Expert Systems with Applications 37 (2010) 2421–2426

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Automated extraction of extended structured motifs using multi-objective genetic algorithm

Mehmet Kaya

Department of Computer Engineering, Firat University, Elazig, Turkey

ARTICLE INFO

Keywords:
Motif discovery
Data mining
Multi-objective genetic algorithm

ABSTRACT

A structured motif is defined as a collection of highly conserved simple motifs with pre-specified sizes and gaps between them. In structured motif extraction, while all simple motifs are unknown, all gap ranges are known earlier. In this paper, we propose a novel method using multi-objective evolutionary algorithm to extract automatically extended structured motifs in which all simple motifs and gap ranges are unknown. The method employs three conflicting objectives; similarity and support maximization and total gap range minimization. To the best of our knowledge, this is the first effort in this direction. The proposed method can be applied to any data set with a sequential character. Furthermore, it allows any choice of similarity measures for finding motifs. We compare our method with the two well-known structured motif extraction methods, EXMOTIF and RISOTTO. Experiments conducted on synthetics and real data set demonstrate that the proposed method exhibits good performance over the other methods in terms of runtime and accuracy.

© 2009 Elsevier Ltd. All rights reserved.

1. Introduction

Automatic discovery of patterns in unaligned biological sequences is an important problem in molecular biology. An important part of gene regulation is mediated by specific proteins, called transcription factors, which influence the transcription of particular gene by binding to specific site on DNA sequences, called transcription factor binding site. Such binding sites are relatively short stretches of DNA, normally 5–25 nucleotides long, and are located in the so-called promoter regions.

The DNA sites involved in gene regulation can be identified by searching for well-conserved regions in a set of noncoding DNA sequences. Such well-conserved regions, also known as consensus regions, are called motifs and can be found by comparison of the noncoding sequences of related genes in different organism. In the first approach, frequently occurring patterns are likely to correspond to the binding sites of a common transcription factor. The second approach is called phylogenetic footprinting (Duret & Bucher, 1997) and requires careful identification of the appropriate genes to use.

Many simple motif extraction algorithms have been proposed primarily for extracting the transcription factor binding sites, where each motif consists of a unique binding site (Bailey & Elkan, 1994; Pavesi, Mereghetti, Mauri, & Pesole, 2004; Roth1, Hughes, Estep, & Church, 1998; Sinha & Tompa, 2003) or two binding sites separated by a fixed number of gaps (Eskin & Pevzner, 2002). Structured motif extraction problems, in which variable numbers of gaps are allowed, have attracted much attention recently, where the structured motifs can be extracted either from multiple sequences or from a single sequence (Carvalho, Freitas, Oliveira, & Sagot, 2008; Pisanti, Carvalho, Marsan, & Sagot, 2006; Zhang & Zaki, 2006). In many cases, more than one transcription factor may cooperatively regulate a gene. Such patterns are called composite regulatory patterns. To detect the composite regulatory patterns, one may apply single binding site identification algorithms to detect each component separately. However, this solution may fail when some components are not very strong. Thus it is necessary to detect the whole composite regulatory patterns directly; whose gaps and other possibly strong components can increase its significance.

Recently, Genetic Algorithms (GAs) have been used for discovering simple motifs in multiple unaligned DNA sequences (Bi, 2007; Congdon et al., 2005, Congdon, Aman, Nava, Gaskins, & Mattingly, 2008; Che, Song, & Rasheed, 2005; Fogel, Weekes, Varga, Dow, & Harlow, 2004; Liu, Tsai, Chen, Chen, & Shih, 2004; Paul & Iba, 2006; Stine, Dasgupta, & Mukatira, 2003). Of these, Fogel et al. (2004) focused on the development and examination of the utility of evolutionary computation for the discovery of known and putative transcription factor binding site motifs in upstream regions of coexpressed genes. Then, Che et al. (2005) proposed a new GA approach called MDGA to efficiently predict the binding sites for homologous genes. The fitness value for an individual is evaluated by summing up the information content for each column in the

