IJCSN International Journal of Computer Science and Network, Volume 3, Issue 4, August 2014
ISSN (Online) : 2277-5420 www.IJCSN.org
Impact Factor: 0.274

An Evolutionary Optimization for Multiple Sequence Alignment

[1] K. Lohitha Lakshmi , [2] P. Rajesh

[1] M tech Scholar Department of Computer Science , VVIT
Nambur , Guntur,A.P.

[2] Assistant Prof Department of Computer Science, VVIT
Nambur, Guntur, A.P.

Abstract - Multiple Sequence Alignment is one of the most useful tools in bioinformatics. It is widely used to identify conservation of protein domains, RNA secondary structure and classification of biological sequences. However, it is recognized as one of the most challenging tasks in bioinformatics. Evolutionary algorithms are providing competitive solutions for engineering optimization. Genetic algorithms are relatively new optimization technique that can be applied to various problems, including those that are NP-hard. We implemented conventional Genetic Algorithm on this problem using a research on Evolutionary Computation System (ECM) using MAT lab. To date, the Genetic Algorithm successfully prevented premature and brought in improvement in Multiple Sequence Alignment for short sequences. However, for the dataset with long sequences, there is no significant improvement. The proposed project work provides evolutionary optimization for MSA with long sequences.

Keywords - *Multiple Sequence Alignment, Genetic Algorithms, optimization*

1. Introduction

Bioinformatics can be defined as conceptualizing biology in terms of Macromolecules (in the sense of physical-chemistry) and then applying "informatics" techniques (derived from disciplines such as applied mathematics, computer science, and statistics) to understand and organize the information associated with these molecules, on a large-scale. In short, bioinformatics leverage the techniques borrowed from computer science to solve problems in molecular biology. This exciting area is a new field, and the pace of research is driven by the large and rapidly increasing amount of data being produced for example, efforts to sequence the genomes of a variety of organisms. The areas where computer science can be applied range from assembly of sequence fragments, analysis of DNA, RNA and protein sequences, prediction and this paper work proposes an algorithm that provides optimal solution for the problem of Multiple Sequence Alignment. (MSA) is one of the most useful tools in bioinformatics. It is widely used to identify conservation of protein domains, RNA secondary structure and classification of biological sequences. However, it is recognized as one of the most challenging tasks in bioinformatics. Evolutionary algorithms are providing competitive solutions for engineering optimization. Genetic algorithms are relatively new optimization technique that can be applied to various problems, including those that are NP-hard. We implemented conventional Genetic Algorithm on this problem using ECM, a research Evolutionary Computation System using MAT lab.

2. An Evolutionary Optimization for MSA

To date, the Genetic Algorithm successfully prevented premature and brought in improvement in Multiple Sequence Alignment for short sequences. However, for the dataset with long sequences, there is no significant improvement. The proposed work provides evolutionary optimization [1] for MSA with long sequences.

2.1. MSA for Short Sequences

Multiple Sequence Alignment (MSA)[2] is a sequence alignment of three or more biological sequences which have an evolutionary relationship. MSA is among the most useful tools in bioinformatics. Thus in recent decades genetic algorithms (GAs) have been proposed to solve MSA problems. They can search through the solution space effectively and generate good alignment results. However, most of these techniques suffer from the problem of premature convergence which leads to a local optimal solution. In this project we make use of the method of reserved area in genetic algorithm [3] and do parameter tuning in order to prevent GA's premature convergence and improve its performance on MSA problem. Multiple sequence alignment (MSA) uses optimally aligning of three or more sequences of symbols

with or without inserting gaps. The objective is to maximize the number of matching symbols between the sequences and also use only minimum gap insertion, if gaps are permitted. All MSA algorithms also require an objective function to determine the relative quality of each possible multiple sequence alignment. Multiple sequence alignments of biological sequences provide a valuable source of information for investigating the properties, characteristics, and functions of novel sequences.

