Modeling Cooperative Gene Regulation Using Fast Orthogonal Search

Ian Minz* and Michael J. Korenberg*

Department of Electrical and Computer Engineering, Queen's University, Kingston, ON, Canada

Abstract: Gene regulation is a complex and relatively poorly understood process. While a number of methods have suggested means by which gene transcription is activated, there are factors at work that no model has been able to fully explain. In eukaryotes, gene regulation is quite complex, so models have primarily focused on a relatively simple species, *Saccharomyces cerevisiae.* Because of the inherent complexity in higher species, and even in yeast, a method of identifying transcription factor (TF) binding motifs must be efficient and thorough in its analysis. Here we propose a method using the very efficient Fast Orthogonal Search (FOS) algorithm in order to uncover motifs as well as cooperatively binding groups of motifs that can explain variations in gene expression. The algorithm is very fast, exploring a motif list and constructing a final model within seconds or a few minutes, produces model terms that are consistent with known motifs while also revealing new motifs and interactions, and causes impressive reduction in variance with relatively few model terms over the cell cycle.

Keywords: Gene regulation, Gene expression, Binding motif, Transcription factor, Fast orthogonal search.

1. INTRODUCTION

While many methods have focused on modeling gene regulation through finding individual motifs to which transcription factors bind, a very significant part of regulation is driven by the cooperation of multiple factors. Recently, methods have been developed to consider the synergistic effects of multiple transcription factors [1-7]. The intriguing MARS-based methods [1, 2] were able to predict several pairs of motifs, as well as three-way interactions, that are likely to be functional in cooperatively binding to transcription factors. The present paper introduces the use of FOS to rapidly find motifs and interacting groups of motifs involved in gene regulation. Because of Fast Orthogonal Search's efficiency, third and potentially higher order cross products are easily searched and several groups of motifs that show high potential for functional cooperative behavior are found.

One strength of the proposed method is that it is able to discover motifs and synergistic pairs and groups of motifs without introducing a great number of parameters. While more complex models can approximate a system's output (gene expression) more accurately over the training data, there is a tendency for the added accuracy to be a side effect of noise fitting. All models built by FOS were done so with few parameters to reduce this problem. Cross validation (CV) was used in order to distinguish motifs which appear infrequently, indicating that those motifs are more likely to be fitting noise rather than explaining real regulatory effects on expression levels.

Three main measures of this method's performance were deemed important in its design: the accuracy of the models, their simplicity, and the running time. The objective was to create concise prediction models with high levels of correlation to the expression levels, while lowering the amount of time required by previous methods. Models able to accurately predict expression levels with very few terms are obviously far more desirable for several reasons. Firstly, it is unlikely that a large number of transcription factors is required, by analogy to the relatively simple regulations that occur in other biological systems. Secondly, a large number of terms in a model could mean a small contribution by each, so that those motifs' contributions would likely be statistically insignificant. By the same token, a large contribution by a small number of motifs or groups of motifs would mean each is more likely to be predictive.

Several previous methods have used gene clustering in order to determine motifs found in similarly-transcribed genes [7-10]. While these methods have been used to discover functional sequences, this approach introduces a loss of information, since genes within a cluster may lack motifs shared by other genes [11]. To avoid this loss of information, lists were created without narrowing down the possibilities for motifs based on gene clustering data.

2. METHODS

A program that accepts an array of gene expression data, genes' promoter region sequences and a list of candidate motifs and outputs a list of potential TF-binding motifs was created using MATLAB. Publicly available gene expression data for 715 genes of the *S. cerevisiae* genome, taken in 77 experiments, were used. In contrast to previous methods which calculate weight matrix scores [1, 2, 5, 11-13] or introduce other complexities, our method simply correlates the word counts present in the promoter regions of genes to the log ratio of those genes' expression levels (the ratio between the test sample and a control). For each search, only one time point was used, and the majority of the models were created using the expression levels of the alpha-arrest experiment at the 14 minute time point, which takes place during the M/G1

*Address correspondence to these authors at Walter Light Hall, 19 Union St., Kingston, ON, Canada K7L 3N6; Tel: 1-647-294-9190;
Fax: 1-613-533-6615; E-mail: ian.minz@utoronto.ca; Tel: 1-613-533-2931;
Fax: 1-613-533-6615; E-mail: korenber@queensu.ca

1875-0362/08

boundary of the cell cycle [11]. This time point was used in order to compare results to those found by previous methods. At different time points, gene expression becomes regulated by different transcription factors, and while results are, in general, from the 14 minute time point, the program can analyze any number of time points, by creating separate models which correlate to the expression levels at the corresponding times.

2.1. Fast Orthogonal Search

The Fast Orthogonal Search algorithm was employed in order to create concise models based on candidate motifs and motif groups that exhibit word count profiles with strong correlation to gene expression levels. Unlike previously reported methods, which use various types of linear regression or regression trees [14], FOS implicitly orthogonalizes the terms prior to adding them to the model, ensuring the minimal amount of redundancy in the model and very fast operation [15,16]. FOS is a system identification algorithm that operates by searching through a list of pre-designated candidate functions and iteratively adding the term that lowers the mean squared error (MSE) of the model by the greatest amount. Portions of a slightly modified Cholesky decomposition are used to rapidly locate this term. The consequence is that a large set of candidates can be quickly explored for the best choice without carrying out a full linear regression for each candidate. Moreover, the implicit orthogonalization of terms added to the model avoids the need to recalculate previously computed quantities for existing terms. Terms are added to the model sequentially until no candidate is able to improve the MSE beyond a certain threshold. This threshold is based on a standard correlation test requiring that a potential model addition be highly correlated with the unexplained residual at that point, and helps to avoid adding terms that are merely fitting noise [17]. Once the terms have been selected, least-square estimates of their coefficients are given directly as a byproduct of the algorithm [15].

