International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

An Efficient Genetic Algorithm based Framework for Sequence Alignment in Computational Biology

N. Senthilkumaran
Centre for Bioinformatics, Department of Computer Science and Applications,
The Gandhigram Rural Institute (Deemed to be University), Gandhigram, Tamil Nadu, India.

Abstract **- The alignment of biological sequences is a key component of bioinformatics, enabling scientists to determine the structural, functional and evolutionary relationships between DNA, RNA and protein sequences. When applied to large data sets with long sequences, traditional techniques of precise alignment, such as dynamic scheduling may be computationally expensive and time consuming. This study proposes a heuristic optimization framework based on genetic algorithms to effectively solve the bioequation problem and to overcome these computational overheads. The proposed method presents candidate alignments as a chromosome population and is based on the concepts of genetics and natural selection. To iteratively explore the field and find high-quality matches, it combines appropriate coding schemes, specialized fitness functions, and tailored genetic operators such as selection, selection-crossing, and mutation. Compared to conventional precision methods, experimental testing shows that the proposed genetic algorithm can produce competitive alignment, while potentially reducing computational and memory requirements. This work provides a scalable framework for aligning biological sequences, particularly in the context of large-scale analysis of genomic data, and shows the potential of evolutionary computing to solve complex bioinformatics problems.**

Keywords - Bioinformatics; Biological Sequence Alignment; Genetic Algorithms; Multiple Sequence Alignment;

I. INTRODUCTION

Computer analysis is now an essential part of modern molecular biology and bioinformatics, owing to the exponential growth of biological data in the post-genome era. The main method in this area is biological sequence alignment, which involves sequencing DNA, RNA, or protein sequences to look for similarities. These similarities may reveal important functional, structural, and evolutionary relationships between organisms. Accurate sequence alignment is a key step in the analysis of biological data, whether to track evolutionary lineage, to identify conserved protein domains, or to identify genetic mutations associated with disease. Biological sequence alignment, and in particular multiple sequence alignment (MSA)[5], poses considerable computational problems, despite its basic function. The search area is expanding rapidly as the number and length of sequences increase, making it more difficult to quickly find ideal matches.

Conventional precision methods, such as the Smith-Waterman algorithm for local alignment and the Needleman-Wunsch algorithm for global alignment, use dynamic scheduling to find the best alignment according to a predefined scoring scheme[8]. These techniques provide the best pairwise alignment solutions, but their memory and processing requirements may increase significantly when working with large datasets. Moreover, the computational complexity of precise methods makes their direct extension to multiple sequence alignment impractical for large numbers of sequences[10]. Researchers are increasingly exploring heuristic optimization methods to overcome these limitations, and genetic algorithms have shown promise. Genetic agents such as selection, mutation, and recombination are used by GAs, motivated by the concepts of biological evolution, to explore complex environments and find better solutions[15].

This paper represents candidate sequence alignment as a population of possible solutions that changes over generations according to a pre-specified fitness function. For the assessment of the quality of alignment, the fitness function may include substitution matrix and alignment score notation. This study explores the use of genetic algorithms to align biological sequences. The proposed method aims to explore complex problem spaces while striking a balance between the quality of alignment and the computational efficiency of evolutionary computing. The framework is designed to provide a scalable strategy for optimizing biosequencing, particularly in the analysis of large genomic and proteomic datasets.

This is how the remainder of the document is organized. Related works in Section II, Section III describes the use of a genetic algorithm for sequence alignment. Results and discussion in Section IV. Lastly, Section V presents the conclusions.

II. RELATED WORKS

Improved alignment of sequences may be deterministically achieved by dynamic algorithms such as Needleman-Wunsch or Smith-Waterman[1]. The first is a known algorithm for global sequence alignment and the second is used for local sequence alignment. However, these algorithms have been designed ideally for the alignment of pair sequences and, because of their computational costs, it is impractical to perform alignment for three or more sequences[2]. Therefore, multiple sequence alignment algorithms[3] have been proposed to smooth out the high computational costs. This program can be based on a variety of heuristics, such as ant colony, simulated extraction, progressive alignment, genetic algorithms, among others. As regards GA, it can be used to solve problems of multiple

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

sequence alignment by means of an evolutionary approach[4]. In this method, individuals in a population are exposed to mutation, recombination and selection to develop a population of potential solutions, the biological significance of which is measured by an objective function[5]. The first tool developed to solve the problem of multiple sequence alignment by means of a genome algorithm was SAGA (sequencing alignment by genetic algorithm).