E-mail address: kaya@firat.edu.tr

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.06.101

alignment of its binding site. Congdon et al. (2005) developed a GA approach to motif inference, called GAMI, to work with divergent species, and possibly long nucleotide sequences. The system design reduces the size of the search space as compared to typical window-location approaches for motif inference. They presented preliminary results on data from the literature and from novel projects. Paul and Iba (2006) presented a GA based method for identification of multiple (l, d) motifs in each of the given sequences. The method can handle longer motifs and can identify multiple positions of motif instances of a consensus motif and can extract weakly conserved regions in the given sequences. Finally, Bi (2007) proposed and implemented a genetic-based expectation maximization motif-finding algorithm (GEMFA) aiming to overcome the drawbacks inherent in expectation maximization motif discovery algorithms.

However, all of the above studies employ single-objective to discover motifs, although different methods of fitness calculation are used. In previous work, we first proposed a novel method to demonstrate advantages of multi-objective approach over single-objective ones to discover motifs efficiently and effectively (Kaya, 2009). Then, we extended our method to extract structured motif (Kaya, 2008). In this paper, we improve the previous work in order to extract extended structured motifs. In extended structured motif extraction, all simple motifs and gap ranges are unknown. For this purpose, we used three different conflicting objectives to find nondominated motifs. The objectives of this paper are to;

$$\textit{Maximize Similarity},\ \textit{Minimize Total Gap Range},\ \textit{Maximize Support}$$

The meanings of these objectives will be discussed in the objectives subsection in detail. Next, we compare the three-objective formulation with the single-objective approach based on the following weighted scalar objective function:

$$\textit{Maximize}\ w_1 * \textit{similarity} - w_2 * \textit{gap} + w_3 * \sup\textit{port} \tag{1}$$

The rest of this paper is organized as follows. Multi-objective optimization is defined in Section 2. Our approach of utilizing multi-objective GA to extract automatically extended structured motif is described in Section 3. Experimental results for synthetic and real data sets are reported and discussed in Section 4. Section 5 includes a summary and the conclusions.

2. Multi-objective optimization

Many real world problems involve multiple measures of performance or objectives, which should be optimized simultaneously. Multi-objective optimization (MOO) functions by seeking to optimize the component of a vector-valued objective function. Unlike single-objective optimization, the solution to a MOO problem is a family of points known as the Pareto-optimal set. A general minimization problem of M objectives can be mathematically stated as

$$\begin{aligned}&\text{Minimize}\quad f(x) = [f_i(x), i = 1, \ldots, M] \qquad (2)\\ &\text{subject to}: g_j(x) \leqslant 0 \quad j = 1, 2, \ldots, \quad J\\ &h_k(h) = 0 \quad k = 1, 2, \ldots, K\end{aligned}$$

where $f_i(x)$ is the ith objective function, $g_j(x)$ is the jth inequality constraint. The MOO problem then reduces to finding x such that $f(x)$ is optimized. In general, the objectives of the optimization problem are often conflicting. Optimal performance according to one objective, if such an optimum exists, often implies unacceptable low performance in one or more of the other objective dimensions, creating the need for a compromise to be reached. A suitable solution to such problems involving conflicting objectives should offer *acceptable*, though possibly sub-optimal in the single-objective sense, performance in all objective dimensions, where acceptable is a problem dependent and ultimately subjective concept. An important concept in MOO is that of domination, where a solution x_i is said to dominate another solution x_j if both the following conditions are true:

- the solution x_i is not worse than x_j in all objectives;
- the solution x_i is strictly better than x_j in at least one objective.

This, in turn, leads to the definition of *Pareto-optimality*, where a decision vector $x_i \in U$, where U stands for the universe, is said to be *Pareto-optimal* if and only if there exists no $x_j, x_j \in U$, such that x_i is dominated by x_j. Solution x_i is said to be *nondominated*. The set of all such nondominated solutions is called the *Pareto-optimal set* or the *nondominated set*. In general, MOO problems tend to achieve a family of alternatives which must be considered the relevance of each objective relative to the others (Zitzler, Deb, & Thiele, 2000). Recently, some researchers have studied on different problems by using multi-objective genetic algorithms. We have already participated in some of these efforts with data mining area (Kaya & Alhajj, 2004a, 2004b).

