ISSN:2229-6093

Rahul Chauhan et al ,Int.J.Computer Technology & Applications,Vol 3 (4), 1510-1515

A Review: Applying Genetic Algorithms for Motif Discovery

Rahul Chauhan

Azad Institute of Engineering & Technology, Lucknow,U.P, India
rahulchauhan0786@gmail.com

Dr. Pankaj Agarwal

Professor, Department of Computer Science & Engineering, IMS Engineering College, Ghaziabad,U.P, india
Pankaj7877@gmail.com

Abstract

This paper explores & reviews the use of genetic algorithms by various researchers as a solution to discover motifs in molecular sequences. This survey talks about the general GA based procedure for motif discovery & reviews the latest developments in DNA motif finding using Genetic algorithms. Although GA approach has not been applied extensively by researchers as compared to other computational methods for motif discovery, however in the recent past many researchers have explored the usefulness of GA for finding motifs in sequences. This paper is an attempt towards exploring the effectiveness of GA based approach for motif discovery.

1. Introduction

Large scale DNA sequencing of various organisms has resulted in the generation of huge amount of biological data and therefore there is always an increasing need to develop computational techniques that can help in finding useful information amongst all the data. Discovering motifs involves determining short meaningful sequences that may be repeated over many sequences in various species.

A DNA motif is defined as a nucleic acid sequence pattern that has some biological significance such as being DNA binding sites for a regulatory protein, i.e., a transcription factor. Normally, the pattern is fairly short (5 to 20 base-pairs (bp) long) and is known to recur in different genes or several times within a gene **[1]**.

Motifs are patterns in biological sequences which can indicate the presence of certain biological characteristics. In general, these could represent patterns in any kind of biological sequences such as DNA sequences, RNA sequences, protein sequences etc. Some of the features of motifs are:

- They are patterns of length 10 to 25 bases, and are repeated over many sequences
- They are statistically over-represented in regulatory regions
- They are small, have constant size, and are repeated very often.

Identification of motifs is becoming very important because they represent conserved sequences which can be biologically meaningful. Some of the areas where motif discovery can be useful include finding binding sites in amino acids, finding regulatory information within either DNA or RNA sequences, searching for splicing information, and protein domains. The motifs can represent patterns which activate or inhibit the transcription process and are responsible for regulating gene expression. Motif identification can be thought of as finding the best local multiple alignment for the sequences under consideration. **[2]**

The challenges present in motif identification include **[2]**:

- The motifs are never exactly the same as the actual conserved sequence. There is always a lot of sequence variability present with respect to a single motif.
- Motifs are very short signals as compared to the size of the DNA sequence under consideration.
- The regulatory sequences containing the motifs may sometimes be located very far away from coding regions that they regulate. This makes it difficult to determine the portion of the DNA sequence that should be analyzed.
- The regulatory sequences may, at times, be present on the opposite strand from the coding sequence they regulate.
- The motif discovery problem is an NP-Complete problem, so there is no

polynomial-time solution present for motif discovery.

An important task in this challenge is to identify regulatory elements, especially the binding sites in deoxyribonucleic acid (DNA) for transcription factors. These binding sites are short DNA segments that are called motifs. **[3]**. Computationally identifying transcription factor binding sites in the promoter regions of genes is an important problem in computational biology and has been under intensive research for a decade. To predict the binding site locations efficiently, many algorithms that incorporate either approximate or heuristic techniques have been developed. However, the prediction accuracy is not satisfactory and binding site prediction thus remains a challenging problem **[4]**.

Although not much work has been done on motif discovery using genetic algorithms, however recently GA based approaches has been one of the key areas of interest for researchers while discovering sequence motifs. This paper reviews some of the significant work done for motif discovery using genetic algorithms.

2. Motif-discovery using GA approach

A genetic algorithm is an approach based on evolutionary computing, which involves using the concept of evolution (reproduction) in order to evolve the solutions.