An computerized scheduler chooses its complex pack of 22 operators, which includes recombination and mutation [6]. However, as compared to a uniform choice of operators, several investigations have demonstrated that the complexity of SAGA is superfluous and that the automatic scheduler does not enhance the ultimate quality of the alignments [7]. MSA-GA is another potential tool that employs GA for multiple sequence alignment of nucleotide and amino acid sequences [8]. Because of its more straightforward computational methodology, this tool is more effective at solving global multiple sequence alignment problems [9].

III. SEQUENCE ALIGNMENT USING GENETIC ALGORITHM

A bioinformatics technique called biological sequencing alignments uses two or more sequences of DNA, RNA, or protein in such a way that similarities can be found[11]. It helps scientists understand the structural, functional, and evolutionary links between biological sequences. Three types of sequence alignment are global sequence alignment, local sequence alignment and multiple sequence alignment (MSA). Figure 1 shows genetic algorithm frame work[12].

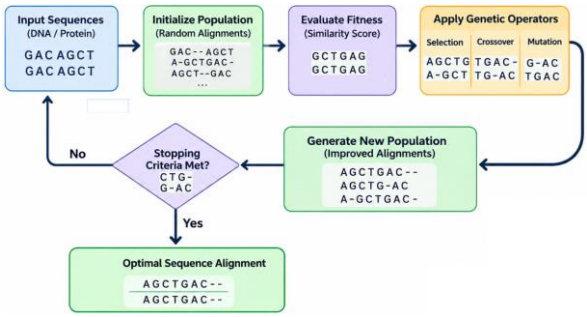


Fig. 1. Genetic algorithm frame work for sequence alignment

Global sequence alignment is the alignment of two sequences together. Ideal for sequences of similar length and common structure. Needleman-Wunsch is a fairly standard algorithm[13]. One example is comparing two genes that are closely related. The local alignment of the sequences determines which region or regions of two sequences are the most similar. Useful if the sequences have different lengths or only share a small part. Smith-Waterman's is a generic algorithm[14]. This algorithm used for the identification of a conserved domain in proteins. Multiple Sequence Alignment (MSA) is the simultaneous alignment of three or more biological sequences. finds differences and preserved areas. Common tools include the MAFFT, MUSCLE and Cluster Omega tools[16]. It is used for comparing the same gene in different species as following Figure 2.

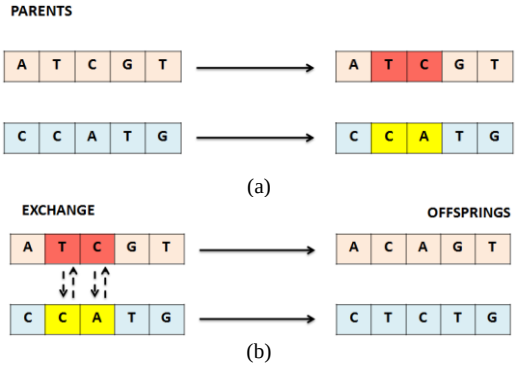


Fig. 2. Crossover operation for DNA sequence
a) cut for operation b) after exchange

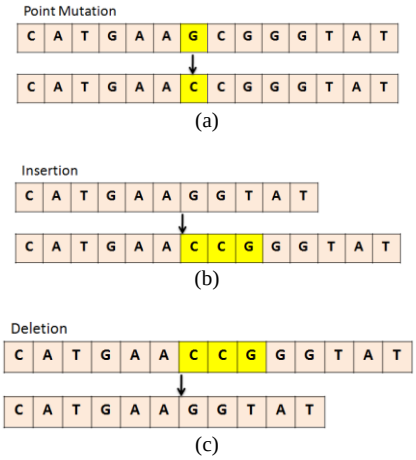
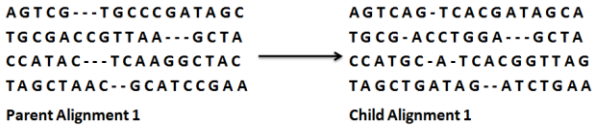


Fig. 3. Mutation operation for DNA sequence a) Point mutation b) insertion for mutation c) deletion for mutation.

