Searching Transcriptional Modules Using Evolutionary Algorithms

Je-Gun Joung[1], Sok June Oh[3], and Byoung-Tak Zhang[1,2]

[1] Biointelligence Laboratory, Graduate Program in Bioinformatics
Seoul National University, Seoul 151-742, Korea
Phone: +82-2-880-5890, Fax: +82-2-883-9120
{jgjoung,btzhang}@bi.snu.ac.kr
[2] School of Computer Science and Engineering,
Seoul National University, Seoul 151-742, Korea
[3] Department of Pharmacology, College of Medicine,
Inje University, Busan 614-735, Korea
juno@bi.snu.ac.kr

Abstract. The mechanism of gene regulation has been studied intensely for decades. It is important to identify synergistic transcriptional motifs. Its search space is so large that an efficient computational method is required. In this paper, we present the method that can search automatically both transcriptional motif list and gene expression profiles for synergistic motif combinations. It uses evolutionary algorithms to find an optimal solution for the problems which have the huge search space. Our approach includes the additional evolutionary operator performing local search to improve searching ability. Our method was applied to four *Saccharomyces cerevisiae* gene expression datasets. The result shows that genes containing synergistic motif combination from our optimization technique are highly correlated than those from k-means clustering. In cell cycle as well as other expression datasets, our results generally coincide with the previous experimental results.

1 Introduction

One of the great challenges in the post-genome era is to understand gene regulation on a genomic scale. The availability of complete genome sequences and large-scale gene expression data of many species has opened up new possibilities of understanding gene regulatory mechanisms. From this point of view, functional combination of transcription factor binding sites (TFBS) is an essential part to understanding gene regulations and a variety of computational techniques will play a key role in generating hypotheses for further study of regulatory mechanisms.

Transcriptional regulation in eukaryotes depends on the activities of hundreds of sequence specific DNA binding proteins known as transcription factors (TFs). In other words, complex expression patterns are mediated by a combination of transcription factors which bind to *cis*-regulatory regions. Each TF recognizes

X. Yao et al. (Eds.): PPSN VIII, LNCS 3242, pp. 532–540, 2004.

a specific site in the promoter region that is a unique family of short sequence elements, usually from 5 to 15 base pairs in length. Although the number of transcriptional factors are limited, the combinatorial transcriptional control makes it possible to regulate genes in response to a variety of signals from the environment.

There have been several works which identify regulatory elements by using microarray data [1][2][3]. These approaches take the following procedure. First, gene expression data is clustered to look for the co-expressed gene group. Then a motif finding algorithm scans shared patterns over the sequences. It is based on the underlying hypothesis that co-expression of genes implies common regulatory mechanism. Pilpel *et al.* [4] attempted to identify synergistic motif combinations that control gene expression patterns in *Saccharomyces cerevisiae*. Their result was obtained through exhaustive computing of expression coherence for genes containing each motif pair. Currently, one of more systemic approaches is generating of probabilistic frameworks integrated by the models of Segal *et al.* [5]. Further, there are several approaches that apply machine learning techniques such as self-organizing maps (SOM), associative clustering (AC) to this problem [6][7]. If we define this problem as an optimization problem, EAs is one of the most appropriate methods for solving it.

In this paper, our goal is to search motif combinations that importantly participate in several regulatory processes of conditions, including the cell cycle. We introduce a technique searching potential transcriptional motif combinations that affect a regulation in several experimental conditions. Our method corresponds to a searching technique through optimum search algorithms based on the mechanism of natural selection in a population and incorporating local search [8]. Here, each individual in the population represents a set of possible motif indexes in the upstream region of a gene. The fitness measure is based on the clustering for gene expression profiles containing common motifs.

For transcriptional motifs, we used the dataset of known and putative transcription regulatory motifs identified by applying motif finding algorithm in each gene of the *Saccharomyces cerevisiae* [4]. We evaluated on four expression profile datasets including cell cycle and protein interactions extracted from MIPS protein complex database [9]. Our results were compared with k-means clustering of measuring coherent motifs in four expression datasets. Also, we examined whether the algorithm recovers previous known synergistic motifs from four expression datasets.

2 Methods

The main point in search problem for motif combination is to find several motif sets correlated in different expression patterns over total expression space. Our approach is designed to perform both searching motif combinations and clustering gene expression profiles.

