Original Research Article **Bioinformatics**

International Journal of Pharma and Bio Sciences

ISSN
0975-6299

NOVEL APPROACH TO FIND THE OPTIMAL ALIGNMENT USING LCMSQ ALGORITHM FOR IDENTIFYING THE VARIOUS STAGES OF LYMPHOMA

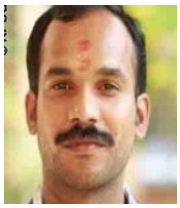
BIPIN NAIR B J[*1], PRANAV V[2], ATHULYA VISWAN[3]

[1,2,3]Department of Computer Science, Amrita School of Arts and Sciences, Amrita Vishwa Vidyapeetham, Amrita University, Mysuru Campus, Karnataka, India

ABSTRACT

Lymphoma is a commonly occurring disease which is seen in coastal areas of Kollam district, Kerala. There are two types of "Lymphoma", one is Hodgkin and another Non-Hodgkin, Both are caused due to due to mutated Lymphocyte (a type of white blood cells) from gamma radiation. We are considering a small population of people living near coastal area where High Background Radiation (HBR) has been repeatedly shown an increase in the frequency of chromosome aberration in the circulating lymphocytes of exposed person of has leads to unconditional growth of cells. In existing system called LCS (Longest continues subsequence) algorithm is more time consuming and less efficient for finding longest common subsequence and for aligning the sequences. So here we are going to develop an innovative approach for finding the optimal sequence alignment to reduce the time and space complexity and increase the efficiency of sequence alignment in the large data set. Here we are using innovative longest continues matching sequence queue method(LCMSQ) to reduce the execution of time and making it cost effective for optimal sequence alignment. The LCMSQ uses the split method when a match occurs, the sequence split into the left and the right part where we consider the right part of the sequence for finding another match. As a result, we would obtain a maximum possible number of matches when traversed. So here we are aligning the lymphoma sequence using the resultant longest sequence queue data. After aligning the sequence, we would find the various stages of lymphoma based on a number of matches, mismatches, and gaps. Through this work, we are comparing the normal sequence and affected sequence of lymphoma for predicting the different stages of lymphoma.

KEYWORDS: *Lymphoma, High Background Radiation, Longest Continues Subsequence*



BIPIN NAIR B J*

Department of Computer Science, Amrita School of Arts and Sciences,
Amrita Vishwa Vidyapeetham, Amrita University, Mysuru Campus, Karnataka, India

***Corresponding Author**

Received on : 14-01-2017
Revised and Accepted on : 24-03-2017
DOI: http://dx.doi.org/10.22376/ijpbs.2017.8.2.b532-541

INTRODUCTION

In Coastal area of Kerala especially among fishermen community, there is no proper awareness about Non-Hodgkin lymphoma. The uncontrolled growth of lymphocytes lead to lymphoma which is a disease with less survival rate. Especially in coastal areas of Kerala which have a high rate of Thorium-containing Monazite sand that emit gamma radiation causing lymphoma to the people nearby. In our work, we compare two DNA's or nucleotides sequence i.e. a normal and an affected sequence of lymphoma. We collect the dataset from PDB and analyze them using an innovative longest continues matching sequence with a split method for finding the maximum matching sequence queue, which can be used to align the sequence. Here we can find the total number of matches, mismatches, gap, insertion, and deletion to regain the affected sequence. We are comparing aligned sequence of each person to find out the various stages of lymphoma. For a large number of sequence, the existing algorithm is not accurate for huge data set. Needle Man-Wunsch algorithm, LCS, waterman smith algorithm, but in the innovative method we can find the best case sequence alignment and predict various stages of lymphoma. In proposed method we are using the new approach for sequence alignment but, Needleman-Wunsch algorithm is one of the first application of dynamic programming to compare biological sequences and referred to as the optimal matching algorithm. It is a technique used for global alignment as well as for comparing all sequence. The main disadvantage of needle man Wunsch algorithm and waterman smith is that it is not feasible for large data set and constructing a matrix along with dynamic backtracking for sequence alignment. In existing work, there is no efficient technique for aligning the lymphoma affected sequence. Existing work is using some of the algorithms like Needle man Wunsch, waterman smith, LCS and pairwise alignment etc. which is less efficient, time consuming for execution and space complexity is higher than the proposed algorithm. The proposed algorithm ensures high efficiency and less time complexity when compared to the existing algorithm because we are using substring matching and split method to reduce and avoid the unwanted sequence searching and finding the best maximum possible sequence matching queue. This algorithm is suitable for aligning the lymphoma affected sequence and predicting the stages of lymphoma. An innovative approach for aligning the abnormal sequence to find the various stages of lymphoma based on the various parameters like matches, mismatches, gap and number of insertion is required to regain the lymphoma and cost of deletion will be considered. The objective is to reduce the time and space complexity and making an efficient way for that comparing the various stages of lymphoma using affected sequence and compare. Finally predicting the best case approach for Lymphoma.

