MDGA: Motif Discovery Using A Genetic Algorithm

Dongsheng Che
Department of Computer Science
University of Georgia
Athens, GA 30602, USA
Phone: 01-706-389-6318
che@cs.uga.edu

Yinglei Song
Department of Computer Science
University of Georgia
Athens, GA 30602, USA
Phone: 01-706-542-6702
song@cs.uga.edu

Khaled Rasheed
Department of Computer Science
University of Georgia
Athens, GA 30602, USA
Phone: 01-706-542-3444
khaled@cs.uga.edu

ABSTRACT

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. In this paper, we develop an approach that can be used to predict binding site motifs using a genetic algorithm. Based on the generic framework of a genetic algorithm, the 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. Initial experiments demonstrated that our approach could achieve high prediction accuracy in a small amount of computation time. A promising advantage of our approach is the fact that the computation time does not explicitly depend on the length of target sequences and hence may not increase significantly when the target sequences become very long.

Categories and Subject Descriptors

I.2.8 [**Computing Methodologies**]: Artificial Intelligence – *problem solving, control method, and search.*

General Terms:

Algorithms, Experimentation.

Keywords

Genetic Algorithms, Motif finding, Transcription factor.

1. INTRODUCTION

Transcription factor binding sites are short sequence fragments in the promoter regions of genes. These short fragments, however, play important roles in gene transcription processes. Specifically, the transcription process is initiated when protein molecules bind to the upstream region on these binding sites. On the other hand, the transcription process might be inhibited when other competing molecules interact with these binding sites on its upstream region. Accurate identification of these binding sites is thus important and may facilitate the understanding of the biological mechanisms involved in the transcription regulating processes of a gene. Experimental methods, such as DNase foot-printing [4] and gel-shift assay [5] remain the most accurate and reliable identification methods, 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.

Homologous genes often have similar transcription factor binding sites. It is thus possible to identify the transcription binding sites for a set of homologous genes by comparing their upstream regions and searching for the parts that have the maximal identity in sequence content. Computationally, the difficulty arises from the fact that the locations of the binding site can vary significantly on the upstream regions of different homologous genes. Searching all possible combinations of starting locations is impractical and requires exponential computation time. To avoid the exhaustive search, many computational tools have been developed to identify the common binding sites of homologous genes based on the stochastic Gibbs sampling algorithm, such as AlignACE [13], BioProspector [11] and Gibbs Motif sampler [10]. Initially, the Gibbs sampling programs randomly select one motif element in each sequence. The programs then run through two steps: the predictive update step, updating the background and the motif matrix based on motifs selected, and the sampling step, in which each starting position for a motif in the given sequence is assigned a probability. A motif element is then assigned to that sequence by performing a weighted sample from all the possible starting positions. These steps are iterated until a local maximum is reached or a maximum number of iterations are made. To avoid becoming trapped in a local maximum, the whole process is usually restarted several times with a different seed. Other deterministic approaches introduced heuristics into their algorithms to reduce the computational time. For example, Consensus [7] uses the greedy algorithm to find the binding sites on one sequence at a time. Bailey and Elkan use an EM-algorithm to find a maximum likelihood estimate of parameters in a similar statistical model [1]. These approaches are practically useful and have significantly reduced the computation time needed for binding site prediction. However, the prediction accuracy is not satisfactory and far from the expectation of biologists.

Genetic algorithms (GAs), like Gibbs sampling, apply a stochastic optimization technique, but operate on a population of candidate

GECCO'05, June 25-29, 2005, Washington, DC, USA

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 [3]. Recently, Liu et al. applied a GA to the motif discovery problem, and a program called FMGA was developed [9]. They used the general GA framework and operators described in SAGA (sequence alignment by genetic algorithm) [12]. 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. They manipulated mutations based on the position weight matrices (PWM) to maintain the conserved motifs. Additionally, they also implemented the crossover with special-designed gap penalties to produce the optimal child pattern. To overcome the problem of local optima, they introduced a rearrangement method based on PWM. Experimental results showed that FMGA was more accurate than Gibbs Motif Sample in terms of motif prediction accuracy and needed less computation time when compared with the MEME program. Unfortunately the FMGA software is not publicly available for experimentation and comparison.

