# ConstrainedMotif: A Periodicity Constraint Based Algorithm to Predict Cell-cycle Associated Promoter Motifs using Time-course Gene Expression Data

Yingren Liu[2*] *liuyingr@comp.nus.edu.sg,*
Karuturi R. Krishna Murthy[1*†] *karuturikm@gis.a-star.edu.sg*
Wing-Kin Sung[1,2] *sungk@gis.a-star.edu.sg*

[1]Genome Institute of Singapore
60, Biopolis Street
Singapore 138672
Republic of Singapore

[2]National University of Singapore
10 Kent Ridge Crescent
Singapore
Republic of Singapore

## Abstract

**Motivation:** *Cell-cycle associated promoter motif prediction is very important to understand the cell-cycle control and process. Modeling genome-wide gene expression as a function of the promoter sequence motif features has drawn great attention recently. The proposed techniques using this approach are not specific to cell-cycle associated motif discovery, hence find aperiodic motif weights across the time-course and lower sensitivity. Motifs are scored based on the successive model error reduction steps which may not reveal all relevant motifs since they are alternatives for the model. Another, drawback is, these methods output a list of sequences which may either contain several instances of a dominating motif box (a set of alternative sequence motifs) such as MCB or only a few instances of an important box.*
**Results:** *To address the above problems, we propose a multi-step constrained optimization based position weight matrix (PWM) motif finding methodology called* ***ConstrainedMotif****. It models the cell-cycle regulated gene expression as a linear function of the motif features while the weights of them are constrained to be periodic across the time-course. The score of a motif is the error reduction in the prediction by that motif alone. The multi-step modeling starts with a set of sequences and output a ranked list of cell-cycle associated PWM motifs. We evaluate this methodology using S. Cerevesiae cell-cycle data published by Spellman et al. The results show that* **ConstrainedMotif** *is more sensitive and most of the instances of the boxes are represented by the respective matching PWMs.*
**Software:** *giscompute.gis.nus.edu.sg/˜karu/cdcAnal/motif*
**Keywords:** *Cell-cycle, Motif prediction, Microarray, Constrained optimization.*

*These authors contributed equally to this work.
†Corresponding author

## 1. Introduction

Predicting cell-cycle associated promoter motifs is very importnant component of the cell-cycle regulated gene expression analysis. Several methods have been developed and they fall into three broad categories: (1) sequence-only approach in which the nonrandomly occurred motifs are discovered using sequence background model; (2) enriched motif prediction among groups of genes discovered using microarray data; and, (3) modeling the gene expression as a function of the motif statistics in the promoter regions of the genes. The gene groups may be either cell-cycle regulated genes whose expression reaches maximum at a specific phase of the cell-cycle or the cell-cycle regulated genes whose time-course expression profiles cluster together. The motif statistics may be either simple counts i.e. number of occurrances of the given motif or may be score based on more complex background model.

Wolfsberg et al. [17] ordered the genes in the ascending order of their phase of expression. For each window of phase of expression ($[\phi_l, \phi_h]$), $\chi^2$-test [5] was used to discover whether the window was enriched in the genes with the motif under investigation. This approach uses the sequence representation of motifs. Spellman et al. [15], Rustici et al. [13] and Peng et al. [10] identified the cell-cycle associated motifs using the same window-on-phase approach. The statistical significance of the motifs was evaluated by Hyper-geometric distribution [7] and Tang & Lewonton statistic [16] in Rustici et al. and Peng et al. respectively. This window-on-phase approach is biased towards motifs whose regulated genes mostly belong to same phase space and it is not shared by other genes controlled by different motifs.

The second approach, followed in several experiments, is to cluster the genes based on their expression profiles and motifs are found by searching over-represented sequences

in various clusters. This approach may not work well for cell-cycle regulated genes because of the nature of the cell-cycle expression data as discussed in Karuturi & Liu [8].

The most recent approach is to model gene expression as a function of the motifs in their promoter sequences. Bussmaker et al. [1] developed REDUCE which models gene expression as a linear function of the motif features. It uses sequence motif features and measures their relevance to a gene as number of times it occurred in its promoter region. To achieve computational feasibility, REDUCE includes motif-by-motif in the model. In this method, a single motif is fitted at each step and the error residual (the difference between the actual and predicted values) of the model is calculated and another best motif is found that reduces the most of the residual is selected and so on. REDUCE fits different models at different time-points. Similar approaches were developed and used by several other researchers later with changes to the initial motif set definition, feature measurement scheme, inclusion of motif-motif interaction terms in the model. Keles et al. [6] used the cell-cycle data published by Cho et al. [2] using both linear and interaction terms in the model with sequence motifs as features. The model was fit to the data using the feature (motif) selection procedure. Conlon et al. [3] proposed *MOTIF REGRESSOR* which uses MDSCAN [9] to select the initial set of sequence motifs present in the most induced or repressed genes. It further narrows down the motif set based on the single motif fit. This filtered set of sequence motifs were used to fit linear model for the gene expression. Using entire data from all microarrays using Spellman et al's $\alpha$-factor experiment they could discover 6 cell-cycle associated motifs (amounts to 60% sensitivity) of which only 3 motifs' (MCB,SCB and MCM1) weights show repeatability over the time-course. Das et al. [4] proposed *MARSMOTIF* which models the interaction terms using linear splines. It chooses initial candidate set of sequence motif pairs and fits the linear-splines model to discover the putative motifs. Using $\alpha 49$ experiment, it could discover 5 cell-cycle associated motifs (i.e. 50% sensitivity).

