An Analysis Of Bayesian And Nonbayesian Functions For Motif Detection Using Genetic Algorithm

V.Bhuvaneswari .A[1], Nusrath [2]

ABSTRACT

A motif, in the context of biological sequence analysis, is a consensus pattern of DNA bases or amino acids which accurately captures a conserved feature common to a group of DNA or protein sequences. Finding motif-patterns of conserved residues-within nucleotide and protein sequence is a key part of understanding function and regulation within biological system. Computational motif discovery has been used with some success in simple organisms like yeast. However, when moves to higher organisms with more complex genomes, more sensitive methods are needed. Genetic Algorithm is an efficient method for detecting motifs, since it has greater freedom of movement between different configurations than simpler algorithms. This paper analyses the genetic algorithm method for the detection of motifs by using Bayesian and NonBayesian functions as fitness function and compares it with the other existing tools.

Keywords : DNA sequence, motif, genetic algorithm.

1. INTRODUCTION

The complete information that defines the characteristics of living cells within an organism is encoded in the form of a moderately simple molecule, deoxyribonucleic acid, or DNA. The building blocks of DNA consists of four nucleotides, abbreviated by their attached organic bases as A, C, G and T. A-T and C-G are complementary bases between which hydrogen bonds can form. A DNA molecule consists of two long chains of nucleotides that are complementary to each other and joined by hydrogen bonds twisted into a double helix. This structure gives rise to the term "base pair" when describing a DNA sequence. The specific ordering of these nucleotides, the "genetic code," is the means by which information is stored that completely defines all functions within a cell. With the recent development of high-throughput sequencing technology, the National Institutes of Health genetic sequence database, Gen-Bank, has sustained an exponential growth rate since 1982.[14]

The central dogma of molecular biology dictates that certain segments of the DNA (i.e., genes) are transcribed into another molecule, RNA, which serves as a transient template to make the basic building blocks of cellular life, proteins. Although all the cells in the same organism possess exactly the same DNA sequences (i.e., genetic information), they display different physiological characteristics in different tissues, developmental stages and environmental conditions. This "differentiation" is caused by the differences among the collections of proteins that are synthesized in different cells or at different cell states. If a protein is being synthesized at a certain state, its coding DNA (called a gene) is termed as "active" or "expressed." Thus, a cell in a particular physiological state can be roughly viewed as a mechanical system where each different gene is switched either on (active) or off (inactive).

[1] Lecturer, School of Computer Science and Engineering, Bharathiar University. e-mail : bhuvanes.v@yahoo.com
[2] MPhil Research Scholar, School of Computer Science and Engineering, Bharathiar University.
e-mail : nusrathaa@gmail.com

In many organisms, the DNA that codes for proteins (genes) is only a small portion of the total genomic DNA. For example, genes make up only about 1.5% of the human genome (International Human Genome Sequencing Consortium, 2001). The noncoding components of DNA, which were initially considered as"Junk" sequences; actually contain the control mechanisms for activating and deactivating the genes, and thus the synthesis and nonsynthesis of proteins. Most of the control sequences for a gene lie in the upstream regulatory region, which is the region of a few thousand base pairs directly before the gene [also called the transcription regulatory region (TRR) or the promoter. Transcribing or activating a gene requires not only the DNA sequence in the TRR, but also many proteins called transcription factors (TFs). When these TFs are present, they bind to specific DNA patterns in the TRR of genes and either induce or repress the transcription of these genes by recruiting other necessary proteins.[7]

One transcription factor can bind to many different upstream regions, thus regulating the transcription of many genes. The binding sites of the same transcription factor show a significant sequence conservation, which is often summarized as a short (5-20 bases long) common pattern called a transcription factor binding motif (TFBM) or binding consensus, although some variability is tolerated.

In prokaryotes (lower organisms without nuclei), there are fewer TFs, their motifs tend to be relatively long and the strength of regulation for a particular gene often depends on how closely a particular site matches the consensus for the motif. The more mismatches to the consensus in a binding site, the less often the TF will bind and therefore the less control it will exert on the target gene. The variability between sites is sometimes crucial to the regulatory process, since TF binding sites that are perfect matches to the optimal pattern would bind the TF too tightly, preventing the subsequent steps of transcription.