The set of candidate functions was in part made up of basis functions of the form $c_m(n)$, the count of appearances of the motif m in the genes $n = 1, \ldots, N$. In addition to these $c_m(n)$, cross-products (including powers) of these functions were added to the set of candidates to represent cooperativity between motifs. Cross-products up to third order were allowed as candidates to reflect the presence of several different motifs in the promoter regions of genes. These cross-products were calculated as follows. For a pair of motifs m_1 and m_2, with word count profiles $c_{m1}(n)$ and $c_{m2}(n)$, a new candidate function $c_{m1,m2}(n)$, representing a dual contribution of these motifs, was created such that

$$c_{m1,m2}(n) = c_{m1}(n)c_{m2}(n) \tag{2.1}$$

This can be extended to consider any number of motifs by determining, for each gene, the product of the word counts of all motifs under consideration.

It is important to note that the candidates are not assumed to be orthogonal functions. The models developed by FOS are of the form:

$$E(n) = \sum_{m=0}^{M} a_m p_m(n) + e(n) \tag{2.2}$$

where $p_0(n) = 1$ and for $m > 0$ the $p_m(n)$ are the non-orthogonal candidate functions which were selected to be included in the model, and the a_m are the associated coefficients which best fit the output (i.e., are least-square estimates). Finally, $E(n)$ is the log expression ratio of gene n, $e(n)$ is the model error, and M is the number of (non-constant) terms included in the model and is not held fixed. Since a_0 is a least-square estimate and $p_0(n) = 1$, it follows that the average value of $e(n)$ over all the genes used to identify the model is zero. Because a constant term is first added to the model prior to adding any other terms, the model can alternatively be written as:

$$E(n) = a_0 + \sum_{m=1}^{M} a_m p_m(n) + e(n) \tag{2.3}$$

The above equation appears to be linear in its parameters and developing it amenable to linear approaches, but this is only so once the model terms $p_m(n)$ have been determined and the problem reduces to estimation of the remaining unknowns, the a_m. In fact finding the model terms and their coefficients that will minimize the MSE is a *nonlinear* least-squares problem. The term selection used by FOS is a nonlinear process, and FOS excels in selecting model terms (and least-square estimates of their coefficients) that are near-optimal in such nonlinear MSE minimizations. This capability enables FOS to obtain much greater improvements by adding cross-product terms than have been observed with linear methods.

FOS is related to a method by Desrochers [18] for selecting terms for nonlinear models, but amongst other differences, the computational and memory storage requirements of the latter method are proportional to the square of the number of candidates, whereas in FOS they depend linearly on the number.

Using fast orthogonal search, we can efficiently build up, implicitly, an economical series representation of the form

$$y(n) = \sum_{m=0}^{M} g_m w_m(n) + e(n) \tag{2.4}$$

where the $w_m(n)$ are mutually orthogonal over the data record, and the g_m are the orthogonal expansion coefficients achieving a least-squares fit. In the particular application of this paper, the model output $y(n) = E(n)$, the log expression ratio of gene n.

The orthogonal expansion coefficients in (2.4) are given by [15,16]

$$g_m = \frac{C(m)}{D(m,m)}, \text{ for } m = 0, \ldots, M \tag{2.5}$$

where

$$D(0,0) = 1 \tag{2.6}$$

$$D(m,0) = \overline{p_m(n)}, \text{ for } m = 1, \ldots, M \tag{2.7}$$

$$D(m,r) = \overline{p_m(n)\,p_r(n)} - \sum_{i=0}^{r-1} \alpha_{ri} D(m,i), \text{ for } m = 1, \ldots, M; r = 0, \ldots, m-1 \tag{2.8}$$

$$\alpha_{mr} = \frac{D(m,r)}{D(r,r)}, \text{ for } m = 1, \ldots, M; r = 0, \ldots, m-1 \quad (2.9)$$

$$C(0) = \overline{y(n)} \quad (2.10)$$

$$C(m) = \overline{y(n)p_m(n)} - \sum_{r=0}^{m-1} \alpha_{mr} C(r), \text{ for } m = 1, \ldots, M \quad (2.11)$$

The over-bar signifies taking the average over all genes n used to identify the model. The α_{mr} and $D(m,r)$ can be calculated by the following pseudocode, which achieves a Cholesky factorization:

$D(0,0) = 1$

FOR $m = 1$ TO M

Calculate $D(m,0)$ from (2.7)

FOR $m = 1$ TO M

FOR $r = 0$ TO $m - 1$

Calculate α_{mr} from (2.9)

Calculate $D(m, r + 1)$ from (2.8)

Once the g_m and α_{mr} are known, the coefficients a_m in (2.2) can be obtained by [15-17]:

$$a_m = \sum_{i=m}^{M} g_i v_i \quad (2.12)$$

where

$$v_m = 1 \quad (2.13)$$

$$v_i = -\sum_{r=m}^{i-1} \alpha_{ir} v_r, \text{ for } i = m + 1, \ldots, M \quad (2.14)$$

Note that the reduction in MSE by adding the M-th non-constant term, $p_M(n)$, is equivalently

$$Q(M) = g_M{}^2 D(M,M). \quad (2.15)$$

A similar observation was made by Desrochers [18].