Since all these methods focussed on modeling genome-wide gene expression, they tend to miss cell-cycle associated motifs. Apart from this, they model gene expression at each time point independently which results in non-repeatability of motif prediction across time-course. Repeatability of prediction across time-course is important because the CDC regulated gene expression is measured for, approximately, two cell-cycles in most of the cell-cycle data and we expect any CDC associated motifs to be active in all cell-cycles at a particular phase. They also prioritize motifs by reducing the error in the model in successive steps which may lead to elimination of some important motifs since the genes that can be modeled by these motifs may be modeled by some alternative motifs in the earlier error reduction steps. The same approach also lead to the discovery of a few of the several possible binding motifs of the given cell-cycle regualted transcription factor. Whereas the single motif model, i.e. gene expression is modeled as a linear function of only one motif, will result in highly redundant representation of a given bindinig motif in the output list of motifs. This may cause some important motifs may be ranked low because of the redundancy of some dominant motif like MCB box in S. Cerevesiae.

We addressed the above problems by proposing a PWM motif detection methodology called ***ConstrainedMotif*** built around constrained linear model for gene expression. In this method, time-course gene expression is modeled as a linear function of the time-course scores of motifs by one-motif-fit-model. The model is constrained to produce periodic motif weights as expected from cell-cycle regulated genes. The methodology starts with a sequence based motif definition and finally produces ranked *position weight matrices (PWMs)* as motifs. Given a set of sequence motifs, each motif-weight time-course (is a series of weights assigned to a motif at different time points in the time-course) is found using the proposed parameter estimation procedure. It then chooses, for a given significance threshold, all significant motifs and generates overlapping clusters of them using both sequence and motif-weight time-course information. These clusters are used to generate the first set of PWMs which are again ranked by fitting the same model. The redundancy in motif representation among these PWMs may be reduced by clustering them and generating the next set of PWMs. This procedure goes on till we are comfortable with redundancy. The constraint of periodicity is important because any cell-cycle associated motif weight should smoothly reach its peak weight at the associated phase of cell-cycle. This peak should repeat for as many cycles as present in the data exactly at that phase of the cycle. The constraint is achieved by penalty based constrained optimization by introducing a penalty parameter $\gamma$. The term that penalizes for non-periodicity is the mean-squared error between motif weight time-course and $p\cos(\omega t - \phi)$, where $\omega = \frac{2\pi}{T}$ is the radial frequency of the cell-cycle and $T$ is the period of the cell-cycle. Both $p$ and $\phi$ are peak height and phase of the motif-weight time-course which are estimated as a part of model estimation.

We evaluated this methodology using cell-cycle data published by Spellman et al. to discover S. Cerevesiae cell-cycle associated motifs. We used 46 transcription factor binding motifs, of which 10 (MCB,SCB,SFF,ECB, SWI5,CPF1,MCM1,RAP1,ACE2,STE12) are shown to be cell-cycle associated, which we treat as true positives and the rest as true negatives. The resultant PWMs are a few and they represent most of these cell-cycle associated motifs as compared to the other methods known.

## 2. ConstrainedMotif Algorithm

This section describes the problem formulation and an algorithm to obtain position weight matrix representation of cell-cycle associated motifs. It also describes a methodology to match a given motif to a set of PWMs. The first subsection describes problem formulation, the second subsection describes an algorithm to obtain cell-cycle associated PWMs and the third subsection presents a methodology to test whether a given motif is cell-cycle associated using a given set of ranked PWMs.

### 2.1. Cell-cycle Regulated Gene Expression Model with Periodicity Constraint

Let $T$ be the period of cell-cycle in a given experiment. Consider $N$ genes $\{g_1, g_2, \ldots, g_N\}$ and $M$ time points $\{t_1, t_2, \ldots, t_M\}$. Let $D = (d_{ij})_{N \times M}$ be a $N \times M$ matrix representing the microarray measurements, where $d_{ij}$ represents the expression value of gene $g_i$ at time $t_j$.

