Cairo University

Graduation Project Documentation

Next Generation Sequencing Data Visualization

Author:
Fatima Hassan
Doaa Hashem

Supervisor:
Prof. Ayman ElDieb
Prof. Mohammed AbouElHoda

A thesis submitted in fulfillment of the requirements for the Bachelor Degree

in the

Systems and Biomedical Engineering Department

July 2, 2016

Acknowledgment

Prima facie, we would like to thank Allah for the wellbeing, energy, and passion to complete this project. We wish to express our sincere appreciation and thanks to our supervisors Prof. Mohamed Abou ElHoda and Prof. Ayman ElDeib not only for their great support and assistance till this project came to light, but also for giving us the opportunity to set a foot onto a career we never thought possible during our undergrad years.

Furthermore, We would like to extend our gratitude to Eng. Mostafa Shokrof and Eng Mohamed Elkalioby , researchers and teaching assistants at Nile University, for their great support, encouragement, scientific discussions, and valuable and noble way of supervision.

Most importantly, none of this would have been possible to happen without the aid and encouragement of our families, and to them we are forever grateful.

Finally, we would like to thank the administrative team and our professors at our department of Systems and Biomedical Engineering for their cooperation.

Abstract

Next Generation Sequencing is a new technology for fast and cost effective reading of genomic data. Whole genome can now be sequenced within one day and a cost of about 1000 USD.

The challenge with this technology is the huge amounts of data that limits its efficient analysis. Management of NGS data includes efficient handling of sequencing reads in terms of transferring them, storing, sorting, and extracting subsets of them. It also includes comparison of them to a reference genome and visualizing the results.

The project aims at two main goals: evaluating basic bioinformatics tools for management of NGS read data, which includes SAMTOOLS, and IGV and light weight plotting of reads and alignments using python.

Moreover, our visualization is Linux based and open source visualization based on Plotly and Pysam packages.

Contents

Chapter 1

Introduction

The world around us is filled with different life forms that urge to be discovered. A single gram of fertile soil, for instance, may contain as many as 2.5 billion unicellular organisms. This wide diversity of life forms shows how much there is still to be learned and explored about life. Wherever you live, the world around you teems with life.

In the domain of biology, scientists or researchers collect and interpret data beginning with the interaction of communities and populations together to cell theories and genetics. Bringing together the fields of biology and computer science has led to a revolutionary peak in the amounts of data interpreted and the massive amounts of detailed knowledge of life science. Figure 1.1 shows the evolutionary change of knowledge basis in the last half century. The field of science or biology has expanded, or more precisely has evolved, in a way unprecedented and unexpected by scientists or researchers, providing us with massive amounts of data , not to be collected, but to be analysed.

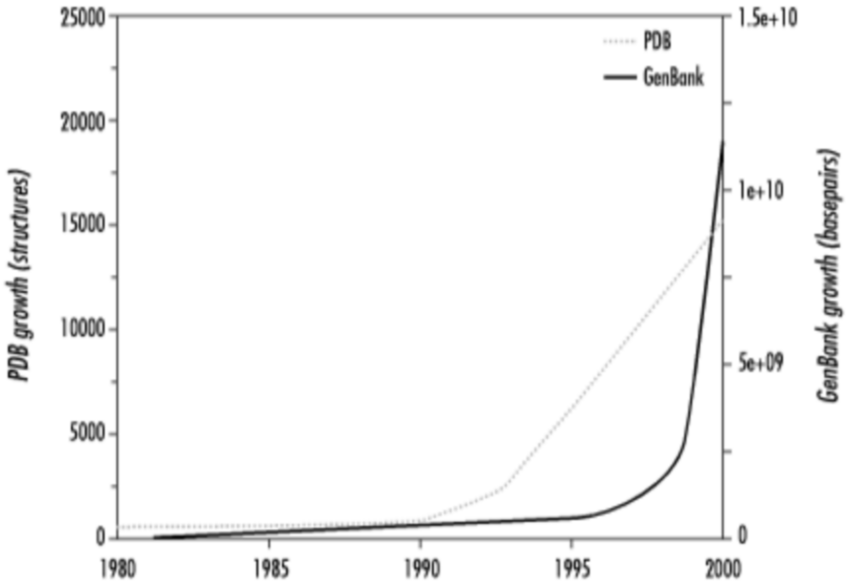

Figure 1.1: caption

Bioinformatics is the intersection between medicine, life science, biology, and computer science. It is where magic happens. Bioinformatics reveals the true value of computer science via the interpretation, collection, design, manipulation, maintenance, and distribution of biomedical data. Bioinformaticians are the tool-builders, as they combine the knowledge in the biological field with the mathematical and statistical approaches, finding solutions and designing algorithms for data analysis, and much more.