In this paper, we propose a new genetic algorithm approach called MDGA to efficiently predict the binding sites for homologous genes. 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. We evaluated the prediction accuracy of MDGA and our experiments showed that 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.

The remainder of the paper is organized as follows. Section 2 describes the proposed approach in detail. Section 3 describes the experiments and results. The paper is concluded in Section 4 with a discussion of the findings and of future work.

2. THE PROPOSED APPROACH

In the proposed genetic algorithm (MDGA), the initial population consists of randomly generated individuals. During the evolution procedure, individuals compete for the opportunity to reproduce. In the remainder of this section, we present a detailed description of the approach.

2.1 Representation

An individual is represented by a list of integers that specify the starting locations of the motif on all the target sequences. We used binary strings to represent individuals where a single starting location occupies 16 bits and the binary encodings of all starting locations are concatenated to form a single binary string.

2.2 Population Initialization

The population is randomly initialized with an integer seed provided by the user on the command line. It contains a fixed number of individuals (100 in our experiments) during the evolution.

2.3 Fitness

The fitness function must be able to provide a measure of similarity among all motifs defined in an individual. One of the popular measures of motif similarity is called 'information content' [14]. After all binding sites are aligned, the information content for a single column can be computed as follows.

$$IC = \sum f_b \log_2 \frac{f_b}{p_b} \qquad (1)$$

where f_b is the observed frequency of nucleotide b on the column and p_b is the background frequency of the same nucleotide. The summation is taken over the four possible types of nucleotides.

We define the fitness score function as the summation of information contents of all columns in the alignment, which reflects the overall similarity of the sequence segments of binding sites defined in an individual. In particular, for a given individual, its binding site motifs can be obtained based on start positions of the motif in each sequence and the motif width (W) given by the user. Motifs are thus aligned and the information content for each column in the alignment can be computed. As the last step, we compute the summation of information contents for all columns to obtain the fitness value as equation 2.

$$fitness = \sum_{i=1}^{W} IC_i \quad (2)$$

To resolve the computational difficulty that may arise when the observed frequency of nucleotide b on the column equals to 0, and thus cannot be evaluated, we used pseudo counts as previous described in [8]. Based on the notion of pseudo counts, the modified f'_b and p'_b of nucleotide b can be written as follows.

$$f'_b = \frac{c_b + d_b}{N - 1 + D} \qquad (3)$$

$$p'_b = \frac{c_{0b} + d_b}{S + D} \qquad (4)$$

where c_b and c_{0b} are the observed counts of nucleotide b on the column and in the background respectively, d_b is the pseudo counts of the nucleotide b, N is the number of sequences, S is the sum of observed counts of all nucleotides in the background, and D is the sum of pseudo counts of all nucleotides.

The 'phase' problem arises from the random selection of start positions described in [8]. Basically, it states that the prediction algorithm may enter and then get locked into non-optimal 'local optima', which are often shifted forms of the optimal pattern. This problem can be alleviated by shifting all starting positions of motifs to the left or right by a small number. We adopt this technique in our algorithm, and allow the starting location of a motif to vary within a small range and the maximum overall information content obtained over the possible starting points is taken to be the fitness.

2.4 Selection

Each iteration, with a certain probability, two parents need to be selected from the population for crossover and generating a child. In order to ensure that every individual has a nonzero probability to be selected for reproduction, we used the Roulette wheel mechanism to choose individuals, where the probability for an individual to be selected is its fitness value normalized with all the individuals in the population.

2.5 Replacement Strategy

The GA used is generational with a generation gap. In each generation, a number of individuals equal to half of the population are generated. The new individuals are merged into the population and the worst one-third of all individuals are eliminated. In the implementation 50 new individuals were created in each generation, merged with the 100 parents and then the worst of the 150 individuals were eliminated.