So far we implemented two sets of programs. One is conventional GA. The other is GA(4) with reserved area (GARS). The only difference between GA and GARS lies in that GARS introduce a concept named reserved area which is a subset of the whole population. In this area, individuals are selected to breed offspring based on their uniqueness values which measure the density around each individual. Also we have made use of the idea of reserved area to address the premature convergence problem which most GAs suffers from in MSA problems. The results show that for the dataset with short sequences, GA with reserved area successfully maintains population diversity and alleviates premature convergence during evolution process.

2.2. MSA for Long Sequences of Variable Length

In proposed work, we use the input sequences of variable length, which is pragmatic in real problems. Variable length input sequences are taken as input and crossover and mutation is applied on these sequences.

The overall performance of GA is improved. However, the reserved area mechanism did not show significant advantage over traditional GAs for long sequences. We think that this happened because uniqueness was calculated based on a single metric, which is the alignment score in our implementation. This cannot always accurately reflect differences between individuals. For long sequences which generally have higher alignment scores, the inaccuracy will be amplified.DNA, RNA and protein sequences contain useful information for biologists to understand the evolution of organism as well as the phenomena of life. Comparing these sequences further improves our understanding of evolutionary relationship and distance between different organisms. Multiple Sequence Alignment (MSA) is a sequence alignment of three or more biological sequences which have an evolutionary relationship.

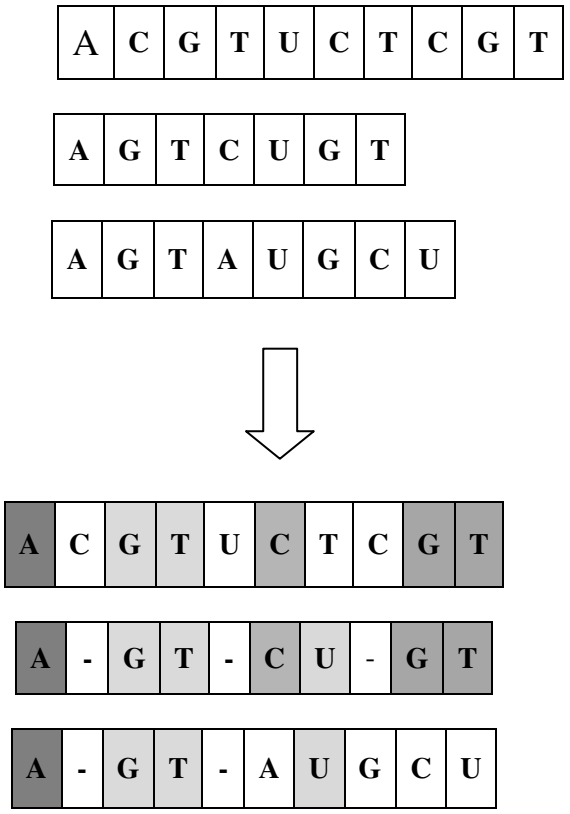

Fig 1. Example for an individual multiple input sequence

MSA is among the most useful tools in bioinformatics. However, it is an NP-complete problem [5], which is costly computational process in terms of both time and space. Most practical algorithms use heuristic methods because finding global optimization is extremely computationally expensive in most cases.

2.3. Flow of Proposed Evolutionary Algorithm for MSA

The proposed evolutionary algorithm specifically developed for aligning the variable length input sequences for large data set to generate the child sequences.
From these child sequences we will get the optimal solution by calculating the fitness function. The predicted characters may be expected in the generated sequences of the optimal solution .The Evolutionary algorithm for MSA is divided into two main modules .the first module applying cross over and mutation on parent sequences for aligning to generate child sequences. The second module used fitness function to trace optimal sequences.