3. The extended structured motif extraction algorithm using multi-objective evolutionary algorithm

3.1. Structured motifs

A structured motif can be defined as an order of collection of simple motifs with gap constraints between each pair of adjacent simple motifs (Zhang & Zaki, 2006). For example, many *retrotransposons* in the *Ty1-copia* group (Policriti, Vitacolonna, Morgante, & Zuccolo, 2004) have as consensus the structured motif: MT[115, 136]MTNTAYGG[121, 151]GTNGAYGAY. Here MT, MTNTAYGG and GTNGAYGAY are three simple motifs; [115, 136] and [121, 151] are variable gap constrains ([minimum gap, maximum gap]) allowed between the adjacent simple motifs. More formally, a structured motif, M, is specified in the form:

$$M_1[\min_1, \max_1]M_2[\min_2, \max_2]M_3, \ldots, M_k$$

where $M_i, 1 \leqslant i \leqslant k$, is a simple motif component, and $\min_i$ and $\max_i$ $(0 \leqslant \min_i \leqslant \max_i)$, are the minimum and maximum number of gaps allowed between M_i and M_{i+1}, respectively. A gap is defined to be the number of intervening positions after M_i, but before M_{i+1}. In the structured motif extraction problem, the components motifs M_i are unknown before extraction. However, some parameters are pre-specified to restrict the structured motifs to be extracted. These are the number of simple motif components, k; the length of each component $M_i, |M_i|$ and the gap range between M_i and $M_{i+1}, [\min_i, \max_i]$.

As extended structured motif extraction is concerned, none of the parameters above are pre-specified. User can only restrict the search space with value ranges of parameters instead of the exact value of each parameter.

We use a well-known high-performance multi-objective genetic algorithm called NSGA II (Deb, Pratap, Agarwal, & Meyerivan, 2002) to find a large number of extended structured motifs from biosequences with respect to three objectives, which will be discussed in the objectives subsection.

3.2. Structure of the individuals

An individual in our method represents the starting locations and ranges of gaps of an extended structured motif on all the target sequences. An individual is divided into n genes, where each gene corresponds to number of simple motif components, the length of each component M_i and the gap range between M_i and M_{i+1}, if any, in the corresponding sequence. The genes are positional, i.e., the first gene deals with the first sequence, the second gene deals with the second sequence, and so on. Each ith gene, $i = 1, \ldots, n$, is subdivided into three fields: weight (w_i) and starting location of first gap in ith sequence (s_i), and the ranges of first and second gaps, $\max_{i1} - \min_{i1}, \max_{i2} - \min_{i2}$, respectively, as shown in Fig. 1. Each individual has also three fields which show the length of each component $M_i, |M_i|$. In Fig. 1, the number of component was set to 3.

The field weight (w_i) is a real-valued variable taking values in the range [0,…,1]. This variable indicates whether or not the extended structured motif is present in the corresponding sequence. More precisely, when w_i is smaller than a user-defined threshold (called *Limit*) the extended structured motif will not be extracted from the ith sequence. Therefore, the greater the value of the threshold *Limit*, the smaller is the probability that the corresponding sequence will be included in discovering that motif.

Note that the above encoding is quite flexible with respect to the length of the motifs and gap ranges. A traditional GA is very limited in this aspect, since it can only cope with fixed-length motifs. In our approach, although each individual has a fixed-length, the genes are interpreted (based on the value of the weight w_i) in such a way that the individual phenotype (the motif) has a variable length. The start of the first population consists of generating, arbitrarily, a fixed number of individuals during the evolution.

3.3. Objectives and selection

The fitness of an individual in our method is assessed on the basis of *similarity*, *gap range* and *support*.

