Abstract:

Over the past decades, the quantity and quality of biological information has skyrocketed, largely because of advances in molecular biology and genomic technology. In Bioinformatics, a sequence alignment is a way of arranging the sequence of DNA, RNA or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequence. In our process, we report binding of cancer cells, normal cells and KRAS genes, detections of high and low density of cancer cells and sequence mutations of cancer cells 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 which filtrate mutated DNA. This approach helps in detecting any abnormal changes and the mutation percentages of those abnormal changes and 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 comparison 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.

Keywords: Alignment, Mutations, Protein, DNA, KRAS gene, Filtering, Bioinformatics.

Introduction:

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.

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 to be 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.

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.

In the below section at first we discuss about the related work may also be called a literature review. The point of the section is to highlight work done by others that somehow ties in with our own work. Then we discuss about Specific Objectives. Specific Objectives are statements that describe results in terms of knowledge, aspiration. In this section we discuss about our work and our goal. Then we discuss about our proposed method. In this section at first we discuss our proposed method parameters. Then we discuss about our proposed model algorithm and then flowchart of our whole work which is overview of our process and then methodology where the description of our process. Then we discuss about our experimental result analysis. In this section we discuss about our process performance, results, compassion with other process and Tentative Analysis where we discuss about the software that we used in our process. Then we discuss about the difficulties of our process. Each method has some difficulties, so we also face some difficulties to implement this method. Then we discuss about our whole work in discussion section. The purpose of the discussion is to interpret and describe the significance of our findings, what was already known about the research problem being investigated and to explain any new understanding or insights that emerged as a result of our study of the problem. Then we discuss conclusion, it is to restate the main argument. And then references, A reference list lists only the sources that we refer to in our writing to consulted for their ideas. The purpose of the reference list is to allow our sources to be found by our reader.

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.

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.

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.

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.

Specific 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.

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)

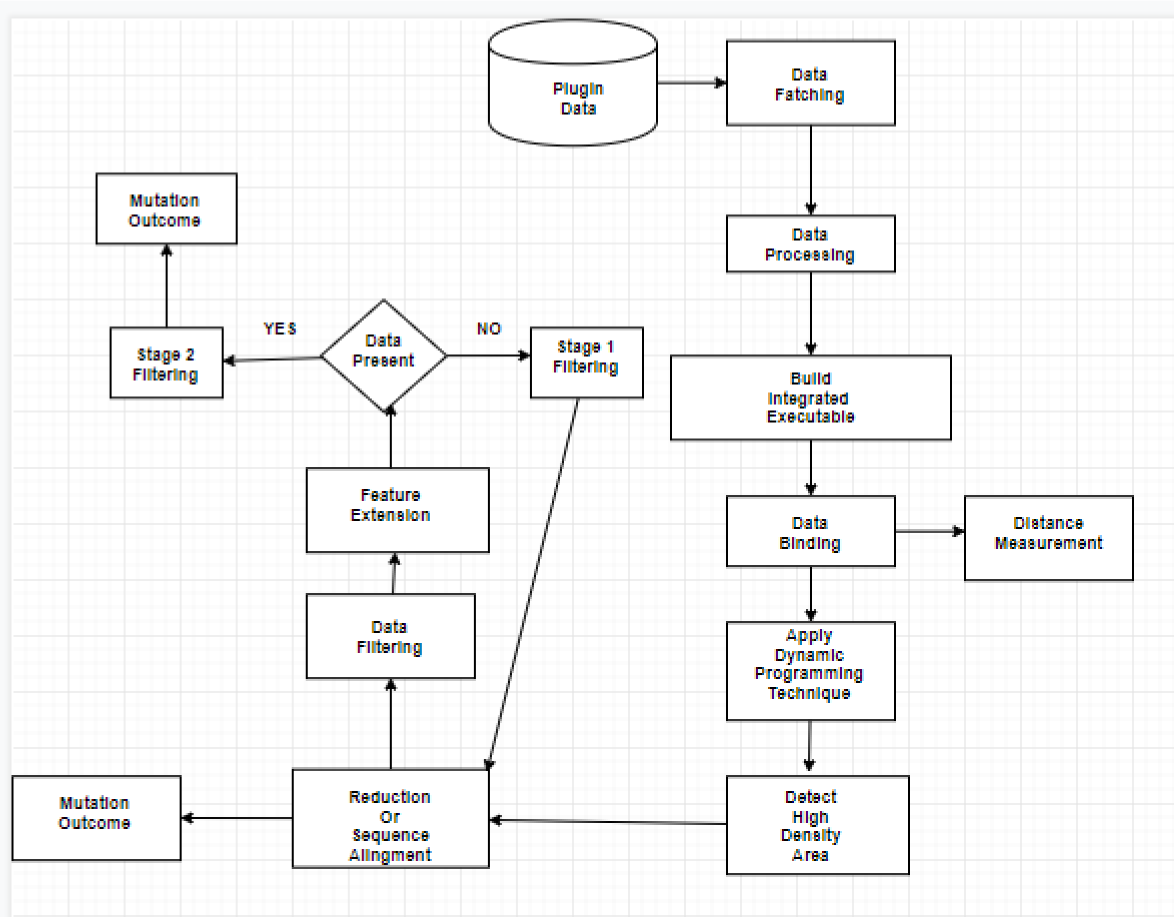
For this method, we use Pymol and Python shell that can control the cancer cells in a feasible tolerance metrics. The specialty of this model, it is a hybrid model which is a mix interpretation of Cells binding, Detection, Reduction or Sequence Alignment and filtering method. This process can make a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

