International Journal of Technology Research and Management
ISSN (Online): 2348-9006
Vol 3 Issue 5 May 2016

# Sequence Comparison Techniques for Biological Sequence Comparison

Jayshree G Patil
M Tech Scholar Department of Information Technology Mahakal Institute of Technology,
RGPV Bhopal Madhya Pradesh India
jaishreepatil12@gmail.com

**Abstract:** *Comparing DNA sequences is one of the basic tasks in computational biology. In bioinformatics, a sequence alignment is a way of arranging the sequences of DNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a matrix. Gaps are inserted between the residues so that identical or similar characters are aligned in successive columns. The comparison and alignment of DNA and protein sequences are important tasks in molecular biology and bioinformatics. There are several algorithms for sequence comparison. The parallel solution is based on the dynamic programming approach and presents less processing time. In this paper we review different algorithms, the Needleman-Wunsch algorithm, and the first algorithm applying the dynamic programming to comparing biological sequences. Later the Smith-Waterman algorithm, based on dynamic programming, is one of the most fundamental algorithms in bioinformatics***.**

**Keywords:** *Biological Sequence Alignment, dynamic programming, DNA, sequence comparison.*

## 1. INTRODUCTION

Deoxyribonucleic acid or DNA is the molecule of life. It is the chemical code specifying our function, appearance and lineage, and it is unique for each individual. DNA can be seen as the molecular blueprint for the cell. In fact, it contains all the instructions needed to direct cellular activities. It is a linear polymer that is made up of nucleotide units. A nucleotide unit consists of a base, a deoxyribose sugar, and a phosphate. There are four types of bases: Adenine (A), Thymine (T), Guanine (G) and Cytosine(C). Bases belonging to different DNA strands tend toform pair-wise bindings: A with T, and G with C. Bases thatcan form a pair are said to be complementary. DNA (deoxyribonucleic acid) is the chemical material in a cell that carries the genetic codes for living organisms. Its structure is a double helix consisting of two sequences of letters from a four-letter alphabet (A, T, C, G), such that A is paired with T, and C with G. The letters represent the nucleotides or bases known as adenine, thymine, cytosine and guanine. Since the bases are paired, they are referred to asbase pairs.

In this paper we review different algorithms, the Needleman-Wunsch algorithm, the first algorithm applying the dynamic programming to comparing biological sequences Later the Smith-Waterman algorithm, based on dynamic programming, is one of the most fundamental algorithms in bioinformatics Considering the parallelization of this implementation, since parallelization of an iterative implementation of the algorithm would not be feasible. There has been significant recent work on the parallelization of dynamic programming algorithms in computational biology including implementations suitable for computational grids. The rest of this paper is organized as follows. In section 2 and 2.1, we provide Smith Waterman algorithm with its weakness. Section 3 and 3.1 provides Needleman-Wunsch algorithm parallel processing techniques with its weakness. The parallel computation techniques for DNA sequence comparison.

## 2. LITERATURE REVIEW

There are several methods for alignment of two biological sequences. The dynamic programming is probably the most popular programming method in sequences alignment. The Needleman-Wunsch algorithm, the first algorithm applying the dynamic programming to comparing biological sequences, was proposed by Needleman and Wunsch. Later, Smith and Waterman improved the Needleman-Wunsch algorithm and proposed the well-known Smith-Waterman algorithm. The time complexity of these algorithm is O(mn), where m, n are the lengths of the two sequences respectively. Because the cores of these algorithms are dynamic programming, all algorithms need to manipulate an (n+1) (m+1) matrix, named dynamic programming matrix. The most time spent in these algorithms is calculating the dynamic programming matrix, so research work on parallelization of two sequences alignment focuses mostly on the calculation of the matrix. However, in order to obtain the optimal result, these algorithms need to store the entire dynamic programming matrix in each parallel processor. As the growth of biological sequence database, the length of sequences often becomes very long, and the size of the matrix becomes verylarge. Thus, not only the execution time of these algorithms needs to be very long, the memory space needed in the algorithm becomes very large. Even in some cases the size of the matrix is bigger than the size of memory space in one processor.

## 3. SMITH-WATERMAN ALGORITHM

A modification of the dynamic programming algorithm for sequence alignment provides a local sequence alignment giving the highest scoring local match between two sequences. The rules for calculating the scoring matrix values are:

i) The scoring system must include negative scores for mismatching,

ii) When a dynamic programming scoring matrix value becomes negative, that value is set to zero, which has the effect of terminating any alignment up to that point.

The alignments are produced by starting at the highest scoring position in the scoring matrix and following a trace-back path from that position up to the position that scores zero.

For two sequences $a=a1\ a2 - - - an$ and
$b=bl\ b2- - - bn$
Where $Hij=H(al\ a2 - - ai,\ bl\ b2 - - bj)$, then,
$Hij=\max \{Hi-1\ j-1+S(aibj),$
$\max \{Hi-xj-Wx),$
$\max (Hij-y-Wy\}, 0\},$
$x>=1 \qquad y>=1$

Where, Hij is the score at the position i in sequence a and position j in sequence b. S(aibj) is the score for aligning the characters at positions i and j, wx is the penalty for a gap of length x in sequence a and wy is the penalty for a gap of length y in sequence b.

When looking for similarities between subsequences of two sequences, as is usually the goal in the methods used to find homologies by database searches, a local alignment method is more appropriate than a global. The simple dynamic programming algorithm described by Smith and Waterman is the basis for this type of alignments. The Smith-Waterman algorithm is perhaps the most widely used local similarity algorithm for biological sequence database searching. In Smith-Waterman database searches, the dynamic programming method is used to compare every database sequence to the query sequence and assign a score to each result. The dynamic programming method checks every possible alignment between two given sequences. This algorithm can be used both to compute the optimal alignment score and for creating the actual alignment. It uses memory space proportional to the product of the lengths of the two sequences, mn, and computing time proportional to mn (m + n). The recursion relations used in the original Smith-Waterman algorithm are the following:

$Hi, j = \max \{Hi−1,j−1, S[ai, bj],Ei,j , Fi,j\}$
Where
$Ei,j = \max 0<k<i\{Hi−k,j − g(k)\}$
$Fi,j = \max 0<l<j\{Hi,j−l − g(l)\}$

Here, Hi,j is the score of the optimal alignment ending at position (i, j) in the matrix, while Ei,j and Fi,j are the scores of optimal alignments that ends at the same position but with a gap in sequence A or B, respectively. S is the match/mismatch value of ai and bj, or amino acid substitution score matrix, while g(k) is the gap penalty function. The computations should be started with Ei,j = Fi,j = Hi,j = 0 for all i = 0 or j = 0, and proceeded with i going from 1 to m and j going from 1 to n.The order of computation is strict, because the value of H in any cell in

the alignment matrix cannot be computed before all cells to the left or above it has been computed. The overall optimal alignment score is equal to the maximum value of Hi,j.

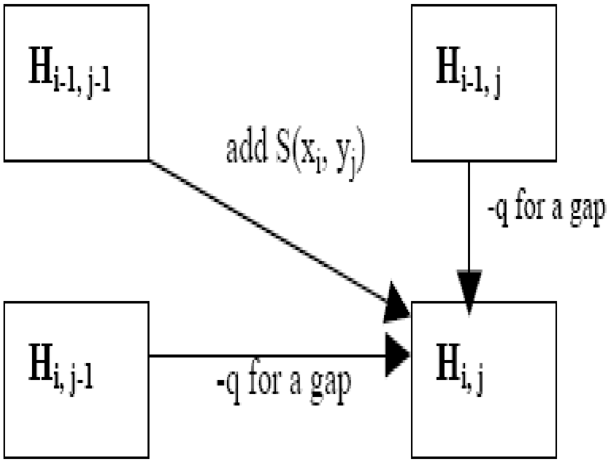


Fig: 1 Dynamic programming illustration.

### 3.1 Weakness of Smith-Waterman Algorithm

Smith-Waterman algorithm requires much larger number of computational steps, since we have to form the matrix and trace-back accordingly. This algorithm also suffers from much larger space complexity due to storage of matrix.

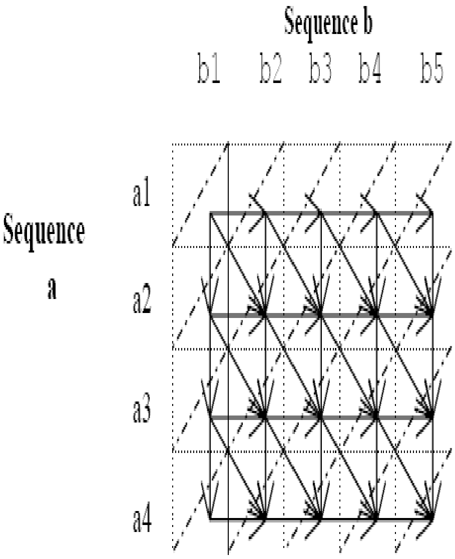
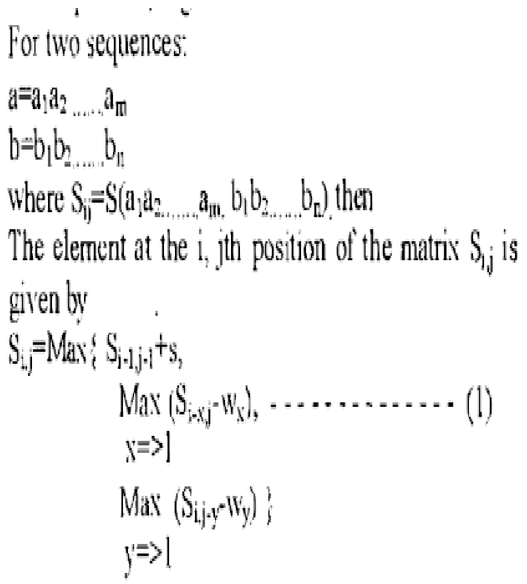


Fig : 2

## 4. NEEDLEMAN-WUNSCH ALGORITHM

Needleman-Wunsch used dynamic programming in order to obtain global alignment between two sequences. Global alignment, as the name suggests takes into account all the elements of the two sequences while aligning the two sequences. We can also call it as an "end to end “alignment. In Needleman-Wunsch algorithm, a scoring matrix of size m*n (m being the length of longer sequence and n being that of the shorter sequence) is first formed. The optimal score at each matrix position is calculated by adding the current match score to previously scored positions and subtracting gap penalties. Each matrix position may have a positive, negative or 0 value.



For two sequences:

$a=a_1a_2 ......a_m$

$b=b_1b_2......b_n$

where $S_{ij}=S(a_1a_2......a_m, b_1b_2......b_n)$ then

The element at the i, jth position of the matrix $S_{i,j}$ is given by

$$S_{i,j}=\mathrm{Max}\{\, S_{i-1,j-1}+s,\quad \underset{x=>1}{\mathrm{Max}}\,(S_{i-x,j}-w_x),\quad \underset{y=>1}{\mathrm{Max}}\,(S_{i,j-y}-w_y)\,\} \qquad \text{-------------- (1)}$$

where Sij is the score at position i in the sequence a and j in the sequence b, S(aibj) is the score for aligning the characters at positions i and j, wx is the penalty for a gap of length x in the sequence a and wy is the penalty for a gap of length y in the sequence b. After the S matrix is filled up, to determine all optimal alignment of the sequences from scoring matrix, a method called trace back is used. The trace back keeps track of the position in the scoring matrix that contributed to the highest overall score found. The positions may align or may be next to a gap, depending on the information in the trace back matrix. There may exist multiple maximal alignments.

### 4.1 Weakness of Needleman-Wunsch Algorithm

A study of the Needleman-Wunsch algorithm reveals two shortcomings of the algorithm:-

For increasing length of the sequences in comparison, the computational complexity becomes quite large and the size of the scoring matrix also might be huge. The implementation of the above algorithm on a computer using static allocation might be difficult. It is not guaranteed that the results given by the Needleman-Wunsch algorithm implementation will give global alignment. This means if there is a requirement of an "end to end" matching, the results might not comply with the requirement. Hence a new approach may be taken up in aligning the two sequences by direct comparison method taking into consideration that computational steps have to be minimized and "end to end' matching is fulfilled. The classical method to obtain global alignment is theNeedleman-Wunsch method. However this method suffers from the drawback that it involves a largenumber of computational steps and has to staticallyallocate a large section of memory for computerimplementation.

## 5. CONCLUSION

The aim of this study is to show the different algorithms based on dynamic programming, is one of the most fundamental algorithms in bioinformatics. This study shows the survey of different algorithms based on dynamic programming with their drawbacks and weakness.

## REFERENCES

[1] Sudha Gunturu*, Xiaolin Li*, and Laurence Tianruo Yang** *"*Load Scheduling Strategies for Parallel DNA Sequencing Applications" 11th IEEE International Conference on High Performance Computing and Communications 2009.

[2] Nasreddine Hireche, J.M. Pierre Langlois and Gabriela Nicolescu Département de Génie Informatique, École Polytechnique de Montréal ''Survey of Biological High Performance Computing: Algorithms, Implementations and Outlook Research'' IEEE CCECE/CCGEI, Ottawa, May 2006.

[3] Friman S´anchez, Esther Salam´ı, Alex Ramirez and Mateo Valero HiPEAC European Network of Excellence Universitat Polit`ecnica de Catalunya (UPC), Barcelona, Spain "Parallel Processing in Biological Sequence Comparison Using General Purpose Processors" 2005 IEEE.

[4] Matteo Canella - Filippo Miglioli Universit`a di Ferrara (Italy) Alessandro Bogliolo Universit`a di Urbino (Italy) Enrico Petraglio - Eduardo Sanchez Ecole Polytechnique F´ed´erale de Lausanne EPFL-LSL,Lausanne (Switzerland)" Performing DNA Comparison on a Bio- Inspired Tissue of FPGAs" Proceedings of the International Parallel and Distributed Processing Symposium (IPDPS'03) 2003 IEEE.

[5] N. F. Almeida Jr ,C. E. R. Alves, E. N. Caceres, S. W.Song "Comparison of Genomes using High- Performance Parallel Computing" Proceedings of the 15th Symposium on Computer Architecture and High Performance Computing (SBAC-PAD'03) 2003 IEEE.

[6] Fa Zhang, Xiang-Zhen Qiao and Zhi-Yong Liu Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, *National Natural Science Foundation of China, Beijing, 100083"*A Parallel Smith- Waterman Algorithm Based on Divide and Conquer" Proceedings of the Fifth International Conference on Algorithms and Architectures for Parallel Processing (ICA3PP.02) 2002 IEEE.