Demonstrating Protein Analysis for cancer Disease using Approximation Algorithms

A. Rajapriya, A. Nagarajan

Abstract: ***Cancer is the most hazardous disease. If it went on last stage then it is very difficult to cure but, if it detect at earlier stage then it may be curable. Cancer detection by using sequence alignment technique is one of the important research topics of bioinformatics. It searches for similarities and identification of organic mutations between protein sequence and DNA sequence. There are many heuristic and non-heuristic algorithm used for sequence alignment. The research focuses on developing a system to check out the different approximation algorithms performance for detection of cancer. This research work demonstrates the use of Smith Waterman Algorithm for similarity matching of protein sequences and hybrid algorithm for cancer disease detection at early stage.***

Index Terms: Protein Analysis, Cancer Disease, Early Prediction, Approximation algorithms

I. INTRODUCTION

Cancer is also called as malignant tumor and it is leading reason of deaths in world. In developing countries cancer disease is emerged as a major health problem in public. According to world cancer report deaths from cancer disease rates could further spread by 50% to 15 million new cases in the year. Cancer rates are increasing at a disturbing rate globally 2020[14]. Detection of cancer is still challenging for doctors and researchers. Detection and treatment at early stage can save many lives. Using bioinformatics technique we can predict whether person is suffering from cancer disease or not. Nowadays to analyze biological data researchers uses Bioinformatics techniques. Bioinformatics is an integration of analytical, mathematical and computer methods. There are many bioinformatics tasks on sequences but most important is computational chain analysis [4]. Proteins are large biological molecules composed of one or more amino acids chain.

Proteins are responsible for lot of functions within the living organism [11]. Protein functions primarily defined by its shape and amino acid sequences. Amino acids can represent by a letter, thus sequence of proteins represented by letters. Alignment finds similarity level or difference level between query sequence and different database sequences. Types of two sequence alignment are as follows [4]. Global sequence alignment: Identify preserved regions similarity and differences between two sequences which are basically equivalent. Local sequence alignment: Identify preserved regions similarity and differences between two sequences which may or may not be related. These two alignment methods are described by different algorithms, for alignment of two sequences which uses scoring matrices Smith waterman and Hybrid algorithm are a nonheuristic dynamic programming algorithm which finds the best optimal local alignment between two protein sequences. As compare to heuristic algorithms SWA and hybrid algorithm guarantees to find the local alignment between two sequences and it reduces the number of false positives as well as raising the quality of overall alignment [10]. This paper is organized as follows: In Section 2, present the related research. Section 3 explains the proposed method and system flow. The last section shows the experimental results obtained and analysis. We disclosed that dangerous cancer disease can be identified using approximation algorithm. This can help patients for further clinical investigation.

II. RELATED WORK

Researchers have showed a number of sequence alignment techniques. Hong Luo, et al. [1] had explained three major approaches of amino acid repeats detection are self comparison strategy, pattern recognition strategy and complexity measurement strategy. They have implemented different algorithm based on different techniques to manage with differen repeat patterns. Stefan Dydel et al. [2] proposed different parallelization scheme which leading to full utilization of processing units regardless of sequence length.

S. Parthasarathy [4] had developed fold recognition program to anticipate possible folds for a new protein sequence based on the 3D-1D profile method. They have demonstrated protein fold prediction tools and discussed in briefly about the algorithms of sequence alignment. Tao Meng, et al. [5] reviews the current progress of using wavelets in biological sequences analysis in cancer genome. They have showed numerical representation of DNA/protein a sequence are crucial in the success of the overall framework and is an active research area. Yan Yang et al. [6] signified and compared two analytical methods for quantification of serum proteins in patient with oral cancer based on iTRAQ labeling and LC-MS/MS analysis. Barry Strengholt et al. [7] showed local sequence alignment algorithms are mainly used for DNA sequence alignment. They have also showed that this system has low cost to use and assets as compare to all other available systems. Shehab et al. [10] presented new algorithm for sequence alignment which overcomes limitation of bioinformatics algorithm by ignoring the unused data. They have reduced the execution time, increased the performance of sequence comparison.

