Correlation Analysis between Regulatory Sequence Motifs and Expression Profiles by Kernel CCA

Je-Keun Rhee[1,2] **Je-Gun Joung**[1,2] **Jeong-Ho Chang**[3] **Byoung-Tak Zhang**[1,2,3]
[1]*Graduate Program in Bioinformatics, Seoul National University, Seoul, Korea*
[2]*Center for Bioinformation Technology, Seoul National University, Seoul, Korea*
[3]*School of Computer Science and Engineering, Seoul National University, Seoul, Korea*
Email : jkrhee@bi.snu.ac.kr, jgjoung@bi.snu.ac.kr, jhchang@bi.snu.ac.kr, btzhang@bi.snu.ac.kr

ABSTRACT: Transcription factors regulate gene expression by binding to gene upstream region. Each transcription factor has the specific binding site in promoter region. So the analysis of gene upstream sequence is necessary for understanding regulatory mechanism of genes, under a plausible idea that assumption that DNA sequence motif profiles are closely related to gene expression behaviors of the corresponding genes. Here, we present an effective approach to the analysis of the relation between gene expression profiles and gene upstream sequences on the basis of kernel canonical correlation analysis (kernel CCA). Kernel CCA is a useful method for finding relationships underlying between two different data sets. In the application to a yeast cell cycle data set, it is shown that gene upstream sequence profile is closely related to gene expression patterns in terms of canonical correlation scores. By the further analysis of the contributing values or weights of sequence motifs in the construction of a pair of sequence motif profiles and expression profiles, we show that the proposed method can identify significant DNA sequence motifs involved with some specific gene expression patterns, including some well known motifs and those putative, in the process of the yeast cell cycle.

1 INTRODUCION

Gene regulation is one of key mechanisms in cellular processes in living organisms. Although genes are basically transcribed by RNA polymerase, the expression level of genes is regulated by many transcription factors (TFs). Transcription factors are proteins which are bound to the promoter region and thereby regulate gene expression. The upstream region of genes has specific transcription factor binding sites (TFBSs) for each transcription factors. The TFBS is also called a regulatory sequence motif. Since gene expression is mainly regulated through the binding of TFs to their specific TFBSs, it is highly likely that there is a close relationship between sequence motif profiles and expression profiles of genes.

At past, the study on gene regulation was mainly based on wet-lab experiments. But as high-throughput techniques such as microarray experiments develop, the schemes to study biological issues are being diversified. Especially the high-throughput technologies make it possible for one to handle and to analyze a lot of diverse data by computational methods. Spellman *et al*. [1] and Cho *et al*. [2] analyzed the massive gene expression profiles during the cell cycle process of the yeast *Saccharomyces cerevisiae* by the application of computational methods to the high-throughput microarray data. Their researches generated valuable data for yeast gene expression pattern analysis. After them, many researches through microarray data analysis have been actively done for the yeast genome. It was shown that one can identify many useful biological facts through the application of a variety of computational methods, including clustering and dimensionality reduction, to the microarray-based gene expression data [3, 4].

In particular, there have been many researches which attempted to find regulatory motifs from gene upstream sequences. One useful method is based on a statistical analysis of upstream sequence of genes. It works by directly searching for significant sequence motifs on gene upstream sequences, using such methods as maximum-likelihood estimation or Gibbs-sampling [5, 6]. Other researches using alternative methods have also been progressed. Recently, there was a research to find regulatory sequence motifs using a SOM (self-organizing map)-based clustering method [7].

In another way, there have been many studies for the identification of regulatory sequence motifs by linking gene expression patterns and DNA sequence motif profiles. The approaches of this kind can enhance one’s understanding of gene regulation, in addition to the finding of regulatory sequence motifs themselves. Tavazoie *et al.* [8] predicted genetic regulation and function of each ORF by regulatory motif analysis. Recently, in addition, Beer and Tavazoie [9] presented that gene expression profiles can be predicted from sequence information by Bayesian networks, to a relatively high accuracy [9]. Although many biological studies about gene regulation are progressed together with vast amount of data and various computational methods, the definite relationship between sequences and expression patterns has been unknown so far.

