Comparative Sequence Analysis of Lymphoma Using A Hybrid Approach of Waterman Smith and Needleman WUNSCH

B.J. Bipin Nair, K.K. Nidheesh, M. Vishnudev

Abstract---* *Bioinformatics is considered as the computational technique where in which it helps to solve the biological based issues. In computer science it helps to improve methods for organizing, analyzing the data, storage and retrieval of the data. In bioinformatics, sequence alignment is one of the major real-time application where in which it helps to compare two or more sequences and the similarity in these sequence can be later used for understanding the relationships. In this work we are developing a computational tool which can predict the 3D structure of different stages of lymphoma. In sequence alignment we are following algorithms such as Needleman Wunch, Smith-Waterman. In each case we are trying to find out best sequence compatibility. Lymphoma is a type of cancer which is mainly affecting in the lymphatic system. The cancer is affecting the white blood cells known as lymphocytes which performs a major role in the immune system. There are mainly two types of Lymphoma namely Hodgkin and Non Hodgkin. The presence of specific type of cell called a Reed-Sternberg cell will identify the presence of Hodgkin, and if that cell is not present it is considered as Non Hodgkin. There are basically four levels of structure prediction namely Primary, Secondary, Tertiary and Quaternary structures. In this work we are making the prediction of these various stages of lymphoma using an hybrid approach of Needleman Wunsch and Waterman smith algorithm.

I. INTRODUCTION

In bioinformatics there are several techniques which are used for analyzing and predicting various kinds of evolutions. It mainly focuses upon the genetic features of a living being. There are different sub areas which are included within bioinformatics. Sequence alignment is considered as a technique where in which it helps in finding the various evolutionary factors. This mechanism helps in understanding the relationship from one generation to the next generation. The sequence alignment is the best technique for finding the DNA, RNA and protein matching properties .Also sequence alignment is used for non-biomedical sequences. It helps to find the structural or evolutionary relationships. Sequence alignment technique mainly includes finding of the best possible alignments which can make better result. Here in our work we are predicting the various stages of lymphoma disease which can help in early detection of disease and help the doctors in the treatment process of the patient.

II. WORK FLOW DIAGRAM

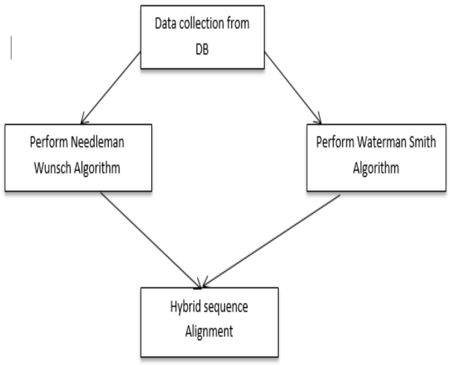

Data collection from DB
Perform Needleman Wunsch Algorithm
Perform Waterman Smith Algorithm
Hybrid sequence Alignment

Fig. 1: Work flow diagram

III. LITERATURE SURVEY

Sudhir Kumar et all[1] says about Evolutionary Genetics which helps in the automatic alignment technique, web related extraction of databases, and a conclusion of the phylogenetic matters, approximate calculation of evolutionary relationship and testing the hypotheses of evolution. In this they provide an overall view of various methods different tools and visual modules for the data and the results which can be obtainable in MEGA. MEGA is a software (Molecular Evolutionary Genetics) which helps in estimating evolutionary distances. In this, Multiple sequence alignment is done using Clustal W algorithm. MEGA facilitates evolutionary molecular basis of DNA and sequences of various amino acids by using pair wise distance matrices method. In future it will helps in extracting phylogenetic relationships. Zhi-min Zhou et all[2] says about Sequence Alignment where in which alignment of the related protein sequence is one of the major problem that exist in the present system.

