**DNA SEQUENCING AND ITS APPLICATIONS**

by

Balasubramanian M (111508012)

B.Tech. (IT) (Semester-III)

June – Dec 2016

A report submitted in conformity with the requirements for T2 Test  
for the subject Science of Living Systems (SLS)

Department of Applied Science

Faculty: Dr. Anup Kale

DNA SEQUENCING AND ITS APPLICATIONS

Balasubramanian M

B.Tech IT

Science of Living Systems (SLS)

Semester-III (June-Dec 2016)

**Abstract**

DNA microarray technologies are used in a variety of biological disciplines. The diversity of platforms and analytical methods employed has raised concerns over the reliability, reproducibility and correlation of data produced across the different approaches. Initial investigations (years 2000–2003) found discrepancies in the gene expression measures produced by different microarray technologies. Increasing knowledge and control of the factors that result in poor correlation among the technologies has led to much higher levels of correlation among more recent publications (years 2004 to present). Here, we review the studies examining the correlation among microarray technologies. We find that with improvements in the technology (optimization and standardization of methods, including data analysis) and annotation, analysis across platforms yields highly correlated and reproducible results.

**Keywords : DNA, gene expression**

**Signature of Faculty: Dr. Anup Kale**

**Date: 09/11/2016**

**Table of Contents**

**CHAPTER NO. TITLE PAGE NO.**

**ABSTRACT iii**

**NOMENCLATURE v**

**CONTENTS iv**

**REFERENCES x**

**1. Introduction .......................................................................................... 6**

**1.1 Brief statement of problem ..................................................... 6**

**1.2 Importance ................................................................................ 6**

**1.3 related literature........................................................................ 6**

**1.4 Scope of the project ................................................................. 6**

**2. Approach used ...................................................................................... 7**

**2.1 Computational approach.......................................................... 7**

**2.2 Theory ........................................................................................ 8**

**3. Result …................................................................................................. 10**

**4. Conclusion …………………………………………………………… 10**

**5. Ethical Issues …...................................................................................... 10**

**List of Symbols, Abbreviations and Nomenclature**

1. DNA – Deoxyribo Nucleic Acid
2. NCBI – National Center for Biotechnology Information.
3. .txt – text format
4. .fasta – A text format used to represent nucleotide or peptide sequences
5. A – Adenine
6. G – Guanine
7. T – Thymine
8. C – Cytosine
9. VNTR – Variable Number of Tandom Repeat
10. PCR - Polymerase Chain Reaction

**DNA SEQUENCING**

**1. Introduction**

**1.1 Brief statement of the problem**

To find the actual suspect involved in a crime. So the actual problem is finding the DNA sequence of all the suspects and then finding the DNA sequence of the hair sample or blood sample left over by the victim and if atleast 5 sequences match then we can find the culprit.

**1.2 Importance**

So if we can write a code to find the DNA sequence of any sample and then if atleast 5 sequences match then we can find the culprit. This can actually be used in real life to find the crime victim i.e. it can be used in forensics. More applications of DNA sequencing include detecting genes which link to various genetic disorders and can also be used to study genomes and the proteins they encode.

**1.3 Related literature**

DNA sequencing provides the means to know the how nucleotide bases are arranged in a piece of DNA. The method was pivotal to the international Human Genome Project. Costing over US$3 billion and taking 13 years to complete, this project provided the first complete Human DNA sequence in 2003. This data has provided for the first time a tool to map out the genetic mutations that underlie specific genetic diseases. It has also opened up a path to more personalised medicine, enabling scientists to examine the extent to which a patient's response to a drug is determined by their genetic profile. The genetic profile of a patient's tumour, for example, can now be used to work out what is the most effective treatment for an individual. It is also hoped that in the future the information from the human DNA sequence will provide a means to work out a person's predisposition to certain diseases, such as heart disease, cancer and type II diabetes, which could pave the way to better preventative care.

