BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 11 2010, pages 1416–1422
doi:10.1093/bioinformatics/btq144

Gene expression

Advance Access publication April 7, 2010

Multilevel support vector regression analysis to identify condition-specific regulatory networks

Li Chen[1], Jianhua Xuan[1,*], Rebecca B. Riggins[2], Yue Wang[1], Eric P. Hoffman[3] and Robert Clarke[2,4]

[1]Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA 22203, [2]Lombardi Comprehensive Cancer Center and Department of Oncology, Georgetown University, Washington, DC 20057, [3]Research Center for Genetic Medicine, Children's National Medical Center, Washington, DC 20010 and [4]Department of Physiology and Biophysics, Georgetown University, Washington, DC 20057, USA

Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: The identification of gene regulatory modules is an important yet challenging problem in computational biology. While many computational methods have been proposed to identify regulatory modules, their initial success is largely compromised by a high rate of false positives, especially when applied to human cancer studies. New strategies are needed for reliable regulatory module identification.

Results: We present a new approach, namely multilevel support vector regression (ml-SVR), to systematically identify condition-specific regulatory modules. The approach is built upon a multilevel analysis strategy designed for suppressing false positive predictions. With this strategy, a regulatory module becomes ever more significant as more relevant gene sets are formed at finer levels. At each level, a two-stage support vector regression (SVR) method is utilized to help reduce false positive predictions by integrating binding motif information and gene expression data; a significant analysis procedure is followed to assess the significance of each regulatory module. To evaluate the effectiveness of the proposed strategy, we first compared the ml-SVR approach with other existing methods on simulation data and yeast cell cycle data. The resulting performance shows that the ml-SVR approach outperforms other methods in the identification of both regulators and their target genes. We then applied our method to breast cancer cell line data to identify condition-specific regulatory modules associated with estrogen treatment. Experimental results show that our method can identify biologically meaningful regulatory modules related to estrogen signaling and action in breast cancer.

Availability and implementation: The ml-SVR MATLAB package can be downloaded at http://www.cbil.ece.vt.edu/software.htm

Contact: xuan@vt.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 6, 2009; revised on March 11, 2010; accepted on April 2, 2010

1 INTRODUCTION

Identifying regulatory modules is one of the key steps to understanding the molecular mechanisms of biological processes, especially important for defining the deregulated pathways in cancer. At the transcriptional level, a regulatory module is defined as a set of genes controlled by one or several transcription factors (TFs) in a condition-specific manner (Segal *et al.*, 2003). TFs can either activate or inhibit gene expression, usually by binding to short, highly conserved, DNA sequences in the promoter (or upstream) region, i.e. transcription factor binding site (TFBS) or binding motif. In higher eukaryotes, TFBSs are often organized in clusters called *cis*-regulatory modules (CRMs). Many computational methods have been developed to facilitate the identification of CRMs from either gene expression data or DNA sequence data. Expression-based methods (Ihmels *et al.*, 2004; Segal *et al.*, 2003; Wang *et al.*, 2005) take advantage of gene expression data but lack of sequence binding constraints. Sequence-based module discovery algorithms, such as CisModule (Zhou and Wong, 2004), CREME (Sharan *et al.*, 2003) and ModuleSearch (Aerts *et al.*, 2003), analyze the promoter regions of a set of coregulated genes to identify overrepresented motif combinations. A major limitation of sequence-based methods is that they do not consider the condition-specific nature of regulatory modules, i.e. they ignore the relationship between binding affinities and gene expression levels.

A living cell is a dynamic system in which gene activities and interactions exhibit temporal patterns and spatial compartmentalization (Qi and Ge, 2006). Recently, several studies have shown that binding of TFs not only depends on their affinity for the binding sites but binding also occurs in a condition-specific manner in response to various environmental changes (Lee *et al.*, 2002; Segal *et al.*, 2008). Thus, a TF may play different regulatory roles to its downstream target genes or may even have different downstream targets under different conditions (Lee *et al.*, 2002). Motivated by this understanding, many computational algorithms were proposed to discover condition-specific regulatory modules by integrating condition-specific gene expression profiles and motif information. Regression models are widely used to combine these two types of information (Das *et al.*, 2006; Gao *et al.*, 2004; Nguyen and D'Haeseleer, 2006; Ruan and Zhang, 2006; Yu and Li, 2005). For example, a least square regression (LS regression) method described by (Nguyen and D'Haeseleer, 2006) identifies significant regulators by combining mRNA expression level and ChIP-on-chip binding data to minimize a fitting error. GRAM (Bar-Joseph *et al.*, 2003) is another regression method based on an iterative search to identify significant regulators and

[*]To whom correspondence should be addressed.

target genes. Bayesian models have also been used for regulatory module identification. A thermodynamic model (Segal *et al.*, 2008) was proposed to predict expression patterns from regulatory sequence data in *Drosophila* segmentation. COGRIM (Chen *et al.*, 2007) is a Bayesian hierarchical model with Gibbs Sampling implementation that integrates gene expression data, ChIP binding data and TF motif information to identify regulatory modules.

While these methods have achieved some degree of success, a high false positive prediction rate is still a major problem mainly due to the noises in motif information and gene expression data. To reduce the false positive rate (FPR), we propose a novel method, namely multilevel regulatory module identification through support vector regression (ml-SVR), to help find significant and stable regulatory modules. The ml-SVR method is particularly effective because of several novel adaptations: (i) a two-stage support vector regression (SVR) method is used to integrate binding motif information and gene expression data, aiming to improve the noise-tolerance capability; (ii) a significance analysis procedure is applied to identify statistically significant regulatory modules; (iii) a multilevel analysis strategy is developed to reduce the FPR for reliable regulatory module identification; and (iv) a weighted voting scheme is implemented for target gene identification, taking into account the entire multilevel analysis.