Consider $K$ motifs $\{m_1, m_2, \ldots, m_K\}$. Let $S = (s_{mi})_{K \times N}$ be the $K \times N$ score matrix, where $s_{mi}$ is the score of motif $m_m$ on gene $g_i$. Here, we define the score $s_{mi}$ to be the number of occurrences of motif $m_m$ in the promoter region of gene $g_i$.

$D$ is transformed such that values in each column has zero-mean and unit-variance under the assumption that the each column follows a Normal distribution [7]. Similarly, $S$ is transformed so that each row has zero-mean and unit-variance.

Our algorithm computes the weight $w_{mj}$ for every motif $m_m$ at all times $t_j$. The weight $w_{mj}$ measures the binding strength of the motif $m_m$ at time $t_j$. The weights of the motifs is represented by the matrix $W = (w_{mj})_{K \times M}$ of size $K \times M$.

Let $A_{y\bullet}$ and $A_{\bullet z}$ represent $y^{th}$ row and $z^{th}$ column respectively of a matrix $A$.

Our aim is to model $D$ as a linear function of $S$ with parameters $W$ as in Bussmaker et al., i.e.

$$D = S^T W + \epsilon$$

where $\epsilon$ is the error between data and the model.

To find an appropriate $W$ we minimize the average squared error

$$\frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \epsilon_{ij}^2$$

which is equal to the mean squared objective function $O$ over $W$

$$\begin{aligned} O &= \tfrac{1}{NM} \sum_{n=1}^{N} \|(D_{n\bullet} - (S^T W)_{n\bullet}\|^2 \\ &= \tfrac{1}{NM} \sum_{n=1}^{N} \sum_{t=1}^{M} (d_{nt} - \sum_{k=1}^{K} w_{kt} s_{kn})^2 \end{aligned}$$

We expect the cell-cycle associated motifs should have periodic weights over time, we constrain $W_{k\bullet}$ to be periodic with period $T$. We achieve this constraint by requiring that the sum of the squared difference between actual weights $w_{it}$ and the best fit $P_i \cos(\omega t - \theta_i)$ is zero, where $\theta_i$ is the phase of the peak weight, $\omega = \frac{2\pi}{T}$.

$$Q = \frac{1}{KM} \sum_{k=1}^{K} \sum_{t=1}^{M} (w_{kt} - P_k \cos(\omega t - \theta_k))^2 = 0$$

The above constrained optimization problem has been transformed to unconstrained optimization problem by using penalty based methods [12] by using a penalty parameter $\gamma$. Now the unconstrained optimization problem for a given $\gamma$ is

$$\begin{aligned} \min_W O^c &= O + \gamma Q \\ &= \tfrac{1}{NM} \sum_{n=1}^{N} \sum_{t=1}^{M} (d_{nt} - \sum_{k=1}^{K} w_{kt} s_{kn})^2 \\ &+ \tfrac{\gamma}{KM} \sum_{k=1}^{K} \sum_{t=1}^{M} (w_{kt} - P_k \cos(\omega t - \theta_k))^2 \end{aligned}$$

Since the above optimization problem has too many variables to find exact solution and we aim to evaluate each motif based on its independent merit to fit the data, we reduce $O^c$ to $O_k^c$ which finds the best fit of each motif ($m_k$) while keeping its weight profile $W_{k\bullet}$ periodic across time-course.

$$\begin{aligned} \min_{W_{k\bullet}} O_k^c &= \tfrac{1}{NM} \sum_{n=1}^{N} \sum_{t=1}^{M} (d_{nt} - w_{kt} s_{kn})^2 \\ &+ \tfrac{\gamma}{M} \sum_{t=1}^{M} (w_{kt} - P_k \cos(\omega t - \theta_k))^2 \end{aligned}$$

For a given $M$ and $N$, $\gamma$ can absorb both $N$ and $M$ which means we can eliminate $M$ and $N$ from $O_k^c$. For computational convenience, the term $P_k \cos(\omega t - \theta_k)$ is rewritten as $A_k \cos(\omega t) + B_k \sin(\omega t)$ where $P_k = \sqrt{A_k^2 + B_k^2}$ and

$\theta_k = \tan^{-1}(B_k \div A_k)$. The resulting optimization problem, denoted by $F_k$, is

$$\min_{W_{k\bullet}} F_k = \sum_{n=1}^{N}\sum_{t=1}^{M}(d_{nt} - w_{kt}s_{kn})^2 + \gamma\sum_{t=1}^{M}(w_{kt} - A_k\cos(\omega t) - B_k\sin(\omega t))^2$$

