University of Kent

Division of Computing, Engineering and Mathematical Sciences
Department of Computing, University of Kent *in Canterbury*

Applications of Evolutionary Algorithms in Bioinformatics

By

Lukasz Ryszard Tomaszewski
lrgt2@kent.ac.uk

Computer Science (Artificial Intelligence) MSc
Module CO837: Natural Computation

LaTeX Word Count: 942

19th Nov – 10th Dec 2021

Contents

1 Introduction **1**

2 Evolutionary Algorithms in Bioinformatics **1**

3 Candidate Solution & the Fitness Function **2**

4 Advantages of EA's in Bioinformatics **2**

5 Disadvantages of EA's in Bioinformatics **3**

6 Conclusion **3**

7 References **4**

1 Introduction

Evolutionary algorithms are used to provide approximated solutions to a variety of problems by assigning a function to each individual component and rating those individual components by a set of chosen variables. In Bioinformatics, this takes the form of the genetic algorithm. This algorithm specifies individuals or chromosomes that correspond to binary strings to which an operator is applied and manipulates the string, from parent to children, this repeats until a set number of iterations have passed or a termination condition is met (shown in fig. 1) The simplest idea is that of single-point crossover in which two parents split their sequence to share with their children where as the mutation operation can occur in which a binary string will change values upon crossover. Specifically in Bioinformatics and Biology, "In the medical field GA-based solutions have been posed for a variety of problems including symptom and ailment classification, visualization as well as identification and diagnoses of diseases." [7], thus proving benefits in using evolutionary algorithms in this application.

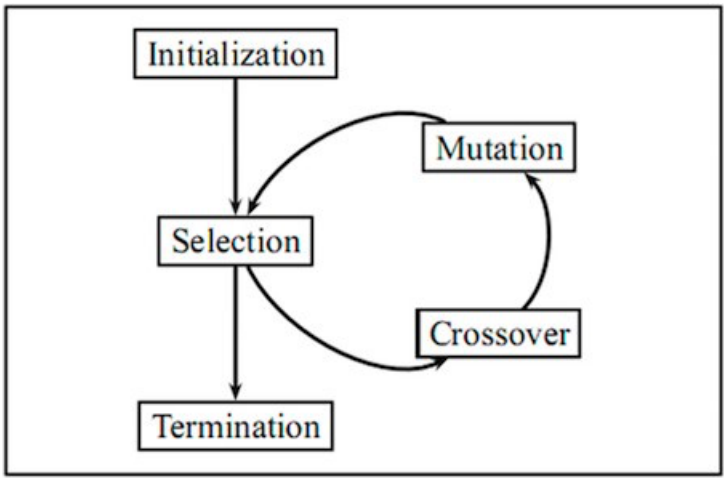

Figure 1: Figure showing the cycle of a genetic algorithm.

2 Evolutionary Algorithms in Bioinformatics

This essay explores the applications of evolutionary algorithms in Bioinformatics in RNA structure discovery, motif discovery [5] and multiple sequence alignment [6]. In the assistance in the discovery of the RNA structure, "an evolutionary algorithm is used to improve each structure based on both free and pseudo-free energies. Finally, a structure with minimum summation of free and pseudo-free energies is considered as the predicted RNA secondary structure." [3]. Through combining an evolutionary genetic algorithm and a random projection strategy to identify the (l-d)-Motifs through deduction of potential candidate Motifs to which can be applied to a fitness function [2], also "helps us distinguish real signal sub sequence patterns from background sequences"[8]. In the area of multiple sequence alignment, where the similarities between multiple sequences are analyzed [8], genetic algorithms are used to provide the accuracy of the protein alignment for the individual methods used [1], further accuracy is theoretically possible using a combination of Markov models and the partition function (particle swarm optimization) [9].

3 Candidate Solution & the Fitness Function

Exploring the assistance in Motif discovery, a candidate solution is derived to the initial population of the genetic algorithm so to improve the range of "candidate motifs" [2]. However in the experiment in [2], "Since the search space is large, randomly generated population will rarely come close to an optimal solution and there is little chance that a random population will converge to the optimal solution."[2].
Further exploration into the use of the candidate solution and fitness function in bioinformatics, in multiple sequence alignment, the population ("characterized by four parameters namely, population size, crossover rate, mutation rate, and elitist selection."[1]) is given a fitness function to produce a population that is based off its individual fitness. Processes such as selection, crossover (random exchange of genetic material between chromosomes) and mutation (random change of a certain bit in a chromosome) are used on the fittest populace, which births form parent to offspring the next generation of the process, which begins again until the termination condition is satisfied. Deeper analysis shows that multiple researchers have applied genetic algorithms, in the software 'SAGA' made by Notredame and Higgins [4], "SAGA uses two types of crossover: one-point (shown in fig. 2) and uniform (shown in fig. 3). It uses 20 different types of complex mutation operators other than two crossover operations to obtain optimum alignment."[1].