Working in the field of bioinformatics requires a basic understanding in the biological and life science problems, as to be able to create the suitable tools, interfaces, and analytical models through which researchers can use to answer complex questions and prove different theories and models of life. The ultimate goal of analytical bioinformatics is to develop predictive methods that allow scientists to model the function and phenotype of an organism based only on its genome sequence. This is a grand goal,

and one that will be approached only in small steps, by many scientists working together.

To be able to advance in bioinformatics field, one should have a fairly deep background in some aspect of molecular biology and an absolute understanding of its central dogma. The existence of genome projects implies our intention to use the data they generate. The implicit goals of modern molecular biology are, simply stated, to read the entire genomes of living things, to identify every gene, to match each gene with the protein it encodes, and to determine the structure and function of each protein. Detailed knowledge of gene sequence, protein structure and function, and gene expression patterns is expected to give us the ability to understand how life works at the highest possible resolution. Implicit in this is the ability to manipulate living things with precision and accuracy.

1.1 Central Dogma

This chapter establishes a brief introduction to molecular biology with an emphasis on genomics and bioinformatics. It is intended for engineers, computer scientists, and/or anybody with a background or strong interest in science, regardless of a background in biology. Finally, it will be subject to the field of bioinformatics and its applications.

1.1.1 DNA

Deoxyribonucleic acid (DNA) is the genetic source that enables cells to have different forms and perform different functions. The primary function of DNA in organisms is to store and transmit the genetic information that tells cells which proteins to make, and when to make them.

DNA Structure

The structure of the DNA is made up of two long chains, or subunits, of nucleotides. A nucleotide has three parts: a sugar molecule called**deoxyribose**, a phosphate group, and a **nitrogen-containing base**. The deoxyribose sugar and the phosphate group are always identical. The nitrogen-containing base, however, may be one of four kinds whose abbreviations represent the nucleotide. They are **adenine**, **guanine**, **cytosine**, and **thymine**. Their abbreviations are **A**, **G**, **C**, **T**, respectively. Bases that have two rings of carbon and nitrogen atoms, such as adenine and guanine, are called **purines**. Bases that have only one ring of carbon and nitrogen atoms, such as cytosine and thymine, are called **pyrimidines**. Complementary base-pairing rules state that cytosine bonds with guanine while adenine bonds with thymine, forming hydrogen bonds.

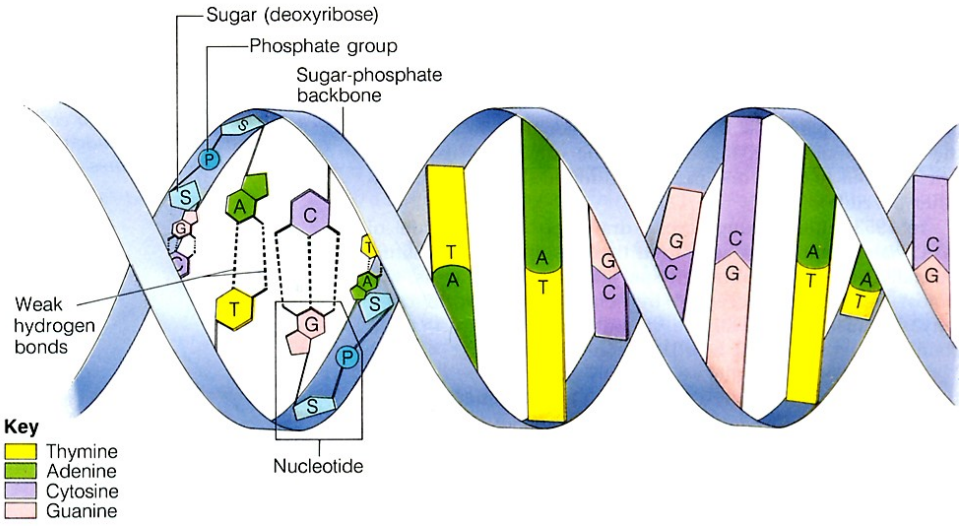

Figure 1.2: caption

The model of DNA is composed of two nucleotide chains that wrap around each other, forming a double spiral called a **double helix**. The individual nucleotides face toward the alternating deoxyribose sugar

and the phosphate molecules to which they are attached by covalent bonds. Moreover, they face toward the helix center and the bases on the other chain of DNA, with which they form hydrogen bonds. The base pairs are of uniform length as in each case one base is a double-ringed purine and the other is a single-ringed pyrimidine. The form of DNA that is most commonly found has a right-hand twist, with each full turn consisting of ten base pairs.