In eukaryotes (higher organisms with nuclei), many more transcription factors are involved in the regulation of a gene, and their binding motifs tend to be shorter. Eukaryotic upstream regions usually contain regulatory modules, a collection of adjacent binding sites (sometimes multiple binding sites) of several transcription factors. Transcription regulation not only relies on the combination of the TFs involved, but also on the number of site copies in the upstream regions[17].

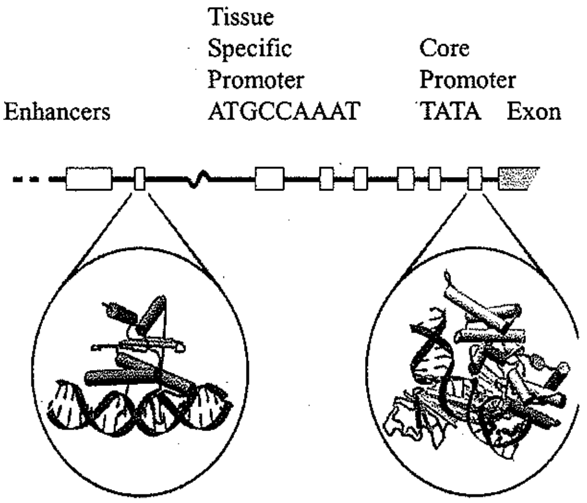


Figure 1: Organization of Eukaryotic DNA Sequence.

Understanding the regulatory networks of higher organisms is one of the main challenges of functional genomics. Gene expression is regulated by transcription factors (TF) binding to specific transcription factor binding sites (TFBS) in regulatory regions associated with genes or gene clusters. Identification of regulatory regions and binding sites is a prerequisite for understanding gene regulation, and as experimental identification and verification of such elements is challenging, much effort has been put into the development of computational approaches. Good computational methods can potentially provide high-quality prediction of binding sites and reduce the time needed for experimental verification. However, the computational approach has turned out to be at least as challenging as the experimental one, and a very large number of different methods have been developed.

Computational discovery of motifs is mainly possible because they occur several times in the same genome, and because they may be evolutionary conserved. However, this apparently simple approach is complicated by the fact that most binding site motifs are short, and they may also show some sequence variation without loss of function. Therefore most motifs are found as random hits throughout the genome,.and it is a challenging problem to distinguish between these false positive hits and true binding sites.

Genetic algorithms are mainly used for its ability to perform adaptive, powerful and robust searches. As an evolutionary computation technique, they operate in parallel over a population of candidate solutions, allowing a simultaneous exploration of different regions of the search space in the solution domain.

The paper is organized as follows. Section 1 gives an introduction to motif and the need for motif prediction using GA. Section 2 literature review of motif discovery algorithms are discussed. Section 3 describes the methodology for motif prediction using genetic algorithm. Section 4 analyses the experimental results of the resultant motif using Genetic Algorithm and with other existing methods. In Section 5 the conclusion of the work is inferred.

1.1 Why Motifs ?

A motif, in the context of biological sequence analysis, is a consensus pattern of DNA bases or amino acids which accurately captures a conserved feature common to a group of DNA or protein sequences. DNA motifs are sometimes termed signals: examples are regulatory sequences, scaffold attachment sites, and messenger RNA splice sites. Examples of protein motifs, which are also known as fingerprints, include enzyme active sites, structural domains, and cellular localization tags. Motif discovery is the act of identifying and characterizing motifs, and underlies a number of important biomedical activities. For example: the identification of regulatory signals has applications for gene finding in sequenced genomes, understanding of regulatory networks, and the design of drugs for regulating specific genes; and protein motifs are routinely used to identify the function of newly-sequenced genes and to understand the basis of a protein's cellular function.[9]

A nucleotide sequence is a string of letters (A,C,G and T) representing the sequence of nucleotide bases(Adenine, Cytosine, Guanine and Thymine) present within DNA molecules. Fig 2 gives an example for motif from three DNA sequences.