Let G be a gene set for the analysis. G consists of $\{g^1, g^2 \ldots, g^N\}$ with total number of genes, N. g^i given by i-th gene has the set of motif indexes, g^i_M

and the expression profile, $v(g^i)$. Here g^i_M consists of $\{m^i_1, m^i_2, \ldots, m^i_z\}$ that is a subset of total motif index set. z is the size of motif indexes for each gene and has variable size. For each gene, the set of motif indexes has different combination and there is the set of common motif indexes for several genes which is M. Let g_M be genes with this set of motif indexes. The algorithm finds best K motif combinations that make diverse gene clusters over expression profiles.

2.1 Individual Representation

Given the number of clusters K, each individual is represented as the set of strings, $S = \{s_1, s_2, \ldots, s_K\}$. Here i-th string s_i is represented as $[l,\ m_1, m_2, \ldots, m_{q_{max}}]$. Each m is the index of motif picked from total motif indexes. l is the size of actual motifs so that only motif combination is considered from m_1 to m_l and the rest are not used to measure fitness. At initial generation, each m is randomly selected from motif list. This representation provides motif combination with variable size. The size l is converged by evolutionary procedure. For K cluster, there are $l_1, l_2, \ldots, l_K$ and let L be a set of them. The motif combination used to measure fitness is $M = \{m_1, m_2, \ldots, m_l\}$. Given the number of clusters K, the aim of learning is to find a set of optimal motif combinations, $\mathcal{M}^* = \{M_1, \ldots, M_K\}^{best}$.

2.2 Algorithm Overview

The individual represents a set of motif combinations and it is a possible solution of motif combinations. These solutions are fitted as the generation goes. At the last generation, the best motif combinations are selected and it is verified with other resource such as protein complexes. Figure 1 presents the summary of algorithm for searching synergistic motif combination.

An initial population is randomly created by the population size *Pop*. Each individual I_i in the population is then evaluated using the fitness function for motif combinations $\mathcal{M}_i$. Then, selection is performed by randomly choosing a pair of individuals from the mating pool. Our selection strategy is roulette wheel selection (RWS) that is based on probabilistic selection [10]. After selection, crossover operation is performed with probability p_c. Two offsprings are produced through the exchange of genetic information between the two parent strings. In this paper, crossover operator creates two new individuals through swapping all strings with the randomly-chosen crossover point. It can be defined as $[s^a_1, s^a_2, \ldots, s^a_K] \times [s^b_1, s^b_2, \ldots, s^b_K]$. Mutation operators were performed by choosing a mutation point with a low probability. They flipped the bit at that point for the actual motif size. Then hill-climbing search is performed for motif indexes. The above steps are repeated until a termination condition is reached. Otherwise, it stops if the number of generations reaches maximum generation.

2.3 Fitness Function

The fitness of each individual is defined by

$$Fitness = \alpha EC + S. \tag{1}$$

```
Procedure TranscriptionalModuleSearching($K$, $Pop$)
  $K$: number of clusters
  $Pop$: size of population
begin
 Initialize population: $I_1, \ldots, I_{Pop}$.
for motif combinations $\mathcal{M}_i$ of each individual $I_i$ do
      fitness evaluation
end
$t := 0$
while (not termination condition) do
      for $i := 1$ to #recombinations do
            Select two parents $I^a$, $I^b$ randomly
            Crossover motif combination sets $S^a$, $S^b$
            Add offsprings $I^{a'}$, $I^{b'}$ to next population
      end
      for $i := 1$ to #mutation do
            Select an offspring $I^c$ randomly
            Mutate size of actual motif $L^c$
            Local search for motif combinations $\mathcal{M}^c$
      end
      $t := t+1$
end
```

Fig. 1. Summary of the algorithm.

EC (expression coherence) represents how well the related genes are clustered in expression space and S (separation) explains how several groups are separated each other. In other words, the fitness induces clustering of genes in expression space. Here α is a parameter to control trade-off of two terms. If α is highly weighted, it emphasizes condensation of each gene group, otherwise it is emphasis on the separation among groups.

When M_k is the k-th motif combination in a certain individual of the algorithm and g_{M_k} defines the genes containing M_k, EC is defined as

$$EC = \frac{1}{K} \sum_{k=1}^{K} C(g_{M_k}), \tag{2}$$

where $C(g_{M_k})$ is a mean of correlation coefficients as follows:

$$C(g_{M_k}) = \frac{1}{P} \sum_{i=1}^{J_k} \sum_{j=i+1}^{J_k} r(v(g^i_{M_k}), v(g^j_{M_k})). \tag{3}$$

Where r is the similarity between gene pair over expression profiles. It is measured by the Pearson correlation coefficient. $v(g^i_{M_k})$ indicates the expression profile of i-th gene containing M_k. P is the total number of possible gene pairs, $(J_k^2 - J_k)/2$. J_k is the total number of genes in k-th group.

