Array Manipulation And Matrix-Tree Method To Identify High Concentration Regions (HCRs)

Rachana Arora , Joby George

Abstract-Sequence Alignment and Analysis is one of the most important applications of bioinformatics. It involves alignment a pair or more sequences with each other and identify a common pattern that would ultimately lead to conclusions of homology or dissimilarity. A number of algorithms that make use of dynamic programming to perform alignment between sequences are available. One of their main disadvantages is that they involve complicated computations and backtracking methods that are difficult to implement. This paper describes a much simpler method to identify common regions in 2 sequences and align them based on the density of common sequences identified.

Index Terms-Sequence Alignment, HCRs, BLAST, MSA

1 Introduction

Any living organisms possess a genetic material that defines its identity. This genetic material can be RNA [1] for simpler organisms like viruses) and DNA[2] (for more complex organisms like humans). Bioinformatics combines the 3 entirely different science fields-computer science, mathematics and biology. The genetic materials of organisms are analyzed using computers, using mathematical techniques. The main composition of these genetic materials is nucleotides, which are a sequence of ATGC base pairs in DNA and AUGC base pairs in RNA. The arrangement of these 4 nucleotides in different fashion leads to wide diversity of living found on the earth. Moreover, genetic disorders arise due to some change in the order in which they are found in the organisms. The result is that any DNA sequences extracted, needs to be aligned with some sequences, either to identify any possible mutation or to identify it's similarity with the already known sequences. Accessing genomic information and synthesizing it for the discovery of new knowledge is the main component of biological research. Various databanks are available that store the extracted sequences of various organisms. Relationships between these sequences are usually discovered by aligning them together and assigning this alignment a score. There are two main types of sequence alignment. Pair-wise sequence alignment only compares two sequences at a time and multiple sequence alignment compares many sequences in one go. Two important algorithms for aligning pairs of sequences are the Needleman-Wunsch[4] algorithm and the Smith-Waterman[5] algorithm. The former one performs what is called global alignment, while latter performs local alignment.

2 Existing Works

Global sequence alignment tries to find the best alignment between an entire sequence S1 and another entire sequence S2. The Needleman-Wunsch[3] algorithm is used for computing global alignments. A two-dimensional table with one sequence along the top and one along the left side is created. At arriving each cell score is computed for it in one of three ways: From the cell above, which corresponds to aligning the character to the left with a space. From the cell to the left, which corresponds to aligning the character above with a space. From the cell diagonally to the above-left, which corresponds to aligning the characters to the left and above (which might or might not match). Once all the cells are filled, a process called backtracking is employed that traces the maximum-cost path from last to first column in the table. With local sequence alignment, it is not constrained to aligning the whole of both sequences; user can just use parts of each to obtain a maximum score. The Smith-Watermann[4] Algorithm is used to perform local alignment. The only difference with previous algorithm is the process of backtracking. Backtracking continues till the minimum score (0) is encountered in any of the cells, and that sub portion is returned as local alignment of sequences. BLAST (Basic Local Alignment Search Tool)[5] is the most popularly used algorithm for performing local alignment, using Smith-Watermann Algorithm. The FASTA[6] is the algorithm for global alignment. The basic format in which the sequences are stored is in .fasta format. The pairwise alignment is the first step towards Multiple Sequence Alignment (MSA)[7]. It includes aligning 3 or more sequences simultaneously. There are various tools available for this purpose, like ClustalW[8], T-Coffee [9], Muscle[10] being the most popular ones. The Muscle algorithm provides most accurate results, while ClustalW gives only approximate results in optimal time.

3 Proposed Work

While the pairwise alignment algorithm provides either global or local alignments, we proposes a methodology that combines both these methods. While identification of HCRs is similar to performing the local search, the alignment of sequences gives the sense of global alignment. The following algorithms are run recursively to identify the HCRs and finally align them optimally by inserting gaps.

Algorithm 1.1. *Tree-Match Algorithm For Aligning Sequences*

1. Read the 2 sequences: Sequence 1 and Sequence 2 to be aligned from a FASTA file.

- *Rachana Arora is currently pursuing Masters Degree program in Computer Science and Engg. M.A College of Engg. Kerala ,India. E-mail: rachanaarora1992@gmail.com*
- *Joby George: Proffessor in Computer Science and Engg. M.A College of Engg. Kerala ,India. E-mail: jobygeo@hotmail.com*

2. Insert them into 2 separate arrays
3. Make the 2 entire sequences to be aligned as the root node of a tree T.
4. Call Algorithm 1.2, the return values are:
5. Longest Aligned portion
6. Start and end index of the aligned regions in sequence 1 and 2, start_1, end_1 for sequence 1 and start_2, end_2 for sequence 2(corresponding array indexes).
7. Split the 2 sequences into subparts:
8. 1st sequence's left subpart will be from array_index 0 to start_1, 2nd subpart will be end_2 to last index. Do the same for sequence 2.
9. Insert them as left and right subtrees of the root node.