Preclinical Evaluation of the Novel BTK Inhibitor Acalabrutinib In Canine Models of B-Cell Non-Hodgkin Lymphoma by Bonnie K. Harrington [1] explained about Evaluated ACA ibrutinib in impulsively happening canine lymphoma, a model of B-cell malignancy similar to human diffuse large B-cell lymphoma using the Trizol method, RMA and Median polish algorithm which gives the result as similar pathway inhibition. Lois B. Travis [2] proposed a method to identify Hodgkin's disease using Radiotherapy and Dosimetry. They explains about Lung Cancer following Chemotherapy and radiotherapy based on Hodgkin's disease. Patrice Gouet [3] used a program called ESPrit to calculate the homology score by columns of residues and sort this calculation by a group of the sequence. It is possible to obtain an output from different files of aligned sequence. Dr.D.Chandrakala [4] proposed a method to calculate the identity and similarity score using the Needleman-Wunsch algorithm to find the matches, mismatches, indel and then represent it using dot-matrix.Julie D.Thompson [5] used the Dynamic programming methods to align two sequences. This will ensure the optimal alignment and also demonstrate the performance of CLUSTAL W.Russell F. Doolittle [6] shows an innovative study about protein evolution based on the reconstruction of past events that give rise to the inventory of protein in existence. Sara A.Shehab [7] proposed a method to test the execution time using the existing algorithm Needleman-Wunsch algorithm, Smith-Waterman, and a proposed new algorithm DASA which shows the same optimal alignment with lesser execution time. Zhou, Z. M [8] proposed an algorithm that calculates the optimal alignment by using dynamic programming. Using multiple algorithms they are comparing the performance of pairwise protein sequence alignment.Sabri, A., Saliman [9] compared local alignment and global alignment using different sequence alignment algorithm and shows the best method to reduce space complexity. Paramita Basak Upama [10] proposed a smart sequence alignment algorithm applying DNA replication which uses the method of Smith-waterman algorithm with trace backing of Needleman-Wunsch algorithm to fill the matrix. Lisa Mullan [11] demonstrated pairwise sequence comparison with other algorithm based on the performance. Arthur M.Lesk1e [12] used a method called variable gap weights for bringing the structural information to address the problem of distantly related sequence alignment. Bipin Nair B J,K Kamarudheen,Ranjith H S [13] the system is used to identify the presence of factor ix gene in DNA sequence using vector and LCCS sequence alignment approach, which is faster and efficient for identifying the LCCS sequence. The impact of genetic operators in solving multiple protein sequence alignment is introduced by Manish Kumar [14]. He proposed a method for sequence alignment by improving the genetic operators of genetic algorithm. It shows better running time and the algorithm increases its solution quality. Unique presentation of Non-Hodkin's Lymphoma is presented by Sumitha R [15] presenting an isolated ENT manifestation of Non-Hodkin's Lymphoma. She explained about five stages of Lymphoma and its treatment based on the drug and shows the survival rate.A.G Hari Narayan proposed a Map Reducing Architecture to find the optimal conformation of a protein using HP model[16]. M.V Judy introduced a new algorithm that shows the hydrophobic interactions. Which has a major role in drug design[17].

