

DNA Sequence Optimizations using Parallel Architecture

Sumashri V. ¹, Kajal Bangera ¹, B. Neelima ²PG Student, Department of Computer Science, NMAMIT College, Nitte, Udupi District, Karnataka, India¹Professor, Department of Computer Science, NMAMIT College, Nitte, Udupi District, Karnataka, India²

ABSTRACT: Sequence alignment has its own importance in the field of bioinformatics. Sequencing is referred to the process of aligning two sequences to find the area of similarity. Needleman Wunsch algorithm is one such algorithm used to deal with the alignment of Deoxyribo Nucleic Acid (DNA). The traditional Needleman Wunsch gives the accurate results but takes time to compute the results. So parallel implementation of the algorithm is carried out for the better performance.

KEYWORDS: DNA sequence, Needleman Wunsch, Sequence alignment.

I. INTRODUCTION

Bioinformatics is an interdisciplinary branch which deals with the biological data to produce meaningful information. It helps in handling and sorting of biological information. It acts as a bridge between biological science and computer science. It detects valuable data from huge raw data. These data is applied in the field of medicine, agriculture, environmental science and nutrition. Bioinformatics deals with a number of areas, sequence alignment is one such area. Sequencing is fundamentally an important area in bioinformatics. It is an important step in analysing the structure and functional or evolutionary relationship between two sequences. Sequence alignment is a process of aligning two sequences in order to find matches between the DNA sequences. Sequence alignment of DNA has number of applications, one such application is detecting the species the organism belongs to or similarity between two species. It can also be used to detect hereditary diseases within the family. Therefore alignment has great importance in the field of biology.

Alignment is easier on similar sequences, but it becomes very difficult to align dissimilar sequences. Sequence alignment is where the DNA or RNA sequences are arranged in such a manner so as to find the common region of similarity. The DNA is made up of Adenine (A), Guanine (G), Cytosine (C) and Thymine (T) nucleotides. These nucleotides are bounded in the following fashion. 'A' bonds with 'T' and 'C' bonds with 'G'. The DNA sequences are combinations of these four characters. The most similar pattern amongst the sequences are said to be aligned.

Sequence comparison is considered important in bioinformatics. It has become a basic tool in molecular biology. It will help in identifying the structural and functional analysis of sequences which are newly determined. Since new biological data are being generated in a high rate, sequence comparison is required to be done to draw evolutionary and functional inference of a new DNA sequence with already existing DNA sequence in the database. The fundamental process in this kind of comparison is sequence alignment.

Sequence alignment algorithms usually compare two sequences at once. This process is required to search the common pattern of characters among two sequences. Similarity can be termed as homology. It can be used as basis for the prediction of function and structure of uncharacterized sequences.

If two sequences are similar then it's understood that they share the common ancestor. If they are not similar then they might have undergone evolutionary changes in the due course of time. It is also possible that the sequences are derived from same ancestor, but have under gone massive changes, that the similarity is not in a recognizable state.

The quality of the alignment is measured through the alignment score. The score is calculated from the number of matches, mismatches and gaps present in the sequence matrix. An optimal alignment algorithm will always aim for the highest alignment score. And it finds the alignments that will maximize the score.

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 9, May 2016

Sequence alignment is of two types. Pairwise alignment and multiple alignment. Alignment between two sequences is pairwise alignment. Aligning more than two sequences is known as multiple sequence alignment. There are number of ways to align two sequences. The alignment can be at the beginning, in the middle or at the end. Two strategies or methods exist for pairwise sequence alignment. They are local and global alignment. Taking small stretch and then progressing is known as local alignment, whereas in global alignment it takes long stretch of pattern that attempts to align each pattern in each sequence. The global sequence alignment will match both the sequences with each other from end to end. Global alignment sequencing is best suited when sequences are of the similar length with a significant number of similarity throughout the sequence. The local sequence alignment will match the sequences that match well with each other, anywhere in the sequence. The local alignment is useful in case of sequences of different lengths and very small chance of similarity.

Pairwise sequence alignment is obtained through three types of methods. They are

- 1) dot matrix method
- 2) dynamic programming method
- 3) word method

Dot matrix or dot plot method is where given two sequences are arranged in matrix order and a dot is placed where the sequences align. If the given two sequences match with each other the line will be drawn diagonally. Dot matrix method is a visual aid method and hence the user will know easily the regions of similarity. Through this method the user can visually identify the similar regions present in the sequences. But this method lacks the sophistication of the dynamic programming method and the word method.

