

CHAPTER 3

BIOINFORMATICS DATA

AND

APPLICATIONS

Chapter 3. BioInformatics Data and Applications

3.1. Introduction

“BioInformatics is a science that involves collecting, manipulating, analyzing and transmitting huge quantities of biological data, and uses computer whenever appropriate²³.”

Thus, BioInformatics is currently one of the several thrust areas of research in Computer Science and Information Technology. BioInformatics is an interdisciplinary research area that applies computer science and information technology to solve biological problems.

BioInformatics further can be divided into several specific sub-areas like:

- Computational Biology
- Systems Biology
- Structural Biology
- Functional Biology

The research work undertaken emphasizes on Computational Biology.

²³ Bryon Bergeron, “BioInformatics Computing”, Pearson Education, Inc. 2003

3.1.1. Several terminologies associated with Bioinformatics, which are referred in this research study:

- **Genomic Sequences:** The entire Information about genome sequences is the information pertaining to details about DNA, RNA and protein sequences of living organisms.
- **Genomics²⁴:** The study of the complete genome of an organism is known as Genomics. The study of a particular gene is known as Genetics
- **Genome²⁵:** The entire DNA sequence that codes for a living being is called its genome. The entire set of genetic instructions found in a cell is known as a Genome. In humans, the genome comprises of 23 pairs of chromosomes. The complete sets of chromosomes contain nearly 3.1 billion basepairs.
- **Chromosomes²⁶:** A chromosome is an organized bundle of DNA found in the cell's nucleus. The numbers of chromosomes vary from one organism to another. 23 pairs of chromosomes exist in Humans. Out of 23 pairs of chromosomes, there are 22 pairs of numbered chromosomes known as autosomes, and single pair of sex chromosomes called X and Y chromosomes. One chromosome is transferred from each parent to each pair, so that an offspring gets half of its chromosomes from the mother and the other half from the father.

²⁴ <http://www.genome.gov/glossary/index.cfm?id=532>

²⁵ <http://www.genome.gov/glossary/index.cfm?id=90>

²⁶ <http://www.genome.gov/glossary/index.cfm?id=33>

- Gene²⁷: The gene is the blueprint of inheritance. It is basic physical unit that is passed from parents to offspring which contain the information necessary to specify traits. Genes are arranged sequentially, one following another, on structures called chromosomes. A chromosome contains a long, single DNA molecule, only a part of which match to a single gene. Approximately 20,000 genes are arranged on the human chromosomes which specify various traits.
- DNA²⁸: The De-oxy RiboNucleic Acid (DNA) is the chemical name for the molecule that carries genetic information in all living beings. It is a double-helix structure of nucleotides, held together by chemical bonds between the nucleotides. Each strand comprises of a backbone made of alternating sugar (deoxyribose) and phosphate groups. Each sugar molecule is attached to one of four bases Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). The two strands have bonding between the bases. Adenine base bonds with thymine and cytosine bonds with guanine. These sequences of the bases help in assembling protein and RNA molecules. The DNA provides the blue-print of living being. It is known as the hereditary material carried from one generation of an organism to another generation.
- RNA²⁹: RiboNucleic Acid (RNA), like DNA, are nucleic acids found in all cells. RNA are involved in regulatory mechanisms of cells such as catalyzing biological reactions, controlling gene

²⁷ <http://www.genome.gov/glossary/index.cfm?id=70>

²⁸ <http://www.genome.gov/glossary/index.cfm?id=48>

²⁹ <http://ghr.nlm.nih.gov/glossary=rna>

expression, or communicating messages during protein synthesis. RNA is single stranded, unlike DNA which is double-helix. An RNA strand has a backbone made of alternating sugar (ribose) and phosphate groups. Each sugar has one of the four nucleotide bases, Adenine (A), Cytosine (C), Guanine (G) or Uracil (U) attached to it. Different types of RNA exist in the cell such as messenger RNA (mRNA), ribosomal RNA (rRNA), and transfer RNA (tRNA). RNA play an important role in protein synthesis and other cell activities, by transmitting genetic information from DNA to protein, that is produced by the cell.