ACT***CAAGTC***TTATCACCC
GCGAAATTCG***CAAGTC***TT
C***CAAGTC***GTCGCTATATA

Figure 2 : CAAGTC is an Example for a Motif

2. Related Works On Motif Discovery

One of the early origins of DNA motif discovery is the computer program written by Korn et al. (1977) [4] that was able to discover sequence similarities in regions immediately upstream of TSS. Both mismatches and flexible gaps are accounted for, but using only pair wise comparisons. This approach was further developed by Queen et al (1982) [12] by comparing multiple sequences simultaneously. In this work , the exact requirements of a motif was also defined clearly, with quorum constraints on sequence support, max number of mismatches in occurrences, and max distances between occurrence positions in the different sequences. In the same year, Stormo et al.(1982) [16] introduced a Perceptron algorithm that calculated the sum of independent weighted match scores for each position of a motif aligned with a sequence. Similar to this, Staden (1984) [15] introduced a position weight matrix with weights corresponding to log-frequencies of nucleotides in aligned motif occurrences.

Jason D. Hughes, Preston W. Estep, Saeed Tavazoie and George M. Church (2000) [3], proposed AlignACE, a stochastic motif discovery algorithm based on the widely adopted Gibbs Sampling method . Compared with the original Gibbs Sampling method, it adds the following major features: both strands of sequences are searched; near-optimum sampling is improved; an iterative masking approach is used to search multiple motifs.

X Lui, D. L. Brutlag and J. S. Lui (2001) [6], proposed BioProspector, a Gibbs Sampling algorithm. Compared with the Lawrence version, it added a Markov model estimated from all promoter sequences in the genome to model adjacent nucleotide dependency. It has 15 parameters. The background frequency model is generated using the whole E.coli genome, and the third-order Markov model is used unless otherwise specified.

Giulio Pavesi, Paolo Mereghetti, Giancario Mauri and Graziano Pesole (2004) [11], proposed Weeder, a consensus based method that enumerates exhaustively all the motifs upto a maximum length and collects their occurrences (with substitution) from input sequences. Each motif is evaluated according to the number of sequences in which it appears and how well conserved it in each sequence, with respect to expected values derived from the oligo frequency analysis of all the available upstream sequences of the same organism.

Shane T. Jensen, X. Shirley Lui, Qing Zhou and Jun S. Lui (2004) [14], proposed a set of Bayesian models useful for developing motif finding tools and generalization of these models that allow for unknown motif width w and unknown motif abundance ratio P0. Bayesian models provided insight to the similarities between the full Bayesian model based approaches and some NonBayesian methods such as CONSENSUS.

Shane T. Jensen and Jun S. Liu (2004) [13], proposed BioOptimizer which works based on a full Bayesian model that can handle unknown site abundance, unknown motif width and two block motifs with variable length gaps.

Zhi Wei and Shane T. Jensen (2006)[18] introduced GAME, which utilizes a genetic algorithm to find optimal motifs in DNA sequences. GAME evolves motifs with high fitness froma population of randomly generated starting motifs, which eliminate the reliance on additional motif-finding programs. In addition to using standard genetic operations, GAME also incorporates two additional operators that are specific to the motif discovery problem.

Leping Li, Yu Liang and Robert L. Bass (2007) [5] proposed GAPWM which derives high quality PWMs for genome wide identification of transcription factor binding sites. Starting from an existing PWM, a set of ChIP sequences, and a set of background sequences, GAPWM derives an improved PWM through genetic algorithm that maximizes the area under the receiver operating characteristic (ROC) curve.

More than a hundred methods have been proposed for motif discovery in recent years, representing a large variation with respect to both algorithmic approaches as well as the underlying models of regulatory regions.

3. Genetic Algorithm

Genetic algorithms are a part of evolutionary computing, which is a rapidly growing area of artificial intelligence. It was inspired by Darwin's theory of evolution. Genetic Algorithms (GAs) are computer programs which create an environment where populations of possible solutions can compete and only the fittest survive.[2]

Genetic Algorithms works based on the assumption that simulating an evolutionary process in a population of potential solutions evolves better solutions. Biological terms are conveniently used to describe this process:

- The chromosomes represent the potential solutions. Every chromosome is typically composed of several genes, the solution parameters.
- A set of chromosomes forms a population. Successive populations are referred to as generations.
- To create new chromosomes (offsprings), two kinds of operators are typically used: Crossovers are used to exchange genes between two chromosomes, while mutations change one or more genes in a single chromosome.
- Based on the principle of survival-of-the-fittest, chromosomes with a good performance (according to an applied fitness function) are more likely to be selected to produce offsprings for the next generation.