EC is the similarity measure among gene expression profiles in the gene group containing the motif combination. On the other hand, S is the dissimilarity measure among K groups that contain each expression profiles. S is defined as follows:

$$S = \frac{1}{P} \sum_{i=1}^{K} \sum_{j=i+1}^{K} d(\hat{v}(g_{M_i}), \hat{v}(g_{M_j})). \tag{4}$$

Where $\hat{v}(g_{M_i})$ is a mean of expression profiles of genes containing motif combination M_i and the distance d is $1 - r$.

3 Experimental Setup

Our method was applied to the following datasets. Motif dataset contains yeast motif information extracted by Pilpel [4]. This motif dataset consists of 37 known motifs and 329 putative motifs that are upstream DNA promoter motifs obtained using the AlignACE program. We generated dataset by checking an occurrence of motifs for each gene.

To verify our method, we used the result of microarray analysis of 800 ORFs involved during the yeast cell cycle [11]. This dataset was used to examine a characteristic of our method in synergistic effect of motifs. As another test for synergistic effect, algorithm was additionally applied to sporulation, heat-shock and diauxic shift expression data for over 6000 ORFs [12].

The parameter setting of algorithm is as follows. The size of individuals is 100 and the maximum generation is 200. Crossover probability p_c is 0.9 and mutation probability p_m is 0.01. In reproduction, we use elitist selection. Elitism ensures that at least one copy of the best individual in the population is always passed onto the next generation. For local search, hill-climbing was performed during 30 iteration. The maximum size of motif combination in representation, l_{max} set 5. In the fitness function, α is set to 0.8 and it is heuristic value from the repeated runs. The size of cluster K sets 5.

4 Results

First, we examined the learning effect of the algorithm from expression profiles of 800 ORFs associated with cell cycle and 37 known motifs. Figure 2 shows the result of comparison between our method and k-means clustering for expression coherence distribution. The curve represents the distribution of pair-wise correlations. The distribution for our result is strongly shifted to the right. This figure explains that the gene pair obtained by our algorithm is highly correlated in expression.

We tested correlations between genes containing motif combination from our method. k-means clustering algorithm was used as a baseline algorithm for comparison. Figure 3 shows the comparison between our method and k-means clustering for the protein interaction ratio. Our test is based on investigation of the

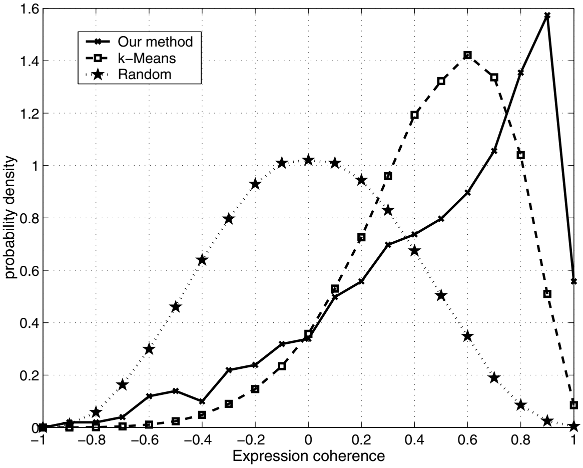

Fig. 2. The comparison between our method and k-means clustering for expression coherence distribution (k=30).

relationship between expression profiles and protein-protein interactions [13][14]. Their results showed many interactions between proteins encoded by genes in the same cluster over expression or subunit.

We define the protein interaction score as the correlation measurement of genes in the same group. The list for yeast protein interactions was extracted from MIPS protein complex database. The interaction list was extracted from 971 complex entries and the size is 92068. Protein interaction score is calculated as the difference between $PID(M)$ and $PID(R)$ as follows.

$$PI\ score = log(PID(M)/PID(R)) \tag{5}$$

PID is protein interaction density (PID) which is the ratio of the number of observed protein interaction pairs to the total number of possible pair-wise combinations of protein pairs. According to analysis of correlation between transcriptome and interactome, clusters with a significantly higher PID tend to be tighter in expression profiles [13].

