A Novel Computational Based Method for Discovery of Sequence Motifs from Coexpressed Genes

Wengang Zhou, Hong Zhu, Guixia Liu, Yanxin Huang,
Yan Wang, Dongbing Han, Chunguang Zhou

College of Computer Science and Technology, Jilin University, Changchun 130012, P. R. China

wgzhou@email.jlu.edu.cn, cgzhou@jlu.edu.cn

Abstract

The transcriptional regulation of gene expression is a key mechanism in the functioning of the cell. It is mostly affected through acting element binding to specific sequence motifs. In this paper, we present a computational based approach to select the most relevant information for searching binding motifs from the long upstream regions. First, we demonstrate that evolutionary computation method can be used for the discovery of binding motifs. Then we propose a novel algorithm IPSO-GA by integrating an improved particle swarm optimization with genetic algorithm to search sequence motifs from coexpressed genes regulated by the NF-kb transcription factor. Experiment results show that the proposed algorithm can find the binding motifs efficiently. Some of these discovered motifs have been determined by the experiment and other potential motifs can more probably present novel binding motifs that are not discovered yet.

Keyword: Transcription Factor, Binding Motif, Particle Swarm Optimization, Genetic Algorithm.

I. Introduction

The regulation of gene expression in the eukaryotic cell happens at several different levels among which the transcriptional one is the most important. Some researchers have proposed that the next phase of genomics is to comprehend the entire functional elements [1]. Two of the most important functional elements are cis-acting or trans-acting element, usually a protein that stimulates or represses gene transcription, and a recognition site, usually a short sequence motif that is located in upstream of the coding region. Nowadays, the availability of several fully sequenced genomes and other experimental data has made the more complete understanding of regulation elements and their binding motifs become possible [2]. It will also permit a deeper comprehension of the potential functions of individual genes.

Several computational methods for discovery of binding motifs have been proposed [3-5] in the literature. A widely used strategy for identification of regulatory elements is that coexpressed genes may share common regulatory elements. So we can identify binding motifs from a set of genes experimentally known or presumed to be coregulated [6], for example, these genes are involved in the same biological process or show similar mRNA expression profiles in microarray experiments. Methods such as consensus, Gibbs motif sampler, BioProspector and ANN-SPEC [7-9] have successfully applied this strategy in finding regulatory elements from lower organisms such as bacteria and yeast.

In this paper, we present a computational based approach to select the most relevant information for searching binding motif from the long sequences of upstream regions. First, we demonstrate that evolutionary computation method can be used for motif discovery. Then we propose a novel hybrid

algorithm IPSO-GA by integrating an improved particle swarm optimization (IPSO) with genetic algorithm (GA) to search sequence motifs from coexpressed genes regulated by the NF-kb transcription factor. Experiment results show that the proposed algorithm can find the binding motifs efficiently. Some of these discovered motifs have been determined by experiment and other potential motifs are previously unknown. These putative motifs can more probably present novel binding sites that are not discovered yet. In general, the results are highly encouraging.

II. Database

TRANSFAC [10] is the largest and most commonly used database on eukaryotic cis-acting regulatory DNA elements. It catalogs eukaryotic transcription factors and their known binding sites that cover the whole range from yeast to human. The TRANSFAC data have been generally extracted from the original literature, occasionally they have been taken from other compilations which is appropriately indicated. TRANSCompel database is originated from the COMPEL [11] and provides the information of composite regulatory elements which contain two closely situated binding sites and actually are minimal functional units providing combinatorial transcriptional regulation for distinct transcription factors.

It contains 256 experimentally validated composite elements from which we choose one transcription factor, nuclear factor kappa B (NF-kb) as test example. This transcription factor is chosen because its binding mechanism is well studied and it represents the families with multiple family members binding to slightly different binding motif, thus increasing the difficulty of motif search. The nine genes regulated by transcription factor NF-kb with known sequence motifs in their 1kb upstream region are shown in table 1. In our experiment, the 1kb regions upstream to the transcription start site for each of these genes are searched for binding motifs discovery.