We have applied the ml-SVR method to simulation data and yeast cell cycle data to assess its performance for gene module identification, in comparison with existing methods. The comparison results clearly demonstrate that the proposed ml-SVR method notably outperforms other methods. We then applied our method to two breast cancer microarray datasets to identify condition-specific regulatory modules, respectively, in response to different estrogen conditions. The experimental results show that our method can successfully identify biologically meaningful modules associated with estrogen signaling and action in breast cancer.

2 METHODS

The ml-SVR method is aimed to identify significant condition-specific regulatory modules by integrating mRNA gene expression data and binding motif information. Figure 1 illustrates the flow chart of the ml-SVR approach, shown as an iterative procedure in a nutshell. This multilevel analysis procedure, as conducted in a coarse-to-fine way, ensures that a condition-specific regulatory module becomes ever more significant as more relevant gene sets are formed at finer levels. At each level, SVR is used to integrate binding motif information and gene expression data. Specifically, a two-stage SVR method is implemented to refine the estimation of transcription factor activity (TFA) and binding strength. Significance analysis of regulatory modules is achieved by evaluating the regression fitting errors compared to a baseline without motif information; an *F*-statistic is calculated from a permutation test to assess the significance (*P*-value) of a regulatory module. Finally, with the multilevel analysis, significant gene modules can be determined and their target genes identified by a voting scheme running through all levels. In the following subsections, we provide a detailed description of each component in the ml-SVR approach.

2.1 Sequence analysis for motif information

ChIP-on-chip, also known as genome-wide location analysis, is a technique that can isolate and identify DNA sequences occupied by specific DNA binding proteins (Aparicio *et al.*, 2004). However, it is not a trivial task to measure the binding strengths for all TFs from ChIP-on-chip experiments due to the limited antibodies available, especially for higher eukaryote studies. An alternative and practical way is to extract binding motif information from the promoter regions of focused genes. We assume that the binding strength for a specific TF to its target gene is proportional to the similarity score of its binding site and the number of occurrences of the binding site in the gene promoter region. We generated a gene-motif binding strength matrix $\mathbf{X} = [x_{gm}]$ using the cut offs that minimize the FPR. The rows in the matrix $\mathbf{X}$ correspond to different genes, and the columns correspond to different binding sites (or motifs). Each element x_{gm} represents the binding strength at motif m in the promoter region of a gene g, which is calculated mathematically as follows:

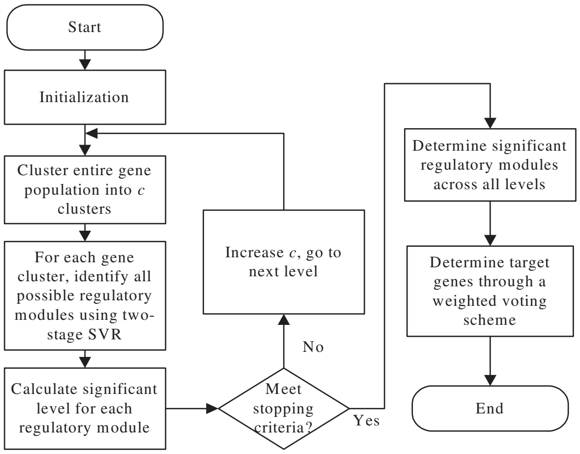
Start
Initialization
Cluster entire gene population into c clusters
For each gene cluster, identify all possible regulatory modules using two-stage SVR
Calculate significant level for each regulatory module
Meet stopping criteria?
No
Yes
Increase c, go to next level
Determine significant regulatory modules across all levels
Determine target genes through a weighted voting scheme
End

Fig. 1. Flow chart of the ml-SVR approach.

$$x_{gm} = \sum_{i=1}^{N} \frac{1}{2}(\mathrm{mss}_{gmi} + \mathrm{css}_{gmi}), \tag{1}$$

where N is the number of occurrences of motif m in the promoter region of gene g; mss_{gmi} and css_{gmi} are the matrix similarity score and core similarity score for motif m and gene g in the i-th hit, respectively (for more details, please refer to Section S1 in the Supplementary Material).

2.2 Two-stage SVR to infer regulatory modules

Suppose that there are G genes and T gene expression profiles. We represent microarray gene expression data as a matrix $\mathbf{Y}_{G\times T} = [y_{gt}]$, $g = 1,\ldots,G$; $t = 1,\ldots,T$, where each element y_{gt} is the log ratio of the expression level of gene g in sample t to that of the control sample. We also assume that there are M motifs on this gene set and the corresponding gene motif binding matrix is $\mathbf{X}_{G\times M} = [x_{gm}]$, $g = 1,\ldots,G$; $m = 1,\ldots,M$, where x_{gm} is the binding strength on motif m in the promoter region of gene g. The relationship between gene expression level and binding strength can be mathematically described by a linear model as follows:

$$\mathbf{Y}_{G\times T} = \mathbf{X}_{G\times M}\mathbf{A}_{M\times T} + \mathbf{N}, \tag{2}$$

where $\mathbf{A}_{M\times T} = [a_{mt}]$, $m = 1,\ldots,M; t = 1,\ldots,T$ is the TF activity matrix and $\mathbf{N}$ the noise matrix. Biologically, the model represents the log ratio of gene expression levels expressed as a linear combination of log ratios of TFAs (denoted as a_{mt}) weighted by their binding strengths (i.e. x_{gm}) (Liao *et al.*, 2003).

