ELSEVIER

Computer Methods and Programs in Biomedicine 65 (2001) 163–174

Computer Methods and Programs in Biomedicine

www.elsevier.com/locate/cmpb

An integrated approach for genome-wide gene expression analysis

Yuh-Jyh Hu *

Department of Computer and Information Science, National Chiao-Tung University, 1001 Ta Hsueh Rd., Hsinchu, Taiwan, 300, ROC

Received 8 December 1999; received in revised form 23 May 2000; accepted 21 June 2000

Abstract

Since efficient and relatively cheap methods were developed for determining biosequences, a lot of biosequence data has been generated. As the main problem in molecular biology is the analysis of the data instead of the data acquisition, part of the study of computational biology is to extract all kinds of meaningful information from the sequences. Computer-assisted methods have become very important in analyzing biosequence data. However, most of the current computer-assisted methods are limited to finding motifs. Genes can be regulated in many ways, including combinations of regulatory elements. This research is aimed at developing a new integrated system for genome-wide gene expression analysis. This research begins with a new motif-finding method, using a new objective function combining multiple well defined components and an improved stochastic iterative sampling strategy. Combinatorial motif analysis is accomplished by constructive induction that analyzes potential motif combinations. We then apply standard inductive learning algorithms to generate hypotheses for different gene behaviors. A genome-wide gene expression analysis demonstrated the value of this novel integrated system.

Keywords: Motifs; Gene regulation; Constructive induction; Decision tree

1. Introduction

The completion of genome sequences, e.g. *Saccharomyces cerevisiae* [1], has produced sufficient data for biological analysis, and the advent of the microarray and the genechip technology [2,3] has resulted in the expanded ability of monitoring simultaneously the levels of mRNA from each gene in a cell. Because this type of experiment provides comprehensive information about regulation of mRNA levels of every gene, it becomes theoretically possible to identify sets of genes that are similarly regulated under a given condition. This allows, (1) inferences about functions of unknown genes that are co-regulated with genes of known functions; (2) discovery of regulatory motifs and combinations of motifs; and (3) greatly improved understanding of the biology of the cellular response to particular environmental

* Tel.: +886-3-5712121.
E-mail address: yhu@cse.ttu.edu.tw (Y.-J. Hu).

0169-2607/01/$ - see front matter
PII: S0169-2607(00)00126-7

stimuli. In order to realize the potential of information emerging from such global gene regulation studies, data from multiple genes must be analyzed in parallel.

As the advance of the microarray and the genechip technology, two experimental methods of monitoring complete yeast genome expression have been described in the literature; the system developed in the laboratory of P. Brown [2] and the Affymetrix GeneChip system [3]. This technology provides a global view of changes in gene expression on a genomic scale. As more is learned about the functions of every gene in the entire genome, we have the ability to gain insight into a cell's response to a changing environment through its global gene expression patterns. Potential applications include predicting drug interaction or drug resistance, exploring the immune system, etc.

It is useful to obtain distinct temporal patterns of gene expression, knowing how genes behave differently under a certain condition. Nevertheless, in addition to the macro view of the changes in gene expression level, biologists are also interested in the following questions, e.g. (1) is there any consensus motif (i.e. pattern) in a given family of genes; (2) if there exist several motifs commonly shared by a set of genes, what is the correlation among them in terms of number of repeats and locations, etc.; and (3) what makes these genes behave differently in the same environment, i.e. why some are up-regulated by the stimuli, but some have no change?

To learn beyond how genes behave over time, it would be valuable if we understand what actually makes these genes behave differently. One way to find possible answers to the questions is to look into the regulatory regions of genes at the sequence level. In this paper, we introduce a method for gene regulation analysis from a different point of view.

2. Background

Biologists have traditionally studied the regulation of genes selected for a particular type of activity or function. Although this approach has allowed the identification of regulatory proteins that affect the expression of those genes, it has tended to focus attention on specific sets of genes and inferences concerning the regulation of those genes that have by necessity been extended to less well understood genes. It is likely given the number of genes for which no function is known (estimated at one third of the yeast genome), that regulatory proteins remain to be identified. In addition, due to the relatively intensive study of exemplary genes, certain aspects of regulation, including positional effects, multiplicity of regulatory motifs, orientation of motifs and the role of combinations of different motifs, although appreciated conceptually, have not been explored comprehensively. Emerging knowledge of genome-wide gene activity, combined with the algorithms to infer motifs and to correlate activity and motifs, could broaden our understanding of gene regulation into under-explored areas.