Table 1. Genes with experimentally validated Oct binding sites

TF	Gene name	Access. no	Species	Binding motif
NF-kb	ELAM-1	C00097	H. sapiens	GGGGATTT
NF-kb	Interferon-beta	C00099	H. sapiens	GGGAAATT
NF-kb	Serum amyloid A2	C00100	H. sapiens	GGACTTTC
NF-kb	Serum amyloid A1	C00101	R. norvegicus	GACTTTCC
NF-kb	Interleukin-6	C00152	H. sapiens	GGATTTTC
NF-kb	Serum amyloid A3	C00153	M. musculus	GAAATGCC
NF-kb	ICAM-1	C00155	H. sapiens	GAAATTCC
NF-kb	GM-CSF	C00156	M. musculus	GAAATTCC
NF-kb	Interleukin-2	C00165	H. sapiens	GTAGTTCC

III. IPSO-GA

Particle swarm optimization (PSO) method is an evolutionary computation technique first developed by Kennedy and Eberhart [12, 13] in 1995. It starts with the random initialization of a population of individuals (particles) in the search space and works on the social behavior of the particles in the swarm. Therefore, it finds the global best solution by simply adjusting the trajectory of each individual towards its own best location and towards the best particle of the swarm at each generation [14, 15]. Each particle in the search space is adjusted by dynamically altering the velocity of each particle according to its own flying experience and the flying experience of other particles.

The position and the velocity of the i_{th} particle in the n-dimensional search space can be represented as $X_i = (x_{i1}, x_{i2}, \cdots, x_{in})$ and $V_i = (v_{i1}, v_{i2}, \cdots, v_{in})$ respectively. Each particle has its best position $P_i = (p_{i1}, p_{i2}, \cdots, p_{in})$ correspond to the personal fitness obtained at time t. The global best

particle is denoted by P_g which represents the fittest particle found so far at time t. The velocity and position vector is calculated according to the following equations:

$$V_i(t+1) = w \cdot V_i(t) + c_1 \cdot rand() \cdot (P_i(t) - X_i(t)) + c_2 \cdot rand() \cdot (P_g(t) - X_i(t)) \quad (1)$$

$$X_i(t+1) = X_i(t) + V_i(t) \quad (2)$$

Where c_1 and c_2 are constants and are known as acceleration coefficients, w is called the inertia weight, *rand()* generates random numbers in the range of [0, 1].

We propose a novel algorithm IPSO-GA by integrating an improved particle swarm optimization (IPSO) with genetic algorithm (GA). An additional local search operation with two proposed operators is introduced in IPSO. Nine offspring individuals are produced by using the two operators for each parent individual before updating its position. Then the best individual is chosen from the total ten individuals to replace the original one. After the update of velocity and position for each particle, we execute the crossover operation which is a basic process in genetic algorithm (GA) in a predefined probability to increase the diversity of population. We use the single point crossover. The details are shown in section 4.

IV. Discovery of Sequence Motif Using IPSO-GA

It is well known that evolutionary computation is fit for solving combinatorial optimization problem. We will specify that identification of binding motif can be formulated as a combinatorial optimization problem. Hence, the application of the proposed novel algorithm IPSO-GA for discovery of binding motifs appears to be promising. Sequence motif consists of a set of windows with one window for each sequence. We can choose one window from each sequence, and then the aim of our algorithm is to find the optimum combination of these windows that can maximize the fitness function. The number of possible combination (search space) is computed according to the following formula:

$$p = (l - d)^S \quad (3)$$

Where l is the length of the sequences, d is the width of the window being used, and s is the number of sequences. The algorithm is implemented in Matlab 6.5. The sequence information is used to calculate a nucleotide likelihood matrix based on which fitness is measured.