Manuscript received February 01, 2019

B.J. Bipin Nair, Department of computer Science, Amrita School of Arts and Sciences, Mysuru, Karnatka, Amrita Vishwa Vidyapeetham, India. (e-mail: bipin.bj.nair@gmail.com)

K.K. Nidheesh, Department of computer Science, Amrita School of Arts and Sciences, Mysuru, Karnatka, Amrita Vishwa Vidyapeetham, India. (e-mail: nidheeshkk1@gmail.com)

M. Vishnudev, Department of computer Science, Amrita School of Arts and Sciences, Mysuru, Amrita Vishwa Vidyapeetham, India. (e-mail: vishnudev.manikhoth@gmail.com)

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

An accurate and fast algorithm can make an effective result in biology research programs. In this work they compare by using two different techniques (dynamic and local) and they are implementing a task on multiple alignment. The Encapsulated Gene-by-gene Matching approach is a method which proposes a graph identical strategy to detect gene orthologs and gene segments. Global alignments and local alignment can be developed by introducing dynamic alignments. In future Fuzzy logic can be combined with the progressive method to ensure optimal alignment. Izzat Alsmadi et al[3] says about String DNA comparison which is done using two algorithms. Results shows that the used algorithms and their implementations are not accurate. In this paper it analyses the two algorithms which are used for comparing DNA sequences. They are Longest Common Substring and Subsequence (LCS, LCSS). Analysis of the sequences are performed based upon the several code implementations under the given two algorithms. The result of those used algorithms in the bioinformatics and DNA sequences comparison shows that they are having variety of implementations. In future it allows a better performance in terms of its efficiency to find the solutions. Other techniques and reduction algorithms can be used to reduce the time complexity in calculation process. K Syed Khamarudheen et al[4] says about identifying the splice junctions in DNA.As per the observations, absence of the factor IX genes can cause a disorder named Haemophilia B, so identifying of the splice junctions in the input sequence can predict the occurrence of this disease in the gene. Clotting of blood will not happen when a person is affected with this disease. Therefore in order to predict the presence of this disease they are using LCS and LCCS algorithm. At the same time they prefer for LCCS algorithm as it runs in lesser time. They have used Machine learning technology in order to determine the splice junctions easily. GastonH gonnet et al[5]says about the matching of protein sequence. There are several sets of datas which are available or we can consider it as a group of data where in which it is matched with the sequence database. In this paper, the methods are trying to manage the sequence which are generated during the genome projects. This kinds of datas helps us in providing the datas for the future purpose in order to meet the new challenges in structural biochemistry areas. David J lipman et al[6] says about developing a new algorithm in which it will search for the best similarity between the formed amino acids. So that it can help in prediction of matching protein alignments with respect to the available datas in the database. In this work they are producing an algorithm which will result in high response and sensitivity. In future it can be used for rapid search of protein database. T.F. Smith et al[7]says about Comparative Bio sequence Metrics where in which the paper mainly focuses upon the sequence alignment algorithms, namely algorithms of Needleman and Wunsch then latter as generalized by Waterman, Smith and Beyer . Both of the algorithms works in the same efficiency and it results in maximum matching compatibility of protein sequences which helps in identifying the relationships. The algorithms are used in simple iterative manner. In future, both the algorithms helps out in predicting the various sequence alignment happening with the protein sequences. Lois B travis et al[8]says about Lung Cancer and analyzing the patients who are affected with Lung cancer after being affected by the hodgkins disease. They are considering the patients rate of affecting with lung cancer even after being affected with hodgkins disease. Chemotherapy and radiotherapy is considering as the best treatment ways for cancer. They are analysing the rate of radiations which will cause side effect with respect to other organs. They concluded that the radiation therapy for hodgkins disease is highly increasing the risk of causing lung cancer. DR.D.Chandrakalaet al[9] says about the Needleman algorithm under the dynamic programming to increase the response sensitivity of the algorithm. Here in this paper they are analysing the sequencing technique by using the Needleman algorithm. Best similarities between different sequences is found out which will result in better understanding about the relationships. They are introducing a system which provides 3 best sequences and implements matching and mismatching concepts which will help in future reference. CeÂdric Notredame et al[10] says about the alignment of n number of sequences are one of the major problem which we are facing now. So in this work they proposed t-coffee method in order to sequence the alignments with high accuracy and quick as compared to other normal techniques. This will results in in combination of global and local alignment of sequence by the usage of T-coffee algorithm and comparing the results. Chirtospher Lee et al[11] says about Multiple sequence alignment on partial order graph and about the Progressive Multiple sequence alignment(MSA).But it leads to the loss of the information which is needed for the alignment. Here a graph is presented which allows to reduce the loss of accurate information and reduces the MSA to linear profile. Here it uses an algorithm called POA(partial order alignment) that will consider each new sequence with the MSA sequence.MSA is the most important tool in bioinformatics. The POA algorithm is used to consider simple alignment of two sequence. Ramnadsatra et al[12] says about accelerating computation of DNA multiple sequence alignment in distributed environment .MSA is the technique for finding the similarities between sequences. It helps in the finding of similar single Nucleotide polymorphism etc. The simple algorithm in MSA is star algorithm. The algorithm helps to find the aligning sequences and choose a star sequence which has the maximum alignment score. In DNA, dynamic programing technique is high and the computation is high when DNA length of sequence is high. This aims at accelerating the computing of star Multiple Sequence Alignment using Message Passing Interfaces. Limin Li et al[13] says about Predicting enzymes target for cancer drug by focusing upon human metabolic reactions. The drugs can influence the metabolic system by enzymes where it can create the reactions that helps to target the drugs involved. The paper shows that a network is created to target the anti-cancer drugs.