Genetic algorithms have been applied extensively in the field of computational biology, mainly for solving multiple sequence alignments (MSA) problems, and have proved quite successful. In the recent years, use of evolutionary approach for motif discovery is one of the interest areas for many researchers. This technique, although somewhat similar in concept to machine learning, is much more successful as machine-learning algorithms perform local searches and thus generates solutions that may be locally optimal, but are seldom globally optimal. Genetic algorithms, on the other hand, perform a global search on the search space without performing an exhaustive search. As a result of this, although the genetic algorithms do not always guarantee an optimal solution, they have a much better chance of finding an optimal solution. **[2]**

Beauty of genetic algorithms is that it allows the use of fitness functions for scoring the solutions. These fitness functions need not be constant for all problems and can use any relevant information to score the solutions including biological information, functional information, etc. Thus, they provide flexibility in evaluating the solutions. Also, the genetic algorithms provide a flexibility of deciding how to represent the problem instance. There are some studies where motifs have been represented in various different formats suitable for the genetic algorithm under consideration. These include representing motifs as regular expressions, position frequency matrices, etc.

There has not yet been as much study in the area of evolutionary approach for motif discovery, as there has been for the other approaches. Some early work focused on using steady-state algorithms for evolving the solution [2**].** Genetic algorithms have been used for evolving not only the content of the motif, but also the position using special crossover operators [5]. These solutions are scored based on pre-determined fitness functions which evaluate how fit each solution is, thereby aiding decisions regarding which solution needs to be maintained or discarded [2**].**

Two methods are used for generating new offspring i.e. crossover and mutations. The parents are selected using specified selection techniques. The most common technique for parent selection is the roulette-wheel method. In this method, the elements of the population are assigned slots on the roulette wheel, where the size of each slot is proportional to the fitness of the element. The roulette wheel is then spun randomly, and the slot (element) where it stops is chosen as the parent. This method allows the parents to be selected based on the fitness; where more fit solutions have a higher chance of getting selected.
The most basic crossover approach is where two parents are chosen from the population, based on some selection technique and then a crossover-site is selected at random, and the right-most strings are swapped, generating two children

Mutation works by randomly choosing some position for one of the elements, and changing the value at that position. If the encoding technique for the problem instance is bits, then mutation works by simply flipping the bit at the selected position.

Genetic algorithms have a large number of advantages. The most important advantage of this is that it is an extremely robust technique. The use of genetic algorithms is not limited to only a particular family of problems. They can be used for solving a wide variety of problems, and have been found to be extremely successful in those areas where the search

space is large. Also, although genetic algorithms do not guarantee the best solution every time, they are very useful in finding acceptable solutions faster than other methods. They have the ability to be hybridized with existing techniques of solving some problems to give better solutions **[2]**.

However, the general evolutionary approach has few disadvantages as well. Often, it has been observed that the final solution converges to a local maximum because of some instances of exceptionally high fitness candidates, which are not necessarily optimal. If this happens, then the algorithm is no longer able to find better solutions because crossovers lead to identical solutions with higher fitness, making it of little use. In most cases, mutation on its own, results in a large reduction in the speed of searching. Thus the overall process becomes very slow, resulting in the need for larger number of generations to be obtained **[2]**.

3. Review of GA methods for Motif Discovery

Genetic algorithms (GAs), like Gibbs sampling, apply a stochastic optimization technique, but operate on a population of candidate solutions to a specific problem domain. Specifically, the structures in the current population are evaluated for their effectiveness as solutions during each generation. Based on these evaluations, a new population of candidate structures is formed using operators like crossover and mutation. This process is iterated until an optimal solution is found or no improvement is achieved after a significant amount of evaluations **[6]**.

CompareProspector [http://compareprospector.stanford.edu] is a sequence motif-finding algorithm which extends Gibbs sampling by biasing the search in promoter regions conserved across species. CompareProspector outperformed many other computational motif-finding programs tested, demonstrating the power of comparative genomics-based biased sampling in eukaryotic regulatory element identification.