METHODOLOGY

Suppose we have two nucleotides or DNA sequence, the normal and affected lymphoma sequence. If a mutation occurs in sequences then the person is affected by lymphoma. For finding the mutated sequence, we have to align the sequence based on the normal sequence. After aligning the sequence we will get the number of matches and mismatches and gaps, based on this we will design the drug for regaining the sequence depending on the various measures. In our method, we can align the sequence and find the various stages based on the comparison of matches, mismatches and gap. The proposed algorithm is used to find the longest continues occurring matching sequence which will be considered as the maximum matches occurred in sequence. In this work the split method is used to reduce the unwanted search the element for reduce the storage space. Here we will start the traversal from sequence 1. We will take the first sequence with the first element then will check whether the element is a substring of sequence 2 with the second sequence, If the sequence is found to be matching then we will add the matching element to the queue and we will update the matching element with "-" symbol. Then split the sequence and take the second part of further operation, remove the sequence before this "-" symbol because we skip left part of the sequence. Here we are finding the highest continues matching sequence and if we are using this technique for finding continues occurring sequence, then the performance of algorithm will increase and we don't want to start with beginning of sequence. We will continue searching for the right part of a sequence. After getting the longest matching sequence we will use longest continuous matching queue (LCMQ) considers for sequence alignment. After the alignment we will predict the stage based on the number matches, if matches are high then we will consider the case as the first stages of lymphoma, the process will continue according to the parameter values for the further stages of lymphoma.

FLOW DIAGRAM

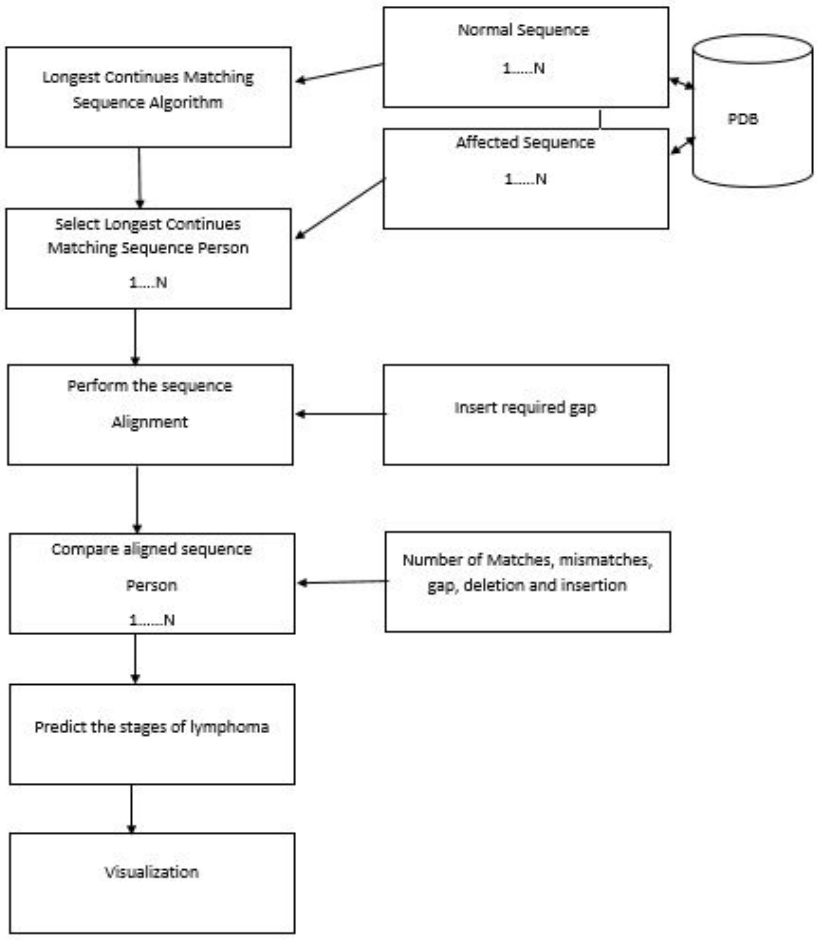


Figure 1
Flow diagram

DATASET

The data set we have taken is from PDB as well as from Kollam Amrita hospital. The dataset contains the 10000 to 1000000 normal and abnormal sequence. The affected sequence is like lymphoma affected persons.