The recent metabolic system helps to find the metabolic reaction cell-line flux state which is based on the cell-line gene. c.v Umesh et al[14] says about the Persual of different Gene-Susceptibility to tuberculosis in different Indian populations. Tuberculosis is the one of the dangerous disease which cannot be controlled and also no other drugs has been found to cure it. Mannose Binding Lectin (MBL) is used as a tool that helps in preventing the attack of tuberculosis. As the data is collected across India and the it is compared with coefficient of variants which gave a conclusion that showed tuberculosis are mostly caused with Mbl2 genes in chromosomes. Ashok et al[15] says about an approach for Visualization of the information's in the Biological Datas where in which in their work they are integrating several platforms from where many information's can be collected with respect to the biological field. Here they are proposing an integrated flow system in order to combine the datas from several areas which helps in analyzing the data and extracting it for future purpose. Shibin.K et al[16]Proposed a hybrid method of both global and local algorithms to compare the DNA sequence of the patients effected with Brainstem Glioma disease. Needleman wunsch and Waterman smith algorithms are used.

IV. PROPOSED SYSTEM

Sequence Alignment is considered as the best technique in order to find the best matching properties among various sequences. In our work we are introducing a hybrid approach of two algorithms namely Waterman Smith and Needleman Wunsch. Using this we are finding the three parameters, namely match, mismatch and gap values. We are making a hybrid approach in order to make an improvement in the sequence alignment and accuracy along with its time efficiency. Comparing these two algorithms and their functionalities ensures the efficiency in matrix operations for attaining the best possible alignments.

Needleman wunsch Algorithm

Step1: Considering two inputs collected from PDB
Step2: Matrix creation by allocating the match, mismatch and gap value.
Step 3: Performing trace back matrix operation.
Step 4: Calculating match, mismatch and gap parameters

Waterman Smith Algorithm

Step1: Accepting two different DNA sequences and performing matrix formation.
Step 2: Initializing the matrix by considering only positive values.
Step3: performing traceback operation and finding the match, mismatch and gap values.
Step 4: Finding the similarity in sequences.

V. DATASET

In our proposed work we have taken 500 DNA sequences of lymphoma disease from PDB (Protien Data Bank).