Dynamic programming method can be called as the quantitative and an exhaustive method to find the optimal alignments. Scoring matrix is used in DNA and RNA alignments, but in practice positive score is assigned for the match, negative score for the mismatch and negative for gap penalty. Dynamic programming method will find the optimal alignment for the given set of sequences. This method is extensible for more than two sequences, but the computation becomes slow for the large or extremely long sequences.

Word method is also known as the k-tuple method. It is known as the heuristic method where the optimal solution is not guaranteed but the method will be better than the dynamic program method. This method will be useful in a very large database search. Word method is implemented in the heuristics methods like Basic Local Alignment Search Tool (BLAST) [3] [4] and FASTA programs to search the required sequence.

The paper is been organized as follow: Part II deals with the related work and gives a brief summary on the algorithm Needleman-Wunsch. It also shows how the algorithm works. Section III gives the design of the new idea. Section VI shows the implementation and results. Finally, Section V concludes the paper.

II. RELATED WORK

This section will include the survey on the sequence alignment algorithm that is the Needleman Wunsch algorithm.

Needleman Wunsch

Needleman-Wunsch algorithm [1] is the dynamic programming model which uses global alignment method to align two sequences. It uses the two dimensional matrix to compute the alignment. Each cell in the matrix will represent pairing of letters from each sequence. Score and pointer are the two values used in the cell. Scoring schema is used to calculate the score and pointer will point to the left, diagonal or up of the cell. Alignment starts from the upper left of the matrix and moves towards along the lower right corner of the matrix. Every letter is mapped to either a letter or a gap since it satisfies global alignment rule. The algorithm consists of three steps namely initialization, fill matrix and trace back.

Initialization: In the first step, the first row and column of the matrix is to be filled, as shown in the figure 1. Pointer of cell will point back to the origin, since it's necessary to calculate the global alignment.

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 9, May 2016

	A	G	A	A
A				
T				
A				
G				

Figure 1: Initialization of the sequences

Fill Matrix: In this step all the cells will be filled with scores and pointers. The score is calculated through simple step as shown in the figure 2.

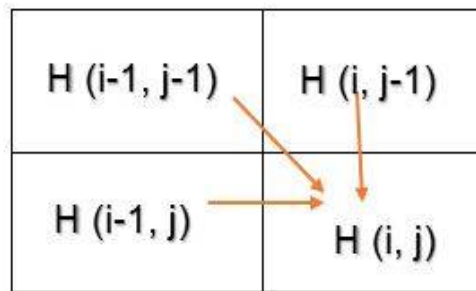


Figure 2: Fill rule of Needleman Wunsch algorithm

Maximum score among match, horizontal and the vertical gap score is used to fill the cell $H(i, j)$. Match score is the total of the diagonal cell score and horizontal gap score is the sum of the cells to the left with gap score whereas vertical gap score is sum of the cells in the top with gap score. The sample fill matrix table is shown in the figure 3.

	A	G	A	A
A	2	-1	2	2
T	-1	-1	-1	-1
A	2	-1	2	2
G	-1	2	-1	-1

Figure 3: Sample fill matrix table

Trace back: It will give the final alignment from the matrix. It is seen it to that the pointer from each of the cells, is pointed back to the origin. The recovery of the alignment process starts from the bottom right corner to the top left corner along the diagonal. Each letter will be aligned to either a letter or to a gap; hyphen is used to represent a gap. Alignment will be done from the end to the start. Because of this, the alignment will be backwards. To get the final alignment, the alignment will be reversed as shown in the figure 4.

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 9, May 2016

	A	G	A	A
A	2	-1	2	2
T	-1	-1	-1	-1
A	2	-1	2	2
G	-1	2	-1	-1

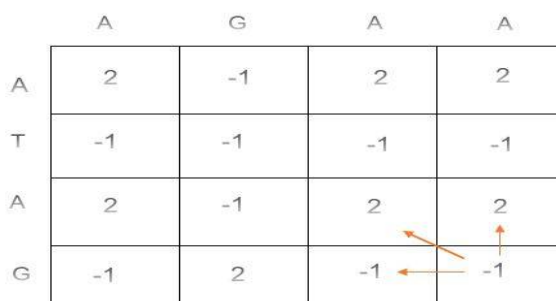


Figure 4: Trace back step of Needleman Wunsch

The Fill matrix has lot of computation to do to calculate the score and this part of the algorithm takes much time for the computation. To get the overall speedup, this part of the algorithm needs to be parallelized. So wavefront method can be introduced to execute the algorithm in parallel. Wavefront method executes the fill matrix as one diagonal at a time.