DNA Replication

During replication, the process of copying DNA in a cell, first, the two nucleotide chains separate by unwinding. They are separated by **helicase** enzymes, and each serves as a template for a new nucleotide chain. Second, **DNA polymerases** bind to the separated chains of nucleotides. One nucleotide at a time, the enzyme constructs a new complementary chain of nucleotides. DNA polymerases begin replication simultaneously at various parts along the separated chains, which allows for faster replication without affecting its accuracy. About one error takes place in every 10,000 paired nucleotides. Any change that takes place in a nucleotide sequence at any one location is a **mutation**, which may have serious effects in new cells. However, the number of errors and mutations in DNA replication is reduced as enzymes proofread DNA and repair errors. At the end of replication, there are two identical copies of the original DNA molecule. Each molecule is made of one chain of nucleotides from the original molecule and one new chain of nucleotides. Once replication is completed, the cell is ready for cell division.

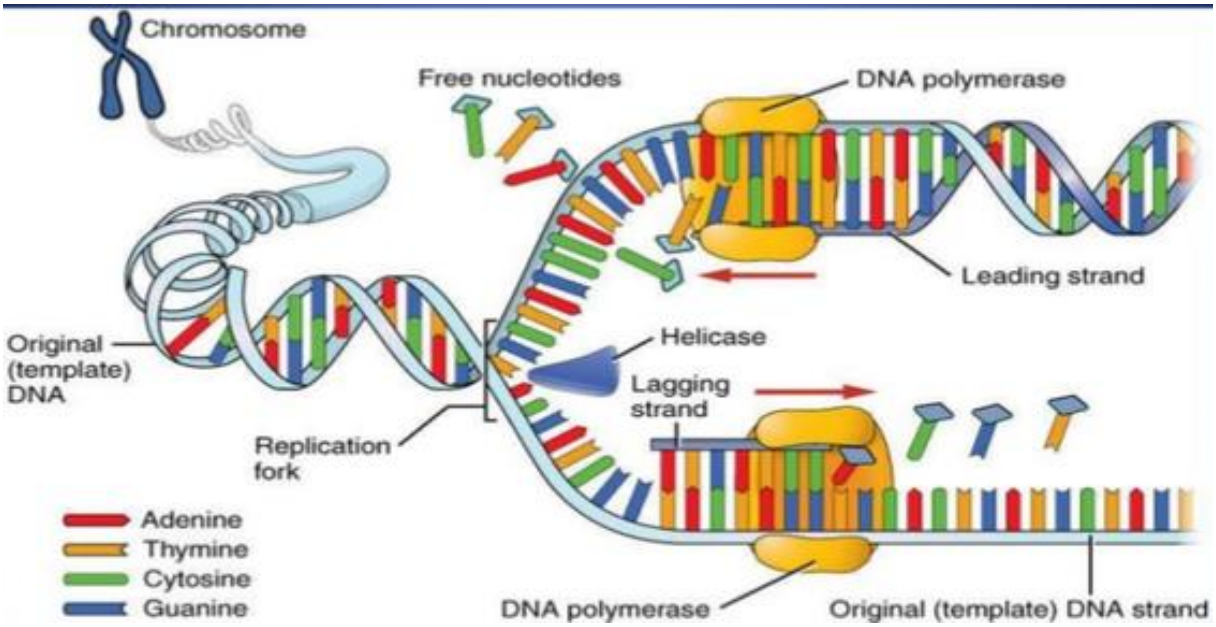

Figure 1.3: caption

1.1.2 RNA

The ribonucleic acid (RNA) is a nucleic acid that is responsible for the movement of genetic information from the DNA in the nucleus to the site of protein synthesis in the cytosol of the cell.