**1.4 Scope of the project**

The main methods of DNA sequencing include MAXAM – GILBERT sequencing and CHAIN TERMINATION METHOD. APPLICATIONS OF DNA SEQUENCING and a code which finds the DNA sequence of a given DNA and also addresses the ethical issues.

**2. Approach Used**

**2.1 Computational approach**

You can find our code on the link : https://github.com/mbala2810/bio\_project

The main 3 files are :

**1. Generator.cpp :**

So here we have written a code which basically generates a random DNA sequence. It uses rand() function which generates random numbers and these numbers are then mod by 4 i.e. the values obtained will be 0, 1, 2 and 3. So if case is 0 then G is inputted in the string, if case is 1 then C is inputted in the string and similarly for 2 and 3 A and T are inputted. So a string is generated which consists of randomly generated A, T, G and C. So the DNA sample is ready. The length of the string is assumed to be 1000. One can obviously increase it. Instead of writing the code for generator.cpp, one could actually obtain the standard DNA sequences from NCBI website. But these are in .fasta format and what we use here is .txt format. But this shouldn't matter.

**2. dna\_check.cpp :**

This basically finds the maximally occuring k-mer in the string. K-mer typically refers to all the possible substrings of length k that are contained in a string. So the input, i.e. the string generated using generator.cpp is actually first written to a file i.e. into .txt format and this code reads a .txt file and stores the string in a datatype called as string and then a function findMaxSubstring() finds the maximally occuring k-mer in the string i.e. the substring that occurs most number of times in that string. Here the length of the substring is 100 i.e. k. So if you increment the length of the DNA you can also increase the value of k and point to be noted is that The amount of k-mers possible given a string of length, L, is L + k – 1.

**3. check.c :**

So suppose we have two DNA sequences generated by dna\_check.cpp, then this code checks the them and tells whether it is same or not, i.e. if they are same it prints FOUND else prints NOT FOUND. So this basically checks whether the sequence is same or not.

**2.2 Theory**

**1. DNA :**

DNA is a molecule that carries the genetic instructions used in the growth, development, functioning and reproduction of all known living organisms and many viruses. Consists of two anti-parallel strands and are composed of simpler monmer units known as nucleotides and a sugar called as deoxyribose and a phosphate group. Four nucleotides are Cytosine(C), Guanine(G), Thymine(T) and Adenine(A) and structure is double helix. Basically DNA is a large string of A, G, T and C. A and T compliment each other whereas C and G compliment each other. DNA STORES BIOLOGICAL INFORMATION. DNA is often called as BLUEPRINT of life.

**2. VNTR :**

It is a location in the Genome where a short nucleotide sequence is organized as tandom repeat. These can be found on many chromosome & often show variation in length between individual.For many tandem repeats, the number of repeated units vary between individuals. Such loci are termed VNTRs. One VNTR in humans is a 17 bp sequence of DNA repeated between 70 and 450 times in the genome. The total number of base pairs at this locus could vary from 1190 to 7650.

**3. RNA :**

RNA stands for ribonucleic acid. It is an important molecule with long chains of nucleotides. A nucleotide contains a nitrogenous base, a ribose sugar, and a phosphate. Just like DNA, RNA is vital for living beings.Three types of RNA are m-RNA, t-RNA and r-RNA.**Messenger RNA** (**mRNA**) is a large family of RNA molecules that convey genetic information from DNA to the ribosome, where they specify the amino acid sequence of the protein products of gene expression. **Transfer ribonucleic acid** (**tRNA**) is a type of RNA molecule that helps decode a messenger RNA (mRNA) sequence into a protein. tRNAs function at specific sites in the ribosome during translation, which is a process that synthesizes a protein from an mRNA molecule. **Ribosomal** ribonucleic acid (**rRNA**) is the RNA component of the ribosome, and is essential for protein synthesis in all living organisms. It constitutes the predominant material within the ribosome, which is approximately 60% rRNA and 40% protein by weight.

**4. DNA Sequencing :**

DNA sequencing is the process of determination of precise order of nucleotides in DNA i.e. the precise order of A, T, G and C in a strand of DNA for sequencing. Main methods are MAXAM-GILBERT sequencing and CHAIN TERMINATION method of sequencing.