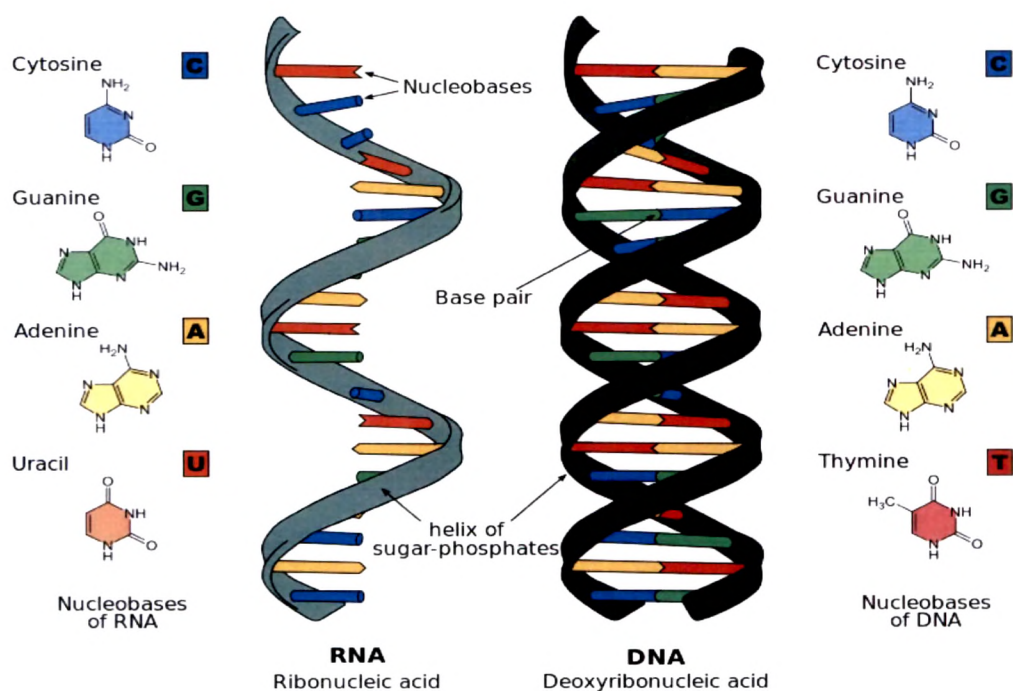


Figure 7. Difference between the structure of RNA and DNA
Source:http://upload.wikimedia.org/wikipedia/commons/3/37/Difference_DNA_RNA-EN.svg

The structural difference between RNA and DNA can be represented as in Figure 7³⁰

- Protein³¹: Proteins are the building blocks of cells. Proteins are molecules made up of several chains of amino acids in a definite order. The sequence of nucleotides in the gene that codes for the protein determines this order. Proteins are essential for the proper functioning of the body. Each protein has unique function that is necessary for the structure, functioning, and regulation of the body's cells, tissues, and organs. They are the basis for body structures like skin or hair and for substances like enzymes, hormones, cytokines or antibodies. There are 20 amino-acids which are used to build protein sequences.
- DNA Sequencing: The process of determining the identity and order of bases in a molecule of DNA. The process of Genome sequencing is done to identify the complete DNA sequence of any organism. The sequenced stream of DNA consists of all the gene information of an organism whose sequencing has been performed. There are different methods of sequencing like Sanger sequencing, Pyrosequencing and Whole Genome Shot-Gun Sequencing etc.
- Read³²: The physical process of DNA sequencing produces a large number of random substrings of the sequence known as

³⁰http://upload.wikimedia.org/wikipedia/commons/3/37/Difference_DNA_RNA-EN.svg

³¹ <http://ghr.nlm.nih.gov/glossary=protein>

³² Highly Scalable Genome Assembly on Campus Grids, Christopher Moretti, Michael Olson, Scott Emrich, and Douglas Thain, MTAGS '09 November 16th, 2009, Portland, Oregon, USA, Copyright 2009 ACM 978-1-60558-714-1/09/11

reads. This is due to current non-availability of sequencing process, which is capable of producing an organism's entire string of millions or billions of bases. Depending on the exact process, these reads can vary in length from 25 to 1000 bases each. Individually, these reads have limited scientific value, but this is the basic raw data acquired from biological sample used for determining DNA sequence. Reads are assembled to form contigs that further are used for genome annotation.

- Contig³³: Derived from the word "contiguous" a Contig is a series of overlapping DNA sequences used to build a physical map that reconstructs the original DNA sequence for a complete or a part of the chromosome. A Contig can also refer to one of the DNA sequences used in building such a map.

