An Improved Approach to Develop Global Alignment Algorithm and its Implementation by Making Software for Multiple Alignment, Phylogenetic Tree and GC Content

Md Rezaul Karim,[1,*] , Md.Rethwan Kabeer,[2] Abdur Rahman Bin Shahid,[1] Sanjib Biswas,[1] and S.M.Muslem Uddin[3]

[1]Department of Computer Science and Engineering, Chittagong University of Engineering and Technology
Chittagong-4349, Bangladesh

[2]Department of Computer Science and Information System, University of Malaya
Kuala Lumpur, Malaysia

[3]Department of Electrical Engineering, University of Malaya
Kuala Lumpur, Malaysia

[*]reza_cse06@yahoo.com

Abstract—**Bioinformatics is the field of science in which biology, computer science, and information technology merge to form a single discipline. Bioinformatics deals with algorithms, databases and information systems, web technologies, artificial intelligence and soft computing, information and computation theory, software engineering, data mining, image processing, modelling and simulation, signal processing, discrete mathematics, control and system theory, circuit theory, and statistics, for generating new knowledge of biology and medicine, and improving & discovering new models of computation (e.g. DNA computing, neural computing, evolutionary computing, immuno-computing, swarm-computing, cellular-computing). In bioinformatics, a sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. A large number of algorithms are used in bioinformatics such as global alignment, local alignment, multiple alignments etc. Global alignments, which attempt to align every residue in every sequence, are most useful when the sequences in the query set are similar and of roughly equal size. We worked with the global alignment algorithm with an improved and efficient manner. The improved algorithm is too much accurate for aligning the genome sequences with better computational time. And using this improved way we also developed the multiple alignment, phylogenetic tree and GC content of the sequences.**

Index Terms— **Global alignments, evolutionary relationships, phylogenetic tree.**

I. Introduction

Global alignment is a necessary part of bioinformatics. It is an alignment where two sequences are aligned over their entire lengths. In the local alignment it is not needed to align over their entire lengths but global alignment needed this. When two sequences are nearly similar then global alignment is very useful [3][7]. A multiple sequence alignment is an alignment of n > 2 sequences using pair wise alignment. Multiple sequence alignment also refers to the process of aligning such a sequence set. Because three or more sequences of biologically relevant length can be difficult and are almost always time-consuming to align by hand, computational algorithms are used to produce and analyze the alignments. Multiple alignments are used in phylogenetic tree [5].

A phylogenetic tree or evolutionary tree is a branching diagram or "tree" showing the inferred evolutionary relationships among various biological species or other entities based upon similarities and differences in their physical and/or genetic characteristics [2]. GC content ensures stable binding of primer/template. 40-60% GC content ensures stable binding of primer. G-C bonds contribute more to the stability (increased melting temperatures) of primer/template binding than do A-T bonds.

In this paper we developed a improved global alignment algorithm and using this algorithm we made a software for multiple alignment, phylogenetic tree and GC content.

II. Related research

Initially dot matrix was used for global alignment which is time-consuming. Then Needleman and Wunsch was introduced algorithm which takes quadratic time [1].

There is much software for doing sequence alignment such as using the “FEAST” nucleotide type sequence is used for local alignment. Using the “Path” protein type sequence is used for local alignment. “Ngila” software is used for global alignment using both (nucleotide and protein) type of sequences. “ClustalW”[4] and “T-Coffee” [6] software is used for multiple alignments for both sequences. For doing phylogenetic tree “PhyloDraw” and also “ATV” is used.

III. Proposed method

A. Overview

The proposed method aims at improving the global alignment algorithm and then has to make software for multiple alignment, phylogenetic tree and GC content. Details of these steps are described in the following sections.

B. Proposed Global Alignment Algorithm

1) *Select Parameter Value*: There are three parameters for global alignment which have to be selected.

μ = Mismatch penalty, σ = Indel penalty, Ω = Matched prize

978-1-4673-1436-7/12/$31.00 ©2012 IEEE

$$\begin{matrix}\rightarrow \\ \downarrow \ \searrow\end{matrix} = \begin{cases} -\sigma \\ \Omega \\ -\mu \end{cases} \qquad (1)$$

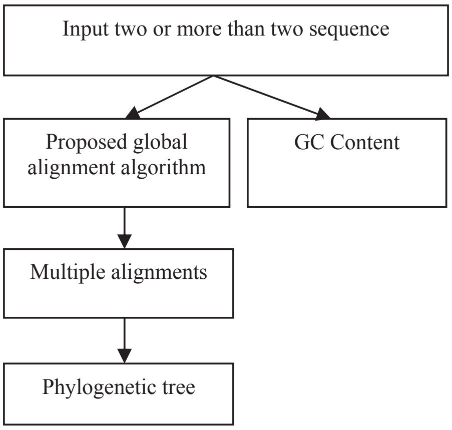

Input two or more than two sequence
Proposed global alignment algorithm
GC Content
Multiple alignments
Phylogenetic tree