Revised Manuscript Received on June 05, 2019

A. Rajapriya, Research Scholar, Department of Computer Applications, Alagappa University- Karaikudi. Tamil Nadu India

Dr. A. Nagarajan, Assistant Professor, [2]Department of ComputerApplications, Alagappa University- Karaikudi. Tamil Nadu India

III. PROPOSED WORK

In this paper a most efficient Smith Waterman Algorithm and Hybrid Algorithm is proposed for detection of cancer disease. It is possible to find similarities and differences between two protein sequences using these algorithms [7]. Smith waterman algorithm is overlying algorithm than previously implemented approximation algorithms. This algorithm explored more number of tuples at a time for comparison. Smith Waterman Algorithm searching method:

1. It compares query string with each sequence stored in database.
2. It performs comparison between two sequences.
3. Uses dynamic programming approach.
4. For optimal alignment it traces back the similarity matrix.

In SWA O (mn) time is required to align two sequences of lengths m and n. As shown in Figure 1 first of all we collect patient's protein sequence for disease detection. If we got mismatching between patient protein sequence and database sequence then we can conclude patient has disease. Then to investigate patient is suffering from cancer disease or not we have applied hybrid algorithm. This approximation algorithm also takes protein as an input and compares with database sequence. For this algorithm we have set one threshold value. Following are the steps of cancer disease detection by using Hybrid Algorithm:

1. Process of sequence searching
2. Matrix of dot plotting
3. Get diagonals
4. Count total number of diagonals, calculates score, gap and get diagonal sequence
5. Merge top diagonals and calculate score
6. Apply dynamic programming to get protein sequence
7. Rescore
8. Graphical representation

This algorithm is faster than smith waterman algorithm because it select sequence value is greater than smith waterman algorithm.

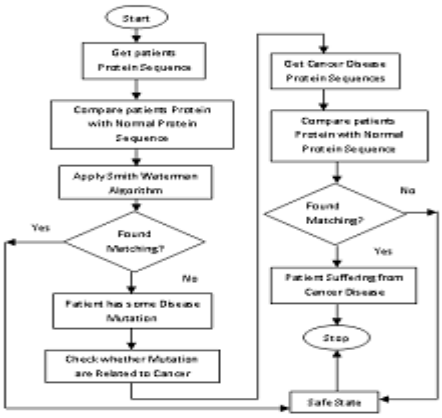

Figure 1. Flow Diagram of the Proposed Method

IV. EXPERIMENTAL RESULTS

We have created one protein analysis application where user can enter their login id, password and protein sequence for further clinical investigation. It compares patient's protein sequence with existing database.

Local alignment of Protein sequences

Sequence 1- ACACACTA

Sequence 2- ACCACACA

Two sequences are arranged in matrix form. Values in the first row and first column are set to 0. Then compare sequence 1 and sequence 2 and find maximum score. Then perform trace backing process which is used to acquire optimal sequence alignment. To do this first of all discover the highest value position. Then start pointing back to the origin, continue until we reach the score 0. To find the aligned sequences the next set of rules is applied to each step in the path [7].

A step diagonally up corresponds with a replacement i.e. match or mismatch

A step towards the left corresponds with a deletion i.e. a gap in a sequence 1.

A step upwards corresponds with an insertion i.e. a gap in a sequence 2.

Table 1. Patient Details

Patient Id	Patient Name	Age	Gender	Protein Sequence
P1	ABC	39	F	ACCGTACCGT
P2	XYZ	25	F	GGGTACCCCTG
P3	MNO	48	M	CCCCTTTTAAAG
P4	JKL	65	F	GACTACTAGCAGT

Table 1 shows the patient details. By using smith waterman algorithm we compare protein sequence with database sequence. We have set threshold value (8) to plot the graph. If score value is more than the threshold value i.e. more than 8 then patient has normal protein sequence. If score value is less than 8 then we can say that patient has disease. Once we found if patient has some mutations then we compare patient protein sequence with existing database using hybrid algorithm. This algorithm first check up the local regions to plot the matrix. Then calculate score, gap and merge best diagonals. If score is greater than threshold value then patient has no cancer disease. Table 2 shows the features of developed application. Proposed application based on sequences of proteins dataset and is not universal method and is more cost effective as compare to other traditional methods.