Algorithm : Proposed model

Data : Automatically classified label dataset
Result : Sequential classification into align

```
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-1] +del(A[i]);
                          | T[i-1, j-1] + ins (B[j]);
                          | T[i-1, j-1] + gap penalty
                          | T[i-1, 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
Step 6: End
```

Flowchart of the whole work:



Methodology:

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. Here in our method at first we plugin data from PDB Japan (Protein Data Bank of Japan). We can save plugin data in our desktop and can run the data later.

Dataset: a collection of related sets of information that is composed of separate elements but can be manipulated as a unit by a computer. Here we use protein data from Protein Data Bank Japan(PDBJ). The Protein Data Bank (PDB) is a Web-based database for the three-dimensional structural data of large biological molecules, such as proteins and nucleic acids and we can get labeled and unlabeled data from this database. We use labeled protein data for our method.

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. After data plugin PyMol software fetch the data for execution.

Data 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. Here the data is processed for execution.

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. In our method after fetch the data build integrates helps to execution.

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

Measurement: Measurement is the process observing and recording the observations that are collected as part of a research effort. Here, we can measure the distance between two or more cells or atoms.

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. Here in our method we use Global alignment method pair-wise method to detect the high and low density area of cancer cells. Which is an 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. Here, in our method reduction or sequence alignment is a mutagenesis process. Where we can change or align amino acid sequence by pair-wise sequence and it change the mutagenesis percentage. Based on mutagenesis percentage we can change the amino acid 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. In our method, we filter data to insert, delete, replace the identified problems.

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.

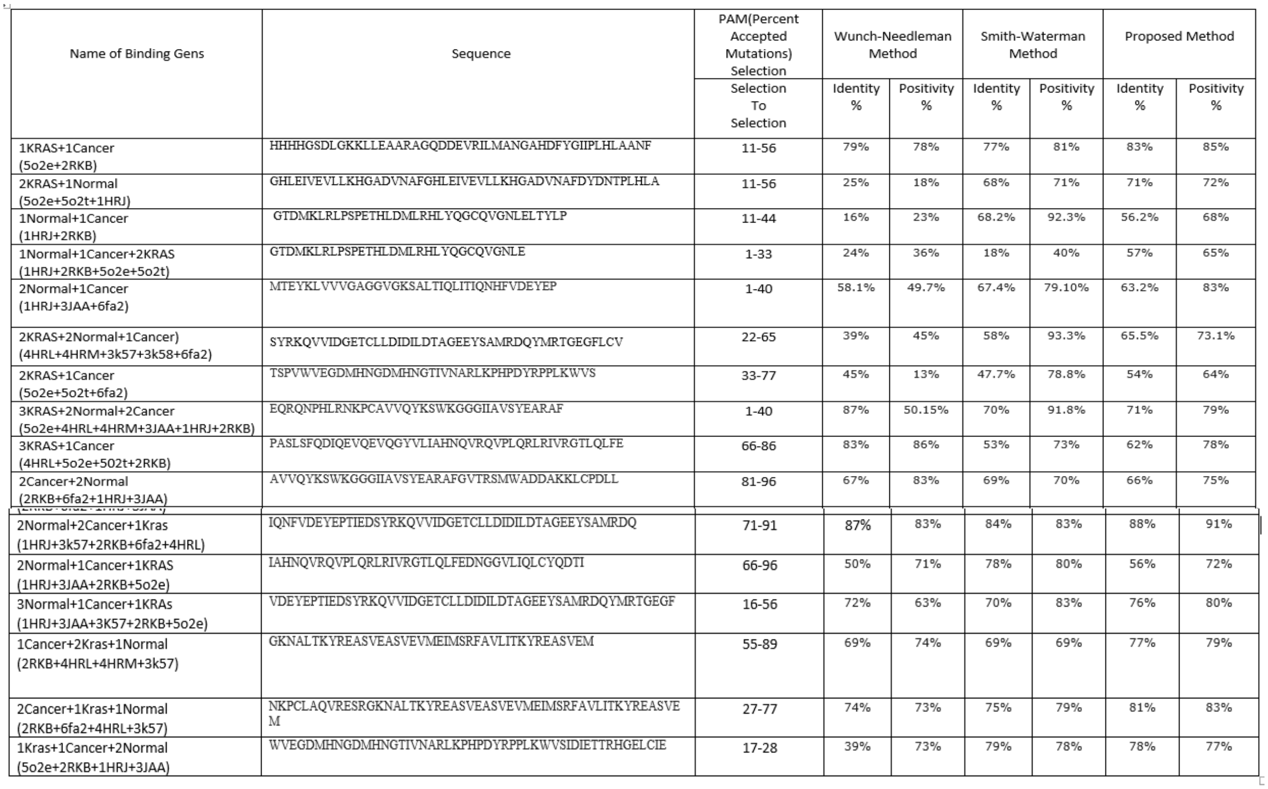
Stage1 Filtering: Data filtering is a process of choosing a smaller part of dataset and using that subset for viewing or analysis. In data filtering, Stage1 filtering is a counting system. If the data is present in hash value, then it goes to Stage1 filtering.