ALGORITHM

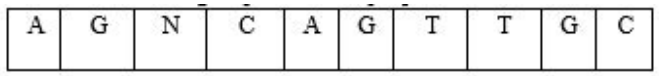
We will take the two sequences which are a normal and an abnormal sequence (i.e.Lymphoma affected sequence).

```
Input: seq1, seq2 for lymphoma
Output: the longest continues matching sequence queue can be identified.
N □ seq1.length
M □ seq2.length
Longest_Match_Length □ 0, I □ 0
Longest_Continues_Match=""
If N ==0 OR M ==0 Then
Return "invalid"
Else
Mid □ N/2
while I < N
do
If I == mid then
If longest_match_length > mid Then
Break
End
End
S1 □ seq1[I…N]
S2 □ seq2
Matching_seq=" "
While S1.length != 0 AND S2.length != 0
Do
First □ 0
If S1 does not end AND S2 not end then
Do
IsNot_Substring □ false
If S1[first] is substring of S2 then
Matching_seq □ S1[first]
Index □ S1[first] in S2
S2[index] □ '-'
Split_seq □ s2.split('-')
S2 □ Split_seq [1]
S1 □ S1[First + 1 … length]
Else
If First +1 < N then
IsNot_Substring □ true
S1□ S1[First+1..length]
End if
While IsNot_Substring == true
Done
If S2 Not end AND S1 is Not end then
IsNot_Substring □ false
If S2[first] is substring of S1 then
Matching_seq □ S2[first]
Index □ S2[first] in S1
S1[index] □ '-'
Split_seq □ s1.split('-')
S1 □ Split_seq [1]
S2 □ S2[First+1 … length]
Else
If First+1 < N then
IsNot_Substring = true
S2 □ S2[First+1..length]
End if
While IsNot_Substring == true
Done
Done
if Longest_Match_Length < Matching_seq.length
Then
Longest_Continues_Match □ Matching_seq
Longest_Match_Length □ Matching_seq.length
End
I □ I + 1
Done
Return Longest_Continues_Match
```

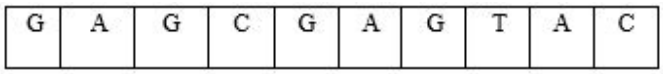
OBSERVATIONS AND RESULT

The existing work based on LCS and Needle Man Wunsch algorithm, which is used for sequence alignment and has high time and space complexity for large data set, when compared to LCMSQ. The LCMSQ algorithm is much faster than existing algorithm I.e. we do not want to construct N*M matrix here as it is consuming more space for creating matrix and backtracking. But the proposed work consumes less memory space for the execution of LCMSQ, which can be applied for the large data set.

Here is an example for finding the longest continues matching sequence which is used in proposed method



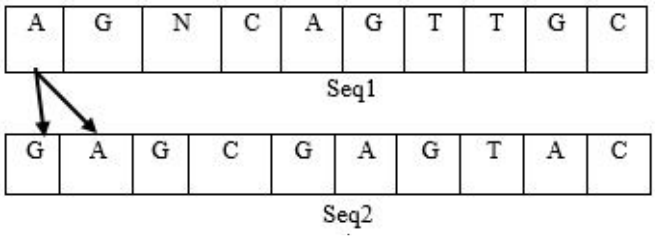
This is normal sequence Seq1



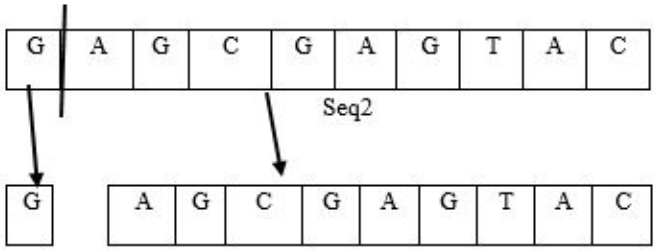
Affected sequence Seq2

We take two sequences seq1 and seq2 for finding the longest continues matches for optimal sequence alignment.

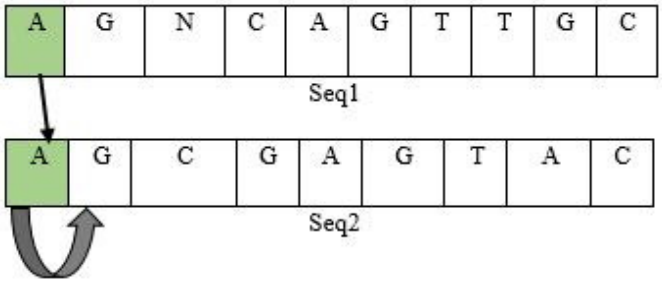
Let us start with the first element of seq1 and check whether the given element is a substring of Seq2.