III. DESIGN

The design of the system with Needleman Wunsch algorithm with wavefront method is shown below in the figure 5.

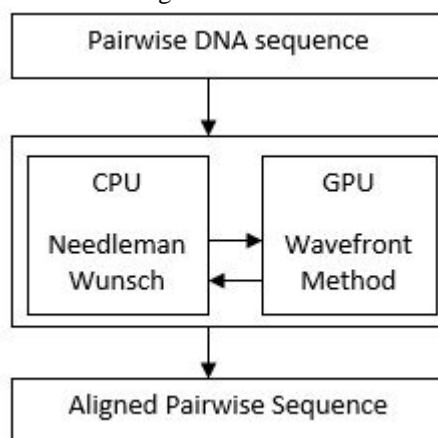


Figure 5: Design of the proposed work

Input for the design is given in the FASTA format. The sample of the FASTA is shown in the figure 6 below.

```

>gi|568815581:c7687550-7668402 Homo sapiens chromosome 17, GRCh38.p2 Primary Assembly
GATGGGATTGGGGTTTTCCCTCCCATGTGCTCAAGACTGGCGCTAAAAGTTTTGAGCTTCTCAAAAGTC
TAGAGCCACCGTCCAGGGAGCAGGTAGCTGCTGGGCTCCGGGGACACTTTCGCTTCGGGCTGGGAGCGTG
CTTCCACGACGGTGACACGCTTCCCTGGATTGGGTAAGCTCCTGACTGAACTTGATGAGTCCTCTCTGA
  
```

Figure 6: FASTA format

The Needleman Wunsch algorithm works in three steps. The initialization, fill matrix and trace back step. The task is divided into different steps. Inputs are given in the form of FASTA files. In sequence reading step, the input file will be read and the database will be loaded. Wave front method [2] is used in the fill matrix of the GPU Needleman Wunsch. It will work on the diagonals and parallelize the alignment [5] [6] as shown in the flow graph figure 7.

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 9, May 2016

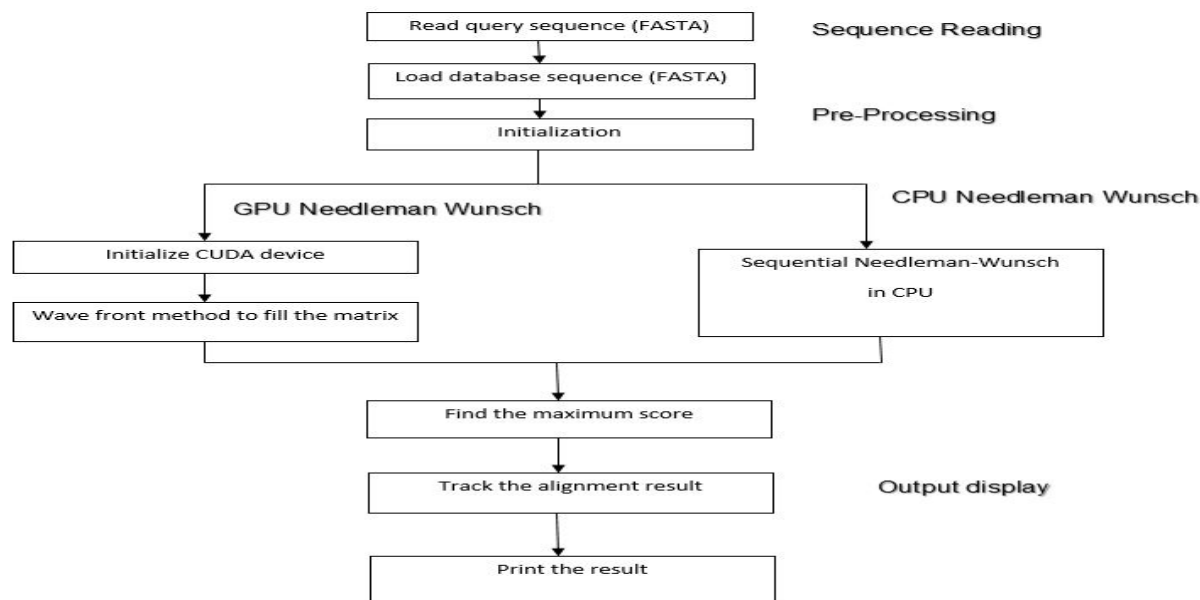


Figure 7: Flow graph of the proposed work.