To solve some problem, the solution should be the best among others. The space of all feasible solutions is called search space (also state space). Each point in the search space represents one feasible solution. Each feasible solution can be "marked" by its value or fitness for the problem. The solution is one point (or more) among feasible solutions - that is one point in the search space.

3.1 Eelements of Genetic Algorithm

3.1.1 Encoding Chromosome

The chromosome should contain information about the solution it represents. Binary encoding, Permutation encoding, Value encoding, Tree Encoding are the different methods used for encoding chromosomes[8].

3.1.2 Reproduction

Reproduction is a process in which individual strings (chromosomes) are copied according to their fitness. There are many methods for selecting the best chromosomes: roulette wheel selection, Boltzman selection, tournament selection, rank selection, and steady state selection.

3.1.3 Crossover Operators

Crossover operators are responsible for creating one or more new chromosomes (offsprings) by combining the attributes of existing chromosomes (parents). Single Point crossover, Two Point crossover, Uniform crossover, Specific crossover, Arithmetic crossover are the different methods used.

3.1.4 Mutation Operators

Mutation operators are used to maintain genetic diversity from one generation of a population of chromosomes to the next. The purpose of mutation in GAs is to allow the algorithm to avoid local minima by preventing the population of chromosomes from becoming too similar to each other, thus slowing or even stopping evolution. This reasoning also explains the fact that most GA systems avoid only taking the fittest of the population in generating the next but rather a random (or semi-random) selection with a weighting toward those that are fitter[8].

3.2 Methodology for Motif Prediction using GA

Genetic algorithms have greater freedom of movement between different configurations than simpler algorithms. It makes them a valuable tool for the discovery of optimal motifs.

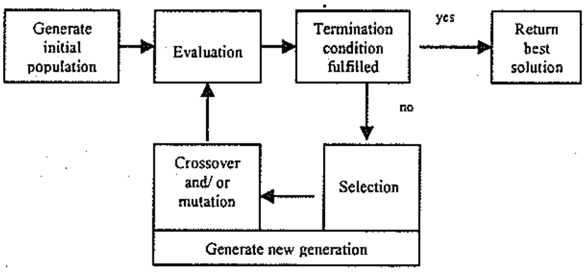

Generate initial population
Evaluation
Termination condition fulfilled
yes
Return best solution
no
Crossover and/ or mutation
Selection
Generate new generation

Figure 3 . System Architecture

3.2.1 Search Space

Search space consists of m upstream sequences, each of length l, where Sij is the nucleotide in position j of sequence i. A motif is a consensus sequence repeating more times in search space and having high fitness value.

3.2.2 Fitness Function

The genetic algorithm uses Bayesian and NonBayesian as scoring functions to calculate fitness.

Bayesian scoring function

The Bayesian scoring function is as follows:[18]

$$\psi_{ent}(A)=|A|\ .(\log(\hat{p}/(1-\hat{p}))-1+\hat{\theta}_{jk}\log(\hat{\theta}_{jk}/\theta_{ok}))$$

where |A| is the number of predicted sites and $\hat{p}=|A|/L$ is the estimated motif abundance out of $L=\sum_j l_i - w + 1$ possible site locations in A. The term $\hat{\theta}_{jk}\log(\hat{\theta}_{jk}/\theta_{ok})$ is the relative entropy between the estimated motif matrix frequencies $\hat{\theta}_{jk}$ and background frequencies θ_{ok}.

NonBayesian scoring function

The NonBayesian scoring function is as follows:[14]

$$\psi_{md}(A)=(\log|A|)\ /\ w\sum_j\sum_k\hat{\theta}_{jk}\log(\hat{\theta}_{jk}\ /\ \theta_{ok})$$

where |A| is the number of predicted sites, w is the width of motif $\hat{\theta}_{jk}$ jk is the frequency of nucleotide k in column j of the motif and θ_{ok} is the frequency of nucleotide k in the background

- In this work the alignment matrix is found as the first step to find scoring function. Each element N_{jk} is the occurrence of base k at position j of all the aligned subsequences in a chromosome.
- From alignment matrix, frequency matrix is found using formula $f_{jk} = N_{jk}/N$, where N is the number of individual subsequences in a chromosome.