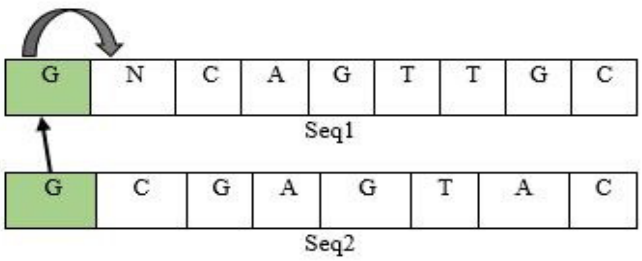
if we are checking whether 'A' is a substring of Seq2, here A is a substring of Seq2 then we are going to split the sequence seq2 into two parts where one is shown on left side and another one is on the right side, we will consider the right part for further operation.



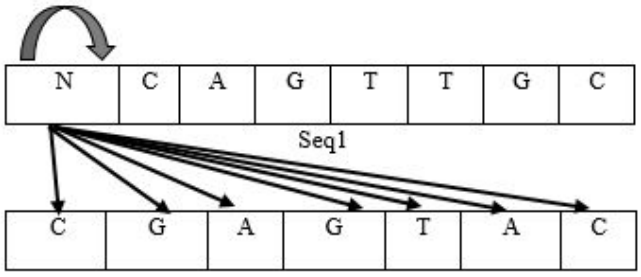
Left part right part of seq2



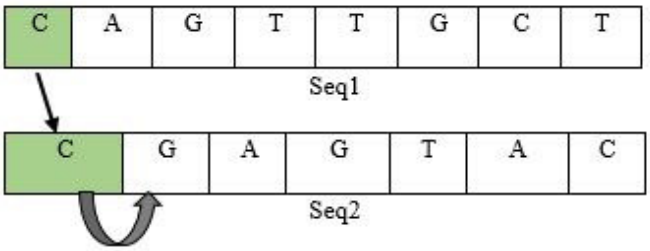
In the above figure if the first element of seq1 and seq2 are matching then we will put in a queue for longest continues matching sequence. If any matches occur then we increment the position to the right side which is represented by a symbol " ↺ "or " ↷ "



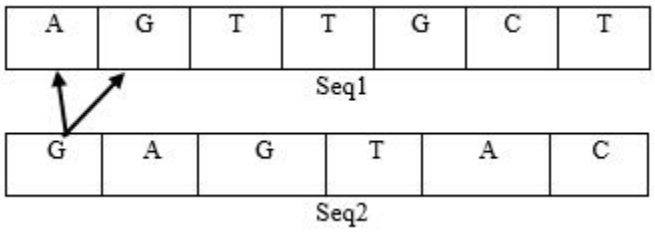
Here ‘G’ is the searching for an element in Seq2. Check the ‘G’ is a substring of Seq1, if ‘G’ is matching with Seq1 then we will store on queue list then we will take the right part of the sequence. If any matches occur then we will fix the next element as searching element.



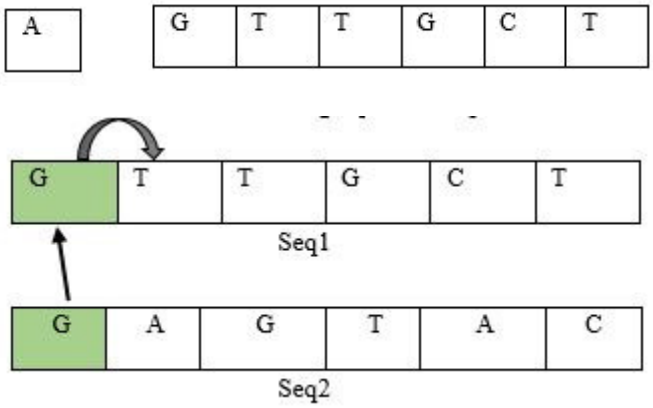
Here ‘N’ is not a substring of Seq2 then starting pointer of seq1 is pointing to next position that is C in Seq1.



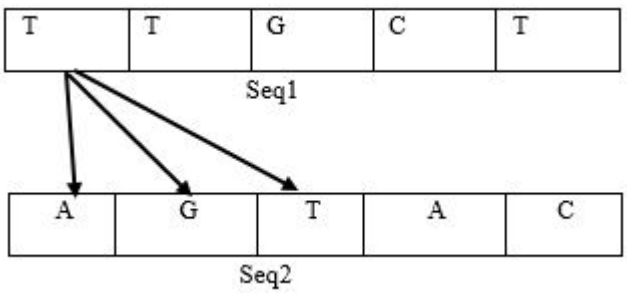
Here we found ‘C’ is a substring of Seq2 then we will split the sequence seq2 and store the right part of the sequence.