A. *Encoding*

In this experiment, we use a fixed sequence motif window size of eight. A single candidate solution represents a set of windows randomly placed over the 1kb upstream sequences with only one window per sequence. Thus the i_{th} particle is initialized as a nine dimensional vector with the following form: $X_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}, x_{i6}, x_{i7}, x_{i8}, x_{i9})$ .where x_{ik} is a random number from 1 to 993 (because the sequence length is one thousand and the window size is equal to eight) and it represents the initial position of the k_{th} window. The details are shown in figure 1.

C00097: GCATGCGCCACCATGCCCAGCTAATTTTGTATTTTTTTTAGAG
C00099: TGCCTTCTGAGTTCTCCATCCCACCCTGGTTGTTTTTTTTCTAT
C00100: TGAGGAAATGACCGGTATAGTCAGGAGCTGGCTTTTTTTTTGC
C00101: ACACGTGGATCTGTGGCACCCTCCACCCACACAAAAGCAAAA
C00152: TGCCACAAGGTCCTCCTTTGACATCCCCAACAAAGAGGTGAG
C00153: TCCTGCTATAGGGCCAGGAAAACAAAGATGAGCATGCCATTT
C00155: CGTGATCCTTTATAGCGCTAGCCACCTGGGGGCCAAGGGGCG
C00156: CAGCCTCAGAGACCCAGGTATCCCATAATGGTACAGATAGCA
C00165: TTGTGGCAGGAGTTGAGGTTACTGTGAGTAGTGATTAAAGAG

Fig.1. The individual X_i = (8, 22, 31, 17, 6, 23, 13, 27, 10) represents the set of red windows

B. Local Search

For the purpose of enlarging search space and intensifying the ability of local search, we have developed two variation operators: window position change and A+T percentage measure. Before updating the position of each individual (particle), we use the two operators to generate nine offspring solutions for each individual and select the best one from the total ten individuals (one parent and nine offspring) to replace the original individual. We first generate a random number from the range [1, 3] which decides to how many times we use the two operators. Each time an operator is required, the choice on the type of variation operators is made according to the user-defined probabilities for each operator. This process is repeated until the maximum times are reached. The main pseudo-code is summarized in figure 2.

```
For each individual P_j with 1≤j≤popsize in the population P
   initialize n=1;
   while n<=9 do
    generate a random number ntime from the range [1, 3];
    for i=1 to ntime
    if  rand() < 0.2
    choose one window at random from parent individual;
    use the window position change operator;
    else use the A+T percentage measure operator;
    end
    increase n;
    calculate the fitness value for the ten individuals;
    select the individual lbest with the maximum fitness;
    replace the original individual P_j with lbest;
    end
end
```

Fig. 2. The main steps of the local search procedure for each individual in the population

One window in the parent solution is chosen randomly for modification. When we use the window position change operator, the window is moved either to the left or to the right across the 1kb upstream region with equal probability. A choice of new window position is chosen at random from the range [1, n], where n represents the maximum number of nucleotides in that direction. In our experiment, n is equal to 993 (sequence length - window width + 1) with the sequence starting from the 5' end. We propose the A+T percentage measure operator by getting some hints from the known binding motif information as shown in table 1. It is obvious that these motifs have higher A+T percentage. Then the windows with higher A+T percentage are more probable binding motif. When a single window is chosen at random from an individual in the population, the average A+T percentage is computed for all windows except the window being modified. Then a percentage similarity to this average is calculated for all possible windows in the full 1000 nucleotide sequence for the window being modified. All the window positions are stored with A+T similarity equal to or greater than a user-defined threshold. We set the threshold to 1 in all the experiments. From this set of window positions, a new location for the window being modified is chosen with equal probability.

C. Fitness Function

We introduce two criteria: similarity and complexity [16] for evaluating the individuals. Since we aim to get the sequence motif, similarity between these short sequences must be satisfied. But complexity of sequences should also be considered so that avoid low complexity solutions, for example, the two sequences 'AAAA' and 'AAAA' are very similar in fact they are identical, but it is not a meaningful motif. So the total fitness of each individual is calculated according to the following formula:

$$Fitness(i) = w_1 \times Similarity(i) + w_2 \times Complexity(i) \quad (4)$$

Where w_1 and w_2 are used to balance the importance of similarity and complexity. Their values are adjusted according to many experiments and fixed during the evolution. In our experiment, w_1=0.6 and w_2=0.4 have generated better results. But we do not make sure these settings will be effective for other transcription factor binding sites motif identification.