A	C	G	T
0.00	0.33	0.58	0.08
0.17	0.08	0.67	0.08
0.00	0.00	1.00	0.00
1.00	0.00	0.00	0.00
0.00	0.00	0.00	1.00
0.00	0.00	0.00	1.00
1.00	0.00	0.00	0.00
0.50	0.08	0.17	0.25

Alignment Matrix

Pos	A	C	G	T
1	0	4	7	1
2	2	1	8	1
3	0	0	12	0
4	12	0	0	0
5	0	0	0	12
6	0	0	0	12
7	12	0	0	0
8	6	1	2	3

Frequency Matrix

3.2.3 Genetic Operators

Each individual motif configuration was represented by a vector $A(a_i, \ldots, a_m)$ where $0 \leq a_i \leq (l_i\_ w + 1)$ (where w is the width of motif).

This algorithm applies a bit inversion mutation to convert $A(a_1, \ldots, a_i, \ldots, a_m)$ into $A(a1, \ldots, \acute{a}_i, \ldots, a_m)$ with a certain mutation probability r. The mutation probability used is 0.001. The effect of mutation move is given below.

$A(a_1, a_2, a_3, a_4, a_5)$		$A(a_1, a_2, a_3, a_4, a_5)$
GATTACA		GATTACA
GATTAGG		GATTACA
GATTACA	Mutation Move	GATTACA
GATTACA		GATTACA
GATTACA		GATTACA

Figure 4: Examples of Mutation Operations

A standard one-point crossover move is also used that allows individual configurations to share and exchange alignment information with each other. Our population of N individual configurations is randomly grouped into N/2 pairs of con- figurations. For each pair of individual configurations $A(a_1, \ldots, a_m)$ and $B(b_1, \ldots, b_m)$, a crossover point c is randomly generated that gives rise to two children configurations $\acute{A}(a_1, \ldots, a_c, b_{c+1}, \ldots, b_m)$ and $B'(b_1, \ldots, b_c, a_{c+1}, \ldots, a_m)$. The effects of crossover are given below.

$A(a_1, \ldots, a_5)$	$B(b_1, \ldots, b_5)$
GATTACA	GAGGACA
GATTAGA	GAGGACA
GATTACA	GAGGAGA
GAGGACA	GATTAGA
GAGGACA	GATTACA

Crossover Move

GATTACA	GAGGACA
GATTAGA	GAGGACA
GATTACA	GAGGAGA
GATTAGA	GAGGACA
GATTACA	GAGGACA
$A(a_1, a_2, a_3, b_4, b_5)$	$B(b_1, b_2, b_3, a_4, a_5)$

Figure 5 . Examples of Crossover Operations

Table 1 : GA Parameters

Population Size	100
Max.Generation	2000
Selection	Rank Selection
Crossover	One Point Crossover
Mutation	Bit inversion
Mutation Rate	0.001
Fitness Function	**Bayesian Scoring Function** $\psi(A) = \lvert A\rvert . (\log (\hat{P}/(1-\hat{P})) -1 + \prod\prod\hat{\theta}_{jk} \log(\hat{\theta}_{jk}/\theta_{ok}))$ **NonBayesian Scoring Function** $\psi(A) = (\log\lvert A\rvert) /w \sum_j \sum_k \hat{\theta}_{jk} \log(\hat{\theta}_{jk}/\theta_{ok})$

4. Experimental Results

4.1 Dataset

In order to evaluate the performance of the Genetic Algorithm, it is tested using nucleotide sequence dataset. This training dataset consists of two types of nucleotide sequences, short dataset and long dataset. The short dataset consists of 18 ecoli sequences that contain cyclic AMP receptor protein (CRP) binding sites. Each sequence is 105 base pairs long and each contains at least one motif that has been experimentally determined via foot printing method. The cyclic AMP receptor protein (CRP) functions as a transcription factor in Escherichia coli. This dataset has been previously analyzed by Lawrence and Reilly using an EM algorithm and by Lui using a Gibbs Sampler. The long dataset consists of 4 Homo sapiens sequences. It includes Homo sapiens oct6, homo sapiens elf3, homo sapiens oct4, homo sapiens dusp3. Each sequence is 1020 base pairs long and contains transcription binding site motifs.

4.2 Comparison of Bayesian and Nonbayesian functions

Genetic Algorithm depends on encoding of chromosomes, fitness function evaluation, mutation rate and crossover.

In this work, the encoding used is value encoding, fitness function is Bayesian or NonBayesian scoring functions and crossover operation is one point crossover.