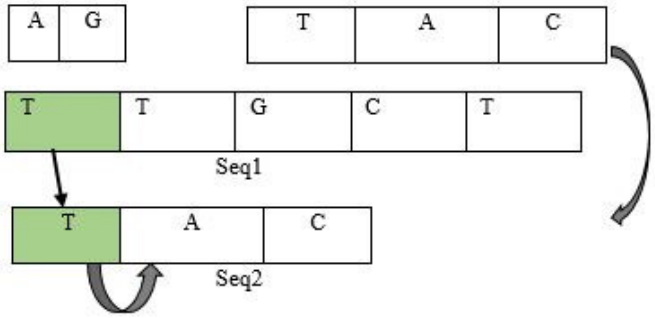
Her ‘G’ is a substring of Seq1 then we are going to split the Seq1 into two part.



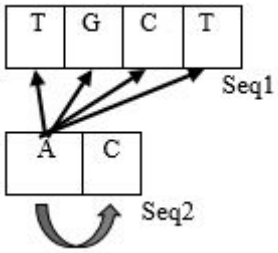
The ‘G’ of seq2 is substring and we store the ‘G’ into queue list.



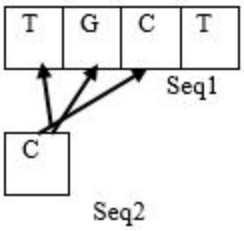
We are checking whether ‘T’ is substring of seq2, yes we found that it is substring then we are going to split the seq2 into two parts



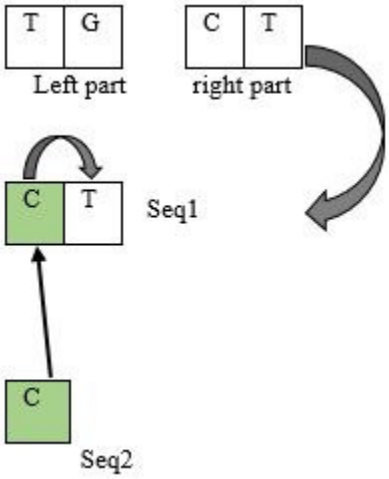
The Same process continue here



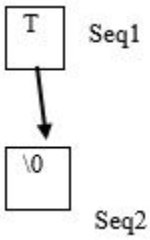
'A' is in the first location of seq2 then check the 'A' is a substring of Seq1 or not, here 'A' is not matching then remove the 'A' from Seq2.



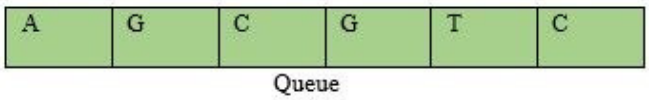
Here 'C' is substring of seq1 and seq1 divide into two parts



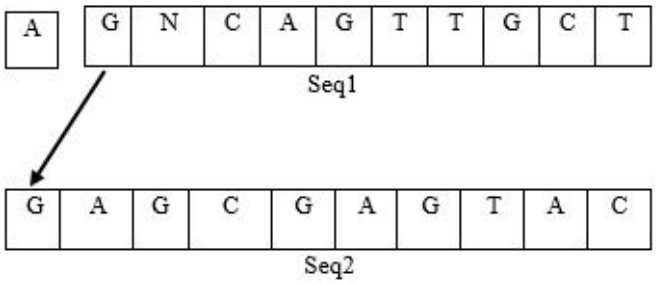
The 'C' is a substring of seq1 and 'C' added to the queue. After this operation seq1 contain the only 'T' and in seq2 contain a null value.



After the first traversal, we will get the longest continues matching sequence queue.



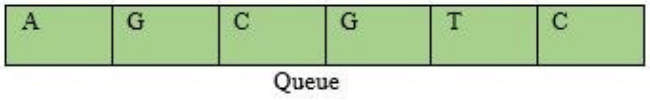
If we start off with the first element of Seq1, we will get the longest continues matching sequence is in the queue. After the first traversal, we start with the next element of Seq1 with Seq2 for finding the maximum longest continues sequence.