Similarity: Several methods for alignment of multiple sequences have been proposed in the literature. Hertz and Stormo (1999) have proposed a relative entropy based score metric for alignment of multiple sequences. The score metric is as follows:

$$I_{align} = \sum_{i=1}^{l} \sum_{b \in \Sigma} f_{b,i} \log_a \frac{f_{b,i}}{p_b} \tag{3}$$

where l is the total length of components, $f_{b,i}$ is the observed frequency of the symbol b at position $i, \sum$ is the alphabet of symbols (for DNA sequences, $\sum = \{A, C, T, G\}$), a is the base of the logarithm, and p_b is the background probability distribution of symbol b. When all the symbols are same, the information content will be the highest, and the sequences will be optimally aligned. However, due to linear summation of each positional information content, the alignment may not be optimal. For example, consider two alignments of 4 DNA sequences of length 2: {AT, AC, AG, AA} and {AC, TC, AG, TG}. The information content of the first alignment using $a = 2$ is

$$I_{align1} = 1.0 * \log_2 \frac{1.0}{0.25} + 4 * 0.25 * \log_2 \frac{0.25}{0.25} = 2$$

$$I_{align2} = 2 * 0.5 * \log_2 \frac{0.5}{0.25} + 2 * 0.5 * \log_2 \frac{0.5}{0.25} = 2$$

Using these scores, we cannot determine the best alignment (alignment 1) that has a conserved position (position 1). For this purpose, we proposed a novel similarity measure. It performs a measure of similarity among all motif instances defining an individual. To calculate it, we first generate a position weight matrix from the motif patterns found by the proposed method in every sequence. Then, the dominance value (dv) of the dominant nucleotide in each column is found as follows:

$$dv(i) = \max_b \{f(b,i)\}, \quad i = 1, \ldots, l \tag{4}$$

where $f(b,i)$ is the score of the nucleotide b on column i in the position weight matrix, $dv(i)$ is the dominance value of the dominant nucleotide on column i, and l is motif length.

Table 1
The position weight matrix of a motif with length 4.

	1	2	3	4
A	0.2	1	0	0
C	0.2	0	1	1
T	0.6	0	0	0
G	0	0	0	0

Table 2
The position weight matrix of a motif with length 5.

	1	2	3	4	5
A	0.25	0	0.75	1	0
C	0.5	0	0.25	0	0
T	0.25	1	0	0	1
G	0	0	0	0	0

We define the similarity objective function of motif M as the average of the dominance values of all columns in the position weight matrix. In other words,

$$Similarity(M) = \frac{\sum_{i=1}^{l} dv(i)}{l} \tag{5}$$

The closer the value of the similarity of a motif, M, to one, the larger the probability that the candidate motif will be discovered as a motif. The following example shows how to compute the similarity measure in given two position weight matrixes with different size:

The first matrix (Table 1) implies that the number of target sequences in a dataset may be 5 and the motif length is found to be 4. In such a case, the dominant nucleotide for column 1 is T and its dominance value is 0.6. The dominance values of the other columns are 1. As for the similarity value, it is computed as:

$$\frac{(0.6 + 1 + 1 + 1)}{4} = 0.9$$

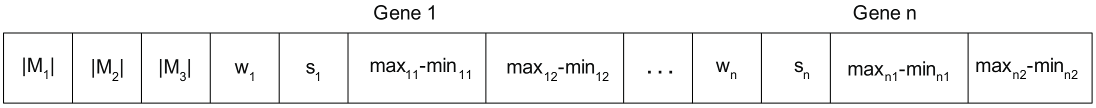
Similarly, in the second matrix (Table 2), the number of target sequences may be 8 and the motif length is determined to be 5. The similarity value is computed as 0.85 for this longer motif.

3.3.1. Gap range

In extended structured motif extraction, less gap range is always desired because the less the gap range of a structured motif, the lesser is the chance that it simply occurred by chance in the given target sequence.

