Chapter 1

INTRODUCTION

Bioinformatics is an interdisciplinary field mainly involving molecular biology and genetics, computer science, mathematics, and statistics. Data intensive, large-scale biological problems are addressed from a computational point of view.

1.1 Introduction

The most common problems are modeling biological processes at the molecular level and making inferences from collected data. A bioinformatics solution usually involves the following steps: Collect statistics from biological data [1]. Build a computational model. Solve a computational modeling problem. Test and evaluate a computational algorithm. This chapter gives a brief introduction to bioinformatics by first providing an introduction to biological terminology and then discussing some classical bioinformatics problems organized by the types of data sources [2]. Sequence analysis is the analysis of DNA and protein sequences for clues regarding function and includes sub problems such as identification of homologs, multiple sequence alignment, searching sequence patterns, and evolutionary analyses. Protein structures are three-dimensional data and the associated problems are structure prediction (secondary and tertiary), analysis of protein structures for clues regarding function, and structural alignment [3]. Gene expression data is usually represented as matrices and analysis of microarray data mostly involves statistics analysis, classification, and clustering approaches. Biological networks such as gene regulatory networks, metabolic pathways, and protein-protein interaction networks are usually modeled as graphs and graph theoretic approaches are used to solve associated problems such as construction and analysis of large-scale networks [4].

Bioinformatics involves the integration of computers, software tools, and databases in an effort to address biological questions. Bioinformatics approaches are often used for major initiatives that generate large data sets. Two important large-scale activities that use bioinformatics are

Genomics and proteomics. Genomics refers to the analysis of genomes. A genome can be thought of as the complete set of DNA sequences that codes for the hereditary material that is passed on from generation to generation. These DNA sequences include all of the genes (the functional and physical unit of heredity passed from parent to offspring) and transcripts (the RNA copies that are the initial step in decoding the genetic information) included within the genome. Thus, genomics refers to the sequencing and analysis of all of these genomic entities, including genes and transcripts, in an organism. Proteomics, on the other hand, refers to the analysis of the complete set of proteins or proteome. In addition to genomics and proteomics, there are many more areas of biology where bioinformatics is being applied (metabolomics, transcriptomic). Each of these important areas in bioinformatics aims to understand biological systems [4].

Many scientists today refer to the next wave in bioinformatics as systems biology, an approach to tackle new and complex biological questions. Systems biology involves the integration of genomics, proteomics, and bioinformatics information to create a whole system view of a biological entity.

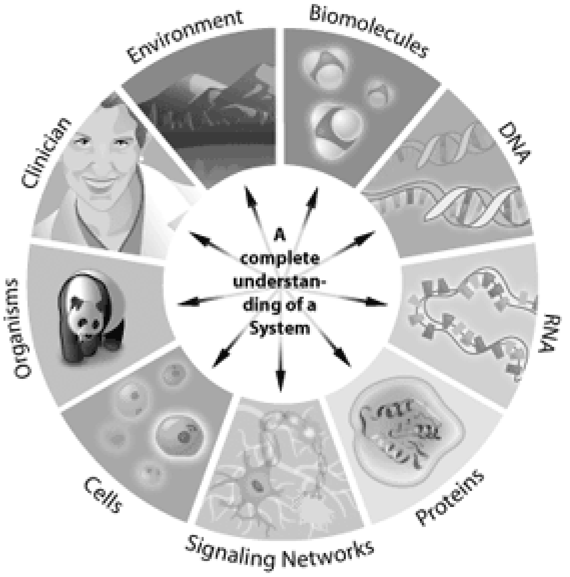


Figure 1. The Wheel of Biological Understanding. System biology strives to understand all aspects of an organism and its environment through the combination of a variety of scientific fields [4].

For instance, how a signaling pathway works in a cell can be addressed through systems biology. The genes involved in the pathway, how they interact, and how modifications change the outcomes downstream, can all be modeled using systems biology. Any system where the information can be represented digitally offers a potential application for bioinformatics. Thus bioinformatics can be applied from single cells to whole ecosystems. By understanding the complete "parts list" in a genome, scientists are gaining a better understanding of complex biological systems. Understanding the interactions that occur between all of these parts in a genome or proteome represents the next level of complexity in the system. Through these approaches, bioinformatics has the potential to offer key insights into our understanding and modeling of how specific human diseases or healthy states manifest themselves [4].

The beginning of bioinformatics can be traced back to Margaret Dayhoff in 1968 and her collection of protein sequences known as the Atlas of Protein Sequence and Structure. One of the early significant experiments in bioinformatics was the application of a sequence similarity searching program to the identification of the origins of a viral gene. In this study, scientists used one of the first sequence similarity searching computer programs (called FASTP), to determine that the contents of v-sis, a cancer-causing viral sequence, were most similar to the well-characterized cellular PDGF gene. This surprising result provided important mechanistic insights for biologists working on how this viral sequence causes cancer. From this first initial application of computers to biology, the field of bioinformatics has exploded. The growth of bioinformatics is parallel to the development of DNA sequencing technology. In the same way that the development of the microscope in the late 1600's revolutionized biological sciences by allowing Anton Van Leeuwenhoek to look at cells for the first time, DNA sequencing technology has revolutionized the field of bioinformatics. The rapid growth of bioinformatics can be illustrated by the growth of DNA sequences contained in the public repository of nucleotide sequences called Gen Bank [5].

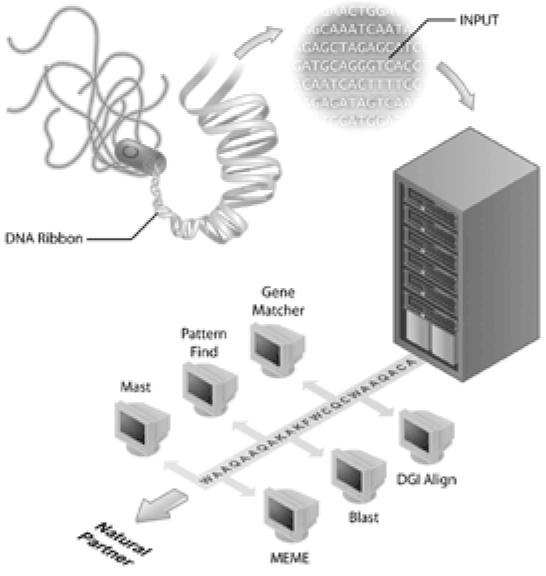


Figure 2. The Use of Computers to Process Biological Information. The wealth of genome sequencing information has required the design of software and the use of computers to process this information [5].

Genome sequencing projects have become the flagships of many bioinformatics initiatives. The human genome sequencing project is an example of a successful genome sequencing project but many other genomes have also been sequenced and are being sequenced. In fact, the first genomes to be sequenced were of viruses and bacteria, with the genome of Hemophilic influenza Rd being the first genome of a free living organism to be deposited into the public sequence data banks. This accomplishment was received with less fanfare than the completion of the human genome but it is becoming clear that the sequencing of other genomes is an important step for bioinformatics today. However, genome sequence by itself has limited information. To interpret genomic information, comparative analysis of sequences needs to be done and an important reagent for these analyses are the publicly accessible sequence databases. Without the databases of sequences (such as Gen Bank), in which biologists have captured information about their sequence of interest, much of the rich information obtained from genome sequencing projects would not be available [5].

The same way developments in microscopy foreshadowed discoveries in cell biology, new discoveries in information technology and molecular biology are foreshadowing discoveries in bioinformatics. In fact, an important part of the field of bioinformatics is the development of new technology that enables the science of bioinformatics to proceed at a very fast pace. On the computer side, the Internet, new software developments, new algorithms, and the development of computer cluster technology has enabled bioinformatics to make great leaps in terms of the amount of data which can be efficiently analyzed. On the laboratory side, new technologies and methods such as DNA sequencing, serial analysis of gene expression (SAGE), microarrays, and new mass spectrometry chemistries have developed at an equally blistering pace enabling scientists to produce data for analyses at an incredible rate. Bioinformatics provides both the platform technologies that enable scientists to deal with the large amounts of data produced through genomics and proteomics initiatives as well as the approach to interpret these data. In many ways, bioinformatics provides the tools for applying scientific method to large-scale data and should be seen as a scientific approach for asking many new and different types of biological questions [5].

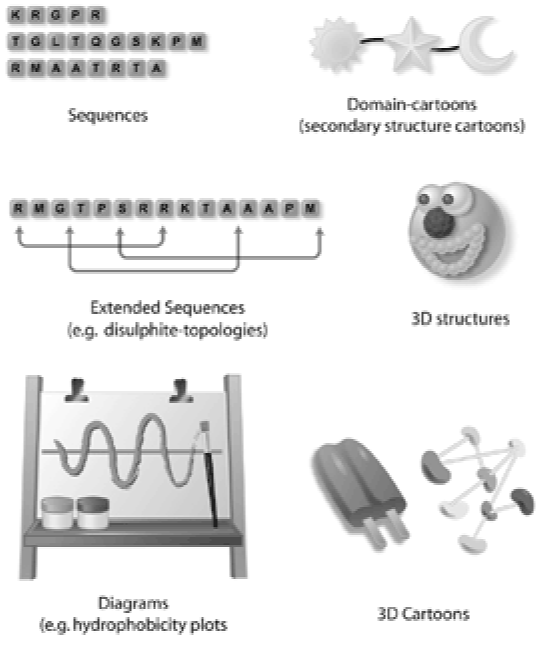


Figure: 3

In Bioinformatics, a sequence alignment is a way of arranging the sequence of DNA, RNA or protein to identify regions of similarity if two sequences in an alignment share a common ancestor, mismatches can be interpreted as point mutations and gaps as indels. In our process, we report binding of cancer cells, normal cells and KRAS genes [6] and detections and sequence mutations of cancer cells which is iterative that can control the cancer cells in feasible tolerant matrices. The goal of this process is to explore the computational approaches to sequence alignment and mutations in a faster and optimal way by using PYMOL software with the help of filtering method by using Anaconda software which filtrate mutated DNA. This approach helps in detecting any abnormal changes and the mutation percentages of those abnormal changes. This process is successful in reading multiple lengths of DNA sequences, detecting high density of cancer cell atoms and generating optimal alignment efficiently. In this process, we have used the idea of both the alignment techniques (Needleman-Wunsch algorithm and Smith-Waterman algorithm for global alignment) which helps in generating proper alignment and comparing with our process. We are hopeful that, the result of our process can make a good help to the biological researchers and others who works in Bioinformatics arena. [6]

Our bodies are made up of trillions of cells. They normally grow, work, divide and die. But when there is a change or damage in these cells which grow out of control called cancer. Cancer cells are different and does not act like normal cells and help to develop tumor. Cancer cells get into the blood and spread more easily to other part of the body [6].

With person's genetic functions, some external agents damage the cells.

- Physical carcinogens: such as Ultraviolet and Ionizing
- Chemical Carcinogens: Tobacco smoke, Arsenic
- Biological Carcinogens: Infections

Here, we work with proteins in our body cell. Proteins are made up of a series of amino acids. Nucleic Acids (RNA and DNA) are made up of a series of nucleotides. The center of an amino acid is the carbon bonded to four different groups. A nucleotide is composed of a five carbon sugar, a nitrogenous base and a phosphate group. DNA fixes which protein is formed. Cancer is a major burden of disease worldwide.

Each year, tens of millions of people are diagnosed with cancer around the world, and more than half of the patients eventually die from it. In many countries, cancer ranks the second most common cause of death. With significant improvement in treatment and prevention of cancer has or will soon become the number one killer in many parts of the world and cancer will remain a major health problem around the globe.

The global cancer burden is estimated to have risen to 18.1 million new cases and 9.6 million deaths in 2018. One in 5 men and one in 6 women worldwide develop cancer during their lifetime, and one in 8 men and one in 11 women die from the disease. Worldwide, the total number of people who are alive within 5 years of a cancer diagnosis, called the 5-year prevalence, is estimated to be 43.8 million [6].

The cancer burden can also be reduced through early detection of cancer and management of patients, where cancer develops. Many cancers have a high chance of cure if diagnosed early and treated adequately.

So far, many methods have been worked out such as Needleman-Wunsch algorithm for global alignment and Smith-Waterman algorithm for local alignment, Dot-matrix method, Word methods, Sequence alignment (Pair wise). Between 30–50% of cancers can currently be prevented by avoiding risk factors and implementing existing evidence-based prevention strategies. But these methods are different from each other and they work differently. But these methods are different from each other and they do not work as a mix interpretation. For implementations, each methods algorithm is to be used separately which are time consuming and costly [6].

We think that protein as a mass noun: a homogeneous substance, something that diet should contain in a certain proportion. But if ever work in a molecular biology lab protein may start to look very different to us. Well, firsthand that protein isn’t just a single substance. Instead, there are lots and lots of different proteins in an organism, or even in a single cell. They come in every size, shape, and type can imagine, and each one has a unique and specific job.

Some are structural parts, givin cells shape or helping them move. Others act as signals, drifting between cells like messages in a bottle. Still others are metabolic enzymes, putting together or snapping apart biomolecules needed by the cell. And, odds are, one of these unique molecular players will become for the duration of research.

Proteins are among the most abundant organic molecules in living systems and are way more diverse in structure and function than other classes of macromolecules. A single cell can contain thousands of proteins, each with a unique function. Although their structures, like their functions, vary greatly, all proteins are made up of one or more chains of amino acids. In this article, we will look in more detail at the building blocks, structures, and roles of proteins [7].

1.2 Types and functions of proteins

Proteins can play a wide array of roles in a cell or organism. Here, we'll touch on a few examples of common protein types that may be familiar to all, and that are important in the biology of many organisms (including us) [7].

1.2.1 Enzymes

Enzymes act as catalysts in biochemical reactions, meaning that they speed the reactions up. Each enzyme recognizes one or more substrates, the molecules that serve as starting material for the reaction it catalyzes. Different enzymes participate in different types of reactions and may break down, link up, or rearrange their substrates.

One example of an enzyme found in body is salivary amylase, which breaks amylose (a kind of starch) down into smaller sugars. The amylose doesn't taste very sweet, but the smaller sugars do. This is why starchy foods often taste sweeter if chew them for longer: giving salivary amylase time to get to work [7].

1.2.2 Hormones

Hormones are long-distance chemical signals released by endocrine cells (like the cells of pituitary gland). They control specific physiological processes, such as growth, development, metabolism, and reproduction.

While some hormones are steroid-based (see the article on lipids), others are proteins. These protein-based hormones are commonly called peptide hormones.

For example, insulin is an important peptide hormone that helps regulate blood glucose levels. When blood glucose rises specialized cells in the pancreas release insulin. The insulin binds to cells in the liver and other parts of the body, causing them to take up the glucose. This process helps return blood sugar to its normal, resting level.

Some additional types of proteins and their functions are listed in the table below [7]: Protein types and functions

Role	Examples	Functions
Digestive enzyme	Amylase, lipase, pepsin	Break down nutrients in food into small pieces that can be readily absorbed
Transport	Hemoglobin	Carry substances throughout the body in blood or lymph
Structure	Actin, tubulin, keratin	Build different structures, like the cytoskeleton
Hormone signaling	Insulin, glucagon	Coordinate the activity of different body systems
Defense	Antibodies	Protect the body from foreign pathogens
Contraction	Myosin	Carry out muscle contraction
Storage	Legume storage proteins, egg white (albumin)	Provide food for the early development of the embryo or the seedling

Proteins come in many different shapes and sizes. Some are globular (roughly spherical) in shape, whereas others form long, thin fibers. For example, the hemoglobin protein that carries oxygen in the blood is a globular protein, while collagen, found in our skin, is a fibrous protein.

A protein's shape is critical to its function, and, as we'll see in the next article, many different types of chemical bonds may be important in maintaining this shape. Changes in temperature and pH, as well as the presence of certain chemicals, may disrupt a protein's shape and cause it to lose functionality, a process known as denaturation [7].

1.2.3 Amino acids

Amino acids are the monomers that make up proteins. Specifically, a protein is made up of one or more linear chains of amino acids, each of which is called a polypeptide. There are 202020 types of amino acids commonly found in proteins [8].