2.6 Crossover and Mutation

Two crossover operators, single-point and double-point, were implemented and tested in our program. A bitwise mutation operator was adopted. Both crossover and mutation occur with certain probabilities and can be specified by the user as command line parameters.

2.7 Program Implementation

We used GAlib2.4.5 (http://lancet.mit.edu/ga) as the platform for implementation. The package contains a flexible working environment in which users are allowed to vary and experiment with almost all the parameter settings. In addition, users can implement their own crossover and mutation operators. The program MDGA was implemented in C and all related parameters were passed as command line arguments. Our experiments can thus be automated with several shell scripts. The following is the pseudo-code of the MDGA program:

```
1.  SET UP Parameters
        SET W = motif width;
        SET G = maximum iteration number;
        SET S = shift range;
        SELECT a crossover strategy;

2. INITIALISE population with random candidates (vectors of
            start positions);

3. EVALUATE each candidate
   {
        DO loop (-S<= i <= S)
        {
            SET all start positions to original start position + i;
            OBTAIN aligned motifs based on new start positions,
                the motif width W, and input sequences;
            SET fitness value = summation of information
                contents of all columns;
            if (Maximum fitness < current fitness)
                Maximum fitness = current fitness
        }
        RETURN Maximum fitness;
   }

4. REPEAT UNTIL (iteration number = G)
   {
        SELECT parents by roulette wheel selection;
        CROSSOVER parents based on crossover
            strategy picked;
        MUTATE the resulting offspring;
        EVALUATE new candidates;
        REPLACE worst individuals;
   }

5. OUTPUT predicted motifs and their consensus motif
```

3. EXPERIMENTAL RESULTS

3.1 CRP Binding Sites

The dataset of cyclic-AMP receptor protein (CRP) consists of 18 sequences of 105 bps each [15]. Twenty-three binding sites have been determined by using the DNA footprinting method, with a motif width of 22 [14].

To test the performance of the MDGA on different probabilistic parameters and crossover operators, we fixed the probability of mutation at value 0.01 and varied the probabilities of crossover for the single-point and two-point crossover operator. We treat it correctly predicted if the start position difference between predicted and experimentally confirmed is less than five. On average, MDGA achieves the best accuracy when the crossover probability is around 0.4 for both crossover operators.

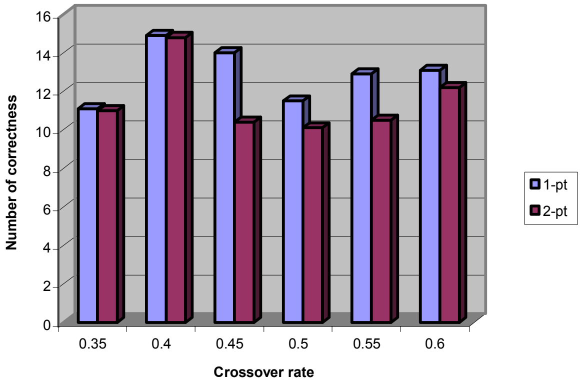

Figure 1: The average number of correctly predicted start positions of CRP dataset using different crossover probabilities for both single-point and two-point crossover operators. For each case, we run MDGA 10 times.

It is surprising that the crossover probability needs to be much less than 1.0 to achieve the optimal average accuracy. In general, the crossover replaces an individual in the original population with the child generated and thus enhances the exploring ability of the algorithm. However, due to the replacements that result from crossovers, the population may deviate from the global optimum when the population starts to converge. An appropriate compromise between the needs to explore the search space and the tendency of converging to the global optimum can thus achieve the best performance in terms of accuracy. It is also evident from Figure 1 that the single-point crossover operator performs as well as or better than the two-point one with all the

Table 1: Comparison of the performance achieved by the Gibbs Sampler, BioProspector and MDGA respectively. FP denotes the starting locations of the binding site(s) measured with the footprint experiments. A single sequence may contain two binding site motifs. GS, BP, GA represent the prediction results obtained with the Gibbs sampler, Bioprospector and MDGA respectively. Additional columns of ER follow each of them to show the deviations of the predicted starting locations of binding sites from the experimental values.