Fundamentally gene regulation is determined by chemical reactions which are, in turn, controlled by the shape and physico-chemical properties of the molecules involved. One instance of this is the interaction between regulatory proteins and their target binding site. Unfortunately this information is not typically available. We expect that the local shape of a binding or receptor site will be primarily determined by the bases involved, acknowledging the fact that non-local base changes can affect local shape. The difficulty of finding the biologically meaningful motif results from the variability in (1) the bases at each position in the motif, (2) the location of the motif in the sequence, (3) the multiplicity of motif occurrences within a given sequence, and (4) the orientation of the motif in the sequence. In addition, the short length of many biologically significant motifs and the fact that motifs often gain biological significance only in combinations, make them difficult to determine using standard statistical methods.

A general framework for these computational methods could be defined on the objective function we choose. The purpose of an objective function is to approximate the correlation between sequence patterns and their biological meanings in terms of a mathematical function. The objective functions are only heuristics. Once the objective

function is determined, the goal is solely to find those patterns of high objective function value. To reach this goal, two important associated issues are the pattern representation and the search strategy.

As the primary DNA sequences are described by a double-stranded string of nucleic bases, the most basic pattern representation is the exact base string. Due to the complexity and flexibility of the motif binding mechanism, there is rarely any motif that can be exactly described by a string of nucleic bases. To obtain more flexibility, other more expressive representations have been developed such as the IUPAC code, the position weight matrix (PWM) and the hidden Markov model (HMM).

Besides those uncertain factors mentioned earlier, the search space of the patterns is also determined by the size of the given set of sequences and the length of the patterns of interest. A systematic search through the sequences for all possible patterns can reach the best results; however, as the search space could be computationally intractable, an exhaustive search strategy could only deal with data sets of relatively small size and detect shorter patterns. By random sampling and iterative improvement, a stochastic approach, on the other hand, could avoid computational explosion, but may not guarantee the optimum results.

3. Design considerations

The current motif detecting approaches alone are not sufficient for the analysis of gene regulation due to the inherent complexity of gene control which often involves motif interactions. We thus propose a novel view of the gene regulation analysis. With the assistance of the genechip technology, genes can be grouped into families according to different temporal patterns. Our analysis method for gene regulation focuses on the search for significant motifs and their combinations involved in the regulation as well as potential hypotheses regarding how the gene regulation is related to the motifs. This type of analysis not only complements the global study of the changes of gene expression by looking into the involvement of the motifs in regulation, but may also suggest further biological tests on the genes through the inference from the hypotheses produced by the analysis.

The aim of our gene regulation analysis is to extract the regularity from the DNA sequences. However, the simple but uninformative string-based representation shows very little valuable information by itself. Meanwhile, the goal of constructive induction is to transform the original representation space into a new one where the regularity is more apparent. As a consequence, these two paradigms become a perfect match.

There are four basic steps in our method of analyzing gene regulation.

1. The first step of analyzing the sequence data is to categorize the genes into families according to their expression patterns. Then by multiple sequence comparison and alignment we could find some consensus information in a given family. Patterns, also called motifs, common to multiple sequences are related to molecular structures, functions and evolution [6]. We use the Affymetrix GeneChip system to collect the data of genome-wide gene expression changes corresponding to specific stimuli under a controlled environment in a fixed period of time. According to the gene expression changes over time, genes can be clustered into families for further analyses.
2. For any family of genes of interest, we extract the control region for each gene and apply some motif-finding algorithm to the family to find significant motifs. As discussed earlier, several methods have been developed for detection of patterns shared by a set of functionally related biosequences [5–15]. In this paper, we introduce a new motif-finding algorithm called detecting motifs from sequences (DMS), and apply it in our gene regulation analysis. A particular challenge of finding regulatory motifs is that they can be quite short and thus not statistically distinctive per se. As genes can be regulated in many ways, redundancy, location and combinations have to be considered to distinguish regulatory motifs. Finding motifs alone is not sufficient for the analysis of gene

regulation. The motifs found will serve as the building blocks for representational transformation in the next step. The change of representation is necessary for revealing more regularity originally implicit in the sequence data.

3. Based on the information of the motifs found by DMS, we transform the original sequences into a higher-level representation. It includes (1) the locations of the motifs, (2) the total number of repeats of each motif, (3) the number of repeats of each motif within a selected location range, (4) the distance between motifs, and (5) combinatorial motifs as Boolean combinations. From the point of view of constructive induction, the objective of this step is to transform the raw string-based representation into a new one for better understanding and additional analyses. The new representation is used to manifest the regularity originally implicit in the raw string-based data. All the information described by the new representation can be either used as a whole or partially used for further analyses.
4. Finally, after the raw sequence data are redescribed in an appropriate higher-level representation, we can apply a suitable standard inductive learning algorithm on the data to generate hypotheses. There are many inductive learning algorithms available. Each has its own advantages and limitations. In our experiments, we applied a decision tree learning algorithm to construct hypotheses represented as decision trees. The hypotheses are easily understood and provide a micro view of the families as they suggest reasons for different behaviors of the genes to complement the macro view of the genome-wide gene expression changes.