3.1.2. Bioinformatics Data

Bioinformatics data deals to a great extent with:

- Data related to Genomics i.e. information about DNA, RNA and protein sequences of living organisms
- Genetic Data i.e. data pertaining to information in one particular gene
- Data generated from HealthCare Systems.

DNA data being the focus for the research studies, the emphasis is put on explaining the DNA data and its components.

³³ <http://www.genome.gov/glossary/index.cfm?id=39>

DNA maintains an organism's hereditary and functional information. The other biological data which usually is used for biological studies is RNA and proteins. The DNA, RNA and proteins, are linear chains of macromolecules, which in turn are composed of smaller molecules. The macromolecules are assembled from a fixed set of well-understood chemicals. Nucleotides form nucleic acids and nucleic acids later on form the DNA. Nucleotides comprise of smaller molecules such as sugar, phosphate and nitrogenous bases (See Figure 8)³⁴.

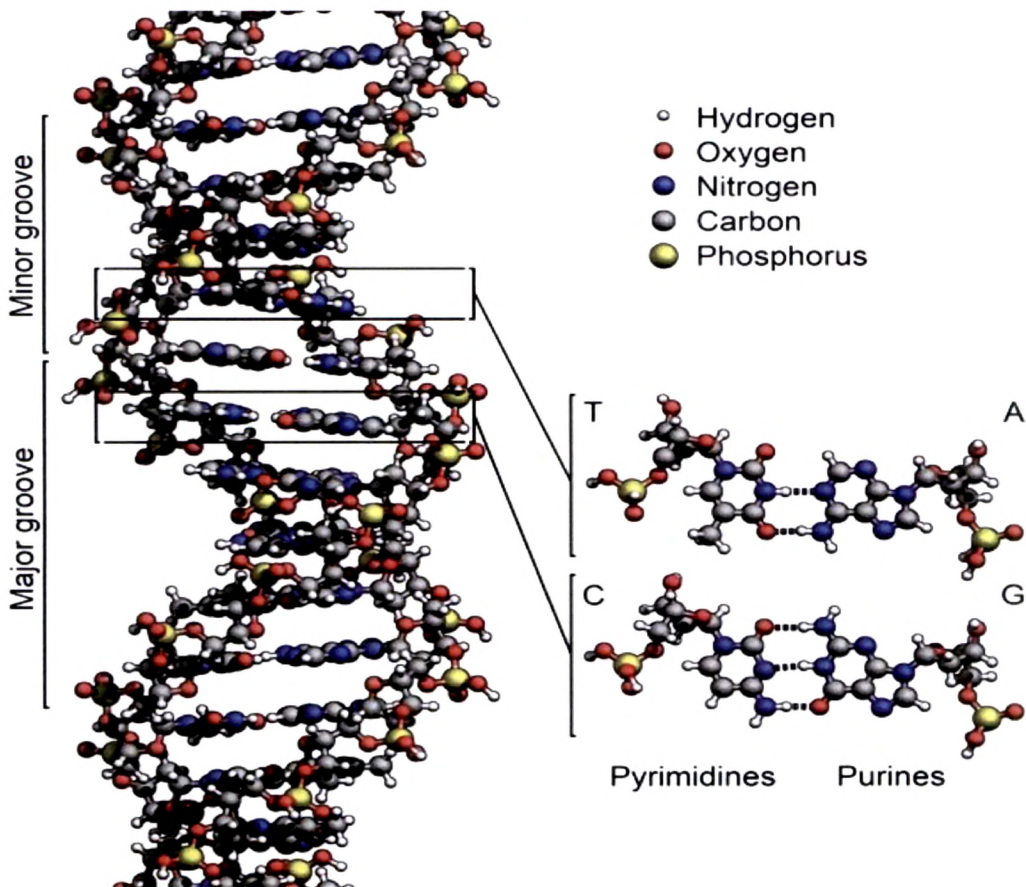
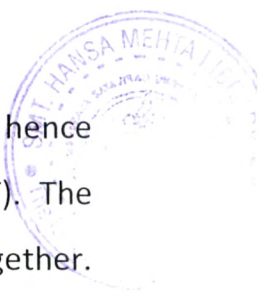


Figure 8. DNA, as a chain of smaller molecules or nucleotide bases
Source: <http://upload.wikimedia.org>

