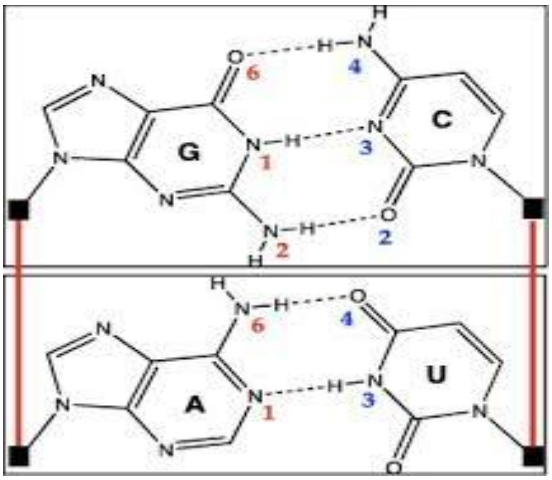

Figure 1.2 Base Pairing

A GC base pair with three hydrogen bonds. Bottom, an AT base pair with two hydrogen bonds. Non-covalent hydrogen bonds between the pairs are shown as dashed lines.

1.8.2 Two super coiling

DNA can be twisted like a rope in a process called DNA super coiling. With DNA in its "relaxed" state, a strand usually circles the axis of the double helix once every 10.4 base pairs, but if the DNA is twisted the strands become more tightly or more loosely wound. If the DNA is twisted in the direction of the helix, this is positive super coiling, and the bases are held more tightly together. If they are twisted in the opposite direction, this is negative super coiling, and the bases come apart more easily. In nature, most DNA has slight negative super coiling that is introduced by chromosomes (C1 and C2).

A deoxyribonucleic acid helix typically will not move with alternative segments of DNA, and in human cells the different bodys even occupy separate areas within the nucleus referred to as chromosome territories. [2]

This physical separation of different chromosomes is very important for the power of deoxyribonucleic acid to operate as a stable repository for data, as one of the few times chromosomes interact is throughout body after they recombine. Chromosomal crossover is once Two deoxyribonucleic acid helices break, swap a section so rejoin. Recombination allows chromosomes to exchange genetic data and produces new combos of genes, which will increase the potency of natural choice and might be vital within the fast evolution of latest proteins. Genetic recombination can additionally be concerned in deoxyribonucleic acid repair, particularly in the cell's response to double-strand breaks. [7] [8]

The most common sort of chromosomal crossover is homologous recombination, where the Two chromosomes concerned share terribly similar sequences. Non-homologous recombination can be damaging to cells, as it can manufacture body translocations and genetic abnormalities. [9]

2. PROPOSED WORK

In This Paper Representations based on two dimensional Cartesian coordinates remain the staple form of graphical methods for their simplicity and intuitive feel. The original plot of a DNA sequence as a random walk on a TwoD grid using the four cardinal directions to represent the four bases was done by Gatesand then rediscovered independently by Nandy and Leong and Morgenthaler. The idea was to read a DNA sequence base by base and plot succeeding points on the graph. According to the Nandy prescription, a point was plotted by moving one step in the negative x- direction if the base was an adenine (A) and in the opposite direction if it was a guanine (G) and a walk of one step in the positive y-direction if the base was a cytosine (C) and in the opposite direction if it was a thymine (T). The Gates method prescribed the bases GTCA and the Leong Morgenthaler method prescribed CTAG reading clockwise starting from the negative x-axis for the walks.

2.1 Proposed Method

The idea behind numerical characterization of a DNA sequence is to devise mathematical descriptors that would capture the essence of the base composition and distribution of the sequence in a quantitative manner which would facilitate sequence identification and comparison of similarities and dissimilarities of sequences.

In the number of different approaches that have been proposed to mathematically characterize and describe the DNA sequences, it is important to compare them critically. We would expect that since all the methods proposed so far have calculated the similarity/dissimilarity indexes for the DNA sequence of exon 1 of the beta globin gene, the trends should be similar although the individual methods may differ in the absolute magnitudes across all methods of each index and could differ in some way in relative ratios.

To eliminate the complexities and other problems associated with previous methods, the graphical representation technique is modified by us. To construct a new graphical representation of DNA sequences in the first quadrant of the Cartesian coordinate plane The unit vectors representing four nucleotides A, G, C, and T are as follows:

(1,0)→A, (√3 /2,1/2)→G, (1/Two, √3 /2)→C,(0,1)→T, that means OA lies on x-axis, the angle between OG and x-axis is 30 degrees, the angle between OC and x-axis is 60 degrees, OT stands up on y-axis, and |OA|=|OG|=|OC|=|OT|=1.

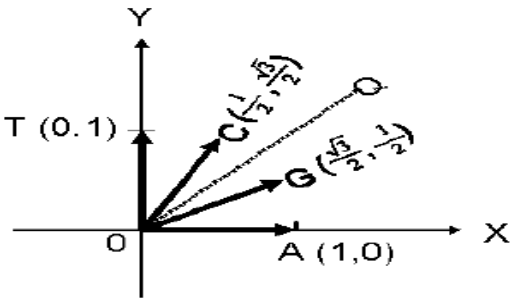

Figure 3 Graph with Unit Vector

we set purines (A and G) under the bisector OQ of the first quadrant and pyrimidines (T and C) above the bisector OQ of the first quadrant. To get the graphical representation of a

DNA strand, we assume that S1 , S 2 , …, Sn stand for a DNA sequence length at n, where Si belongs to {A,T,C,G}. The sequence of the points, P1 , P2 , …, Pn will then be constructed as the vector P i-1 Pi that corresponds to S i , where P 0 is the origin, and |Pi-1 P i | =1. If S i =A, then Pi-1 P i is parallel to x-axis; if S i =T, then Pi-1 P i is parallel to y-axis; if S i =G, then the angle between Pi-1 P i and the ray from Pi-1 parallel to x-axis is 30 degrees; if S i =C then the angle between Pi-1 P i and the ray from P 1 - i parallel with x-axis is 60 degrees.