4. System description

In this section, we will detail the new motif-finding algorithm, DMS, explain how the original string-based DNA sequences are transformed into higher-level representation, and describe how to apply a decision tree learning algorithm to derive comprehensible hypotheses.

4.1. DMS: detecting motifs from sequences

As the sequence segments, such as binding sites for a particular protein, are generally not accurately represented by a single consensus sequence pattern, some positions are more conserved than others and the preference for each of the four nucleic bases can be different. Thus, we adopt the weight matrix as our motif representation. By running an iterative sampling optimization process, DMS outputs a user-specified number of motifs. This number is only used to determine how many motifs will be reported in the output, i.e. the user has the option to see all possible motifs or only part of them. It does not affect the process of the algorithm in any way.

The weight matrix method approach has been used in various pattern-identification problems [6–9,14]. It is usually built from the base frequency of example biosequences. If we divide every element in the matrix by the total number of sequences, we get a normalized matrix.

Based on the normalized motif matrix, we can calculate the match score of any 6-base sequence by dividing the sum of the value for each position by the width of the motif. The success of these analyses confirms the fact that the frequencies of bases at positions within sites are related to the importance of the bases to the functioning within the sites [4]. The challenge is to find a matrix that well represents the motif in terms of the objective function.

We propose a new motif-finding algorithm, DMS. Unlike other approaches, DMS uses a new type of objective function that consists of multiple components. They are the motif consensus quality, the motif multiplicity significance and the motif coverage. The consensus quality is only used to guide the search for well conserved motif candidates, the motif multiplicity significance reflects the value of multiple copies of a single motif, and the motif coverage addresses the importance of a motif's being commonly shared by a given family of sequences.

The consensus quality of a matrix is derived from the entropy [16]. The lower the entropy, the better conserved the motif. The entropy is calculated from

the probability that each base occurs at each position in the motif, P_m. More precisely, the entropy for a particular column n in the matrix is given by:

$$E(n) = -\sum_{i=b1}^{b4} P_{mi} \log_2 P_{mi}$$

where $b1$, …, $b4$ are the bases A, G, C, T. If the bases are uniformly distributed over a position, then the maximum value of 2 is obtained. If only a single base appears in a position then the minimum value of 0 is obtained. Thus, we define the consensus quality of column n as:

$$C(n) = 2 - E(n)$$

The final consensus quality of a matrix b, is defined as the average of all position quality.

$$\mathrm{Con}(b) = \frac{1}{w}\sum_{1}^{w} C(n)$$

where w is the width of the motif.

The multiplicity significance is derived from the measure of precision as defined in the information retrieval paradigm. It is simple and empirically effective. We define the multiplicity significance of a motif b as:

$$\mathrm{Mul}(b) = \frac{\mathrm{occ}_S(b)}{\mathrm{occ}_G(b)}$$

where $\mathrm{occ}_S(b)$ is b's occurrences in a given family S, and $\mathrm{occ}_G(b)$ is b's occurrences in genome.

The motif coverage is defined as the ratio of the number of the sequences containing b to the total number of sequences given.

$$\mathrm{Cov}(b) = \frac{\mathrm{cont}_S(b)}{|S|}$$

where $\mathrm{cont}_S(b)$ is the number of sequences in S that contain b, and $|S|$ is the total number of sequences in S.

Given a set of N biosequences, DMS carries out an iterative improvement search that attempts to find all potential motifs, e.g. d matrices, which maximizes the consensus quality defined above. These d matrices are motif candidates. DMS then ranks these motifs according to a merit measure based on the combination of the consensus quality, the multiplicity significance and the motif coverage. Given the d motifs, we first normalize the consensus quality, the multiplicity significance and the motif coverage of each motif b, using the maximum value, as defined below:

$$\mathrm{Con}_{\mathrm{norm}}(b) = \frac{\mathrm{Con}(b)}{\mathrm{MAX(Con)}}$$

$$\mathrm{Mul}_{\mathrm{norm}}(b) = \frac{\mathrm{Mul}(b)}{\mathrm{MAX(Mul)}}$$

$$\mathrm{Cov}_{\mathrm{norm}}(b) = \frac{\mathrm{Cov}(b)}{\mathrm{MAX(Cov)}}$$

where MAX(Con) is the maximum consensus quality of the d motifs, MAX(Mul), the maximum multiplicity significance of the d motifs, and MAX(Cov), the maximum motif coverage of the d motifs.