The genetic algorithm was executed for 2000 generation using Bayesian function. The generation Vs fitness graph shows convergence after 50 generations.

The graph obtained by genetic algorithm using Bayesian scoring function is given in figure 4 below.

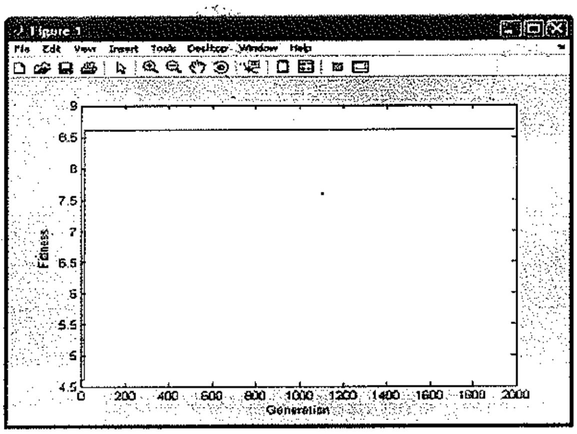


Figure 4. Generation Vs Fitness Graph (Bayesian)

Also the genetic algorithm was executed for 2000 generation using NonBayesian function. The generation Vs fitness graph shows convergence after 300 generations.

The graph obtained by genetic algorithm using NonBayesian scoring function is given in figure5 below.

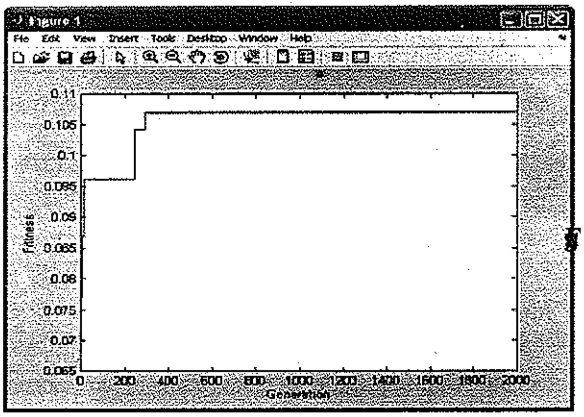


Figure 5. Generation Vs FitnessGraph -NonBayesian

The genetic algorithm moves around the space of possible motifs and finds best among them. The proposed algorithm was executed in short dataset for various motif lengths by using Bayesian and NonBayesian scoring functions. The table 2 & 3 given below shows the resultants motif using GA. The results are also compared with the existing methods like Gibbs Sampling and Enumeration method.

Table 2: Resultant Motif for E.Coli Sequence

Motif Length	Gibbs Sampling (AlignACE)	Enumeration (Weeder)	Genetic Algorithm Bayesian	Genetic Algorithm NonBayesian
6	nil	ccgatg	ccgatg	ccgatg
8	nil	tgcaagtg	tgcaagtg	tgcaagtg
10	nil	cgcccagctg	cgcccagctg	nil
11	gcggctgggcc	nil	gcggctgggcc	nil
12	nil	ggatcaacgtgg	ggatcaacgtg	nil
20	gcaggggccgcg-taccgcga	nil	nil	nil

Table 3: Resultant Motif for Homosapien Sequence

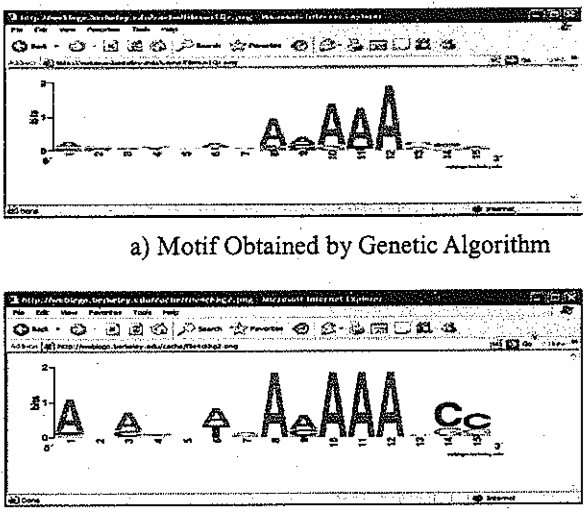
Motif Length	Gibbs Sampling (AlignACE)	Enumeration (Weeder)	Genetic Algorithm (Bayesian)	Genetic Algorithm NonBayesian
6	nil	ttgtga	ttgtga	ttgtga
8	nil	aattgtga	aattgtga	aattgtga
10	cacatcacaa	gtcacacttt	gtcacacttt cacatcacaa	gtcacacttt cacatcacaa
11	aaaatgagag	nil	aaaatgagacg	nil
12	aaacttgtaagt	aaacttgtaagt	aaacttgtaagt	nil
15	agatcacaca-aagcg	nil	agatcacaca-aagcg	nil