In this paper, we analyze the effect to regulation over gene upstream sequences using Kernel canonical correlation analysis (kernel CCA). Kernel CCA is a method for investigating relationships between two different data [11, 12, 13]. It works by first (implicitly) mapping each data point to higher dimension space than the original input space and then by analyzing the relationship between each projected component using kernel trick.

We apply the kernel CCA to a paired set of gene upstream sequence profiles and gene expression profiles of the yeast *Saccharomyces cerevisiae*. By the application, we inquire whether there is a significant relation between the two profiles. We also search for significant sequence regulatory motifs for specific expression patterns by analyzing the contributing values or weights of those motifs in relating the two different profiles. Eventually, we show that our method can identify significant sequence motif, some well known and others putative, which affects gene regulation.

2 METHODS

2.1 Investigation of the Relationship between Expression Profiles and Sequence Motif Profiles

Canonical correlation analysis (CCA) [10] is a classical multivariate statistical method for finding linearly correlated features from a pair of different data sets. The Kernel CCA is a version of nonlinear CCA where the kernel trick is utilized to find nonlinearly correlated features from two data sets [11, 12, 13]. In other words, while CCA is limited to linear features, kernel CCA can capture non-linear relations. Kernel CCA has shown its effectiveness in several applications including text retrieval [11], biological data analysis [14], and so on.

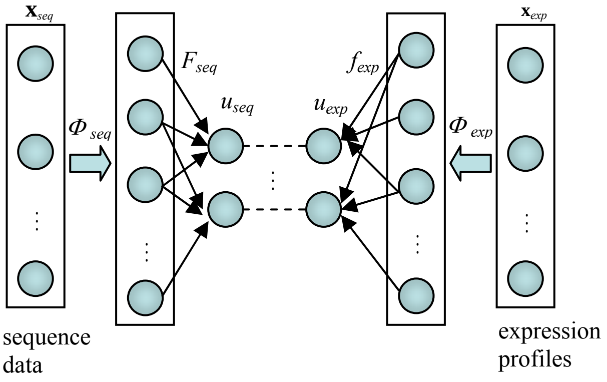

Figure 1: The basic scheme of Kernel CCA. The sequence data and expression data are transformed to Hilbert space by ϕ function. By taking inner products, we can find *uexp* and *useq* which maximizes the correlation between the two data sets.

Figure 1 illustrates a basic scheme of Kernel CCA in our application of the integrated analysis of DNA sequence motif data and gene expression data. By using Kernel CCA, we try to find maximal correlated features between the gene expression and the sequence motif profile. Here, a gene x is represented by two separate profiles in terms of its transcriptional behavior and its upstream sequence, that is, by $\mathbf{x}_{exp}=(e_1, e_2, ..., e_N)$ and $\mathbf{x}_{seq}=(m_1, m_2, ..., m_M)$ respectively. The value e_i ($1 \le i \le N$) is the expression value of the gene in the i-th sample or experimental condition from a microarray data and m_j ($1 \le j \le M$) denotes the occurrence frequency of the j-th sequence motif in the upstream region of the gene. For the detection of the correlated features between the two data sets, $\mathbf{x}_{exp}$ and $\mathbf{x}_{seq}$ are first mapped to Hilbert space, H, by function ϕ. That is, each x is projected into two direction f_{exp} and f_{seq} in Hilbert space according to its representation,

$$u_{exp} = \left\langle f_{exp}, \phi_{exp}(\mathbf{x}_{exp}) \right\rangle \qquad (1)$$

$$u_{seq} = \left\langle f_{seq}, \phi_{seq}(\mathbf{x}_{seq}) \right\rangle \qquad (2)$$

where $\langle \cdot,\cdot \rangle$ denotes the dot product. Kernel CCA aims at finding maximally correlated features between $\{\mathbf{x}_{exp}\}$ and $\{\mathbf{x}_{seq}\}$,