3.3.2. Support

Here, the meaning of this objective is the same with that in the data mining field. Sometimes, a candidate motif may not appear in every sequence. In other words, one or more sequence may not include any candidate motif. In this case, the aim should become to discover optimal motif, leaving these corrupted sequences out. So,



				Gene 1					Gene n		
$\|M_1\|$	$\|M_2\|$	$\|M_3\|$	w_1	s_1	$\max_{11}$-$\min_{11}$	$\max_{12}$-$\min_{12}$	. . .	w_n	s_n	$\max_{n1}$-$\min_{n1}$	$\max_{n2}$-$\min_{n2}$

Fig. 1. Representation of an individual.

support is the number of sequences composing candidate motifs. The greater the support value, the stronger is the motif covered by most of the sequences in the dataset. Overall, the optimal motif discovery problem is converted into the following three-objective optimization problem:

$$\textit{Maximize similarity}(M),\ \textit{Minimize total gap range}(M), \textit{Maximize}\ \sup \textit{port}(M)$$

After applying the nondominated sort algorithm, the individuals are assigned a crowding distance and selection is performed using the crowded tournament selection. The crowding distance is computed as the sum of the difference of the objective values of the solutions preceding and following the current solution (corresponding to the individual under consideration) for each objective. This provides a measure of the density of the solution surrounding a particular point in the population.

In crowded tournament selection, a pair of individuals is selected randomly, and the one with the lower rank is selected. If the ranks of the individuals are equal, then the individual with a larger crowding distance is chosen. The large crowding distance ensures that the solutions are spread along the Pareto-optimal front.

3.4. Genetic operators

In our method, the usual one-point crossover operator is stochastically applied with a predefined probability, using two individuals of the selected pool. The mutation operator is used to foster more exploration of the search space and to avoid unrecoverable loss of genetic material that leads to premature convergence to some local minima. In general, mutation is implemented by changing the value of a specific position of an individual with a given probability, denominated mutation probability. Our method developed three mutation operators tailored for our genome representation:

3.4.1. Shift the starting location towards the right

The value in the starting location of a randomly selected gene is increased by one.

3.4.2. Shift the starting location towards the left

The value in the starting location of a randomly selected gene is decreased by one.

3.4.3. Random-changing

The mutation produces a small integer number that is then added to or subtracted from the current content of any of length, weight or starting location. This is implemented in such a way that the lower and upper bounds the domain of the field are never exceeded.

To sum up, the process employed in this study can be summarized by the following algorithm,

The Algorithm

```
Input: Population size N; Maximum number of generations G;
Crossover probability $p_c$; Mutation rate $p_m$.
Output: Nondominated set
P:=Initialize (P)
while the termination criterion is not satisfied do
C:=Select From (P)
$C^l$:=Genetic Operators (C)
P:=Replace ($PUC^l$
end while
return (P)
```

First, an initial population P is generated in Step 1. Pairs of parent solutions are chosen from the current population P in Step 3. The set of the selected pairs of parent solutions is denoted by C in Step 3. Crossover and mutation operations are applied to each pair in C to generate the offspring population C^l in Step 4. The next population is constructed by choosing good solutions from the merged population PUC^l. The pareto-dominance relation and a crowding measure are used to evaluate each solution in the current population P in Step 3 and the merged population PUC^l in Step 5. Elitism is implemented in Step 5 by choosing good solutions as members in the next population from the merged solution PUC^l.

4. Experimental results

We conducted some experiments in order to analyze and demonstrate the efficiency and effectiveness of our method. Further, the superiority of the proposed approach has been demonstrated by comparing it with two existing structured motif extraction methods, namely EXOTIF (Zhang & Zaki, 2006), RISOTTO (Pisanti et al., 2006). In our experiments, we concentrate on testing the accuracy and the time requirements as well as changes in the main factors that affects the proposed multi-objective process, namely finding nondominated sets, support, gap range and similarity. All of the experiments have been conducted on a C2D 1.83 GHz CPU with 1GB of memory and running Windows XP. Further, in all the experiments conducted in this study, the process started with a population of 200 individuals and the results were obtained with the average of 20 runs. As the termination criteria, the maximum number of generations has been fixed at 3000. Finally, while the crossover probability is chosen to be 0.8, the mutation rate of 0.3 was used for each kind of mutation. The standard single-objective GA was also executed using the same parameter values; and the weight values $w^1 = 20$, $w^2 = 1$ and $w^3 = 1$. As data sets are concerned, we used synthetic and real data sets.