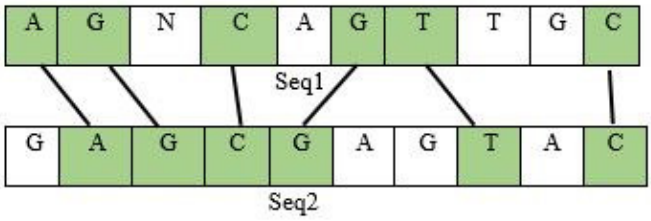
The traversal of Seq1 will continue,
N = Seq1.length
Mid =N/2
If I == mid then
If longest_match_length > mid then
Break
End if
End if

we will traverse the sequence until and unless 'I' reach in the mid position. If we reached in mid position we would check the longest_match_length and if it is found to be greater than mid then we stop the execution, if not, It will continue till the end of the sequence. For checking the above conditions, after mid value, the matching continues and sequence is N/2. If the longest_match_length is greater than N/2 we don't want to check because of the longest_match_length greater than N/2.

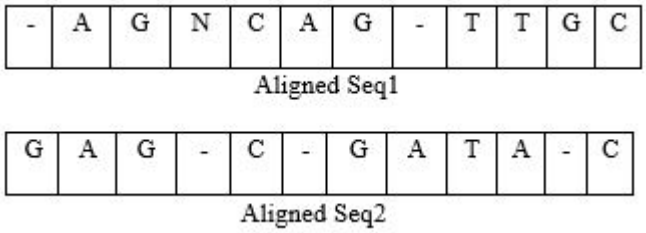
After execution of the proposed algorithm will get the longest continues sequence match we get is



We are going to align the sequence based on the longest continues matching sequence.



The matching sequence is colored.
The aligned sequence will be as follows



After alignment, we are going to find the matches, gap, and a number of operations like deletion and insertion which is required from the aligned sequence. In this work, we will analyze around 5 to 6 person for identifying the various stages of lymphoma. We will take the normal and the abnormal sequence from PDB and then perform the proposed algorithm to find the longest continues matching sequence for sequence alignment.

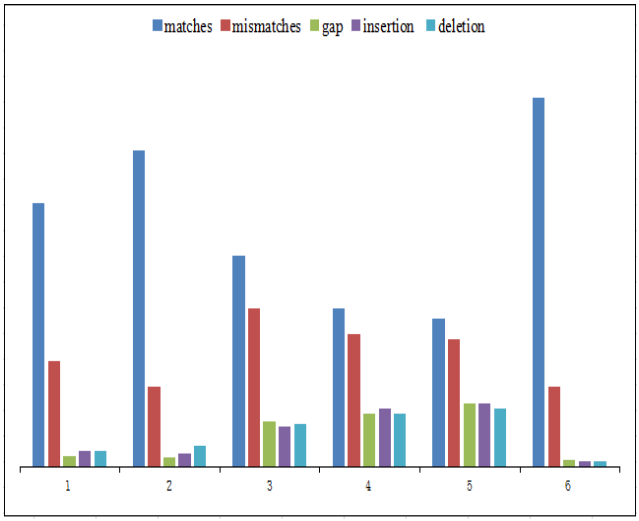


Figure 2
Identifying the stages of lymphoma.

Figure 2 is depicting the number of matches, mismatches, gap, insertion, and deletion occurred in 6 sample data (the above graph is depicting the measurement from 1 to 6 peoples sequence details). We will align the sequences of 6 people's data set, and then calculates the various parameters which is mentioned in the graph. From the result we are concluding highest matches (100000), mismatches (10000), gap (250), insertion (220) and deletion (30). These are the kind of measurements used to find the various stages of lymphoma.

Table1
Comparison of Needle man Wunsch,LCS and LCMSQ

Algorithm	Space complexity	Time complexity
Needleman Wunsch	O(M.N)	O(M+N)
LCS	O(MN)	O(M+N)
LCMSQ	O(N log N)	O(N)

CONCLUSION

Lymphoma occurs due to the mutation in lymphocyte. An affected sequence provides the wrong pattern than a normal sequence, which can cause to lymphoma. In our work, we propose a novel approach to identify the various stages of Lymphoma. We have analyzed our proposed longest continues matching queue sequence (LCMQ) with the splitting algorithm which is an efficient method for sequence alignment. By using this we have identified that the proposed algorithm is comparatively efficient than the needle man Wunsch global sequence alignment. It takes less execution time and space complexity to get the final optimal alignment. In future, we can improve the efficiency of the algorithm using various recursive approach and visualize the various stage of lymphoma using specific tools. We can compare the drug interaction of lymphoma affected protein and drug structure.

CONFLICT OF INTEREST

Conflict of interest declared none.