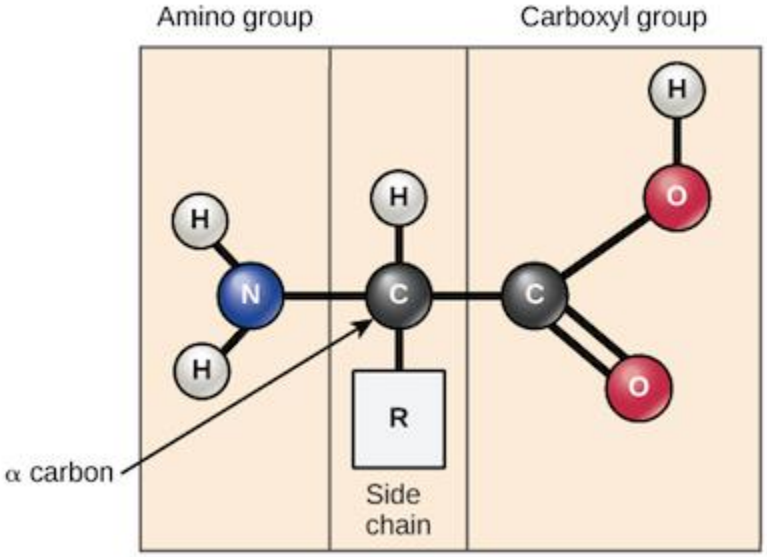


Figure: 4

Image of an amino acid, indicating the amino group, carboxyl group, alpha carbon, and R group. Amino acids share a basic structure, which consists of a central carbon atom, also known as the alpha (α) carbon, bonded to an amino group (\text {NH}_2NH2N, H, start subscript, 2, end subscript), a carboxyl group (\text {COOH}COOHC, O, O, H), and a hydrogen atom.

Although the generalized amino acid shown above is shown with its amino and carboxyl groups neutral for simplicity, this is not actually the state in which an amino acid would typically be found. At physiological pH (7.27.27, point, 2 - 7.47.47, point, 4), the amino group is typically protonated and bears a positive charge, while the carboxyl group is typically deprotonated and bears a negative charge.

Every amino acid also has another atom or group of atoms bonded to the central atom, known as the R group, which determines the identity of the amino acid. For instance, if the R group is a hydrogen atom, then the amino acid is glycine, while if it's a methyl (\text {CH}_3CH3C, H, start subscript, 3, end subscript) group, the amino acid is alanine. The twenty common amino acids are shown in the chart below, with their R groups highlighted in blue [8].

Chart depicting the 20 common amino acids in their predominant protonation forms at physiological pH (7.2-7.4).

The properties of the side chain determine an amino acid's chemical behavior. For example, amino acids such as valine and leucine are nonpolar and hydrophobic, while amino acids like serine and glutamine have hydrophilic side chains and are polar. Some amino acids, such as lysine and arginine, have side chains that are positively charged at physiological pH and are considered basic amino acids. Aspartate and glutamate, on the other hand, are negatively charged at physiological pH and are considered acidic [8].

1.3 Motivation: In recent years, genome projects conducted on a variety of organisms generated massive amounts of sequence data for genes and proteins, which requires computational analysis. Sequence alignment shows the relations between genes or between proteins, leading to a better understanding of their homology and functionality. Sequence alignment can also reveal conserved domains and motifs [9].The motivation and justification factors shown by the study lead to preferring naturalist research for Bioinformatics, Because it depends on real data. The method empowers Bioinformatics techniques to handle the true properties and reducing assumptions for un-modeled or uncover biodata phenomena.

1.4 Principle Outcomes and Objectives: We have developed a model by using PyMol software and based on Global Alignment (Pair-Wise) method that has an algorithm that works on cancer cells. Without detoxifying the protein sequence, the cells will be bind together then it detects the high density area of cancer cells then it aligns the sequence the process will re-connect the protein sequence in a pattern wise manner once after another possible cancer cell then we use Python shell for filtering process which filtrate the mutated genes that can reduce and minimize the death of normal cells. Hopefully, by following our method it will be possible to reduce the cancer cells and minimize the death of normal cells [9].

1.5 Research Challenges (difficulties): We have used the basic match values instead of substitution matrices mentioned in several papers.
We tried using Global which contains the different match values for different nucleotide alignment as well as protein alignment. Our system was unable to read this matrix; if it would have worked then the system could have taken protein sequences as inputs. This method is an algorithm analysis.This method uses Pymol academic edition but working with full edition can give better results. Pymol cannot be used in all versions of windows.

The main limitation is, we still do not know that using this algorithm protein sequence of human gene will be omitting cancer cells and merged with new proteins sequences practically. Development has been focused on capabilities, not on easy-of-use for new users. Although this method is a chemical reaction process, so that pymol simulation of algorithm method does not appear accurate results in all cases [9].

1.5 Contribution

Bioinformatics is now very effective for research and development in biology and medicine, the premise of bioinformatics is integration of diverse sets of data and to synthesize knowledge from the interpretation of such data [10].

Thus, it may be easier to fully understand first the sources of biological information and what is needed for their integration. We first note that biological information had traditionally been obtained through the classical disciplines.

Each of these disciplines contributed fundamental and important information that revolutionized our understanding of the origin, distribution and evolution of organisms, traits, as well as their functions or deformities.

Each of these disciplines contributed fundamental and important information that revolutionized our understanding of the origin, distribution and evolution of organisms, traits, as well as their functions and/or deformities. Knowledge gained from these individual disciplines helped not only the basic sciences, but also helped medical, public health, and even economic policy decision experts. Thus, even though general organismal level knowledge had been the source of bioinformatics observations, their relevance to human health and heredity is often direct and immensely important.

The integration of knowledge gains is basically governed by some common themes of biological investigations together with their respective interdisciplinary features that define many of the currently practiced bioinformatics activities [10].

This thesis report is as follows in bellow

1. Proposed Method
2. Our Objectives & Analysis
3. Proposed Method Flowchart
4. Performance Analysis & Comparison
5. Future work
6. Conclusion
7. References

Chapter 2

Background Study

To perform local sequence alignment between two nucleotide or amino acid sequences and find out structural or functional similarity.

2.1 Theory

The most commonly asked question in molecular biology is whether two given sequences are related or not, in order to identify their structure or function. The most simpler way to answer this question is to compare their sequences [10].

Sequence is a collection of nucleotides or amino acid residues which are connected with each other. Speaking biologically, a typical DNA/RNA sequence consists of nucleotides while a protein sequence consists of amino acids.

Sequencing is the process to determine the nucleotide or amino acid sequence of a DNA fragment or a protein. There are different experimental methods for sequencing, and the obtained sequence is submitted to different databases like NCBI, Gene bank etc.

2.2 Methods of Sequencing:

Sequences stored in the database were obtained from different experimental methods. Most commonly used methods for DNA sequencing are Sanger Method and Maxam-Gilbert Method. Similarly, Edman Degradation method and Mass Spectrometry technique are used for protein sequencing [10].

2.2.1 Sanger Method (dideoxy chain termination method): Here 4 test tubes are taken labelled with A, T, G and C. Into each of the test tubes, DNA has to be added in denatured form (single strands). Next a primer is to be added which anneals to one of the strand in template. The 3' end of the primer accommodates the dideoxy nucleotides [ddNTPs] (specific to each tube) as well as deoxy nucleotides randomly. When the ddNTP's gets attached to the growing chain, the chain terminates due to lack of 3'OH which forms the phospho dieter bond with the next nucleotide. Thus small strands of DNA are formed. Electrophoresis is done and the sequence order can be obtained by analyzing the bands in the gel based on the molecular weight. The primer or one of the nucleotides can be radioactively or fluorescently labeled also, so that the final product can be detected from the gel easily and the sequence can be inferred [10].

2.2.2 Maxam-Gilbert (Chemical degradation method): This method requires denatured DNA fragment whose 5' end is radioactively labeled. This fragment is then subjected to purification before proceeding for chemical treatment which results in a series of labeled fragments. Electrophoresis technique helps in arranging the fragments based on their molecular weight. To view the fragments, gel is exposed to X-ray film for autoradiography. A series of dark bands will appear, each corresponding to a radio labeled DNA fragment, from which the sequence can be inferred [11].

2.2.3 Edman Degradation reaction: The reaction finds the order of amino acids in a protein by cleaving each amino acid from the N-terminal without distrubing the bonds in the protein. After each clevage, chromatography or electrophoresis is done to identify the amino acid.

2.2.4 Mass Spectrometry: It is used to determine the mass of particle, composition of molecule and for finding the chemical structures of molecules like peptides and other chemical compounds. Based on the mass to charge ratio, one can identify the amino acids in a protein [11].

2.3 Sequence Alignment and importance:

Sequence Alignment or sequence comparison lies at heart of the bioinformatics, which describes the way of arrangement of DNA/RNA or protein sequences, in order to identify the regions of similarity among them. It is used to infer structural, functional and evolutionary relationship between the sequences. Alignment finds similarity level between query sequence and different database sequences. The algorithm works by dynamic programming approach which divides the problem into smaller independent sub problems. It finds the alignment more quantitatively by assigning scores.

When a new sequence is found, the structure and function can be easily predicted by doing sequence alignment. Since it is believed that, a sequence sharing common ancestor would exhibit similar structure or function. Greater the sequence similarity, greater is the chance that they share similar structure or function [11].

2.4 Methods of Sequence Alignment:

There are mainly two methods of Sequence Alignment:

2.4.1 Global Alignment: Closely related sequences which are of same length are very much appropriate for global alignment. Here, the alignment is carried out from beginning till end of the sequence to find out the best possible alignment.

2.4.2 Local Alignment: Sequences which are suspected to have similarity or even dissimilar sequences can be compared with local alignment method. It finds the local regions with high level of similarity.

These two methods of alignments are defined by different algorithms, which use scoring matrices to align the two different series of characters or patterns (sequences). The two different alignment methods are mostly defined by Dynamic programming approach for aligning two different sequences [11].

2.4.3 Dynamic programming

Dynamic programming is used for optimal alignment of two sequences. It finds the alignment in a more quantitative way by giving some scores for matches and mismatches (Scoring matrices), rather than only applying dots. By searching the highest scores in the matrix, alignment can be accurately obtained. The Dynamic Programming solves the original problem by dividing the problem into smaller independent sub problems. These techniques are used in many different aspects of computer science. Needleman-Wunsch and Smith-Waterman algorithms for sequence alignment are defined by dynamic programming approach [12].

Smith Waterman algorithm was first proposed by Temple F. Smith and Michael S. Waterman in 1981. The algorithm explains the local sequence alignment, it gives conserved regions between the two sequences, and one can align two partially overlapping sequences, also it’s possible to align the subsequence of the sequence to itself. These are the main advantages of Local Sequence Alignment.

This algorithm mainly differs within two aspects from the Needleman-Wunsch algorithm. First aspect being, local alignment differs by having only a negative score for the mismatch and when the matrix value becomes negative it has to be set to zero (need to take maximum value of the scorings compared with zero). They are also predefined scoring matrices for nucleotide or protein sequences [12].

2.5 Scoring matrices:

In optimal alignment procedures, mostly Needleman-Wunsch and Smith-Waterman algorithms use scoring system. For nucleotide sequence alignment, the scoring matrices used are relatively simpler since the frequency of mutation for all the bases are equal. Positive or higher value is assigns for a match and a negative or a lower value is assigned for mismatch. These assumption based scores can be used for scoring the matrices. There are other scoring matrices which are predefined mostly, used in the case of amino acid substitutions [12].

Mainly used predefined matrices are PAM and BLOSUM.

2.5.1 PAM Matrices: Margaret Dayhoff was the first one to develop the PAM matrix, PAM stands for Point Accepted Mutations. PAM matrices are calculated by observing the differences in closely related proteins. One PAM unit (PAM1) specifies one accepted point mutation per 100 amino acid residues, i.e. 1% change and 99% remains as such [13].

2.5.2 BLOSUM: BLOcks SUbstitution Matrix, developed by Henikoff and Henikoff in 1992, used conserved regions. These matrices are actual percentage identity values. Simply to say, they depend on similarity. Blosum 62 means there is 62 % similarity [13].

2.5.3 Gap score or gap penalty: Dynamic programming algorithms use gap penalties to maximize the biological meaning. Gap penalty is subtracted for each gap that has been introduced. There are different gap penalties such as gap open and gap extension. The gap score defines a penalty given to alignment when we have insertion or deletion. During the evolution, there may be a case where we can see continuous gaps all along the sequence, so the linear gap penalty would not be appropriate for the alignment. Thus gap open and gap extension has been introduced when there are continuous gaps (five or more). The open penalty is always applied at the start of the gap, and then the other gaps following it is given with a gap extension penalty which will be less compared to the open penalty. Typical values are –12 for gap opening, and –4 for gap extension [13].

2.5.4 Assumed scoring schemas: If the residues (nucleotide or amino acids) are same in both the sequences the match score is assumed (Si,j) as +5 which is added to the diagonally positioned cell of the current cell (i, j position). If the residues are not same, the mismatch score is assumed as -3. This score should be added to the diagonally positioned cell of the current cell. The gap penalty score is assumed as -4 which is added to left and above positioned cells of the current cell. These scores are not unique, they can be user defined also, but the mismatch and gap penalty should be the negative values [13].

Related Work:

Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K. W., & Vogelstein, B et al. [1] shows in their work, selective identification of somatic mutations in pancreatic cancer cells through a combination of next-generation sequencing of plasma DNA using molecular barcodes and a bioinformatics variant filter. They use the following parameters- Molecular barcode adapters, Linear amplification, CV78 (Variant Filter). But their main limitations are this method is not effective for removing artifacts. In addition, preexisting somatic mutations in normal cells make it difficult to discriminate. This method only works on selective area in pancreatic cancer cells.

Naser, W. M., Shawarby, M. A., Al-Tamimi, D. M., Seth, A., Al-Quorain, A., Al Nemer, A. M., & Albagha, O. M. et al. [3] shows in their work, Novel KRAS Gene Mutations in Sporadic Colorectal Cancer. In this article, they report 7 novel KRAS gene mutations discovered while retrospectively studying the prevalence and pattern of KRAS mutations in cancerous tissue obtained from 56 Saudi sporadic colorectal cancer patients from the Eastern Province. Genomic DNA was extracted from formalin-fixed, paraffin-embedded cancerous and noncancerous colorectal tissues. KRAS gene mutations were detected in the cancer tissue of 24 cases (42.85%). Of these, 11 had exon 4 mutations (19.64%).

Kamburov, A., Lawrence, M. S., Polak, P., Leshchiner, I., Lage, K., Golub, T. R., ... & Getz, G. et al [4] shows in their work, Large-scale tumor sequencing projects enabled the identification of many new cancer gene candidates through computational approaches. Here, they describe a general method to detect cancer genes based on significant 3D clustering of mutations relative to the structure of the encoded protein products. The approach can also be used to search for proteins with an enrichment of mutations at binding interfaces with a protein, nucleic acid, or small molecule partner. They applied this approach to systematically analyze the Pan Cancer compendium of somatic mutations from 4,742 tumors relative to all known 3D structures of

Human proteins in the Protein Data Bank.

In background study our algorithm related with two algorithms which is Wunch-Needleman algorithm and another is Smith-Waterman algorithm.

Islam, N. (2012) et al [2] shows in their work faster and efficient algorithm for sequence alignment Information in helped us to understand the importance of sequence alignment. We learned the basic strategies used for aligning and finding similarity. It helps to highlighting the major factors to concentrate on while aligning, explained the workings of both local and global alignment algorithm. It helps us learn the heuristic method FASTA algorithm. By the help of these papers we have learned and built this system.

Baugh, E. H., Lyskov, S., Weitzner, B. D., & Gray, J. J. (2011) et al [5] shows in their work, Real-time PyMOL visualization This is a follow-along guide for the Introduction to PyMOL classroom tutorial taught by DeLano Scientific, LLC. It covers the basics of PyMOL for medicinal chemists and other industrial scientists, including visualization of protein ligand interactions, creating figures, and working with session files. This tutorial was created for PyMOL version 1.2 or greater running under Windows, Mac, or Linux.

Likic, V. (2008). The Needleman-Wunsch algorithm for sequence alignment et al [6] shows in their work, the Needleman-Wunsch algorithm works in the same way regardless of the length or complexity of sequences, and guarantees to find the best alignment. The Needleman-Wunsch algorithm is appropriate for finding the best alignment of two sequences which are (i) of the similar length; (ii) similar across their entire lengths.