If $\mathbf{X}$ and $\mathbf{Y}$ are known, the solution to the linear model [Equation (2)] can then be easily obtained by a simple regression (Bussemaker *et al.*, 2001). However, since both motif information and gene expression data are noisy, a simple regression will inevitably introduce a large number of false positive predictions. To alleviate this problem, we propose a two-stage SVR method to specifically address the noises in motif information and gene expression data. SVR has been shown to have good robust properties against noise through the regularization term in its cost function (Smola and Scholkopf, 1998); the regularization term is intended to keep the estimated TF activity (in matrix $\mathbf{A}$)

as smooth as possible so as to combat the noise in gene expression data (**Y**). The ε-insensitive loss function is used in SVR to ensure the existence of the global minimum and a high tolerance to noise, which is defined by

$$L_\varepsilon(y_{gt}) = \begin{cases} 0, & \text{if } |y_{gt} - \hat{y}_{gt}| < \varepsilon \\ |y_{gt} - \hat{y}_{gt}| - \varepsilon, & \text{otherwise} \end{cases}, \tag{3}$$

where $\hat{y}_{gt}$ is the estimated value of expression log ratio y_{gt}. To combat the noise in motif information, we use a similar strategy as in the two-stage approach proposed by Yu *et al.* (2005) to update the binding strength matrix **X** based on **Y** and the estimated **A.** In this way, we can reduce the number of false binding motifs, which are initially present in the binding strength matrix **X** but with no support from gene expression data (**Y**) and estimated TF activity (**A**).

The two-stage SVR method is implemented as an iterative procedure, which updates matrices **A** and **X** alternately until converged. In the implementation, we normalize (or standardize) the gene expression data to 0 mean and 1 standard deviation. We also standardize the estimated TF activity at each iteration step of our algorithm. The final algorithm of our two-stage SVR approach can be summarized as follows:

(1) Estimate **A** using **X** and **Y**. For each column vector $\mathbf{y}_t$ in matrix **Y**, regress $\mathbf{y}_t$ against **X** based on $y_{gt} = f(\mathbf{x}_g) = \sum_{m=1}^{M} x_{gm} a_{mt}$; calculate regression coefficient a_{mt} using ε-insensitive SVR.

(2) Update **X** using **A** and **Y**. For each row vector $\mathbf{y}_g$ in matrix **Y**, regress $\mathbf{y}_g$ against **A** based on $y_{gt} = f(\mathbf{a}_t) = \sum_{m=1}^{M} x'_{gm} a_{mt}$; calculate regression coefficient x'_{gm} using ε-insensitive SVR; update **X** by $\mathbf{X} = \mathbf{X} + \eta(\mathbf{X}' - \mathbf{X})$, where η is a parameter in the range of (0, 1). (Note that η is set to 0.2 in our experiments.)

(3) Repeat Step (1) and Step (2) until convergence. The convergence criterion is defined as the average correlation coefficient of TF activities between two successive iterations is larger than a predefined threshold r_0. (Note that r_0 is set as 0.9 in our experiments.)

2.3 Significance analysis of regulatory modules

A significance analysis procedure is designed to test if a selected motif set is statistically associated with the regulation of a given gene set, aiming to identify active regulators for that set. The null and alternative hypotheses (H_0 and H_1, respectively) are given as follows:

H_0: the motif set is not actively involved in regulating a given gene set;

H_1: the motif set is actively involved in regulating a given gene set.

We use a summary statistic to represent the fitting results as described below:

$$F = \frac{\mathrm{RSS}_0 - \mathrm{RSS}_1}{\mathrm{RSS}_1}; \mathrm{RSS}_0 = \sum_{g,t} (y_{gt} - \bar{y}_t)^2,\ \bar{y}_t = \frac{1}{|G|} \sum_{g \in G} y_{gt}$$
$$\mathrm{RSS}_1 = \sum_{g,t} (y_{gt} - \hat{y}_{gt})^2,\ \hat{y}_{gt} = \sum_m a_{mt} x_{gm} + b_t \quad , \tag{4}$$

where RSS_0 is the residual sum of squares without motif information, and RSS_1 is the residual sum of squares with motif information. The above equation is proportional to the typical F-statistic used to compare two models (Lomax, 2007). To calculate the P-value, we use the permutation method described below to form the null distribution. For a given motif set, we randomly select a gene set G_0 with the same size of G from the entire gene population, and then repeat B times to generate the corresponding null statistic score F^{0b}, for $b = 1, 2, \ldots, B$ ($B = 1000$ in our experiments). The P-value can be obtained for each gene set by calculating the probability that a null gene set has a statistic more extreme than the observed statistic. Mathematically, the P-value is calculated by the following equation:

$$p = \mathrm{Pr}_{H_0}(F^{0b} > F) = \frac{\#\{b : F^{0b} > F, b = 1, \ldots, B\}}{B}. \tag{5}$$

2.4 Multilevel analysis for regulatory module identification

Assuming that most genes involved in a regulatory module are coexpressed under a given condition, we can use a clustering method to form the gene set for regression analysis. However, simple gene clustering based on gene expression data alone often results in many false positives for gene module identification. In addition, motif information is noisy and incomplete due to the current status of limited biological knowledge. Thus, false positives would be included based on a fixed gene set and available motif information. To reduce the false positives, we developed a multilevel analysis strategy to search for regulatory modules showing significance consistently from coarse level to fine levels. With this strategy, a condition-specific regulatory module and its enriched motifs will appear increasingly significant in finer levels, as the irrelevant genes are gradually eliminated (see Supplementary Fig. S1 in the Supplementary Material for an illustration of the multilevel strategy). Technically, a multilevel gene clustering procedure, such as self-organizing map clustering (Kohonen, 1997), is used to form the gene clusters to gradually reduce the irrelevant genes for multilevel analysis. The multilevel analysis strategy, incorporating the two-stage SVR approach described previously, is the backbone of the ml-SVR approach proposed in this article for reliable regulatory module identification. The final ml-SVR procedure is illustrated in Figure 1, which can also be summarized as follows:

(1) Set cluster number $c = 1$ and cluster level $l = 1$. For all possible enriched motif sets, calculate their P-values on current gene set G through the two-stage SVR analysis and significance analysis described in Subsections 2.2 and 2.3.

(2) Increment c by 1 and l by 1. Cluster the gene population into c clusters, denoted as $\{G_1^l, G_2^l, \ldots, G_c^l\}$.

(3) For each gene cluster, calculate P-values for all possible enriched motifs by the two-stage SVR analysis and significance analysis (Subsections 2.2 and 2.3).