4.1. MAXAM – GILBERT SEQUENCING :

Found by A.M. Maxam and W. Gilbert in 1977. Sequencing is performed by chain breakage at specific nucleotides.

Procedure :

1. Denature a double-stranded DNA to single-stranded by increasing temperature.
2. Radioactively label one 5' end of the DNA fragment to be sequenced by a kinase reaction using gamma-32P.
3. Cleave DNA strand at specific positions using chemical reactions. For example, we can use one of two chemicals followed by piperdine. Dimethyl sulphate selectively attacks purine (A and G), while hydrazine selectively attacks pyrimidines (C and T). The chemical treatments outlined in Maxam-Gilbert's paper cleaved at G, A+G, C and C+T. A+G means that it cleaves at A, but occasionally at G as well.
4. Now in four reaction tubes, we will have several differently sized DNA strands.
5. Fragments are electrophoresed in high-resolution acrylamide gels for size separation.
6. These gels are placed under X-ray film, which then yields a series of dark bands which show the location of radiolabeled DNA molecules. The fragments are ordered by size and so we can deduce the sequence of the DNA molecule.

4.2 CHAIN TERMINATION METHOD :

Also known as Sanger sequencing, it is a method of DNA sequencing first commercialized by Applied Biosystems, based on the selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during in vitro DNA replication. Found by Sanger in 1977.

Procedure :

1. Get enough quantity of DNA (run PCR) and aliqot DNA into four different reaction tubes.
2. Prepare reaction mixture which consists of Primer, taq PM, template(ssDNA), dNTPs and ddNTPs( ddATP, ddGTP, ddCTP, ddTTP) and run PCR again.
3. Perform Gel Electrophoresis and deduce DNA sequence.

**5. APPLICATIONS :**

1. Forensics : to help identify individuals because each individual has different genetic sequence.

2. Medicines : can be used to help detect genes which are linked to various genetic disorders such as muscular dystrophy.

3. Agriculture : the mapping and sequencing of genome of micro organism has helped make them useful for crops and food plants.

4. Molecular Biology : Sequencing is used in molecular biology to study genomes and proteins they encode.

5. Evolutionary Biology : since DNA is informative micro molecule in terms of transmission from one generation to another, DNA sequencing is used in evolutionary biology to study how different organisms are related and how they evolved.

**6. DISADVANTAGES :**

1. Whole genome cannot be sequenced at once.

2. Very slow and time consuming.

**3. Result**

So when the code is run we can find whether the DNA sequence is same or not and if we can atleast generate 5 such matching sequences we can definitely say that that person is the culprit. So this basically can be used extensively in forensics. The code can be extended to match atleast 5 sequences by just putting them in a loop. Hence, this can be used in real life to find the victim among many suspects. Just the DNA sample of the suspects are required and some left over hair or blood or sweat sample in the area of the crime committed is enough to find the victim since it is very very rare to find two persons having the same DNA.

**4. Conclusions**

Every individual has different DNA but the DNA sequence can be same in one in ten. But, if we match more than 5 DNA sequences then the probability of finding the victim increases. DNA is very important in terms of transfer of information from one to another and it basically stores the genetic information and is called blueprint of life.

**6. Ethical issues**

Human genetics have been included within the field of bioethics since the early 1970s and the growth in the use of DNA sequencing (particularly high-throughput sequencing) has introduced a number of ethical issues. One key issue is the ownership of an individual's DNA and the data produced when that DNA is sequenced. As DNA sequencing becomes more widespread, the storage, security and sharing of genomic data has also become more important.

**REFERENCES**

1. [www.slideshare.net](http://www.slideshare.net/)

2. <https://binf.snipcademy.com/lessons/dna-sequencing-techniques/maxam-gilbert>

3. http://www.whatisbiotechnology.org/science/sequencing

4. [www.wikipedia.org](http://www.wikipedia.org/)

5. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2682332/