There are many methods to calculate similarity [17, 18]. In this paper, we first get a likelihood matrix by calculating the frequency of A, T, G, and C at each column. The greatest value of each column in the likelihood matrix is subtracted from 1 and the absolute value of the result is stored. Then we calculate the sum of each column in the subtraction matrix and a difference of 1 minus the sum is got. The sum of the difference value for each column is the final similarity.

The average complexity for all windows represents the total complexity score for each individual solution. Complexity of a window can be calculated according to the following formula:

$$Complexity = \log_{10}{}^{d!/\prod n_i!} \quad (5)$$

Where n_i is the number of nucleotides of type i, $i \in \{A,T,G,C\}$. For example, n_A=1, n_T=2, n_G=1 and n_C=0 respectively correspond to the sequence 'ATTG'. In the worst case, a window sequence has only one type of nucleotides, then its complexity is equal to zero.

D. *Update and Crossover*

The velocity V of each individual is initialized in the first generation as a seven-dimensional vector and each element in the vector has the range [1, 8]. Then V can be updated according to the Eq. (2) and X can be updated according to Eq. (1). We limit the maximum and the minimum velocity for each particle in order to avoid missing the global optima or entrapping in the local optima soon. So any window in the individual can move two window widths at most one time either to the left or to the right. In addition, the position X should be modified when the particle escapes from the boundary. The pseudo code for modifying velocity and position is shown in figure 3.

```
For each particle in the population p
   Calculate V according to Equation (1);
   for k=1 to 9
   if V_ik >16 then V_ik =16; if V_ik <-16 then V_ik =-16;
   end
   end
   Update X according to Equation (2);
   for k=1 to 9
   if  X_ik <1  then  X_ik =1; elseif  X_ik >993 then X_ik =993;
   end
   end
end
```

Fig. 3. Pseudo code for the modification of velocity V and position X

From the observation of simulation experiment, we notice that there may be some duplicate individuals in each generation. These duplicate solutions can induce the premature convergence on the local optima, so it is important to maintain the variance of the population. We propose a single point crossover operation to increase the diversity of population. After updating the X value for all the individuals, two individuals chosen at random with equal probabilities from the population are executed the crossover operation in a user-defined probability. Then a random crossover point is selected from the range [1, 9]. The two individuals' information is exchanged after this crossover point. The details are shown in figure 4.

Particle I			Particle J		
Xi=(120,877,63,598,316,207,915,108,419)			Xj=(367,82,710,581,182,676,236,524,153)		
Seq. 1	(120,127)	GTGGTGGG	Seq. 1	(367,374)	ATGTTAAA
Seq. 2	(877,884)	TAGAGAGA	Seq. 2	(82,89)	TCCACTTG
Seq. 3	(63,70)	GCAGCCCA	Seq. 3	(710,717)	GTTATGGG
Seq. 4	(598,605)	ACGTGCGT	Seq. 4	(581,588)	AAGAAGAA
Seq. 5	(316,323)	CACACACT	Seq. 5	(182,189)	ACCTCTGG
Seq. 6	(207,214)	AATCCCAC	Seq. 6	(676,683)	ATCCCACT
Seq. 7	(915,922)	GGCAGGGA	Seq. 7	(236,243)	TTACCGTT
Seq. 8	(108,115)	GAAGAGTC	Seq. 8	(524,531)	GTAGGTAG
Seq. 9	(419,426)	GGTGGACA	Seq. 9	(153,160)	AGTCTGAA
Crossover point = random number from the range [1, 9] = 7					
Xi'=(120,877,63,598,316,207,**236**,**524**,**153**)			Xj'=(367,82,710,581,182,676,**915**,**108**,**419**)		
Seq. 1	(120,127)	GTGGTGGG	Seq. 1	(367,374)	ATGTTAAA
Seq. 2	(877,884)	TAGAGAGA	Seq. 2	(82,89)	TCCACTTG
Seq. 3	(63,70)	GCAGCCCA	Seq. 3	(710,717)	GTTATGGG
Seq. 4	(598,605)	ACGTGCGT	Seq. 4	(581,588)	AAGAAGAA
Seq. 5	(316,323)	CACACACT	Seq. 5	(182,189)	ACCTCTGG
Seq. 6	(207,214)	AATCCCAC	Seq. 6	(676,683)	ATCCCACT
Seq. 7	(236, 243)	**TTACCGTT**	Seq. 7	(915,922)	**GGCAGGGA**
Seq. 8	(524,531)	**GTAGGTAG**	Seq. 8	(108,115)	**GAAGAGTC**
Seq. 9	(153,160)	**AGTCTGAA**	Seq. 9	(419,426)	**GGTGGACA**