To get optimal $w_{kt}, A_k, B_k$, partial derivates of $F_k$ with respect to variables $w_{kt}, A_k, B_k$ are equated to zero, i.e.

$$\frac{\partial F_k}{\partial w_{kt}} = 0; \frac{\partial F_k}{\partial A_k} = 0, \text{ and } \frac{\partial F_k}{\partial B_k} = 0$$

Performing certain algebraic manipulation along with the orthogonality assumption

$$\sum_{t=1}^{M}\cos(\omega t)\sin(\omega t) = 0$$

we get the following closed form solution for $w_{kt}, A_k, B_k$,

$$w_{kt} = \frac{(\sum_{n=1}^{N} d_{nt}s_{kn}) + \gamma A_k\cos(\omega t) + \gamma B_k\sin(\omega t)}{\gamma + 1}$$

$$A_k = \frac{\sum_{n=1}^{N}\sum_{t=1}^{M} d_{nt}s_{kn}\cos(\omega t)}{\sum_{t=1}^{M}\cos^2(\omega t)}$$

and

$$B_k = \frac{\sum_{n=1}^{N}\sum_{t=1}^{M} d_{nt}s_{kn}\sin(\omega t)}{\sum_{t=1}^{M}\sin^2(\omega t)}$$

From the above equations, one can calculate both $A_k$ and $B_k$ independently and then calculate $w_{kt}$. The error reduction obtained by including motif $m_k$, denoted by $\Delta F_k$ is

$$\Delta F_k = \frac{1}{\gamma+1}\sum_{t=1}^{M}\left(\left(\sum_{n=1}^{N} d_{nt}s_{kn}\right)^2 + A_k^2\gamma\cos^2(\omega t) + B_k^2\gamma\sin^2(\omega t)\right)$$

Let us define a random variable $Z_k$ which is zero mean and unit variance normal variate as

$$Z_k = \frac{\Delta F_k}{\sigma_k}$$

where $\sigma_k$ is standard deviation of $\Delta F_k$ which is a zero mean normal variate based on the assumption that $d_{nt}$, $s_{kn}$ are also zero mean normal variates. It is given by the following formula

$$\sigma_{kl}^2 = \frac{1}{N(\gamma+1)^2}\left\{\left(1 + \frac{\gamma\cos^2(\omega k)}{\sum_{z=1}^{M}\cos^2(\omega z)} + \frac{\gamma\sin^2(\omega k)}{\sum_{z=1}^{M}\sin^2(\omega z)}\right)^2 + \sum_{t\neq k, t=1}^{M}\gamma^2\left(\frac{\cos(\omega k)\cos(\omega t)}{\sum_{z=1}^{M}\cos^2(\omega z)} + \frac{\sin(\omega k)\sin(\omega t)}{\sum_{z=1}^{M}\sin^2(\omega z)}\right)^2\right\}$$

and

$$\sigma_l^2 = \frac{1}{M^2}\sum_{k=1}^{M}\sigma_{kl}^2$$

It is calculated from the equation of $\Delta F_k$ using the principle that the variance of the weighted sum of two zero mean normal variables $Y_1 = N(0, \sigma_1^2)$ and $Y_2 = N(0, \sigma_2^2)$, $Y = aY_1 + bY_2$ is also a zero mean normal variate with the resultant variance of $a^2\sigma_1^2 + b^2\sigma_2^2$.

The estimation of $Z_k$ gives $z - score$ of the motif $m_k$ from which the statistical significance of it can be found using $z - test$.

$$PValue(Z_k) = \int_{Z_k}^{\infty}\frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}dx$$

The next subsection describes an algorithm which uses the above formulation to derive the position weight matrix representation of cell-cycle associated motifs.

## 2.2. Algorithm to Predict Cell-cycle Associated PWM Motifs

Position Weight Matrix (PWM) of a motif of length $P$ is a $4 \times P$ dimensional matrix of real numbers ranging from zero to one. The rows represent one of the four bases $\{A, C, G, T\}$, and the columns represent position ($\{1, 2, 3, ..., P\}$) in the motif. An element $p_{ij}$ in a PWM denotes the probability of having base $b_i$ in position $j$ of the corresponding motif.