(4) Repeat Steps (2) and (3) until the stopping criterion is met, that is, the number of genes is less than a threshold t_0 for all gene clusters.

(5) Use p_M^{lc} to denote the P-value of a candidate motif set $M$for cluster G_c^lat level l. Output the significantly enriched motif sets if they satisfy $\min_c (p_M^{lc}) < p_0^l, \forall l$, where p_0^l is the threshold of P-value at level l. The total number of levels is L. Assign the final weighted average P-value as $p_M = \sum_l {}^{l}/_{\Delta} \min(p_M^{lc}), \quad \Delta = 1 + 2 + \cdots + L$.

(6) Use a voting scheme to determine the gene members of a regulatory module with the enriched motif set M: first initialize a gene weight vector **w** as 0 and then update **w** by the following equation:

$$\forall l, c, \quad \mathbf{w}_{G_c^l} = \mathbf{w}_{G_c^l} + \sum_{m \in M} \mathbf{X}_{G_c^l m'} \quad \mathit{if}\ p_M^{lc} < p_0^l.$$

(7) Finally, the genes whose weights are greater than a threshold w_0 are chosen as the members of a corresponding regulatory module. In our implementation, we set w_0 as the mean of **w** plus one standard deviation, which gives us a reasonable number of target genes for further study (see the Supplementary Material, Section S7, for a discussion on the choice of threshold w_0).

3 RESULTS

3.1 Simulation data

We first tested our method on a synthetic yeast microarray dataset. The microarray dataset was simulated using the network generator software SynTReN (Van den Bulcke *et al.*, 2006), where network topologies are generated from yeast regulatory networks using a neighbor addition strategy. The network consists of 29 TFs and 260 target genes. The mRNA expression profiles were generated

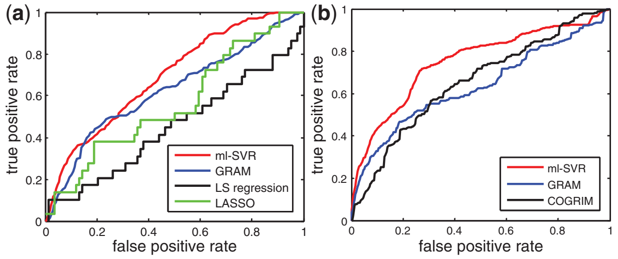

Fig. 2. Comparison of ROC curves for ml-SVR and other methods on simulation data. (**a**) Transcription factor identification. (**b**) Target gene identification.

for 260 genes at 50 different conditions based on the network. In our algorithm, we used a ChIP-on-chip data (Lee *et al.*, 2002) as our binding information data that includes 113 regulators and their binding *P*-values to all genes. The purpose of this study is to first identify true regulators and then their downstream target genes. To fulfill this purpose, we applied the ml-SVR approach to the simulation data and identified significant regulatory networks associated with TFs. The detailed experimental procedure as well as the parameter settings of ml-SVR can be found in the Supplementary Material (Section S2).

To evaluate our proposed ml-SVR approach, we compared its performance with similar methods including LS regression (Nguyen and D'Haeseleer, 2006), LASSO (Tibshirani, 1996), GRAM (Bar-Joseph *et al.*, 2003) and COGRIM (Chen *et al.*, 2007). Among these three existing methods, only GRAM can simultaneously identify significant regulators and target genes. LS regression and LASSO can only identify significant regulators with known target genes by assuming the binding information is known from ChIP-on-chip data. COGRIM is derived from a Bayesian hierarchical model, which assumes the TFs and their activities are known so as to infer new target genes based on binding information. As a common practice (Segal *et al.*, 2003) but faulty (Liao *et al.*, 2003), mRNA expression level of each TF is often used to approximate the TF activity for COGRIM. Therefore, in this study we compared ml-SVR with GRAM and LS regression for TF identification, while we compared ml-SVR with GRAM and COGRIM for target gene identification.

Figure 2a shows the receiver operator characteristic (ROC) curves of TF identification for ml-SVR, GRAM, LS regression and LASSO, respectively. From the figure, we can see that ml-SVR outperforms GRAM and LS regression methods in identifying significant TFs. The mean area under the ROC curve (AUC) value of ml-SVR is 0.6912 (with a standard deviation of 0.0196), which is greater than the AUC values of GRAM (0.6245), LS regression (0.5530) and LASSO (0.5620). It should be noted that in this comparison experiment, the overall performances of all three methods are relatively low; this is indeed a relatively difficult case since some non-linear relationships between TFs and target genes were included by SynTReN in the simulation data. Nevertheless, the FPR is much reduced by ml-SVR as compared to GRAM and LS regression. When the true positive rate (TPR) is fixed at 80%, the FPR for ml-SVR is 55.48% while 74.64% for GRAM, 94.05% for LS regression and 71.42% for LASSO, showing a substantial improvement in FPR reduction.

For the 29 known TFs, we compared the performance of target gene identification for ml-SVR, GRAM and COGRIM. Figure 2b shows the average of ROC curves of target gene identification for all TFs using ml-SVR, COGRIM and GRAM, respectively. The ml-SVR approach gave us the best performance with a mean AUC value of 0.7358 (and a standard deviation of 0.0090). The performances of COGRIM and GRAM are similar with the AUC values of 0.6434 and 0.6438, respectively, which are much lower than that of ml-SVR. Also seen from Figure 2b, the FPR for ml-SVR is 42.12% given TPR = 80%, which shows a reduction of ∼25% when compared to 68.79% for GRAM and 66.04% for COGRIM. This comparison result demonstrates the advantage of ml-SVR over other methods for identifying significant TFs and their target genes. For more ROC analysis results, please refer to Figure S3 in the Supplementary Material to see the detailed performance of target gene identification for several individual TFs.