$$\gamma(f_{exp}, f_{seq}) = \max \frac{\mathrm{cov}(u_{exp}, u_{seq})}{(\mathrm{var}(u_{exp}) + \lambda_{exp} \| f_{exp} \|^2)^{\frac{1}{2}} (\mathrm{var}(u_{seq}) + \lambda_{seq} \| f_{seq} \|^2)^{\frac{1}{2}}}, \qquad (3)$$

where λ_{exp} and λ_{seq} are regularization parameters. The f_{exp} and f_{seq} can be found by solving the following Lagrangean:

$$\begin{aligned} L_0 &= E[(u_{exp} - E[u_{exp}])(u_{seq} - E[u_{seq}])] \\ &- \frac{\rho_{exp}}{2} E[(u_{exp} - E[u_{exp}])^2] \\ &- \frac{\rho_{seq}}{2} E[(u_{seq} - E[u_{seq}])^2], \end{aligned} \qquad (4)$$

where ρ_{exp} and ρ_{seq} are Lagrangean multipliers. Again, the Lagrangean of Equation (4) can be rewritten in terms of kernel matrices,

$$\begin{aligned} L &= \alpha_{exp}^T \mathbf{K}_{exp} \mathbf{K}_{seq} \alpha_{seq} \\ &- \frac{\rho_{exp}}{2} \alpha_{exp}^T (\mathbf{K}_{exp} + \lambda_{exp} \mathbf{I})^2 \alpha_{exp} \\ &- \frac{\rho_{seq}}{2} \alpha_{seq}^T (\mathbf{K}_{seq} + \lambda_{seq} \mathbf{I})^2 \alpha_{seq}, \end{aligned} \qquad (5)$$

where $\mathbf{I}$ denotes the identity matrix. $\mathbf{K}_{exp}$ is the kernel matrix for expression profile data and $\mathbf{K}_{seq}$ is the kernel matrix for sequence motif profile data. By mediating regularization parameter λ, Lagrangean value is maximized. Finally, the solution of the kernel CCA can be given by solving a generalized eigenvalue problem,

$$\begin{aligned} &\begin{pmatrix} 0 & \mathbf{K}_{exp}\mathbf{K}_{seq} \\ \mathbf{K}_{seq}\mathbf{K}_{exp} & 0 \end{pmatrix} \begin{pmatrix} \alpha_{exp} \\ \alpha_{seq} \end{pmatrix} = \\ &\rho \begin{pmatrix} (\mathbf{K}_{exp} + \frac{n\lambda_{exp}}{2}\mathbf{I})^2 & 0 \\ 0 & (\mathbf{K}_{seq} + \frac{n\lambda_{seq}}{2}\mathbf{I})^2 \end{pmatrix} \begin{pmatrix} \alpha_{exp} \\ \alpha_{seq} \end{pmatrix}. \end{aligned} \qquad (6)$$

When given α_{exp} and α_{seq} as the solution of the above generalized eigenvalue problem with the largest eigenvalue, canonical correlation scores (CC score) for $\mathbf{x}_{seq}$ and $\mathbf{x}_{exp}$ are estimated by $u_{seq} = \mathbf{K}_{seq}\alpha_{seq}$ and $u_{exp} = \mathbf{K}_{exp}\alpha_{exp}$. The CC scores are the low dimensional mapping of genes in terms of two separate representations and can be used to show the salient correlation between the two. Once we have obtained α vector, the weights of motif and expression profile, $\mathbf{W}_{seq}$ and $\mathbf{W}_{exp}$, are also obtained as follows:

$$\mathbf{W}_{exp} = \mathbf{X}_{exp}^T \alpha_{exp} \qquad (7)$$

$$\mathbf{W}_{seq} = \mathbf{X}_{seq}^{T}\alpha_{seq} \tag{8}$$

As the weight of specific sequence motif gets high value, the motif can be said to be strongly correlated with the expression patterns of those genes with the high-valued CC scores. That is, if the absolute value of the weight of a specific motif is high, the motif is a candidate to be investigated for its effectiveness in regulating genes of which upstream region contain it.