VI. EXPERIMENTAL RESULT

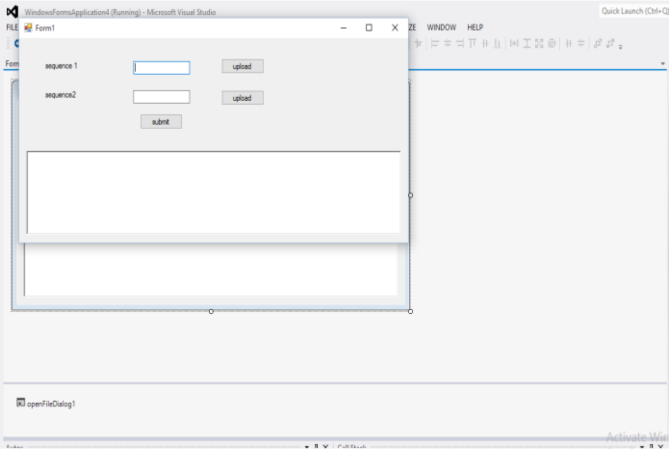
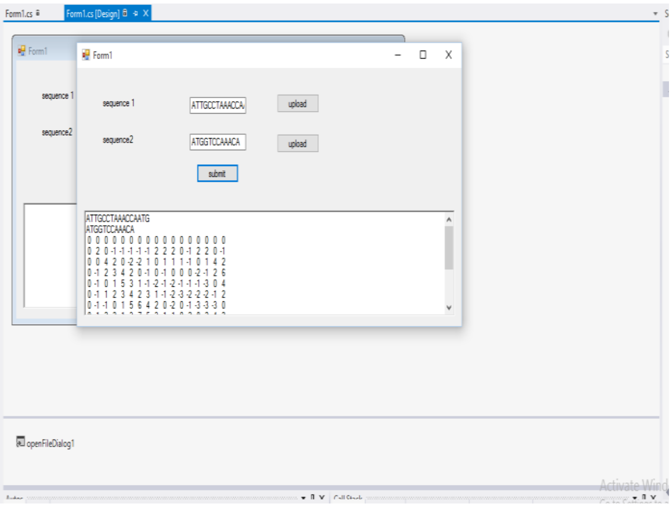

WindowsFormsApplication4 (Running) - Microsoft Visual Studio
Form1
sequence 1
sequence2
upload
upload
submit
openFileDialog1

Fig. 2: Loading the sequence

Carrying out of Needleman Wunsch algorithm, we are taking two inputs as sequence 1 and sequence 2 of lymphoma sequence which is collected from PDB and comparing using global alignment technique.

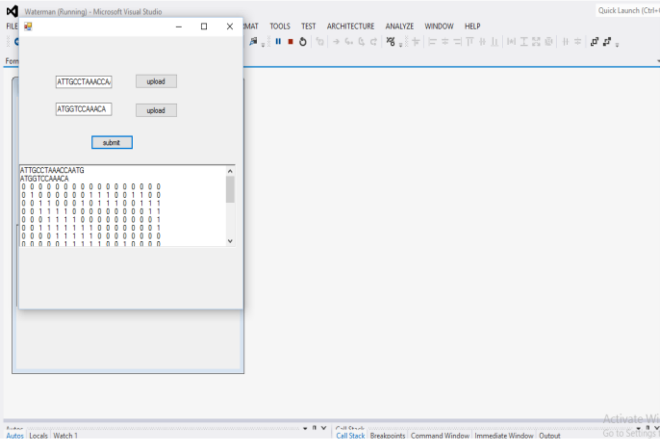


sequence 1 ATTGCCTAAACCA
sequence2 ATGGTCCAAACA

```
ATTGCCTAAACCAATG
ATGGTCCAAACA
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 2 0 -1 -1 -1 -1 -1 2 2 2 0 -1 2 2 0 -1
0 0 4 2 0 -2 -2 1 0 1 1 1 -1 0 1 4 2
0 -1 2 3 4 2 0 -1 0 -1 0 0 0 -2 -1 2 6
0 -1 0 1 5 3 1 -1 -2 -1 -2 -1 -1 -1 -3 0 4
0 -1 1 2 3 4 2 3 1 -1 -2 -3 -2 -2 -2 -1 2
0 -1 -1 0 1 5 6 4 2 0 -2 0 -1 -3 -3 -3 0
```

Fig. 3: Similarity Matrix of Needleman wunsch

After processing, the inputs matrix formation is done by comparing these sequences. From this matrix trace back operation is done with matrix. Then we find the match, mismatch and gap values from that.