3.2 Yeast cell cycle data

We also applied the ml-SVR method to a yeast cell cycle microarray dataset (Spellman *et al.*, 1998). This microarray dataset includes 77 samples collected with three different synchronization experimental conditions. For the binding information, we used the ChIP-on-chip data from Lee *et al.* (2002), which provides significance levels (*P*-values) of 113 TFs binding to their target genes. Among the 113 TFs, 19 regulators have been identified as cell cycle-related TFs. We preprocessed the dataset and finally obtained 6099 open reading frames (ORFs) that have both expression measurements and binding information (see Section S3 in the Supplementary Material for more details). The goal of this study is to identify the cell cycle-related condition-specific TFs and their target genes.

To demonstrate the feasibility of applying ml-SVR to real microarray data, we compared the performances of ml-SVR, GRAM and LS regression for TF identification using 19 known cell cycle-related regulators as the ground truth. The parameters in our algorithm are same as those in the simulation study. Figure S4 in the Supplementary Material shows the ROC curves of TF identification by ml-SVR, GRAM, LS regression and LASSO. The mean AUC value for ml-SVR is 0.9284 (with a standard deviation of 0.0127). The AUC values for GRAM, LS regression and LASSO methods are 0.6691, 0.8761 and 0.7704, respectively. The improvement of ml-SVR over GRAM is substantial in terms of FPR reduction. Again, when the TPR is fixed at 80%, the FPR for ml-SVR is 11.31% while it is 74.06% for GRAM, 18.68% for LS regression and 52.18% for LASSO, showing a substantial improvement in FPR reduction. These results clearly show that ml-SVR outperforms the GRAM and LS regression methods for the identification of cell cycle-related TFs.

For target gene identification of all cell cycle-related TFs, since the ground truth target genes are not known for all TFs, we assessed their Gene Ontology (GO) functional enrichment as an alternative using software BiNGO (Maere *et al.*, 2005). The GO function enrichment score is defined as the negative logarithm of Benjamin-corrected *P*-value from an overrepresentative analysis in BiNGO. The average GO functional enrichment scores are 3.53 for ml-SVR, 3.41 for COGRIM and 2.86 for GRAM, which indicates that our method can identify more functionally coherent gene clusters associated with specific TFs.

3.3 Breast cancer data

3.3.1 Estrogen-induced condition A breast cancer cell line microarray dataset (Creighton *et al.*, 2006) was used to identify

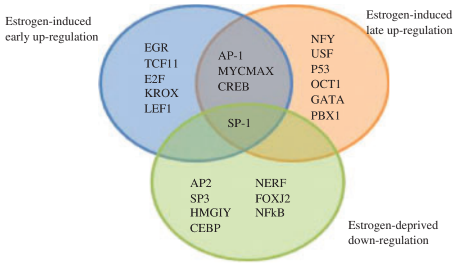

Fig. 3. Venn diagram of significantly enriched motifs in estrogen-induced and estrogen-deprived conditions.

condition-specific regulatory modules associated with estrogen signaling in breast cancer. Estrogen plays a significant role in breast cancer development and progression. The original profiling study was designed to examine how estrogen-induced gene expression patterns observed *in vitro* correlate with the expression patterns in breast tumors *in vivo*. Three estrogen-dependent breast cancer cell lines (MCF-7, T47D and BT-474) were treated with 17β-estradiol (E2) from 0 to 24 h, and then profiled for gene expression using Affymetrix GeneChip Arrays. As reported in the paper (Creighton *et al.*, 2006), eight E2-induced gene clusters were formed and among them, the expression pattern in four clusters [i.e. Cluster A, B, C and D as denoted in Creighton *et al.* (2006)] clearly showed upregulation along the time from early to late, which provides us an important starting point to study regulatory mechanisms related to estrogen signaling and action in breast cancer. The ml-SVR approach was applied to identify significant regulatory networks (see Section S4 in the Supplementary Material for the detailed experimental procedure).

The identified significant motifs in each cluster are shown in Table S1 (see the Supplementary Material), along with their average *P*-values across all levels, number of probe sets in the module and the description of the corresponding TFs. The significant motifs are defined by average *P*-values $\leq$0.05. Figure S5 in the Supplementary Material shows an example of using the multilevel strategy to determine that SP1 and AP1 are significant TFs while ATF3 and E2F are not significant. Among all listed motifs and their corresponding TFs, we found that several TFs are tightly related to estrogen signaling as reported in previous studies (Bjornstrom and Sjoberg, 2005), to name just a few here, AP-1, SP-1 and CREB. From the table, we can also see that the significantly enriched motifs are different in each cluster, reflecting the condition-specific nature of transcriptional regulation. Since the target genes in Clusters A and B are upregulated within 4 h, we assigned the significantly enriched motifs in these two clusters to the early upregulation condition. The target genes in Clusters C and D showed sustained induction by E2 at 8 h and 12–24 h, respectively. We assigned the significantly enriched motifs in Clusters C and D to the late upregulation condition.

Figure 3 shows the significantly enriched motifs in two different conditions, i.e. early and late conditions. We can see that AP-1, SP-1, MYCMAX and CREB are significantly enriched in both early and late conditions, suggesting their important roles in estrogen signaling and action. AP-1 and SP-1 are known to form TF complexes with estrogen receptor (ER) to regulate genes with the appropriate binding site(s); the TF CREB is phosphorylated after the MAPK signaling pathway has been activated by 17β-estradiol and the phophorylation of CREB leads to the expression of genes that contain CRE binding motifs (Bjornstrom and Sjoberg, 2005). EGR, TCF11, E2F, KROX and LEF1 are only significantly enriched in the early upregulation condition. Since many of their transcriptional functions are not known, we annotated their target genes biological function through GO analysis; their significant GO terms are related to 'ribosome biogenesis', 'RNA metabolism' and 'protein folding' (*P*-value <0.01). This may suggest some potential functions of these binding TFs. For example, a change in the ability to fold proteins adequately induces the unfolded protein response, which we have previously implicated in antiestrogen resistance (Gomez *et al.*, 2007; Gu *et al.*, 2002). Similarly, NFY, USF, P53, OCT1, GATA and PBX1 are only significantly enriched in the late upregulation condition. Significant GO terms of their target genes include 'cell cycle', 'cell proliferation', 'mitosis' and 'DNA replication' (*P*-value <0.01). Among them, previous studies (Imbriano *et al.*, 2005) have shown that nuclear transcription factor Y (NFY) and p53 are related to cell cycle arrest; Octamer transcription factor-1 (Oct-1) is a member of the POU family of TFs and is involved in the transcriptional regulation of a variety of gene expression related to cell cycle regulation, development and hormonal signals (Kakizawa *et al.*, 2001); Upstream stimulatory factor 1 (USF) is a transcription coactivator that plays a role in regulation of cell proliferation and associated with breast neoplasms (Xing and Archer, 1998); Pre-B-cell leukemia homeobox 1 (PBX1) is a transcription activator that promotes TF activity and cell growth, which may play an important role in Wnt receptor signaling (Hayward *et al.*, 2005).