Sequence	FP	GS	ER	BP	ER	GA	ER
1	17, 61	59	-2	63	2	62	1
2	17, 55	53	-2	57	2	56	1
3	76	74	-2	78	2	77	1
4	63	59	-4	65	2	64	1
5	50	11	-39	52	2	51	1
6	7, 60	5	-2	9	2	8	1
7	42	40	-2	26	-16	43	1
8	39	37	-2	41	2	40	1
9	9, 80	7	-2	11	2	10	1
10	14	12	-2	16	2	15	1
11	61	59	-2	63	2	62	1
12	41	47	6	43	2	42	1
13	48	46	-2	50	2	49	1
14	71	69	-2	73	2	72	1
15	17	15	-2	19	2	18	1
16	53	49	-4	55	2	54	1
17	1, 84	25	24	68	-16	56	-28
18	78	74	-4	80	2	77	1

crossover probabilities. This may suggest that the diversity introduced by the single-point crossover is sufficient to explore the search space of the problem.

To compare the performance of MDGA to other approaches that use the Gibbs sampling algorithm to sample the search space, we used MDGA and two other computational tools, the Gibbs Sampler and the BioProspector, to predict the locations of binding site motifs on the set of 18 sequences. It can be seen from Table 1 that MDGA outperforms the two other approaches in terms of prediction accuracy. MDGA failed to predict the correct starting location of the binding site for sequence 17, however, all of the three prediction programs failed on this sequence. Both binding sites on sequence 17 may thus have a much lower similarity in sequence content to those of other sequences.

The higher prediction accuracy achieved by MDGA can possibly be ascribed to the better sampling and exploring capability of the genetic algorithm adopted in MDGA. Compared with a genetic algorithm, the Gibbs sampling algorithm explores the search space by allowing only one of the variables to vary for a single sampling step and thus may fail to correct a strong mistake made initially or during the convergence process.

3.2 YDR02c Binding Sites

The YDR02c sequence dataset was downloaded from http://jura.wi.mit.edu/fraenkel/download/release_v24/fsafiles/. It consists of 15 target genes of transcription factor YDR02c selected by the Chromatin-Immunoprecipitation-micorarray (ChIP-chip) procedure in yeast [6]. The binding site motif pattern has not been experimentally confirmed. We used the MDGA program with different levels of one-point crossover probabilities (range from 0.2 to 0.8) and a fixed motif width of 10. The results showed that the consensus motif pattern is "TCCGGGTAAA" for the highest fitness function value. The sequence logo for this 'best' motif is given in Figure 2.

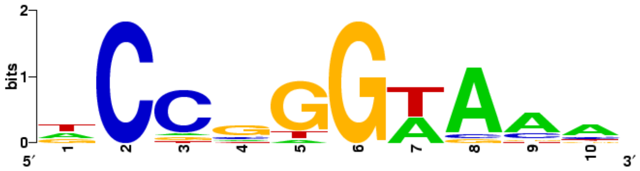

Figure 2: The consensus binding sequence for YDR02c as predicted by application of MDGA program to YDR02c binding regions identified by CHIP-on-chip. The result is presented as sequence logo in which height of the letters in bits is proportional to their frequency [6]

We also used other motif-finding programs with the same motif width. Comparing all motif patterns predicted by these six programs, we conclude that all motif patterns are very similar, suggesting that all programs could detect motif patterns of this dataset in general. Furthermore, we observed that the motif predicted by MDGA is exactly the same as that of the AlignACE program, indicating this pattern could be the true motif pattern from a statistical point of view (Table 2).

Table 2: Comparisons of the conserved motifs predicted by AlignACE, BioProspector, Consensus, Gibbs Sampler, MEME and MDGA programs respectively. AlignACE and MDGA predicted the same pattern, which is "TCCGGGTAAA".