In Figure 3, The alignment problem is formulated as an evolutionary optimization problem in the methodology of applying genetic algorithms to the alignment of biological sequences. By simulating natural selection principles and using genetic operators such as selection, crossing and mutation, the proposed framework will methodically search the universe for optimal or near optimal alignment[17]. The algorithm aims to reduce computational complexity while increasing the quality of alignment by refining candidate solutions over generations[18].

IV. RESULTS AND DISCUSSION

By choosing a cutting point at random and switching the string from that point with a predetermined probability, a crossover operation is carried out over the two biological sequence strings. This operator first selects a column at random from the parent alignments and creates a cut point there, as seen in Figures 4 and 5. Two new offspring, often referred to as children, are then created by switching the various components of the parents. Gaps may be added to the children produced by this kind of procedure[19].



(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

Fig. 4. One point crossover for parent - child alignment 1

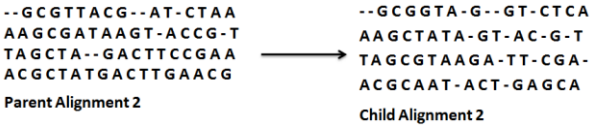


Fig. 5. One point crossover for parent - child alignment 2

This algorithm is kept from becoming stuck in a local minimum by mutation. It aids in the recovery of missing genetic materials and randomly distributes genetic information among other people. A few bits of a chromosome are randomly flipped during a mutation procedure.

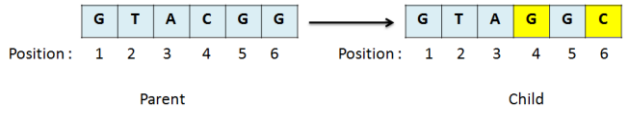


Fig. 6. Exchange Mutation operator

Exchange mutation operator is explained in Figure 6 in which, the position of two nucleotide (position 4 and 6) are exchanged which are randomly chosen.

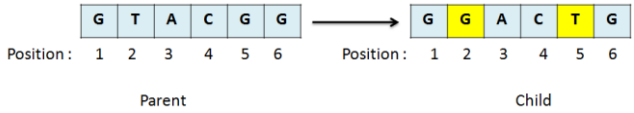


Fig.7. Reverse Mutation operator

Figure 7 provides a clear illustration of the reverse mutation operator. Here, two randomly selected positions, 2 and 5, have limited the sequence S. This sequence's nucleotide order will be inverted in the same manner as the preceding operation[20].



Fig.8. Position mutation operator

Three nucleotides were selected at random to occupy the various positions in the position mutation operator, which are not necessarily consecutive (2 < 4 < 6). The nucleotide that is currently at position 2 will move to position 4, the nucleotide that was at position 4 will move to position 6, and the ucleotide that is currently at position 2 will move back to position 2. The procedures mentioned above are shown in Figure 8.



Fig.9. Inverse mutation operator

The chromosomes were divided into two sections to create the two nucleotide sections shown in Figure 9. Every nucleotide in every section of a child is duplicated and positioned inversely in the same section. The key finding of the results is that, in order to solve complex problems such as MSA, it is not necessary to employ complex operators; instead, it is necessary to precisely define operators that can enhance the population of genetic algorithms[21].

Genetic algorithms are an effective and adaptable tool for solving a variety of bioinformatics problems. The evolutionary aspect of GAs can be used to solve any problem where solutions can be ranked in a meaningful way to measure fitness. This work elaborates on the advantages of employing a GA and its variants based technique for BI performance enhancement. Future in-depth research and GA implementation will be carried out with a variety of goals, including energy, time, and resources.

V. CONCLUSION

Genetic algorithm (GA) provides a versatile framework of heuristic optimization for examining complex search environments and finding better matches. The proposed method aims to improve the quality of alignment while reducing the risk of early convergence towards local optimum by combining the concepts of natural selection, crossing and mutation with biologically-based scoring systems such as the sum of pairs and substitution matrices. But the design of the fitness function, genetic operators, and parameterization all have an effect on how efficient it is.