2.2 Data Representation

We used microarray data by Spellman for getting expression profiles of all ORFs (open reading frames) in yeast [1]. It presents the expression profile during the cell cycle. It consists of total 18 time points in alpha factor synchronization case.

The sequence data is experimented in two cases. First, we used motif data extracted by Pilpel [15]. These are composed of 42 motifs. We extracted motif information in each ORF using AlignACE program [5]. At first, we analyzed relationship between these data through Kernel CCA methods.

Next, we analyzed the relationship with expression profiles using raw sequence data. We extracted gene upstream sequences as the size of 985 bases from each gene. From base sequences, we calculated frequency of *n*-mer base combination in each gene. For example, if *n* is 4, every gene has 256 (= 4^4) base combinations, and if *n* is 5, each gene has 1024 (= 4^5) base combinations. And then, the frequency of each base combination is counted. And we analyzed data in the similar manner.

But these data themselves described above cannot be a Kernel matrix. Therefore we should convert these matrices to the Kernel matrix. We applied a linear kernel or a polynomial kernel to the sequence frequency matrix by $k(\mathbf{x}_{seq}, \mathbf{x}'_{seq}) = (\mathbf{x}_{seq}^{T}\mathbf{x}'_{seq})^{d}$. If *d* is 1, it is linear kernel and if other number, it is polynomial kernel.

In the gene expression data, we used a Gaussian RBF kernel: $k(\mathbf{x}_{exp}, \mathbf{x}'_{exp}) = \exp\left[-\frac{d(\mathbf{x}_{exp}, \mathbf{x}'_{exp})}{2\sigma^2}\right]$, where σ is a parameter and function *d* is a Euclidean distance. We analyzed the relationship of them using kernel CCA.

3 RESULTS

First of all, we analyzed the relationship of expression data with motif data extracted by AlignACE. We used total 551 ORFs related to cell cycle. The degree of polynomial kernel sets 3 in the sequence data. The parameter σ is 0.5 in Gaussian RBF kernel applying to expression data, and the regularization parameter is 0.1. After applying kernel CCA methods, we plotted using the CC1 score. The CC1 score means the first canonical correlation score. The plot of CC1 score shows close relationship between sequence information and expression profiles (Figure 2).

In Figure 2, each point corresponds to one gene and the diagonal shapes of points mean that correlation has been detected. The right side one of Figure 2 is to magnify parts crowded with points.

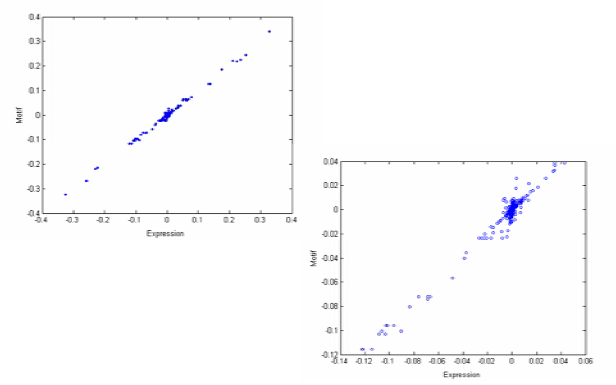

Figure 2: CC1 Score (Expression profile vs. Motif data). This plot shows diagonal shapes that mean close relationship between the expression data and the sequence data

Motif	Weight	Function
SWI5	0.89026	Binding site in transcription factor that activates transcription of genes expressed in G1 phase and at the G1/M boundary
SFF’	0.45399	Binding site in transcription factor FKH1 of the forkhead family that regulates the cell cycle
MCB	0.29633	Binding site in transcription factor MBF that activates in late G1 phase
LYS14	0.21796	Transcriptional activator involved in regulation of genes of the lysine biosynthesis pathway
ALPHA2	0.16532	Silenced copy of ALPHA2, encoding a homeobox-domain containing protein that associates with Mcm1p in haploid cells to repress a-specific gene expression and interacts with A1p in diploid cells to repress haploid-specific gene expression

Table 1: High scored motifs. The table shows the top 5 rank of the results. The highly ranked motifs have been known as significant motifs like SWI5, SFF’, and MCB motif.