RNA Structure

RNA is comparable with DNA in a couple of points. RNA is made up of repeating nucleotides. However, the sugar molecule of every RNA nucleotide is **ribose**, whereas DNA nucleotides contain deoxyribose sugar. A second difference is that **uracil**, instead of thymine, usually pairs with adenine in RNA.

RNA exists in three types: **Messenger RNA (mRNA)** carries the RNA nucleotides in the form of a single, uncoiled chain from the DNA to the cytosol. **Transfer RNA (tRNA)** consists of folded nucleotides that bind to specific amino acids. There are about 45 varieties. **Ribosomal RNA (rRNA)** makes up the ribosomes where proteins are made. Its the most abundant form of RNA.

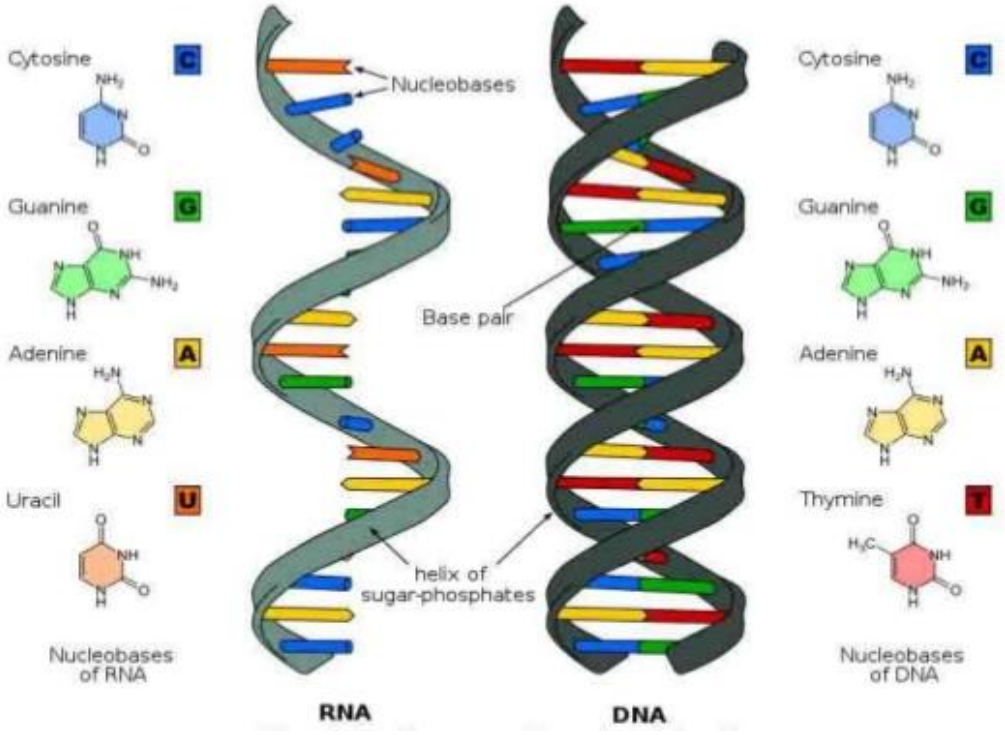

Figure 1.4: caption

Transcription

Its the process by which the genetic information is copied from DNA to RNA. During transcription, **RNA polymerase**, the primary transcription enzyme, synthesizes RNA copies of specific sequences of DNA. RNA polymerase binds to specific DNA regions marking the beginning of DNA chain that will be transcribed called **promoters**. The nucleotide chain separates at the region where the RNA polymerase bound to the DNA chain allowing for transcription to take place. Only one of the separated DNA chains are used for transcription.

RNA polymerase then binds to the first nucleotide in the DNA chain to be transcribed. Then it begins adding complementary RNA nucleotides one at a time to the newly forming RNA molecule until the RNA polymerase reaches a DNA region called the **termination region**, a specific sequence of nucleotides marking the end of transcription. All three types of RNA molecules are formed in the transcription process. MRna then moves through nucleus pores to the cytosol of the cell for protein synthesis.

