BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 12 2014, pages 1643–1650
doi:10.1093/bioinformatics/btu066

Genome analysis

Advance Access publication February 14, 2014

Condition-specific target prediction from motifs and expression

Guofeng Meng* and Martin Vingron

Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany

Associate Editor: Gunnar Ratsch

ABSTRACT

Motivation: It is commonplace to predict targets of transcription factors (TFs) by sequence matching with their binding motifs. However, this ignores the particular condition of the cells. Gene expression data can provide condition-specific information, as is, e.g. exploited in Motif Enrichment Analysis.

Results: Here, we introduce a novel tool named condition-specific target prediction (CSTP) to predict condition-specific targets for TFs from expression data measured by either microarray or RNA-seq. Based on the philosophy of guilt by association, CSTP infers the regulators of each studied gene by recovering the regulators of its co-expressed genes. In contrast to the currently used methods, CSTP does not insist on binding sites of TFs in the promoter of the target genes. CSTP was applied to three independent biological processes for evaluation purposes. By analyzing the predictions for the same TF in three biological processes, we confirm that predictions with CSTP are condition-specific. Predictions were further compared with true TF binding sites as determined by ChIP-seq/chip. We find that CSTP predictions overlap with true binding sites to a degree comparable with motif-based predictions, although the two target sets do not coincide.

Availability and implementation: CSTP is available via a web-based interface at http://cstp.molgen.mpg.de.

Contact: meng@molgen.mpg.de

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

Received on May 1, 2013; revised on November 20, 2013; accepted on January 28, 2014

1 INTRODUCTION

Gene expression is regulated at different levels, and one important level is transcriptional regulation (Levine and Tjian, 2003). Transcription factors (TFs) bind to promoter regions of genes and induce or repress expression of the target gene. This may further drive downstream cellular processes (Thomas and Chiang, 2006). Understanding of regulatory processes at the transcriptional level is essential in understanding the regulation of biological processes. It remains a great challenge for computational biologists to develop effective methods for identifying transcriptional regulatory interactions, given available sequences and expression data.

To investigate transcriptional regulation, one popular strategy is, for a given TF, to find genes with the binding motif in their promoters and those genes are supposed to be regulated by the TF. Many computational tools, such as MATCH (Kel *et al.*, 2003), TFBS (Lenhard and Wasserman, 2002) and TRAP (Roider *et al.*, 2007), have been developed based on this strategy. These tools rely on the description of the TF binding sites in the form of a position weight matrix (PWM). PWM-based methods have been widely used in recovering transcriptional targets in many biological processes.

The shortcomings of PWM-based methods become apparent when comparing their results with experimental data like gene expression measurements or TF location as determined by chromatin immunoprecipitation with subsequent array hybridization or sequencing (ChIP–chip/Chip-seq). All types of discrepancies can be observed: PWM matches may not appear to be bound as determined by ChIP (Consortium, 2012). Targets predicted based on the PWM method may not be affected in the studied biological process (Ong and Corces, 2011). On the other hand, TFs may bind in places where there is no binding motif and expression may change even though a gene's promoter lacks a binding site. All this is not surprising, though, given the fact that PWM-based prediction is a static sequence-based prediction that remains ignorant of the condition of the cell. Thus, the problem of prediction of transcriptional targets is an ill-posed one as long as it only considers sequence and ignores the condition of the cell.

One way to include the condition of the cell, in a rudimentary manner, is by determining common regulatory factors of co-expressed genes. Here, the condition of the cell enters through the co-expression. This approach forms the basis of what is called 'overrepresentation analysis' or 'motif-enrichment' (Ho *et al.*, 2007; Marstrand *et al.*, 2008; Roider *et al.*, 2009; Zambelli *et al.*, 2009). In overrepresentation analysis, TF binding motifs are searched in the promoter of each of the co-expressed genes and then the fraction of genes possessing a certain TF binding motif is judged based on statistical considerations. A TF that is significantly enriched is assumed to regulate its targets among the co-expressed genes. An underlying assumption is that co-expressed genes are co-regulated. One caveat, of course, is that due to the cascades of gene regulation, a co-expressed group may be due to many TFs acting serially, rather than one TF being responsible for all observed expression changes.