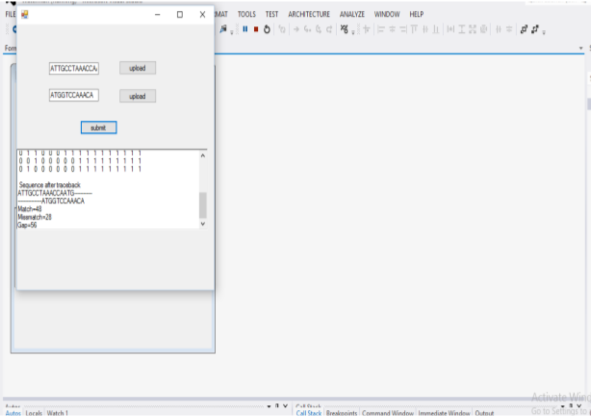
ATTGCCTAAACCA
ATGGTCCAAACA

```
ATTGCCTAAACCAATG
ATGGTCCAAACA
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 1 0 0 0 0 0 0 1 1 1 0 0 1 1 0 0
0 0 1 1 0 0 0 1 0 1 1 1 0 0 1 1 1
0 0 1 1 1 1 0 0 0 0 0 0 0 0 0 1 1
0 0 0 1 1 1 1 0 0 0 0 0 0 0 0 0 1
0 0 1 1 1 1 1 1 1 0 0 0 0 0 0 0 1
0 0 0 0 1 1 1 1 1 0 0 0 0 0 0 0 0
0 0 0 0 0 1 1 1 1 1 0 0 1 0 0 0 0
```

Fig. 4: Waterman smith similarity matrix

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Carrying out of Waterman smith algorithm by accepting two sequences which is collected from the protein data bank and making the comparison approach using local alignment method. Formation of matrix results in trace baking operation which is used in calculating the match, mismatch and gap parameters.



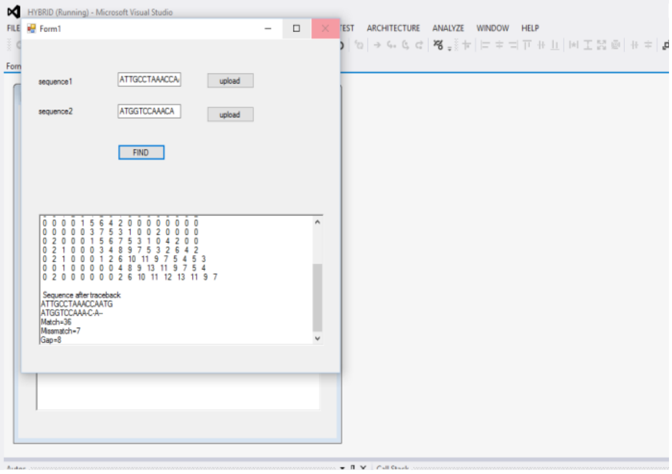
ATTGCCTAAACCA
ATGGTCCAAACA

```
0 1 1 0 0 0 1 1 1 1 1 1 1 1 1 1 1
0 0 1 0 0 0 0 0 1 1 1 1 1 1 1 1 1
0 1 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1

 Sequence after traceback
ATTGCCTAAACCAATG--------
----------ATGGTCCAAACA
Match=48
Mismatch=28
Gap=56
```

Fig. 5: Calculating match, mismatch and gap value using Waterman smith

Calculation of match, mismatch and gap value using waterman smith algorithm results in finding the sequential similarities and dissimilarities.



sequence1 ATTGCCTAAACCA

sequence2 ATGGTCCAAACA

```
0 0 0 0 1 5 6 4 2 0 0 0 0 0 0 0 0
0 0 0 0 0 3 7 5 3 1 0 0 2 0 0 0 0
0 2 0 0 0 1 5 6 7 5 3 1 0 4 2 0 0
0 2 1 0 0 0 3 4 8 9 7 5 3 2 6 4 2
0 2 1 0 0 0 1 2 6 10 11 9 7 5 4 5 3
0 0 1 0 0 0 0 0 4 8 9 13 11 9 7 5 4
0 2 0 0 0 0 0 0 2 6 10 11 12 13 11 9 7

 Sequence after traceback
ATTGCCTAAACCAATG
ATGGTCCAAA-C-A--
Match=36
Missmatch=7
Gap=8
```

Fig. 6: Hybrid approach

In hybrid approach we have combined the functionalities of both algorithms in order to improve its efficiency in calculation. Along with that alignment is made more accurate with its results and time efficiency is improved in calculation.