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. If the data is not present in hash value, then it goes to Stage2 filtering.

Experimental Result Analysis:

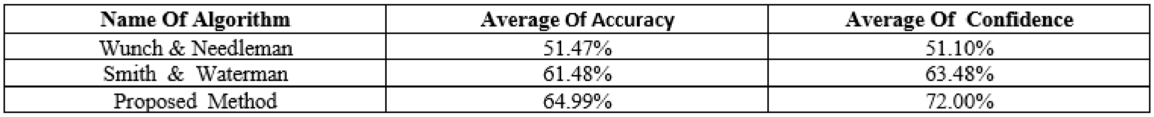
We use PyMol software helps to analyses data which can control the cancer cells in a feasible tolerance metrics and Python shell to filtrate the mutated gene. So we implement a method which is a mix interpretation and we have received our aspirations of reduction was 41.8% before our method was implemented and after implementing our method the result we got 55.7% in alignment or mutagenesis process. After comparing the experimental result of sequences, with our method we got the average accuracy 64.99% which is better result from another method. In filtering process after implementing our method it is possible to filtrate the mutated genes and replace the identified problems.

Here the performance analysis and comparison result and average



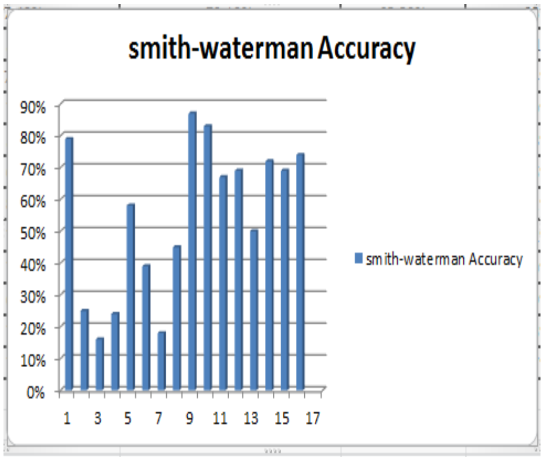
Name of Binding Gens	Sequence	PAM(Percent Accepted Mutations) Selection	Wunch-Needleman Method		Smith-Waterman Method		Proposed Method	
		Selection To Selection	Identity %	Positivity %	Identity %	Positivity %	Identity %	Positivity %
1KRAS+1Cancer (5o2e+2RKB)	HHHHGSDLGKKLLEAARAGQDDEVRILMANGAHDFYGIIPLHLAANF	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)	NKPCLAQVRESRGKNALTKYREASVEASVEVMEIMSRFAVLITKYREASVEM	27-77	74%	73%	75%	79%	81%	83%
1Kras+1Cancer+2Normal (5o2e+2RKB+1HRJ+3JAA)	WVEGDMHNGDMHNGTIVNARLKPHPDYRPPLKWVSIDIETTRHGELCIE	17-28	39%	73%	79%	78%	78%	77%

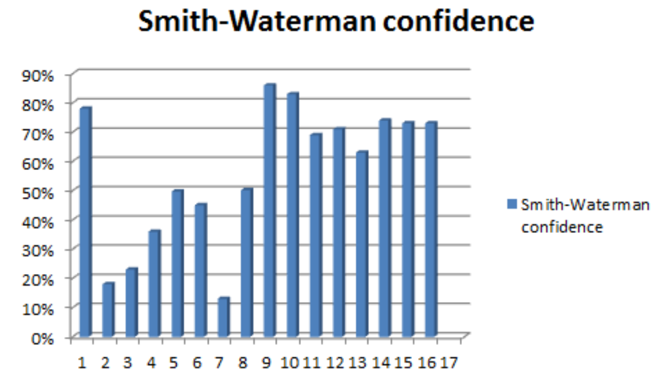
Average Of All Algorithm's Accuracy & Confidence