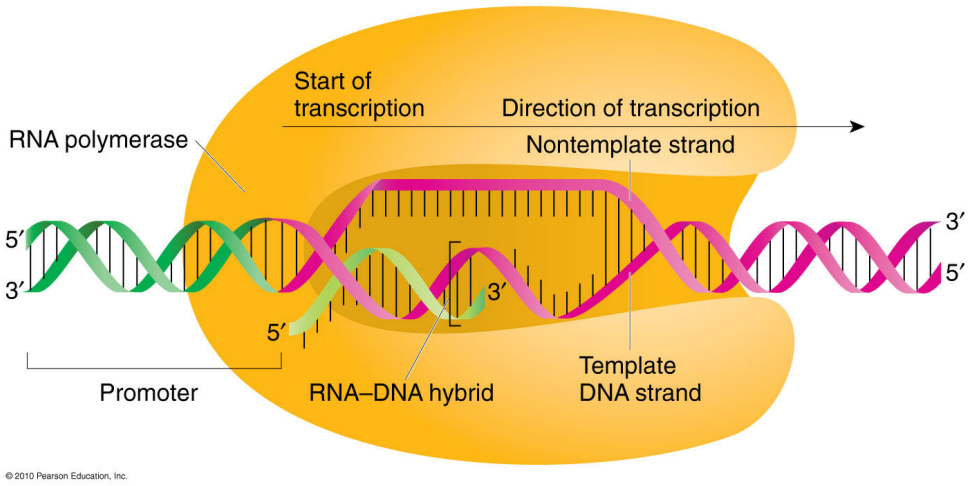

Figure 1.5: caption

1.1.3 Proteins

Proteins Structure

The amount and kind of proteins that are produced in a cell determine the structure and function of the cell. In this way, proteins carry out the genetic instructions encoded in an organism's DNA. The structure of a protein is made up of one or more polypeptides that consist of different amino acids arranged in a particular sequence. The function of a protein depends on its three-dimensional structure which is determined by its amino-acid sequence.

During protein synthesis, the **genetic code**, or the correlation between a nucleotide sequence in a mRNA and an amino-acid sequence, is used. This code translates the nucleotide sequence into proteins. Each **codon** (the combination of three mRNA nucleotides) codes for a specific amino acid.

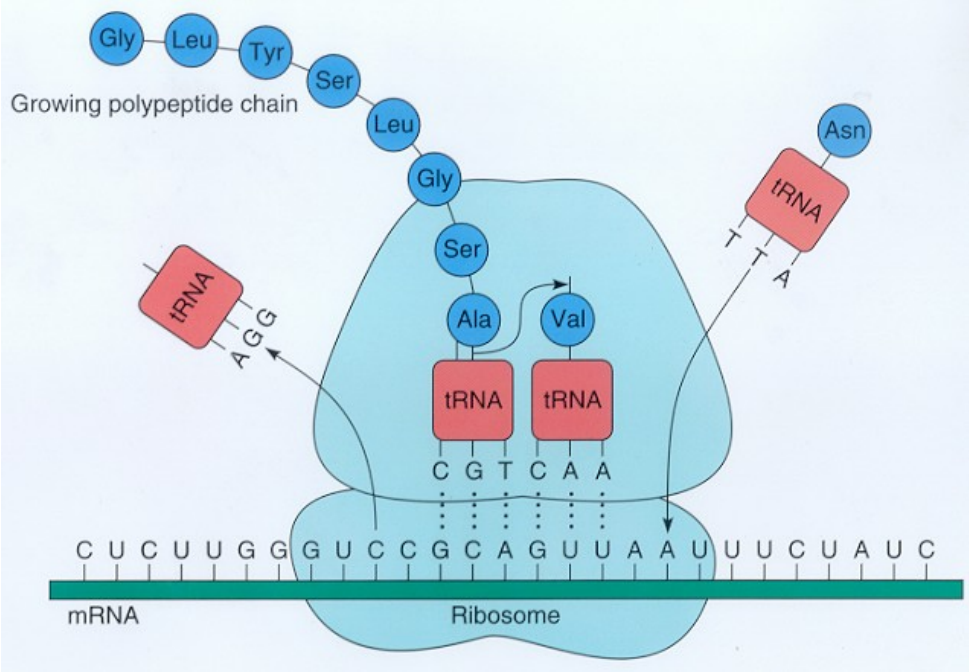

Figure 1.6: caption

Translation

It's the process of assembling polypeptides in mRNA. It begins when mRNA leaves the nucleus and migrates to a ribosome in the cytosol, the site of protein synthesis. Amino acids floating freely in the cytosol are transported to the ribosomes by tRNA molecules. The same base-pairing rules followed during transcription are followed during translation.

A tRNA molecule has two binding regions. One region binds to a specific amino acid. The other region, called the anticodon, is formed of three nucleotides that are complementary and pair with three mRNA codons .