Fig. 4. Crossover operation. Two particles Xi and Xj represent a set of short sequence fragments which are extracted from the nine upstream sequences (C00097, C00099, etc.)

V. Experiment Results

The proposed IPSO-GA is run twenty times with a population size of 30 in our experiment. It is terminated when the maximum generation of 600 is arrived each time. The three parameters in the Eq. (1) are set to as follows: $c_1 = c_2 = 2$, $w = 0.6$. These parameters are tuned repeatedly with respect to the observation from the experiment results. In table 2, the solution with highest fitness in all the runs is shown. It is similar with the known sequence motifs in table 1. The different ones compared with known binding motifs are marked with bold font.

Table 2. The best solution found in all runs during evolution

Accession no	Location	Binding motif
C00097	(747,754)	GGGGATTT
C00099	(861,868)	GGGAAATT
C00100	(288,295)	GGACTTTC
C00101	(946,953)	GACTTTCC
C00152	(889,896)	GGATTTTC
C00153	(20,27)	**GAAATGCT**
C00155	(859,866)	**GAAATTCC**
C00156	(829,836)	**GAGATTCC**
C00165	(28,35)	GTAGTGAT

Other better solutions are grouped by similar binding motif for all sequences and are shown in table 3. These putative motifs for each sequence are slightly different from the known motif. It is

reasonable because one transcription factor can bind to many different sites and only part of these sites has been validated. Capturing the information included in these putative motifs is as important as capturing the known motifs. These results will be helpful to identify binding motifs in vivo.

Table 3. Putative binding sites discovered by IPSO-GA

Accession no	Putative motif	
C00099	GCTCAATT,	GGGAGAAG
C00100	GGACCCGC,	GTGATTTT
C00097,C00101	GACTCCCA,	AGTGATTG
C00152	CACTTTTC,	GGAAACTC
C00153	TAAACACA,	CAAGTTCC
C00155	GAGATTCC,	GCATTTCT
C00156,C00165	GAGCTTGC,	GACCTTAT

The problem of sequence motif discovery from coexpressed genes may be dealt with exhaustive search method if the sequence length is relative short. However, for the nine sequences with 1000 nucleotides, the search space of exhaustive method will be nearly 10^{27} (according to Eq. 3). We can get the near optimum solutions by using our algorithm in most of the runs with a small search space about 10^5 .

VI. Conclusion

In this paper, we present a computational based approach to select the most relevant information for searching binding motif from the long sequences of upstream regions. First, we demonstrate that evolutionary computation method can be used for motif identification. Then we propose a novel hybrid algorithm IPSO-GA by integrating an improved particle swarm optimization (IPSO) with genetic algorithm (GA) to search sequence motifs from coexpressed genes regulated by the NF-kb transcription factor. Experiment results show that the proposed algorithm can find the binding motifs efficiently. Some of these discovered motifs have been determined by experiment and other potential motifs are previously unknown. These putative motifs can more probably present novel binding sites that are not discovered yet. In general, the results are highly promising.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant No. 60433020 and the Key Laboratory of Symbol Computation and Knowledge Engineering of the Ministry of Education.