In selecting $p_M(n)$, we only need to carry out the above pseudocode for $m = M$, which allows us to avoid repeating calculations done for previous values of m. Once $p_M(n)$ has been chosen, by searching the list of available candidates to find the one that maximizes $Q(M)$ of (2.15), the abbreviated pseudocode can be repeated to recalculate $C(M)$ and g_M. This allows the α_{mr}, $D(M,M)$, $C(M)$ and g_M to be properly set prior to searching for $p_{M+1}(n)$ [15]. This implementation reduces memory storage requirements, an important consideration in searching for regulatory motifs and cooperating groups of motifs, but more efficient code is available [19] with increased memory storage.

In summary, to model the gene expression level by motif appearances in the promoter regions of the genes, the gene expression data were treated as a time series, while profiles of the raw word counts of each motif in each gene, and cross-products thereof, were treated as the candidate functions. Previous methods used basis functions such as splines in order to more specifically choose terms. Using spines introduces a threshold where only above (or below) a certain word count does a motif have any biological function [1]. Having functions based on word counts already introduces a switch-like characteristic that the splines are meant to emphasize. When splines were incorporated into the FOS program, only very minimal improvements were found in the models, at the expense of greater complexity and longer running time. It was deemed more valuable to allow a greater variety of candidate motif sequences in lieu of simply increasing the number of candidates by multiplying them by various spline functions.

2.2. Input Motif Lists

Several motif lists, falling into two categories, were used as lists of candidate motifs from which to build models. Initially, it was desired that all motifs of lengths 5bp to 10bp be searched. However, due to the massive numbers of possible candidate motifs (approximately 250,000 possible 10bp-long motifs were found in the promoter regions of the 715 genes used), lists of degenerate "skeleton" motifs were created. By only designating a certain number of "prongs" in each motif list, the number of different motifs in these longer lists was decreased to allow for much more computationally feasible searching. The skeleton lists will hereafter be referred to by names of the form NbpPp, where N is the number of base pairs and P is the number of prongs. An example of a skeleton motif is the following: `actnntcngn`. This motif is 10 base pairs in length and each of the 4 spaces, denoted by "n", can be any of the four bases. Due to its length and number of required prongs (six), it would be in the 10bp6p list.

Three different lists of skeletons were created for each of the 10mers and 9mers. The number of prongs used in each list was 5, 6, or 7. The six skeleton lists created are (i) 10bp7p, (ii) 10bp6p, (iii) 10bp5p, (iv) 9bp7p, (v) 9bp6p, and (vi) 9bp5p. On average, these lists ended up with approximately 4000 skeleton motifs each. While this method of grouping motifs together to create one skeleton representing a family of motifs does result in information loss since FOS will then not search each 10bp motif for activity, it allows for the possibility that transcription factors are able to bind to any of a family of motifs. Some details about the binding of TFs to DNA are still unknown, and the possibility that spaces in the motifs are necessary must not be overlooked. Methods have analyzed and found potentially functional degenerate motifs. Methods have analyzed oligomers with fixed spacing [10]; however, no previous methods have used lists composed of skeleton motifs with irregularly placed prongs. It has been noted that human TFs are likely to bind to much more degenerate motifs than in yeast [5], and so it will become more important to search these lists in analyzing human gene expression data.

These lists were created starting with a complete list of all 10bp (or 9bp) possible motifs found in the promoter regions of the set of genes. All motifs were compared to the first motif in this list, and a histogram showing the number of matches at each base position was generated. For 10mers and 7-pronged skeletons, three positions are allowed to be arbitrary. The three positions at which the least motifs shared a base with the first motif were then allowed to be any of the four bases and a new skeleton motif was built with prongs (i.e., fixed bases) matching the first motif at the other positions. All motifs fitting this skeleton were removed from the list, and the skeleton was added to the list of skeleton motifs.

This process was repeated until all motifs in the 10mer list were represented in the skeleton list. This process was analogously repeated for the five other lists of various forms of skeleton motifs.

In addition to the skeleton motifs, lists of every occurring 5mer, 6mer and 7mer were searched for active motifs. While most other methods have inputted lists of motifs found to be significant by previous methods, we make no assumptions and the program requires no a priori information.

2.3. Reduction in Variance

The reduction in variance (RIV) was used as a measure of the created models' abilities to explain variations in the gene expression data, and was calculated by the following equation:

$$\%RIV = \left[1 - \frac{\overline{\left[E(n) - \hat{E}(n) - \overline{E(n) - \hat{E}(n)}\right]^2}}{\overline{\left[E(n) - \overline{E(n)}\right]^2}}\right] \times 100\% \qquad (2.16)$$

where

$$\hat{E}(n) = \sum_{m=0}^{M} a_m p_m(n) \qquad (2.17)$$

and where $E(n)$ is the log expression ratio of gene n at the single time point studied, and the over-bar signifies taking the average over all the genes used to fit the model. Note that in the numerator, inside the mean-square term, the mean

$$\overline{\left[E(n) - \sum_{m=0}^{M} a_m p_m(n)\right]} = \overline{e(n)} = 0 \qquad (2.18)$$

over the genes used to identify the model. However, the %RIV will also be used to evaluate the identified model over novel gene expression data, in which cases the over-bar will denote the average over the new data where the mean in question is not necessarily zero.