1.2 Gene Expression

Gene expression, or the activation of a gene, occurs in transcription and translation. Regulating a gene expression allows the cells to control which portion of the genome, the complete genetic material, will be expressed and when. Gene expression has evolved mechanisms to ensure that each protein is produced only when it's needed.

1.3 DNA Technology

The application of molecular genetics for practical uses is called **genetic engineering**. DNA technology is involved in genetic engineering; it can cure diseases, treat genetic disorders, improve food crops and increase agricultural yields, and much more. This process takes place by manipulating genes using cloning vectors, which are the carriers used in this procedure to clone a gene and transfer it from one body to another.

DNA technology intersects diverse aspects of technology trends nowadays. At one of its aspects, **DNA fingerprints** is a pattern of bands made up of specific fragments from an individual's DNA. DNA fingerprints have many uses. They can be compared to establish whether two individuals are related, for instance, or to analyse blood and stains in crime scenes. DNA fingerprints are very accurate because they compare segments of DNA that vary the most from one person to another. These repeated patterns of DNA segments are found throughout the genome.

Figure 1.7: caption

Moreover, many viral diseases can only be treated effectively by prevention, using vaccines. Vaccines are solutions that contain a harmless version of a virus. DNA technology can be used to produce effective vaccines, instead of disease-causing agents called pathogens that are chemically or physically treated so that they can no longer cause disease. DNA technology can also be used to alter the pathogen's genome so that it no longer causes a disease. These altered pathogens can therefore be used as vaccines.

Furthermore, genome research all over the world uses DNA technology to achieve the two main goals of the Human Genome Project (HGP). The two goals were to determine the order or sequence of the entire human genome and to map the location of every gene on each chromosome producing linkage maps. This global research continued in an effort to understand how genomes are organized, how gene expression is controlled, how cellular growth and differentiation are under genetic control, and how evolution occurs. The project ha revealed the entire nucleotide pairs of DNA in the human genome. The full sequence was completed and published in April 2003 costing billions of dollars, yet it has made it possible to establish the identity of and determine the function of many hidden secrets of DNA sequence.

Nevertheless, in 1990, doctors began to develop and test gene therapies on humans. Gene therapy is well suited for treating genetic disorders that result from a deficiency of a single enzyme or protein. Treating a genetic disorder by introducing a gene into a cell or by correcting a gene defect in a cell's genome is called **gene therapy**.

DNA technology has also intersected with the field of safety and environmental issues, where the procedure of genetic engineering and the safety of genetically engineered products are regulated by the Food and Drug Administration (FDA), the National Institutes of Health Recombinant DNA Advisory Committee, the Department of Agriculture (USDA), and the Environmental Protection Agency (EPA)

in the US to set standards for safety and products and allow special permits and labels to be used or sold.

1.4 Next Generation Sequencing

DNA codes for the instructions that when executed metabolize our body and takes care of all necessary or needed functions by cells. Bases or nucleotides in DNA strands could code for regular instructions, genes that turn on or off specific regularities, or even could code for a certain disease. Understanding the hidden messages of DNA sequences helps us understand more of life science.

Sequencing the genomes, or reading the the bases or nucleotides sequence of the DNA strand, gives us the opportunity to understand which genetic part codes for what function, detect mutations, and much more.

1.4.1 Why Sequencing

Gaining knowledge of the DNA sequence enables us to study rare genetic diseases in children by sequencing the genomes of ancient humans to learn more of human origins, evolution, and patterns of migration. Sequencing is a core master in the field of cancer as well. It is used the sequencing methodology to identify drug-resistance in HIV. It is also used in the study of vast number of microbes, bacteria, and such that live in our bodies especially our guts. We use sequencing to characterize the human gut microbiome. What we eat has a significant impact on our gut microbial communities. This has significant implications for research on human nutrition, obesity and famine. Sequencing has also been used in plant and agricultural research and many other fields and application in diverse fields of life science.

1.4.2 Polymerase Chain Reaction

Polymerase Chain Reaction (PCR) is a simple but very powerful technique through which we create multiple copies of DNA strands. The technique simply denotes two properties of the DNA structure. First, DNA sticks to itself; DNA is always double-stranded. Second, that the DNA is directional. And the direction that it goes, we always talk about it going from the 5' direction to the 3' direction. So the beginning of the DNA sequence will be the 5' end and the end will be the 3' end.