4.1. Synthetic data set

We randomly generated (with a uniform distribution over the four letters size DNA alphabet) a synthetic dataset with planted structured motifs. Each dataset consists of 70 sequences of size 1000 where we planted one motif.

In all the experiments, we varied one parameter while keeping the other fixed. We set the default gap range to [0, 20], the default simple motif component length to (Bailey & Elkan, 1994; Sinha & Tompa, 2003) and the default number of component to 3.

The first experiment deals with some non-dominated solutions reported in Table 3. Here, the values of similarity, total gap and similarity of some non-dominated solutions are given with respect

Table 3
Comparisons of the objective values of the nondominated solutions.

	Similarity	Total gap	Support (%)
Our method	⩾ 0.90	32	32
		24	29
		18	24
		11	20
	0.77	⩽ 20	28
	0.83		22
	0.85		19
	0.89		15
	0.93		12
	0.84	28	⩾ 30
	0.79	24	
	0.75	21	
	0.72	19	
Single-objective GA	0.89	29	31

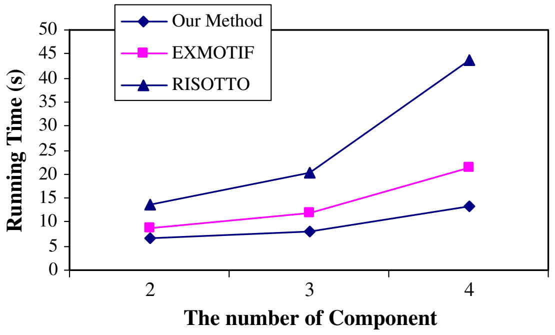


Fig. 2. Comparisons of running times with respect to the number of component.

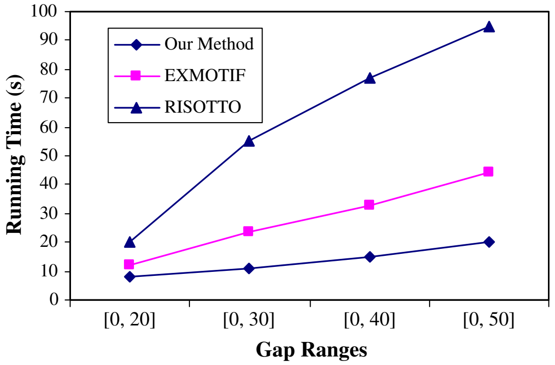


Fig. 3. Comparisons of running times with respect to the gap ranges.

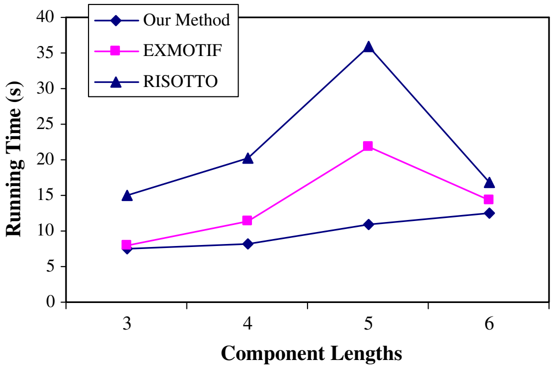


Fig. 4. Comparisons of running times with respect to the compound lengths.

to pre-specified objective value. As can be easily seen from Table 3, as the support value decreases, the average gap decreases as well, for *Similarity* $\geqslant$ 0.90. However, for *Total Gap* $\leqslant$ 20, as the support value decreases, the similarity value raises. Thus, the tradeoff between the similarity and the average gap is clearly observed for five different solutions. Table 3 gives the solution found by the standard single-objective GA process as well.

The next experiment is dedicated to investigate the time as a function of the number components. The results are given in Fig. 2. We observe that as number of components increases the time gap between our method and the other approaches increases.

Figs. 3 and 4 shows the effect of increasing gap ranges, from [0, 20] to [0, 50]. We find that as the gap range increases the time for our method increases at a slower rate compared to the other two approaches.