Combining all the objective functions introduced above, we propose the final merit measure of a motif b as defined below:

$$\mathrm{Merit}(b) = \frac{1}{1/3(1/\mathrm{Con}_{\mathrm{norm}}(b)) + 1/\mathrm{Mul}_{\mathrm{norm}}(b) + 1/\mathrm{Cov}_{\mathrm{norm}}(b))}$$

The value of merit is in the range between 0 and 1. It reflects the synergy of the consensus quality, the multiplicity significance and the motif coverage.

There are three steps in DMS that are detailed in the following subsections.

4.1.1. Translating subsequences into matrices

As the motif location(s) is unknown, we begin by allowing each subsequence of length w to be a candidate motif. Like most current algorithms, the length w is specified by the user. We convert this particular subsequence into a probability matrix in two steps, adopting an idea from [9]. First we fix the probability of every base in the subsequence to some value $0 < X < 1$, and assign probabilities of the other bases according to $(1 - X)/(4 - 1)$ (i.e. 4 nucleic bases). Following Bailey and Elkan, we set X to 0.5. This gives us a set of seed probability matrices to be used as starting points for iterative improvement. For a given family of sequences, we can either exhaus-

tively translate every subsequence into a matrix for analysis or we can select a random subset of the sequences and only generate candidate starting points from this subset. Because significant motifs are generally well conserved and thus occur in most sequences, this subsetting strategy is effective without losing generality.

4.1.2. Filtering possible motif occurrences

Rather than making the common assumption that each motif occurs only once per sequence, we allow for the possibility that a motif may occur multiple times in a single sequence. For each matrix and each sequence, we find the position that maximizes the match score and put it in the list of potential motif positions. Then we set the threshold for deciding if a motif occurs at any position as the mean of match scores. Finally we add to the list of motif positions any other position whose match score is greater than this threshold. Occurrence overlap is allowed. This process defines a set of potential motif positions.

4.1.3. Finding and ranking motif candidates

After the likely motif positions are determined, DMS performs an iterative optimization procedure to find the motif probability matrix. Unlike current approaches, such as the Gibbs sampler, that search all possible positions within a sequence, DMS only considers the potential motif positions determined in the previous step. This strategy significantly constrains the search space. For initialization, DMS randomly selects a position from the set of potential motif positions that are determined in the previous step to form the initial probability matrix.

A sequence is then chosen at random for optimization. DMS optimizes the consensus quality of the matrix by checking every potential motif position within the selected sequence. The position that gains the highest consensus quality is chosen to update the matrix. The process is repeated until no improvement is noted. In each optimization cycle, the order of sequences is randomly shuffled. The randomization in each trial cycle is important to remove implicit biases, such as the order of the sequences that can be harmful in search algorithms. At this point, in each sequence, the subsequence that contributes to the last updated matrix is determined. We then compute the mean of the match scores of the subsequences that form the matrix, and isolate all subsequences with a match score over the mean as possible motif repeats in each sequence. All these motif repeats in sequences are used to form the final motif matrix, and it becomes a motif candidate.

The same procedure is performed on all matrices to produce the candidate motifs. Finally, DMS ranks the candidate motifs according to its merit measure.

4.2. Representational transformation

Two types of additional motif information become available after DMS detects motifs from the given family of sequences, the motif repeats and the motif locations. Based on the information available, we can transform the original sequence data into a higher-level representation as a vector space. This vector representation was chosen because it is the most widely used representation for standard inductive learners in the machine learning community. It increases the applicability of machine learning techniques. Each sequence is transformed into a vector that contains the motif information associated with the sequence, including the number of motif repeats in the entire sequence and the number of motif repeats within a selected segment of the sequence. Given only one family of sequences, the particular segment is selected based on the background knowledge, i.e. specified by the user. If two or more families of sequences are provided, the sequence segment can be either specified by the domain expert or determined by DMS. It is computationally prohibited to find the optimal segment that gains the maximum discrimination between families by checking all possibilities. Therefore, we divide the sequences into equal intervals. For each interval, we compute the information gain according to the number of motif repeats in that interval. DMS thus selects the interval that attains the highest information gain.

For example, assuming three motifs are found, M_1, M_2 and M_3, an original nucleic sequence can be represented as a vector, (M_{total_1}, M_{total_2},

M_{total_3}, $M_{100-150\ bp\_1}$, $M_{300-450\ bp\_2}$, $M_{50-100\ bp\_3}$). The first three elements are the total number of repeats of M_1, M_2 and M_3. The fourth element presents the number of repeats of M_1 within the range 100–150 bp in the upstream region of the sequence. The last two elements present the number of repeats of M_2 and M_3 within the range 350–400 and 50–100 bp, respectively. For instance, a sequence can be transformed into (5, 2, 3, 3, 1, 2). This means the sequence has totally five copies of M_1, two copies of M_2 and three copies of M_3. There are three copies of M_1 located in 100–150 bp upstream of the sequence, one copy of M_2 in 350–400 bp upstream, and two copies of M_3 in 50–100 bp upstream.