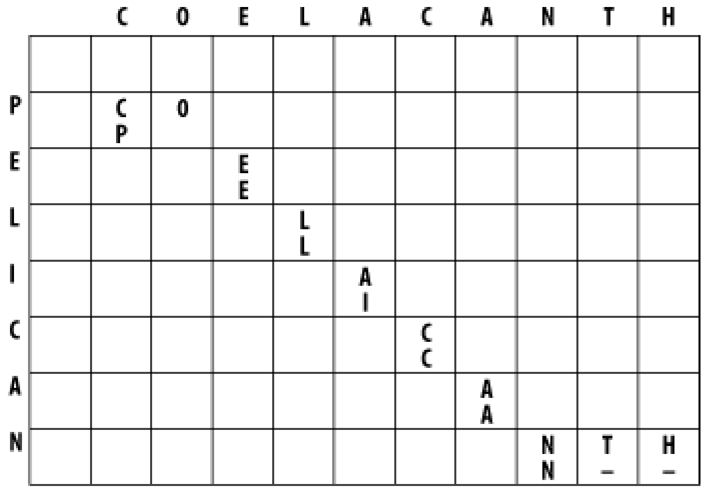
Wunch-Needleman Algorithm: The global alignment algorithm described here is called the Needleman-Wunsch algorithm. We will explain it in a way that seems natural to biologists, that is, it tells the end of the story first, and then fills in the details. (This is why biologists make terrible comedians; they always tell the punch line first.) We will align the words COELANCANTH and PELICAN using a simple scoring scheme: +1 for letters that match, -1 for mismatches, and -1 for gaps. The alignment will eventually look like one of the following, which are equivalent given our scoring scheme [14]:

```
COELACANTH     COELACANTH

P-ELICAN--     -PELICAN--
```

Note that every letter of each sequence is aligned to a letter or a gap. In local alignments, discussed later, this isn't the case.

The alignment takes place in a two-dimensional matrix in which each cell corresponds to a pairing of one letter from each sequence. To get an intuitive understanding of the alignment algorithm, look at Figure 3-1, which shows where the maximum scoring alignment lies in the matrix. The alignment starts at the upper left and follows a mostly diagonal path down and to the right. When two letters are aligned, the path follows a diagonal trajectory. There are several places in which the letters from COELACANTH are paired to gap characters. In this case, the graph is followed horizontally. Although not shown here, the path may be also be followed vertically when the letters from PELICAN are paired with gap characters. Gap characters can never be paired to each other. Note that the first row and column are blank. The reason for this will become clear shortly [14].



		C	O	E	L	A	C	A	N	T	H
P		C P	O								
E				E E							
L					L L						
I						A I					
C							C C				
A								A A			
N									N N	T –	H –

Figure: 2.1: Trace back

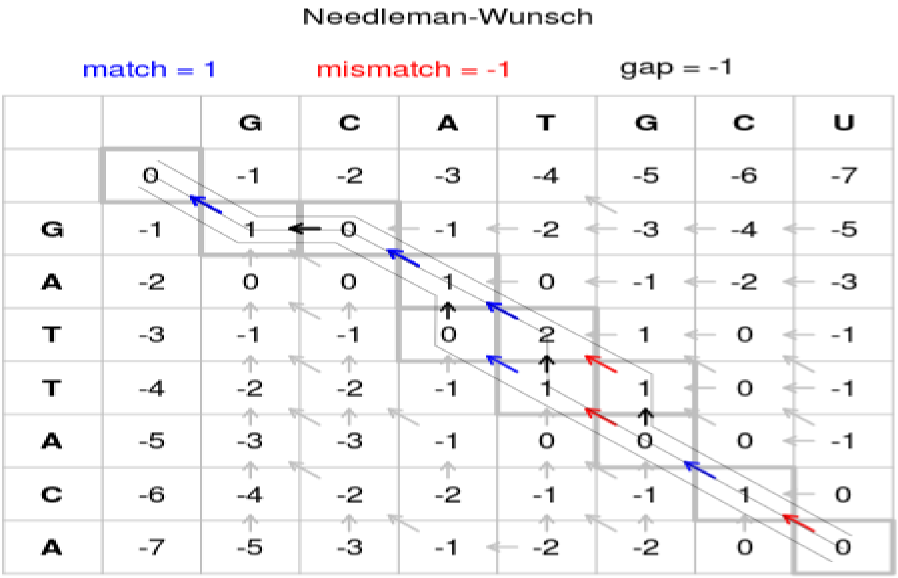
In reality, you don't store letters in the matrix as shown in Figure 3-1. Each cell of the matrix actually contains two values: a score and a pointer. The score is derived from the scoring scheme

Here, this means +1 or -1, but when aligning biological sequences, the values come from a scoring matrix (a topic of the next chapter). The pointer is a directional indicator (an arrow) that points up, left, or diagonally up and left. The pointer navigates the matrix, and its use will become clearer later in the chapter. Now, let's look at the algorithm in detail. There are three major phases: initialization, fill, and trace-back [14].

The Needleman-Wunch algorithm is an algorithm used in bioinformatics to align protein or nucleotide sequence [15].

It was one of the first applications of dynamic programming to compare biological sequences. The algorithm was developed by Saul B. Needleman and Christian D. Wunsch and published in 1970. The algorithm essentially divides a large problem (e.g. the full sequence) into a series of smaller problems, and it uses the solutions to the smaller problems to find an optimal solution to the larger problem. It is also sometimes referred to as the optimal matching algorithm and the global alignment technique. The Needleman–Wunsch algorithm is still widely used for optimal global alignment, particularly when the quality of the global alignment is of the utmost importance. The algorithm

Assigns a score to every possible alignment, and the purpose of the algorithm is to find all possible alignments having the highest score [15].


Needleman-Wunsch

match = 1 mismatch = -1 gap = -1

		G	C	A	T	G	C	U
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6	-4	-2	-2	-1	-1	1	0
A	-7	-5	-3	-1	-2	-2	0	0

Figure: 2.2

Score: 0 // 2 gaps, score -2

Alignment: +GCATGCUResults:

```
Sequences    Best alignments
---------    ----------------------
GCATGCU      GCATG-CU      GCA-TGCU      GCAT-GCU
GATTACA      G-ATTACA      G-ATTACA      G-ATTACA
```

Interpretation of the initialization step:

One can interpret the leftmost column in the above figure like this (putting a "handle" before each sequence, annotated as + here):

```
Alignment:  +GCATGCU
            +GATTACA
Score:      0  // Handle matches handle, doesn't win any score
```

```
Alignment:  +GCATGCU
           +GATTACA
Score:      0  // 1 gap,  score -1

Alignment:  +GCATGCU


         +GATTACA
Score:      0  // 3 gaps, score -3

Alignment:  +GCATGCU
        +GATTACA
Score:      0  // 4 gaps, score -4

...
```

The same thing can be done for the uppermost row.

Smith-Waterman Algorithm: The Smith–Waterman algorithm performs local sequence alignment; that is, for determining similar regions between two strings of nucleic acid sequences or protein sequences. Instead of looking at the entire sequence, the Smith–Waterman algorithm compares segments of all possible lengths and optimizes the similarity measure [15].

The algorithm was first proposed by Temple F. Smith and Michael S. Waterman in 1981.[1] Like the Needleman–Wunsch algorithm, of which it is a variation, Smith–Waterman is a dynamic programming algorithm. As such, it has the desirable property that it is guaranteed to find the optimal local alignment with respect to the scoring system being used (which includes the substitution matrix and the gap-scoring scheme). The main difference to the

Needleman–Wunsch algorithm is that negative scoring matrix cells are set to zero, which renders the (thus positively scoring) local alignments visible. Trace back procedure starts at the highest scoring matrix cell and proceeds until a cell with score zero is encountered, yielding the highest scoring local alignment [16].

2.6 Working of Smith-Waterman Algorithm:

Initialization of Matrix

The basic steps for the algorithm are similar to that of Needleman-Wunsch algorithm. The steps are:

1. Initialization of a matrix.

2. Matrix Filling with the appropriate scores.

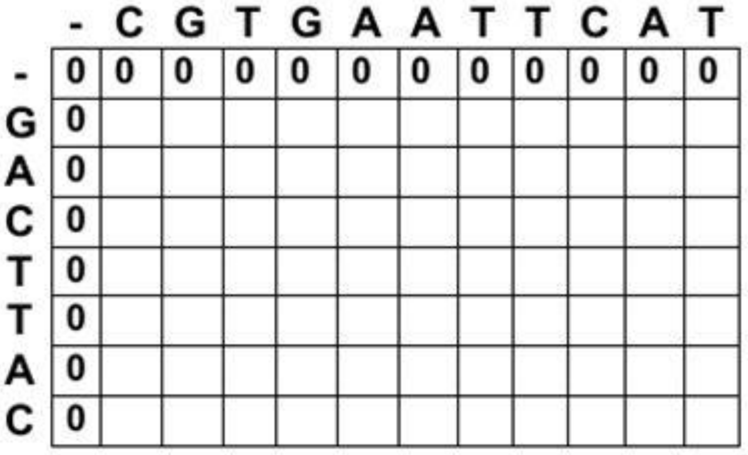
3. Trace back the sequences for a suitable alignment.

To study the Local sequence alignment consider the given below sequences.

CGTGAATTCAT (sequence #1 or A)

GACTTAC (sequence #2 or B)

The two sequences are arranged in a matrix form with A+1columns and B+1rows. The values in the first row and first column are set to zero as shown in Figure 1 [16].



	-	C	G	T	G	A	A	T	T	C	A	T
-	0	0	0	0	0	0	0	0	0	0	0	0
G	0											
A	0											
C	0											
T	0											
T	0											
A	0											
C	0											

Figure 2.3: Initialization of Matrix

2.6.1 Matrix Filling

The second and crucial step of the algorithm is filling the entire matrix, so it is more important to know the neighbor values (diagonal, upper and left) of the current cell to fill each and every cell [16].

$$M_{i,j} = Maximum\ \left[M_{i-1,j-1} + \ S_{i,j,}\ M_{i,j-1} + W,\ M_{i-1,j} + W, 0\right]$$

As per the assumptions stated earlier, fill the entire matrix using the assumed scoring schema and initial values. One can fill the 1st row and 1st column with the scoring matrix as follows.

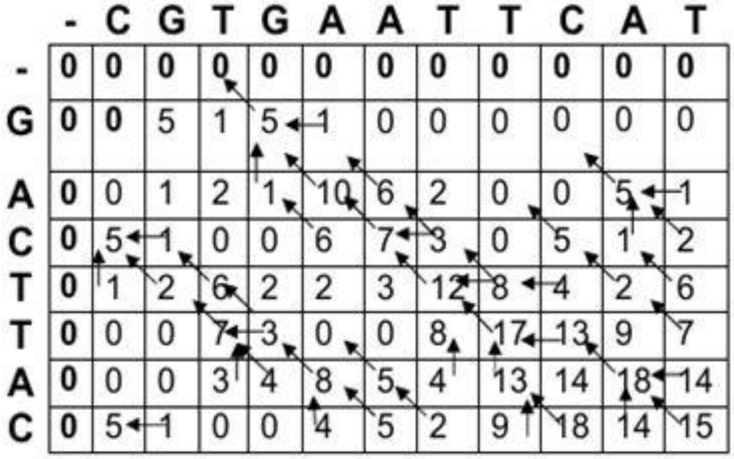
The first residue (nucleotides or amino acids) in both sequences is ‘C’ and ‘G’, the matching score or the mismatching score is going to be added the neighboring value which is diagonally located i.e. 0. The upper and left values are added to the gap penalty score from the matrix. So the scoring schema equation can be shown as follows [17].

$$\begin{aligned} M_{1,1} &= Maximum\ [M_{0,0} + S_{1,1},\ M_{1,0} + W,\ M_{0,1} + W,\ 0] \\ &= Maximum\ [0(-3), 0 + (-4), 0 + (-4), 0] \\ &= Maximum\ [-3, -4, -4, 0] \\ &= 0 \end{aligned}$$

From the above calculations the maximum value obtained is 0. Finding the maximum value for $M_{i,j}$ position, one can notice that there is no chance to see any negative values in the matrix, since we are taking 0 as lowest value.

After filling the matrix, keep the pointer back to the cell from where the maximum score has been determined. In the similar fashion fill all the values of the matrix of the cell [17].

For the example the matrix can be filled is shown in Figure [17].



	-	C	G	T	G	A	A	T	T	C	A	T
-	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	5	1	5	1	0	0	0	0	0	0
A	0	0	1	2	1	10	6	2	0	0	5	1
C	0	5	1	0	0	6	7	3	0	5	1	2
T	0	1	2	6	2	2	3	12	8	4	2	6
T	0	0	0	7	3	0	0	8	17	13	9	7
A	0	0	0	3	4	8	5	4	13	14	18	14
C	0	5	1	0	0	4	5	2	9	18	14	15

Figure 2.4: Matrix filling with back pointers

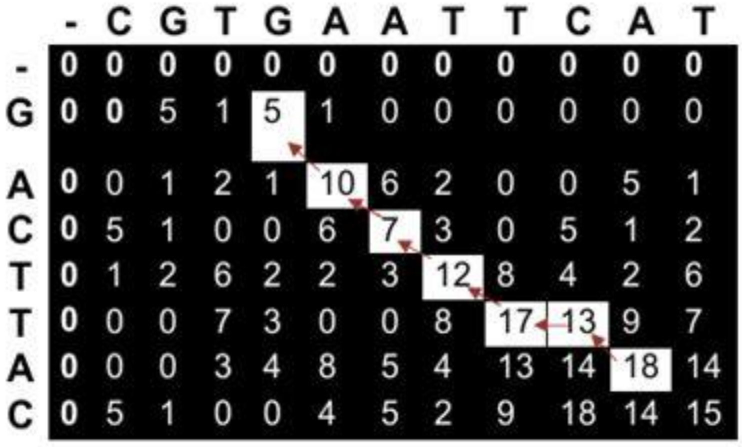
Each cell is back pointed by one or more pointers from where the maximum score has been obtained [17].

2.7 Trace backing the sequences for an optimal alignment:

The final step for the appropriate alignment is trace backing, prior to that one needs to find out the maximum score obtained in the entire matrix for the local alignment of the sequences. It is possible that the maximum scores can be present in more than one cell, in that case there may be possibility of two or more alignments, and the best alignment by scoring it [18].

In this example we can see the maximum score in the matrix as 18, which is found in two positions that lead to multiple alignments, so the best alignment has to be found.

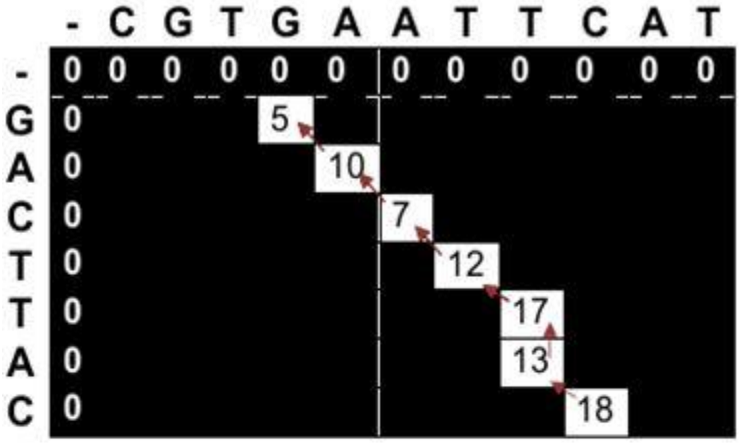
So the trace back begins from the position which has the highest value, pointing back with the pointers, thus find out the possible predecessor, then move to next predecessor and continue until we reach the score 0 [18].