In this article, we extend overrepresentation analysis and put forward a strategy 'CSTP' to perform `condition-specific target prediction` within a specific biological process. In CSTP, we determine the co-expressed genes for each studied gene and then apply overrepresentation analysis to predict their potential regulators. Based on a philosophy of guilt by association, we make a key assumption that a regulator for the co-expressed

*To whom correspondence should be addressed.

group will in particular regulate its centroid gene, which is called core gene. Thus, the result of a CSTP prediction is a set of predicted TF–target relationships between TFs that have been identified by overrepresentation analysis and core genes that are used as the centroid of co-expressed groups. Because the regulators are inferred from such a group of co-expressed genes, there is no strict requirement that the predicted target actually possesses the respective binding site in its promoter. Thus, one can in principle predict regulatory interactions even in the absence of a binding motif in the promoter. The prediction also reflects the condition of the cell via the co-expression patterns. CSTP has been implemented as a web-based tool that takes user-provided expression data and outputs regulatory interactions between predicted TFs and genes.

2 METHODS

2.1 Prediction with CSTP for transcriptional regulation

The main idea underlying CSTP prediction is to infer transcriptional regulators for each studied gene by recovering the regulators of its co-expressed genes, as these are assumed to also regulate the gene in question. Therefore, we indirectly infer regulators of one gene through overrepresentation analysis of its co-expressed genes. If any TF is found to be overrepresented, this TF is predicted to regulate expression of the studied gene. This process is repeated for all genes and finally, the transcriptional regulation for all studied genes can be recovered with CSTP.

2.2 Core genes

Core genes are genes that are treated as centroids to find co-expressed gene groups. Only their regulators will be predicted with CSTP. Therefore, core genes should be genes that are transcriptionally regulated in the studied biological process. Experimental and computational methods can be used to find core genes. One way is to find genes with differential expression between different conditions. In this work, all the core genes in investigated experiments are determined by differential expression analysis between treated samples and control (or untreated samples). Differential expression analysis of microarray data is performed with `R` package *SAM* (Tusher *et al.*, 2001), which can assign a *q*-value to each gene for its significance of differential expression. A cutoff is chosen, above which differentially expressed genes are treated as core genes. Other methods can also be used to find core genes, such as choosing genes with enough expression variances, which is especially useful to time-series data without control. With the implementation of CSTP, users are required to determine which genes are to be treated as core genes and can choose methods based on their needs.

2.3 Co-expressed gene groups

For each core gene, we find its co-expressed gene group as follows. First, an expression vector, which describes the gene expression profile, is assigned to each gene with its expression values at different conditions as elements. For time-series data, each condition is one time point. Then, the expression similarities of gene pairs are measured with Pearson's correlation r between their expression vectors. Only those gene pairs with r greater than a certain threshold are considered to be co-expressed. Finally, the co-expressed gene group of core gene i is determined by selecting all of its co-expressed genes. Those co-expressed gene groups will be used for overrepresentation analysis. Based on our previous evaluation of overrepresentation analysis (Meng *et al.*, 2010), groups with an insufficient number of genes (e.g. <60 genes) are removed and not further analyzed. For those groups with >200 genes, genes will be sorted by descending r values, and only the top 200 genes of the list are used for analysis.

2.4 Overrepresentation analysis

Overrepresentation analysis is used to recover the transcriptional regulators of a co-expressed gene group. We built a local version of an overrepresentation analysis tool based on the method of `oPOSSUM` (version 2.0) (Ho *et al.*, 2007). It combines phylogenetic foot-printing and motif similarity to predict the TF binding sites and evaluates their enrichment with Fisher's exact test and *z*-score. We use 474 PWMs from three different sources, including the `Jaspar` Database (Bryne *et al.*, 2008), published work (Badis *et al.*, 2009) and the Transfac Database (June 2007). Parameters, used by overrepresentation analysis, are determined based on the evaluation in our previous work (Meng *et al.*, 2010), in which overrepresentation analysis was evaluated in 33 microarray experiments with clear molecular contexts, e.g. TF knockout versus control. All the parameters, including promoter length (from −2000 to +2000 bp), sequence conservation (top 30%), gene number (200), expression similarity (>0.6) and cutoff ($z>40$ and $P<0.01$) are chosen to make sure of the optimal performance of overrepresentation analysis. Detailed description for overrepresentation analysis is available in the 'Methods' section of Supplementary Materials.