Given multiple families of genes, after the transformation, the original sequence data is represented as sets of vectors. These vectors are used as the training examples for GALA [17,18] to further analyze motif combinations. From the point of view of GALA, each element of a vector is a primitive attribute. The purpose of applying GALA here is to find combinations of attributes as new attributes to improve the quality of the hypotheses that will be later generated by a standard inductive learning algorithm. GALA applies Boolean operators to construct new attributes represented as Boolean combinations. Boolean combinations are understandable. Comprehensibility allows domain experts to explore the new attributes either for further improvement or for justification.

4.3. Hypothesis generation

The final hypothesis for different gene behaviors in the same environment is produced by the standard inductive learning algorithm C4.5 [17]. The input to C4.5 is a set of feature vectors transformed from the original sequence data combined with the combinatorial motifs, i.e. Boolean combinations of motifs, generated by GALA. With the input as the training examples, C4.5 produces a classification hypothesis that could be used to explain why these families of genes behave differently under the same condition as well as suggest further biological tests on the genes.

5. Status report

In this section, we first report the experimental results of DMS on ten regulatory families in yeast genome. Second, we show the analysis results of the yeast genome gene expressions under the oxidative stress.

5.1. Finding motifs in real regulons

Yeast metabolism has been widely studied, and in some cases the transcription factors involved in the regulation of members of a common pathway are known. Those families of co-regulated genes provide ideal data sets on which to test the systems designed to detect regulatory motifs.

From the study of the literature, van Helden et al. defined ten families of genes that have known common regulatory site(s) or motif(s). There are many additional motifs involved in regulation generally, but the known ones in these regulons define ten learning tasks for comparing the various algorithms. It is assumed for this exercise that the regulation of a gene is determined by motifs in the upstream region. The 800 bp upstream region was used for each gene, as this is the same sized region used by van Helden et al. in their experiments.

There are two parameters used by DMS. One is the motif width, and the other is the random subset size (refer to Section 4.1.1). To maintain consistency, for those families with more than ten members, we set the subset size to be 10; otherwise, we set the subset size to be equal to the family size. The motif width is set to that of the published motif in each family. Except for these two parameters, we did not tune DMS or modify the sequence data in any way, e.g. by pre-specifying the expected number of motif matches/occurrences. To test the stability of DMS, we ran DMS on each family five times, using different random seeds. The results showed that DMS identified all the published motifs in all regulatory families in each run. By 'identified' we mean that the published motifs are found and ranked in the top 40 motifs according to the merit measure as defined earlier. Most of the published motifs are ranked the top except for some weak motifs or short

motifs, e.g. the motif in HAP family, CCAAY, is ranked the 34th, and the less conserved motif in PHO family, GCACGTTTT, is ranked the 18th. We also varied the subset size (e.g. 15, 20, 25) to check its effect, and found that the results of different runs were similar.

5.2. Analysis of global gene expression

Regulation of the stress response occurs at many levels including transcriptional, translational and post-translational mechanisms. These forms of stress are sensed by different signaling systems, and a signal in each case is transduced through a MAP kinase regulatory cascade that activates specific sets of transcription factors [20,21]. Because the stress response is highly conserved and there is a base of available information against which the results can be evaluated, several stress induction experiments have been performed, using the Affymetrix GeneChip system.

In the Affymetrix GeneChip system, the yeast gene expression probe array interrogates over 6200 yeast genes using 20 complementary 25-mers per gene. These are arrayed at high concentrations on four silicon chips, together totaling an area of approximately 1^2 in. [3]. The oligonucleotide probes are synthesized in situ using a photolithography process and sequential rounds of masking, photo-deprotection, and synthesis. Specificity of hybridization is internally controlled by hybridization of a set of antisense oligonucleotide probes arrayed at neighboring positions on the chips. PolyA RNA samples are converted to double-stranded cDNA and transcribed into tagged RNAs in the presence of biotinylated precursors, providing a 20–200-fold amplification. RNA species are fragmented and hybridized to the microarrays. Bound, biotinylated RNAs are stained with streptavidin–phycoerythin conjugate. The Affymetrix system reads out results using a scanning confocal fluorescence microscope with an argon laser light source, and the Affymetrix proprietary software is used to process the image file, resulting in the values for each gene proportional to the level of expression of the gene.