$PID(M)$ is protein interaction density in the genes containing motif combination M. $PID(R)$ is protein interaction density in random set. At k-means clustering, the number of cluster k set 30. After k-means clustering algorithm was performed, the gene set containing motif combination correlated with cluster showed lower protein interaction ratio than set from our method.

Finally we present synergistic motif combination for motif dataset in four different expression profile datasets. Figure 4 shows the result of the putative synergistic motif combinations searched by our algorithm. Motif combinations given from our method are significantly related with particular biological condition.

The result from the cell cycle data presents well known MCB and SCB motif combination as well as several combination. SBF (SCB-binding factor) and MBF

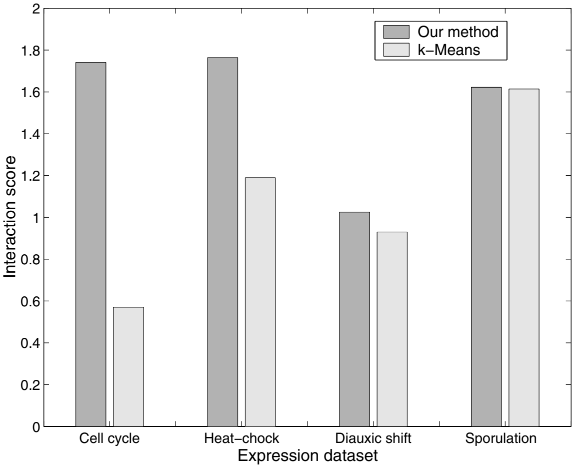

Fig. 3. The comparison between our method and k-means clustering for protein interaction score (k=30). This score indicates the measure of the correlation in specific gene group.

(MCB-binding factor) regulate genes necessary for the transition through G1/S [15]. These two are sequence-specific transcription factors and the two regulatory elements (MCB and SCB) are well known synergistic motifs. These motifs lead to combinatorial control and synergistic effects.

In the sporulation, m155 is MSE which is bound in specific transcriptional factor Ndt80 [16]. Ndt80 is strongly induced during the middle stages of the sporulation pathway. It is important to perform sporulation procedure. Moreover, MSE is related to about 150 genes through meiosis. In the heat-shock dataset, several motifs (RAP1, m301, m303, 306 and 308) are RPE (ribosomal protein element) which is negatively correlated to this condition.

5 Conclusions

We have proposed an effective method to search for the motif combination of genes that are co-regulated in the set of experiments. Our approach is based on the evolutionary algorithm known to be highly effective for several combinatorial optimization problems. It is designed to perform both searching motif combinations and clustering gene expression profiles. Through clustering, our method prevents the generation of similar motif combinations.

Our results show that the method can find several significant motif combinations from highly coherent gene expressions. Genes with motif combination found by our approach were highly correlated than those which are obtained by previous approach. Consequently, the comparison for coherence supports that the searching result has high confidence for finding synergistic motif combinations. Our approach can be used to study the regulatory mechanism at the genome level.

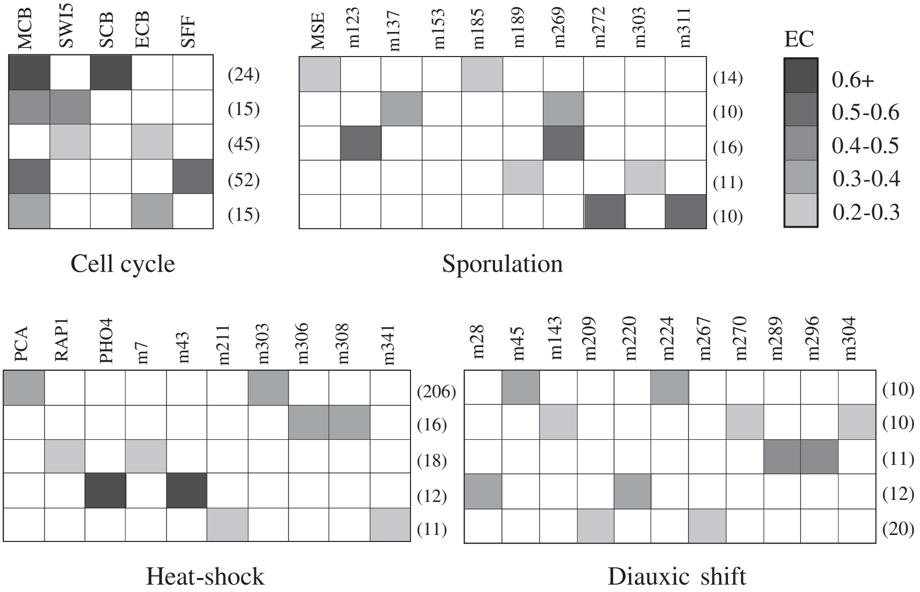

Fig. 4. The putative synergistic motif combinations in several conditions. In each grid, a gray square indicates that particular motif is present in gene set. In top of the grid, motifs are listed and the value on the right side shows the number of genes in each cluster.