REFERENCES

[1] Zhimin Zh, Zhong WC. Dynamic programming for protein sequence alignment. Int BioScience Bio Technol. 2013

[2] Wei-C C, Yu JC, Chien CC, Der TL, Jan MH. Optimizing a map reduce module of preprocessing high throughput DNA sequencing data. IEEE International Conference on BigData. 6-9, 2013.

[3] Althaus, E., Caprara, A., Lenhof, H.-P. and Reinert, K. A branch-and-cut algorithm for multiple sequence alignment. *Math. Programm.*, vol. 105, pp. 387-425, 2006.

[4] Althaus, E. and Canzar, S. LASA: A tool for non-heuristic alignment of multiple sequences. *Bioinformatics research and development chapter*,Springer, pp. 489-498, 2008.

[5] Lee, Z. J., Su, S. F., Chuang, C.,C. and Liu, K. H. Genetic algorithm with ant colony optimization (GA-ACO) for multiple sequence alignment. *Applied Soft Computing*, vol.8, pp 55-78, 2008.

[6] Mount, W. D. Bioinformatics: sequence and genome analysis. *Cold Spring Harbor Laboratory Press, New York*, 2004.

[7] Isokawa, M., Wayama, M. and Shimizu, T. Multiple sequence alignment using a genetic algorithm. Genome Informatics 7: 176-177, 1996.

[8] Naznin F, Sarker R, Essam D. Vertical decomposition with Genetic Algorithm for Multiple Sequence Alignment. BMC Bioinformatics, 12: 353-79, 2011.

[9] Gondro C, Kinghorn BP. A simple genetic algorithm for multiple sequence alignment. Genet Mol Res, 6(4): 964-82, 2007.

[10] Zara-Mirakabad F., Ahrabian H., Sadeghi M., Hashemifar S., Nowzari-Dalini A., and Goliaei B. Genetic algorithm for dyad pattern finding in DNA sequences. *Genes and Genetic Systems* 84: 81-93, 2009.

[11] Ooi C.H., and Tan, P. Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. *Bioinformatics* 19:37-44, 2003.

[12] Montaseri S., Zare-Mirakabad F., and Moghadam-Charkari N. RNA-RNA interaction prediction using genetic algorithm. *Algorithms for Molecular Biology*, 9:17, 2014.

[13] Moore J. H., Hahn L. W., Ritchie M. D., Thornton, T. A. and White, B. C. Routine discovery of complex genetic models using genetic algorithms. *Applied Soft Computing* 4:79-86, 2004.

[14] Dowell R. and Eddy S. Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints. *BMC Bioinformatics* 7:400, 2006.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

[15] Radenbaugh, A. J., Austin T. *Genetic algorithm for multiple sequence alignment:GAMSA.* San Jose State University, Department of Computer Science, Course 286: Advanced Bioinformatics, 2006.

[16] Wang, C, & Lefkowitz, E.J. Genomic multiple sequence alignments: Refinement using a genetic algorithm. *BMC Bioinformatics, 6:200,* 2005.

[17] P. Agarwal, R. Chauhan, Alignment of multiple sequences using GA method, Int. J. Emerg. Technol. Comp. Appl. Sci. 4, 411–421, 2013.

[18] L.A. Anbarasu, P. Narayanasamy, V. Sundararajan, Multiple molecular sequence alignment by island parallel genetic algorithm, Curr. Sci. 78, 858–863, 2000.

[19] D.E. Arenas, H. Ochoterena, V.K. Rodriguez, Multiple sequence alignment using a genetic algorithm and GLOCSA, J. Artif. Evol. Appl., 963150, 2009.

[20] M. Botta, G. Negro, Multiple sequence alignment with genetic algorithms, in: F. Masulli, L.E. Peterson, R.Tagliaferri (Eds.), Computational Intelligence Methods for Bioinformatics and Biostatistics (CIBB 2009), Lecture Notes in Computer Science, vol.6160, Springer, Berlin, Heidelberg, pp. 206–214, 2010.

[21] H. Carrillo, D. Lipman, The multiple sequence alignment problem in biology, SIAM J. Appl.Math. 48, 1073–1082, 1988.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)