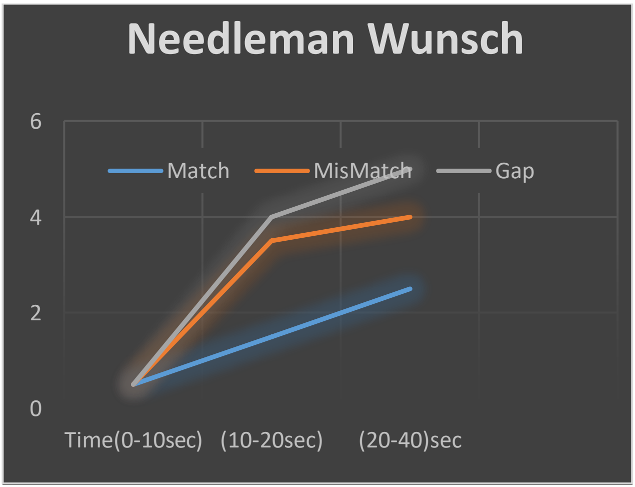


Fig. 7: Efficiency of Needleman wunsch

In need leman wunsch algorithm calculation time is more and the result shows that the match value is in a lower range where the other two parameters are in a higher range, it shows the lower efficiency of the algorithm to predict the relationships among various sequences .

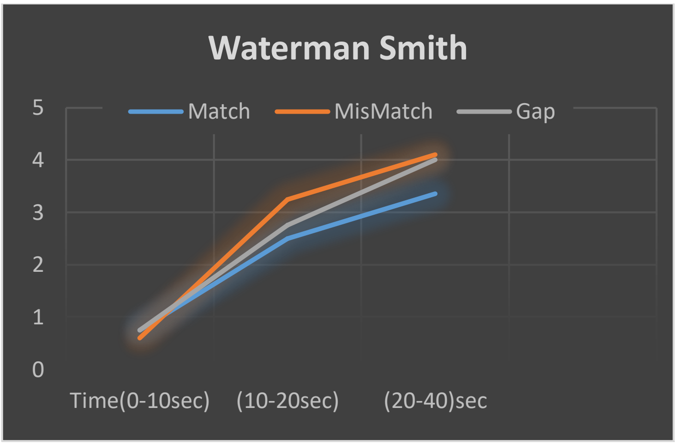


Fig. 8: Efficiency of water man smith

In waterman smith match, mismatch values are in an average range value. As compared to need leman wunsch the value for the match property got increased and the gap value got slight variations. Increase in the match property helped in understanding the overall relationship among sequences.

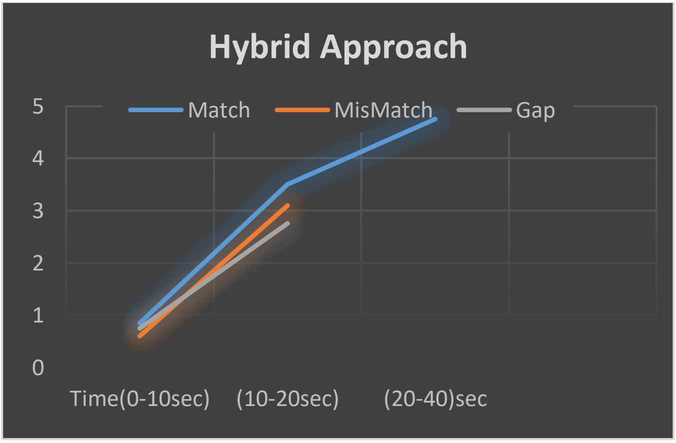


Fig. 9: Efficiency of Hybrid approach

In hybrid approach we are comparing both the algorithms characteristics and making out a combined approach for calculating the various parameters. Here through the hybrid method we have got a result where it made improvements in matching property of the sequences with less calculation time. Although, the mismatch and gap values got decreased through hybrid approach and helped in better analyzing and much improvement in matching of DNA sequences.