As the future works, we consider that the method is improved in several parts. Our searching strategy will be tested by the diverse local searches such as simulated annealing. Also, we consider a co-evolutionary searching technique in an aspect of efficient optimization. The co-evolutionary searching technique can make search both sides. The algorithm will try to find motif combinations in one side as well as clustering gene expressions in the other side.

Acknowledgments

This research was supported in part by NRL, the Korean Ministry of Science and Technology (Korean Systems Biology Research Grant, M10309000002-03B5000-00110).

References

1. S. Sinha, M. Tompa, A statistical method for finding transcription factor binding sites, *Proc. Int Conf. Intell. Syst. Mol. Biol.*, Vol. 8, pp. 344–354, 2000.
2. S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church, Systematic determination of genetic network architecture, *Nature Genetics*, Vol. 22, pp. 281–285, 1999.

3. A. Brazma, I. Jonassen, J. Vilo and E. Ukkonen, Predicting gene regulatory elements *in silico* on a genomic scale, *Genome Research*, 8(11), pp. 1202–1215, 1998.
4. Y. Pilpel, P. Sudarsanam and G. Church, Identifying regulatory networks by combinatorial analysis of promoter elements, *Nat. Genet.*, Vol. 29, pp. 153–159, 2001.
5. E. Segal, R. Yelensky and D. Koller, Genome-wide discovery of transcriptional modules from DNA sequence and gene expression, *Bioinformatics*, Vol. 19, pp. i273–i282, 2003.
6. J. Kasturi, R. Acharya, Clustering of diverse genomic data using information fusion, *2004 ACM symposium on Applied computing*, pp. 116–120, 2004.
7. J. Sinkkonen, J. Nikkila, L. Lahti and S. Kaski, *Associative Clustering by Maximizing a Bayes Factor*, Technical Report A68, Helsinki University of Technology, 2003.
8. P. Moscato, *On Evolution, Search, Optimization, Genetic Algorithm and Martial Arts: Towards Memetic Algorithms*, Technical Report C3P Report 826, Caltech Concurrent Computation Program, California Institute of Technology, 1998.
9. H. W. Mewes, et al., MIPS: a database for genomes and protein sequences, *Nucleic Acids Res.*, 30(1), pp. 31–34, 2002.
10. B. L. Miller and D. E. Goldberg, *Genetic Algorithms, Selection Schemes and the Varying Effect of Noise*, IlliGAL report, No. 95009, 1995.
11. P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein and B. Futcher, Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization, *Molecular Biology of the Cell*, Vol. 9, pp. 3273–3297, 1998.
12. M. B. Eisen, P. T. Spellman, P. O. Brown and D. Botstein, Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci.*, USA, Vol. 95, pp. 14863–14868, 1998.
13. H. Ge, Z. Liu, G. Church and M. Vidal, Correlation between transcriptome and interactome mapping data from *Saccharomyces cerivisia*, *Nature Genet.*, Vol. 29, pp. 482–486, 2001.
14. R. Jansen, D. Greenbaum and M. Gerstein, Relating whole-genome expression data with protein-protein interactions, *Genome Res.*, Vol. 12, pp. 37–46, 2002.
15. I. Simon, J. Barnett, N. Hannett, C. T. Harbison, N. J. Rinaldi, T. L. Volkert, J. J. Wyrick, J. Zeitlinger, D. K. Gifford, R. S. Jaakkola and R. A. Young, Serial regulation of transcriptional regulators in the yeast cell cycle, *Cell*, Vol. 106, pp. 697–708, 2001.
16. S. L. Jason, S. David, T. Roger, W. Cynthia and J. N. M. Glover, Structure of the sporulation-specific transcription factor Ndt80 bound to DNA, *EMBO*, 21(21), pp. 5721–5732, 2002.