Figure 2: Figure showing the parent and children of single-point crossover (crossover point is randomly chosen)

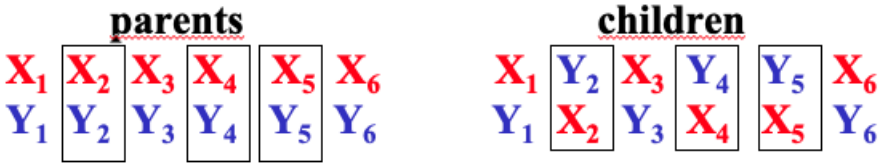

Figure 3: Figure showing the parent and children of uniform crossover (each gene value is crossed-over with probability p_c)

4 Advantages of EA's in Bioinformatics

Generic advantages of using the genetic algorithm are improved analysis of a variety of data where the fitness function defines only certain data succeeds into the next repetition in the form of a new data population. The algorithm not only can analyse data regardless of size but also analyse each individual in the total population in terms of a variety of variables. More specific to

Bioinformatics, this is important as the size of the data to too vast to analyse conventionally, by applying the algorithm and selecting the variables, the system will analyse the data and output the fittest population based off the variables and terminate accordingly.

5 Disadvantages of EA's in Bioinformatics

In Bioinformatics, there appears to be some controversy surrounding the use of evolutionary algorithms as implementing the algorithm, designing a fitness function that is specifically related to the data population and choosing select variables can be difficult, the rates of the crossover and mutation operators can also swing the result to favour a more bias output, more so as this method requires expensive computation compared to conventional methods. The fitness function can only guide the algorithm to achieve a result, a failure in the design would alter a bias result and could be theoretically why evolutionary algorithms are fully trusted yet. Generic disadvantages of using the genetic algorithm is that one must choose a fixed number of iterations by generations or choose a stopping criterion (which can be unknown), more so the choice of operation is in question with crossover wielding a high probability and mutation wielding low probability.

6 Conclusion

In conclusion, this essay highlights the structure and the use of the application of evolutionary algorithms (specifically the genetic algorithm) in Bioinformatics, while still controversially used proves to be highly useful in examining large amounts of data which is vastly apparent in Bioinformatics and analyzing the data via a variety of pre-chosen variables. Only used in a few areas within bioinformatics and biology, there is still more uses for these evolutionary algorithms as they have proven to provide solutions based off the strength of its own data that fit the chosen variables, it can be said that these algorithms have a very broad reach in every area of research and experimental testing. It could be said that with an expanse of computational bioinformatics, the use of evolutionary algorithms may rise as the disadvantages outlined in section 5 lack trepidation.

7 References

[1] Gautam, G. Biswanath, C. A review on multiple sequence alignment from the perspective of genetic algorithm. *Genomics*, 109(5-6):419–431, 2017. https://www.sciencedirect.com/science/article/pii/S0888754317300551?via%3Dihub.

[2] Vojislav, S. Hongwei, H., Zhenhua, Z. and Lifang, L. Optimizing genetic algorithm for motif discovery. *Mathematical and Computer Modelling*, 52(11-12):2011–2020, 2010. https://www.sciencedirect.com/science/article/pii/S0895717710002748.

[3] Ganjtabesh, M. Montaseri, S. and Zare-Mirakabad, F. Evolutionary algorithm for rna secondary structure prediction based on simulated shape data. *PLoS ONE*, 2016. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0166965.

[4] Holm, L. Notredame, C. and Higgins, D. Coffee: an objective function for multiple sequence alignments. *Bioinformatics*, 14(5):407–422, 1998. https://www.tcoffee.org/Publications/Ps_pdf/coffee.pdf.

[5] Piserchia, Z. Applications of genetic algorithms in bioinformatics. *University of California, Riverside*, 2018. https://escholarship.org/uc/item/9087560g.

[6] Radenbaugh, A. Applications of genetic algorithms in bioinformatics. *San Jose State University*, 2008. https://scholarworks.sjsu.edu/etd_theses/3495.

[7] Massop, B. Swerhun, M., Foley, J. and Mago, V. A summary of the prevalence of genetic algorithms in bioinformatics from 2015 onwards. *CoRR*, abs/2008.09017, 2020. https://arxiv.org/abs/2008.09017.

[8] K-C. Wong. Evolutionary algorithms: Concepts, designs, and applications in bioinformatics. *Nature-Inspired Computing: Concepts, Methodologies, Tools, and Applications*, 2015. https://www.researchgate.net/publication/280695580_Evolutionary_Algorithms_Concepts_Designs_and_Applications_in_Bioinformatics.

[9] Wang, N. Zhan, Q. and Jin, S. Probpfp: a multiple sequence alignment algorithm combining hidden markov model optimized by particle swarm optimization with partition function. *BMC Bioinformatics*, 20:573, 2019. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-3132-7#citeas.