The results of significant motifs searched by weight function are shown in Table 1. In Table 1, SWI5 motif has the highest weight. SWI5 is one of typical transcription factors in yeast cell cycle. It has been known as acting in G1 phase and M/G1 boundary in cell cycle [16, 17]. SFF’ motif is a binding site of FKH1 transcription factor. FKH1 forms SFF protein with NDD1. In SFF, FKH1 is a DNA binding component and NDD1 acts transcriptional regulatory role. SFF is also an important protein in cell cycle, and works mainly late S phase [18].

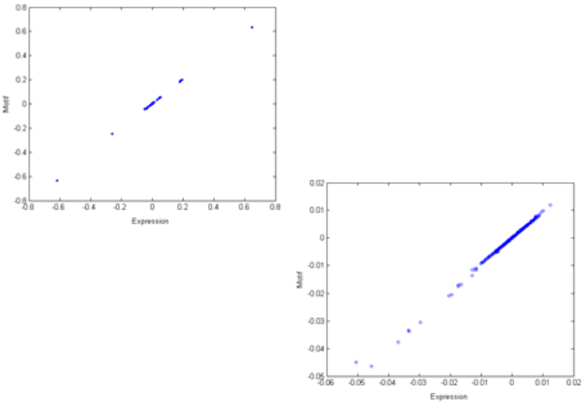

Figure 3: CC1 Scores using raw upstream sequence data. This plot shows the expression data also closely related with the motif data using 5-mer base sequences.

MCB motif is a very famous motif in yeast cell cycle. It is a binding site in MBF protein. MBF protein is composed of MBP1 and SWI6. In MBF protein, the DNA binding component is MBP1 and the regulatory component is SWI6. It is well known that the MBF protein regulates transcription of genes in late G1 phase [18, 19]. ALPHA2 protein is also inferred that it is related to cell cycle. That protein can be associated with MCM1 protein and represses other genes [20, 21]. MCM1 protein is also known as a significant protein in cell cycle [18, 22].

In our results, motifs of relatively high weight were known motifs related to cell cycle. Therefore our methods using Kernel CCA could be validated.

At the second experiment, we used the raw upstream gene sequence data. The sequence window size, *n*, is 5, so there are total 1024 (=4^5) attributes. Then the linear kernel was applied to the sequence data, and Gaussian RBF kernel was applied to the expression data at parameter σ value of 0.3. The regularization parameter is 0.1. Also, in this case, CC1 score and its plot show close relationship between the sequence data and expression profiles (Figure 3).

Like Figure 2, the right side graph is one enlarged the specific clouded part. Motifs having higher weight are shown in table 2.

The results show the 5'-GCGTG-3' sequence has the highest weight score. It is similar to MCB sequence (5'-ACGCGT-3'). The sequence ranked in the highest position appears only one base shift from MCB sequence. As we explained previously, MCB is an important motif in cell cycle. Motifs ranked in high position can be mostly supposed to be important motif in cell cycle.

Until now, we only analyzed the first component. We also observed the second component results. Table 3 shows the second component results. From the second component, we found the sequence identical to MCB motif. The fifth rank sequence (5'-CGCGT-3') is exactly same to known MCB motif sequence (5'-ACGCGT-3'). That is, it could be re-confirmed MCB motif strongly affecting gene expression in cell cycle. And also, the sequence similar to SCB motif has high score. In our results, the fourth rank sequence is 5'-CCACG-3'. It is also only one base shift from known SCB sequence (5'-CACGAAA-3').

Sequence	Weight	Motif information
GCGTG	0.079567	MCB-like sequence (ACGCGT)
CGTGT	0.075340	MATalpha2-like (CRTGTWWWW)
TGCGT	0.063041	-
TTGCG	0.057494	-
CATCA	0.054292	-
CATGA	0.050729	-
GCATG	0.049969	-
GATCA	0.049208	-
ATGTG	0.048790	-
TTAGA	0.047648	-
TGTCA	0.046667	-
CATGT	0.046299	MATalpha2-like (CRTGTWWWW)
CCGGA	0.044133	MCM1-like sequence (CCNNNWWRGG)
CTAGA	0.042840	-
TAAGG	0.042387	MCM1-like sequence (CCNNNWWRGG)

Table 2: High scored motifs in the first component using 5-mer base combination data. This table shows top 15 results. The sequence inferred to MCB motif is top ranked.