3.3.2 Estrogen-deprived condition We previously derived a series of breast cancer variants that closely reflect clinical phenotypes of endocrine sensitive and resistant tumors (Brunner *et al.*, 1997; Clarke *et al.*, 1989). We selected two cell lines for this study: MCF-7 and MCF-7 stripped. MCF-7 stripped denotes estrogen-deprived MCF-7 human breast cancer cells, which were grown in the absence of estrogen for 96 h. Three independent total RNA samples were extracted for each cell line (MCF-7 and MCF-7 stripped) and the samples were arrayed using Affymetrix GeneChip HG-U133A. Raw data are available in GEO (http://www.ncbi.nlm.nih.gov/geo/; accession number: GSE 20700). We analyzed the enriched motifs and their targets for the genes significantly downregulated in MCF-7-stripped cells as compared to MCF-7 cells. Downregulated genes are identified by SAM analysis (Tusher *et al.*, 2001) with FDR <0.05. Again, we applied the ml-SVR approach for this study to identify significant regulatory networks (for more details about the procedure, please see Section S5 in the Supplementary Material).

Supplementary Table S2 shows the identified significant motifs, their average *P*-values across all levels, number of probe sets in the module and the description of the corresponding TFs. As in the previous subsection, significant motifs were selected when their average *P*-values $\leq$0.05. From these motifs and their corresponding TFs, we found several TFs that have known associations with breast cancer, such as SP-1 and NFκB (Bjornstrom and Sjoberg, 2005). For the their target genes, the significant GO terms functions are related to cell cycle, intracellular membrane-bound, DNA replication, etc. (*P*-value <0.01).

In Figure 3, we show a Venn diagram of significantly enriched motifs in both estrogen-induced and estrogen-deprived conditions. We can see that SP-1 is significantly enriched in both conditions, while AP-1 is only enriched in the estrogen-induced condition and

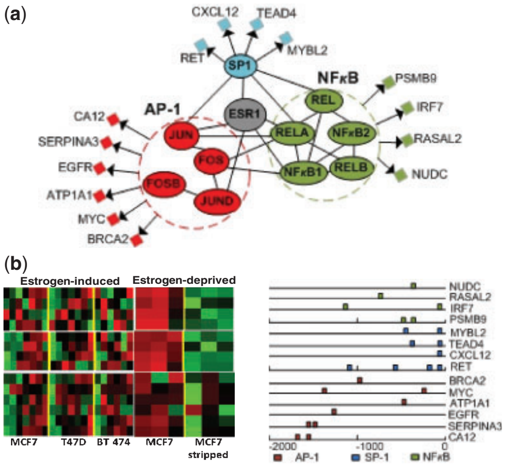

Fig. 4. (**a**) Identified transcription regulatory modules of AP-1, SP-1 and NFκB in breast cancer study. (**b**) Gene expression patterns of target genes of AP-1, SP-1 and NFκB in estrogen-induced and estrogen-deprived conditions (left), and the binding sites of AP-1, SP-1 and NFκB on the promoter regions of their target genes (right).

NFκB is only enriched in the estrogen-deprived condition. A number of publications have reported that elevated AP-1 and NFκB activities are each associated with tamoxifen-resistant breast cancer (Pratt *et al.*, 2003; Riggins *et al.*, 2005; 2008; Zhou *et al.*, 2007). We depicted these three transcription regulatory modules with their target genes in Figure 4a. The interactions among the TFs ESR1, AP-1, SP-1and NFκB are extracted from Human Protein Reference Database (Mishra *et al.*, 2006). Figure 4b shows the binding sites for AP-1, SP-1 and NFκB in the promoter regions of their target genes, and their expression patterns in both conditions. In the next section, we provide a detailed description of the SP-1 network to establish its function role in estrogen signaling and action. The detailed description of AP-1 and NFκB can be found in the Supplementary Material (Fig. S6).

SP-1 motifs are significantly enriched under both estrogen-induced and estrogen-deprived conditions, but the role of this TF in estrogen and antiestrogen signaling is less clear. Kim *et al.* (2003) have reported that in breast cancer cells, E2 and antiestrogens can both stimulate transcription on G/C-rich promoters via ER/SP-1 complexes. Table S3 (see the Supplementary Material) shows the SP-1 target genes common to the estrogen-induced and estrogen-deprived conditions. Among these genes, some of them have been confirmed to be regulated by SP-1 in previous studies, and may have direct relevance to breast cancer, estrogen signaling and antiestrogen resistance. For instance, it has been shown that the TF SP1 can bind to the promoter of CXCL12 (Luker and Luker, 2006), and that estrogen-stimulated proliferation of ER+ T47D breast cancer cells can be blocked by a specific antagonist of the receptor for CXCL12 (Pattarozzi *et al.*, 2008). MYBL2 (B-MYB) is a ubiquitous protein required for mammalian cell growth, and a study by Sala *et al.* (1999) showed that B-MYB functions as a coactivator of SP1, binding to the 120 bp B-MYB promoter fragment. Moreover, it has recently been shown that MYBL2 mRNA expression is significantly increased in breast cancer cells resistant to the tamoxifen analogue Toremifene (Pennanen *et al.*, 2009). Finally the RET proto-oncogene, more commonly associated with multiple endocrine neoplasia and medullary thyroid carcinoma, is also known to be transcriptionally regulated by SP1 (Andrew *et al.*, 2000). Boulay *et al.* have reported that RET is induced by estrogens; RET signaling enhances the proliferative effect(s) of estrogen in ER+ MCF7 and T47D breast cancer cells, and RET is coexpressed with ER in primary breast tumors (Boulay *et al.*, 2008). We have also observed RET mRNA overexpression in tamoxifen-resistant SUM44 breast cancer cells (Riggins *et al.*, 2008). These results demonstrate that our method can successfully identify relevant TF targets that play key, functional roles in estrogen signaling and action in breast cancer.