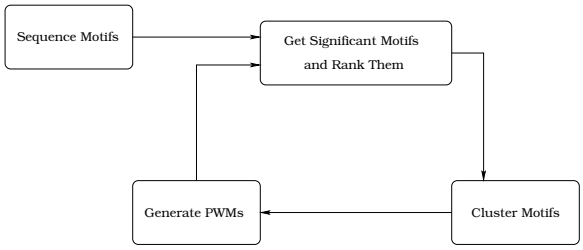


**Figure 1. The Overall Methodology to generate less redundant PWMs**

Our methodology is summarized in Figure 1. The algorithm proceeds in several motif list reduction steps to arrive at the final list of PWMs as follows: In the first step, we collect all possible sequence motifs of 4-7 bases long. Then, using the above formulation we get a set of significant motifs $m_k$ (whose z-test p-value $< 0.01$) ranked in the descending order of the reduction, $\Delta F_k$. This ranked list of significant motifs are clustered into various overlapping groups which in turn give rise to PWMs. These PWMs are scored against the list of promoters used to fit the model, and the significant PWMs are discovered. The significant PWMs are clustered and a new set of PWMs are generated. This procedure goes on for a fixed number of iterations or till the PWMs cannot be clustered anymore. The details of each of these steps are described below.

**Scoring and Ranking Motifs:** The score $s_{kn}$ of a sequence motif $m_k$ on gene $g_n$ is number of times the motif has occurred in the promoter region of the gene. When the motif $m_k$ is not a sequence motif but a PWM, the score between PWM $k$, denoted by $PWM^k$, and the gene $g_n$ is defined as follows.

Let the bases A,C,G and T be numbered as 1,2,3 and 4 respectively. Let $L_n$ and $L_k$ represent the length of $PrS^n$ and $PWM^k$ respectively, where $PrS^n$ is the promoter sequence of gene $g_n$. We denote $PrS^n_i$ be the base present at position $i$ of $PrS^n$.

$$s_{kn} = \frac{1}{L_n} \sum_{i=1}^{L_n} \left( \prod_{j=1}^{L_k} PWM^k_{(PrS^n_{(i+j-1)})j} \right)$$

After having obtained the normalized matrix $S$, using the above scoring functions, we proceed to find significant motifs for a given significance threshold for $Z_\bullet$. The significant motifs are ordered in the descending order of their reduction scores $\Delta F_\bullet$. We will cluster the motifs in this ranked list as described below.

**Clustering Motifs:** A given pair of motifs $m_a$ and $m_b$ are said to be similar in the context of our algorithm if the pearson correlation, denoted by $\rho_{ab}$, of their weight profles ($W_{a\bullet}$ and $W_{b\bullet}$) is above certain threshold $T_w$ and the sequence level similarity, denoted by $SS_{ab}$ between them is also more than a certain threshold $T^{ab}_s$.

The similarity score between a pair of sequence motifs is different from that of a pair of PWMs. For sequence motifs $m_a$ and $m_b$, $SS_{ab}$ is defined as the highest number of consecutive positions at which both sequences match in their content for any offset between them. $T^{ab}_s$ is the maximum similarity that could be achieved by two arbitrary sequence motifs of lengths of $m_a$ and $m_b$ with probability $\leq 0.05$.

For PWM motifs $PWM^a$ and $PWM^b$, $SS_{ab} = \max_i -SS^i_{ab}$ where $SS^i_{ab}$ is the error between $PWM^a$ and $PWM^b$ for an offset of $i$ positions of left most position of $PWM^b$ from that of $PWM^a$. Only those values of $i$ which gaurantee at least 3 position overlap between $PWM^a$ and $PWM^b$ are considered. $T^{ab}_s = T_s = -1.2$

$$SS_{ab} = \max_{3-|PWM^b| \leq i \leq |PWM^a|-3} -SS^i_{ab} =$$

$$-\sum_{tu} (PWM^a_{tu} - PWM^b_{t(u+i)})^2$$

where $|PWM^a|$ and $|PWM^b|$ are number of columns in $PWM^a$ and $PWM^b$ respectively.

Once all pairs whose $SS_{ab} >= T^{ab}_s$ and $\rho_{ab} >= T_w$ are identified using the above similarity measures, we proceed to generating clusters of motifs. The clustering takes the motifs in ranked order and proceeds from top rank to the prespecified rank. At each rank position, the clustering algorithm takes the motif at that rank as a seed for the new cluster. Cluster all significant motifs in the ranked list whose similarities with the seed motif satisfy the threshold constraints ($T_w$ and $T^{ab}_s$). The list of motifs, input for clustering, upto a given rank are called seed motifs. A seed motif may be part of the cluster of another seed motif. If the seed motif is a sequence then the motifs (others will also be sequences) that paired with it will not be removed from the seed list. Whereas, if the seed motif is a PWM, then the other PWMs in the list that paired with it are removed from the seed PWM list. The clusters with fewer sequence motifs will be ignored since we expect that each single relevant sequence motif is represented in several approximate forms in our list of significant sequence motifs.

These clusters are then used to generate one PWM for each cluster. The process of generating PWM from a cluster of motifs is described below.