SCB motif is a representative important motif with MCB motif in yeast cell cycle. SCB motif is a binding site of SBF protein. SBF protein is constituted SWI4 and SWI6. SWI4 is a DNA binding component and SWI6 has a regulatory role [18]. And other sequences in higher position can also infer to significant motifs. Therefore, as we inspect closely the second component as well as the first component, we could find other meaningful results.

4 DISCUSSION

We could obtain meaningful results through applying kernel methods, mapping raw data to higher dimensional space, to classical statistical methods. By the methods called Kernel CCA, it was possible to analyze the relationship between promoter sequences and expression profiles, and, as the result, we showed close relationship between two data through CC score. Besides, we could find significant motifs affecting gene expression in yeast cell cycle. It was also possible to confirm that our results are agreed with previous works. Furthermore, we can also find putative important regulatory motifs.

Sequence	Weight	Motif information
GGCGA	0.019679	-
CGGAA	0.019447	-
GAACG	0.019151	-
CCACG	0.018992	SCB-like sequence (CACGAAA)
CGCGT	0.017870	MCB-like sequence (ACGCGT)
ACCTG	0.017713	-
GCACT	0.017208	-
CCTCG	0.016666	-
GTGTT	0.016595	MATalpha2-like (CRTGTWWWW)
GGACC	0.015884	-
TGGCC	0.015816	-
GTCCG	0.015422	-
CGGAG	0.015207	-
CAGGC	0.015203	-
GGCGT	0.014963	-

Table 3: Top 15 ranked motifs in the second component. The motif inferred to MCB or SCB motif is located in high position

Previously, many works are mainly performed using one type data. But as amounts of various data and information increase, it has been important to integrate many data. Actually, if we use diverse information for specific research purpose, it has been known the fact that the results are much better. Kernel CCA has the advantage of using data of different types together. Such as our works, each different kernel is applied to each data, and we can analyze the relationship between different data in higher dimension space. And we can find more significant features using weight function applied to each attribute.

But it remains some difficult aspects in our method. Using extracted motif data, we could get relatively good results. But it is slightly hard to interpret results using raw upstream sequence data. We could experiment with only 5-mer case or under that for computing power. Since the size of most known motifs are over 5-mer, it is short window size. But if we expand the window size, the base combination number also increases exponentially (4^6=4096, 4^7=16384, etc.). In that case, it is hard to learn data due to many attribute number. And our methods are possible to apply only the priorly decided base size. That is, it doesn't reflect individually different motif sizes. Also, it is difficult to perform using all genes at PC level.

If the computing power is supported, our methods can be new motif finding algorithms. Although the sequence size is decided constantly, if the experiment is performed repeatedly about various sizes, it will be able to find the motif sequence of various sequence length. That is, we can be to find the new motifs from gene expression profiles.

We used yeast as the model organism and analyzed using public cell cycle microarray data. Our methods can easily apply to other organisms. Of course, it is also possible to apply to other expression data not cell cycle. In the yeast case, although many related researches have been preceded, it is lots of lack and hardness of related works in many organisms including human. If our methods apply to research using other data, it also helps the understanding in complex organism.

Furthermore, if we expand our methods, we will be able to find the synergistic motif combination. Generally, the transcription of gene is regulated by many complex mechanisms. That is, gene regulation is affected by combination of many transcription factors. Therefore it will be valuable works to find the synergistic motif combination.

Our results make it possible to estimate the expression patterns from upstream sequences. Conversely, given an expression profiles, we will also be able to predict the key role motifs in the upstream region of a gene. From these, it is possible to predict important transcription factors regulating the specific gene. Therefore it helps the regulatory mechanism prediction, and it will make complex gene regulatory process more brightly.