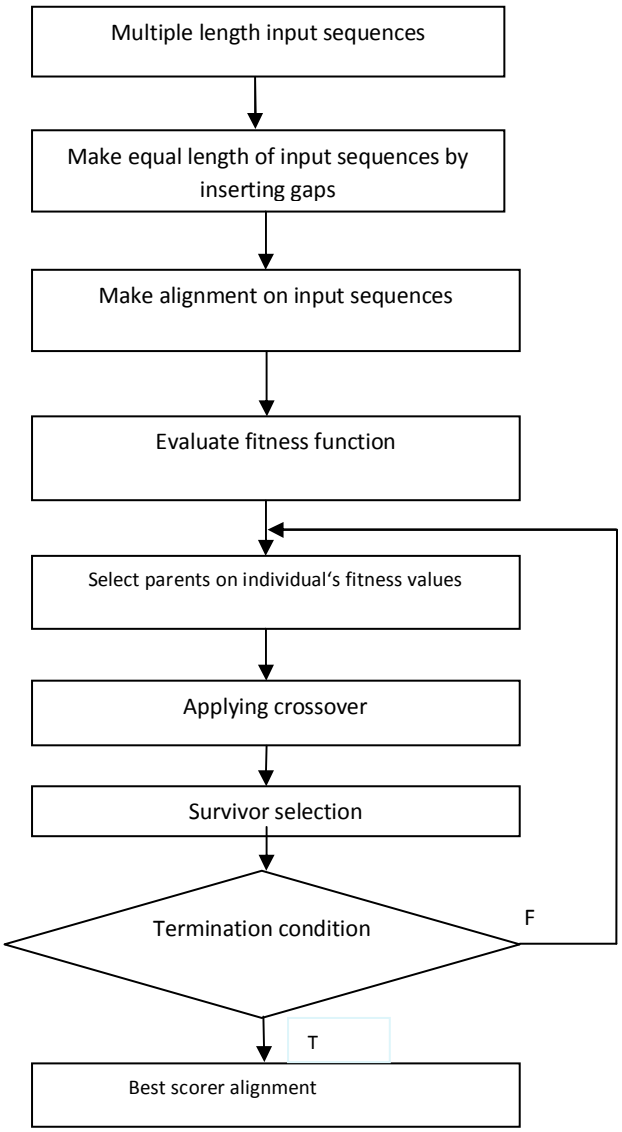

Fig 2. Flow of proposed evolutionary algorithm for MSA

Selection of sequences in optimal solution is based on fitness score so called sum of pair score.

3. Implementation

3.1 Implementation with MATLAB

Introduction

MATLAB (matrix laboratory) is a fourth-generation high-level programming language and interactive environment for numerical computation, visualization and programming.
MATLAB is developed by Math Works.

It allows matrix manipulations, plotting of functions and data , implementation of algorithms, creation of user interfaces, interfacing with programs written in other languages, including C, C++, Java, and Fortran; analyze data; develop algorithms; and create models and applications.

MATLAB is widely used as a computational tool in science and engineering encompassing the fields of physics, chemistry, math and all engineering streams. It is used in a range of applications including:

- Signal Processing and Communications
- Image and Video Processing
- Control Systems
- Test and Measurement
- Computational Finance
- Computational Biology

Multiple sequence alignment is an optimization problem that appears in many and diverse scientific fields. During the last decade, lots of increasing interest in the field of computational biology for methods that can efficiently solve this problem for sequences such as biological macromolecules, DNA and proteins. It was one of the most important and challenging tasks in computational biology because the time complexity for solving MSA grows exponentially with the size of the considered problem. Multiple sequence alignment is a natural extension of two-sequence alignment. In multiple sequences Alignment, the emphasis is to find optimal alignment for a group of sequences. Several applicable techniques were observed in this research, from traditional method such as dynamic programming to the extend of widely used stochastic optimization method such as Genetic Algorithms (GAs) and Simulated Annealing. A framework with combination of Genetic Algorithm and Simulated Annealing is presented to solve Multiple Sequence Alignment problems.

The methods mentioned above lack the ability to effectively search the huge solution space. Thus in recent decades genetic algorithms (GAs) have been proposed to solve MSA problems. They can search through the solution space effectively and generate good alignment results.

However, most of these techniques suffer from the problem of premature convergence which leads to local optima. In this paper work we make use of the method of reserved area in genetic algorithm and do parameter tuning in order to prevent GA's premature convergence and improve its performance on MSA problem and also implemented on various length input sequences.

3.2 Result Analysis

Step – 1:

Reading multiple sequences of various length. These sequences are used as input for proposed algorithm.

Commands Used:

>> mouse1= getgenbank ('AF179942')
>> mouse2= getgenbank ('AF179933')
>> mouse3= getgenbank ('AF179939')

Now sequences related to gene structure of three mouse parents are taken as input sequences.