Liu et al. applied a GA to the motif discovery problem, and a program called FMGA was developed based on genetic algorithms (GAs) for finding potential motifs in the regions located from the -2000 bp upstream to +1000 bp downstream of the transcription start site **[7]**. They used the general GA framework and operators described in SAGA (sequence alignment by genetic algorithm) **[8]**. In FMGA, each individual is encoded as a set of candidate motif patterns generated randomly, one motif pattern per sequence. The fitness score for a single sequence is computed as the best matching percentage of all subsequences in that sequence, and the overall fitness score is the summation of individual fitness scores for all sequences. The mutation in GA is performed by using position weight matrices to reserve the completely conserved positions. The crossover is implemented with specially designed gap penalties to produce the optimal child pattern. This algorithm also uses a rearrangement method based on position weight matrices to avoid the presence of a very stable local minimum, which may make it quite difficult for the other operators to generate the optimal pattern. The authors reported that FMGA performs better in comparison to MEME and Gibbs sampler algorithms. Unfortunately the FMGA software is not publicly available for experimentation and comparison.

Shahar Michal et. al **[9]** used genetic programming to predict RNA consensus motifs based solely on the data set. Their system—dubbed GeRNAMo (Genetic programming of RNA Motifs)—predicts the most common motifs without sequence alignment and is capable of dealing with any motif size. Program only requires the maximum number of stems in the motif and, if prior knowledge is available, the user can specify other attributes of the motif (e.g., the range of the motif's minimum and maximum sizes), thereby increasing both sensitivity and speed. They described several experiments using either ferritin iron response element (IRE), signal recognition particle (SRP), or microRNA sequences showing that the most common motif is found repeatedly and that the system offers substantial advantages over previous methods.

Taichung **[10]** proposed a new approach FMGA for finding potential motifs in the regions located from the -2000 bp upstream to +1000 bp downstream of transcription start site (TSS). The approach is developed based on the genetic algorithm (GA). The mutation in the GA is performed by using position weight matrices to reserve the completely conserved positions. The crossover is implemented with special-designed gap penalties to produce the optimal child pattern. The work presents a rearrangement method based on position weight matrices to avoid the presence of a very stable local minimum, which may make it quite difficult for the other operators to generate the optimal pattern. As per the author his approach shows superior results by comparing with Multiple EM for Motif Elicitation (MEME) and

Gibbs Sampler, which are two popular algorithms for finding motifs.
A novel DNA motif discovery approach using a genetic algorithm is proposed by Xi Li et al **[11]** to explore the ways to improve the algorithm performance. Publicly available motif models such as Position Frequency Matrix (PFM) to initialize the population were taken into account. By considering both conservation and complexity of the DNA motifs, a novel fitness function is developed to better evaluate the motif models during the evolution process. A final model refinement process is also introduced for optimizing the motif models. According to the author experimental results demonstrated a comparable (superior) performance compared to known approaches.

Fatemeh Zare-Mirakabad et al **[12]** presented a genetic algorithm for the dyad motif finding problem. The genetic algorithm uses a multi-objective fitness function based on the sum of pairs, the number of matches, and the information content. The individuals required for the population pool in the genetic algorithm were optimized by Gibbs sampling method. Also, new crossover and mutation operators were designed. The algorithm is implemented and tested on the different types of real datasets. The results were compared with other well-known algorithms.

Genetic-enabled **EM** motif-**F**inding **A**lgorithm

(GEMFA)**:** The genetic algorithm (GA) is one of the prevalent intelligent computing methods able to tackle the local optimal issue. The GEMFA algorithm, originally presented in the *IEEE CIBCB'07* conference **[13]**, is an EM variant iteratively driven by GA. In each iteration, GA generates a population of new local alignments as start seeds, and then uses these seeds as input to the EM motif algorithms and produces refined local alignments. Obviously, GEMFA tightly integrates GA with EM, and cooperatively evolves a population of alignments towards a global optimal solution. *GEMFA is a de novo motif-finder designed to perform multiple local alignment of DNA or protein sequences*. In the web version, GEMFA adds the WEM motif-finder (very fast than regular EM, i.e. DEM) as an option.