4 DISCUSSION

Identification of transcription regulatory modules has become increasingly important to understand the molecular mechanisms associated with cancer. Previous methods (Das *et al.*, 2006; Gao *et al.*, 2004; Nguyen and D'Haeseleer, 2006; Ruan and Zhang, 2006; Yu and Li, 2005) focused on how to model the relationship of TF binding and gene expression levels, assuming either active TFs or target genes are known. However, it is a challenging problem in many cancer studies due to significant noise in data sources: inaccurate motif binding information, noisy gene expression data and incomplete knowledge of the biological problem under study. The ml-SVR method is intended to address these problems and simultaneously identify significant TFs and their target genes through a multilevel strategy. SVR is utilized because its performance for combining binding motif information and gene expression data is robust in the presence of noise; note that it can also be extended to model the non-linear relationship between binding information and expression data through kernel functions. Clustering is used to group genes in multiple levels, in a coarse-to-fine way, to avoid hard split of the genes, which may be undesirable considering the noises.

There are several issues for further investigation. The method described here assumes that coexpressed genes should be coregulated to some degree; hence, genes are clustered based on their expression profiles alone. Recently, Gong *et al.* (2008) proposed to cluster genes based on their gene expression data and binding motif information together, which may provide more accurate gene clusters for analysis. Another important issue that needs to be addressed is how to determine an appropriate motif set for SVR fitting. In our experiment, we only focused on each individual TF and their modules. However, finding the cooperative TFs is also important for many biological studies. Due to the large number of motifs under study (typically in a range of 50 to 500), it is not feasible to consider all possible motif combinations when the order of the motif set increases. In our recent work (Chen *et al.*, 2008), we developed a stepwise forward greedy search strategy, using a modified loss function to find the cooperative motifs in a given gene set. Finally, the parameters in the algorithm need to be further optimized in the future work.

5 CONCLUSION

We have proposed a multilevel two-step SVR method to identify significant condition-specific regulatory networks. Binding motif information and gene expression data are integrated by SVR followed by significance analysis to find the active motif sets.

A multilevel analysis strategy is further developed to help reduce false positives for reliable regulatory module identification. The simulation study and the experiment on yeast cell cycle data demonstrated the effectiveness of our method in identifying TF and target genes. Furthermore, we studied two breast cancer cell line datasets and the results showed that our method can successfully identify condition-specific regulatory modules associated with estrogen signaling in breast cancer.

ACKNOWLEDGEMENTS

We would also like to thank Alan Zwart for his work in the acquisition of breast cancer cell line microarray data.

Funding: National Institutes of Health Grants (NS29525-13A, EB000830, CA109872, CA096483 and CA139246); DoD/CDMRP grant (BC030280).

Conflict of Interest: none declared.

REFERENCES

Aerts,S. *et al.* (2003) Computational detection of cis-regulatory modules. *Bioinformatics*, **19** (Suppl. 2), ii5–ii14.

Andrew,S.D. *et al.* (2000) Sp1 and Sp3 transactivate the RET proto-oncogene promoter. *Gene*, **256**, 283–291.

Aparicio,O. *et al.* (2004) Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. *Curr. Protoc. Cell Biol.*, **Chapter 17**, Unit 17 17.

Bar-Joseph,Z. *et al.* (2003) Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol.*, **21**, 1337–1342.

Bjornstrom,L. and Sjoberg,M. (2005) Mechanisms of estrogen receptor signaling: convergence of genomic and nongenomic actions on target genes. *Mol. Endocrinol.*, **19**, 833–842.

Boulay,A. *et al.* (2008) The Ret receptor tyrosine kinase pathway functionally interacts with the ERalpha pathway in breast cancer. *Cancer Res.*, **68**, 3743–3751.

Brunner,N. *et al.* (1997) MCF7/LCC9: an antiestrogen-resistant MCF-7 variant in which acquired resistance to the steroidal antiestrogen ICI 182,780 confers an early cross-resistance to the nonsteroidal antiestrogen tamoxifen. *Cancer Res.*, **57**, 3486–3493.

Bussemaker,H.J. *et al.* (2001) Regulatory element detection using correlation with expression. *Nat. Genet.*, **27**, 167–171.

Chen,G. *et al.* (2007) Clustering of genes into regulons using integrated modeling-COGRIM. *Genome Biol.*, **8**, R4.

Chen,L. *et al.* (2008) Identification of condition-specific regulatory modules by multi-level motif and mRNA expression analysis. *The 2008 International Conference on Bioinformatics and Computational Biology*. Las Vegas, Nevada.

Clarke,R. *et al.* (1989) Progression from hormone dependent to hormone independent growth in MCF-7 human breast cancer cells. *Proc. Natl Acad. Sci.*, **86**, 3649–3653.

Creighton,C.J. *et al.* (2006) Genes regulated by estrogen in breast tumor cells in vitro are similarly regulated in vivo in tumor xenografts and human breast tumors. *Genome Biol.*, **7**, R28.

Das,D. *et al.* (2006) Adaptively inferring human transcriptional subnetworks. *Mol. Syst. Biol.*, **2**, 2006 0029.

Gao,F. *et al.* (2004) Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics*, **5**, 31.