Mouse1InputSequence

```
Frame 1

000001   atctccatcatgaatcaggcagctgtaactctgagatctctagactgcatgggtccagcaacct
000065   tagaacacagagcgagagaagccagcatccagagatgtgcaggcaaggtgggatgaggcttttta
000129   catcaagccatcag

Frame 2

000001   atctccatcatgaatcaggcagctgtaactctgagatctctagactgcatgggtccagcaacct
000065   tagaacacagagcgagagaagccagcatccagagatgtgcaggcaaggtgggatgaggcttttta
000129   catcaagccatcag

Frame 3

000001   atctccatcatgaatcaggcagctgtaactctgagatctctagactgcatgggtccagcaacct
000065   tagaacacagagcgagagaagccagcatccagagatgtgcaggcaaggtgggatgaggcttttta
000129   catcaagccatcag
```

Fig 4. Mouse1 input sequence

Mouse2 Input Sequence

```
Frame 1

000001   agtcaggaggtccagacctgtcctcatgtttcccacgatctccctgcctctccaccctgaagaa
000065   gcaacaaaagcacctgtgacaacact

Frame 2

000001   agtcaggaggtccagacctgtcctcatgtttcccacgatctccctgcctctccaccctgaagaa
000065   gcaacaaaagcacctgtgacaacact

Frame 3

000001   agtcaggaggtccagacctgtcctcatgtttcccacgatctccctgcctctccaccctgaagaa
000065   gcaacaaaagcacctgtgacaacact
```

Fig 5. Mouse1 input sequence

Mouse3 Input Sequence

```
Frame 1

000001   gacaccagcccataagacgatcccattcacattcagaaagcatcttcagcgtagcctctctgga
000065   aatgctctcagagacaaccccagagctgtatctcctaa

Frame 2

000001   gacaccagcccataagacgatcccattcacattcagaaagcatcttcagcgtagcctctctgga
000065   aatgctctcagagacaaccccagagctgtatctcctaa

Frame 3

000001   gacaccagcccataagacgatcccattcacattcagaaagcatcttcagcgtagcctctctgga
000065   aatgctctcagagacaaccccagagctgtatctcctaa
```

Fig 6. Mouse1 input sequence

Step – 2:

The input sequences are placed in a vector to apply MSA algorithm. Then the sequences are aligned by applying mat-lab function.

seqs=
{'atctccatcatgaatcaggcagctgtaactctgagatctctagactgcatgggtc
cagcaaccttagaacacagagcgagagaagccagcatccagagatgtgcaggc
aaggtgggatgaggcttttacatcaagccatcag','agtcaggaggtccagacct
gtcctcatgtttcccacgatctccctgcctctccaccctgaagaagcaacaaaagc
acctgtgacaacact','gacaccagcccataagacgatcccattcacattcagaaa
gcatcttcagcgtagcctctctggaaatgctctcagagacaaccccagagctgtat
ctcctaa'}

ma=multialign (seqs)
Show alignment (ma)

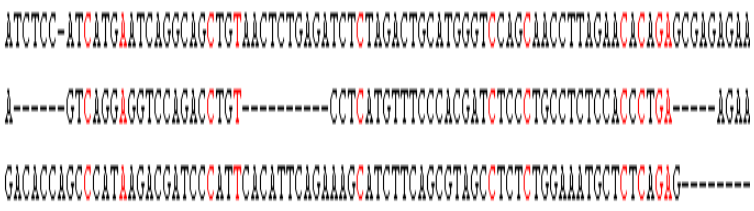

Fig 7. Multiple sequence alignment output sequence

Step - 3:

Calculating the fitness value of input sequences. The fitness function is applied on input sequences.