³⁴ <http://upload.wikimedia.org>



The four nitrogenous bases that are used to form nucleotides and hence DNA are Adenine(A), Cytosine(C), Guanine (G) and Thymine (T). The nitrogen bases are found in pairs, with A & T and G & C paired together. The sugar used in DNA is de-oxy-ribose. The sequence of the nitrogenous bases can be arranged in an infinite ways, to form a "double helix" structure. Every organisms DNA consists of the same four nitrogen bases. The sequence and number of bases is what creates diversity. 99% of DNA sequence in all human beings is identical. DNA does not actually make the organism, it only makes proteins. The DNA is transcribed into mRNA and mRNA is translated into protein, and the protein then forms the organism. The change in DNA sequence, changes the way in which the protein is formed. This leads to either a different protein, or an inactive protein.

These macromolecules of DNA, RNA or proteins, which are linear chains of well-defined components, can be represented as sequences of symbols or strings for computational purpose. These sequences of symbols can be compared to find similarities, which suggest that the molecules are related by form, function or phylogeny.

The diagrammatic representation i.e. 3-D view of sequence of DNA is as shown in Figure 9³⁵, which represents DNA as a group of molecules, bonded with each other to form double-helix structure. This 3-Dimensional structure can also be represented in 1-Dimensional linear form which looks as shown in Figure 10.

³⁵ www.rpi.edu

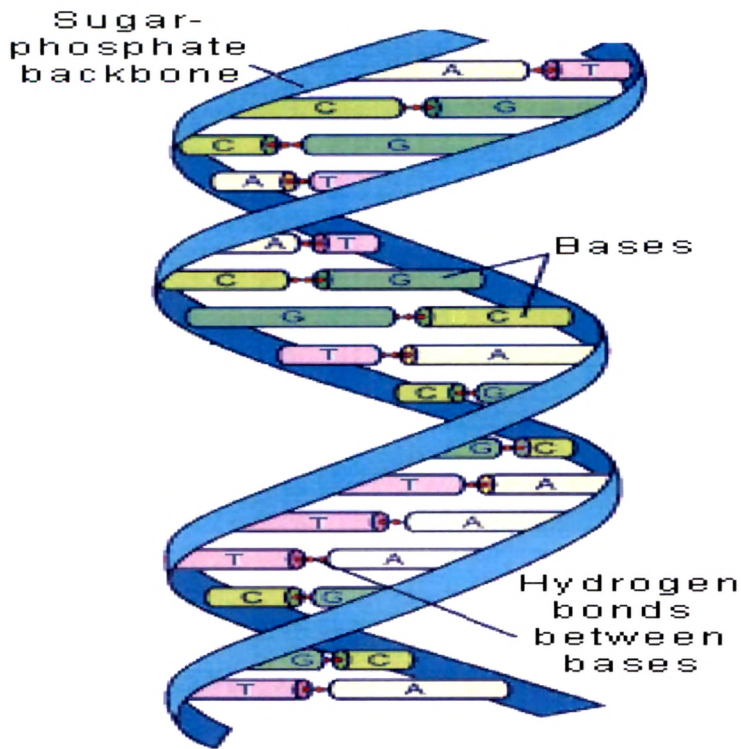


Figure 9. Three Dimensional view of Nucleotide Sequence

Source: www.rpi.edu

ACGTATATATTAGTTTCGCGGGAGGACGTCACACTAAAACGATTAAGACGTA
GCGAACA.....

Figure 10. One Dimensional view of Nucleotide Sequence

Representing DNA sequence in linear form becomes easier to store in computer in form of files or databases and also becomes easier to process it.

- This type of linear representation can be obtained by considering the alphabets that represent the nucleotides that form the DNA.

3.1.3. Source of Bioinformatics Data

Various Biological experiments are conducted in different laboratories and Research Centres all over the world. For global knowledge and discussion of these new findings, the central repository of data is maintained at National Centre for Biotechnology Information (NCBI). The sequencing information of various organisms can be acquired from NCBI (<http://ncbi.nlm.nih.gov>). NCBI's nucleotide database is an arsenal of sequences from several sources like GenBank, RefSeq, TPA and PDB.

