

MDGA: Motif discovery using genetic algorithm

Dongsheng Che, Haibo Zhao and Yinglei Song

Department of Computer Science, University of Georgia, Athens, GA 30602

ABSTRACT

In this project, we developed an approach that can be used to computationally detect the binding site motifs on the upstream regions of genes. Based on the generic framework of a genetic algorithm, the approach explores the search space comprised of all possible combinations of starting locations of the binding site motifs in all the target sequences with an evolving population. Individuals in the population may stochastically participate in the evolution procedure which includes crossovers and mutations. Experiments demonstrated that, compared with other approaches that use the Gibbs sampling algorithm, an improved level of prediction accuracy was observed using our approach. More importantly, the running time needed in our approach is independent of the length of target sequences and thus can be significantly reduced when the target sequences become very long.

INTRODUCTION

Transcription factor binding sites are relative short fragments in the upstream regions of genes. Accurate identification of these binding sites should facilitate our understanding the mechanisms of how proteins regulate the transcription of genes. Experimental methods, such as DNase footprinting (Galas and Schmitz 1978) and gel-shift assay (Garner and Reczin 1981) are reliable, but time-consuming and expensive. Accordingly, many computational tools have been developed. Some programs use stochastic Gibbs

Sampling method, such as AlignACE (Roth et al., 1998), BioProspector (Liu et al., 2001) and Gibbs Motif sampler (Liu et al., 1995). Consensus uses the greedy algorithm to find motif sites one sequence at a time (Hertz and Stormo, 1999). Bailey and Elkan (1994) use an EM-algorithm to find a maximum likelihood estimate of parameter in a similar statistical model. However, prediction accuracy is far from what biologists expect.

Recently, genetic algorithms were applied to motif discovery. In FMGA (2004), the mutation is performed by using position weight matrices to reserve the completely conserved positions, and the crossover is implemented with special-designed gap penalties to produce the optimal child pattern. FMGA shows superior results by comparing with MEME and Gibbs Motif Sampler programs. Detailed information, however, was not obtained because the paper is not freely accessible.

In this paper, we describe our genetic algorithm approach called MDGA. MDGA explores the search space comprised of all possible combinations of starting locations of the binding site motifs in all the target sequences with an evolving population. The representation and evolutionary techniques are described in Materials and Methods. Initial results of several applications indicated that our method is useful for motif discovery.

MATERIALS AND METHODS

We followed steps in the generic framework of a genetic algorithm. More specifically, a population comprised of 100 randomly generated individuals was created as the starting point of the evolution. During the evolution procedure, individuals competed for the opportunity to reproduce and as a result, one individual in the population must be selected to be replaced by the child generated in the crossover. In the following subsections, we present a detailed description of the approach.

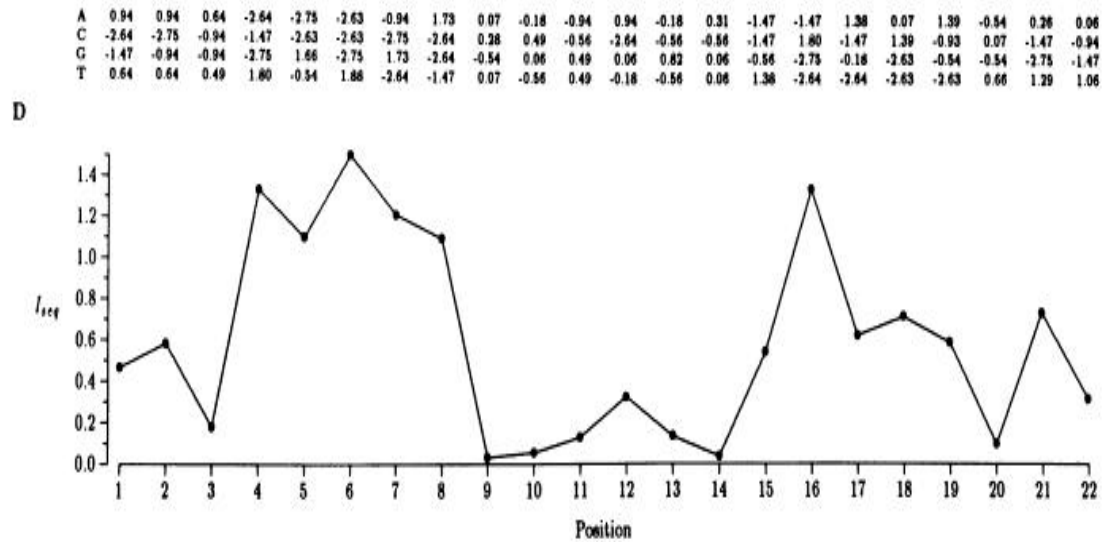


Figure 1: An example for evaluating the fitness value of an individual, the above part lists the frequencies of all bases computed for each column in the multiple alignment of motifs. The lower part plots the computed information content for each column.