Note: Each node will have subparts of both the sequences.

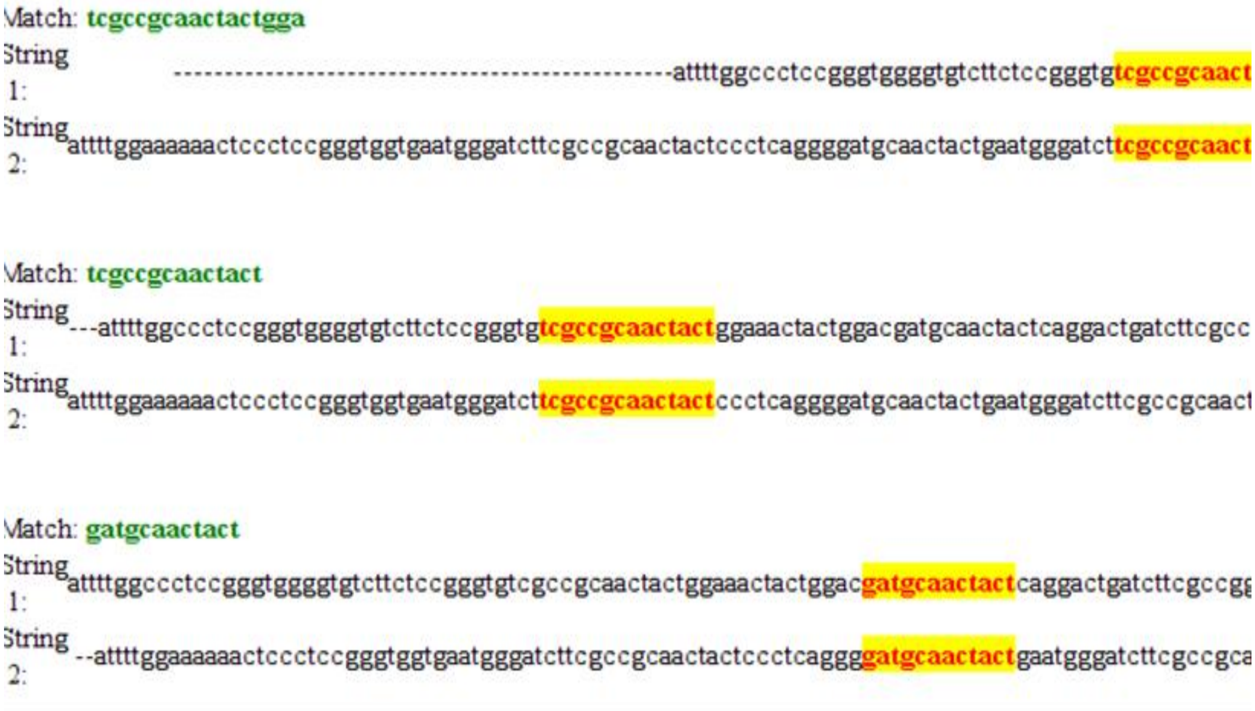
10. Call Algorithm 1.2 recursively, until all the subparts gets aligned.
11. Display all aligned subsections that have length>3.

Algorithm 1.2: *Match-Regions Identification*

1. Store both the sequences in 2 separate arrays.
2. Split the array elements into character pairs. Name split array of Sequence 1 as seq1.
3. Split sequence 2 in two different formats:
4. In second split, there will be a pair of characters from the beginning(seq2b).
5. Start splitting with single element in 1st position(seq2a) and from next position, a pair of characters
6. Create a 2-dimensional matrix and insert seq1 as the 1st row of the matrix
7. Insert seq2a as the 1st column and seq2b as last column.
8. 6. i=1,j=1
9. If seq1[i][j]==seq2a[i][j] OR seq1[i][j]==seq2b[i][j]:
    1. goto 7 else goto 13
10. Start_pos1= i, end_pos1=j; Start_pos2= I, end_pos2=j
11. matchedString+=seq1[i][j]
12. j++
13. Go to 6 until seq1[i][j]!=seq2a/b[i][j].
14. If seq1[i][j]!=seq2a/b[i][j]
15. Split the pair and check if 1st element in seq1[i][j]==1st element in seq2b OR 2nd element in seq1[i][j]==2nd element in seq2a[i][j]
16. Change values in end_pos1&end_pos2 accordingly
17. Else
18. Store value in matchedString, start_pos1, start_pos2end_pos1,end_pos2 to an separate array match[i][k++]
19. Increment i and goto 6
20. Repeat until i<seq1.length and j<MAX(seq2a.length,seq2b.length)
21. Return array match to Algorithm 1.1