4.3 Sequence Logo

Sequence Logo is used to visualize the appearance of the motifs. The sequence logo for motifs of length 15 for Ecoli and Homosapien sequence obtained using genetic algorithm and AlignACE (Gibbs Sampling) tool is given below.

a) Motifs in Ecoli sequences

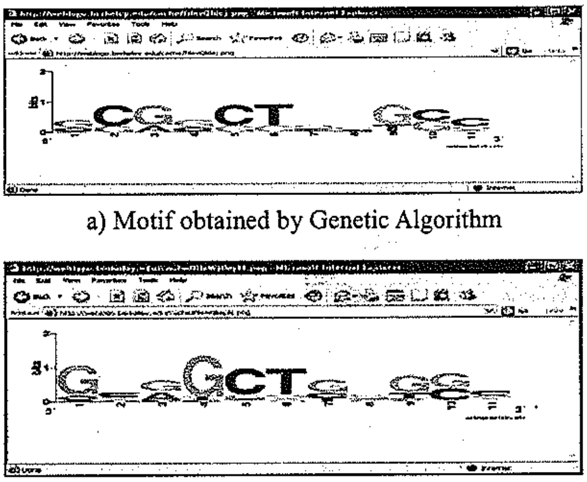
b) Motifs in Homo sapiens sequences



a) Motif Obtained by Genetic Algorithm

b) Motif obtained by Align ACE

Figure 6: Sequence Logo of Ecoli Motif



a) Motif obtained by Genetic Algorithm

b) Motif Obtained by Align ACE

Figure 7: Sequence Logo of Homo Sapiens Motif

5. Discussion

Motif discovery is an important problem in computational biology since the binding of transcription factors to upstream region motifs is crucial to the mechanism of gene regulation. The proposed genetic algorithm when executed both in Ecoli dataset and in Homo sapiens dataset detected motif of various lengths for Bayesian scoring function and NonBayesian scoring functions. The results are concluded as follows.

- The resultant motifs given in Fig.6(a) shows that both Bayesian function and NonBayesian functions detects the same motifs in Ecoli dataset.

  Also in Ecoli dataset Fig 6(a), the genetic algorithm using Bayesian scoring function detects motifs for all motif lengths as that of AlignACE and Weeder. But the genetic algorithm using NonByesian scoring function detects only very short motifs.

- The resultant motifs given in Fig.6(b) shows that both Bayesian function and NonBayesian functions detects the same motifs in Homo sapiens dataset.

- In Homo sapiens dataset, the Bayesian scoring function gave motifs till motif length 12. But NonByesian scoring function detected only very short motifs.

The comparison between Bayesian and NonBayesian scoring functions shows that Bayesian scoring function is good for motif detection, since NonBayesian scoring function detects only very short motifs.

6. Conclusion

The field of motif discovery brings together researchers from several disciplines, in particular from biology, statistics and informatics. Additionally, research in the field is fairly recent and moving at a fast pace. This has resulted in a broad range of computational methods that are described with different vocabulary and different focus, making it difficult to spot similarities as well as differences between methods.

This work using geneticic algorithm detected optimal motifs by moving around the solution space by applying an evolutionary process to an entire population for possible solutions. As GA runs parallel the motifs are detected in a single run than the other traditional methods. The future work is to compare the performance of GA with some hybrid approach and testing the same for protein datasets.

References

[1] Gary D. Stormo, *"DNA Binding Sites: Representation and Discovery"*, Bioinformatics journal, Vol. 16 No.1, 2000.

[2] Goldberg. D,*"Genetic Algorithms in Search, Optimization and Learning "* , Addison-Wesley, Reading, Massachusetts, 1989.