Dongsheng Che et al [4] proposed a new genetic algorithm approach called MDGA to efficiently predict the binding sites for homologous genes. Experimental methods, such as DNase foot-printing **[14]** and gelshift assay **[15]** remain the most accurate and reliable identification methods for predicting the binding sites, but they are time-consuming and expensive. Alternative approaches that can efficiently predict the locations of binding sites with high accuracy are thus highly desirable due to the large amount of sequencing data that have been accumulated during the past decade.

Based on the generic framework of a genetic algorithm, the MDGA approach explores the search space of all possible starting locations of the binding site motifs in different target sequences with a population that undergoes evolution. Individuals in the population compete to participate in the crossovers and mutations occur with a certain probability. In MDGA, an individual is formed by a set of possible starting locations of the binding sites on different homologous sequences. The fitness value for an individual is evaluated by summing up the information content for each column in the alignment of its binding sites. The fitness function penalizes the individuals that have lower similarity in the alignment of their binding sites and thus eventually selects individuals with highly conserved binding sites. As per the authors it is capable of achieving a higher level of prediction accuracy than approaches based on the Gibbs sampling algorithm. Moreover, experiments also showed that the computation time needed for MDGA does not explicitly depend on the sequence length and may remain unchanged even when the sequence becomes very long

In her masters project Medha Pradhan **[2]** presented an evolutionary approach for motif discovery. The population is clustered during every generation of the algorithm and then evolved locally within the clusters to allow the search space to maintain solution diversity. The motifs considered were those indicating presence of promoter elements. In this work motifs that represent regulatory elements in biological sequences were identified. The input to the algorithm consists of two sets of sequences. The first is a set of promoter sequences which has the likelihood of containing regulatory elements. The second is a set of background sequences representing some random portion of the DNA which may not include any promoter sequences. The algorithm processes these input sequences to determine regions in the first set which could classify as motifs. This is done by finding subsequences that are over-represented in the first set of input sequences, as compared to the background sequences. These over-represented portions indicated the presence of motifs.

Very few genetic algorithm based approaches make use of such background sequence sets to determine motifs. In most cases, a pre-determined background model is used. The use of such pre-defined background models affects the efficiency of the system, since it does not provide the flexibility of adapting to the input data.

The basic approach followed in this project for motif-discovery can be summarized as follows:
1. Determine a representation for the motifs.
2. Evaluate the motifs using a well-defined fitness function.
3. Cluster the population based on some clustering metric.
4. Run the genetic algorithm to repeat the above steps until a motif close to the consensus sequence is identified.

The representation for motifs being used in this project is a combination of both position frequency matrices (PFM) and position weight matrices (PWM). The project uses a population clustering technique. It involves using a clustering algorithm to divide the population. The clusters are then subject to the mating process in the genetic algorithm, ensuring that mating occurs within the clusters. Such intra-cluster mating allows solution diversity to be maintained during evolution. After this, the new population is again subject to clustering in the new iteration. This provides the advantage of allowing the solution to move from one cluster to another (based of fitness), thus promoting some degree of inter-cluster mating, providing another means of maintaining solution diversity. It uses the Leader algorithm as a clustering method.

The genetic algorithm used here starts with the initialization phase. This phase is then followed by an iterative process for clustering the population, mating the selected parents (using elitism, mutation and crossover), and evaluating the new offspring generated. Mating is only allowed within clusters, which helps to maintain the solution diversity. The number of offspring generated for each cluster is proportional to the mean fitness of all the solutions in the cluster. Mutation and crossover are done with a mutation to crossover ratio of 7:3. After the reproduction phase, the entire process is again repeated.

Four different data sets have been used, of which two are synthetic data sets in which known motifs have been introduced. In the first set, only one motif is embedded. In the second set, multiple motifs have been embedded. The remaining two data sets are a muscle-specific data and a liver-specific data. The fourth data set (synthetic data set with embedded multiple motifs) was created by inserting motifs at various locations, and checking if the algorithm is able to detect the motifs.

The foreground sequences were taken from the Eukaryotic Promoter Database (EPD). The motifs are then randomly inserted into more than half of the selected sequences. Two different types of motifs (HFH-1 and HLF) were used in two different data sets. The set of background sequences is comprised of randomly selected sequences from EPD.