	-	C	G	T	G	A	A	T	T	C	A	T
-	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	5	1	5	1	0	0	0	0	0	0
A	0	0	1	2	1	10	6	2	0	0	5	1
C	0	5	1	0	0	6	7	3	0	5	1	2
T	0	1	2	6	2	2	3	12	8	4	2	6
T	0	0	0	7	3	0	0	8	17	13	9	7
A	0	0	0	3	4	8	5	4	13	14	18	14
C	0	5	1	0	0	4	5	2	9	18	14	15

Figure 2.5: Trace back of first possible alignment

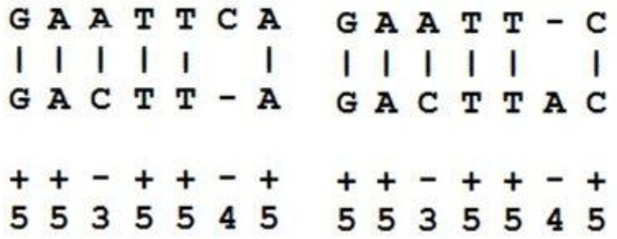
It is possible to find two pointers pointing out from one cell, where both ways (alignments) can be considered, best one is found by scoring and finding maximum score among them.



	-	C	G	T	G	A	A	T	T	C	A	T
-	0	0	0	0	0	0	0	0	0	0	0	0
G	0				5							
A	0					10						
C	0						7					
T	0							12				
T	0								17			
A	0								13			
C	0									18		

Figure 2.6: Trace back of second possible alignment

Thus a local alignment is obtained and one can see the possible alignments as in Figure 5.



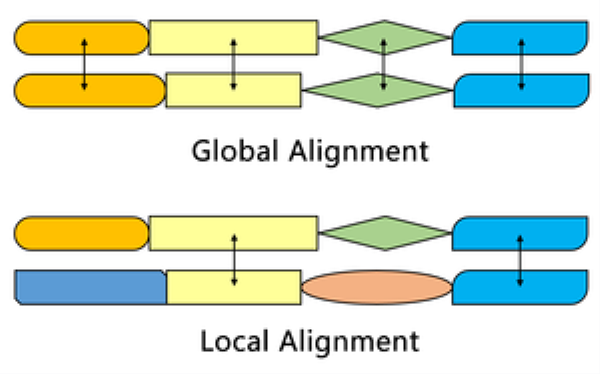
```
G A A T T C A    G A A T T - C
| | | | |   |    | | | | |   |
G A C T T - A    G A C T T A C

+ + - + + - +    + + - + + - +
5 5 3 5 5 4 5    5 5 3 5 5 4 5
```

The two alignments can be given with a score, for matching as +5 , mismatch as -3 and gap penalty as -4, sum up all the individual scores and the alignment which has maximum score after this can be taken as the best alignment.

By summing up the scores both of the alignments are giving the same as 18, so one can predict both alignments are the best [19].

2.8 Comparison with the Needleman–Wunsch algorithm



Global and local sequence alignment

The Smith–Waterman algorithm finds the segments in two sequences that have similarities while the Needleman–Wunsch algorithm aligns two complete sequences. Therefore, they serve different purposes. Both algorithms use the concepts of a substitution matrix, a gap penalty function, a scoring matrix, and a trace back process. Three main differences are [20]:

One of the most important distinctions is that no negative score is assigned in the scoring system of the Smith–Waterman algorithm, which enables local alignment. When any element has a score lower than zero, it means that the sequences up to this position have no similarities; this element will then be set to zero to eliminate influence from previous alignment. In this way, calculation can continue to find alignment in any position afterwards.

The initial scoring matrix of Smith–Waterman algorithm enables the alignment of any segment of one sequence to an arbitrary position in the other sequence. In Needleman–Wunsch algorithm, however, end gap penalty also needs to be considered in order to align the full sequences [20].

Chapter 3

METHODOLOGY

3.1 Proposed Method:

To reduce above problems, we build a model combination of (Cells Binding, Cancer cells detection, Reduction or alignment of cancer cells and filtering of mutated DNA) that will solve identified problems. It does not have to use algorithm separately and works together. This method will reduce time and cost.

In this method we use the following parameters:

1. KRAS gene
2. Short Read Alignment
3. Filtering
4. Mutations in DNA (substitution, deletion, insertion)

For this method, we use Pymol and Anaconda software that can control the cancer cells in a feasible tolerance metrics. The specialty of this model, it is a hybrid model which is a mix interpretation of Cells binding, Detection, Reduction or Sequence Alignment and filtering method. This process can make a good help to the biological researchers and others who works in Bioinformatics arena [21].

3.2 Related Algorithm:

Algorithm: Needleman- Wunch model

1. A scoring function (σ): defines the score to give to a substitution mutation eg. -1 for a match, -1 for mismatch

2. A gap penalty: defines the score to give to an insertion or deletion mutation, eg. -1
3. A recurrence relation: defines what actions we repeat at each iteration(step) of the algorithm; for N-W this is:

$$T(i, j) = \max \begin{cases} T(i-1, j-1) + (S(i), S(j)) \\ T(i-1, j) + \text{gap penalty} \\ T(i, j-1) + \text{gap penalty} \end{cases}$$

4. Fill up a matrix (table) T using the recurrence relation
5. The traceback step: use the filled in matrix T to work out the best alignment

Algorithm: Smith- Waterman model

Let A = a1, a2,……a_n and B = b1,b2,……b_n be the sequences to be aligned, where n and m are the lengths of A and B respectively.

1. Determine the substitution matrix and the gap penalty scheme.
   - s(a, b) - Similarity score of the elements that constituted the two sequences
   - W_k - The penalty of a gap that has length k

2. Construct a scoring matrix H and initialize its first row and first column. The size of the scoring matrix is (n + 1) * (m + 1). Note the 0-based indexing.

$H_{k0} = H_{0l} = 0$ for $0 < k < n$ and $0 < l < m$

3. Fill the scoring matrix using the equation below.

$$H_{ij} = \max \begin{cases} H_{i-1, j-1} + s(a_i, b_j), \\ \max_{k>1} \{ H_{i-k,j} - W_k \}; \quad (1 < i < n, 1 < j < m) \\ \max_{l>1} \{ H_{i,j-1} - Wl \}; \\ 0 \end{cases}$$

Where

$H_{i-1,j-1} + s(a_i, b_j)$ is the score of aligning a_i and b_j,

$H_{i-k,j}$-Wk is the score if a_i is at the end of a gap of length k,

$H_{i,j-1}$-Wl is the score if b_j is at the end of a gap of length l,

0 means there is no similarity up to a_i and b_j.

4. Traceback. Strarting at the highest score in the scoring matrix H and ending at a matrix cell.

3.3 Algorithm: Proposed model

Data : Automatically classified label dataset

Result : Sequential classification into align and diagonal

Step 1: Import dataset

Step 2: Build integrates executable

Binding different type of label data

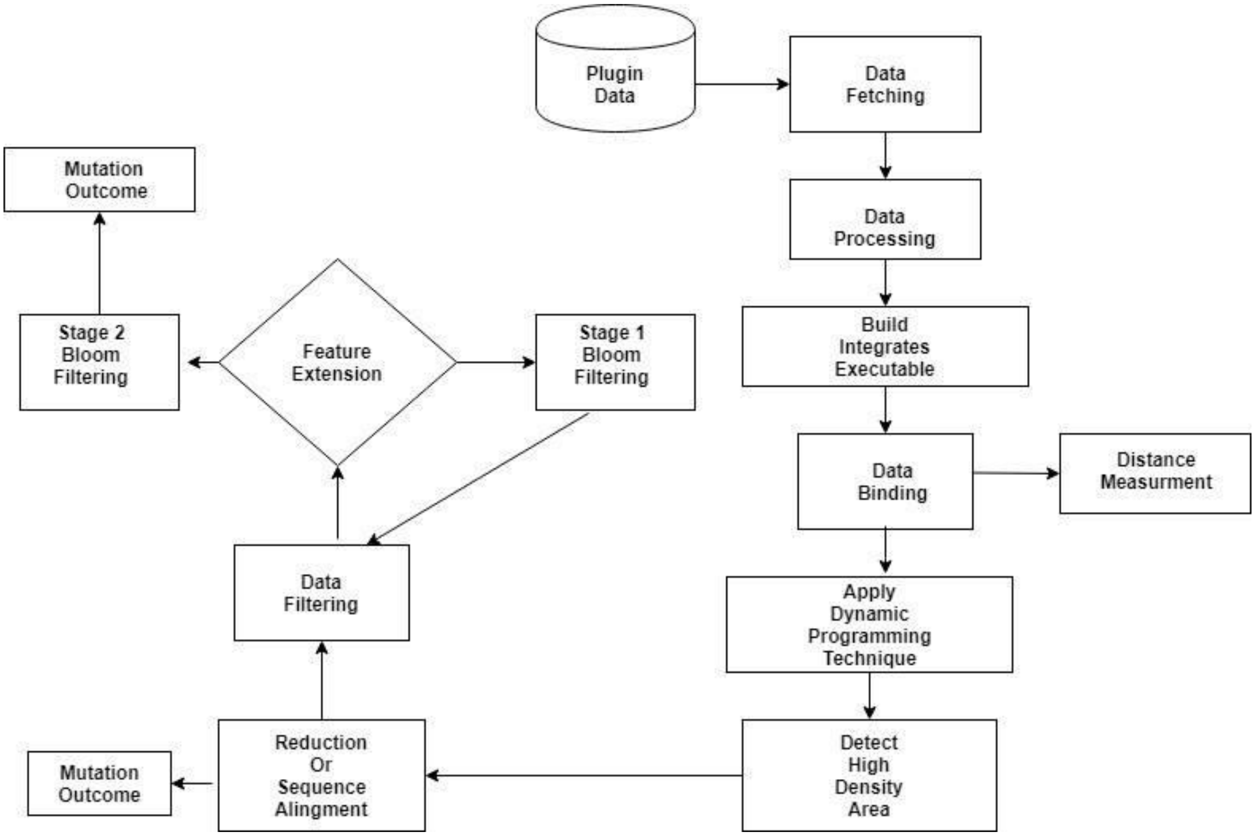
A = a1,a2,…………ai and B = b1,b2,…………bj

Where i and j are the length of A and B respectively

Select and detect the changes of sequences

```
            Distance measurement
Step 3: Mutation Detection
            Detect high density area with N-W algorithm pair wise method
            Define scoring function(σ), gap penalty and recurrence relation
            that computed T[i, j] :
                              | T[i,j] = max{ 0;
                              | T[i-1, j-1] + sub(A[i],B[j]);
              T = (i, j)      |  T[i-1, j] +del(A[i]);
                              | T[i, j-1] + ins (B[j]);
                              | T[i-1, j] + gap penalty
                              | T[i, j-1] + gap penalty
Step 4: Mutagenesis Process
             Mutant identification and adding
             Align the dataset by applying the mutant and respective adding
Step 5: Data Filtering
            Generate spaced seeds
            For each spaced seeds create hash value/table
            Check the value and insert into Stage1 filtering or Stage2 filtering
            Testing the model with best classifier
            Real time testing of data
Step 6: End
```

3.4 Flowchart of the whole work:



Plug-in Data: A plug-in or plugin is a software component that adds a specific feature to an existing computer program. When a program supports plug-ins, it enables customization[11].

Data Fetching: In computer technology systems, fetch has a several meanings related to getting, reading, or moving data objects. It is a feature is an object that can have a geographic location and other properties.

Processing: The carrying out of operations on data, especially by a computer, to retrieve, Transform, or classify information. Generally, it is the collection and manipulation of items of data to produce meaning information. In this sense, it can be considered a subset of information processing.

Build Integrates Executable: Each check-in is then verified by an automated build, allowing teams to detect problems early. Data Integration platform has built-in processors to run and schedule any executable application. Each integration is verified by an automated build (including test) to detect integration errors as quickly as possible.

Data Binding: In computer programming, data binding is a general technique that binds data sources from the provider and consumer together and synchronizes them.

Measurement: Measurement is the process observing and recording the observations that are collected as part of a research effort.

Global Alignment (Pair-wise) Method: Global alignments attempt to align every residue in every sequence, and are most useful when the sequences in the query set are similar and of roughly equal size. It is a general global alignment technique which is based on dynamic programming.

Detection High Density Area: The action or process of identifying the presence of something concealed which detect the high density atomic area in cancer cells.

Reduction / Sequence Alignment: This program allows you to reduce the redundancy in a set of aligned or unaligned sequences.

Data Filtering: Filtering data in a spreadsheet means to set conditions so that only certain data is displayed; it is done to make it easier to focus on specific information in a large dataset or table of data. It is a process of choosing a smaller part of dataset and using that subset for viewing or analysis. In some other cases, data filters work to prevent wider access to sensitive information.

Feature Extension: The Extension features files specify the different requirements for extension feature availability. An extension feature can be any component of extension capabilities. Most notably, this includes extension APIs, but there are also more structural or behavioral features, such as web accessible resources or event pages [23].

Stage1 Filtering: Data filtering is a process of choosing a smaller part of dataset and using that subset for viewing or analysis. In data filtering, stage1 filtering is a counting system.

Stage2 Filtering: Data filtering is a process of choosing a smaller part of dataset and using that subset for viewing or analysis. In data filtering, stage2 filtering is records to all observed data.

Chapter 4

PERFORMANCE ANALYSIS & COMPARISION

4.1 Experimental Analysis:

We use PyMol software helps to analyses data which can control the cancer cells in a feasible tolerance metrics and Python shell to filtrate the mutated gene. So we implement a method which is a mix interpretation and we have received our aspirations of reduction was 41.8% before our method was implemented and after implementing our method the result we got 55.7% in alignment or mutagenesis process. After comparing the experimental result of our method with other method we got better result. In filtering process after implementing our method it is possible to filtrate the mutated genes and replace the identified problems [24].

4.2 Dataset: A data set (or dataset) is a collection of data. Most commonly a data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set in question. The data set lists values for each of the variables, such as height and weight of an object, for each member of the data set. Each value is known as a datum. The data set may comprise data for one or more members, corresponding to the number of rows [25].

4.3 Labeled Dataset:

A data set (or dataset) is a collection of data. Most commonly a data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set in question. Labeled data is a group of samples that have been tagged with one or more labels. Labeling typically takes a set of unlabeled data and augments each piece of that unlabeled data with meaningful tags that are informative.

After obtaining a labeled dataset, machine learning models can be applied to the data so that new unlabeled data can be presented to the model and a likely label can be guessed or predicted for that piece of unlabeled data.

Labeled data is a designation for pieces of data that have been tagged with one or more labels identifying certain properties or characteristics, or classifications or contained objects. Labels make that data specifically useful in certain types of machine learning known as supervised machine learning setups [26].

4.4 Unlabeled Dataset: Unlabeled data is a designation for pieces of data that have not been tagged with labels identifying characteristics, properties or classifications. Unlabeled data is typically used in various forms of machine learning. In types of machine learning called unsupervised machine learning, the machine learning program operates by evaluating sets of unlabeled data. Because the data does not have labels, the machine learning program has to identify each data piece on its properties and characteristics.

Labeled data is a group of samples that have been tagged with one or more labels. Labeling typically takes a set of unlabeled data and augments each piece of that unlabeled data with meaningful tags that are informative [26].

4.5 Working procedure of Binding cells: At first we take one cancer cell (2RKB), two normal cells (1HRJ), (4WHC) and two KRAS genes (5O2S), (5O2T).

Here we bind together those cells by using PYMOL software.