References

[1] F. Collins, “A vision for the future of genomics research”, in Nature, vol. 422, 2003, pp. 835-847.

[2] D. Lockhart, “Genomics, gene expression and DNA arrays”, in Nature, vol. 405, 2000, pp. 827-836.

[3] Z. Qin, “Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites”, in Nature Biotechnology, vol. 21, 2003, pp. 435-439.

[4] V. Olman, “Identification of Regulatory Binding-sites using Minimum Spanning Trees”, in Proceedings of Pacific Symposium on Biocomputing, 2003, pp. 327-338.

[5] D. Cora, "Computational identification of transcription factor binding sites by functional analysis of sets of genes sharing overrepresented upstream motifs", in BMC Bioinformatics, vol. 5, 2004.
[6] Y. Liu, "Eukaryotic regulatory element conservation analysis and identification using comparative genomics", in Genome Research, vol. 14, 2004, pp. 451-458.
[7] X. L. Liu, "Bioprospector: Discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes", in Proceedings of Pacific Symposium on Biocomputing, 2001, pp. 127-138.
[8] G. Z. Hertz, "Identification of consensus patterns in unaligned DNA sequences known to be functionally related", in Comput. Appl. Biosci, vol. 6, 1990, pp. 81-92.
[9] C. T. Workman, "ANN-SPEC: a method for discovering transcription binding sites with improved specificity", in Proceedings of Pacific Symposium on Biocomputing, 2000, pp. 467-478.
[10] V. Matys, "TRANSFAC: transcriptional regulation, from patterns to profiles", in Nucleic Acids Research, vol. 31, 2003, pp. 374-378.
[11] O. V. Kel-Margoulis, "COMPEL: a database on composite regulatory elements providing combinatorial transcriptional regulation", in Nucleic Acids Research, vol. 28, 2003, pp. 311-315.
[12] J. Kennedy, "Particle swarm optimization", in Proceedings of the IEEE International Conference on Neural Networks,1995, pp. 1942-1948.
[13] Y. Shi, "Parameter selection in particle swarm optimization", in Proceedings of the Seventh Annual Conference on Evolutionary Programming, 1998, pp. 591-600.
[14] M. Clerc, "The particle swarm-explosion, stability and convergence in a multidimensional complex space", in IEEE Transactions on Evolutionary Computation, vol. 6, 2002, pp. 58-73.
[15] Y. Shi, "Empirical study of particle swarm optimization", in Proceedings of the IEEE Congress on Evolutionary Computation, 1999, pp. 1945-1950.
[16] G. B. Fogel, "Discovery of sequence motifs related to coexpression of genes using evolutionary computation", in Nucleic Acids Research, vol. 32, 2004, pp. 3826-3835.
[17] T. K. Rasmussen, "Improved Hidden Markov Model training for multiple sequence alignment by a particle swarm optimization", in Biosystems, vol. 72, 2003, pp. 5-17.
[18] B. Chang, "Particle Swarm Optimization for Protein Motif Discovery", in Genetic Programming and Evolvable Machines, vol. 5, 2004, pp. 203-214.

Wengang Zhou is a postgraduate of Jilin University of China. His current research interests include computational intelligence and bioinformatics. He has published more than ten conference papers.

Hong Zhu is a postgraduate of Jilin University of China. Her current research interests include grid computing and swarm intelligence.

Guixia Liu is an associate professor of Jilin University of China. Her current research interests include neural networks and computational molecular biology.

Yanxin Huang is a lecturer of Jilin University of China. His current research interests include fuzzy systems, neural networks and bioinformatics.

Yan Wang is a doctor of Jilin University of China. His current research interests include bayesian networks, evolutionary computation and neural networks.

Chunguang Zhou is a professor of Jilin University of China. He received the P.H. Degree from Qiyu University of Japan. He has coauthored more than 100 papers and published one book.

Dongbing Han is a lecturer of Jilin University of China. His current research interests include pattern recognition, and image processing.