To get the numerical sequence of the points, P1 , P Two , …, Pn corresponding to S1 , S2 , …, Sn , we introduce a two dimensional array x(i) i=1, 2, …, n, and y(i), i=1, 2, …, n, and Pi =(x(i), y(i)). If S i =A, then Pi = Pi-1 +(1,0); if Si =G, then Pi = Pi-1 +(√3 /2,1/2); if Si =C, then Pi = Pi-1 +(1/2, √3 /2) and if Si =T, then Pi = Pi-1 +(0,1), where i=1, 2, …, n, and P 0 =(0,0).

3. IMPLEMENTED WORK CODONE

This is the simple DNA sequence we take as input:-

CAGGTTACAATTGGAGCCATTTCATCTTCTGACTGAG
GAAATCAGGTCCGACAGCGTAGATGTATACACGCTC
TATTCACAAAATTGGTAACGATTCT

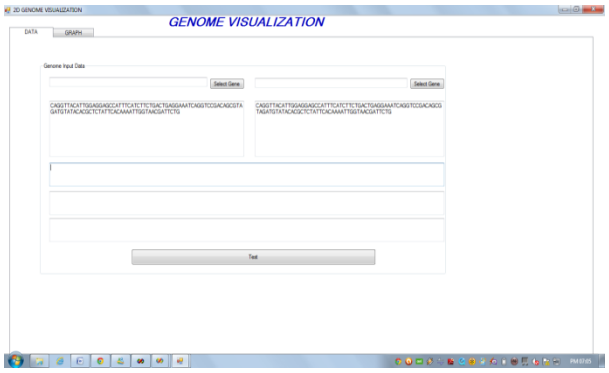

Figure 4 Sequence input

4. CONCLUSIONS AND FUTURE WORK

Graphical representation of DNA sequence may provide a simple way of viewing, sorting and comparing various gene structures. The previous studies of the 2-D graphical representations of DNA sequence used four quadrants and placed four nucleotides along with four axes of coordinate system usually had complex structure. In order to overcome this, we have presented and formulated a new graphical representation of DNA strand by using one quadrant Cartesian coordinates system. This tool is much easy to use than other available tool. Due the complexity of other tools they are not of great use for the users, especially for them who belong to biological background only.

In future DNA analysis tool can be extended to include the Features of performing classification of different DNAs using the alignment of sequences.

5. REFERENCES

[1]. Damian Panasa, Piotr Waz, Dorota Bieli, Ashesh Nandy, Subhash C. Basak "2D–Dynamic Representation of DNA/RNA Sequences as a Characterization Tool of the Zika Virus Genome", MATCH Commun. Math. Comput. Chem. 77 (2016) 321-332.

[2]. D. Bielinska-Waz and P. Waz, "2D-dynamic Representation of DNA Sequences as Graphical Tool in Bioinformatics", AIP Conference Proceedings 1773, 060004 (2016).

[3]. Sai Zou, Lei Wang and Junfeng Wang, "A 2D graphical representation of the sequencesof DNA based on triplets and its applications" , EURASIP Journal on Bioinformatics and Systems Biology 2014.

[4]. Lei Wang,1,2 Hui Peng,1,2 and Jinhua Zheng1, "ADLD: A Novel Graphical Representation of Protein Sequences and Its Application" , Hindawi Publishing Corporation Computational and Mathematical Methods in Medicine Volume 2014, Article ID 959753.

[5]. Chiou-Yi Hor, Chang-Biau Yang, Chia-Hung Chang, Chiou-Ting Tseng and Hung-Hsin Chen, "A Tool Preference Choice Method for RNA Secondary Structure Prediction by SVM with Statistical Tests", Evolutionary Bioinformatics 2013:9 163–184.

[6]. Dorota Bielinska Waz, Timothy Clark , Piotr Waz, Wiesław Nowak, Ashesh Nandy "2D-dynamic representation of DNA sequences", Chemical Physics Letters 442, pp. 140–144, 2011

[7].Jinglu Hu, Yang Chen , "Accurate Reconstruction for DNA Sequencing by Hybridization Based on a Constructive Heuristic", IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) July 2011

[8].Yi-Ming Sun, Wei-Li Liao, Hsien-Da Huang, Baw-Jhiune Liu, Cheng-Wei Chang, Jorng-Tzong Horng, Li-Ching Wu, "A Human DNA Methylation Site Predictor Based on SVM" , Bioinformatic and Bioengineering, IEEE International Symposium, 2009

[9]. Jing Yang, Zhi-xiang Yin, Kai-feng Huang, "The Working Operation Problem on Triple-stranded DNA Structure Model", 2009

[10]. Designation of the two strands of DNA JCBN/NC-IUB Newsletter 1989, Accessed 07 May 2008