3. RESULTS

For each motif list, ten-fold cross-validation was used to determine regulatory motifs that repeatedly are found by FOS. For each set of conditions (motif list, order of interactions allowed, pre-screening [pre-ranking] on or off), models

Table 1. Summary of Modeling Results

Pre Screened?	# of Top Terms	Motif List	%RIV on Modeling Genes			%RIV on Testing Genes		
			1st	2nd	3rd	1st	2nd	3rd
Yes	10	10bp7p	11.77	12.50	12.50	3.36	2.69	2.69
Yes	15	10bp6p	13.72	19.33	20.08	3.72	5.33	6.24
Yes	5	10bp6p	10.70	11.25	11.21	6.27	2.14	2.09
Yes	15	10bp5p	11.68	17.94	19.15	2.61	4.66	1.15
Yes	10	10bp5p	10.94	14.16	14.17	2.35	5.53	4.58
Yes	10	9bp7p	10.37	11.22	11.22	0.83	0.90	0.90
Yes	10	9bp6p	19.11	20.23	20.27	8.61	9.76	9.48
Yes	10	9bp5p	6.58	8.69	11.55	4.30	2.37	0.88
Yes	10	7mers	24.79	24.79	24.79	14.32	14.32	14.32
Yes	15	6mers	20.54	28.19	28.75	15.35	13.01	12.25
Yes	10	6mers	20.05	21.00	21.77	16.27	14.87	15.08
Yes	10	5mers	17.85	21.52	23.10	13.77	15.18	5.87
Yes	15	5mers	18.16	24.94	28.84	12.52	8.61	-0.59
No	10	10bp7p	15.96	17.03		-10.65	-9.40	
No	15	10bp6p	16.59	20.23		-4.47	0.59	
No	10	10bp6p	16.59	18.09		-4.47	-2.44	
No	5	10bp6p	16.59	17.00		-4.47	-5.49	
No	15	10bp5p	14.91	20.47		-7.27	0.45	
No	10	10bp5p	14.91	17.66		-7.27	1.25	
No	10	9bp7p	13.41	16.08		-20.89	-18.27	
No	10	9bp6p	21.41	21.90		-2.31	3.16	
No	10	9bp5p	21.51	22.11		-2.53	-3.15	
No	10	7mers	32.75	32.75		-17.07	-17.07	
No	10	6mers	22.75	22.58		12.00	12.55	
No	10	5mers	18.18	22.35		9.64	13.24	
No	15	5mers	18.18	25.29		9.64	9.00	

were built on 90% of the genes and tested for their fit on the other 10%. This was repeated so as to build a model holding out each of the sets of 10% of the points, making ten models per set of conditions. A histogram of motif appearances in the models was generated, and those appearing above a certain number of times were considered significant. The number of motif appearances out of 10 is called its CV score. Generally, known motifs appeared in 8-10 out of 10 models, so a threshold of 7 appearances was used in order to maintain a high level of rigor in considering motifs to be significant. Those motifs found only 1-3 times are quite possibly false positives, which are found due to their counts' similarities to the noise in the gene expression levels. Those with 7-10 appearances, however, are highly unlikely to be accidentally found.

In general, very concise models were built by FOS. Even with low numbers of terms, the models' reductions in variance (RIVs) ranged from 6.6-32.8%. Thus, FOS was able to create accurate models with very few terms. Often, 3-5 motifs or groups (i.e., cross-products) of motifs were able to accurately explain the variation in the genes' expression levels. Table **1** shows the RIVs associated with the models built from various lists. A number of runs were pre-screened, while a number were not. The pre-screening process involves running FOS on the entire list and pre-ranking the top individual motifs. Then FOS is re-run with only these top individual motifs and cross-products as candidates. This process limits the number of higher order interactions by only considering the possibility of cooperation between motifs that have a high significance level on their own. The number of top terms considered by FOS was generally taken to be 10 (not counting their cross-products), but several trials were executed with 5 or 15 terms. When pre-screening was not used, all possible single motifs were searched, though the prescreening process was still done to limit the number of higher order interactions. Reduction in variance is reported in 6 columns per motif list. FOS was run once when allowing only single motifs (i.e., no cross-products) to be searched. FOS was run again with 2nd order interactions of motifs being considered in addition to the single motifs, and then once more with 3rd order interactions being added onto the 1st and 2nd order interactions. The model RIV refers to the RIV calculated based on the 90% of genes used for determining the model, while the test RIV refers to the RIV calculated based on the 10% of genes withheld during modeling.

The models that achieved a very high RIV but did not perform well when tested on the points that were held out show that in these cases, FOS very accurately modeled the training expression data, but did not capture a model that explains the expression in general. In other words, these models have included terms that have only fit the noise in the data. A more accurate measure of the models' accuracy is their ability to explain variations in other genes' expression levels (i.e. the testing genes – those genes not used to build the model). A high RIV on the testing genes implies that a model has found terms which explain all genes' expression profiles. Strong evidence of the value of the pre-screening process appears from the largely negative %RIV on the testing genes achieved when bypassing the pre-screening process. The majority of the runs that were not pre-screened resulted in models performing poorly on the testing genes. The 5-mers and 6-mers achieved better results than most, but their performance is still enhanced by pre-screening.