4 Experimental Results

In the implementation of this paper, the pairwise alignment method of 2 input sequences is performed. The HCRs are identified and at the same time near global alignment results are produced. The experimental setup runs even on a simple PC that has Java platform. The machine has 4GB RAM and runs on Intel Core i5 processor. The data sequences required are synthetically created, and within a length of 100 base pairs, the results are obtained in a click of a button, while time for aligning increases as the length of sequences increases. However, the HCRs are retrieved in the order of seconds, for almost any possible length of sequences. The following figures depicts the output results on aligning two synthetic sequences in FASTA format. The HCRs are the longest matched areas in both the sequences, while the complete alignment of the sequences is done preserving these HCRs.

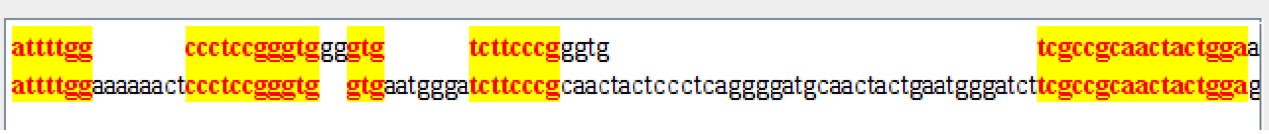


```
Match: tcgccgcaactactgga
String 1:   --------------------------------------------attttggccctccgggtggggtgtcttctccgggtgtcgccgcaact
String 2: attttggaaaaaactccctccgggtggtgaatgggatcttcgccgcaactactccctcaggggatgcaactactgaatgggatcttcgccgcaact

Match: tcgccgcaactact
String 1: ---attttggccctccgggtggggtgtcttctccgggtgtcgccgcaactactggaaactactggacgatgcaactactcaggactgatcttcgcc
String 2: attttggaaaaaactccctccgggtggtgaatgggatcttcgccgcaactactccctcaggggatgcaactactgaatgggatcttcgccgcaact

Match: gatgcaactact
String 1: attttggccctccgggtggggtgtcttctccgggtgtcgccgcaactactggaaactactggacgatgcaactactcaggactgatcttcgccgg
String 2: --attttggaaaaaactccctccgggtggtgaatgggatcttcgccgcaactactccctcagggggatgcaactactgaatgggatcttcgccgca
```

__Fig. 1:__ List of HCRs identified from the given sequences



```
attttgg        ccctccgggtggggtg       tcttcccgggtg                                    tcgccgcaactactggaa
attttggaaaaaactccctccgggtg  gtgaatgggatcttcccgcaactactccctcaggggatgcaactactgaatgggatcttcgccgcaactactggag
```

__Fig. 2:__ Final aligned sequences, with gaps inserted and depicting the HCRs

5 Conclusions And Future Work

This paper proposes a new methodology for identifying the HCRs among 2 sequences and aligning them preserving these HCRs. The future work includes identifying HCRs among different species of organisms and assigning the unclassified sequences to the known species. Another work proposed is to include the algorithm as part of the pair-wise alignment step in MSA.

REFERENCES

[1] RNA Building Blocks, Available at: https://www.khanacademy.org/partner-content/nova/rnawondermolecule/a/rna-the-basics

[2] Introduction to DNA, Available at: http:// seqcore.brcf.med.umich.edu/doc/educ/dnapr/pg1.html

[3] S. B. Needleman and C. D. Wunsch, “A general method applicable to the search for similarities in the amino acid sequence of two proteins.,” Journal of molecular biology, vol. 48, no. 3, pp. 443–53, Mar. 1970.

[4] M. S. Waterman, “Identification of Common Molecular Subsequences Identification of Common Molecular Subsequences,” pp. 195–197, 1981.

[5] Altschul et al, "Basic Local Alignment Search Tools,” Journal of Molecular Biology, pp. 403-410

[6] FASTA- http:// bioinformatics.unl.edu/ pages/ tools/fasta.html

[7] [Iain M Wallace et al, Evaluation of Iterative Algorithms for Multiple Sequence Alignment, Bioinformatics Oxford Journal, Vol 21 No 8 2005

[8] Thompson JD, Higgins DG, Gibson TJ," CLUSTAL W- Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice" Nucleic Acids Res 22:4673-4680. 1994.

[9] CeAdric Notredame, Desmond G. Higgins and Jaap Heringa, Journal of Molecular Biology 2000, “T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment”, pp.205-217

[10] Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research, 32, 1792-1797.