In PCR, we start by melting the DNA strands, or heating it up to 72◦ C to break its bonds and acquire two single strands of DNA. Next, we cool it down slightly to 54◦C for primers to bond onto the single strands for the DNA to be copied. A **primer** as we recall is simply a short sequence of bases usually 15 to 20 bases long that is complementary to the DNA strand we wish to copy.

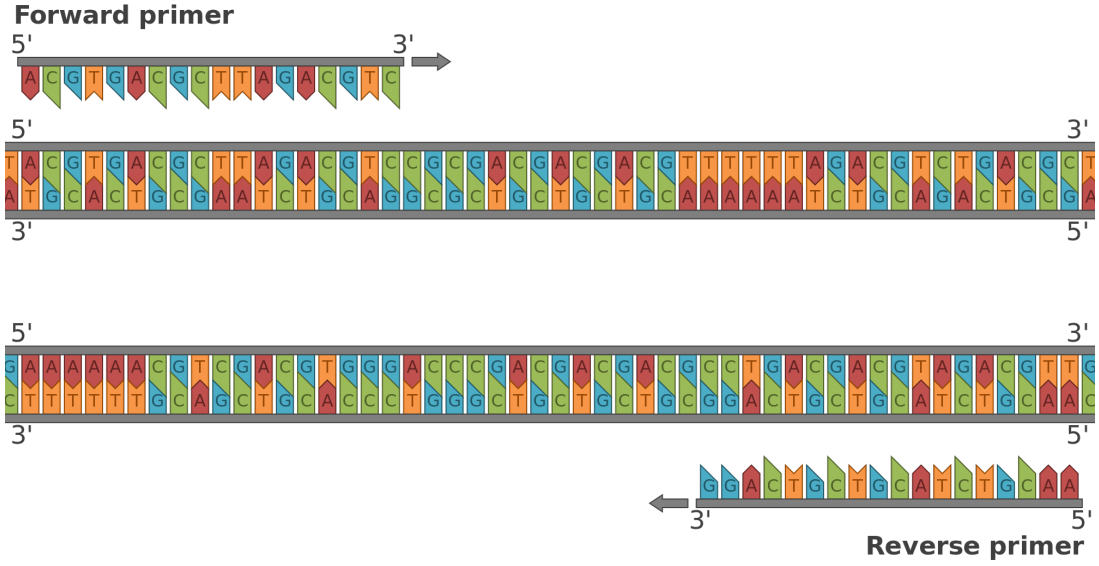

Figure 1.8: caption

Now that our strands are ready to be copied, DNA polymerase of any species are added and raw bases. The DNA polymerase binds the corresponding bases of each single strand starting from the end of the primer bound to the single-strand until the end of the molecule, so that the DNA polymerase can incorporate them into the new double-stranded DNA that it's creating.

We then repeat the whole process again, resulting in four double-stranded DNA molecules. Iterating the process, each time results in double the number of DNA strands and that is the reason why it's been named a chain reaction.