Representation

An individual is comprised of 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.

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.

Fitness

The fitness function must be able to provide a measure of the similarity among all motifs defined in an individual. Motifs are thus aligned and the information content for each column in the alignment is computed. The overall information content is the summation of the information content for all columns in the alignment. To further enhance the exploring ability of the algorithm, we allow the starting location of a motif to be variable within a small range and the maximum value of overall information content obtained over the possible starting points is the fitness. The information content for a single column can be computed with equation (1) as follows.

$$IN = \sum f_b \log \frac{f_b}{p_b} \quad (1)$$

where f_b is the frequency of nucleotide b on the column and p_b is the background frequency of the same nucleotide. Figure 1 provides an example for the fitness evaluation.

Selection

For each generation, 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 implemented 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.

Survival Strategy

The individual with the lowest fitness value in the population is replaced with the child generated from the crossover.

Crossover & Mutation

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

Program Implementation

We used GAlib2.4.5 (<http://lancet.mit.edu/ga>) as the platform for implementation. The package contains a flexible working environment such that users are allowed to vary and perform experiments on 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 are passed as command line arguments. Our experiments can thus be automated with several shell scripts.

RESULTS

CRP binding sites: The cyclic-AMP receptor protein (CRP) dataset (Stormo and Hartzell 1989) consists of 18 sequences of 105 bps each. 24 binding sites have been determined by using DNA footprinting method, with the motif width of 22.

To test the performance of the MDGA on different probabilistic parameters and crossover operators, we fixed the mutation probability at value 0.01 and varied the probabilities of crossover for the single-point and double-point crossover operator. It can be seen from Figure 2 that, on average, the MDGA achieves the best accuracy when the cross over probability is around 0.4 for both crossover operators.