The research work undertaken deals with Genomic DNA data. It primarily deals with the DNA sequencing data like reads and contigs, as well as other genomic data like chromosomes. The data used in the research studies is primarily downloaded from <http://www.ncbi.nlm.nih>. For applying algorithms of identifying duplicate reads, the DNA sequencing data is downloaded from Sequence Read Archive (SRA) data of NCBI.

Sequence Read Archive (SRA)³⁶ is the National Institutes of Health's (NIH) primary repository or archive of high-throughput DNA sequencing data. The SRA repository is part of the international partnership of archives (INSDC) at the NCBI, the European Bioinformatics Institute (EBI) and the DNA Database of Japan (DDBJ). Data submitted to any of these three organizations is accessible to all of them. The Sequence Read Archive (SRA) stores raw sequencing data from "next-generation" sequencing technologies including Roche's 454, IonTorrent, Illumina, SOLiD, Helicos and Complete Genomics. The data used for study primarily consist of sequences generated through Roche's 454 Genome Sequencers.

³⁶ <http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi>

3.1.4. Challenges in Bioinformatics

The field of Bioinformatics comprises of various challenges to Computer Scientists, few of which are:

- Creating and maintaining the databases for the exponentially growing data from biological researches.
- Generating pattern-matching techniques for sequence analysis.
- Generating 2-D and 3-D structures of nucleotides and polypeptides to study their functional behaviour by Modelling the linear sequences.
- Study the complex behaviour of organisms and effect of mutations through Data visualization.

3.1.5. Bioinformatics Application

Various types of Computational applications in the field of Bioinformatics, deal with:

- DNA sequencing
- DNA sequence Base Calling
- De Novo or Reference based Assembling the Reads to form Contigs and Scaffolds
- Sequence Similarity
- Sequence Alignment
- Genome Assembly
- Identifying Repeat Regions
- Identifying Coding Regions (Cds)

- Identifying Gene Expression
- Detecting the Protein Translation
- Protein Structure Prediction
- Protein Structure Analysis
- Approximate Search
- Data Reduction/Compression
- Multiple Sequence Alignment
- Motif discovery from multiple sequences

3.1.6. DNA Sequencing

DNA Sequencing is the process of determining the identity and order of bases in a molecule of DNA.

The popular DNA sequencing methods are:

- Maxam-Gilbert sequencing
- Chain-termination methods / Sanger sequencing
- Pyrosequencing
- Shotgun sequencing
- Sequencing by Synthesis
- Massively Parallel Signature Sequencing (MPSS)
- Single Molecule Fluorescent sequencing
- Single Molecule Real Time (SMRT) sequencing
- Nanopore DNA sequencing

The most popular conventional method of sequencing is Sanger sequencing, which the process, of which has be explained briefly. The pyrosequencing method is another most widely used method these days.

Sanger Sequencing³⁷ :

A common method for sequencing DNA involves: purifying DNA from a sample, making a copy of that DNA *in vitro*, separating the new DNA molecules by their size, and identifying the base at the end of each DNA molecule by measuring the intensity of the fluorescent signal. This entire process is commonly known as Sanger sequencing after Fred Sanger, the biochemist who developed the method for using dideoxy ribonucleotide triphosphates (ddNTPs) to create DNA molecules of random sizes. Automated DNA sequencing instruments use capillary electrophoresis to separate the differently sized molecules of DNA. Capillary electrophoresis separates DNA molecules in a small capillary tube instead of in an agarose gel. Automated DNA sequencing instruments also contain a laser that excites the fluorescent dye attached to each DNA base, instruments that capture and measure the intensity of fluorescence, and software for processing the fluorescent signal and creating a chromatogram.

A key point to note is that the DNA bases that are measured are produced by

Synthesizing new DNA in vitro, and might contain differences from the original sequence in the sample due to errors during DNA synthesis.

³⁷ Analyzing DNA Sequences and DNA Barcoding - ©Northwest Association for Biomedical Research 262
(F:\Research\Downloads_Jan2013\AAMotherPapers\AdvL9.pdf)

Scientists use chromatogram-viewing programs like Finch TV to view and analyze their chromatograms and associated DNA sequence data. They use sequence assembly programs to reconstruct a model of the original sequence.

Pyrosequencing:

The method of pyrosequencing is the widely used process of sequencing which gives more accurate results in optimal time. Appendix-C presents the detailed explanation of the process of pyrosequencing. The data generated through the process of pyrosequencing is used for this research work.