Motif-finding program	Predicted motif
AlignACE	TCCGGGTAAA
BioProspector	TACCGGGTAA
Consensus	CCGGGTAAAA
Gibbs Sampler	TATTTTGATG
MEME	GTCCGGGTAA
MDGA	TCCGGGTAAA

3.3 AZF1 Binding Sites

The AZF1 sequence dataset was also downloaded from http://jura.wi.mit.edu/fraenkel/download/release_v24/fsafiles/. The AZF1 dataset contains 24 sequences with variable sequence lengths, ranging from 175 to 1228. Earlier literature showed that that the consensus motif for AZF1 binding sites is "TTTTTCTT", and it was further predicted as the pattern of "TTTTTCTTTTCCTGTTTC" [6].

We used the MDGA program to experiment with this dataset with different numbers of generations. As shown in Figure 3, the fitness values are stable when the number of generations is 2000 or bigger. We compared the motif pattern predicted by the MDGA program with the true motif and found that the motif patterns predicted by the MDGA program were similar to the true motif pattern when the fitness value was greater than 0.0173, indicating that 2000 generations was adequate for MDGA to find the motif pattern for this dataset.

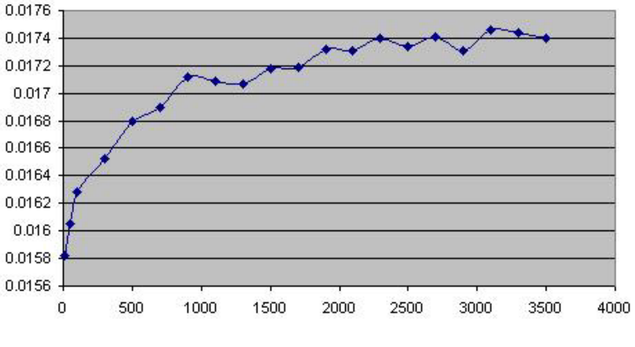

Figure 3: The relationship between fitness value and number of generations. The horizontal axis denotes the number of generations, and the vertical axis denotes the converged fitness values. For each case, we run MDGA 10 times.

Although the accuracy of prediction is the most important issue for motif-finding problem, computation time is another issue needs to be aware when target sequences are long. They can be as large as 3000 base pairs (bps) since potential motif can be from –2000 bp upstream to +1000 bp downstream. Hence, motif identification on a large dataset with long sequences could lead to intensive computation if using exhaustive search method, such as the Gibbs sampling based approach. To test whether the genetic algorithm can gain computation time advantage over other algorithms in such case, we compared the execution time of MDGA and Gibbs sampling based program AlignACE. Based on the fact that the MDGA program with 2000 generations could accurately predict the AZF1 dataset in most cases, we measured the computation time of MDGA under this generation. In addition, we measured and recorded the execution time of the AlignACE program under the same conditions. Our experiments show that, averaged over 30 experiments, computation time was 13.0 seconds for the MDGA program, while it was 20.5 seconds for the AlignACE program as shown in Figure 4. T-test shows that the execution time is significant different between MDGA and AlignACE. The ½ speedup of MDGA shall be useful for motif identification in a large dataset containing long target sequences.

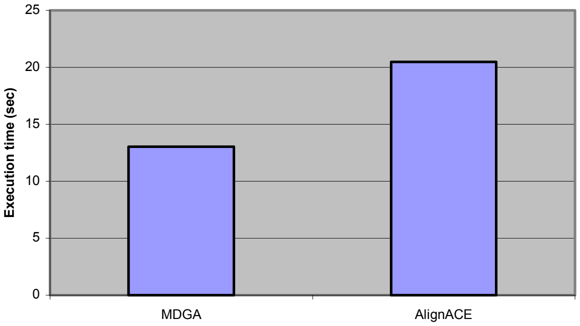

Figure 4: Comparison of computation time between MDGA and AlignACE on AZF1 binding site datasets. T-test shows that the execution time between MDGA and AlignACE is significant different at p <0.00001.