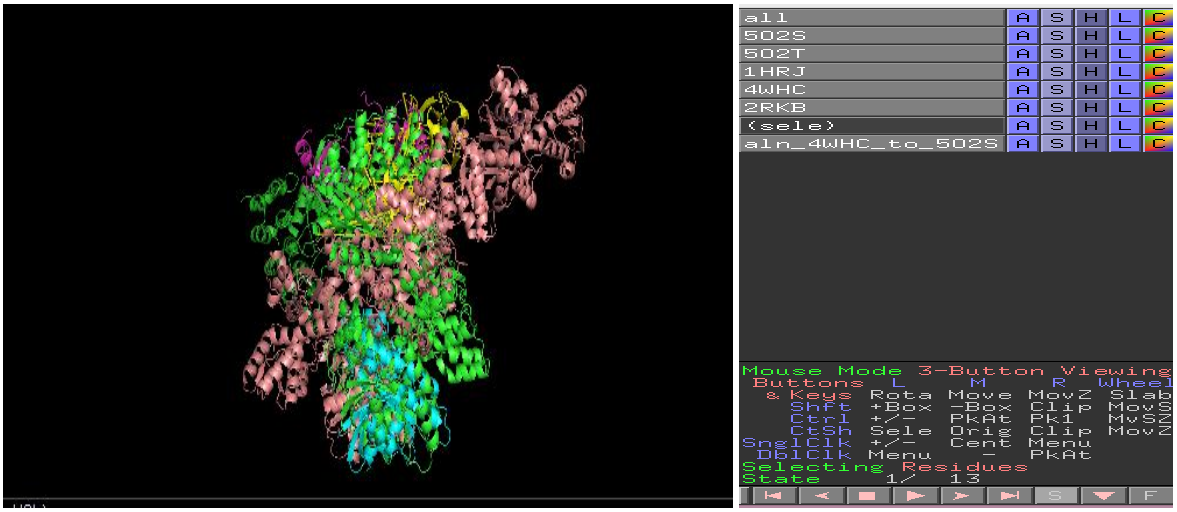
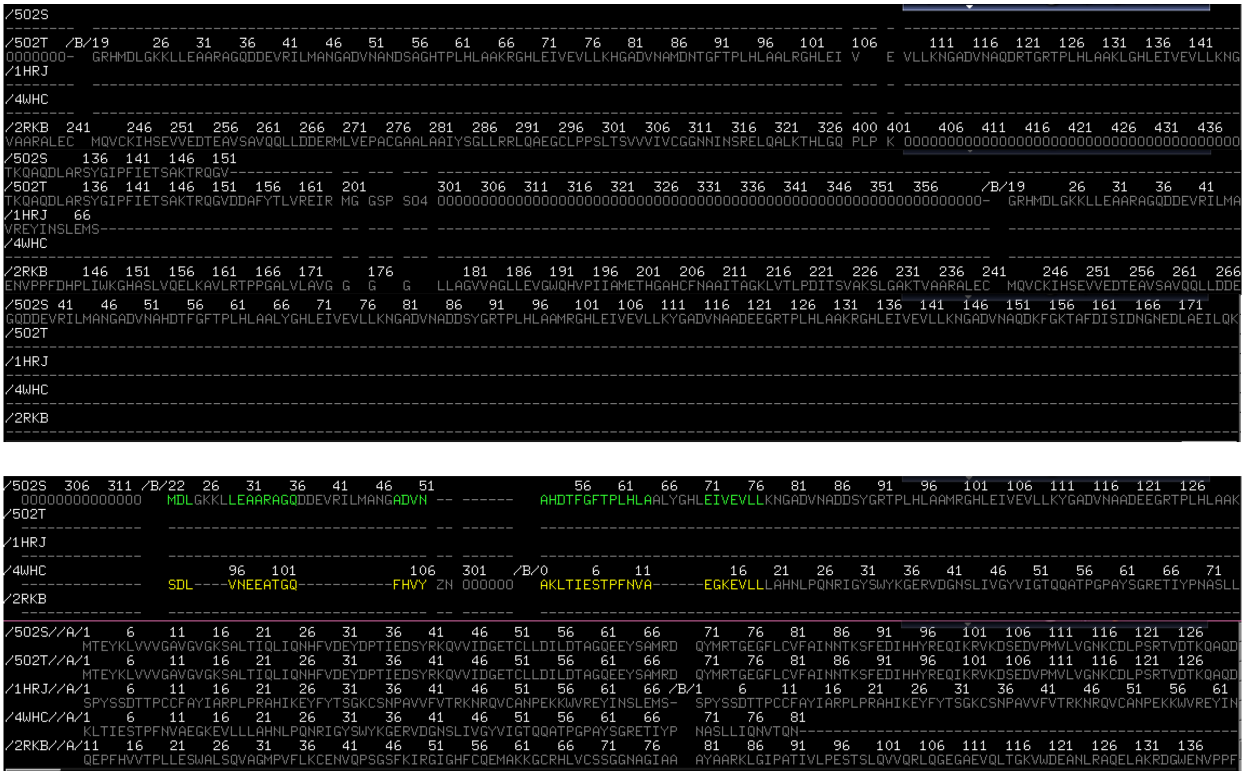


Fig 4.1: Binding cell (2+2+1)

Here Sequence of 2+2+1 binding:



```
/502S
---------  ---------------------------------------------------------------------------- --- - -------------------------------
/502T  /B/19      26   31   36   41   46   51   56   61   66   71   76   81   86   91   96   101   106      111  116  121  126  131  136  141
0000000-  GRHMDLGKKLLEAARAGQDDEVRILMANGADVNANDSAGHTPLHLAAKRGHLEIVEVLLKHGADVNAMDNTGFTPLHLAALRGHLEI V   E VLLKNGADVNAQDRTGRTPLHLAAKLGHLEIVEVLLKNG
/1HRJ
---------  ---------------------------------------------------------------------------- --- - -------------------------------
/4WHC
---------  ---------------------------------------------------------------------------- --- - -------------------------------
/2RKB  241     246  251  256  261  266  271  276  281  286  291  296  301  306  311  316  321  326 400 401    406  411  416  421  426  431  436
VAARALEC  MQVCKIHSEVVEDTEAVSAVQQLLDDERMLVEPACGAALAAIYSGLLRRLQAEGCLPPSLTSVVVIVCGGNNINSRELQALKTHLGQ PLP K 00000000000000000000000000000000000000
/502S    136  141  146  151
TKQAQDLARSYGIPFIETSAKTRQGV------------ -- --- --- ----------------------------------------------------  ----------------------
/502T    136  141  146  151  156  161  201        301  306  311  316  321  326  331  336  341  346  351  356        /B/19     26   31   36   41
TKQAQDLARSYGIPFIETSAKTRQGVDDAFYTLVREIR MG GSP SO4 0000000000000000000000000000000000000000000000000000000-  GRHMDLGKKLLEAARAGQDDEVRILMA
/1HRJ   66
VREYINSLEMS--------------------------- -- --- --- ----------------------------------------------------  ----------------------
/4WHC
-------------------------------------- -- --- --- ----------------------------------------------------  ----------------------
/2RKB    146  151  156  161  166  171      176        181  186  191  196  201  206  211  216  221  226  231  236  241      246  251  256  261  266
ENVPPFDHPLIWKGHASLVQELKAVLRTPPGALVLAVG G   G    G   LLAGVVAGLLEVGWQHVPIIAMETHGAHCFNAAITAGKLVTLPDITSVAKSLGAKTVAARALEC  MQVCKIHSEVVEDTEAVSAVQQLLDDE
/502S 41   46   51   56   61   66   71   76   81   86   91   96   101  106  111  116  121  126  131  136  141  146  151  156  161  166  171
GQDDEVRILMANGADVNAHDTFGFTPLHLAALYGHLEIVEVLLKNGADVNADDSYGRTPLHLAAMRGHLEIVEVLLKYGADVNAADEEGRTPLHLAAKRGHLEIVEVLLKNGADVNAQDKFGKTAFDISIDNGNEDLAEILQK
/502T
-----------------------------------------------------------------------------------------------------------------------------------------
/1HRJ
-----------------------------------------------------------------------------------------------------------------------------------------
/4WHC
-----------------------------------------------------------------------------------------------------------------------------------------
/2RKB
-----------------------------------------------------------------------------------------------------------------------------------------
```

```
/502S  306  311 /B/22  26   31   36   41   46   51                56   61   66   71   76   81   86   91   96   101  106  111  116  121  126
  00000000000000   MDLGKKLLEAARAGQDDEVRILMANGADVN -- ------   AHDTFGFTPLHLAALYGHLEIVEVLLKNGADVNADDSYGRTPLHLAAMRGHLEIVEVLLKYGADVNAADEEGRTPLHLAAK
/502T
  --------------   ------------------------------ -- ------   --------------------------------------------------------------------------------
/1HRJ
  --------------   ------------------------------ -- ------   --------------------------------------------------------------------------------
/4WHC                    96   101                106   301   /B/0     6    11         16   21   26   31   36   41   46   51   56   61   66   71
  --------------   SDL----VNEEATGQ-----------FHVY ZN 000000   AKLTIESTPFNVA------EGKEVLLLAHNLPQNRIGYSWYKGERVDGNSLIVGYVIGTQQATPGPAYSGRETIYPNASLL
/2RKB
  --------------   ------------------------------ -- ------   --------------------------------------------------------------------------------
/502S//A/1    6    11   16   21   26   31   36   41   46   51   56   61   66      71   76   81   86   91   96   101  106  111  116  121  126
         MTEYKLVVVGAVGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAGQEEYSAMRD  QYMRTGEGFLCVFAINNTKSFEDIHHYREQIKRVKDSEDVPMVLVGNKCDLPSRTVDTKQAQD
/502T//A/1    6    11   16   21   26   31   36   41   46   51   56   61   66      71   76   81   86   91   96   101  106  111  116  121  126
         MTEYKLVVVGAVGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAGQEEYSAMRD  QYMRTGEGFLCVFAINNTKSFEDIHHYREQIKRVKDSEDVPMVLVGNKCDLPSRTVDTKQAQD
/1HRJ//A/1    6    11   16   21   26   31   36   41   46   51   56   61   66 /B/1    6    11   16   21   26   31   36   41   46   51   56   61
         SPYSSDTTPCCFAYIARPLPRAHIKEYFYTSGKCSNPAVVFVTRKNRQVCANPEKKWVREYINSLEMS-  SPYSSDTTPCCFAYIARPLPRAHIKEYFYTSGKCSNPAVVFVTRKNRQVCANPEKKWVREYIN
/4WHC//A/1    6    11   16   21   26   31   36   41   46   51   56   61   66      71   76   81
         KLTIESTPFNVAEGKEVLLLAHNLPQNRIGYSWYKGERVDGNSLIVGYVIGTQQATPGPAYSGRETIYP  NASLLIQNVTQN------------------------------------------------
/2RKB//A/11   16   21   26   31   36   41   46   51   56   61   66   71   76      81   86   91   96   101  106  111  116  121  126  131  136
         QEPFHVVTPLLESWALSQVAGMPVFLKCENVQPSGSFKIRGIGHFCQEMAKKGCRHLVCSSGGNAGIAA  AYAARKLGIPATIVLPESTSLQVVQRLQGEGAEVQLTGKVWDEANLRAQELAKRDGWENVPPF
```

Figure: 4.2: Detection sequences

4.6 Working Procedure of Detection:

Here we identify the density of cancer atoms in cancer cell

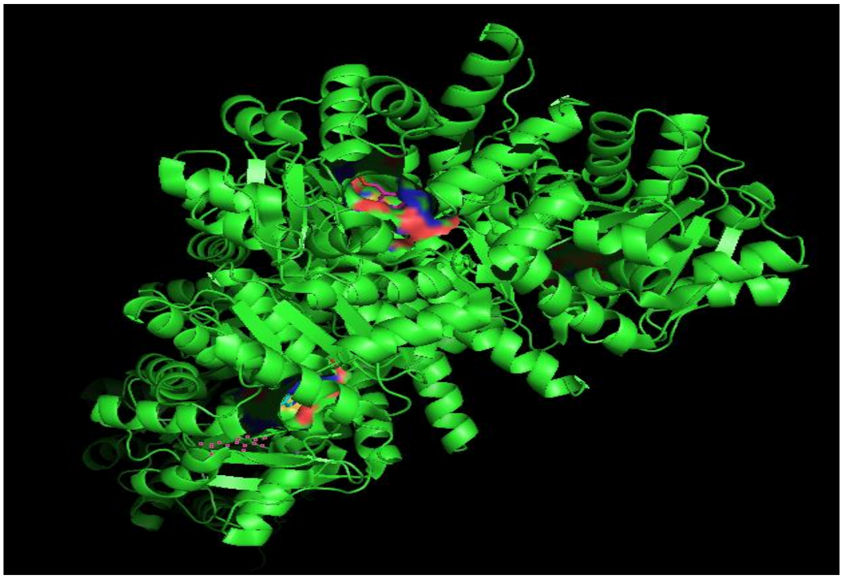


Figure 4.3: Detection of cells

After identify density of cancer cell, we bind that cancer cell with one KRAS (5O2T) and normal cell (1HRJ) then we got our result

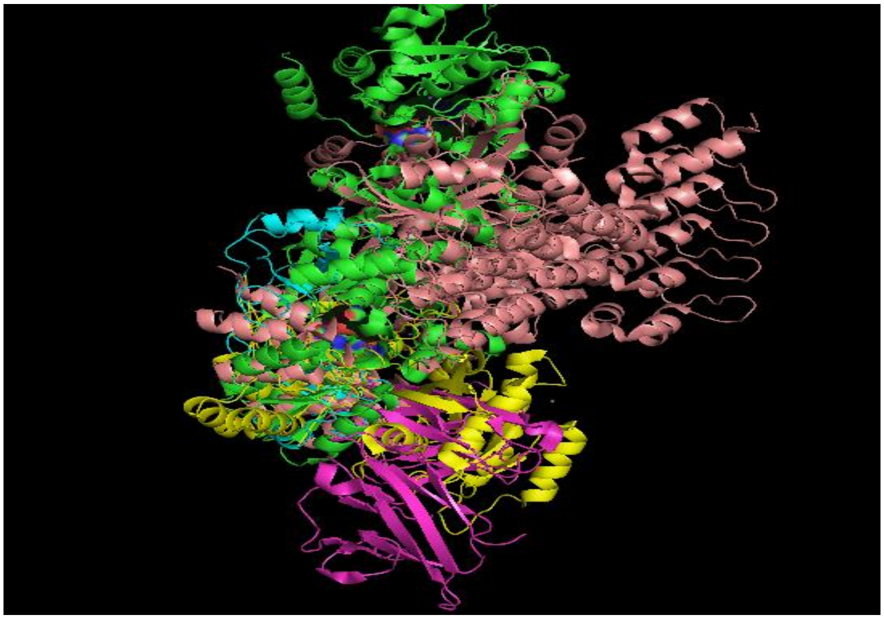
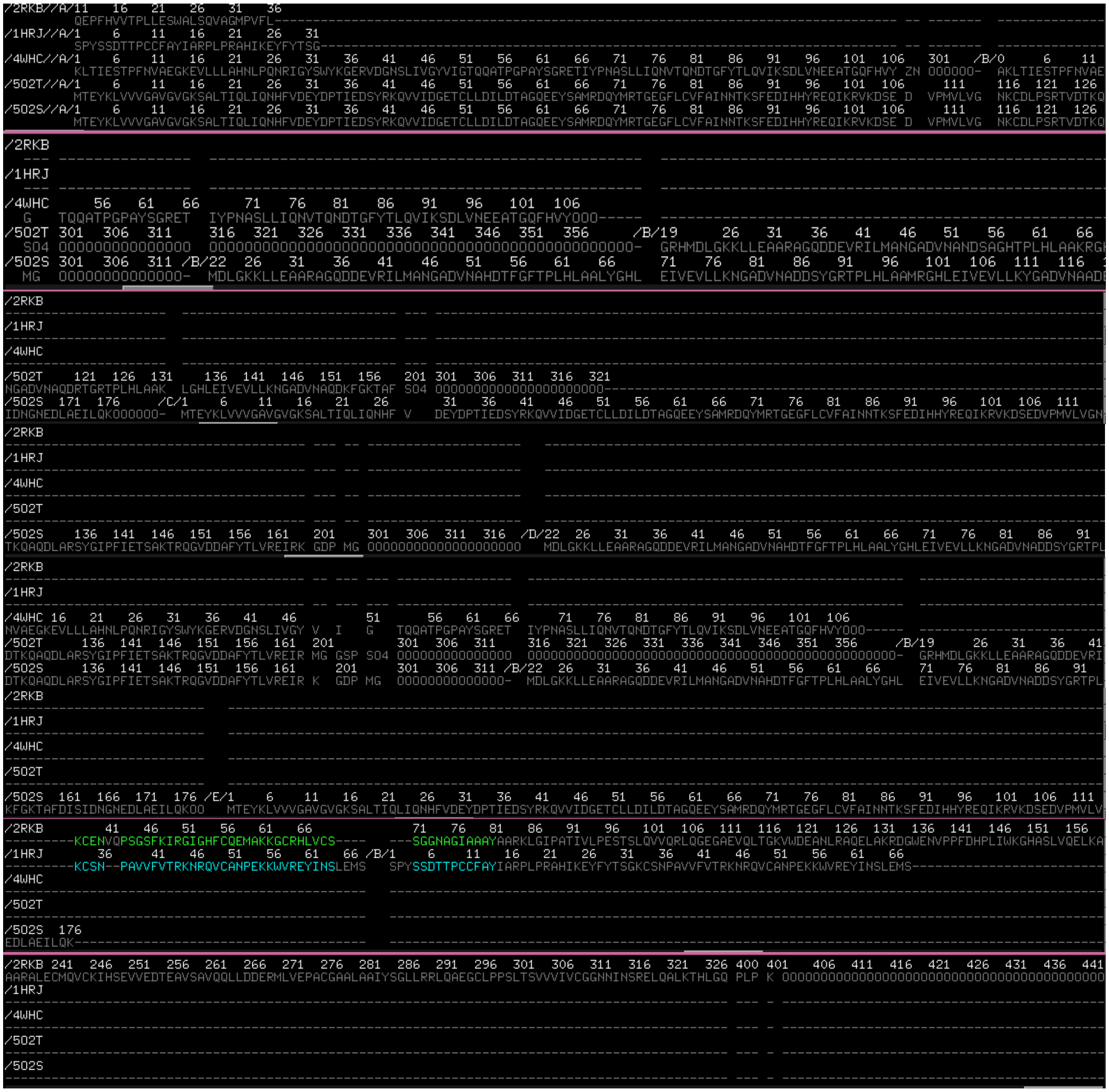