The research study included the working of pyrosequencing process which was conducted on a super computer attached with Roche's GS-FLX sequencer, at Anand Agriculture University.

The study of DNA sequencing process using pyrosequencing process is presented as a paper titled ***"PYROSEQUENCING" Sequencing Technique In Roche's GS-FLX System*** at 1st IFIP International Conference on Bioinformatics (IFIP-2010). The details are given in List of Publications and Presentations.

Popular Genome Sequencing Equipments available are:

The next generation of sequencing platforms include

- Roche 454 GS System³⁸ (GS20, GSFLX and GS Titanium Cluster implementing pyrosequencing)

³⁸ www.454.com

- Illumina's Solexa Genome Analyzer³⁹ (Uses sequencing by synthesis)
- Applied Biosystems' SOLiD System⁴⁰ (3730xl implementing Sanger sequencing⁴¹)
- Helicos BioSciences' Heliscope⁴² (Uses Single Molecule Fluorescent sequencing)
- Pacific Biosciences' PacBio RSII⁴³ (PacBio RSII uses SMRT sequencing method)
- Complete Genomics' Polonator⁴⁴ (G.007 sequencer uses sequencing by ligation method)
- The research study primarily uses DNA sequencing data generated through Roche's 454 **Genome Sequencer using pyrosequencing techniques.**

3.1.7. Data and Application into consideration

In this research work, the data generated using pyrosequencing process through Roche sequencers is taken into consideration.

The Reads and Contigs generated through pyrosequencing process is primary data dealt with in this study.

The applications taken into consideration are identifying duplicate reads, generated during pyrosequencing process.

³⁹ www.illumina.com

⁴⁰ www.appliedbiosystems.com

⁴¹ A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes - Susanne M. D. Goldberg, Justin Johnson, Dana Busam, Tamara Feldblyum *et. al.*

⁴² www.helicosbio.com

⁴³ www.pacificbiosciences.com

⁴⁴ www.polonator.org

In addition to this, given any DNA sequence the Short Tandem Repeat Regions are detected from these DNA sequences.

Both the above issues are tackled by applying Wavelet Transforms for data reduction of these sequences. Thus, data reduction or compression of DNA sequences is another issue taken into consideration.

Examples for Data Set for Algorithm 1 on Data Reduction:

The data samples for Reads, Contigs and Chromosomes, which are used for data reduction in Algorithm 1 are presented in Appendix B.

The details of algorithms and various test results are presented Chapter 5.

Examples of DataSet for Algorithm 2 for Recognizing Identical Reads:

The examples of Identical Reads from SRR0065619:

Read No. 25

gnl|SRA|SRR065619.22.5 GM0L9AE07IGRU2.5 length=124
CGTTACTATGGTGCACCGACTCCTGTAAATCAGCTAGCCGGGGTATCTGCTCC
AGAAGCGCGCCTTTTAGTCATTACTCCTTATGACAAATCCTGAGACTGCCAAG
GCACACAGGGGATAGGNN

Read No. 707

gnl|SRA|SRR065619.520.5 GM0L9AE07IM08Y.5 length=123
CGTTACTATGGTGCACCGACTCCTGTAAATCAGCTAGCCGGGGTATCTGCTCC
AGAAGCGCGCCTTTTAGTCATTACTCCTTATGACAAATCCTGAGACTGCCAAG
GCACACAGGGGATAGGN

Read No. 1334

gnl|SRA|SRR065619.965.5 GM0L9AE07HYHRE.5 length=123
CGTTACTATGGTGCACCGACTCCTGTAAATCAGCTAGCCGGGGTATCTGCTCC
AGAAGCGCGCCTTTTAGTCATTACTCCTTATGACAAATCCTGAGACTGCCAAG
GCACACAGGGGATAGGN

If one takes any other Read sequence from the file containing Reads for SRR0065619, it will never be similar to the above mentioned three sequences. Thus, these three sequences are identical to each other and so only one copy can be preserved or stored for analysis purpose i.e. the Read No. 25, the other sequences i.e. Read No. 707 and 1334 need not be preserved. Only the Reference to the original sequence can be maintained. More details on identical reads in various SRA files and there statistical details are discussed in Chapter 5.

Examples of DataSet for Algorithm 3:

The Repeat Sequences in DNA data may be described appropriately through the following diagram. The research study emphasizes on one aspect of Repeats and those are Tandem Repeats, primarily the Microsatellites.