3 RESULTS

3.1 The procedure of CSTP

The whole pipeline of CSTP is described in Figure 1. (a) Core genes are determined by differential expression analysis to

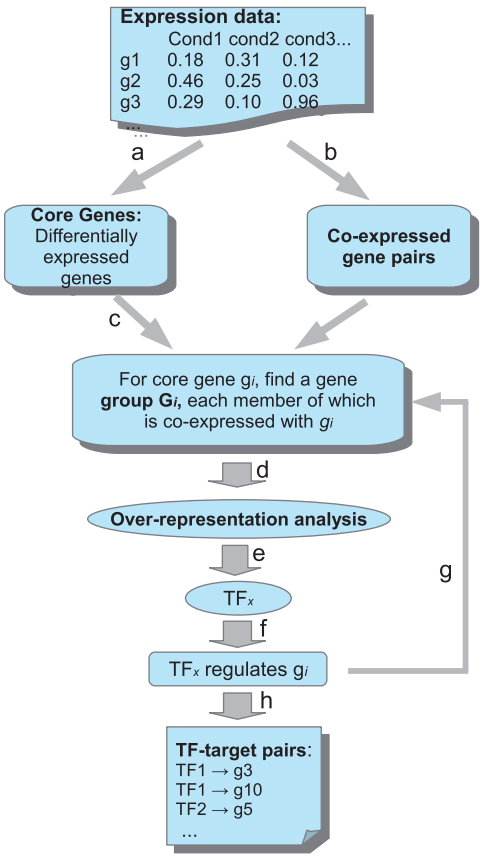

Fig. 1. The pipeline for CSTP prediction. See text for explanation

expression data. (b) Their expression similarities with other genes are measured with Pearson's correlation r. (c) For core gene i, genes with correlation r greater than a cutoff (e.g. >0.6) are treated to be co-expressed with gene i. The co-expressed gene group (G_i) of core gene i is determined by selecting its co-expressed genes. (d) Overrepresentation analysis is performed on co-expressed gene group G_i and (e) if TF x is predicted to be overrepresented, (f) TF x will be assumed to regulate the expression of gene i. (g) This process can be repeated for all core genes. (h) Finally, the transcriptional regulation of all the core genes is recovered with this pipeline. CSTP is being provided through a web-based interface at http://cstp.molgen.mpg.de.

3.2 CSTP identifies TF targets in estrogen stimulation process

Before proceeding to a more systematic evaluation, we present an example application of CSTP. We took microarray data from Lin *et al.* (2007) as an example. In this experiment, MCF7 cells were treated either with estradiol (estrogen) or control, and the expression levels of genes were measured with microarray at three time points: 12, 24 and 48 h. Each time point was repeated with three chips. Therefore, 18 microarrays were available to find differentially expressed and co-expressed genes.

Core genes were determined by differential expression analysis between estrogen-treated samples and control. At a cutoff of $q<0.001$, 1191 and 864 genes were observed to be up- and downregulated, respectively. Those differentially expressed genes were used as core genes to find their transcriptional regulators. Among them, we also observed some TF genes, which included 16 upregulated TF genes and 25 downregulated TF genes (Supplementary Fig. S1). Interestingly, expression of estrogen receptor alpha gene (ESR1) was significantly downregulated, which suggests that there may be some feedback loops to smooth estrogen stimulation effects by downregulating the expression of estrogen receptor (ER) gene.

Co-expressed gene groups of core genes were determined with the same microarray data and then CSTP was applied to recover the transcriptional regulation. As a result, 4373 TF-target pairs were predicted with CSTP. Of 2055 core genes, CSTP identified regulators for 1823 of them (88.7%). Among the core genes, there were 16 TF genes that were upregulated and 9 of them were predicted to take regulatory roles. For 25 downregulated TFs, 13 of them were also predicted to take regulatory roles. Of all TFs, MYC regulated the highest number of core genes, namely, 1106 of 2055. ERs, themselves TFs, were also predicted to regulate 807 genes.

Before usage of overrepresentation analysis, we have evaluated its performance with simulation data by randomly assigning a group of genes to be co-expressed, and estimated the false-positive prediction ratio for each TF binding motifs. With those ratios, we checked the probability to observe TFs to regulate such number of target genes. Testing based on a binomial distribution suggested that predicted TF–target pairs were impossible to be random (see Supplementary Methods and Supplementary Table S1).

3.3 Expression similarity between predicted TF and target genes

Above, we applied CSTP to predict TF–target regulation in the estrogen simulation process. If those predictions reflect true transcriptional regulation, expression of target genes is supposed to be dependent on that of TF genes. Therefore, TF and its target gene were expected to have similar or inversely similar expression patterns. Using ER as an example, we investigated this hypothesis.