Fig. 1 Overview of the proposed method

2) *Filling Array Index:* There is a two dimensional array like matrix which has the value using some calculation. In the calculation there has three options using matched, mismatched and indel penalty. The calculation of matrix array is like,

$$Matrix\ [i, j] = \max \begin{cases} Matrix\ [i-1, j-1] + \Omega, if\ a[i] = b[j] \\ Matrix\ [i-1, j-1] - \mu, if\ a[i] \neq b[j] \\ Matrix\ [i-1, j] - \sigma \\ Matrix\ [i, j-1] - \sigma \end{cases} \quad (2)$$

C. Multiple Alignments

Multiple alignments are the alignment for more than two sequences using the pair wise alignment. For doing the multiple alignments I have used my proposed global alignment algorithm. In the multiple alignment process we have used existing algorithm. In the existing algorithm pair wise alignment is used like local alignment global alignment. In this pair wise alignment section I have used my proposed global alignment algorithm [5].

D. Phylogenetic Tree

Phylogenetic tree is the genetic or evolutionary relationships between species. For doing the phylogenetic tree I have also used the existing algorithm. There are two type of phylogenetic tree. One is rooted tree and another is unrooted tree. From the multiple alignments a distance matrix has to be calculated and then from the distance matrix the most closely related sequence has a common parent and then the next closely related sequence has a common parent which is also common with the previous parent and this process is repeated until the last sequence [2].

E. GC Contents

GC content is usually expressed as a percentage value, but sometimes as a ratio (called G+C ratio or GC ratio). GC-content percentage is calculated as

$$\frac{G+C}{A+T+G+C} \times 100 \qquad (3)$$

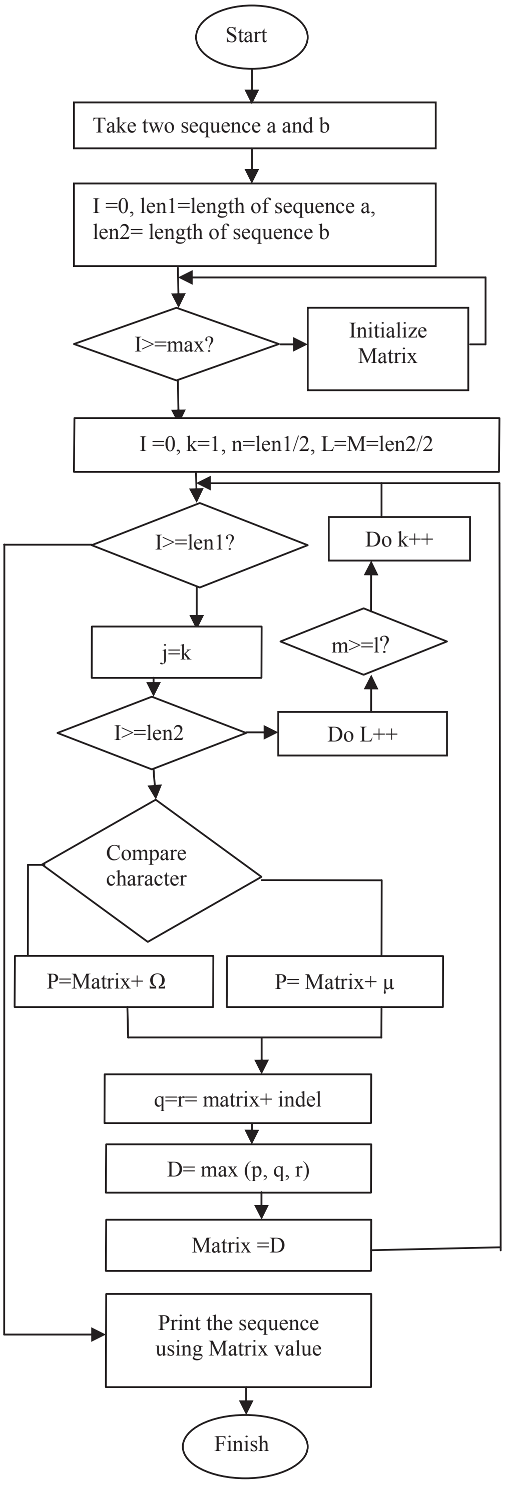
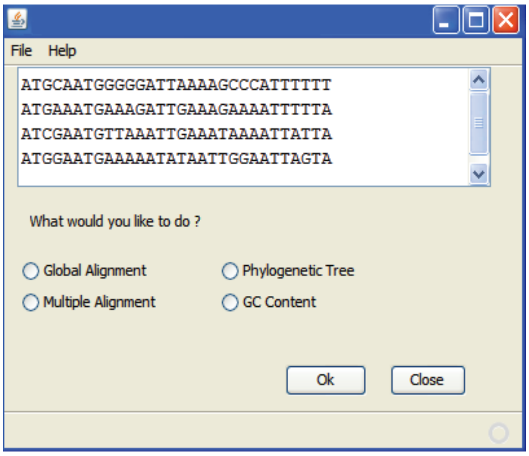