In the last experiment, we plot the effect of increasing component lengths. We find that the time first increases and then decreases for EXMOTIF and RISOTTO. This is because there are a large number of motif occurrences for length 4 and length 5, but relatively few occurrences for length 6. However, this is not true for our method because it extracts a non-dominated solution in every length.

4.2. Real data set

In this experiment, some empirical tests on a real data set in order to evaluate the accuracy and applicability of the proposed approach and compared our method with EXMOTIF structured motif extraction algorithm. We have used GAL4, CAT8, HAP1, LEU3, LYS, PPR1 and PUT3 gene families (Helden, Rios, & Vides, 2000). For this experiment, as the number of component is set to 3, the length of each component is adjusted in between 2 and 5. To extract the extended structured motifs with more quality, we restricted the gap range to [0, 20].

As a result of this experiment, is has been observed in Table 4 that our method can successfully predict the related motifs with very high accuracy. Also, our solution outperforms the EXMOTIF based solution because EXMOTIF cannot extract extended structured motifs.

5. Discussion and conclusions

In this paper, we contributed to the ongoing research by proposing a multi-objective evolutionary algorithm based method for extracting extended structured motifs with respect to criteria we defined. These criteria are similarity of instances, total gap range and support exhibiting the strongness of motif.

The proposed method includes five contributions. First, the algorithm is equally applicable to any variety of sequential data. Second, it allows arbitrary similarity measure. Although we used relatively a simple and strong similarity measure in the paper, it can be easily changed or extended. Another contribution is that a large number of nondominated sets are obtained by its single run. Thus, the decision maker can understand the tradeoff between the similarity, motif length and support by the obtained motifs. Next, by our method, more than one instance may be discovered in the same sequence and multiple-motifs may be extracted. Fourth, the nondominated motifs are obtained without giving motif length. Finally, our method outperforms the two well-known

Table 4
Comparisons of motifs predicted by our method and EXMOTIF.

Family	Known motif	Predicted motif	EXMOTIF
GAL4	CGGRnnRCYnYnCnCCG	CGG[3,4]YY[5,5]CCG	CGG[11,11]CCG
CAT8	CGGnnnnnnGGA	CGG[6,6]GGA	CGG[6,6]GGA
HAP1	CGGnnnTAnCGGCGGnnnTAnCGGnnnTA	CGG[6,7]GGCG[5,6]CGG	CGG[6,6]CGG
LEU3	RCCGGnnCCGGY	CCGG[2,3]CCGG	CCG[4,4]CGG
LYS	WWWTCCRnYGGAWWW	TCC[3,3]GGA	TCC[3,3]GGA
PPR1	WYCGGnnWWYKCCGAW	CGG[5,6]TCCGA	CGG[6,6]CCG
PUT3	YCGGnAnGCGnAnnnCCGA	CGG[3,3]GCAA[3,3]CCG	CGG[10,11]CCG

structured motif extraction methods in terms of runtime and accuracy.

The experiments conducted on synthetic and real data sets illustrated that multi-objective GA is more appropriate and can be used more effectively to achieve nondominated solutions than the classical algorithms described in the literature. The results of two data sets are consistent and hence encouraging. In the future, we are planning to investigate the possibility of applying multi-agent systems to motif discovery problem and to develop and improve different corresponding algorithms for this purpose.

Acknowledgement

This work was supported by the TUBITAK (The Scientific and Technological Research Council of Turkey) under Grant No: 106E199.

References

Bailey, T. L., & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the second international conference ISMB* (pp. 28–36), USA.

Bi, C. (2007). A genetic-based em motif-finding algorithm for biological sequence analysis. In *Proceedings of the 2007 IEEE symposium on computational intelligence in bioinformatics and computational biology* (pp. 275–282).

Carvalho, A. M., Freitas, A. T., Oliveira, A. L., & Sagot, M. F. (2008). An efficient algorithm for the identification of structured motifs in DNA promoter sequences. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 3*(2).