We calculated the Pearson's correlation between expression of ESR1 and its target genes to describe their expression similarity. The smoothed histogram of those correlation values is shown in Figure 2. We found that most of the target genes had either strong positive or negative correlations with the ESR1 gene in expression pattern, which is in line with our expectations. In the same histogram, we also show the distribution of expression similarity with core genes that are not predicted to be regulated by ER. Genes without predicted ER regulation also showed expression similarity with the ESR1 gene, which might be due to indirect regulation by ER. However, genes with predicted ER regulation had stronger correlation peaks than genes without predicted ER regulation. Therefore, we used a Wilcoxon rank sum test to evaluate their correlation differences. The absolute values of correlation were used. Results show a significant difference at $P<2.2e{-}16$, which suggests that genes with predicted ER regulation have a significantly stronger expression similarity or inverse similarity to ER than those without predicted ER regulation.

We also randomly chose genes from the genome and checked their expression similarity to the ESR1 gene. We observed an expression peak near $r=0$, while no correlation peak near $r=1$ or $r=-1$ (see Fig. 2), which was different from that of genes with or without predicted ER regulation. This result further validates expression similarity or inverse similarity between ESR1 and its target genes, supporting the validity of the CSTP

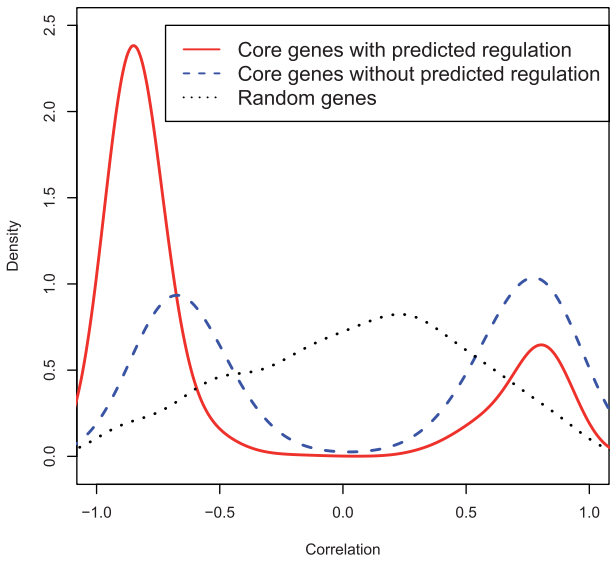

Fig. 2. Distribution of correlation coefficients between ER and its targets genes (solid) and other core genes (dashed) and random genes (dotted)

predictions. To be confident, we also performed the same evaluation on other TFs and their predicted target genes. As above, most TF genes and their predicted target genes have more positive or negative similarity, which is in line with evaluation results for ER (see Supplementary Fig. S2).

3.4 Validation of ER predictions based on other datasets

In the work of Lin *et al.* (2007), they not only performed microarray analysis but also investigated whole-genome ER binding sites with ChIP–PET. Finally, 234 genes were found to be associated with high-confident ER binding sites, 139 of which were also differentially expressed in response to estrogen simulation.

We used those ER binding genes that were determined by ChIP–PET, to evaluate the performance of CSTP. Based on CSTP predictions, we classified all the core genes into two groups: genes with predicted ER regulation and genes without predicted ER regulation. Then we checked their overlap ratio to genes with confirmed ER binding sites, respectively. Fisher's exact test was used to evaluate their ratio difference. Eighty of 807 predicted ER-regulated genes were validated with true ER binding sites, whereas only 59 of the 1252 genes without predicted ER regulation had ER binding sites. The *P*-value of Fisher's exact test was $4.47e-6$, which suggests a significant ratio differences between the two groups. This indicates that those genes with predicted ER regulation were more likely to have true ER binding sites.

Besides the ChIP–PET results, we also used independent ChIP-seq data to validate our prediction. One example was from Kong *et al.* (2011), where the binding sites of ER were detected with ChIP-seq in MCF7 cells. The same evaluation method as above was used, and the *P*-value of Fisher's exact test was $7.2e-4$, in line with results based on ChIP–PET.