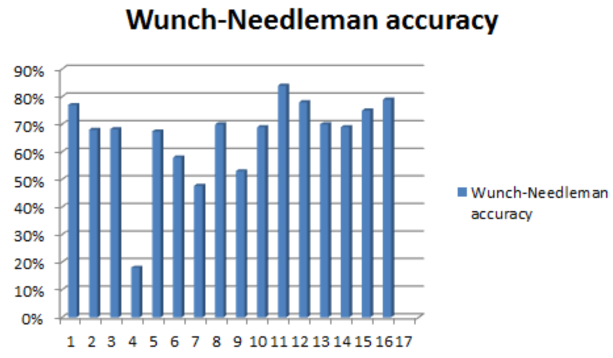


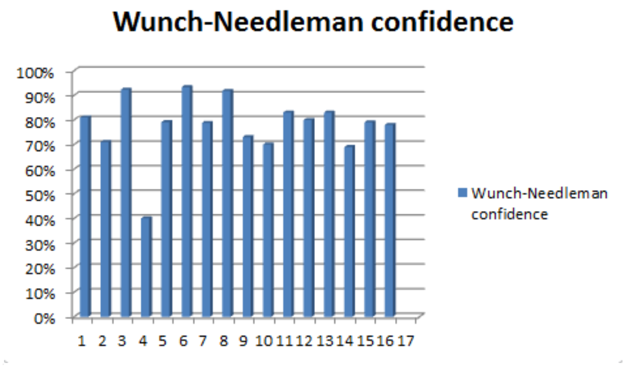
Name Of Algorithm	Average Of Accuracy	Average Of Confidence
Wunch & Needleman	51.47%	51.10%
Smith & Waterman	61.48%	63.48%
Proposed Method	64.99%	72.00%

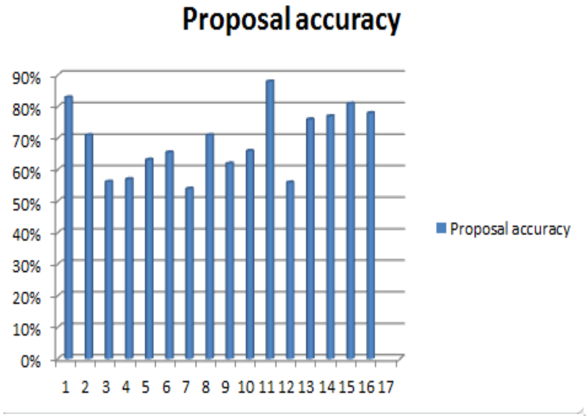
Graphical representation of all method accuracy:

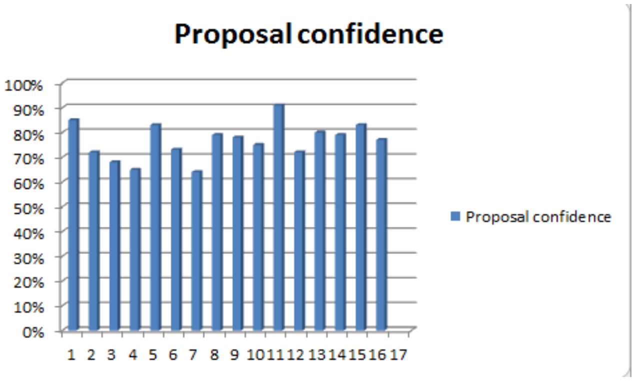












In our method Pymol software helps to analyses data which is open-source model molecule visualization tool which supports both UNIX and Windows, using OpenGL and Python and a small set of Open-source external dependencies. The PyMOL Python API provides a solid way to extend and interface. PyMOL was created in an efficient but highly pragmatic manner, with heavy emphasis on delivering powerful features to end users. Though PyMOL will undoubtedly continue to expand and improve over the next decade. The another software is Anaconda that aims to simplify package management and deployment which is a open-source distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.). The Python shell is used by over 13 million users and includes more than 1400 popular data-science packages. We expect that its impact will primarily inspire other development efforts having more time and resources, and which will undoubtable achieve greater heights.

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.

Discussions:

Life on Earth originated and then evolved from a universal common ancestor approximately 3.8 billion years ago. 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. 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. 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 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.

Conclusion:

In this thesis, the methods of aligning DNA sequences optimally and relatively efficiently are studied. It shows that, by the help of Global Alignment method and its 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. Pymol software rapidly becoming a more affordable option and it is inevitable. Python shell is also a popular and affordable option. We are hopeful that, the result of our process can make a good help to the biological researchers and others in bioinformatics

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.