A Note on Bioinformatics using Genetic Algorithms

Dinesh Prasad Sahu
SC&SS JNU
New Delhi, India
110067

Mohammad Sajid
SC& SS JNU
New Delhi, India
110067

Shiv Prakash
SC& SS JNU
New Delhi, India
110067

ABSTRACT
The Bioinformatics (BI) is a sequence alignment, usually three sequences which can be RNA, DNA and proteins. Because the three or more given sequences can be of large lengths, aligning them by hand can be time consuming and in some cases traditionally impossible. Thus BI comes into use thereby aligning each sequence, and revealing the similar part of the given gene. BI finds great use in bioinformatics where it is used to predict the protein structure, its function, family or its domain. These problems are related to AI and can be classified in the NP complete domain of problems. Thus with the goal of identifying maximum similarities among the sequences, we can use various approaches and techniques like Genetic Algorithms (GA) and its variant in BI.

General Terms
Bioinformatics, NP complete, GA

Keywords
Bioinformatics Alignment, NP complete, GA

1. INTRODUCTION
The Bioinformatics (BI) is a sequence alignment, usually three sequences which can be RNA, DNA and proteins [1, 2, 3]. Because the three or more given sequences can be of large lengths, aligning them by hand can be time consuming and in some cases traditionally impossible [4, 5]. Thus BI comes into use thereby aligning each sequence, and revealing the similar part of the given gene. BI finds great use in bioinformatics where it is used to predict the protein structure, its function, family or its domain. These problems are related to AI and can be classified in the NP complete domain of problems. Thus with the goal of identifying maximum similarities among the sequences, we use various approaches and techniques in BI [2, 3].

2. PROBLEM DEFINITION
To start with, BI refers to the problem by which we can align three or more sequences of maximum symbols. To facilitate aligning process, we can insert gaps between the symbols. Its main objective is to note the number of symbols giving exact match between sequences, taking care that we have minimum gap insertions. BI finds use in several fields including molecular biology, computer science and geology. The DNA molecules are chains of nucleotides which are of four types namely A, T, G and C. We know the primary structure of a protein to be a linear chain of various amino acids. So the objective of our project is to make a tool that can be used for aligning given multiple DNA sequences.

3. RELATED WORK
There are various approximation methods which are put to use in BI along with the following [2, 3, 4] techniques:

1. Dynamic Programming- The motivating factor for use of DP with the two given sequences is that it secures us with an optimal alignment of the given sequences (provided we have a specific scoring scheme).
2. Progressive alignment- In this we first align the almost alike looking sequences using DP and then in progression we keep on adding the sequences which differ in a few ways from original alignment to our initial alignment. Examples include CLUSTALW and CLUSTALX programs.
3. Iterative alignment- Here an initial aligning of sequences is carried out. In an iterative way, revision of the alignments is done to get a more feasible result. It aims at improving the overall alignment score. Several algorithms under this include Mult Alin, PRRP, and DIALIGN.
4. Statistical modeling- We have basically used genetic algorithms in our process to align sequences because we have to first try to generate as much different alignments of the sequences as we can do using the techniques of rearranging the sequences, which can help manage gaps and the genetic recombining events. SAGA (Serial Alignment by Genetic Algorithm) can be one such way.

This algorithm can be improved by applying concept of GA. GA [6] is often used to solve the problem. GA uses various operators to implement its operation which are as follows.

- **Crossover**: Crossover [7-10] is the operator responsible for mating the chromosomes. Some popular crossover operators, in use, are uniform crossover, arithmetic, Heuristic, Intermediate, Scattered, OR, PMX, CX etc. [6].
- **Mutation**: It prevents the population of chromosomes from being too similar to each other. Thus, it prevents the solution to be caught in local optima. In spite of climbing several peaks simultaneously by observing various chromosomes, the possibility of attaining false peak cannot be denied. Mutation works by randomly altering the gene pattern of the chromosomes, thus helping in coming out of the local optima. The probability of mutation is often very low. The result of mutation can result in either a weaker individual or a stronger one but it can be surely established that it subtly changes genes of the chromosomes thereby coming out of the local optima.

- **Selection**: Selection operator selects those chromosomes which satisfy the requirements against the fitness function. It ensures that the chances of survival for the fittest individuals are more than the weaker ones. Since selection process decides the chromosomes selection for mating, it dramas a vital protagonist in exhibiting the presentation of the GA. A good selection leads to the faster convergence of the results. A number of selection approaches available remain Roulette wheel selection, Tournament selection, Sigma selection, Rank based selection, Random selection, etc.

4. CONCLUSION AND FUTURE SCOPE

In this work, the benefit of using a GA and its variants based technique for the performance improvement of BI is elaborated**.** In future detail study and implementation using GA will be performed with many objectives such energy, time and resources.

5. ACKNOWLEDGEMENTS

The assistance for this work is provided by the University Grant Commission and JNU New Delhi, India. Authors would like to thanks to Professors and anonymous reviewers for their valuable suggestions.

6. REFERENCES

[1] Chintapalli, R., Kumar A., Parayitam, L., "Bioinformaticsa quick tour", ICACT 2013, India, 21 September 2013.

[2] Fa Zhang, Xiang-Zhen Qiao, Zhi Yong Liu, "A parallel Smith waterman algorithm based on divide and conquer", Algorithms and architectures for parallel processing, China, 25 October 2002.

[3] Ozer B, Gezici G, Meydan C, Sezerman U, "Bioinformaticsusing structural properties", HIBIT 2010, Antalya,20 April 2010.

[4] Needleman, Saul B, Wunsch, Christian D, "A general method applicable to the search for similarities in the amino acid sequence of two proteins", Journal of Molecular Biology.

[5] Hamidi S, Naghibzadeh M, Sadri J, "Protein Bioinformaticsbased on secondary structure similarity", ICACCI 2013, Mysore, 25 august 2013.

[6] C. Kumar, S. Prakash, T. Kumar and D. P. Sahu, "Variant of genetic algorithm and its applications", International Journal of Artificial Intelligence and Neural Networks, vol. 4(4), pp. 8-12, 2014.

[7] D.P. Sahu, K. Singh, S. Prakash, "Maximizing Availability and Minimizing Markesan for Task Scheduling in Grid Computing using NSGA II", 2nd International conference for computer and communication technology, LNCS, Springer, 1-6, 2015.

[8] D.P. Sahu, K. Singh, S. Prakash, "Task Scheduling in Grid Computing using NSGA II", 2nd International conference fo networking, information and communication technology, SVCE, Vidyanagar, Bangalore-562157, India, IEEE, 1-3, 2015.

[9] D.P. Sahu, K. Singh, S. Prakash, "Resource Allocation and Provisioning in Computational Mobile Grid", International Journal of Applied Evolutionary Computation, vol. 6(2), pp. 57-83, 2015.