Figure 4.4: Data binding

And their sequence is given below:



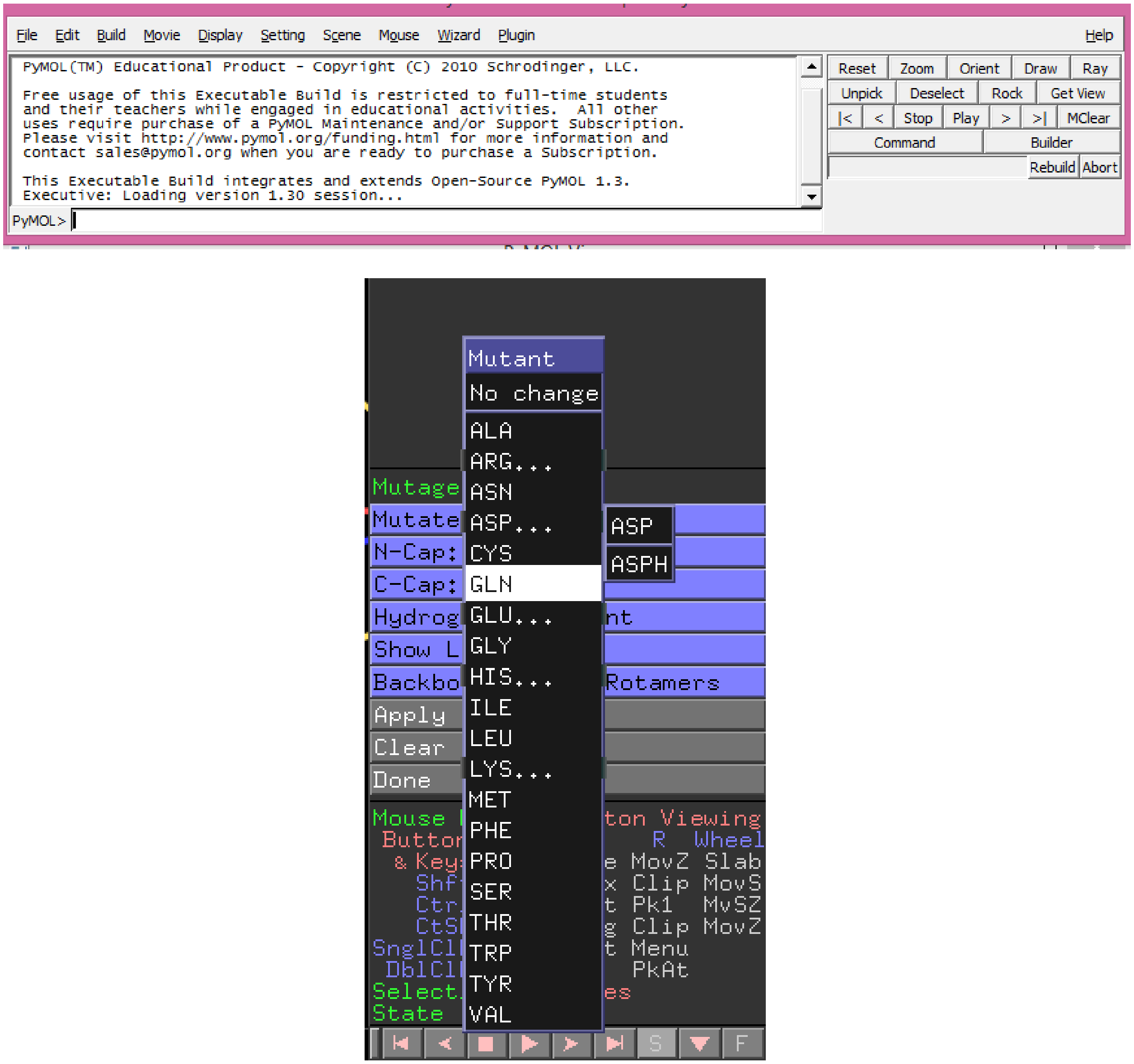
```
/2RKB//A/11   16    21    26    31    36
         QEPFHVVTPLLESWALSQVAGMPVFL-------------------------------------------------------------------------------- -- -------  --------------
/1HRJ//A/1     6     11    16    21    26    31
         SPYSSDTTPCCFAYIARPLPRAHIKEYFYTSG-------------------------------------------------------------------------- -- -------  --------------
/4WHC//A/1     6     11    16    21    26    31    36    41    46    51    56    61    66    71    76    81    86    91    96   101   106    301   /B/0      6     11
         KLTIESTPFNVAEGKEVLLLAHNLPQNRIGYSWYKGERVDGNSLIVGYVIGTQQATPGPAYSGRETIYPNASLLIQNVTQNDTGFYTLQVIKSDLVNEEATGQFHVY ZN OOOOOO-  AKLTIESTPFNVAE
/502T//A/1     6     11    16    21    26    31    36    41    46    51    56    61    66    71    76    81    86    91    96   101   106       111     116  121  126
         MTEYKLVVVGAVGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAGQEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHHYREQIKRVKDSE D  VPMVLVG  NKCDLPSRTVDTKQ
/502S//A/1     6     11    16    21    26    31    36    41    46    51    56    61    66    71    76    81    86    91    96   101   106       111     116  121  126
         MTEYKLVVVGAVGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAGQEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHHYREQIKRVKDSE D  VPMVLVG  NKCDLPSRTVDTKQ
/2RKB
   --- -----------------  -------------------------------------------------------  ----------------------------------------------------------
/1HRJ
   --- -----------------  -------------------------------------------------------  ----------------------------------------------------------
/4WHC     56    61    66      71    76    81    86    91    96   101   106
  G   TQQATPGPAYSGRET  IYPNASLLIQNVTQNDTGFYTLQVIKSDLVNEEATGQFHVYOOO-----  ----------------------------------------------------------
/502T 301   306   311      316   321   326   331   336   341   346   351   356      /B/19      26    31    36    41    46    51    56    61    66
  SO4 OOOOOOOOOOOOOOOOO  OOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOO-  GRHMDLGKKLLEAARAGQDDEVRILMANGADVNANDSAGHTPLHLAAKRGH
/502S 301   306   311 /B/22  26    31    36    41    46    51    56    61    66      71    76    81    86    91    96   101   106   111   116
  MG  OOOOOOOOOOOOOOOO-  MDLGKKLLEAARAGQDDEVRILMANGADVNAHDTFGFTPLHLAALYGHL  EIVEVLLKNGADVNADDSYGRTPLHLAAMRGHLEIVEVLLKYGADVNAAD
/2RKB
----------------------  ---------------------------  ---  --------------------------------------------------------------------------------------
/1HRJ
----------------------  ---------------------------  ---  --------------------------------------------------------------------------------------
/4WHC
----------------------  ---------------------------  ---  --------------------------------------------------------------------------------------
/502T    121  126  131     136  141  146  151  156   201 301  306  311  316  321
NGADVNAQDRTGRTPLHLAAK  LGHLEIVEVLLKNGADVNAQDKFGKTAF  SO4 OOOOOOOOOOOOOOOOOOOOOOO---------------------------------------------------------------
/502S  171  176      /C/1    6    11   16   21   26       31   36   41   46   51   56   61   66   71   76   81   86   91   96   101  106  111
IDNGNEDLAEILQKOOOOOO-  MTEYKLVVVGAVGVGKSALTIQLIQNHF V   DEYDPTIEDSYRKQVVIDGETCLLDILDTAGQEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHHYREQIKRVKDSEDVPMVLVGN
/2RKB
--------------------------------------- --- -- --------------------   ------------------------------------------------------------------------
/1HRJ
--------------------------------------- --- -- --------------------   ------------------------------------------------------------------------
/4WHC
--------------------------------------- --- -- --------------------   ------------------------------------------------------------------------
/502T
--------------------------------------- --- -- --------------------   ------------------------------------------------------------------------
/502S     136  141  146  151  156  161    201     301  306  311  316  /D/22  26   31   36   41   46   51   56   61   66   71   76   81   86   91
TKQAQDLARSYGIPFIETSAKTRQGVDDAFYTLVREIRK GDP MG OOOOOOOOOOOOOOOOOOOO   MDLGKKLLEAARAGQDDEVRILMANGADVNAHDTFGFTPLHLAALYGHLEIVEVLLKNGADVNADDSYGRTPL
/2RKB
--------------------------------------- -- --- --- ---------------  --------------------------------------------------  ------------------------
/1HRJ
--------------------------------------- -- --- --- ---------------  --------------------------------------------------  ------------------------
/4WHC 16   21   26   31   36   41   46         51       56   61   66      71   76   81   86   91   96   101  106
NVAEGKEVLLLAHNLPQNRIGYSWYKGERVDGNSLIVGY V  I   G   TQQATPGPAYSGRET  IYPNASLLIQNVTQNDTGFYTLQVIKSDLVNEEATGQFHVYOOO-----  ------------------------
/502T      136  141  146  151  156  161  201         301  306  311      316  321  326  331  336  341  346  351  356      /B/19       26   31   36   41
DTKQAQDLARSYGIPFIETSAKTRQGVDDAFYTLVREIR MG GSP SO4 OOOOOOOOOOOOOOO  OOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOO-  GRHMDLGKKLLEAARAGQDDEVRI
/502S      136  141  146  151  156  161       201     301  306  311 /B/22  26   31   36   41   46   51   56   61   66      71   76   81   86   91
DTKQAQDLARSYGIPFIETSAKTRQGVDDAFYTLVREIR K   GDP MG  OOOOOOOOOOOOOO-  MDLGKKLLEAARAGQDDEVRILMANGADVNAHDTFGFTPLHLAALYGHL  EIVEVLLKNGADVNADDSYGRTPL
/2RKB
--------------------------   ------------------------------------------------------------------------------------------------------------------
/1HRJ
--------------------------   ------------------------------------------------------------------------------------------------------------------
/4WHC
--------------------------   ------------------------------------------------------------------------------------------------------------------
/502T
--------------------------   ------------------------------------------------------------------------------------------------------------------
/502S  161  166  171  176 /E/1    6    11   16   21   26   31   36   41   46   51   56   61   66   71   76   81   86   91   96   101  106  111
KFGKTAFDISIDNGNEDLAEILQKOO   MTEYKLVVVGAVGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAGQEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHHYREQIKRVKDSEDVPMVLV
/2RKB        41   46   51   56   61   66             71   76   81   86   91   96   101  106  111  116  121  126  131  136  141  146  151  156
---------KCENVQPSGSFKIRGIGHFCQEMAKKGCRHLVCS----   ---SGGNAGIAAAYAARKLGIPATIVLPESTSLQVVQRLQGEGAEVQLTGKVWDEANLRAQELAKRDGWENVPPFDHPLIWKGHASLVQELKA
/1HRJ       36     41   46   51   56   61   66 /B/1    6    11   16   21   26   31   36   41   46   51   56   61   66
---------KCSN--PAVVFVTRKNRQVCANPEKKWVREYINSLEMS   SPYSSDTTPCCFAYIARPLPRAHIKEYFYTSGKCSNPAVVFVTRKNRQVCANPEKKWVREYINSLEMS-------------------------
/4WHC
-----------------------------------------------   ------------------------------------------------------------------------------------------
/502T
-----------------------------------------------   ------------------------------------------------------------------------------------------
/502S  176
EDLAEILQK--------------------------------------   ------------------------------------------------------------------------------------------
/2RKB 241  246  251  256  261  266  271  276  281  286  291  296  301  306  311  316  321  326 400 401   406  411  416  421  426  431  436  441
AARALECMQVCKIHSEVVEDTEAVSAVQQLLDDERMLVEPACGAALAAIYSGLLRRLQAEGCLPPSLTSVVVIVCGGNNINSRELQALKTHLGQ PLP K OOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOO
/1HRJ
--------------------------------------------------------------------------------------------- --- - -----------------------------------------
/4WHC
--------------------------------------------------------------------------------------------- --- - -----------------------------------------
/502T
--------------------------------------------------------------------------------------------- --- - -----------------------------------------
/502S
--------------------------------------------------------------------------------------------- --- - -----------------------------------------
```

Figure 4.5: Sequences of detection

4.7 Working Procedure of Reduction:

At first we go wizard for mutagenesis, after click in mutagenesis then we got mutate and click on mutate we found mutant. By using mutant we can change sequence. For changing sequence we choose any mutant and click on apply for apply these mutant [27].

Procedure is given below in the figure:



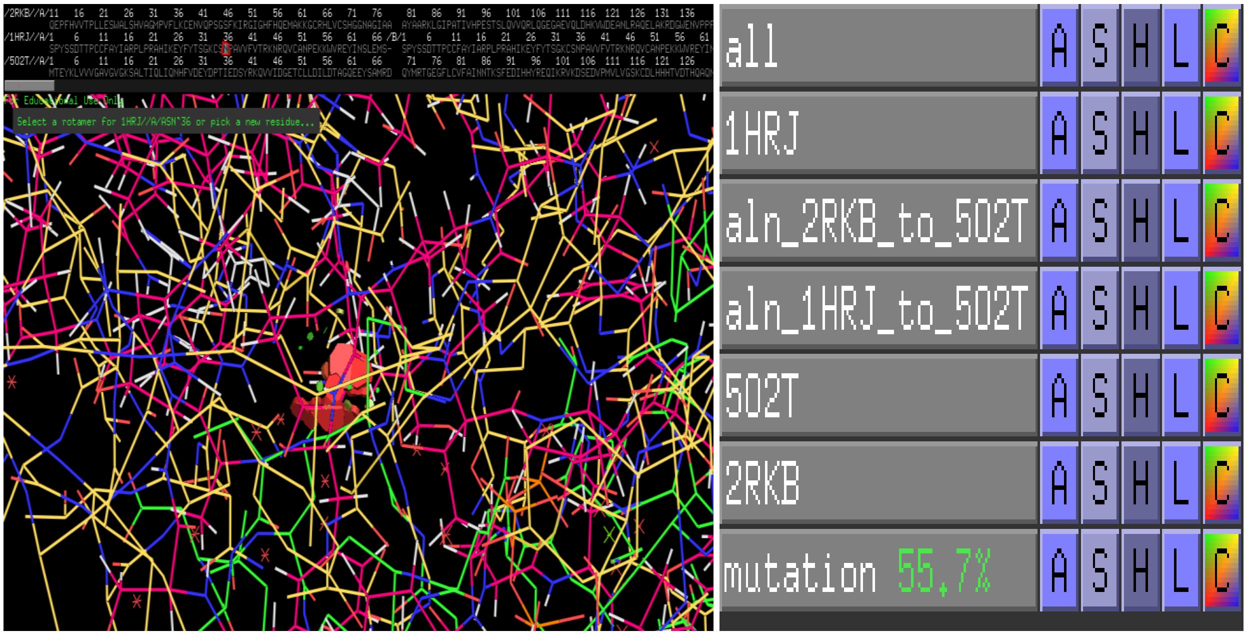
```
PyMOL(TM) Educational Product - Copyright (C) 2010 Schrodinger, LLC.

Free usage of this Executable Build is restricted to full-time students
and their teachers while engaged in educational activities.  All other
uses require purchase of a PyMOL Maintenance and/or Support Subscription.
Please visit http://www.pymol.org/funding.html for more information and
contact sales@pymol.org when you are ready to purchase a Subscription.

This Executable Build integrates and extends Open-Source PyMOL 1.3.
Executive: Loading version 1.30 session...
PyMOL>
```

Figure 4.6: Mutagenesis Process

Then, firstly we click on any sequence for changing or moving sequence. When we click on any sequence we get several side chain orientations (rotamers) are possible, each of the rotamers available for this residue in PYMOL.

