International Journal of Information and Computation Technology.
ISSN 0974-2239 Volume 3, Number 5 (2013), pp. 465-470

http://www. irphouse.com /ijict.htm

Changing Mutation Operator of Genetic Algorithms for Optimizing Multiple Sequence Alignment

Manish Kumar[1] and Haider Banka[2]

Department of Computer SCIENCE and Engineering,
Indian School of Mines, Dhanbad, INDIA.

Abstract

In this paper, we have proposed three mutation operators to produce next generation in Genetic Algorithm. Our research work focuses on solving Multiple Sequence Alignment (MSA) problem by using GA with different mutation operators like gap shift, space merging, full gap column remover. To know the population evolution and quality of the sequence aligned, several studies and tests have been performed on both conventional GA and GA with improved mutation operator over BALIBASE datasets. From our experiment, it is found that GA with improved mutation operator outperforms conventional GA, in terms of quality of the sequences aligned.

Keywords: Multiple sequence alignment; genetic algorithm; mutation operator; Bioinformatics.

1. Introduction

Bioinformatics is conceptualising biology which uses “informatics techniques" to understand and organise the information associated with the molecules, on a large scale. In short, bioinformatics is a management information system for molecular biology and has many practical applications in the field of Molecular medicine, microbial genome applications, agriculture, and animals. Multiple sequence alignment of different biological sequence (DNA/RNA/PROTEIN) is one of the common tasks in bioinformatics. In a multiple sequence alignment, homologous residues among a set of sequences are aligned together in columns (Feng et al,1985). Homologous is meant in both the structural and evolutionary sense. MSA is the problem of lining up the characters of string in the best possible way.

Mutation is viewed as a background operator to maintain genetic diversity in the population. Mutation prevents the algorithm to be trapped in a local minimum and plays an important role in recovering the lost genetic materials. It is an insurance policy against the irreversible loss of genetic material. A genetic algorithm is search technique used in computing to find true or approximate solutions to optimization and search problems (Michalewicz 1996). Genetic algorithms are a particular class of evolutionary algorithms that use techniques inspired by evolutionary biology such as inheritance, mutation, selection, and crossover (also called recombination). (Meng et al,1999)

Current multiple sequence alignment algorithms work well for sequences with high similarity but do not scale well when either the length or number of the sequences is large or if the similarity is low. Research is going on to develop an evolutionary programming (EP) algorithm for multiple sequence alignment.(Kumar et al,1999). An approach for MSA with GA was introduced by Zhang and Wong in 1994. In their research the GA simply evolves the number and position of gaps within conserved segments of an alignment. But, the assumption was that, such conserved segments always exist is never realistic or biologically sound. Therefore, there method can only compare long, highly similar sequences. Similarly, some researcher have also applied GA to MSA with a tool know as sequence alignment by Genetic Algorithm (SAGA).

In this paper we have compared the traditional mutation operator of GA with the new proposed mutation operators, in order to show how the quality of the sequences (which is aligned) improves in terms of scores. The tests for the alignment problem are performed with BALIBASE datasets.(http://bips.u-stra sbg.fr/en/fr/products/datasets/ balibase)

The rest of the paper is organized as follows section 2 gives a basic concept of sequence alignment and GA along with the concepts of datasets used in the experiment, section 3 gives a brief idea about the proposed Mutation operators, section 4, where the experimental results are discussed, section 5 summarizes our findings and come to conclusion.

2. Preliminaries

2.1 Sequence Alignment

In sequence alignment two or more strings are aligned together in order to get the highest number of matching character (Pal et al,2006). Gaps may be inserted into a string in order to shift the characters into better matches. Typically a scoring function is used to rank different alignment so that biologically plausible alignment scores higher. The task of optimal sequence alignment is to find the best possible alignment for a given scoring function and a set of sequences.(Wang et al, 2005).

2.2 Genetic Algorithms

Genetic Algorithm is a directed search algorithms based on the mechanics of biological evolution (L B Booker et al,1987) The genetic algorithms process starts

with an initial population composed of random chromosomes, which form the first generation. Crossover is used to combine genes from the existing chromosomes and create new ones. Then, the best chromosomes are selected to form the next generation. This selection is based on a fitness function which assigns a fitness value to every chromosome. The ones with the best fitness value "survive" to give offspring for the new generation, and the process is repeated until satisfactory solutions evolve. (Komas et, al 1996)

2.3 Balibase Dataset

Balibase is a database of manually-refined multiple sequence alignments specifically designed for the evaluation and comparison of multiple sequence alignment programs.