3.1. Significant Motifs and Groups of Motifs

FOS returned a number of motifs through analysis of the various motif lists. A histogram of the cross-validation scores of these motifs for the 14-minute time point (at the M/G1 boundary) is shown in Figs. (**1**) (5-7mers), (**2**) (9bp skeleton motif lists) and (**3**) (10bp skeleton motifs). In all cases here pre-screening was used to select the top 10 terms, which, together with cross-products up to 3rd order, were then searched by FOS to find the final models. One motif that is consistently found by FOS is the well-known MCB element. There are several forms of the element whose appearances predict expression levels very accurately. The most effective form in terms of reduction in variance was found to be the 5mer `acgcg`. While another form of the motif, `cgcgt`, occasionally appears in models, the `acgcg` is found nearly every time among the models built from the

7mers	
cctcgac	10
gacgcgt	10
aacgcgt	10
tgcgaag	10
tgagaac	9
tttgctc	8
aacttct	7
cgagtgt	6
6mers	
acgcgt	10
aaacaa	10
ctaagc	10
gttccg	9
5mers	
acgcg	10
aaaca	10
agggg, ctcga	10
gtaat, gtaat	9
gtaat	9
tgttc	9
cgcgt, tgttc, gcccg	8
tgttc, tgttc	8
gtaat, gtaat, gtaat	7

Fig. (1). The number of appearances of some of the most commonly found motifs and groups of motifs is shown for the 14-minute time point (at the M/G1 boundary). Motifs are separated by list.

9bp5p	
naaacannn	10
ncnncgnct	8
nngncgctn, ncnncgnct, nngntggnt	8
nntngttnt	7
ncnncgnct, ncnncgnct	7
9bp6p	
ncctcgncn	10
acgcntcnn	10
nngntggtt	10
natntancg, natntancg	9
nnngttccg	8
ngcncnccg	7
natntancg	7
9bp7p	
ngacnaagc	10
tntacgcnt, atcnactan	9
ggnggnccc	8
acnttggcn	7

Fig. (2). The number of appearances of some of the most commonly found 9bp skeleton motifs and groups of motifs is shown for the 14-minute time point (at the M/G1 boundary). Motifs are separated by list.

10bp5p	
nncnnancgc	10
ttngnnngtn	9
nncnnancgc, nncnnancgc	8
nnntnacgng	7
10bp6p	
nannntaatt, ntnnnagatg	10
nannntaatt	9
gncgcntcnn	9
nannntaatt, gncgcntcnn	8
nannntaatt	8
aaannaacnn	7
10bp7p	
anctnaattn	10
nggntangtt	9
nntntttctt, ngggagntgn	8
gncgcgtcnn	8
nntntttctt	8
gncgcgtcnn	7
cnngggcatn	7

Fig. (3). The number of appearances of some of the most commonly found 10bp skeleton motifs and groups of motifs is shown for the 14-minute time point (at the M/G1 boundary). Motifs are separated by list.

5mer list. Cross-validation shows that this motif appears 10 out of 10 times, without fail. Another 5mer that appears consistently is aaaca, known as the STE12 element. Two motifs not previously reported are the gtaat and tgttc motifs, which FOS finds to appear quite regularly. Both consistently appear 8 to 9 times out of 10 through CV experiments, which hints that these motifs have a high probability of being functional. Among the 5mers, FOS finds three pairs of motifs as well as two third order sets that consistently appear in the models. One of these pairs, agggg/ctcga, is found by FOS every run. CV results show it to appear 10 times under most conditions. This pair of motifs is highly significant, and it is almost certain that the cooperativity between those two motifs is strongly influential on gene expression at the particular time point (alpha-arrest) used. Both elements in the pair have been previously found: the first is the known STRE element, and the second corresponds to a portion of the motif cctcgac, which has been reported in the literature [11]. While both have been found, their synergistic functionality has not been previously mentioned. The gtaat/gtaat squared term is found 9 times through CV. Slightly less significant is the tgttc/tgttc squared term, which appears 8 times under the set of conditions considered here. Two 3rd order groups are found with regularity. While not appearing quite as often as some of the single motifs, there is reason to believe that the cgcgt/tgttc/gcccg and gtaat/gtaat/gtaat groups, which appear 8 and 7 times through CV, respectively, might be functional in cooperative regulation.

The 6mers show highly significant motifs as well, with a great deal of corroboration with the 5mers. Three motifs, acgcgt, aaacaa and ctaagc are found in every model built. The first of these is the most well-known form of the MCB element. The second motif is the STE12 element. It does not seem that the last motif appears in the literature, but there is agreement with other motif findings (skeletons) by FOS. The motif gttccg is also consistently found, with a CV score of 9. A small number of pairs of motifs are found, but no higher-order combinations of 6mers are found often enough to seem worthy of note.

Analysis of the 7mers shows two motifs of the MCB family that are found by FOS without fail. The aacgcgt and gacgcgt elements both appear in every model created with the 7mer list and cross-validation shows the two to appear 10 times on every run. Two other motifs are also found to be in every model: cctcgac and tgcgaag. Another 7mer found to be highly significant is the motif tgagaac. The striking similarities in the last two mentioned motifs indicate that there is likely a degenerate motif to which an important TF can bind. The two similar motifs tgcgaag and tgagaac could be grouped together as tgmgaas

(where m can be either a or c, and s can be either c or g). No significant pairs of motifs or groups of third order are found among the 7mers, likely because of the fact that requiring seven straight bases to be held is a fairly rigid constraint.