[3] Hughes. J. D, Estep. P. W. Tavazoie. S and Church. G. M, *"Computational Identification of Cis-regulatory Elements Associated with Groups of Functionally Related Genes in Saccharomyces* Cerevisiae, J. Mol. Biol. Vol. 296, PP. 1205-1214, 2000.

[4] Korn L.J, Queen CL, Wegman MN, *"Computer Analysis of Nucleic Acid Regulatory Sequences"*, Proc Natl Acad Sci U S A, 74(10):44015,1977.

[5] Leping Li, Yu Liang and Robert L. Bass, GAPWM, *"A Genetic Algorithm Method for Optimizing a Position Weight Matrix"*, Bioinformatics ,2007.

[6] Liu.X, Brutlag.D and Liu. J, *" Bioprospector: Discovering Conserved DNA Motifs in upstream Regulatory Regions of Co-expressed Genes"*, Pacific Symposium on Biocomputing 6, 127-138, 2001.

[7] Lodish . H, Baltimore. D, Berk. A, Zipursky. S.L, Matsudaira. P and Darnell. J, *"Regulation of transcription initiation"*, In Molecular Cell Biology, 3rd ed. (J. Darnell, H. Lodish and D. Baltimore, eds.) 405-481. Scientific American Books, New York, 1995.

[8] Melanie Mitchell, *"An Introduction to Genetic Algorithms"*, MIT Press, ISBN-10:0-262-63185-7.

[9] Michael A. Lones and Andy M. Tyrrell, *"The Evolutionary Computation Approach to Motif Discovery in Biological Sequences "*, GECCO, 2005.

[10] Patrik D'haeseleer, *"How does DNA sequence motif discovery work?"* , Nature Biotechnology, Vol . 24, No.8, 2006.

[11] Pavesi .G , Mereghetti P , Mauri .G , Pesole G , *"Weeder Web: Discovery of Transcription factor binding sites in a set of sequence from co-regulated genes "*, Nucleic Acids Res, 32:W199-203.

[12] Queen .C, Wegman .MN, Korn . LJ, *" Improvements to a program for DNA analysis: a procedure to find homologies among many sequences"*, Nucleic Acids Res, 10:449-56, 1982.

[13] Shane T. Jensen and Jun S. Liu, *"BioOptimizer: a Bayesian scoring function approach to motif discovery"* , Bioinformatics Vol. 20 No. 102004, 1557-1564, 2004.

[14] Shane. T, Jensen. X, Shirley Liu, Qing Zhou and Jun S. Liu, *"Computational Discovery of Gene regulatory Binding Motifs: A Bayesian Perspective"*, statistical science, Vol . 19, No 1, 188-204 , 2004.

[15] Staden .R, *"Computer methods to locate signals in nucleic acid sequences"*, Nucleic Acids Res, 12(1 Pt 2):505-19, 1984.

[16] Stormo. GD, Schneider . TD, Gold. L and Ehrenfeucht.A, *"Perceptron' algorithm to distinguish translational initiation sites in E. coli"*, Nucleic Acids Res, 10(9):2997-3011, 1982.

[17] Werner. T, *"Models for prediction and recognition of eukaryotic promoters"*, Mammalian Genome 10, 168-175, 1999.

[18] Zhi Wei and Shane T. Jensen, *"GAME : detecting cis-regulatory elements using a genetic algorithm"*, Bioinformatics, Vol 22 No. 13 PP. 1577-1584,2006.

Author's Biography

Ms V Bhuvaneswari received her Bachelors Degree (B.Sc.) in Computer technology from Bharathiar University, India 1997, Masters Degree (MCA) in Computer

Applications from IGNOU ,India . and M.Phil in Computer Science in 2003 from Bharathiar University, India . She has qualified JRF, UGC-NET, for Lectureship in the year 2003 She is currently pursuing her doctoral research in School of Computer Science and Engineering at Bharathiar University in the area of Data mining . Her research interest include Bioinformatics , Soft computing and Databases. She is currently working as Lecturer in the School of Computer Science and Engineering, Bharathiar University, India. She has for her credit more than 15 publications in International/ National Conferences.

Ms.Nusurath received her Master degree in M.Sc Computer Science from , University of Calicut, India. She has completed her M.phil in School of Computer Science & Engineering, Bharathiar University India in 2008. Her area of research interest includes Data Mining and Bioinformatics. She has presented a paper in National Conference.