The stress responses in yeast are mediated by a network of interacting pathways [20]. The control by different types of stress response genes is complicated by the response of sequence elements to multiple types of stress, and by the presence of multiple, degenerate motifs mediating those effects. Analysis of the oxidative stress response using genome-wide gene expression is expected to result in the identification of candidate regulatory motifs and proteins which can be tested and further our understanding of this important cellular defense mechanism. For the oxidative stress experiment, cells were stressed according to the standard protocols with H_2O_2. Total RNAs were prepared from cells, PolyA RNA isolated, cDNA synthesized, and biotinylated cRNA probes produced by in vitro transcription. Different RNAs were used to hybridize to the yeast gene probe array. Gene expression level changes were measured at 0, 5, 10 and 20 min. The results showed that approximately 500 genes changed in transcriptional level by greater than two folds over the time course. Sets of genes that behave similarly in terms of expression level changes over a time course are likely to have common regulatory elements. Thus, the first step is to identify sets or clusters of genes that are similarly regulated. Genes were described as vectors of expression levels, and those behaved similarly over the time course were grouped automatically using a variation on the standard k-means clustering algorithm. With this clustering algorithm, we clustered the entire yeast genome into families according to how their expression changes over time. Two families were selected for further analyses. In one of the families there were 58 up-regulated genes under oxidative stress whose final expression level exceeds 1000; in the other there were 100 genes whose expressions remained nearly unchanged over time at the average expression level of 50. Certainly, there are other clusters also worth further exploration. Due to the space limit, in this paper, we only show one example to demonstrate the value of the integrated analysis system.

After we extracted a 500 bp upstream region from each gene, we applied DMS to these sequences to find candidate motifs. As we focused on short motifs in our current study, for DMS we set the motif width to be 7 and 5. According to

the merit value, 25 interesting motifs were selected. In the 25 motifs, we found two known motifs, the YAP binding site and the stress element. Based on these selected motifs, we transformed each raw DNA sequence into a vector representation. Each vector indicates for each motif the number of total motif occurrences in a sequence, and the number of motif occurrences within a specific range in that sequence. Note that the motif occurrences are determined by DMS with the mean match score as the threshold as explained earlier. From the point of view of inductive learning, after the transformation, we have a set of pre-classified data, i.e. oxidative stress up-regulated genes and no-change genes. We first applied GALA [17,18] to the new data set to analyze motif combinations, and used the inductive learning program C4.5 [19] to generate a hypothesis, represented as the decision tree shown in Fig. 1. The output of GALA is a list of combinatorial motifs represented as Boolean combinations. These Boolean combinations will be used by C4.5 to construct a decision tree hypothesis. For example, node 1, 2 and 4 described as Boolean combinations in the decision tree shown in Fig. 1 are part of the output of GALA. We modified the original output of C4.5 by directly putting in the motif Boolean combinations learned by GALA to increase readability. The decision tree describes features that are found and not found in genes that are positively or not regulated by the oxidative treatment. This description can be applied to other genes in order to predict their behavior under oxidative stress.

There are four condition nodes in the hypothesis. Each node describes a condition with two outcomes, true or false. In each node, '*' means an 'AND' and '+' means an 'OR'. For example, the root (i.e. node 1) describes a condition (if there is no Motif 3 or no Motif 12) AND (if there is no Motif 8 or no Motif 25 located in 300–400 bp upstream) AND ((if there is no Motif 4 and no Motif 21) or (if there is no Motif 11 and there is one or less Motif 22)). To classify a gene's behavior, the decision tree was traced from the top, i.e. node 1, to the bottom, i.e. a class. Note that when classifying a new gene, to keep the consistency, we used the same threshold as used by DMS during its search for motif occurrences to determine motif occurrences (refer to Section 4.1.2).

```
NODE 1 : ((M3 ≤ 0) + (M12 ≤ 0))*((M8 ≤ 0) + (M25 in [300,400] ≤ 0))*
         ((M4 ≤ 0)*(M21 ≤ 0) + (M11 ≤ 0)*(M22 ≤ 1))
         == TRUE
             NODE 2 : ((M4 ≤ 0) + (M10 ≤ 1)*(M13 in [400,500] ≤ 0))*
                      ((M4 > 0) + (M2 in [0,100] ≤ 0)*(M19 in [0,100] ≤ 0))
                      == TRUE
                          Class: No-change  { 0 up-regulated genes}
                                            {92 no-change genes   }
                      == FALSE
                          NODE 3 : (M21 ≤ 0)
                                   == TRUE
                                       Class: No-change {1 up-regulated genes}
                                                        {2 no-change genes   }
                                   == FALSE
                                       Class: Up-regulated {5 up-regulated genes}
                                                           {0 no-change gene    }
         == FALSE
             NODE 4 : (M3 in [300,400] ≤ 0)*(M17 in [300,400] ≤ 0)*(M4 ≤ 1)*
                      (M10 ≤ 0)*(M24 ≤ 1)*(M25 in [300,400] ≤ 0)
                      == TRUE
                          Class: No-change    {0 up-regulated gene}
                                              {6 no-change genes  }
                      == FALSE
                          Class: Up-regulated {52 up-regulated genes}
                                              { 0 no-change gene    }
```

Fig. 1. The hypothesis of oxidative stress genes (represented by a decision tree).