IV. IMPLEMENTATION AND RESULTS

Work done so far is presented here. Sequential Needleman-Wunsch algorithm is necessary before building the parallel Needleman-Wunsch algorithm. The sequential Needleman-Wunsch algorithm is executed with the inputs of FASTA format. The genes of Homo sapiens and chimpanzee are aligned. It is also executed in OpenMP. And the results obtained are comparatively better than the sequential Needleman Wunsch algorithm. The FASTA files are taken from the National Centre for Biotechnology Information (NCBI). Both the inputs are from the Homo sapiens, chromosome number 6 and 17 are taken as the inputs.

Parallel Programming models:

Pure parallel models use either shared memory approach or distributed memory approach. There are some application programming interfaces (APIs) to encourage the programmer to program in these approaches.

Shared memory:

In shared memory model, a common address space is shared between tasks. Mechanisms such as semaphores and locks can be used to control access to the shared address space. It is an efficient means of passing data between programs. Sometimes memory is also shared between different sections of code of the same program. Depending on perspective; programs may run on a single processor or on multiple separate processors.

OpenMP:

It is an application programming interface that enables shared memory parallelism on multi-platform by using its components namely compiler directives, runtime library routines, environment variables. OpenMP [8] is designed to utilize the shared address space most effectively. OpenMP [9] suits well for those environments where all processors effectively utilize the common memory with low communication overhead in the network. For huge-size problems, large shared memory systems are needed, but it is very expensive to manufacture systems with large address space.

The result of the sequential and parallel Needleman Wunsch algorithm is shown in the figure 8 below.

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Special Issue 9, May 2016

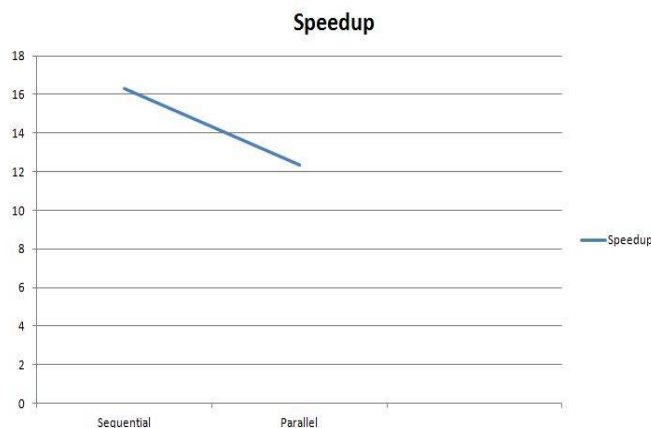


Figure 8: Graph showing the speed up ratio of the sequential and parallel implementation.

V. CONCLUSIONS

Sequence alignment is done using different algorithms. Here we make use of Needleman-Wunsch algorithm which uses dynamic programming approach to align two sequences. The sequential algorithm is successfully implemented in CPU. The program was tested for about ten thousand DNA characters, and the time consumed for the execution was pretty high. The code is also executed in OpenMP. Performance of OpenMP program is better than the sequential execution. As the part of the future work, the Needleman-Wunsch algorithm will be implemented in GPU. Wavefront method will be used in the algorithm to speed up the process.

REFERENCES

- 1] 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. 443-453, March 1970.
- 2] Lipton and D. Lopresti, "A Systolic Array for Rapid String Comparison", Chapel Hill Conference on Very Large Scale Integration, H. Fuchs, ed., Rockville, MD: Computer Science Press, 1985, pp. 363-376.
- 3] T. R. P. Siriwardena ; D. N. Ranasinghe, "Accelerating global sequence alignment using CUDA compatible multi-core GPU", 5th International Conference on Information and Automation for Sustainability (ICIAFs), 2010, pp. 201-206.
- 4] Jung, Sungbo "GPU data-parallel computing of sequence alignment using CUDA", ProQuest Dissertations and Theses; 2008.
- 5] Altschul,"Basic Local Alignment Search Tool", *J.Mol.Bio*, vol 215, pp 403-410, 1990.
- 6] Alba Melo and Edans, "CUDAAlign:Using GPU to Accelarate the comparision of Megabase Genomic Sequences", PPOPP, 2010.
- 7] Alba Melo and Edans "CUDAAlign 3.0: Parallel Biological Sequence Comparision in Large GPU Clusters", IEEE International Symposium on Cluster, Cloud and Grid Computing, 2014.
- 8] Alina Kiessling "An Introduction to parallel programming with OpenMP", 2009.
- 9] Ruud van der Pas "An Introduction to OpenMP" 2005.