2.4 PAM Matrix

A PAM is a set of matrices used to score sequence alignment by assessing the similarity of two aligned protein sequences.

3. Proposed Methods

3.1 Proposed Mutation Operators

With a view to improve the results of Genetic Algorithm in terms of score, we have proposed three different mutation operators namely Gap shift operator, Space merging operator, full gap column remover operator. To improve the results, we will check and apply these mutation operators in each operation of MSA problem as applicable.

3.1.1 Gap shift operator

In Gap shift mutation operator, a gap is randomly chosen in the alignment and it is moved to some other position in the alignment so that better alignment can be formed. We will move the gap at different location until the fitness of the alignment improves than the original one.

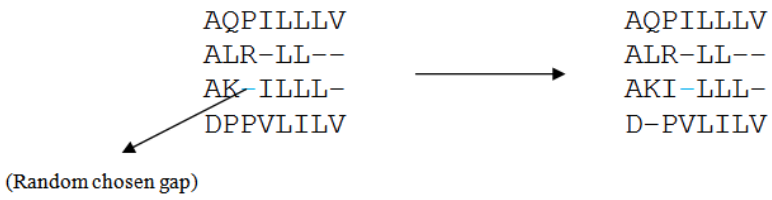


Figure 1: Gap shift mutation operator.

Figure 1 shows that a random gap is chosen from column 3 in the alignment and interchanged with column 4 in the same row.

3.1.2 Space merging operator

This operator will merge two or more gaps together by randomly selecting the gaps and moving them in some different position in the same alignment, so that fitness may improve.

3.1.3 Full gap column remover operator

This operator removes all the column(s) containing only gaps.

4. Experimental Results

4.1 Implementation

In this section we have solved a multiple sequence alignment problems by Genetic Algorithm using data from BALIBASE dataset. Experiment is performed with the help of selection, crossover and mutation operators, in order to produce new solution with defined number of generation.

In the experiment the population size is taken as 10 , tournament selection is employed with tournament size 2,crossover rate and Mutation rate taken as 0.8 and 0.01 respectively. 100 runs of GA is carried out and optimal score in each run is calculated as the results.

In the experiment, the sum of pair for the sequences is calculated (De silva 2009), which is used as a tool to calculate fitness.

$$\text{Sum} - \text{of} - \text{pair (SP)SCORE} = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} scoring\ matrix(\ l_i, l_j)$$

The score is calculated by scoring all the pair wise comparison between each residue in each column of an alignment and adding the scores together.

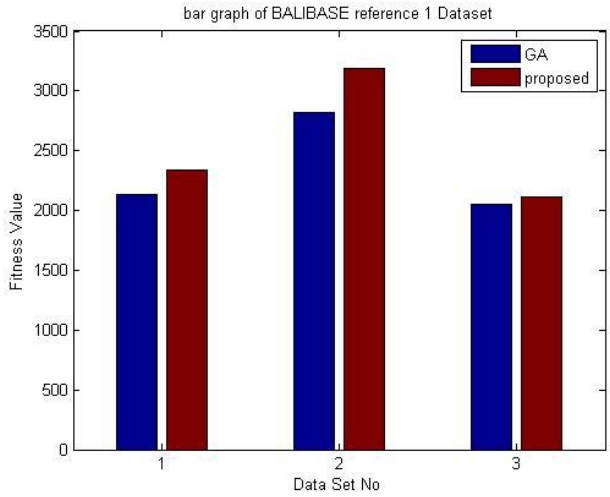
This score will act as a measure to evaluate fitness of the population at each generation. Score for each column for the given sequences is calculated as per the data available in the PAM 250 Matrix.(http://www.dtu.dk/gymnasier/data/PAM250). The result of this experiment is shown on Table 1.