To verify the usefulness of the motif combinations generated by GALA, we performed two iterations of 10-fold cross validation by running C4.5 on the same data set with and without using the motif combinations generated by GALA. To perform one 10-fold cross validation, we first randomly shuffle the total data, i.e. the 58 oxidative stress up-regulated genes and the 100 no-change genes, to remove the ordering bias. We then divide the data into ten equal-sized sets, i.e. each set contains 10% of the total data; the distribution of the oxidative stress up-regulated genes and the no-change genes in each set will be at random. Each set of data will be iteratively used as the validation data to test the accuracy of the predictor, and the remaining nine sets of data will be used for training the predictor. The final predictive accuracy of the predictor is the average of the accuracy of the total ten runs of experiments. Our experimental results showed that motif combinations significantly improved the predictive accuracy by about 6% (82.24% with combinations compared with 76.13% without combinations) in paired t-test (confidence level $\gg 99\%$).

6. Lessons learned

Computational tools for detecting subtle similarities and classifying sequences have become an essential component of the research process. This is essential to our understanding of life and evolution, as well as to the discovery of new drugs and therapies. Bioinformatics is emerging as a strategic discipline at the frontier between biology and computer science, impacting medicine, biotechnology, and society in many ways.

Large databases of biological information create both challenging data-mining problems and opportunities, each requiring new ideas. In this regard, conventional computer science algorithms have been useful, but are increasingly unable to address many of the most interesting sequence analysis problems. This is due to the inherent complexity of biological systems, brought about by evolutionary stochastic process, and to our lack of a comprehensive theory of life's organization at the molecular level. Machine-learning approaches, on the other hand, are ideally suited for domains characterized by the presence of large amounts of data, noisy patterns, and the absence of general theories. The fundamental idea behind these approaches is to learn the theory automatically from the data, trough a process of inference, model fitting, or learning from examples. Thus, they form a viable complementary approach to conventional methods. It is the confluence of all three factors — data, computer and theoretical framework — that is fueling the machine-learning expansion, in bioinformatics and elsewhere [22].

We propose using multiple objective functions to detect meaningful motifs from sequences. Our experimental results demonstrated the synergy of the information content and the multiplicity significance helps maintain the balance between the consensus quality and the over-representation of motifs. The strategy of using multiple complementary objective functions alleviates the limitations of current approaches.

The advent of the genechip technology has provided a macro view of the gene expression on a genomic scale. With this kind of technology, we are finally able to realize the potential of information emerging from such global gene regulation studies. According to the time course, we are able to obtain the distinct temporal patterns of gene expression, and thus to understand different gene behaviors in the same controlled environment. To learn beyond how genes behave in the course of time on a genomic scale, we propose a novel view of the gene regulation analysis. With the assistance of the genechip technology, genes could be grouped into families according to different temporal patterns. Our analysis of gene regulation is focused on the search for significant combinatorial motifs involved in the regulation as well as potential hypotheses of how the gene regulation is related to the motifs. This type of analyses not only complement the global study of the changes of gene expression by looking into the involvement of combinatorial motifs in regulation, but also suggest to biologists further biological tests on the genes through the inference from the hypotheses produced by the analyses.

7. Future plans

Our yeast genome-wide expression studies and the experimental results provide a useful model system in which to explore computational approaches for several reasons.

1. The complete genome sequence of yeast is known and several array systems are available with which the yeast genome can be further studied using the computational approaches; thus, the contribution of these computational approaches can be verified.
2. Because many protein functions and even regulatory motifs are conserved between yeast and metazoan systems, much of what is learned in the studies should be applicable to more poorly understood complex systems.
3. The stress response itself is particularly conserved and so information is particularly likely to be transferable in many respects from the model to human systems.
4. Besides the applications in yeast studies, the computational tools can be extended, if necessary, to analyze the human genome when it is available.