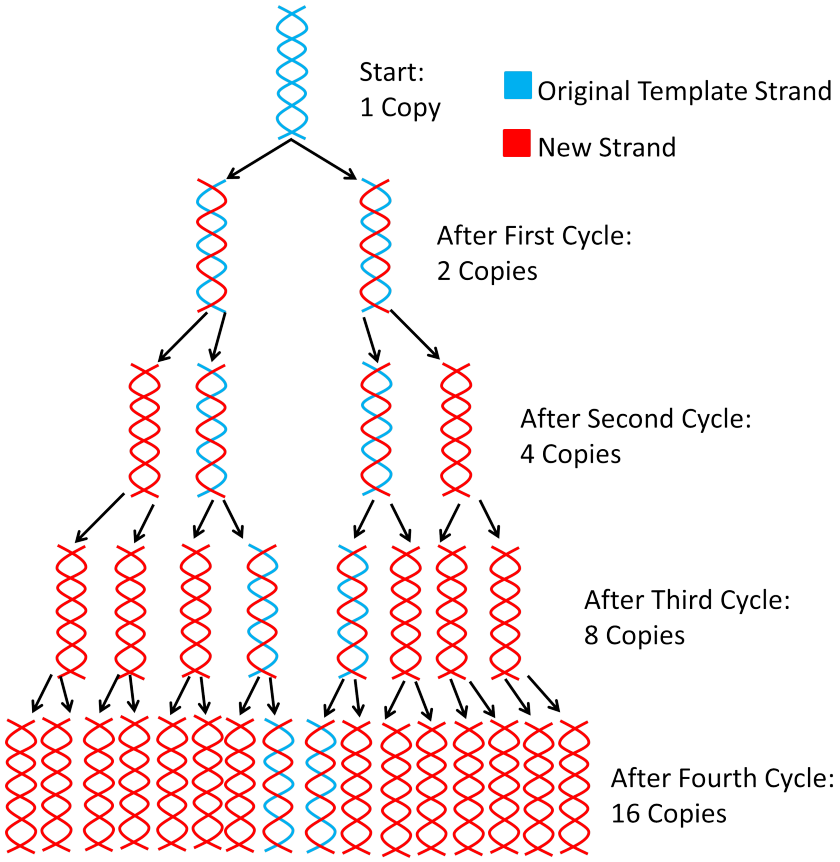

Figure 1.9: caption

1.4.3 Impact of Next-Generation Sequencing

Next-Generation Sequencing (NGS), or high-throughput sequencing, has created a revolution in the area of genomic research in all fields of life science. Despite the great evolution the Human Genome Projected fuelled into the field, it is considered quite expensive in cost and time analyses.

Nowadays, a whole human genome can be sequenced within one day at a cost of about 1000 dollars. NGS is a term used to portray novel sequencing technologies. NGS technology has given us the opportunity to sequence either DNA or RNA in a quicker and a more cheap manner than of Sanger's sequencing methodology, and as such have revolutionized the study of genomics and molecular biology. It has beat the obsolete first generation sequencing methodology of the British scientist Fred Sanger by providing unprecedented speed, cost, and high-throughput frequency which has opened a whole new world of possibilities in the biological world. What came unexpected was not the sequencing process, but the enormous amounts of data produced yet to be analysed. The major advances in NGS allows fast and cost effective reading of genomic data. The ability to generate genomic data in such a manner is now transforming the nature of biological community.

1.4.4 Next-Generation Sequencing Procedure

Considering the huge impact of NGS on the biological area, let's dig a foot onto how NGS takes place. All DNA sequencing methodologies rely on DNA extraction from the test or subject in the laboratory. DNA extracted is then chopped into tiny snippets of reads that are usually 100 base pair long and could be even longer to about 1000 bps. Each snippet is now a read template waiting to be sequenced. Recall that the sequencing technique has been introduced to be able to read the base sequence of the DNA, and to do so we will chemically attach all the reads to a flat surface or slide at one base, given that all reads are single stranded.

Now, given the possibility of creating a complementary strand to each single-stranded read on the slide and watching from a distance a DNA polymerase creating complementary strands to the reads, we can acquire the sequence of bases in the reads being complemented. And thus, we add DNA polymerases and labelled raw bases with terminators that terminate the work of DNA polymerase after being added and emit a fluorescent light with a certain color indicating the base type. Bases are labelled in such a way that they emit a fluorescent light in four different manners, therefore if the base C emits an orange color, we can conclude that we have a Cytosine base every time we snapshot an orange color. If we could take a picture or a snapshot of the emitted light, remove the terminator, add another base, and iterate the cycle, then we will obtain a series of photos containing different colors. Coding these colors to their corresponding base type, we can conclude the sequence of the reads' templates.

NGS has a number of error sources during the process. NGS reads or bases produce little light to be seen in a photograph, therefore before we start the process we need to amplify the source of light emitted. Amplification is done through creating clustered copies of each read on the slide where all the clustered copies stand next to the original template. So when the process begins the emitted light is not produced from only one read, instead a entire cluster of reads emit the same color making it possible to be seen and recorded in the photograph. Unterminated bases cause out of sync templates, and as we iterate cycles, the number of templates falling out of sync increase causing uncertainty in deciding the actual detected color of the cluster.

This concludes why we have to chop the DNA into tiny snippets before we start sequencing, as the number of errors increase as we we continue to iterate. And very long strands may result in tremendous amounts of errors in the NGS process.

Nevertheless, one major advantage of NGS is that we can sequence billions of reads' templates on just a single slide enabling massive parallel processing. The open discussion facts lie now onto the analyses of the huge number of collected data.

Chapter 2

Materials and Methods

Appendix A

Another example

A.1 Appendix A

Bla bla.

Bibliography

List of Figures