The main objective of the research work is to use the proposed mutation operators in solving MSA problem. Experiment has been performed on different datasets of BALIBASE (1aab, 1wit, 1boa) (refer Table 1 and Figure 2) by using Gap shift mutation operator to solve MSA problem and the result of it has been compared with the simple mutation operator. The result given in Table 1.is for gap shift mutation operator when compared with simple GA , Similar result can also be found out by using other two mutation operators i.e. Space merging, full gap column remover.

Table 1: Comparative result of Scope of GA with gap shifting mutation operators on three different datasets.

Gap shift mutation operator	Simple mutation operator	Gap shift mutation operator	Simple mutation operator	Gap shift mutation operator	Simple mutation operator
For 1aab Dataset		For 1wit dataset		For 1boa dataset	

2095	1881	2757	2759	1821	1722
2213	1881	2902	2763	1874	1722
2213	2012	2902	2763	1877	1736
2292	2048	2902	2763	1917	2049
2292	2053	2902	2763	1942	2049
2297	2116	3086	2790	2034	2049
2297	2116	3121	2790	2043	2049
2306	2137	3184	2790	2061	2049
2334	2137	3184	2790	2061	2049
2335	2137	3184	2821	2111	2052



(1=1aab dataset, 2= 1wit dataset, 3=1boa dataset)

Figure 2: Bar graph comparison of GA with simple mutation operator and our proposed method.

5. Conclusion

This paper has presented three different mutation operator to be applied in solving MSA problem .In order to know the population evolution and quality of the sequence aligned we have introduced three mutation operator to solve MSA problem. As we can see from our experiment, changing mutation operator in solving MSA problem has brought an improvement in the results of MSA in terms of score. We also believe that if the option of selecting mutation operators is made randomly and according to the need of the problem, then the results would improve , as testing and trying every mutation operator for each operation of MSA is a bit time consuming. Using the operators would help in yielding solution that are closer to the optimum solution.

6. Acknowledgments

This work was supported by the Council of Scientific and Industrial Research (CSIR) New Delhi, India, under the Project Grant No. [22 (0586)/12/EMR-II].

References

[1] C. Wang and E. J. Lefkowitz, (2005) "Genomic multiple sequence alignments: refinement using a genetic algorithm," *BMC Bioinformatics*, vol. no **6**

[2] **D. F. Feng, M. S. Johnson, R. F. Doolittle (1985) "Aligning amino acid sequences: Comparison of commonly used methods"** ***Journal of Molecular Evolution*****, vol. no** 21**, pp 112-125.**

[3] Da Silva, F.J.M.; Sanchez Perez, J.M.; Pulido, J.A.G.; Rodriguez, M.A.V.,(2009) "Optimizing Multiple Sequence Alignment by Improving Mutation Operators of a Genetic Algorithm," *Ninth International Conference on Intelligent Systems Design and Applications* , pp.1257-1262, Pisa.

[4] Kumar Chellapilla, and G. B. Fogel, (1999) "Multiple sequence alignment using evolutionary programming."*Proceedings of the 1999 Congress on Evolutionary computation*, vol. no **1** pp. *445-452*. Washington, DC.

[5] Kosmas Karadimitriou and Donald H. Kraft (1996) "genetic algorithms and the multiple sequence alignment problem in biology" *Proceedings of the Second Annual Molecular Biology and Biotechnology Conference*, Baton Rouge, LA.

[6] Meng, Q.C.; Feng, T.J.; Chen, Z.; Zhou, C. J.; Bo, J. H.,(1999) "Genetic algorithms encoding study and a sufficient convergence condition of Gas." *Proceedings of 1999 IEEE International Conference on Systems, Man, and Cybernetics*, vol. no **1**, pp.649-652,Tokyo.

[7] **S. K. Pal, S. Bandyopadhyay, and S. S. Ray, (2006)"Evolutionary computation in bioinformatics: A review," IEEE Transactions on Systems Man and Cybernetics Part C-Appl and Rev, vol. no** 36**, pp. 601-615.**

[8] Z. Michalewicz, (1996), *Genetic algorithms + data structures = evolution programs - Third, Revised and Extended Edition, 3 ed: Springer.*

[9] L B BOOKER**,** D.E. Goldberg, J. H. Holland (1987) "classifer system and genetic algorithm, Technical report no 8", university of Michigan.