Table 2. Features of Developed Applications

Features	Proposed Applications
Using Dataset Sequences of Proteins	Yes
Techniques which are used for Predict the Diseases	Used Approximation algorithms to classify Malignant Mutations
Universal Method or Technique	No
The result of application can used to support	Bioinformatics, Bio medical and Clinical Invigilation
Cost Effective for Analysis	Lower cost comparing to existing methods

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

V. CONCLUSION

Due to the high speed and low cost Smith Waterman Algorithm is more efficient as compared to all other sequence alignment algorithm. On regular regions it works more effectively. This algorithm overcomes problems which are associated with bioinformatics tools. To work on large database Hybrid Algorithm is faster than Smith waterman algorithm. At a time it searches large sequence as compare to smith waterman algorithm. Future work will develop new more complex and faster sequence alignment algorithms for protein analysis.

ACKNOWLEDGMENT

This research work has been supported by RUSA PHASE 2.O, Alagappa University.

REFERENCES

1. Hong Luo and Harm Nijveen, “Understanding and identifying amino acid repeats”, Briefings in Bioinformatics, vol 15, No. 4, 2013.
2. Stefan Dydel and Piotr Bala “Large Scale Protein Sequence Alignment using FPGA Reprogrammable Logic Devices”, In: Springer Verlag Berlin Heidelberg 2004.
3. M. P. Sudha, P. Sripriya, “Sequence Alignment in DNA using Smith Waterman and Needleman Algorithms”. International Journal of Computer Science and Information Technologies, Vol. 5(4), 2014.
4. S. Parthasarathy, “Sequence Alignment Algorithms – Application to Bioinformatics Tool Development”,
5. Tao Meng, Ahmed T. Soliman, Mei-Ling Shyu, “Wavelet Analysis in Current Cancer Genome Research: A Survey”, IEEE, 2013.
6. Yan Yang, Junwei Huang, Bahareh Rabii. “Quantitative Proteomic Analysis of Serum Proteins from Oral Cancer Patients: Comparison of Two Analytical Methods.” Int. J. Mol. Sci. 2014.
7. Barry Strengholt, Matthijs Brobbel, “Acceleration of the Smith-Waterman algorithm for DNA sequence alignment using an FPGA platform” 2013.
8. Shannon I. steinfadt, Michael scherger, “A local sequence alignment algorithm using an associative model of parallel computation”.
9. Ayad Ghany Ismaeel, “Novel Method for Mutational Disease Prediction using Bioinformatics Techniques and Back propagation Algorithm”, International Journal(ESTIJ),Vol.3, No.1, 2013.
10. Sara a shehab, Arabi keshk, Hany mahgoub, “Fast Dynamic Algorithm for Sequence Alignment based on Bioinformatics” International Journal of Computer Application, Vol 37-No.7, January 2012.
11. “http://en.wikipedia.org/wiki/Protein”, last accessed on 13/01/2019, 09:15
12. “http://en.wikipedia.org/wiki/Smith-Waterman_algorithm”, last accessed on 26/01/2019, 01:15.
13. AmericanCancerSociety, http://www.cancer.org/acs/groups/content/@research/documents/document/acspc-040951.pdf.
14. Murat Cınar, Mehmet Engin et al.,"Early prostate cancer diagnosis by using artificial neural networks and support vector machines", Expert Systems with Applications: An International Journal, April 2009.
15. HamzaTurabieh, “GA-based Feature Selection with ANFIS Approach to BreastCancer Recurrence”, IJCSI International Journal of Computer Science Issues, Volume 13, Issue 1, ISSN (Print): 1694-0814. January 2016.
16. Rajamani.R and Rathika.M, “Analysis of Liver Cancer using Adaptive Neuro Fuzzy Inference System (ANFIS)”, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 7, July 2015.