As per the author, it was observed that the fitness of the evolved motif increased as the number of generations increased. The experiment with 50 generations did not yield a very good solution. The experiment with 100 generations yielded a much better solution. The percentage of successful runs more than doubled with the increase in number of generations. In case of single motifs, the algorithm performs reasonably well. In case of sequences having multiple motifs, the algorithm is most successful in identifying the most strongly conserved motif

Motif discovery in biological sequence analysis remains a challenge in computational biology. The Expectation Maximization (EM) algorithm is one of the most popular methods used in motif discovery. However, EM heavily depends on initialization and suffers from local optima.

4. Conclusion

Despite considerable efforts to date, DNA motif finding remains a complex challenge for biologists and computer scientists. Researchers have taken many different approaches in developing motif discovery tools and the progress made in this area of research is very encouraging. Performance comparison of different motif finding tools and identification of the best tools have proven to be a difficult task because tools are designed based on algorithms and motif models that are diverse and complex and our incomplete understanding of the biology of regulatory mechanism does not always provide adequate evaluation of underlying algorithms over motif models.

5. References

1. Rombauts S, Dehais P, Van Montagu M, Rouze P. PlantCARE, a plant cis acting regulatory element database. Nucleic Acids Res. 1999;27:295–296.
2. Pradhan, Medha, "Motif Discovery in Biological Sequences" (2008). Master's Projects. Paper 106. http://scholarworks.sjsu.edu/etd_projects/106
3. Das and Dai; licensee BioMed Central Ltd (2007).
4. MDGA: Motif Discovery Using A Genetic Algorithm by Dongsheng Che, Yinglei Song and Khaled Rasheed; Department of Computer Science University of Georgia Athens, GA 30602, USA, 2005
5. Lones, M.A., Tyrrell. "The Evolutionary Computation Approach To Motif Discovery in Biological Sequences." Proc. Genetic and Evolutionary Computation Conf. (GECCO) Workshop Program, Jun 2005, 1-11.
6. Eiben, A.E. and Smith, J.E. Introduction to Evolutionary Computing. Springer-Verlag, New York. 2003.
7. Liu, F.F.M., Tsai, J.J.P., Chen, R.M., Chen, S.N. and Shih, S.H. FMGA: finding motifs by genetic algorithm. *IEEE Fourth Symposium on Bioinformatics and Bioengineering (BIBE 2004)*, May 2004, 459-466
8. Notredame, C. and Higgins, D.G. SAGA: Sequence alignment by genetic algorithm. *Nucleic Acids Res.* 24, 8 (Apr. 1996), 1515–1524
9. Shahar Michal et. Al ; finding a common motif of rna sequences using genetic programming: the gernamo system; IEEE/ACM transactions on computational biology and bioinformatics, vol. 4, no. 4, october-december 2007
10. Taichung;FMGA: Finding Motifs by Genetic Algorithm; Fourth IEEE Symposium on Bioinformatics and Bioengineering (BIBE'04); Taiwan (2004)
11. Xi Li et. Al.; An improved genetic algorithm for DNA motif discovery with public domain information;published in ICONIP'08 Proceedings of the 15th international conference on Advances in neuro-information processing - Volume Part I Pages 521-528 Springer-Verlag Berlin, Heidelberg ©2009.
12. Genetic algorithm for dyad pattern finding in DNA sequences. by Fatemeh Zare-Mirakabad, Hayedeh Ahrabian, Mehdi Sadeghi, Somaieh Hashemifar, Abbas Nowzari-Dalini, Bahram Goliaei Genes genetic systems (2009) Volume: 84, Issue: 1, Pages: 81-93
13. *A genetic-based EM motif-finding algorithm for biological sequence analysis.Proceedings of 2007 IEEE Symposium on Computational Intelligence in Bioinformatics & Computational Biology, pp. 275-282 (2007).*
14. Galas, D.J. and Schmitz, A. A DNA footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res*., 5, 9 (Sep. 1978), 3157–3170.
15. Garner, M. M., and A. Revzin. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. *Nucleic Acids Res*. 9, 13 (July, 1981), 3047-3060.