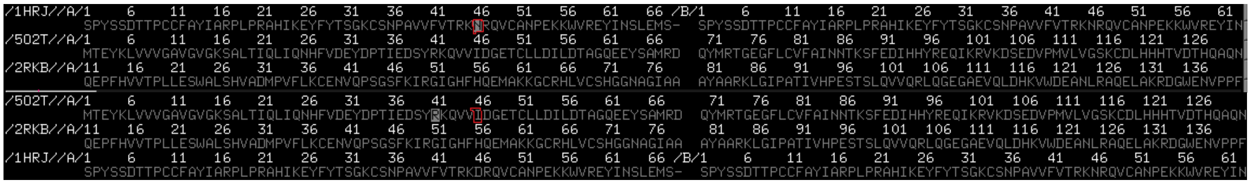


```
/2RKB//A/11   16   21   26   31   36   41   46   51   56   61   66   71   76      81   86   91   96   101  106  111  116  121  126  131  136
              QEPFHVVTPLLESWALSHVADMPVFLKCENVQPSGSFKIRGIGHFHQEMAKKGCRHLVCSHGGNAGIAA  AYAARKLGIPATIVHPESTSLQVVQRLQGEGAEVQLDHKVWDEANLRAQELAKRDGWENVPPF
/1HRJ//A/1    6    11   16   21   26   31   36   41   46   51   56   61   66 /B/1    6    11   16   21   26   31   36   41   46   51   56   61
              SPYSSDTTPCCFAYIARPLPRAHIKEYFYTSGKCSNPAVVFVTRKNRQVCANPEKKWVREYINSLEMS-  SPYSSDTTPCCFAYIARPLPRAHIKEYFYTSGKCSNPAVVFVTRKNRQVCANPEKKWVREYIN
/502T//A/1    6    11   16   21   26   31   36   41   46   51   56   61   66      71   76   81   86   91   96   101  106  111  116  121  126
              MTEYKLVVVGAVGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAGQEEYSAMRD  QYMRTGEGFLCVFAINNTKSFEDIHHYREQIKRVKDSEDVPMVLVGSKCDLHHHTVDTHQAQN

For Educational Use Only
Select a rotamer for 1HRJ//A/ASN`36 or pick a new residue...
```

Figure 4.7: shows the percentage of mutations

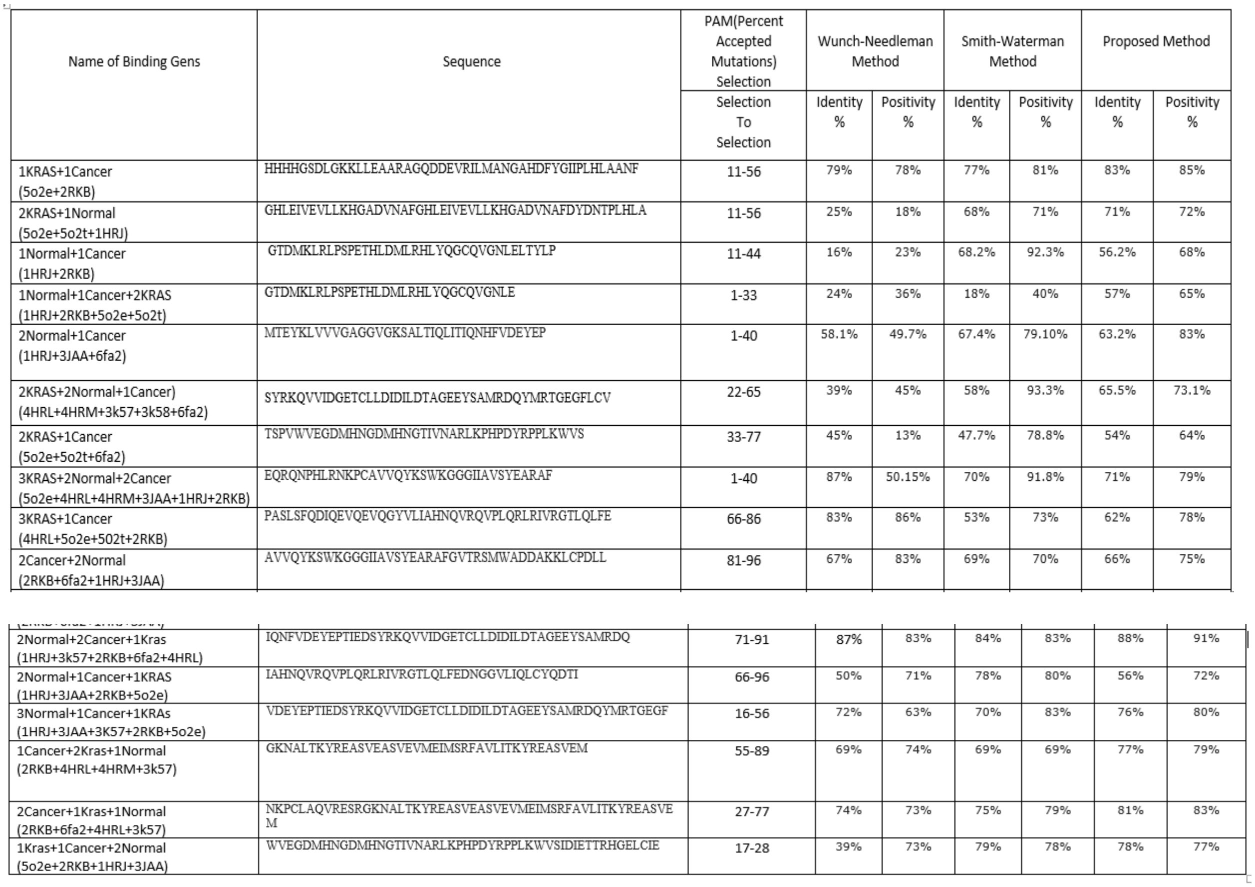
And their reduction sequence:



```
/1HRJ//A/1     6    11   16   21   26   31   36   41   46   51   56   61   66 /B/1    6    11   16   21   26   31   36   41   46   51   56   61
               SPYSSDTTPCCFAYIARPLPRAHIKEYFYTSGKCSNPAVVFVTRKNRQVCANPEKKWVREYINSLEMS-  SPYSSDTTPCCFAYIARPLPRAHIKEYFYTSGKCSNPAVVFVTRKNRQVCANPEKKWVREYIN
/502T//A/1     6    11   16   21   26   31   36   41   46   51   56   61   66      71   76   81   86   91   96   101  106  111  116  121  126
               MTEYKLVVVGAVGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAGQEEYSAMRD  QYMRTGEGFLCVFAINNTKSFEDIHHYREQIKRVKDSEDVPMVLVGSKCDLHHHTVDTHQAQN
/2RKB//A/11    16   21   26   31   36   41   46   51   56   61   66   71   76      81   86   91   96   101  106  111  116  121  126  131  136
               QEPFHVVTPLLESWALSHVADMPVFLKCENVQPSGSFKIRGIGHFHQEMAKKGCRHLVCSHGGNAGIAA  AYAARKLGIPATIVHPESTSLQVVQRLQGEGAEVQLDHKVWDEANLRAQELAKRDGWENVPPF
/502T//A/1     6    11   16   21   26   31   36   41   46   51   56   61   66      71   76   81   86   91   96   101  106  111  116  121  126
               MTEYKLVVVGAVGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAGQEEYSAMRD  QYMRTGEGFLCVFAINNTKSFEDIHHYREQIKRVKDSEDVPMVLVGSKCDLHHHTVDTHQAQN
/2RKB//A/11    16   21   26   31   36   41   46   51   56   61   66   71   76      81   86   91   96   101  106  111  116  121  126  131  136
               QEPFHVVTPLLESWALSHVADMPVFLKCENVQPSGSFKIRGIGHFHQEMAKKGCRHLVCSHGGNAGIAA  AYAARKLGIPATIVHPESTSLQVVQRLQGEGAEVQLDHKVWDEANLRAQELAKRDGWENVPPF
/1HRJ//A/1     6    11   16   21   26   31   36   41   46   51   56   61   66 /B/1    6    11   16   21   26   31   36   41   46   51   56   61
               SPYSSDTTPCCFAYIARPLPRAHIKEYFYTSGKCSNPAVVFVTRKDRQVCANPEKKWVREYINSLEMS-  SPYSSDTTPCCFAYIARPLPRAHIKEYFYTSGKCSNPAVVFVTRKNRQVCANPEKKWVREYIN
```

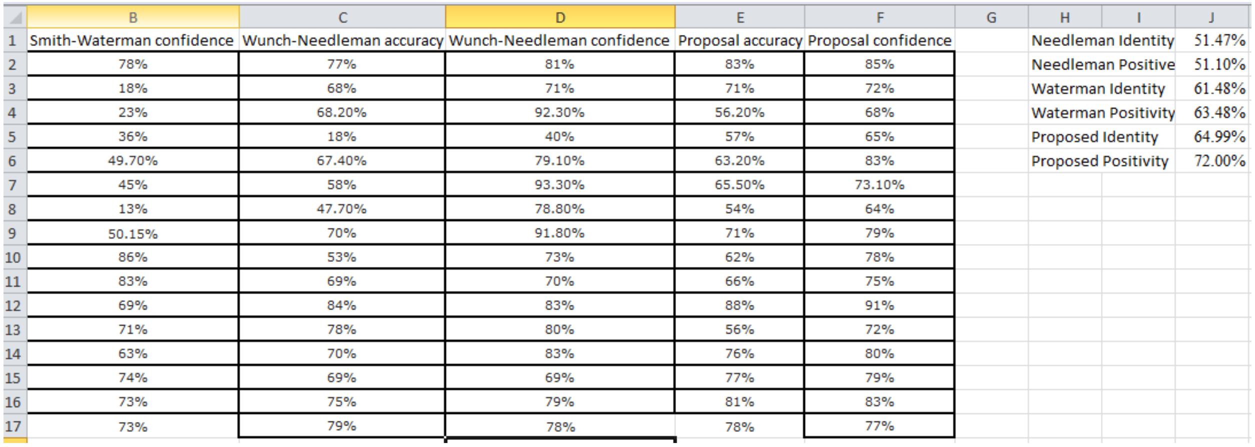
Figure 4.8: Sequences of reduction

4.8 Performance result analysis of all method and comparison result with our method:



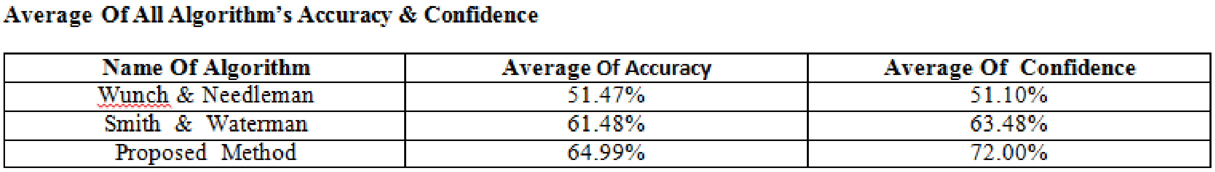
Name of Binding Gens	Sequence	PAM(Percent Accepted Mutations) Selection	Wunch-Needleman Method		Smith-Waterman Method		Proposed Method	
		Selection To Selection	Identity %	Positivity %	Identity %	Positivity %	Identity %	Positivity %
1KRAS+1Cancer (5o2e+2RKB)	HHHHGSDLGKKLLEAARAGQDDEVRILMANGAHDFYGIIPLHLAANF	11-56	79%	78%	77%	81%	83%	85%
2KRAS+1Normal (5o2e+5o2t+1HRJ)	GHLEIVEVLLKHGADVNAFGHLEIVEVLLKHGADVNAFDYDNTPLHLA	11-56	25%	18%	68%	71%	71%	72%
1Normal+1Cancer (1HRJ+2RKB)	GTDMKLRLPSPETHLDMLRHLYQGCQVGNLELTYLP	11-44	16%	23%	68.2%	92.3%	56.2%	68%
1Normal+1Cancer+2KRAS (1HRJ+2RKB+5o2e+5o2t)	GTDMKLRLPSPETHLDMLRHLYQGCQVGNLE	1-33	24%	36%	18%	40%	57%	65%
2Normal+1Cancer (1HRJ+3JAA+6fa2)	MTEYKLVVVGAGGVGKSALTIQLITIQNHFVDEYEP	1-40	58.1%	49.7%	67.4%	79.10%	63.2%	83%
2KRAS+2Normal+1Cancer) (4HRL+4HRM+3k57+3k58+6fa2)	SYRKQVVIDGETCLLDIDILDTAGEEYSAMRDQYMRTGEGFLCV	22-65	39%	45%	58%	93.3%	65.5%	73.1%
2KRAS+1Cancer (5o2e+5o2t+6fa2)	TSPVWVEGDMHNGDMHNGTIVNARLKPHPDYRPPLKWVS	33-77	45%	13%	47.7%	78.8%	54%	64%
3KRAS+2Normal+2Cancer (5o2e+4HRL+4HRM+3JAA+1HRJ+2RKB)	EQRQNPHLRNKPCAVVQYKSWKGGGIIAVSYEARAF	1-40	87%	50.15%	70%	91.8%	71%	79%
3KRAS+1Cancer (4HRL+5o2e+502t+2RKB)	PASLSFQDIQEVQEVQGYVLIAHNQVRQVPLQRLRIVRGTLQLFE	66-86	83%	86%	53%	73%	62%	78%
2Cancer+2Normal (2RKB+6fa2+1HRJ+3JAA)	AVVQYKSWKGGGIIAVSYEARAFGVTRSMWADDAKKLCPDLL	81-96	67%	83%	69%	70%	66%	75%
2Normal+2Cancer+1Kras (1HRJ+3k57+2RKB+6fa2+4HRL)	IQNFVDEYEPTIEDSYRKQVVIDGETCLLDIDILDTAGEEYSAMRDQ	71-91	87%	83%	84%	83%	88%	91%
2Normal+1Cancer+1KRAS (1HRJ+3JAA+2RKB+5o2e)	IAHNQVRQVPLQRLRIVRGTLQLFEDNGGVLIQLCYQDTI	66-96	50%	71%	78%	80%	56%	72%
3Normal+1Cancer+1KRAs (1HRJ+3JAA+3K57+2RKB+5o2e)	VDEYEPTIEDSYRKQVVIDGETCLLDIDILDTAGEEYSAMRDQYMRTGEGF	16-56	72%	63%	70%	83%	76%	80%
1Cancer+2Kras+1Normal (2RKB+4HRL+4HRM+3k57)	GKNALTKYREASVEASVEVMEIMSRFAVLITKYREASVEM	55-89	69%	74%	69%	69%	77%	79%
2Cancer+1Kras+1Normal (2RKB+6fa2+4HRL+3k57)	NKPCLAQVRESRGKNALTKYREASVEASVEVMEIMSRFAVLITKYREASVE M	27-77	74%	73%	75%	79%	81%	83%
1Kras+1Cancer+2Normal (5o2e+2RKB+1HRJ+3JAA)	WVEGDMHNGDMHNGTIVNARLKPHPDYRPPLKWVSIDIETTRHGELCIE	17-28	39%	73%	79%	78%	78%	77%