We also checked the differential expression of those genes with predicted ER regulation. We searched public databases for experiments where the activity of the TF ER was manipulated and found three independent microarray experiments (Stein *et al.*, 2008; Kininis *et al.*, 2007; Saleh *et al.*, 2011). We performed differential expression analysis and used a cutoff of $P<0.01$ to define differentially expressed genes. With the same methods as above, we checked ratio differences between genes with and without predicted ER regulation. The results are shown in Table 1. Three independent microarray experiments all supported the hypothesis that core genes with predicted ER regulation were more likely to be differentially expressed in the experiments where activity of ER was modified, which also supported transcriptional regulation between ER and its predicted target genes.

With ChIP-seq/PET and microarray data, we have validated performance of CSTP in recovering true ER regulation. To be reliable, we further evaluated its performance with other TFs, such as MYC, FOXC1, SOX3 and E2F6, which were all predicted to take regulatory roles in estrogen simulation process. However, most of those TFs have not been studied with either ChIP-seq/chip or microarray/RNA-seq experiments in the appropriate cell lines. In a search of public databases, we only found data for MYC and E2F6. The same analyses were performed as for ER, and the results are shown in Table 2. Like for ER, we found that target genes with predicted MYC regulation were significantly enriched with both true binding sites and differentially expressed genes, which validated the performance of CSTP. With E2F6, there was only one ChIP–chip experiment available, and we failed to find significant overlap with CSTP prediction. Owing to the limited experiment number, it is hard to conclude further. Overall, even with this exception, evaluation results with independent high-throughput data significantly supported the claim that CSTP predictions reflected true transcriptional regulation.

3.5 Validation of CSTP based on other biological processes

Above, we have described how we validated the performance of CSTP in recovering the transcriptional regulation in the estrogen stimulation process. We further evaluated CSTP in other biological processes. Searching through the public databases, two microarray experiments were found that could help testing our approach. One was carried out by Kazmin *et al.* (2006), where LNCaP cells were treated with androgen or control and gene expression levels were measured with microarrays at two time points; in this experiment, the activity of androgen receptor (AR) was manipulated with androgen stimulation. Another one was from Peltonen *et al.* (2010), in which MCF7 cells were treated with different compounds to activate p53 activity. For both experiments, we defined core genes by differential expression analysis to their expression data and predicted transcription regulation with CSTP.

In the biological processes studied with the above two microarray experiments, AR and TP53 were known as the initial and

Table 1. *P*-values for overlap between predicted ER targets and experimental targets

Experiment	Type	*P*
Lin *et al.* (2007)	ChIP–PET	$4.47E-6$
Kong *et al.* (2011)	ChIP-seq	$7.20E-4$
Stein *et al.* (2008)	Microarray	$1.76E-3$
Kininis *et al.* (2007)	Microarray	0.09
Saleh *et al.* (2011)	Microarray	0.05

Table 2. *P*-values for overlap between predicted MYC and E2F6 targets derived from the ER experiment with their experimental targets

Experiment	Type	*P*
MYC		
Song *et al.* (2011)	ChIP-seq	4.99E−5
Hua *et al.* (2009)	ChIP–Chip	9.06E−3
Raha *et al.* (2010)	ChIP-seq	2.10E−9
Cappellen *et al.* (2007)	Microarray	2.37E−3
Musgrove *et al.* (2008)	Microarray	4.1E−20
E2F6		
Xu *et al.* (2007)	ChIP–Chip	0.18

Table 3. *P*-values for overlap between predicted Ar and TP53 targets with their experimental targets

Experiments	Type	*P*-value
Ar		
Wyce *et al.* (2010)	ChIP–chip	6.63E−5
Altintas *et al.* (2013)	Microarray	1.93E−3
Gonit *et al.* (2011)	Microarray	2.18E−4
Vellaichamy *et al.* (2009)	Microarray	1.88E−7
TP53		
Botcheva *et al.* (2011)	ChIP-seq	4.02E−6
Smeenk *et al.* (2011)	ChIP–chip	7.21E−5
Grinkevich *et al.* (2009)	Microarray	4.77e−2
Herbert *et al.* (2010)	Microarray	0.52
Girardini *et al.* (2011)	Microarray	0.99

presumably most important regulators. Therefore, we used them to assess the confidence of our prediction. The same evaluation methods were used as for ER. The results are shown in Table 3. For AR, we found one ChIP–chip experiment and three microarray datasets and significant overlaps were observed, which supported our hypothesis that core genes with predicted AR regulation were more likely to have true AR binding sites or to be differentially expressed after stimulation by AR (see Table 3). Similar results could be observed for TP53. Two ChIP-seq/chip datasets suggested that core genes with predicted p53 regulation were enriched with true p53 binding sites. In one of the three microarray experiments, we observed significant overlap to genes with differential expression pattern. However, we did not observe significant overlap in the remaining two microarray experiments. One possible explanation was that p53 regulated other regulators and the second-wave regulation concealed the signals of p53. Overall, CSTP evaluation with independent biological processes validated the performance of our methods.