Gomez,B.P. *et al.* (2007) Human X-box binding protein-1 confers both estrogen independence and antiestrogen resistance in breast cancer cell lines. *FASEB J.*, **21**, 4013–4027.

Gong,T. *et al.* (2008) Exploring transcriptional modules by integrative gene clustering guided by transcription factor binding information. *The 2008 International Conference on Bioinformatics and Computational Biology*. Las Vegas, Nevada.

Gu,Z. *et al.* (2002) Association of interferon regulatory factor-1, nucleophosmin, nuclear factor-kappaB, and cyclic AMP response element binding with acquired resistance to Faslodex (ICI 182,780). *Cancer Res.*, **62**, 3428–3437.

Hayward,P. *et al.* (2005) Notch modulates Wnt signalling by associating with Armadillo/beta-catenin and regulating its transcriptional activity. *Development*, **132**, 1819–1830.

Ihmels,J. *et al.* (2004) Defining transcription modules using large-scale gene expression data. *Bioinformatics*, **20**, 1993–2003.

Imbriano,C. *et al.* (2005) Direct p53 transcriptional repression: in vivo analysis of CCAAT-containing G2/M promoters. *Mol. Cell Biol.*, **25**, 3737–3751.

Kakizawa,T. *et al.* (2001) Silencing mediator for retinoid and thyroid hormone receptors interacts with octamer transcription factor-1 and acts as a transcriptional repressor. *J. Biol. Chem.*, **276**, 9720–9725.

Kim,K. *et al.* (2003) Domains of estrogen receptor alpha (ERalpha) required for ERalpha/Sp1-mediated activation of GC-rich promoters by estrogens and antiestrogens in breast cancer cells. *Mol. Endocrinol.*, **17**, 804–817.

Kohonen,T. (1997) *Self-Organizing Maps*. Springer, NY.

Lee,T.I. *et al.* (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, **298**, 799–804.

Liao,J.C. *et al.* (2003) Network component analysis: reconstruction of regulatory signals in biological systems. *Proc. Natl Acad. Sci. USA*, **100**, 15522–15527.

Lomax,R.G. (2007) *Statistical Concepts: A Second Course*. Lawerence Erlbaum Associates, Mahwah, NJ .

Luker,K.E. and Luker,G.D. (2006) Functions of CXCL12 and CXCR4 in breast cancer. *Cancer Lett.*, **238**, 30–41.

Maere,S. *et al.* (2005) BiNGO: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, **21**, 3448–3449.

Mishra,G.R. *et al.* (2006) Human protein reference database–2006 update. *Nucleic Acids Res.*, **34**, D411–D414.

Nguyen,D.H. and D'Haeseleer,P. (2006) Deciphering principles of transcription regulation in eukaryotic genomes. *Mol. Syst. Biol.*, **2**, 2006 0012.

Pattarozzi,A. *et al.* (2008) 17beta-estradiol promotes breast cancer cell proliferation-inducing stromal cell-derived factor-1-mediated epidermal growth factor receptor transactivation: reversal by gefitinib pretreatment. *Mol. Pharmacol.*, **73**, 191–202.

Pennanen,P.T. *et al.* (2009) Gene expression changes during the development of estrogen-independent and antiestrogen-resistant growth in breast cancer cell culture models. *Anticancer Drugs*, **20**, 51–58.

Pratt,M.A. *et al.* (2003) Estrogen withdrawal-induced NF-kappaB activity and bcl-3 expression in breast cancer cells: roles in growth and hormone independence. *Mol. Cell Biol.*, **23**, 6887–6900.

Qi,Y. and Ge,H. (2006) Modularity and dynamics of cellular networks. *PLoS Comput. Biol.*, **2**, e174.

Riggins,R.B. *et al.* (2005) The nuclear factor kappa B inhibitor parthenolide restores ICI 182,780 (Faslodex; fulvestrant)-induced apoptosis in antiestrogen-resistant breast cancer cells. *Mol. Cancer Ther.*, **4**, 33–41.

Riggins,R.B. *et al.* (2008) ERRgamma mediates tamoxifen resistance in novel models of invasive lobular breast cancer. *Cancer Res.*, **68**, 8908–8917.

Ruan,J. and Zhang,W. (2006) A bi-dimensional regression tree approach to the modeling of gene expression regulation. *Bioinformatics*, **22**, 332–340.

Sala,A. *et al.* (1999) B-MYB transactivates its own promoter through SP1-binding sites. *Oncogene*, **18**, 1333–1339.

Segal,E. *et al.* (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.*, **34**, 166–176.

Segal,E. *et al.* (2008) Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature*, **451**, 535–540.

Sharan,R. *et al.* (2003) CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments. *Bioinformatics*, **19** (Suppl. 1), i283–i291.

Smola,A.J. and Scholkopf,B. (1998) A tutorial on support vector regression. *NeuroCOLT2 Technical Report*.

Spellman,P.T. *et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. Royal Stat. Soc. B*, **58**, 267–288.

Tusher,V.G. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Van den Bulcke,T. *et al.* (2006) SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics*, **7**, 43.

Wang,W. *et al.* (2005) Inference of combinatorial regulation in yeast transcriptional networks: a case study of sporulation. *Proc. Natl Acad. Sci. USA*, **102**, 1998–2003.

Xing,W. and Archer,T.K. (1998) Upstream stimulatory factors mediate estrogen receptor activation of the cathepsin D promoter. *Mol. Endocrinol.*, **12**, 1310–1321.

Yu,T. and Li,K.C. (2005) Inference of transcriptional regulatory network by two-stage constrained space factor analysis. *Bioinformatics*, **21**, 4033–4038.

Zhou,Q. and Wong,W.H. (2004) CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc. Natl Acad. Sci. USA*, **101**, 12114–12119.

Zhou,Y. *et al.* (2007) Enhanced NF kappa B and AP-1 transcriptional activity associated with antiestrogen resistant breast cancer. *BMC Cancer*, **7**, 59.