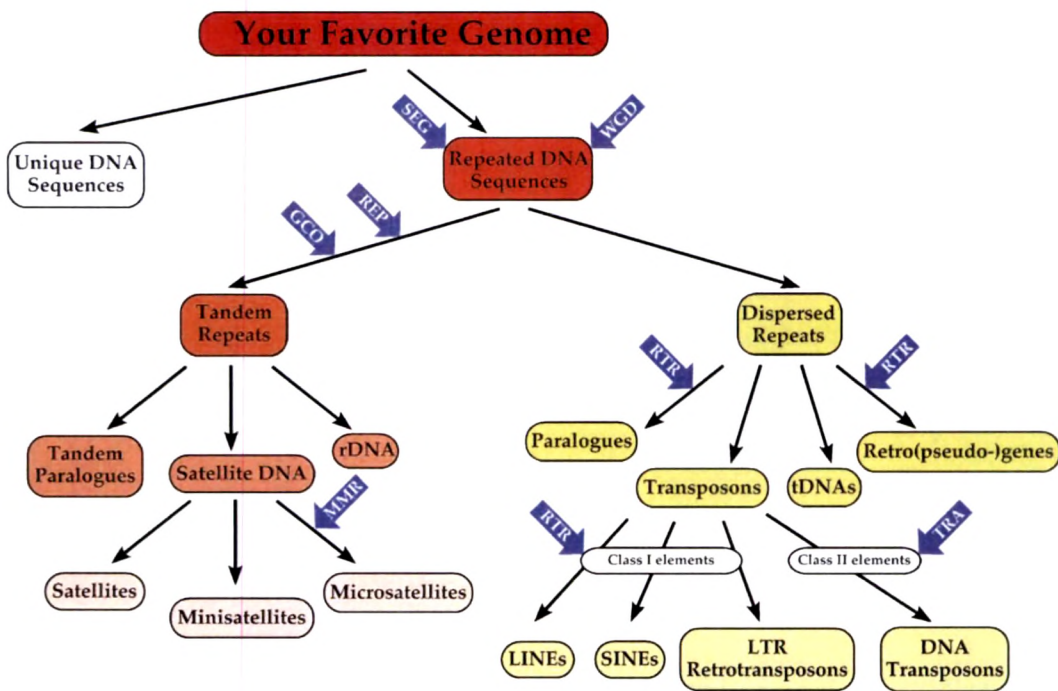


Figure 11. Repeated DNA Sequences in Eukaryotic Genomes

(Source : Comparative Genomics and Molecular Dynamics of DNA Repeats in Eukaryotes, Guy-Franck Richard, Alix Kerrest, and Bernard Dujon, MICROBIOLOGY AND MOLECULAR BIOLOGY REVIEWS, Dec. 2008, p. 686–727 Vol. 72, No. 4 1092-2172/08/\$08.00_0 doi:10.1128/MMBR.00011-08 Copyright © 2008, American Society for Microbiology.)

Repeated DNA sequences in eukaryotic genomes and mechanisms of evolution consists of the two main types of repeated elements (the tandem repeats and the dispersed repeats) are shown in Figure 11, along with subtypes. The arrows in Blue colour point to molecular

mechanisms that are involved in propagation and evolution of repeated sequences such as

REP: replication slippage

GCO: gene conversion

WGD: whole-genome duplication

SEG: segmental duplications

RTR: reverse transcription

TRA: transposition.

The datasets for Short Tandem Repeats are presented as follows:

```
AATTAACCGTTTTGGTCCTACCAACAAT[ACACACACACACACACACACACACACACAC]GTGCAG  
CAAGTTCAGATAATTCAACATTATATGCAGCCATAATATCAAATTCTGAATCTTTAATGCTGGTCAGA  
GGATTTTGAGGAGCCCCGCCCAATTTGGAGAGGGAA[GTGTGTGT]CGC
```

In the DNA sequences stated above, the nucleotide bases **AC** are repeated 16 times (shown in brackets), similarly the nucleotide bases **GT** are also repeated 4 times.

The algorithm using Wavelet Transforms identifies whether the repeat sequence exists or not and that what kind of polymer it is. The details of algorithm alongwith the examples and results are is discussed in Chapter 5.