4. CONCLUSION

We have observed from our initial experiments that MDGA is capable of achieving better prediction accuracy than other approaches such as Gibbs Sampler, which explores the search space using a Gibbs sampling algorithm. It is not surprising that, a genetic algorithm based approach is capable of achieving higher prediction accuracy since; in general, the genetic algorithm explores search spaces with a strategy better than that of the Gibbs sampling approach. Another possible advantage of MDGA over the other approaches is its shorter running time when target sequences contain a large number of nucleotides. For example, for a single iteration, Gibbs sampling based program AlignACE needs to exhaustively evaluate the alignment scores of all possible short subsequences on a given target sequence and the running time thus increases exponentially with the length of the target sequences. In contrast, the MDGA does not perform exhaustive search during the evolution and its running time remains independent of the target sequence length. However, we expect the need for a slight increase in population size to avoid the degradation of prediction accuracy when target sequences become very long.

Moreover, MDGA follows the generic framework of a genetic algorithm and therefore its performance can probably be improved using more intelligent operators for crossover and mutation. On the other hand, the fitness evaluation may also be improved if we are able to additionally incorporate terms that reflect the structural similarities among motifs. In addition, more experiments to compare the performance of MDGA with that of other approaches could be useful.

Currently, most of the computational approaches for binding site prediction do not consider the statistical interdependence of nucleotides within a binding site. The statistical interdependence, however, can be biologically important and it is therefore interesting to study its possible effect on the mechanisms of gene transcription. Approaches based on evolutionary computation can possibly reveal this type of effect since the selection process in an evolutionary computation is likely to be similar to that in the long evolutionary history of a biological gene.

MDGA identifies motifs based on the assumption that each sequence contains a motif. This assumption might not be the case in reality since there could be zero to more motifs for each target sequence. The motif identification problem could become more challenging if there are only a limited number of homologous sequences available for comparison, or sequence similarity of the selected homologous sequences is reduced. Completely different strategies for crossover and mutation may need to be designed for binding site predictions in such situations.

ACKNOWLEDGMENTS

The MDGA software for this work used the Galib genetic algorithm package, written by Matthew Wall at Massachusetts Institute of Technology.

REFERENCES

[1] Bailey, T.L. and Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, Stanford, CA, AAAI Press, Bethesda, MD, 1994, 28–36.

[2] Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: A sequence logo generator, *Genome Research*, 14, 6 (June 2004), 1188-1190.

[3] Eiben, A.E. and Smith, J.E. Introduction to Evolutionary Computing. Springer-Verlag, New York. 2003.

[4] 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.

[5] 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.

[6] Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N., Macisaac, K.D., Danford, T.D., Hannett, N.M., Tagne, J.-B., Reynolds, D.B., Yoo, J., Jennings, E.G., Zeitlinger, J., Pokholok, D.K., Kellis, M., Rolfe, P.A., Takusagawa, K.T., Lander, E.S., Gifford, D.K., Fraenkel, E. and Young, R.A. Transcriptional Regulatory Code of a Eukaryotic Genome. *Nature*, 431, 7004 (Sep. 2004), 99-104.

[7] Hertz, G.Z. and Stormo, G.D. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15, 7 (July, 1999), 563–577.

[8] Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F. and Wooton, J.C. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262, 8 (Oct. 1993), 208-214.

[9] 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.

[10] Liu, J.S., Neuwald, A.F. and Lawrence, C.E. Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Am. Stat. Assoc*., 90, 432 (Nov. 1995), 1156–1170.

[11] Liu, X., Brutlag, D.L. and Liu, J.S. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput*., 6, 2001, 127–138.

[12] Notredame, C. and Higgins, D.G. SAGA: Sequence alignment by genetic algorithm. *Nucleic Acids Res.* 24, 8 (Apr. 1996), 1515–1524

[13] Roth, F.R., Hughes, J. D., Estep, P. E. and Church G. M. Finding DNA regulatory motifs within unaligned non-coding sequences clustered by whole-Genome mRNA quantitation. *Nature Biotechnology* 16, 10 (Oct. 1998), 939-45.

[14] Stormo, G.D. Computer methods for analyzing sequence recognition of nucleic acids. *Annu. Rev. BioChem.* 17, 1988, 241-263.

[15] Stormo,G.D. and Hartzell,G.W. Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl Acad. Sci. USA*, 86, 4, (Feb. 1989), 1183–1187.