Figure 4.9: Performance result analysis



	B	C	D	E	F	G	H	I	J
1	Smith-Waterman confidence	Wunch-Needleman accuracy	Wunch-Needleman confidence	Proposal accuracy	Proposal confidence		Needleman Identity		51.47%
2	78%	77%	81%	83%	85%		Needleman Positive		51.10%
3	18%	68%	71%	71%	72%		Waterman Identity		61.48%
4	23%	68.20%	92.30%	56.20%	68%		Waterman Positivity		63.48%
5	36%	18%	40%	57%	65%		Proposed Identity		64.99%
6	49.70%	67.40%	79.10%	63.20%	83%		Proposed Positivity		72.00%
7	45%	58%	93.30%	65.50%	73.10%				
8	13%	47.70%	78.80%	54%	64%				
9	50.15%	70%	91.80%	71%	79%				
10	86%	53%	73%	62%	78%				
11	83%	69%	70%	66%	75%				
12	69%	84%	83%	88%	91%				
13	71%	78%	80%	56%	72%				
14	63%	70%	83%	76%	80%				
15	74%	69%	69%	77%	79%				
16	73%	75%	79%	81%	83%				
17	73%	79%	78%	78%	77%				

Figure: 4.10: Excel result presentation of all method



Average Of All Algorithm's Accuracy & Confidence

Name Of Algorithm	Average Of Accuracy	Average Of Confidence
Wunch & Needleman	51.47%	51.10%
Smith & Waterman	61.48%	63.48%
Proposed Method	64.99%	72.00%

Figure: 4.11: Average of all algorithms

4.9 Graphical representation of Performance Analysis:

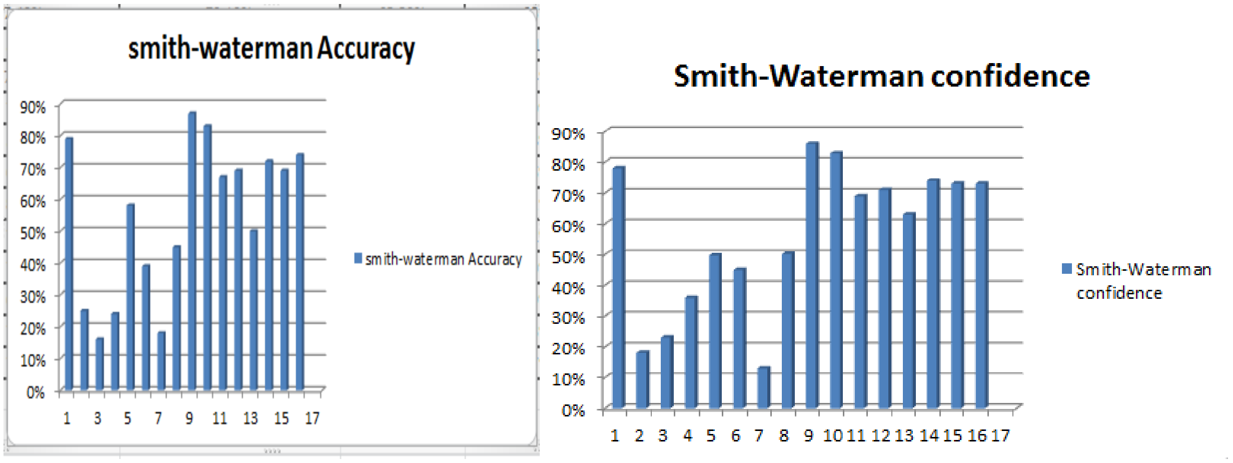


Figure 4.12: Graph of Smith-Waterman method

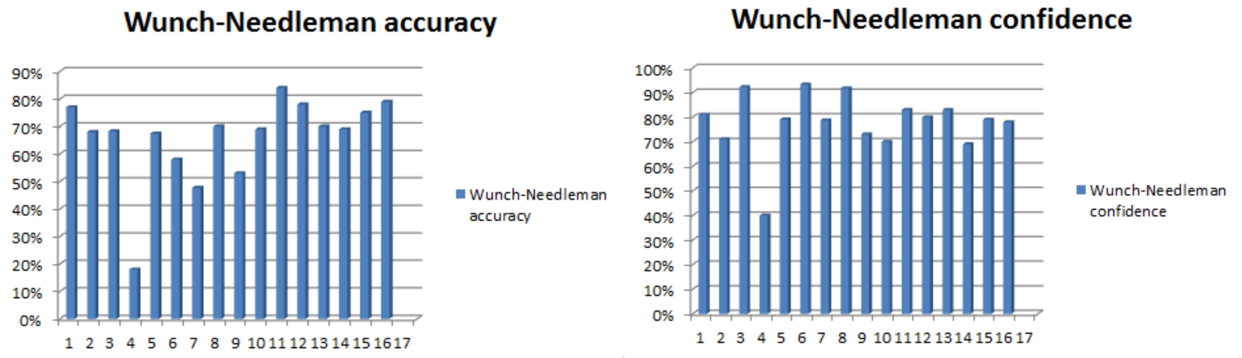


Figure 4.13: Graph of Needleman-Wunch method

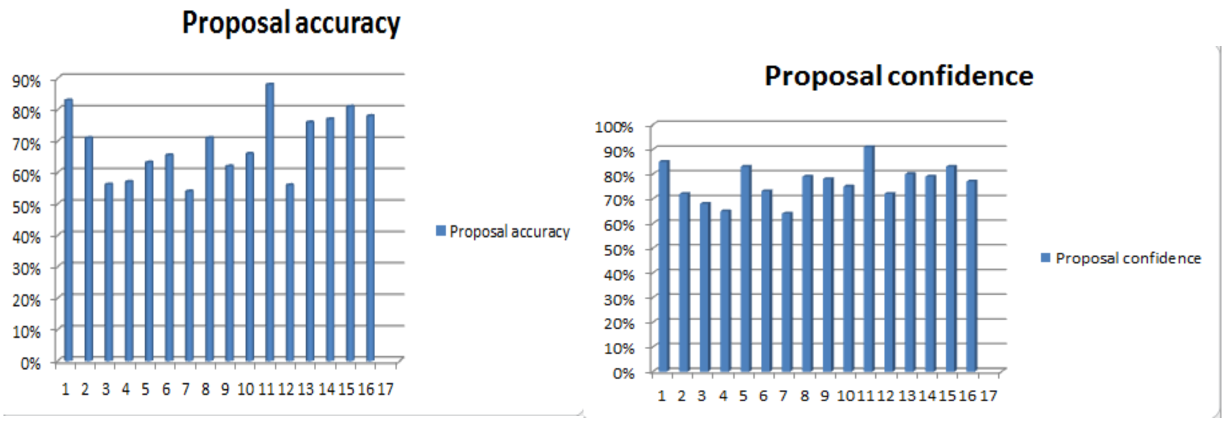


Figure: 4.14: Graph of Proposed method

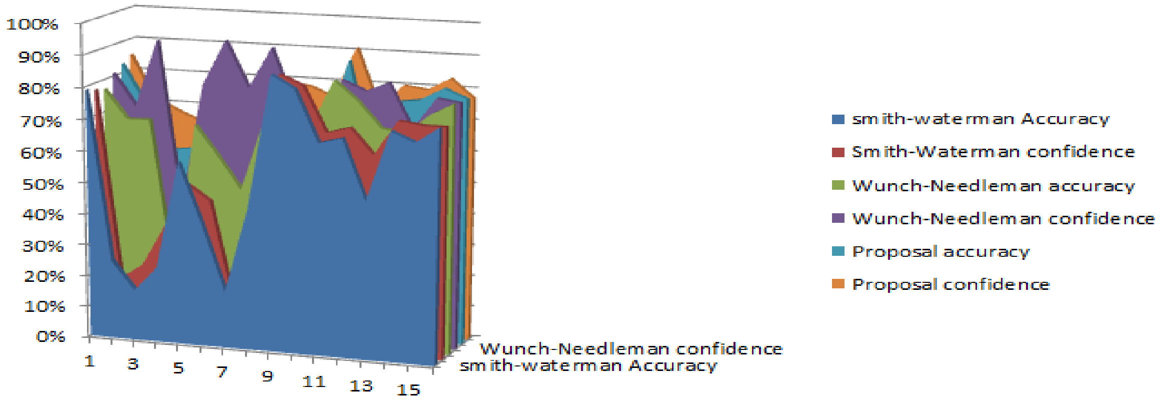


Figure 4.15: All accuracy graph

Chapter 5

Discussion

Life on Earth originated and then evolved from a universal common ancestor approximately 3.8 billion years ago [28]. Repeated speciation and the divergence of life has occurred throughout this time due to shared sets of biochemical and morphological traits, or by shared DNA sequences[13]. These homologous traits and sequences are more similar among species that share a more recent common ancestor, and can be used to reconstruct evolutionary histories, using both existing species and the fossil record. Existing patterns of biodiversity have been shaped both by speciation and by extinction. These similarities were mostly done by the help of sequence alignment. DNA sequencing have become one of the most significant research areas in Bioinformatics [29]. In this paper, we discuss the sequence alignment pair-wise method. Sequence alignment turns out to be helpful while detecting and identifying known genes or unknown genes, all sorts of mutations (insertion, deletion) i.e. detect the DNA nucleotides responsible for the changes. So that, after data plugin from PDB we bind different type of cells (normal cells, cancer cells, KRAS [29] genes) together and we observe the changes of sequence of these cells. Comparison against other known protein sequences will help to understand the changed functionality and structural arrangement. Then we detect the high and low density area of cancer cells which spread through the normal cells with the help of Needleman-Wunch algorithm. After detection we change or align amino acid sequence by pair-wise sequence method and it change the mutagenesis percentage. After mutagenesis process we filter data to insert, delete, replace the identified problems. After that we compare our experimental result with other methods. Sequence alignment plays an important role in biological research laboratory and drug design as a result proper drug can be designed with reduced side effects [32].

Future Work:

Global Alignment (Pair-Wise) Method is based on pair-wise sequence alignment. So far, multiple alignments with mutated gene have not been worked out and our method is also pair-wise method. So in future, we want to work with multiple sequence alignment method based on our method [30]. We want to work with NEOBIO is a library of bioinformatics algorithms implemented in Java. By using this live software, we hope that it will be more easy to implement different methods and also our method for comparison [31].

CONCLUSION

In this thesis, the methods of aligning DNA sequences optimally and relatively techniques several DNA sequences with defects or vague identity could be recognized by aligning with existing genetic data. In bioinformatics, sequence alignment of such type is greatly required for generating correct outputs. As we have mentioned, aligning plays an important role in drug design, forensics, DNA defects etc. The demand for faster and optimizing algorithm would also be at high peak for bioinformatics due to increasing need of better drugs and treatment. In our opinion, this study demonstrated the benefit of using hybrid model for the cells binding, detection, reduction of cancer cells and filtration of mutated genes which is easy-to-use kit format and easy to understand or that are not technically demanding and that require equipment's that is readily available at most academic institutions.

References

[1] Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K. W., & Vogelstein, B. (2011). Detection and quantification of rare mutations with massively parallel sequencing. Proceedings of

the National Academy of Sciences, 108(23), 9530-9535.

[2] Islam, N. (2012). Faster and efficient algorithm for sequence alignment (Doctoral dissertation, BRAC University).

[3] Naser, W. M., Shawarby, M. A., Al-Tamimi, D. M., Seth, A., Al-Quorain, A., Al Nemer, A. M., & Albagha, O. M. (2014). Novel KRAS gene mutations in sporadic colorectal cancer. *PloS one*, *9*(11), e113350

[4] Kamburov, A., Lawrence, M. S., Polak, P., Leshchiner, I., Lage, K., Golub, T. R., ... & Getz, G. (2015). Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proceedings of the National Academy of Sciences*, *112*(40), E5486-E5495.

[5] Moretta, A., Bottino, C., Mingari, M. C., Biassoni, R., & Moretta, L. (2002). What is a natural killer cell?. *Nature immunology*, *3*(1), 6.

[6] Likic, V. (2008). The Needleman-Wunsch algorithm for sequence alignment. Lecture given at the 7th Melbourne Bioinformatics Course, Bi021 Molecular Science and Biotechnology Institute, University of Melbourne, 1-46.

[7] DeLano, W. L., & Bromberg, S. (2004). PyMOL user’s guide. *DeLano Scientific LLC, San Carlos, California, USA*

[8] Kalsi, S., Kaur, H., & Chang, V. (2018). DNA cryptography and deep learning using genetic algorithm with NW algorithm for key generation. *Journal of medical systems*, *42*(1), 17.

[9] Luscombe, N. M., Greenbaum, D., & Gerstein, M. (2001). What is bioinformatics? A proposed definition and overview of the field. *Methods of information in medicine*, *40*(04), 346-358.

[10] Sierk, M. L., Smoot, M. E., Bass, E. J., & Pearson, W. R. (2010). Improving pairwise sequence alignment accuracy using near-optimal protein sequence alignments. BMC bioinformatics, 11(1), 146.

[11] Jančík, S., Drábek, J., Radzioch, D., & Hajdúch, M. (2010). Clinical relevance of KRAS in human cancers. *BioMed* Research International, *2010*.

[12] Kinjo, A. R., Suzuki, H., Yamashita, R., Ikegawa, Y., Kudou, T., Igarashi, R., ... & Nakamura, H. (2012). Protein Data Bank Japan (PDBj): maintaining a structural data archive and resource description framework format. Nucleic acids research, 40(Database issue), D453.

[13] Lesk, A. (2019). Introduction to bioinformatics. Oxford university press.

[14] Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., ... & Hornik, K. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, *5*(10), R80.

[15] Kitano, H. (2002). Computational systems biology. *Nature*, *420*(6912), 206.

[16] Eisenmann, J. C., & Wickel, E. E. (2009). The biological basis of physical activity in children: revisited. Pediatric Exercise Science, *21*(3), 257-272.

[17] Medvedev, P., Stanciu, M., & Brudno, M. (2009). Computational methods for discovering structural variation with next-generation sequencing. *Nature methods*, *6*(11s), S13.

[18] Jenson, S. K., & Domingue, J. O. (1988). Extracting topographic structure from digital elevation data for geographic information system analysis. *Photogrammetric engineering and remote sensing*, *54*(11), 1593-1600.

[19] Murata, M., Richardson, J. S., & Sussman, J. L. (1985). Simultaneous comparison of three protein sequences. *Proceedings of the National Academy of Sciences*, *82*(10), 3073-3077.

[20] Ovesný, M., Křížek, P., Borkovec, J., Švindrych, Z., & Hagen, G. M. (2014). ThunderSTORM: a comprehensive ImageJ plug-in for PALM and STORM data analysis and super-resolution imaging. *Bioinformatics*, *30*(16), 2389-2390.

[21] Burrows, M., & Wheeler, D. J. (1994). A block-sorting lossless data compression algorithm.

[22] Alharthy, A., & Bethel, J. (2002). Heuristic filtering and 3D feature extraction from LIDAR data. International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences, *34*(3/A), 29-34.

[23] Krebs, C. J., & Krebs, C. J. (1994). Ecology: the experimental analysis of distribution and abundance (Vol. 4). New York: HarperCollins College Publishers.

[24] Jarrett, K., Kavukcuoglu, K., & LeCun, Y. (2009, September). What is the best multi-stage architecture for object recognition?. In *2009 IEEE 12th international conference on computer vision* (pp. 2146-2153). IEEE.

[25] Pace, N. R. (1991). Origin of life-facing up to the physical setting. *Cell*, *65*(4), 531-533.

[26] Alharthy, A., & Bethel, J. (2002). Heuristic filtering and 3D feature extraction from LIDAR data. International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences, *34*(3/A), 29-34.

[27] Kitano, H. (2002). Computational systems biology. *Nature*, *420*(6912), 206.

[28] Jančík, S., Drábek, J., Radzioch, D., & Hajdúch, M. (2010). Clinical relevance of KRAS in human cancers. *BioMed* Research International, *2010*.

[29] Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., ... & Hornik, K. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, *5*(10), R80.

[30] Moretta, A., Bottino, C., Mingari, M. C., Biassoni, R., & Moretta, L. (2002). What is a natural killer cell?. *Nature immunology*, *3*(1), 6.

[31] Lesk, A. (2019). Introduction to bioinformatics. Oxford university press.

[32] Jenson, S. K., & Domingue, J. O. (1988). Extracting topographic structure from digital elevation data for geographic information system analysis. *Photogrammetric engineering and remote sensing*, *54*(11), 1593-1600.