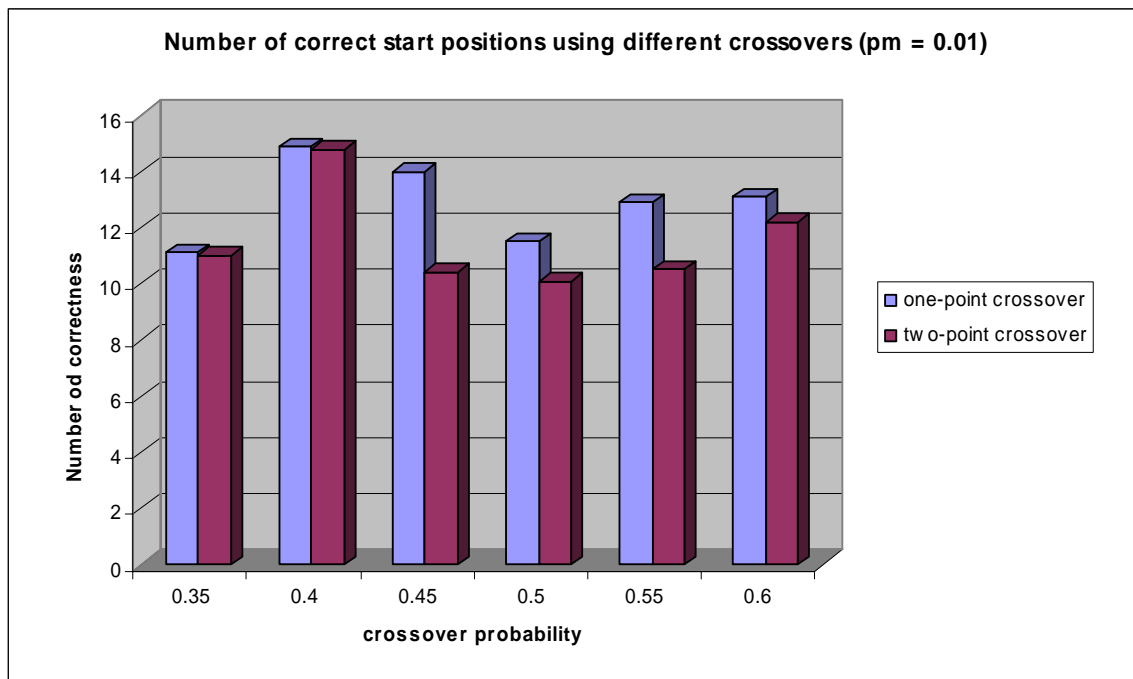


Figure 2: The average prediction accuracy of the algorithm on different crossover probabilities for both single-point and double-point crossover operators (we ran 10 times for every case).

It is not out of our expectation that, the crossover probability needs to be less than 1.0 to achieve the optimal average accuracy. Since, 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 as a result of crossovers, the population may deviate from the global optima 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 optima can thus achieve the best performance in terms of accuracy. It is also evident from the Figure that the single-point crossover operator performs as well as or better than the double-point one in all the 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 the MDGA to other approaches that use the Gibbs sampling algorithm to sample the search space, we used the MDGA and other two computational tools, the Gibbs Sampler and the BioProspector, to predict the locations of binding site motifs on a set of 18 sequences. It can be seen from Table 1 that the MDGA outperforms both other approaches in terms of prediction accuracy. The MDGA failed to predict the correct starting location of the binding site for sequence 17, however, all of the tree 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.

Number	Footprint	Gibbs Sampler	BioProspector	MDGA
1	17, 61	59	63	62
2	17, 55	53	57	56
3	76	74	78	77
4	63	59	65	64
5	50	11	52	51
6	7, 60	5	9	8
7	42	40	26	43
8	39	37	41	40
9	9, 80	7	11	10
10	14	12	16	15
11	29, 61	59	63	62
12	41	47	43	42
13	48	46	50	49
14	71	69	73	72
15	17	15	19	18
16	53	49	55	54
17	1, 84	25	68	56
18	78	74	80	77

Table 1: Comparisons of the performance achieved by the Gibbs Sampler, BioProspector and MDGA respectively, the column for footprint are the starting locations measured experimentally. A single sequence may contain two binding site motifs.

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 Chromatin-Immunoprecipitation-microarray (ChIP-chip) procedure in yeast. The binding site motif pattern has not been confirmed experimentally. We run MDGA program with different levels of one-point crossover probabilities (range from 0.2 to 0.8) and motif width of 10. The result showed the motif pattern is “TCCGGGTAAA” with the highest fitness function value. We also run other motif-finding programs using the same motif width. We found that the motif pattern predicted by MDGA is exactly same as that of AlignACE program, indicating this pattern could be the true motif pattern in terms of the statistical view.

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

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”.

DISCUSSIONS

We have observed from our experiments that, on average, the 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. Another possible advantage

of the MDGA over other approaches is its shorter running time when target sequences contain a large number of nucleotides. For example, for a single iteration, the Gibbs Sampler needs to exhaustively evaluate the alignment scores of all possible short subsequences on a given target sequence and the running time thus increases linearly 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 a slight increase in population size to avoid the degradation of prediction accuracy when target sequences become very long.

Moreover, the MDGA follows the generic framework of a genetic algorithm and its performance can be further improved if multiple operators for crossover and mutation can be developed. 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 into the fitness function. In addition, more experiments also need to be carried out to compare the performance of the MDGA with that of other approaches.

Finally, motif detection could become more difficult if the number of motifs could vary with the target sequence, completely different strategies for crossover and mutation may need to be designed to cope with such situations.

REFERENCES

Bailey, T.L. and Elkan, C. (1994) 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, pp. 28–36.

Galas, D.J. and Schmitz, A. (1978) A DNA footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res.*, **5**, 3157–3170.

Garner, M. M., and A. Revzin. 1981. 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**:3047-3060.

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

Liu, F.F.M., Tsai, J.J.P., Chen, R.M., Chen, S.N. and Shih, S.H. (2004) FMGA: finding motifs by genetic algorithm. *IEEE Fourth Symposium on Bioinformatics and Bioengineering (BIBE 2004)*, May 2004, pp. 459-466.

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

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

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

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