Turning to the results of the modeling based on the skeleton motifs, we begin with the least degenerate list, the 9bp7p list. The motif ngacnaagc is repeatedly found, with CV scores of either 9 or 10. This motif is a near fit to tgcgaag, and especially the 6mer ctaagc, both found by FOS. Among this list, FOS also finds one pair of skeleton motifs, tntacgcnt/atcnactan, which is quite significant. The first of the two closely fits the 6mer version of the MCB element, acgcgt. Two other individual skeletons are found to repeatedly appear as well (ggnggnccc and acnttggcn). Among the 9bp6p list, three notable individual motifs come up 10 times each: acgcntcnn, fitting the MCB shape (acgcgt), the skeleton nngntggtt, and ncctcgncn, the latter closely fitting the cctcgac element also found among the 7mers. Among the 9bp5p list, the STE12 motif aaaca is found in the naaacannn skeleton.

Because of the high level of corroboration between the motifs found by FOS and those found by previous methods, this method has been shown to exude a convincing ability to model the control of gene expression. In particular, FOS finds some of the most well-known motifs, such as the MCB and STE12 elements, consistently so it is clear that FOS is able to correctly determine motifs that have controlling effects on gene expression. A high level of corroboration also exists between the different lists searched by FOS, with many bases being shared among the high-ranking motifs found in different lists. With confidence that motifs found by FOS are correctly discovered, the fact that FOS also finds unknown motifs and groups of motifs with high levels of consistency suggests that it may be able to predict other motifs that could be further analyzed for functionality.

3.2. Running Time

One of the advantages of employing FOS as the method of model-building is its speed. Speed is determined by several algorithm and data-related factors: the length of the time series, the MSE threshold below which no terms are added and the number of candidate functions. Because 715 cell-cycle genes were used, the time series length was constant for all runs. The threshold, Δ_{min}, that had to be exceeded to add the M-th non-constant term was chosen to be

$$\Delta_{\min} = \frac{10.9}{N} \cdot \left[\overline{E^2(n)} - \sum_{m=0}^{M-1} {g_m}^2 D(m,m) \right] \qquad (3.1)$$

where the $E(n)$ are the log expression ratios of the genes n used to identify the model, N is the number of genes, $M - 1$ is the number of (non-constant) terms already in the model and the over-bar signifies taking the average. The g_m and $D(m,m)$ are calculated as described by Korenberg [15]. The g_m are the coefficients of the orthogonal functions that are *implicitly* created from the non-orthogonal $p_m(n)$, and the orthogonality means that existing g_m do not have to be recalculated when new terms are added to the model. The $D(m,m)$ are equal to the mean-square of the corresponding orthogonal functions implicitly created.

The above threshold corresponds to a 99.9% confidence interval, and varies from the threshold corresponding to a 95% confidence interval only in the coefficient, which is 10.9, rather than 4 [17]. The only factor that was variable was the number of candidate functions, which corresponds to the total of the number of motifs and motif interactions searched. When the pre-ranking function is turned on, only the top 1st order motifs and their cross-products were included as candidates. For the majority of runs, the top 10 motifs from the pre-ranking were used. While other cutoffs were tried, cross-validation showed that allowing more than the top 10 introduced higher levels of noise fitting, while less than 10 sometimes excluded motifs that are necessary for cooperative binding. Cross-product terms using the top 10

Table 2. Running Times Using Various Motif Lists

Motif List	Order Allowed	# of Motifs	Running Time (seconds)		
			Prescreening	Modeling	Total
5mers	1st	1024	4.95	0.04	4.99
5mers	1st + 2nd	1024	4.95	0.24	5.19
5mers	1st, 2nd + 3rd	1024	4.95	2.60	7.55
6mers	1st	4096	18.44	0.03	18.47
6mers	1st + 2nd	4096	18.44	0.21	18.65
6mers	1st, 2nd + 3rd	4096	18.44	2.17	20.61
7mers	1st	16384	80.83	0.03	80.86
7mers	1st + 2nd	16384	80.83	0.33	81.16
7mers	1st, 2nd + 3rd	16384	80.83	3.53	84.36
10bp7p	1st + 2nd	4041	16.67	0.09	16.76
9bp7p	1st	3706	14.84	0.03	14.87

motifs would then add 55 and 220 distinct 2nd and 3rd order terms, respectively. Pre-screening motifs would involve running FOS on the entire set of individual motifs in a given list, meaning anywhere from 1024 (5mers) to 16384 (7mers) motifs would be searched.

Table **2** shows the measured running times for a selected set of runs. All runs were executed on a notebook with a 2.0 GHz Pentium Centrino processor. Clearly, the length of the pre-screening process is highly dependent on the length of the candidate motif list. However, once only the most significant candidates are found, the addition of candidate pairs and groups of motifs does not greatly affect the overall running time. Another item to note is the linear relationship between the prescreening time and the number of candidate motifs. Only fairly concise lists of motifs (≤16384) were searched (in ≤ 81 seconds) in the pre-screening, but more lengthy lists could be used without running time blowing up. Memory constraints, however, do become a factor as the number of candidate functions increase, and thus more powerful computers become desirable.

3.3. Cell Cycle

Over the course of the cell cycle, genes' transcription levels change drastically, allowing for physiological changes to take place within the cell. Various TF-binding motifs found by FOS were used to show their individual contributions to the reduction in variance across the 715 genes over the course of the cell cycle. The reduction in variance of the well-known MCB (`acgcgt`) element was previously shown to have a sinusoidal-like plot across the cell cycle [1, 2]. FOS also shows a sinusoidal-like plot, with no contribution from MCB at 4 time points across the two cell-cycles, which was also found by MARSMotif [1, 2] and REDUCE [11]. FOS clearly is able to capture the peak activity of MCB that occurs at the 21 minute time point, exactly agreeing with MARSMotif. The plot of the `acgcgt` motif's contribution to the RIV across the two cell-cycles is shown in Fig. (**4**).