REFERENCES

1. Harrington BK, Gardner HL, Izumi R, Hamdy A, Rothbaum W, Coombes KR, Covey T, Kaptein A, Gulrajani M, Van Lith B, Krejsa C. Preclinical Evaluation of the Novel BTK Inhibitor Acalabrutinib in Canine Models of B-Cell Non-Hodgkin Lymphoma. PLoS ONE. 2016 Jul 19;11(7).
2. Travis LB, Gospodarowicz M, Curtis RE, Clarke EA, Andersson M, Glimelius B, Joensuu T, Lynch CF, van Leeuwen FE, Holowaty E, Storm H. Lung cancer following chemotherapy and radiotherapy for Hodgkin's disease. Journal of the National Cancer Institute. 2002 Feb 6;94(3):182-92.
3. Gouet P, Courcelle E, Stuart DI. ESPript: analysis of multiple sequence alignments in PostScript. Bioinformatics. 1999 Apr 1;15(4):305-8.
4. Chandrakala D, Kumar TS, Preethi S, Sowmya D. Optimization of Process Parameters of Global Sequence Alignment Based Dynamic Program-an Approach to Enhance the Sensitivity of Alignment.
5. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic acids research. 1994 Nov 11;22(22):4673-80. [VI] Lee C, Grasso C, Sharlow MF. Multiple sequence alignment using partial order graphs. Bioinformatics. 2002 Mar 1;18(3):452-64.

6. Doolittle RF. Similar amino acid sequences: chance or common ancestry. Science. 1981 Oct 9;214(4517):149-59.
7. Shehab SA, Keshk A, Mahgoub H. Fast dynamic algorithm for sequence alignment based on bioinformatics. Proceedings of the International Journal of Computer Applications (0975–8887) Volume. 2012 Jan.
8. Satra R, Kusuma WA, Sukoco H. Accelerating computation of DNA multiple sequence alignment in distributed environment. Telkomnika Indonesian Journal of Electrical Engineering. 2014 Dec 1;12(12):8278-85.
9. Sabri A, Saliman NF, Al Junid SA, Majid ZA, Tahir NM. A Comparison of Optimal Path Trace Back Sizing Using Genetic Algorithm (GA). Int. J. Simulation--Systems, Sci. Technol. 2013 Dec 1;14(6).
10. Upama PB, Khan JT, Yasmin Z, Zemim F, Sakib N. A Noble Approach on Bioinformatics: Smart Sequence Alignment Algorithm applying DNA Replication (SSAADR). International Journal of Applied Information Systems. 2014;8(1):23-8.
11. Mullan L. Pairwise sequence alignment—it's all about us!. Briefings in bioinformatics. 2006 Mar 1;7(1):113-5.
12. Lesk1e2 AM, Levitt M, Chothia1a4 C. Alignment of the amino acid sequences of distantly related proteins using variable gap penalties. Protein Engineering. 1986;1(1):77-8.
13. Nair BB, Khamarudheen KS, Ranjitha HS. AN APPROACH FOR IDENTIFYING THE PRESENCE OF FACTOR IX GENE IN DNA SEQUENCES USING POSITION VECTOR ANN. JTAIT 2016 May 1;87(3):396.
14. KUMAR M. THE IMPACT OF GENETIC OPERATORS IN SOLVING MULTIPLE PROTEIN SEQUENCE ALIGNMENT.
15. ABERNA GOVAR AND SR. UNIQUE PRESENTATION OF NON-HODGKIN'S LYMPHOMA. Int J Pharma Bio Sci. 2013;4(4):478–82.
16. Judy MV, Ravichandran KS, Murugesan K. A multi-objective evolutionary algorithm for protein structure prediction with immune operators. Computer methods in biomechanics and biomedical engineering. 2009 Aug1;12(4):407-13.

Reviewers of this article

DR.C.N.Ravikumar

HEAD & Professor, Sri Shivarathreeshwara Science and Technology University, Campus Roads, Mysuru, Karnataka 570006, India

Prof. Y. Prapurna Chandra Rao

Assistant Proffessor, KLE University, Belgaum, Karnataka

Prof.Dr.K.Suriaprabha

Asst. Editor , International Journal of Pharma and Bio sciences.

Prof.P.Muthuprasanna

Managing Editor , International Journal of Pharma and Bio sciences.

We sincerely thank the above reviewers for peer reviewing the manuscript