Che, D., Song, Y., & Rasheed, K. (2005). MDGA: motif discovery using a genetic algorithm. In *Proceedings of the GECCO'05* (447–452), USA.

Congdon, C. B., Aman, J. C., Nava, G. M., Gaskins, H. R., & Mattingly, C. J. (2008). An evaluation of information content as a metric for the inference of putative conserved noncoding regions in DNA sequences using a genetic algorithms approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 5*(1), 1–14.

Congdon, C. B., et al. (2005). Preliminary results for GAMI: A genetic algorithms approach to motif inference. In *Proceedings of the CIBCB'05* (1–8), USA.

Deb, K., Pratap, A., Agarwal, S., & Meyerivan, T. (2002). A fast and elitist multi-objective genetic algorithm: NSGA II. *IEEE Transaction on Evolutionary Computation, 6*, 182–197.

Duret, L., & Bucher, P. (1997). Searching for regulatory elements in human noncoding sequences. *Current Opinions in Structural Biology, 7*(3), 399–406.

Eskin, E., & Pevzner, P. (2002). Finding composite regulatory patterns in DNA sequences. *Bioinformatics, 18*(1), 354–363.

Fogel, G. B., Weekes, D. G., Varga, G., Dow, E. R., & Harlow, H. B. (2004). Discovery of sequences motifs related to coexpression of genes using evolutionary computation. *Nucleic Acids Research, 32*(13), 3826–3835.

Helden, J. V., Rios, A. F., & Vides, J. C. (2000). Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Research, 28*(8), 1808–1818.

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

Kaya, M. (2009). MOGAMOD: Multi-objective genetic algorithm for motif discovery. Expert Systems with Applications, 36(2P1), 1039–1047.

Kaya, M. (2008). A novel approach to extract structured motifs by multi-objective genetic algorithm. In *IEEE international symposium on computer-based medical systems, (CBMS'08)*.

Kaya, M., & Alhajj, R. (2004a). Integrating multi-objective genetic algorithms into clustering for fuzzy association rules mining. In *IEEE international conference on data mining (ICDM 2004)*, Brighton, UK, 1–4 November.

Kaya, M., & Alhajj, R. (2004b). Multi-objective genetic algorithm based approach for optimizing fuzzy sequential patterns. In *16th IEEE international conference on tools with artificial intelligence*, Boca Raton, FL, USA, 15–17 November.

Liu, F. M. M., Tsai, J. J. P., Chen, R. M., Chen, S. N., & Shih, S. H., (2004). FMGA: Finding motifs by genetic algorithm. In *Proceedings of the BIBE'04 Taiwan*, (pp. 459–466).

Paul, T. K., & Iba, H. (2006). Identification of weak motifs in multiple biological sequences using genetic algorithm. In *Proceedings of the GECCO'06* (pp. 271–278), USA.

Pavesi, G., Mereghetti, P., Mauri, G., & Pesole, G. (2004). Weeder Web: Discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Research, 32*, W199–W203.

Pisanti, N., Carvalho, A. M., Marsan, L., & Sagot, M. F. (2006). RISOTTO: Fast extraction of motifs with mismatches. In *Seventh latin American theoretical informatics symposium*.

Policriti, A., Vitacolonna, N., Morgante, M., & Zuccolo, A. (2004). Structured motif search. In *Symposium on research in computational molecular biology* (pp. 133–139).

Roth1, F. P., Hughes, J. D., Estep, P. W., & Church, G. M. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Natural Biotechnology, 16*, 939–945.

Sinha, S., & Tompa, M. (2003). YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Research, 31*, 3586–3588.

Stine, M., Dasgupta, D., & Mukatira, S. (2003). Motif discovery in upstream sequences of coordinately expressed genes. In *CEC'03* (pp. 1596–1603), USA.

Zhang, Y., & Zaki, M. (2006). EXMOTIF: Efficient structured motif extraction. *Algorithms for Molecular Biology, 1*, 21.

Zitzler, E., Deb, K., & Thiele, L. (2000). Comparison of multiobjective evolutionary algorithms: Empirical results. *Evolutionary Computation, 2*, 173–195.