3.4. Noise Levels

Noise levels are very significant in the process of modeling gene expression purely by word counts of potential regulatory elements. Expression is enhanced and repressed by the binding of transcription factors to the regulatory motifs found in the DNA upstream of coding regions; however, several factors other than motif appearances have an effect on the changes in transcription level. Therefore, it is impossible to account for all the variations in the expression data solely by the counts of motif occurrences. Das *et al.* [1, 2] estimate noise to account for ~50% of the variations in the data. Some noise is introduced by inaccuracies in the collecting of the microarray data. A more important source of noise comes from the fact that concentrations in the TFs have great effects on the expression levels. Although a gene may have plenty of binding sites, should the corresponding TF concentration be low, binding will not occur and transcription will occur at a lower rate than would be predicted by the number of motif appearances.

3.5. An Alternative Cross-Validation Method

The reporting of most results came from a method of cross-validation that modeled on different sets of 90% of genes and tested the model on the remaining withheld 10%. However, it was deemed necessary to entirely withhold a larger subset of genes for a more rigorous test of FOS's modeling abilities. The method used in determining significant motifs did, in fact, use all genes. Here, an alternative method of cross-validation is proposed. Rather than breaking the entire set of genes into ten sections and sequentially withholding 10% per model, in the new experiments 20% of genes were set aside, and the previous method was used to build 10 models out of the first 80%. This means that ten sets of 90% of the first 80% of genes were used for modeling. A histogram of significant motifs was created to keep track of how many times each appeared in these 10 models. Interactions were only considered significant if the testing RIV (the average RIV on the ten sets of 8% of genes) was increased. Once a set of significant motifs and interactions was created, FOS was once again run, but this time on the entire set of 80% of genes, in order to calculate the weighting coefficient to be associated with each motif or motif group. This was considered the final model and this would be used to predict the expression levels of the final 20% of genes, which were now completely isolated from the modeling process.

This method was used to create a model for 6-mers. For three sets of conditions (allowing no interactions, up to 2nd order interactions and up to 3rd order interactions), ten models were created, using only 80% of genes. Here, the testing RIV (i.e. the average RIV of the ten models' ability to predict the remaining sets of 8% of genes – all still part of the initial 80%) was found to drop slightly upon the addition of

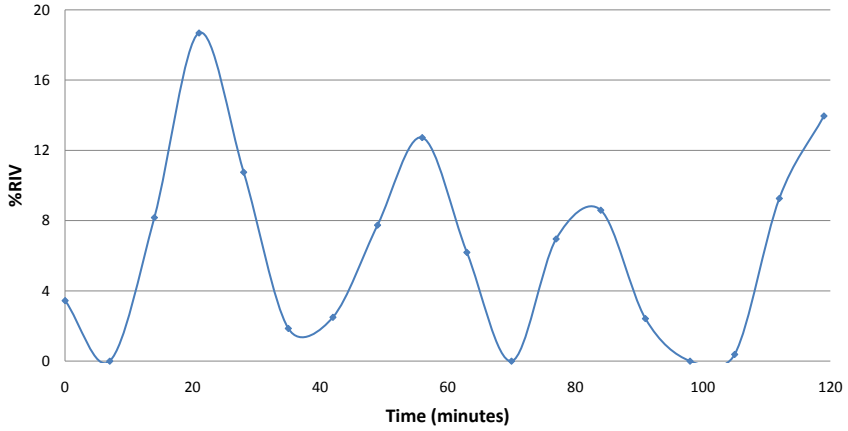

Fig. (4). Percent reduction in variance of the `acgcgt` (MCB) element over two cell-cycles.

two interactions found by FOS. These interactions were thus not included in the final modeling. Four motifs were found to appear 9 or 10 times out of the 10 models built in the first step (`acgcgt`, `aaacaa`, `ctaagc`, `gttccg`). FOS was run on the first 80% of genes to determine the coefficients for these four motifs. This model, when used to predict the expression levels of the final 20% of genes, yielded a RIV of 15.41%. This testing shows that the 10 times cross-validation is capable of yielding terms that correspond to a high RIV over rigorously separated novel genes.

4. DISCUSSION

In recent years, the increase in the ability to gather genetic data has led to a vast library of knowledge. With many fully sequenced genomes and the availability of microarray data that can quite accurately measure transcription levels, a number of methods have been developed to analyze these data. In this paper, it has been demonstrated that FOS compares favorably with previous methods and easily finds functional motifs to which transcription factors can bind. Because of the efficiency of FOS, the discussed method can execute in very little time and build concise models with high degree of accuracy.

The implicit orthogonalization carried out by FOS is a key characteristic to its success in a number of applications. In terms of modeling gene expression profiles with motif counts, the implicit orthogonalization ensures that once a term is added to the model, the next term added will be the one which best explains the output while taking into account that the previous term has been added. Other methods not using this type of orthogonalization will find terms that explain the data, but might not conveniently find the terms that explain the remaining portions of the data after adding terms. The fact that FOS does this makes it a powerful tool in quickly building accurate and concise models.