Fig. 2 Proposed global alignment algorithm

IV. Experimental Results

We experiments various sequence with different length. We create a input window in which we can put two or more input sequences (fig. 3). From the input if anyone wants to find global alignment, multiple alignment, phylogenetic tree or GC content of that input than he have to choose that type. If anyone wants to see global alignment output than the output is like as fig.4.



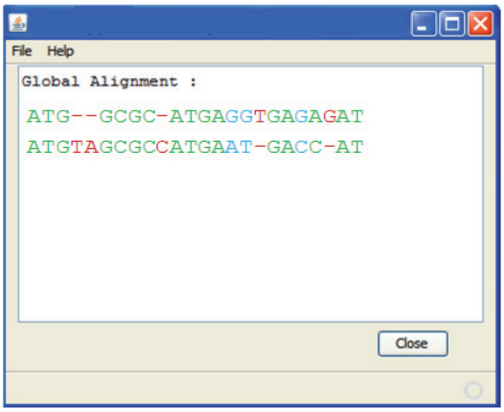
```
ATGCAATGGGGGATTAAAAGCCCATTTTTT
ATGAAATGAAAGATTGAAAGAAAATTTTTA
ATCGAATGTTAAATTGAAATAAAATTATTA
ATGGAATGAAAAATATAATTGGAATTAGTA
```

What would you like to do ?

○ Global Alignment ○ Phylogenetic Tree

○ Multiple Alignment ○ GC Content

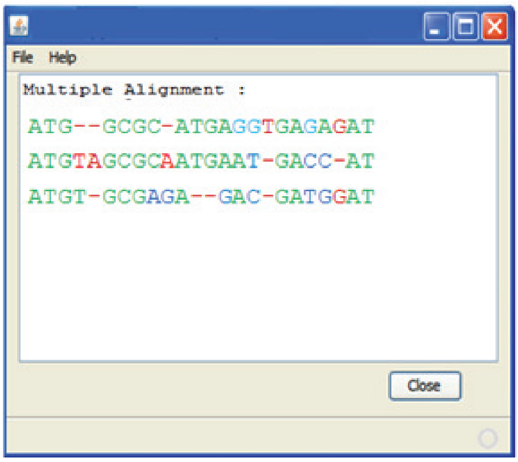
Fig.3 Input Window



```
Global Alignment :
ATG--GCGC-ATGAGGTGAGAGAT
ATGTAGCGCCATGAAT-GACC-AT
```

Fig.4 Output of global alignment

If anyone wants to see multiple alignments output than the output sequences is like as,



```
Multiple Alignment :
ATG--GCGC-ATGAGGTGAGAGAT
ATGTAGCGCAATGAAT-GACC-AT
ATGT-GCGAGA--GAC-GATGGAT
```

Fig. 5 Output of Multiple Alignment

If anyone wants to see the Phylogenetic Tree output than the output sequences is like as fig.6.

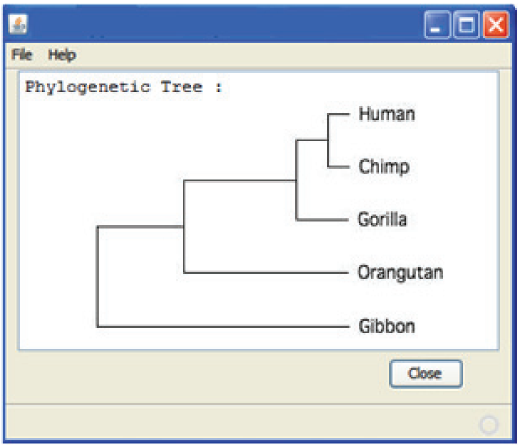
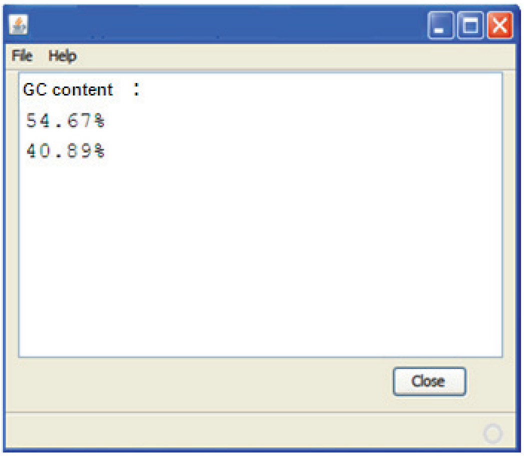


Fig. 6 Output of Phylogenetic Tree

And if anyone wants to see the GC content output than the output is like as fig.7.



```
GC content :
54.67%
40.89%
```

Fig. 7 Output of GC content .