ACKNOWLEDGEMENT

This work was supported by NRL, the Korean Ministry of Science and Technology

REFERENCES

[1] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. Molecular Biology of the Cell, 9: 3273--3297, 1998.

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

[3] J. Kasturi and R. Acharya. Clustering of diverse genomic data using information fusion. Bioinformatics, 21: 423--429, 2005.

[4] L. J. Heyer, S. Kruglyak, and S. Yooseph. Exploring expression data: identification and analysis of coexpressed genes. Genome Research, 9(11): 1106 -- 1115, 1999.

[5] J. D. Hughes, P. W. Estep, S. Tavazoie, and G. M. Church. Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. J. Mol. Biol., 296: 1205--1214, 2000.

[6] T. L. Bailey, and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc. Int. Conf. Intell. Syst. Mol. Biol., 2: 28--36, 1994.

[7] S. Mahony, D. Hendrix, A. Golden, T. J. Smith, and D.

S. Rokhsar. Transcription factor binding site identification using the self-organizing map. Bioinformatics, 21: 1807--1814, 2005.

[8] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. Systematic determination of genetic network architecture. Nature Genetics. 22: 281--285, 1999.

[9] M. A. Beer and S. Tavazoie. Predicting gene expression from sequence. Cell, 117: 185--198, 2004.

[10] H. Hotelling. Relations between two sets of variates. Biometrika, 28: 312--377, 1936.

[11] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis; An overview with application to learning methods. Technical Report CSD-TR-03-02, Royal Holloway University of London, 2003.

[12] S. Akaho. A kernel method for canonical correlation analysis. International meeting of Psychometric Society (IMP2001), 2001.

[13] F. R. Bach and M. I. Jordan. Kernel independent component analysis, Technical Report UCB//CSD-10-1166, UC Berkeley, 2001.

[14] Y. Yamanishi, J.-P. Vert, A. Nakaya, and M. Kanehisa. Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. Bioinformatics, 19 Suppl. 1: i323--i330, 2003.

[15] Y. Pilpel, P. Sudarsanam, and G. M. Church. Identifying regulatory networks by combinatorial analysis of promoter elements. Nature genetics, 29: 153--159, 2001.

[16] P. R. Dohrmann, G. Butler, K. Tamai, S. Dorland, J. R. Greene, D. J. Thiele, and D. J. Stillman. Parallel pathways of gene regulation: homologous regulators SWI5 and ACE2 differentially control transcription of HO and chitinase. Genes and Development. 6(1): 93--104, 1992.

[17] P. R. Dohrmann, W. P. Voth, and D. J. Stillman. Role of negative regulation in promoter specificity of the homologous transcriptional activators Ace2p and Swi5p. Molecular Cell Biology. 16(4): 1746--1758, 1996.

[18] I. Simon, J. Barnett, N. Hannett, C. T. Harbison, N. J. Rinaldi, T. L. Volkert, J. J. Wyrick, J. Zeitlinger, D. K. Gifford, R. S. Jaakkola, and R. A. Young. Serial regulation of transcriptional regulators in the yeast cell cycle. Cell, 106, 697--708, 2001.

[19] C. Koch, T. Moll, M. Neuberg, H. Ahorn, and K. Nasmyth. A role for the transcription factors Mbp1 and Swi4 in progression from G1 to S phase. Science, 261(5128): 1551--1557, 1993.

[20] A. K. Vershon and A. D. Johnson. A short, disordered protein region mediates interactions between the homeodomain of the yeast alpha 2 protein and the MCM1 protein. Cell, 72(1): 105--112, 1993.

[21] H. Zhong, R. McCord, and A. K. Vershon. Identification of target sites of the alpha2-Mcm1 repressor complex in the yeast genome. Genome Reserch. 9(11): 1040--1047, 1999.

[22] D. Lydall, G. Ammerer, and K. A. Nasmyth. New role for MCM1 in yeast: cell cycle regulation of SW15 transcription. Genes & Development. 5(12B): 2405--2419, 1991.