The fact that no a priori information is needed means that this method can be employed on a complete set of motifs, working from scratch without requiring a separate method to narrow down the possible candidates. Because of the efficiency of the method, large numbers of candidates can be searched quickly. In view of the corroboration of some FOS-found motifs by those previously reported, it can be assumed likely that motifs frequently returned by FOS have a high probability of being active binding sites for transcription factors.

In summary, the method finds a number of motifs that have been found as well as a number which have not been previously reported. The ability of FOS to be a predictor of TF-binding motifs suggests that it will become an important algorithm in further analyzing higher order species. For the yeast genome, adding certain pairs of motifs to models leads to high levels of reduction in variance. In analyzing the 5mers, for example, the RIV on the testing genes is improved when allowing 2nd order interactions to be included in the models. When 3rd order interactions are included, however, the testing RIV falls, indicating that perhaps 3rd order interactions do not take place. Yeast is relatively quite simple, and it would not be surprising if the transcription process were less intricately regulated than in more complex species. While, for the most part, groups of 3 or more interacting motifs are not chosen by FOS, the method has the ability to find higher order cooperation between motifs. It is likely that this ability to uncover higher order synergy between elements will become crucial when analyzing more complex species such as mammals.

ACKNOWLEDGEMENT

We are very grateful to Dr. Debopriya Das of the Lawrence Berkeley National Laboratory for providing the data sets used in this paper, and for his constructive comments on the manuscript.

REFERENCES

[1] D. Das, N. Banerjee, and M.Q. Zhang, "Interacting models of cooperative gene regulation", *Proc. Natl. Acad. Sci. USA*, vol. 101, pp. 16234-16239, 2004.

[2] D. Das, and M.Q. Zhang, "Predictive models of gene regulation – Application of regression methods to microarray data" in *Microarray Data Analysis: Methods and Applications*, M.J. Korenberg, Ed. Humana Press, 2007, pp. 95-110.

[3] Y. Pilpel, P. Sudarsanam, and G.M. Church, "Identifying regulatory networks by combinatorial analysis of promoter elements", *Nat. Genet.*, vol. 29, pp. 153-159, 2001.

[4] N. Banerjee, and M.Q. Zhang, "Identifying cooperativity among transcription factors controlling the cell cycle in yeast", *Nucleic Acids Res.*, vol. 31, pp. 7024-7031, 2003.

[5] D. Das, Z. Nahle, and M.Q. Zhang, "Adaptively inferring human transcriptional subnetworks", *Mol. Syst. Biol.*, vol. 2, pp. E1–E14, 2006.

[6] H.-K. Tsai, H.H.-S. Lu, and W.-H. Li, "Statistical methods for identifying yeast cell cycle transcription factors", *Proc. Natl. Acad. Sci. USA*, vol. 102, pp. 13532-13537, 2005.

[7] M. Kato, N. Hata, N. Banerjee, B. Futcher B, and M.Q. Zhang, "Identifying combinatorial regulation of transcription factors and binding motifs", *Genome Biol.*, vol. 5, pp. R56.1-13, 2004.

[8] S. Keles, M. van der Laan, and M.B. Eisen, "Identification of regulatory elements using a feature selection method", *Bioinformatics*, vol. 18, pp. 1167-1175, 2002.

[9] 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 ceycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization", *Mol. Biol. Cell*, vol. 9, pp. 3273-3297, 1998.

[10] E.M. Conlon, X.S. Liu, J.D. Lieb, and J.S. Liu, "Integrating regulatory motif discovery and genome-wide expression analysis", *Proc. Natl. Acad. Sci. USA*, vol. 100, pp. 3339-3344, 2003.

[11] H.J. Bussemaker, H. Li, and E.D. Siggia, "Regulatory element detection using correlation with expression", *Nat. Genet.*, vol. 27, pp. 167-171, 2001.

[12] D.Y. Chiang, A.M. Moses, M. Kellis, E.S. Lander, and M.B. Eisen, "Position specific variation in the rate of evolution in transcriuption factor binding sites", *Genome Biol.*, vol. 4, pp. R43, 2003.

[13] X. Yu, J. Lin, T. Masuda, N. Esumi, D.J. Zack, and J. Qian, "Genome-wide prediction and characterization of interactions between transcription factors in Saccharomyces cerevisiae", *Nucleic Acids Res.*, vol. 34, pp. 917-927, 2006.

[14] T.M. Phuong, D. Lee, and K.H. Lee, "Regression trees for regulatory element identification", *Bioinformatics*, vol. 20, pp. 750-757, 2004.

[15] M.J. Korenberg, "A robust orthogonal algorithm for system identification and time-series analysis", *Biol. Cybern.*, vol. 60, pp. 267-276, 1989.

[16] M. J. Korenberg, "Fast orthogonal identification of nonlinear difference equation and function expansion models", *Proceedings of the 28th Midwest Symposium on Circuits and Systems*, pp. 270-276, 1987.

[17] M.J. Korenberg and L.D. Paarmann, "Orthogonal approaches to time-series analysis and system identification", *IEEE SP Magazine*, pp. 29-43, July 1991.

[18] A.A. Desrochers, "On an improved model reduction technique for nonlinear systems", *Automatica*, vol. 17 (2), pp. 407-409, 1981.

[19] K.M. Adeney and M.J. Korenberg, "Iterative fast orthogonal search for MDL-based training of generalized single-layer networks", *Neural Netw*, vol. 13, pp. 787-799, 2000.

Received: September 08, 2008 Revised: October 22, 2008 Accepted: October 29, 2008