VII. CONCLUSION

We have developed a computational tool by combining two algorithms need leman wunsch and waterman smith. Hybrid approach of sequence alignment helped in improving the accuracy of results by maintaining calculation time. From the obtained results we have made predictions of various stages in lymphoma disease and this can make an improvement in easy analysis and early detection of lymphoma. As the stage prediction is done successfully doctors can prescribe medicines and can help the patients in curing the disease as fast as possible. So this work can make a better change in medical industry in accurate and fast detection of various stages of lymphoma. from the sequence as well as if we predict the protein structure will be the future work.

REFERENCES

1. Kumar, S., Tamura, K., &Nei, M. (2004). MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Briefings in bioinformatics*, *5*(2), 150-163.
2. Zhou, Z. M., & Chen, Z. W. (2013). Dynamic programming for protein sequence alignment. *International Journal of Bio-Science and Bio-Technology*, *5*(2), 141-150.
3. Alsmadi, I., & Nuser, M. (2012). String matching evaluation methods for DNA comparison. *International Journal of Advanced Science and Technology*, *47*(1), p13-32.
4. KHAMARUDHEEN, K., & HS, R. (2016). AN APPROACH FOR IDENTIFYING THE PRESENCE OF FACTOR IX GENE IN DNA SEQUENCES USING POSITION VECTOR ANN. *Journal of Theoretical & Applied Information Technology*, *87*(3).
5. Gonnet, Gaston H., Mark A. Cohen, and Steven A. Benner."Exhaustive matching of the entire protein sequence database." *Science* 256.5062 (1992): 1443-1445.
6. Lipman, David J., and William R. Pearson. "Rapid and sensitive protein similarity searches." *Science* 227.4693 (1985): 1435- 1441.
7. Smith, Temple F., Michael S. Waterman, and Walter M. Fitch. "Comparative biosequence metrics." *Journal of Molecular Evolution*18.1 (1981): 38-46.
8. Travis, Lois B., et al. "Lung cancer following chemotherapy and radiotherapy for Hodgkin's disease." *Journal of the National Cancer Institute* 94.3 (2002): 182-192
9. Chandrakala, D., et al. "Optimization of Process Parameters of Global Sequence Alignment Based Dynamic Program-an Approach to Enhance the Sensitivity of Alignment." (2016).
10. Notredame, Cédric, Desmond G. Higgins, and JaapHeringa. "TCoffee: A novel method for fast and accurate multiple sequence alignment." *Journal of molecular biology* 302.1 (2000): 205- 217.
11. Lee, Christopher, Catherine Grasso, and Mark F. Sharlow. "Multiple sequence alignment using partial order graphs." *Bioinformatics* 18.3 (2002): 452-464.
12. Satra, Ramdan, WisnuAnantaKusuma, and HeruSukoco. "Accelerating computation of DNA multiple sequence alignment in distributed environment." *Telkomnika Indonesian Journal of Electrical Engineering* 12.12 (2014): 8278-8285.
13. Li, Limin, et al. "Predicting enzyme targets for cancer drugs by profiling human metabolic reactions in NCI-60 cell lines." *BMC bioinformatics* 11.1 (2010): 501.
14. Umesh, C. V., et al. "Perusal of Mbl2 Gene-Susceptibility to Tuberculosis in Different Indian Populations." (2015).
15. Ashok, Sreeja, and M. V. Judy. "Process Flow for Information Visualization in Biological Data." *Proceedings of the International Congress on Information and Communication Technology*. Springer Singapore, 2016.
16. Nair, B. B., Shibin, K., &Shamcy, O. (2017, April). An hybrid method for comparing brainstem glioma sequences using needlemanwunsch and waterman smith algorithms. In *Convergence in Technology (I2CT), 2017 2nd International Conference for* (pp. 867-872).IEEE.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication