Comparative study of Genetic Algorithm and Dynamic Programming of DNA Multiple Sequence Alignment

Nabeel H. Kaghed
Ministry of higher education and scientific research
Head of supervision and scientific evaluation apparatus
nhkaghed@itnet.uobabylon.edu.iq
Eman S. Al shamery
Department of software, college of Information Technology, University of Babylon,
emanalshamery@itnet.uobabylon.edu.iq
Fanar Emad Khazaal Al-Khuzaie
Department of software, college of Information Technology, University of Babylon,

fanar@itnet.uobabylon.edu.iq

Abstract

Multiple Sequences Alignment(MSA) is the one of the most important Research themes in bioinformatics.
In this research the goal is to identify the best between the two methodologies(dynamic programming and Genetic Algorithm) . The execution time of dynamic programming (DP)algorithm is Growing specially when the number of join operations in a query is huge , DP suffers from the large storage and computational complexity, especially when the number of sequences is three or more .
This research presents a comparison between the implementation of dynamic programming and execution of Genetic Algorithm (GA) implementation . The database has been used in the form of Deoxyribonucleic acid (DNA) sequences , and protein sequences . The results have shown that the use of genetic algorithm is better than the dynamic programming solution.

Keywords :- Dynamic Programming, Genetic Algorithm, multiple sequence alignment.

الخلاصة

محاذاة السلاسل المتعددة هي واحدة من موضوعات البحوث الأكثر أهمية في المعلوماتية الحيوية. يهدف البحث الى تحديد الأفضل من بين المنهجيتين (البرمجة الديناميكية والخوارزمية الجينية). ان زمن تنفيذ خوارزمية البرمجة الديناميكية يتزايد خاصة عندما يكون عدد من العمليات المرتبطة في الاستعلام ضخمة، تعاني البرمجة الديناميكية من التخزين الكبير والتعقيد الحسابي، وخاصة عندما يكون عدد السلاسل ثلاثة أو أكثر. يقدم هذا البحث مقارنة بين تنفيذ البرمجة الديناميكية وتنفيذ الخوارزمية الجينية . وقد تم استخدام قاعدة البيانات في شكل سلاسل الحامض النووي، وسلاسل البروتين.وقد أظهرت النتائج أن استخدام الخوارزمية الجينية هو أفضل من حل البرمجة الديناميكية.

الكلمات المفتاحية :- البرمجة الديناميكية, الخوارزمية الجينية , محاذاة السلاسل المتعددة .

1.Introduction

Similarities shared by all creatures in the basic unit of life which is the cell ,chemical energy is stored in ATP, genetic information is encoded by DNA ,and the information is transcribed into Ribonucleic acid (RNA). [Deonier, *et al.*, 2005]

Alignment algorithms (heuristic and dynamic) used two different kinds of sequence alignment , Local and Global.

Local is explore best portion matching , while global is explore best match of sequences in whole .[Sonali Vijan and Rajesh Mehra , 2011]

There are many reasons for align which are to infer homology, and to study the evolutionary relationships between the sequences.

The Pairwise sequence alignment is used to find the best match between two sequences , whether (local or global).

There are basic methods which produce Pairwise sequence alignment , dot matrix method , dynamic programming method, and Word method ,where each of the methods has strengths and weaknesses.

Multiple sequence alignment is extension of pairwise but it is to align all sequences in query set.[Jesper Mojbeak,2010]

Multiple sequence comparison indicates the search for symmetry in three or more sequences. [Segun *et al.*, 2009]

Alignment of Multiple sequence is a hard computational case. [Deonier, *et al.*, 2005]

After this introduction , which has clarified the methods used in the comparison , note that it can be used not only for DNA but for protein and RNA as well.

The remaining of the research includes the results and conclusion.

2. Motivation

If all the DNA in human body was put end to end, it would reach to the sun and get back over 600 times. In human body there are approximately 3 billion bases in the DNA code.

[Shishir Kumar Gangwar and Birhanu Worabo, 2011].Consequently, there is a lot of difficulty in applying Multiple sequence alignment which is motivated to use the automatic methods for solution, where that despite having a lot of algorithms to process aligning it is still open problem to look for the best in terms of storage and time which are the two basic criteria for a comparison among the methods.

[Sonali Vijan and Rajesh Mehra , 2011] . Bioinformatics algorithms are used in solving a lot of computer problems such as in security [Scott *et al.*, , 2008] and networks [Santosh *et al.*, 2010] .

3.Related Work

The exact algorithm is used when talking about an algorithm that always finds the optimal solution to an optimization problem, but the best known exact algorithms require exponential time**.** Iterative algorithms are based on the idea that the solution to a given problem can be computed by modifying the already existing sub-optimal solution. [Meghna and Geetika,2013] advantages of DP such as (Needleman-Wunsch ,Smith-Waterman) is ability of finding the optimal alignment solution among the sequences, when disadvantage is taking more time to make the alignment which decreases the method performancing. [Arabi *et al.,* 2012] many iterative stochastic approaches are offered to alleviate these troubles. For example, evolutionary computation techniques especially (GAs) have been successfully applied to the MSA problem .when GA has been used to solve complex problems like MSA It can search for large solution spaces more efficiently.[Yang *et al.,*, 2008]

4. Theoretical Background Of Methodologies :-

4.1 Dynamic Programming :-

Dynamic Programming method consumes long time of execution but gives highly accurate alignment.

In mathematics, computer science and economics, dynamic programing is a method for solving complex problems by dividing them down into simpler sub problems. The idea behind dynamic programming is to solve these sub problems then combine the solutions of the sub problems to reach the overall solution. [Jesper Mojbeak , 2010]

The key point of dynamic programming is to find all possibilities , because of the lengths of the sequences and the size of storage, where it is difficult to apply in the dynamic programming in pair sequence alignment and very difficult in multiple sequence alignment according to the criteria that have been mentioned.[Deonier *et al.*, 2005]. in the next section the algorithm of Dynamic Programming fig.1 .

Name :MSA by Dynamic Programming Algorithm .

Input : set of sequences .

Output : alignment among the sequences .

1- set of sequences K as k1,k2,…..kn.

2- make pair wise alignment for each 2 sequences A,B by using a score matrix M[i,j]

Where i= (length of A) +1 , j= (length of B)+1

we have match ,mismatch ,and gap values
-M[0,0]=0
- fill the rest cells of first row by gap redoubled values.

-fill the rest cells of the first column row by gap redoubled values.

-fill the rest of cells by maximum value of :-

$$M[i,j]= \begin{cases} M[i,j-1]+gap \\ M[i-1,j-1]+p(i,j) \\ M[i-1,j]+gap \end{cases}$$

Where p(i , j) is the function if S[i]=S[j] then return +1 or if S[i] != S[j] then return -1.

-then trace back from the last higher cell of matrix to the first cell by choosing the highest total score path.

3- repeat :- apply sum of pair score function for each column :-

$$s(a_1 \ldots a_k) = \sum_{i,j} s^*(a_i , a_j)$$

until to get higher score.

4- final alignment with maximum score .

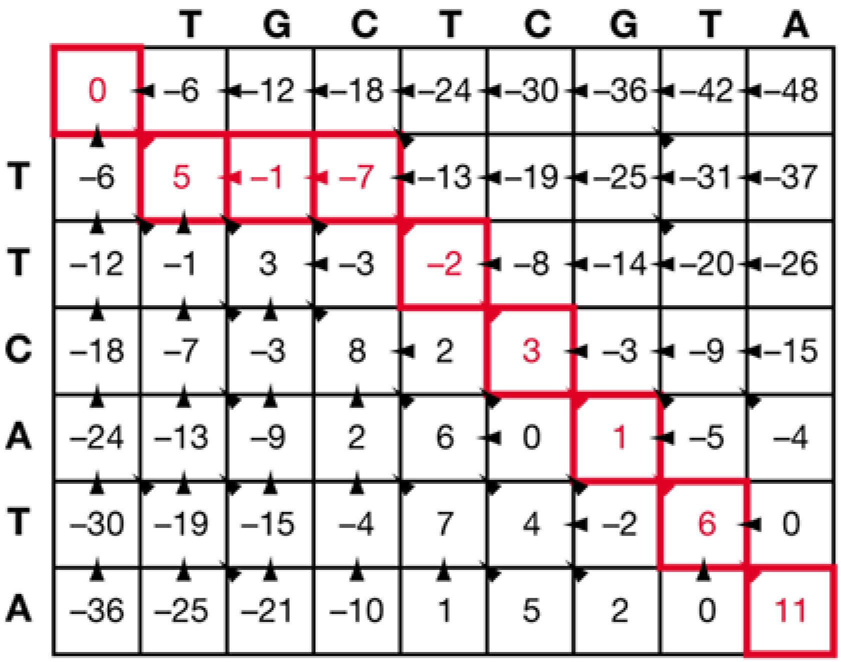
Fig.1 the algorithm of Dynamic Programming .

Case study :-

through the figure 2 , it is clear how to initialize and fill the score matrix for two sequences , also it seems clear how the trace back operation is done(with read bold square) to get path and includes the best result.

The final step is aligning by [set symbols of the first sequence with the second sequence symbols and with Gaps, according to the following (trace the direction of the arrow, if it is diagonal set the symbol in the first sequence and the corresponding second sequence or if the direction is vertical or horizontal set Gap instead of the symbol)[Eric,1997]

using in this case study the following values:-
Gap= -6
match =5
mismatch=-2
and to be sure to calculate the result as follows:
T _ _ T C A T A
T G C T C G T A
5 -6 -6 5 5 -2 5 5 = 11



		T	G	C	T	C	G	T	A
	0	–6	–12	–18	–24	–30	–36	–42	–48
T	–6	5	–1	–7	–13	–19	–25	–31	–37
T	–12	–1	3	–3	–2	–8	–14	–20	–26
C	–18	–7	–3	8	2	3	–3	–9	–15
A	–24	–13	–9	2	6	0	1	–5	–4
T	–30	–19	–15	–4	7	4	–2	6	0
A	–36	–25	–21	–10	1	5	2	0	11

Fig.2 score matrix

Of multi sequences can be calculated the sum of pairs (SP) as follows[Yang *et al.,* 2008]
A C G _ T
_ C G T T
A C _ T T
column1= -7,column2= 15, column3= -7,column4= -7, column5= 15.
SP-score = (-7)+ (15) +(-7) +(-7) +(15)= 9.
- The major problem with the SP method is that finding the optimal MSA which is time consuming.

- given k sequences of length n , time complexity is :- $O(k^2 2^k n^k)$

Also it is not always possible to combine optimal pairwise alignments into a multiple alignment since some pair-wise alignments may be an incompatible. [Mohammed *et al.*, 2013]

4.2 Genetic Algorithm

Simple GA (SGA) started with the generation of population consists of chromosomes , a fixed size encoded solution. Each chromosome represents a possible solution and the space of all feasible solutions is called search space.

The role of GA is to alter the generated chromosomes using various operators to get the optimal chromosome with best fitness value in the search space. The goal is to maximize the similarity among sequences in the minimal number of gaps. Iteration continues till the termination condition is satisfied. [Wen-Yang *et al.*, 2003].

The algorithm terminates on reaching specified number of generations or at convergence.

An optimal solution may not be reached if termination is due to the maximum number of generations.

GA is not a requirement to store each generation , only the last one which contains the best solution. [Yang *et al.*, 2008].

In this research the fitness value has been used by depending on calculation of the base dominant of each column (dom (x_i)) and the number of gaps (Gap_i) of all columns.

In the next section the algorithm of Genetic Algorithm figure 3.

Name : MSA by Simple Genetic Algorithm

Input :population with N size, number of generation G, crossover probability Pc , mutation rate Pm

Output : alignment among the sequences.

1. [Start] Generate random population of *n* chromosomes (suitable solutions for the problem)

   Select the longest sequence ,and complete the other sequences by gaps until equal with the longest sequence.

2. [Fitness] Evaluate the fitness *f(x)* of each chromosome *x* in the population by :-

   $$\left(\sum_{i=1}^{n} (\text{dom}(x_i)) - \sum_{i=1}^{n} \text{Gap}_i\right)$$

   Where :-

   n= number of columns

   dom= refers to dominant base (x) of each column(i)

   Gap = refers to Gaps for each column(i)

   if column has only Gaps the value of it was zero.

3. [New population] Create a new population by repeating the following steps until the new population is complete

   a- [Selection] Select two parent chromosomes from a population according to their fitness (the better fitness, the bigger chance to be selected)

   b- [Crossover] With a crossover probability cross over the parents from the crossover pool, and select crossover point .

   c-[Mutation] mutation operations are applied to each pair , With a mutation probability mutate , here we have change one gen by another randomly .

   d- [Accepting] set new offspring in the new population

4. [Replace] take new generated population for a further run of the algorithm

5. [Test] If the end condition(either access to a specified number of generations OR convergence between generations) is satisfied, stop, and return the best solution in current population.

6. [Loop] Go to step 2.

Fig.3 the algorithm of Genetic Algorithm

Case study

Table 1 shows example For Population Initialization. Fig.4 shows one-point crossover operation between two parents (two chromosomes) ,and the result is two children .

Table1 example For Population Initialization.

sequence	Sequence length	No. of Gap	Gap position	alignment
TCTAGATG	8	4	5|3|6|9|	TC-T--AG-ATG
CTATGATGTA	10	2	12|10|	CTATGATGT-A-
GTTCTAT	7	5	8|4|6|1|12|	-GT-T-C-TAT-
ACGATGTA	8	4	7|4|11|5|	ACG--A-TGT-A
ACGTAT	6	6	7|4|11|5|8|12|	ACG--T--AT--

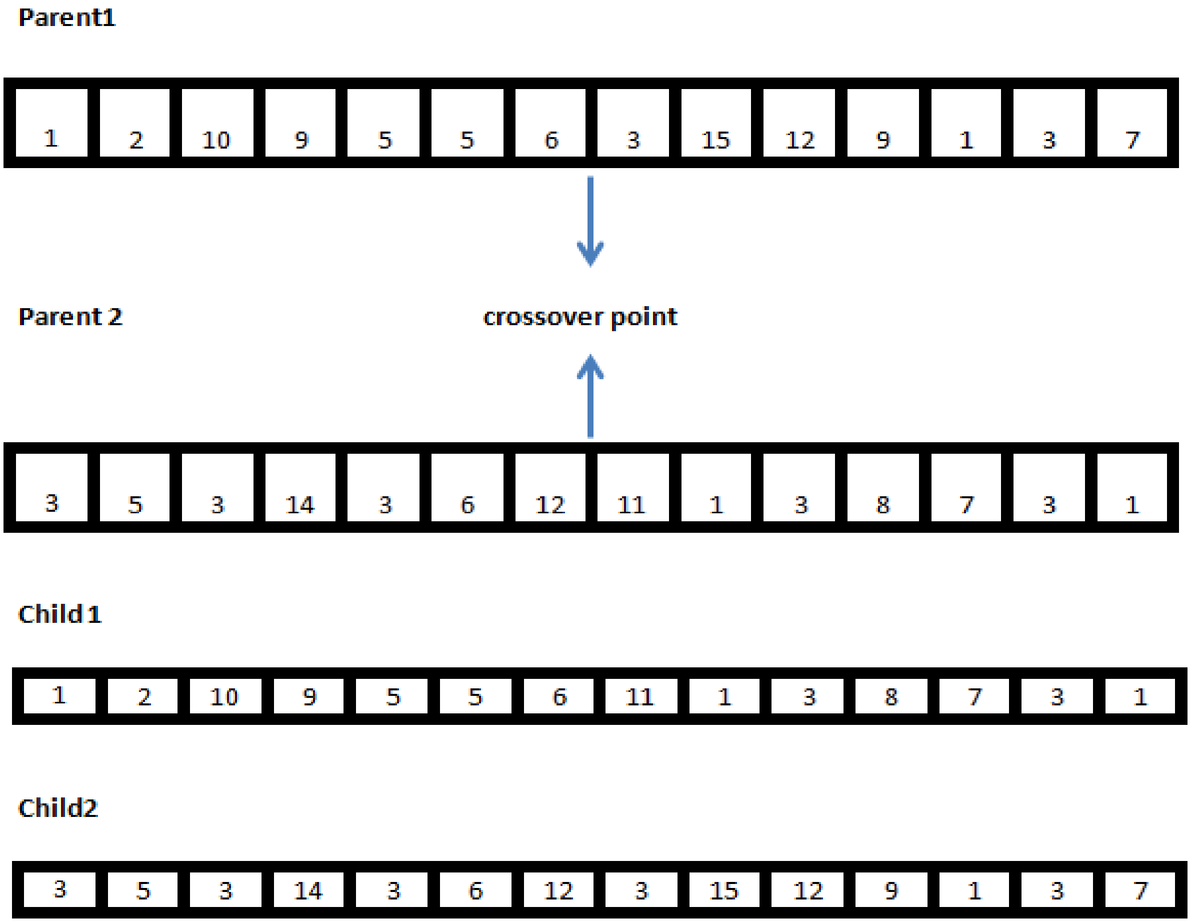


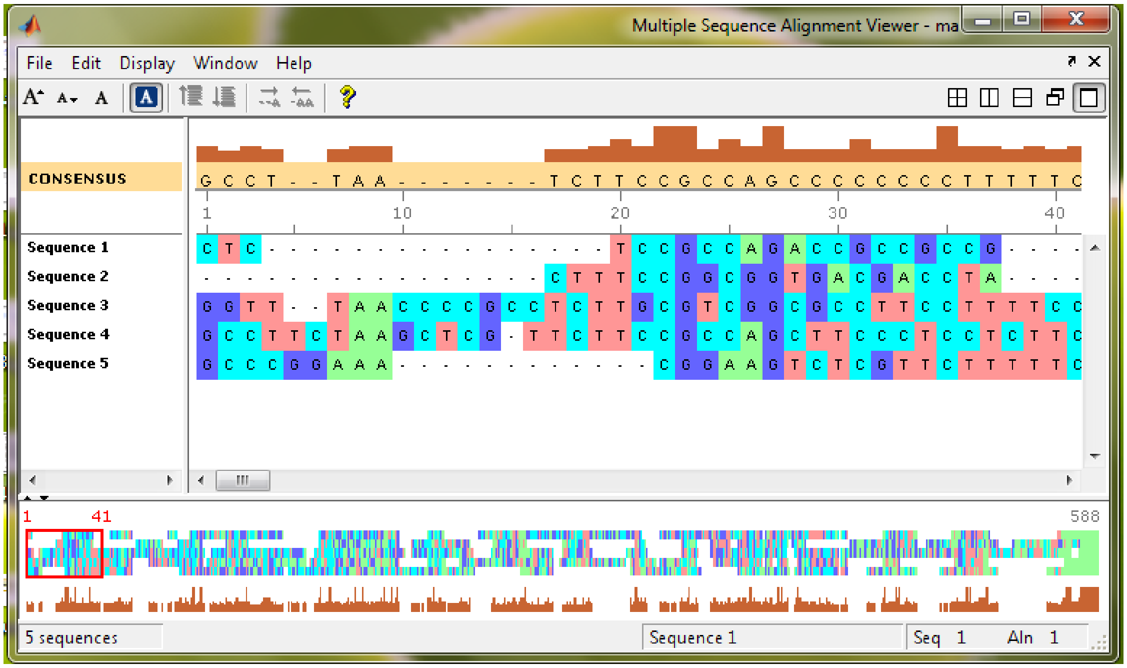
Fig.4 crossover technique – one point crossover.

5. Experimental Results

In this research MATLAB language has been used which has read sequences as a FASTA format files , and has extracted the required information from the files for the purpose of aligning.

The genetic algorithm as well as the dynamic programming algorithm have been programmed for confirmation any of them is the most efficient in terms of time.

Note that it is already known that the dynamic programming gives a perfect solution, while genetic algorithm gives an approximate solution and in less time, see tables 2,3.



```
CONSENSUS   G C C T - - T A A - - - - - - - T C T T C C G C C A G C C C C C C C C T T T T T C
Sequence 1  C T C - - - - - - - - - - - - - - - - T C C G C C A G A C C G C C G C C G - - - -
Sequence 2  - - - - - - - - - - - - - - - - C T T T C C G G C G G T G A C G A C C T A - - - -
Sequence 3  G G T T - - T A A C C C C G C C T C T T G C G T C G G C G C C T T C C T T T T C C
Sequence 4  G C C T T C T A A G C T C G - T T C T T C C G C C A G C T T C C C T C C T C T T C
Sequence 5  G C C C G G A A A - - - - - - - - - - - - C G G A A G T C T C G T T C T T T T T C
```

Fig. 5 Alignment of DNA sequences .

In genetic algorithm , the method used for selection of parents is binary groups.

the methods for mating have been applied (one point crossover, two point crossover), the mating done according to the probability of mating (0.8)

Then follow the mating procedure by mutation action on the children and also according to the probability of a mutation (0.5)

Then stop condition has been made sure which was representing either access to a specified number of generations or convergence between generations.

Various samples of DNA sequences have been taken for the purpose of global alignment , lengths were as follows :- 361, 411,511,401, 439

the longest sequence was 511 while 361 was the shortest , and all belong to the human, were taken from the genbank (www.ncbi.nlm.nih.gov/genbank) , and also multiple samples of protein have been taken for global alignment, the lengths of sequences were as follows :- 163, 239, 264, 381 ,239 . In this research , results of comparing between execution of three sequences and five sequences have been registered in tables below to clear the differences in (cpu time and memory size), as shown in tables 4 and 5.

Table 2 . shows the difference between DP and GA in cpu time.

Time of DP for 3 seq. of DNA	Time of GA with 1X for 3 seq. of DNA	Time of GA with 2X for 3 seq. of DNA	Time of GA with 1X for 3 seq. of protein	Time of GA with 2X of 3seq. of protein
80 seconds	4.406 seconds	4.584 seconds	3.838 seconds	3.493 seconds

Table 3 . shows the difference between DP and GA in memory size.

Memory size of 3 seq. using DP	Memory size of 3 seq. of DNA using GA with 1X	Memory size of 3 seq. of DNA using GA with 2X	Memory size of 3 seq. of protein using GA with 1X	Memory size of 3 seq. of protein using GA with 2X
150.305 KB	128.624 KB	128.800 KB	127.900 KB	127.800 KB

Table 4 . cpu time for difference length of sequences by using GA .

Time of GA with 1X for 5 seq. of DNA	Time of GA with 2X for 5 seq. of DNA	Time of GA with 1X for 5 seq. of protein	Time of GA with 2X of 5 seq. of protein
7.721 seconds	7.934 seconds	6.293 seconds	6.546 seconds

Table 5 . memory size for difference length of sequences by using GA .

Memory size of 5 seq. of DNA using GA with 1X	Memory size of 5 seq. of DNA using GA with 2X	Memory size of 5 seq. of protein using GA with 1X	Memory size of 5 seq. of protein using GA with 2X
128.900 KB	129.012 KB	128.836 KB	128.972 KB

Many biological sequences are selected of (DNA and protein) to get the best results in execution , dynamic algorithm is applied for only three sequences with short lengths because of that (when increasing lengths of the used sequences OR / AND increasing the number of the used sequences) leads to increasing execution time and storage space.

There are some limitations with GA ; there has been tussle between speed and accuracy, sometimes GAs result is unsatisfactory compromise, which is either low quality of solution or high convergence speed. Based on the two sequence alignment algorithms the table 6 gives the summarized observation of two algorithms.

Table 6. the summarized observation of Dynamic Programming and Genetic Algorithm.

Algorithm	Advantages	Disadvantage
Dynamic Programming	It is guaranteed in mathematical sense to provide an optimal alignment for a given set of scoring function.	This approach however results in exponential time complexity, since it requires time proportional to the product of the sequence lengths. It becomes slow as there are large computation steps. The memory requirement also increase as alignment sequences get large.
Genetic Algorithm	GA improves the accuracy of MSA . These can be repeated a number of times or until convergences. it can be implemented to produces approximate solutions to the MSA problem. Using only a small amount of computer resources.	conflict between speed and accuracy , in sometimes GA results are unsatisfactory because of (either low quality of solution or high speed of convergence).

6.Conclusions

Accordingly to the case studies, this paper has proved that GA is better than DP in time of execution . By GA the performance increased , memory location was decreased, and the implementation reduced the time . This research has focused on transactions which directly affect the performance of genetic algorithms such as the selection , fitness function , crossover , and replacement.

Development of these transactions increases the efficiency of genetic algorithms and makes it very impressive .This can be considered as a future work.

7.References

Meghna Mathur and Geetika. (2013)." Multiple Sequence Alignment Using MATLAB ". Department of CSE/IT ITM University , Gurgaon, India.

Mohammed M. Saleh, Ahmed M. Alzohairy, Osama Abdo Mohamed, Gaber H. Alsayed .(2013)." A Comprehensive Study by Using Different Alignment Algorithms to Demonstrate the Genetic Evolution of Heat Shock Factor 1 (HSF1) in Differen Eukaryotic Organisms ", Egypt.

Arabi E.keshk , Lamiaa Fathi Hussein,& Mohammed Ossman, .(2012)." Fast Longest Common Subsequences for Bioinformatics Dynamic Programming " , Menofia University.

Sonali Vijan and Rajesh Mehra.(2011) ."Biological Sequence Alignment for Bioinformatics Applications Using MATLAB".

Shishir Kumar Gangwar and Birhanu Worabo. (2011)." AMAZING FACTS ABOUT HUMAN DNA AND GENOME", science and nature journal.

Jesper Mojbeak .(2010) . " Exact Multiple Sequence Alignment using Forward Dynamic Programmin a thesis in Bioinformatics ". Bioinformatics Research Center , Aarhus University.

Santosh Kumar Singh , Krishna Chandra Roy , and Vibhar Pathak .(2010)." CHANNELS REALLOCATION IN COGNITIVE RADIO NETWORKS BASED ON DNA SEQUENCE ALIGNMENT", Suresh Gyan Vihar University, Jaipur. India .

Segun A. Fatumo, Ibidapo O. Akinyemi and Ezekiel .F. Adebiyi. (2009)." Aligning Multiple Sequences with Genetic Algorithm".

Scott E. Coul , Joel W. Branch , Boleslaw K. Szymanski , and Eric A. Breimer .(2008).
" Sequence Alignment for Masquerade Detection". United States.

Yang Chen, Jinglu Hu,& Kotaro Hirasawa.(2008)." Multiple Sequence Alignment Based on Genetic Algorithms with Reserve Selection " .

Nguyen Thu Hang.(2008) ."COMPARISON OF MULTIPLE SEQUENCE ALIGNMENT PROGRAMS IN PRACTISE a thesis in Bioinformatics". The Bioinformatics Research Center (BiRC) ,University of Århus.

Deonier, R.C.; S. Tavaré,& M.S. Waterman.(2005)." Computational genome analysis : an introduction". Department of Biological Sciences ,University of Southern California, Los Angeles.

Wen-Yang Lin, Wen-Yuan Lee, and Tzung – Pei Hong.(2003)." Adapting Crossover and Mutation Rates in Genetic Algorithms". Taiwan.

Eric C. Rouchka . (1997) http://www.avatar.se/molbioinfo2001/dynprog/adv_dynamic.html