[X fval reason] = ga (@rastriginsfcn, 2)

Optimization terminated: average change in the fitness value less than options.TolFun.

x = 0.0015 -0.0166
Fval = 0.0553
Reason = 1

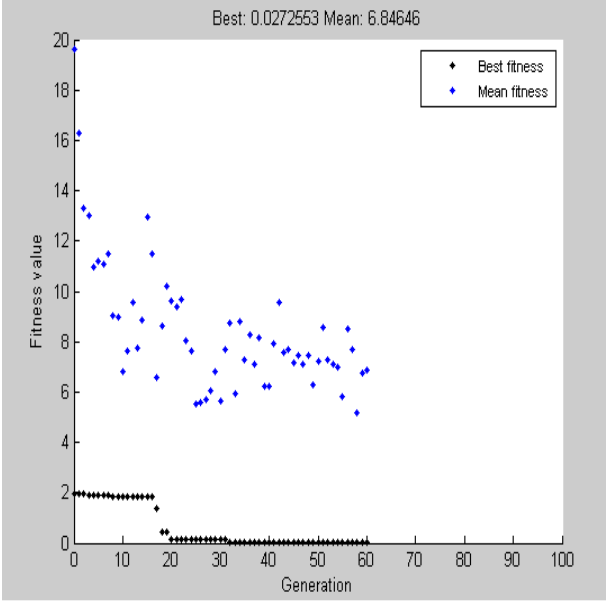

Fig 8. Output graph of Fitness function

The sequence with the best fitness value is the result of MSA.

4. Conclusion & Future Work

Here we have presented a detailed implementation of multiple sequence alignment using Evolutionary Algorithm. Also in current work we allow the input sequences of various length .The variable length input sequences are pragmatic to real world. Biological sequences generated by multiple alignments provides valuable source of information for investigating properties, characteristics of the future generated sequences. The proposed algorithm provides an evolutionary optimization for Multiple Sequence Alignment with long sequences.

5. Acknowledgments

Author is thankful to Assistant Professor Mr. P. Rajesh ,Professor Miss. Karteeka for their valuable suggestions and help.

References

[1] D. B. Fogel. An introduction to simulated evolutionary optimization. IEEE Transactions on Neural Networks, 5:3{14, 1994.

[2] K. Chellapilla and G. B. Fogel. Multiple sequence alignment using evolutionary programming. In Proceedings of the 1999 Congress on Evolutionary Computation (CEC'99), pages 445{452, 1999.

[3] Yang Chen, Jinglu Hu, Kotaro Hirasawa, and Songnian Yu. GARS:an improved genetic algorithm with reserve selection for global optimization. In Proceedings of Genetic and Evolutionary Computation Conference (GECCO'07), pages 1173{1178, 2007.

[4] L. Davis. Handbook of Genetic Algorithms. Van Nostrand Reinhold, New York, 1991.

[5] Wang L and Jiang T. On the complexity of multiple sequence alignment. Comput Biol, 4:337{48, 1994.

[6] Gibson TJ Thompson JD, Higgins DG. Clustal w: improving the sensitivity of progressive multiple se-quence alignment through sequence weighting, positions-speci_c gap penalties and weight matrix choice. Nucleic Acids Res, 22:4673{4680, 1994.

[7] Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. Basic local alignment search tool. J Mol Biol, 215:403{410, 1990.

[8] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. Fast and elitist multiobjective genetic algorithm:NSGA-II. IEEE Transactions on Evolutionary, 6:182{197, 2002.

[9] Andreas W., Indra M., and Gerhard S. An enhanced rna alignment benchmark for sequence alignment programs. Algorithms Mol Biol, 2:19, 2006.

First Author: K.Lohita Lakshmi, is a M.Tech scholar at VVIT(Vasireddy Venkatadri Institute of Technology),Nambur. She got her Master of Computer Applications Degree from Venkateswara University and she got her M.Tech from Nagarjuna University. She is very much interested in Data Mining, Bioinformatics and Computer networks.

Second Author: P.Rajesh received the M.Tech degree in computer science and engineering (CSE) from Jawaharlal Nehru Technological University Hyderabad in 2009. He is currently pursuing Ph. D degree in the department of computer science and engineering from Jawaharlal Nehru Technological University Hyderabad and working as an assistant professor in CSE department at Vasireddy Venkatadri Institute of technology, Guntur, Andhra Pradesh. His research interests are in the area of Data mining, Information security, Privacy preserving data publishing and sharing.