**Generating PWMs:** After having generated clusters of motifs, we need to find one PWM for each cluster. The first step in generating a PWM from a cluster of motifs is to align the motifs in the cluster with respect to the seed motif. Then we collect the count statistics of A,C,G and

T at each position in the alignment. Finally, for each position, the probabilities of occurrences of any of the bases is estimated to be the proportion of occurrences of that base. The collection of count statistics are simple for a cluster of sequence motifs is straight forward. But, for a cluster of PWM motifs, the count statistics are performed by maintaining the count statistics of A,C,G and T for each PWM from the earlier PWM generation steps and normalize the counts accordingly.

### 2.3. Criteria to Match Box Motifs to PWMs

A box motif $Sq$ is given as a set of sequence motifs representing a putative binding site. $Sq$ is said to have matched to a PWM, if at least 30% of the sequences in $Sq$ are represented by that PWM. A sequence motif $Sq_i \in Sq$ is said to be represented by the $PWM^j$ if

$$\log_2 \left( \frac{Sim(Sq_i, PWM^j)}{\frac{1}{|SQ(|Sq_i|)|} \sum_{sq \in SQ(|Sq_i|)} Sim(sq, PWM^j)} \right) \geq T_r$$

where $SQ(l)$ is the set of all sequences of length $l$. $|SQ(l)|$ is the number of sequences in $SQ(l)$. $|Sq_i|$ is the number of bases in $Sq_i$. $Sim(Sq_i, PWM^j)$ is the similarity between sequence $Sq_i$ and the position weight matrix $PWM^j$. This is defined as the highest probability of generating $Sq_i$ from $PWM^j$. This is achieved by appropriately aligning $Sq_i$ with $PWM^j$ and multiplying the probability of generating the respective bases in $Sq_i$ at the position.

The next section presents the results of evaluation of the above algorithm on the cell-cycle regulated gene expression data published by Spellman et al. for S. Cerevesiae.

## 3. Evaluation of ConstrainedMotif

We evaluated the above methodology by predicting cell-cycle associated motifs for S. Cerevesiae. We used time-course microarray gene expression data of 798 putative cell-cycle regulated genes published by Spellman et al. The data contains three sets of experiments conducted by different methods of cell synchronization: (1) $\alpha$-factorization ($\alpha$49); (2) CDC15 block and release; and, (3) CDC28 block and release. Table 1 summarizes the data statistics. The promoter sequences for upto 800 bases upstream from the start codon of the genes were collected from SCPD database [14] for all these genes.

The above procedure has been applied to all three datasets and three lists of PWMs were obtained. To get these lists, the motif clustering and PWM generation procedure was carried out for two times. The PWMs output by the algorithm at the end of the second round were ranked and output as the final list. The PWMs were ranked in the descending order of the error reduction ($\Delta F_k$) they achieved on the respective dataset. To test the goodness of the results, we collected a list of 46 motifs related to S. Cerevesiae cell-cycle. Of these, 10 motifs are known to be cell-cycle associated. They are: MCB, SCB, SFF, ACE2, CPF1, RAP1, SWI5, ECB, STE12, MCM1. We use these motifs, except MCM1 since it is long and our method mainly finds motifs of length 4-8 bases, as true positives. The remaining 36 motifs were considered as true negatives.

| Statistics | $\alpha$49 | CDC15 | CDC28 |
|---|---|---|---|
| T (in minutes) | 60 | 120 | 90 |
| M | 18 | 24 | 17 |
| Number of cycles covered | 2 | 2.3 | 1.8 |
| samples Interval (mins) | 7 | 10 | 10 |

**Table 1. The summary of the time-course data used for various synchronization experiments.**

*Sensitivity* is defined as the fraction of true positives selected and *specificity* is defined as the fraction of the selected motifs are the true positives.

Each of the 46 motifs is represented by a set of sequences. We used the method described in Section 2.3 to evaluate the goodness that a PWM matches a set of sequences with the threshold $T_r = 3.16$. The rank of the motif after analyzing a dataset is the rank of the best ranked PWM representing it. We assigned overall rank to a motif as the highest of the ranks obtained from all three datasets. This approach gaurantees that the given motif is represented in the set of top PWMs, equal to the overall rank of the motif, the motif is represented with the given $T_r$. Then all 46 motifs were ranked according to their overall rank. Figure 2 shows sensitivity-specificity plots for various values of $\gamma$. It shows that the specificity for a given sensitivity is relatively and significantly better when $\gamma = 5$ than when $\gamma = 0$. Note that, $\gamma = 0$ is equivalent to the weight profiles one would get if REDUCE were used at all time points. The choice of $\gamma = 7$ performs better than the other two choices at low sensitivity level. The setting of $\gamma = 5$ outperforms remaining two choices at moderately higher sensitivity and outperforms $\gamma = 0$ at all sensitivity levels except at very high sensitivity level. Apart from doing better than the case for $\gamma = 0$, the sensitivity (78%) achieved here is better than that of Conlon et al. (60% sensitivity) and Das et al. (50% sensitivity) at similar precision levels as summarized in table 2.