We have also implemented the existing Needleman and Wunsch algorithm [1] and we have compaired that with this proposed algorithm.

TABLE I
TIME COMPLEXITY COMPARISON WITH EXISTING METHOD [1]

Sequence length	**Existing global alignment algorithm (sec)**	**Proposed global alignment algorithm (sec)**
1000	0.049000	0.045000
2000	0.180000	0.172000
3000	0.352000	0.334000
4000	0.765000	0.703000
5000	0.894000	0.817000

From the table 1 we are seeing that for different sequences length there are time variation between existing global alignment algorithm and proposed algorithm and the proposed algorithm requires less time than existing algorithm.

From the data the graph representations are given below in which we are seeing that for small sequences the time needed between algorithms is almost same. But when the sequence length is going to high then proposed global algorithm needs less time than existing algorithm.

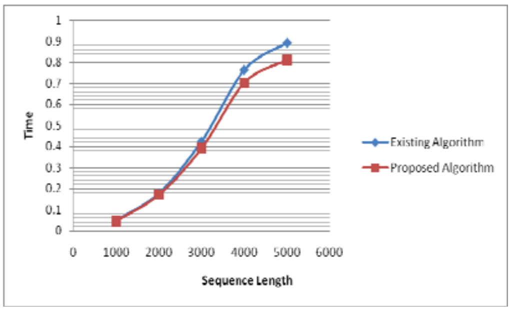


Fig. 8 Curved graph of the tme complexity comparison

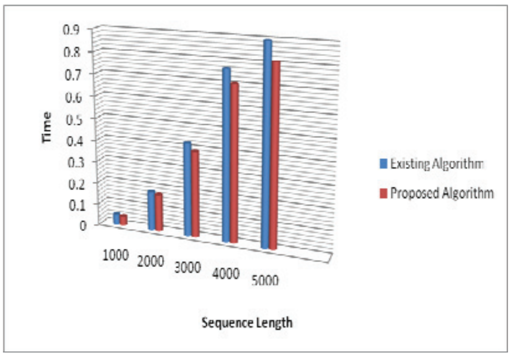


Fig.9 Bar graph of the time complexity comparison

V. Conclusions

The goal of this paper is to develop an improved way by which the global alignment algorithm works better and using these algorithm develop programs for multiple alignment, phylogenetic tree and GC content. Its main focus is on new developments in genome bioinformatics and computational biology. For the alignment we first download the genomic sequence from the online database like NCBI. And after getting the sequences we align two or more sequences with our alignment algorithm and compare with the existing one. We also develop the multiple alignments for a large number of sequences at a time and get better alignment. By the similarities between the genomes we get the phylogenetic tree and GC content. For the following purpose we use java language and implement it.

References

[1] A Class Note on Global Alignment by Kun-Mao Chao 1;2;3, National Taiwan University, Taipei, Taiwan, September 30, 2008.

[2] Constructing Phylogenetic Trees using Multiple Sequence Alignment by Ryan M. Potter, University of Washington, 2008.

[3] Sequence Alignment Algorithms by Sérgio Anibal de Carvalho Junior, university of London, 5th September 2003.

[4] CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting,position-specific gap penalties and weight matrix choice Julie D.Thompson, Desmond G.Higgins+ and Toby J.Gibson* European Molecular Biology Laboratory, Postfach 102209, Meyerhofstrasse 1, D-69012 Heidelberg, Germany, September 23, 1994.

[5] Strategies for Multiple Sequence Alignment, Bio Techniques 32:572-591 (March 2002), Hugh B. Nicholas Jr., Alexander J. Ropelewski, and David, W. Deerfield II Carnegie Mellon University, Pittsburgh, PA, USA.

[6] T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment CeÂdric Notredame1,2,3*, Desmond G. Higgins4 and Jaap Heringa1, 1National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK and 2Information Genetique et Structurale, CNRS-UMR 1889 31 Ch. Joseph Aiguier 13402 Marseille, France.

[7] An Introduction to Bioinformatics Algorithm by Neil C. Jones and Pavel A. Pevzner.

[8] Bioinformatics For Dummies, 2nd Edition by Jean-Michel Claverie & Cedric Notredame.

[9] Improved algorithms for DNA sequence alignment and revision of scoring matrix by Bandyopadhyay, S.S.; Paul, S.; Konar, A. IEEE conference, 2005.

[10] Approximate global alignment of sequences by Kahveci, T., Ramaswamy, V., Han Tao and Tao Li . BIBE 2005. Fifth IEEE Symposium on Publication Date: 2005.