Besides the cross validation introduced earlier, the candidate motifs, their combinations and the associated hypotheses need to be tested for biological activity. They will be validated in the following way in the biology laboratory. First, mutagenesis can be performed to test the role of candidate simple and complex motifs in transcriptional regulation. Examples of genes that contain candidate regulatory motifs or combinations of motifs which are predicted to be novel or of particular interest can be subcloned onto a shuttle vector with the endogenous gene copy disrupted. Motifs can be mutagenized at important positions and cells carrying the construct can be tested by Northern blot analysis to determine whether the motifs contribute, as predicted, to the regulation of the gene. Second, reporter fusion constructs can be used to model and systematically test the roles of candidate complex motifs in transcriptional regulation. For example, effects of motif location, copy number, and combinations of motifs can be tested.

Our analysis method is part of the ongoing yeast genome project. The goal of the project is to develop computational tools for interpreting and tracking the enormous amount of data emerging from genome sequence projects and from genome-wide gene expression studies. The objectives include (1) identification of host functions regulated at the transcriptional level by exposure to stress, (2) identification of sequence motifs and combinations of motifs, (3) testing of candidate sequence motifs and combinations of motifs for biological activities, (4) identification of regulatory proteins, and (5) enhancing our understanding of regulatory networks and cell functions including metabolism that are involved in the stress response.

We have already obtained several preliminary but promising results of the yeast genome-wide gene expression analysis. As the microarray and the genechip technologies become more mature, and more biological background knowledge becomes available, we expect to improve the performance of this novel analysis method, and finally achieve our ultimate goal.

References

[1] B. Dujon, The yeast genome project: what did we learn?, Trends Genet. 12 (1996) 263–270.
[2] J. DeRisi, V. Iyer, P. Brown, Exploring the metabolic and genetic control of gene expression on a genomic scale, Science 278 (1997) 680–696.
[3] L. Wodicak, H. Dong, M. Mittmann, M. Ho, D. Lockhart, Genome-wide expression monitoring in *Saccharomyces cerevisiae*, Nat. Biotechnol. 15 (1997) 1359–1367.
[4] G. Stormo, Computer methods for analyzing sequence recognition of nucleic acids, Annu. Rev. Biophys. Biophys. Chem. 17 (1988) 241–263.
[5] J. van Helden, B. Andre, J. Collado-Vides, Extractinf regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies, J. Mol. Biol. 281 (1998) 827–842.
[6] C. Lawrence, S. Altschul, M. Boguski, J. Liu, A. Neuwald, J. Wootton, Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignments, Science 262 (1993) 208–214.
[7] G. Hertz, G. Hartzell III, G. Stormo, Identification of consensus patterns in unaligned DNA sequences known to be functionally related, Comput. Applic. Biosci. 6 (2) (1990) 81–92.

[8] G. Hertz, G. Stormo, Identification of consensus patterns in unaligned DNA and protein sequences: a large-deviation statistical basis for penalizing gaps, in: Proceedings of the Third International Conference on Bioinformatics and Genome Research, 1995, pp. 201–216.
[9] T. Bailey, C. Elkan, Unsupervised learning of multiple motifs in biopolymers using expectation maximization, Machine Learn. 21 (1995) 51–80.
[10] R. Hughey, A. Krogh, Hidden Markov models for sequence analysis: extension and analysis of the basic method, Comput. Applic. Biosci. 12 (2) (1996) 95–107.
[11] S. Eddy, Multiple alignment using hidden Markov models, in: Proceedings of International Conference on Intelligent Systems for Molecular Biology, 1995, pp. 114–120.
[12] C. Lawrence, A. Reilly, An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences, Proteins Struct. Funct. Genet. 7 (1990) 41–51.
[13] Y. Hu, Biopattern discovery by genetic programming, in: Proceedings of the Third Annual Genetic Programming Conference, 1998, pp. 152–157.
[14] R. Staden, Computer methods to locate signals in nucleic acid sequences, Nucleic Acids Res. 12 (1984) 505–519.
[15] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[16] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948) 379–656.
[17] Y. Hu, D. Kibler, Generation of attributes for learning algorithms, in: Proceeding of the 13th National Conference on Artificial Intelligence, 1996, pp. 806–811.
[18] Y. Hu, Constructive induction: covering attribute spectrum, in: Feature Extraction in Construction and Selection: A Data Mining Perspective, Kluwer Academic Publishers, Dordrecht, 1998, pp. 257–272.
[19] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, 1993.
[20] E. Craig, The heat-shock response of *Saccharomyces cerevisiae*, in: The Molecular and Cellular Biology of the Yeast *Saccharomyces*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1992, pp. 501–537.
[21] V. Cid, A. Duran, F. del Rey, M. Snyder, C. Nombela, M. Sanchez, Molecular basis of cell integrity and morphogenesis in *Saccharomyces cerevisiae*, Microbiol. Rev. 59 (1995) 345–386.
[22] P. Baldi, S. Brunak, Bioinformatics: The Machine Learning Approach, Kluwer Academic Publishers, Dordrecht, 1998.