Similar conclusions can be drawn from Figure 3, which shows the variation of sensitivity along with the list of the 46 motifs ranked according to the overall rank. Figure 4 shows the variation of sensitivity as the rank of the best

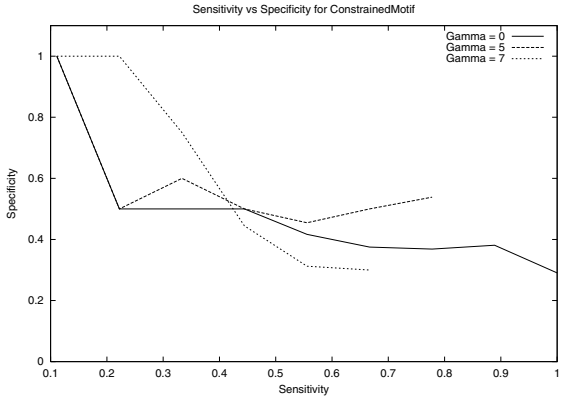


**Figure 2. Sensitivity vs. specificity plots for various values of $\gamma$ (written as Gamma in the plot) using maximum rank approach for motif ranking.**

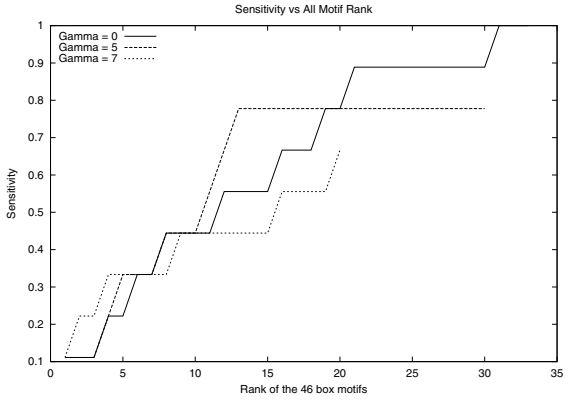


**Figure 3. Sensitivity in the ranked list of 46 motifs based on maximum-rank approach for various values of $\gamma$ (written as Gamma in the plot).**

| ConstrinedMotif (Gamma ($\gamma$) =7) | ConstrinedMotif (Gamma ($\gamma$) = 0) | Conlon et al. | Das et al. |
|---|---|---|---|
| 78% | 22% | 60% | 50% |

**Table 2. The sensitivity of various methods at an approximate specificity of 58%.**

matching PWMs with the given $T_r = 9$. It clearly shows that $\gamma > 0$ filters out a lot of redundant or incorrect PWMs by about 10 times as compared to the case of $\gamma = 0$. The relative performance of settings of $\gamma = 5$ and $\gamma = 7$ is as discussed earlier. The relative performance of the setting $\gamma = 5$ is clearly superior to the setting of $\gamma = 0$ upto 80% sensitivity.

## 4. Conclusions and Future Directions

We presented a methodology called ***ConstrainedMotif*** to predict cell-cycle associated motifs. The application of the proposed methodology to the cell-cycle microarray data published by Spellman et al. yielded good sensitivity and specificity when $\gamma = 5$ as compared to the prediction with $\gamma = 0$. This relative performance is consistent with the way the motifs were ranked using the three datasets. This shows that our formulation is effective. The proposed methodology alleviates the problem of redundant representation of a motif in the predicted list of motifs, by adopting PWM definition of motifs. The methodology gaurantees that the most of the sequences of a motif are represented by the corresponding matching PWM. The methodology also removes the problem of nonrepeated motif weights along the time-course of the experiment. The results showed that our method has sensitivity, at 58% specificity, upto 78% which is much better than the ones reported for cell-cycle associated motif prediction as shown in table 2

We can identify a numerous future directions starting from the formulation, the algorithm and the results presented in this paper. One direction is to improve the motif scoring scheme and the idenitification of cluster and the alignemnt of sequence motifs (and PWMs). This may improve the performance of the algorithm further. Finding motif-motif interaction and very long PWMs using short-to-medium length PWMs using this approach may be easily carried out. By looking for pairs of PWMs which has similar phase of peak weight, i.e. $\theta_\bullet$, resulting in better significance value and aligning them appropriately may be informative of both motif-motif interaction and long PWMs. Another direction could be is to use promoter-position dependent motif features for gene expression modeling and motif prediction. Finally, the above methodology may be extended to a general time-course data and its application to predict the experiment specific motifs and their time of activity.

## 5. Acknowledgements

We thank Jian Hua Liu, Joshy George and Edison T. Liu for their valuable and timely suggestions and help during this work.

## References

[1] Bussemaker H.J., Li H. and Siggia E.D., Regulatory Element Detection using Correlation with Expression, *Nature Genetics*, 27:167-171, 2001.

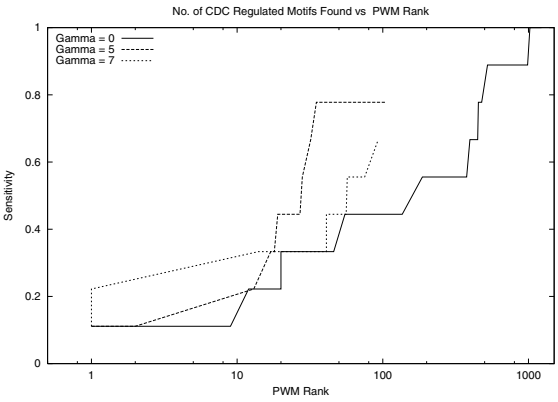


**Figure 4. Sensitivity in the ranked list of PWMs for various values of $\gamma$ (written as Gamma in the plot) using maximum rank approach.**


[2] Cho R.J., Campbell M.J., Winzeler E.A., Steinmetz L., Conway A., Wodicka,L., Wolfsberg,T.G., Gabrielian,A.E., Landsman,D., Lockhart,D.J. and Davis,R.W., A genome-wide transcriptional analysis of the mitotic cell cycle, *Mol. Cell*, 2:65-73, 1998.

[3] Conlon E.M., Liu X.S., Lieb J.D. and Liu J.S., Integrating Regulatory Motif Discovery and Genome-Wide Expression Analysis, *PNAS*, 100(6):3339-3344, 2003.

[4] Das D., Banerjee N. and Zhang M.Q, Interacting Models of Cooperative Gene Regulation, *PNAS*, 101(46):16234-16239, 2004.

[5] Kanji G.K., *100 Statistical Tests*, SAGE Publications, 1999.

[6] Keles S., Van Der Laan M. and Eisen M.B., Identification of Regulatory Elements Using a Feature Selection Method, *Bioinformatics*, 18(9):1167-1175, 2002.

[7] Merran E., *Statistical Distributions*, John Wiley and Sons, New York, 2000.

[8] Karuturi R. K. and Liu J.H., Improved Fourier Transform Method for Unsupervised Cell-cycle Regulated Gene Prediction. in the Proc. of *IEEE Computational Systems Bioinformatics Conference (CSB04)*, Augst 8-12, Stanford, CA, 2004.

[9] Liu X.S., Brutlag D.L. and Liu J.S., An Algorithm for Finding Protein DNA Binding Sites with Application to Chromatin-Immunoprecipitation Microarray Experiments, *Nature Biotechnology*, 20:835-839, 2002.

[10] Peng X., Karuturi R.K., Lance M.D., Lin Kui, Yonghui J., Kondu P., Wang L., Wong L., Liu T.E., Balasubramanian M., Liu J.H., Identification of Cell Cylce-regulated Genes in Fission Yeast, *Molecular Biology of the Cell*, 16(3):1026-1042, 2005.

[11] Phuong T.M., Lee D. and Lee K.H., Regression Trees for Regulatory Element Identification, *Bioinformatics*, 20(5):750-757, 2004.

[12] Rao S.S., *Optimization: Theory and Applications*, John Wiley and Sons, New York, 1979.

[13] Rustici G., Mata J., Kivinen K., Lio P., Penkett C.J., Burns G., Hayles J., Brazma A., Nurse P. and Bahler J., Periodic Gene Expression Program of the Fission Yeast Cell Cycle, *Nature Genetics*, 36:809-817, 2004.

[14] SCPD: The promoter Database of Sachcharomyces Cerevisiae. *htt://cgsigma.cshl.org/jian*

[15] Spellman P.T., Sherlock G., Zhang M.Q., Iyer V.R., Anders K., Eisen M.B., Brown P.O., Bostein D. and Futcher B., Comprehensive Identification of Cell Cycle-regulated Genes of the S. cerevisiae Saccharomyces cerevisiae by Microarray Hybridization, *Molecular Biology of the Cell*, 9:3273-3297, 1998.

[16] Tang H. and Lewontin R.C., Locating Regions of Differential Variability in DNA and Protein Sequences, *Genetics*, 153:485-495, 1999.

[17] Wolfsberg T.G., Gabrielian A.E., Campbell M.J., Cho R.J., Spouge J.L. and Landsman D., Candidate Regulatory Sequence Elements for Cell Cycle-Dependent Transcription in Saccharomyces cerevisiae, *Genome Research*, 9:775-792, 1999.