3.6 CSTP with comparable performance with PWM-based methods

The common computational methods to recover transcriptional regulation are PWM-based methods with which promoters of genes are searched for TF binding motifs by matching to annotated PWMs. CSTP is different from standard PWM-based methods, as that it does not insist on clear TF binding sites in promoters of predicted target genes. Although we have confirmed that CSTP can recover true TF–target relations, we are still not clear whether it could achieve a comparable performance with standard PWM-based methods. Therefore, we performed comparison in the three biological processes mentioned above.

MATCH, an implementation of PWM-based methods from TRANSFAC (Kel *et al.*, 2003), is used to predict TF binding sites in the promoter regions of core genes, which can be clustered into two groups based on the presence of predicted TF binding sites. As what we did above, Fisher's exact test is used to check their overlap with true TF binding sites. Because performance of MATCH hinges on the selection of promoter ranges, we first checked the performance of MATCH at different promoter lengths. We define promoters at a range from x to $+500$ bp around transcription start sites, where x varies from $-100\,000$ to -2000 bp. Figure 3a shows the *P*-values for their overlap to the true ER binding sites at different x values. At $x = 2000$ bp, we observed the most overlapping significance. With the increasing of promoter length from $x = 2000$ bp, the significance decreases, which may result from increasing false-positive prediction. At the promoter length from 10 000 to 30 000 bp, we also observed peaks with weak significance, which was in line with previous reports about enriched ER binding sites at distal promoters (Lin *et al.*, 2007).

We compared the performances of MATCH and CSTP in the estrogen stimulation process. The results are shown as a Venn diagram in Figure 3b. We found that only part of the predictions was shared by both CSTP and MATCH. However, neither of them recovered all of the true ER binding sites. We used Fisher's exact test to evaluate their overlap ratio with ChIP–PET results and found that the significance of CSTP was $P < 4.47e-6$, which was more significant than that of MATCH at $P < 7.69e-3$. The same evaluation was performed for ER with ChIP-seq data from the work of Kong *et al.* (2011) as shown in Figure 3c. The significance of CSTP is at $P < 7.21e-6$, which is better than that of MATCH at $P < 8.92e-4$.

We also compared the performance of CSTP and MATCH on data from the other biological processes used above. For the androgen stimulation process, AR binding sites were predicted based on data from Kazmin *et al.* (2006) and the true AR binding sites were determined by a ChIP–chip experiment (Wyce *et al.*, 2010). Figure 3d shows the results of our evaluation. CSTP prediction overlaps with true AR binding sites with a significance of $P < 6.63e-5$, while the significance for MATCH was $P < 0.55$. The same analysis was performed for p53 with data from Peltonen *et al.* (2010), whose true binding sites were determined with two ChIP-seq experiments (Botcheva *et al.*, 2011; Smeenk *et al.*, 2011). More significant overlap was observed with CSTP predictions, which were supported by both ChIP-seq results (see Fig. 3e and f).

To be confident, we also performed the same evaluation with three other tools: TFBS (Lenhard and Wasserman, 2002), Trap (Roider *et al.*, 2007), FIMO (Grant *et al.*, 2011), all of which are implementations of PWM-based methods. Their evaluation results are similar to that of MATCH (see Supplementary Fig. S3). In summary, these results indicate that CSTP can achieve comparable performance to standard PWM-based methods.

3.7 Condition-specific target prediction with CSTP

Prediction of CSTP is based on the expression patterns in the biological processes under study. Thus, we expect the prediction for the same TFs in different biological processes to be different, i.e. CSTP prediction is expected to be condition-specific. To validate this, we took ER as an example and investigated the prediction in three independent biological processes.

First, we checked the overlap of ER target genes that were predicted based on three sets of data from Lin *et al.* (2007), Kazmin *et al.* (2006) and Kininis *et al.* (2007) respectively. As shown in Figure 4a, we found only three genes were shared as predicted targets of ER in all three biological processes. The

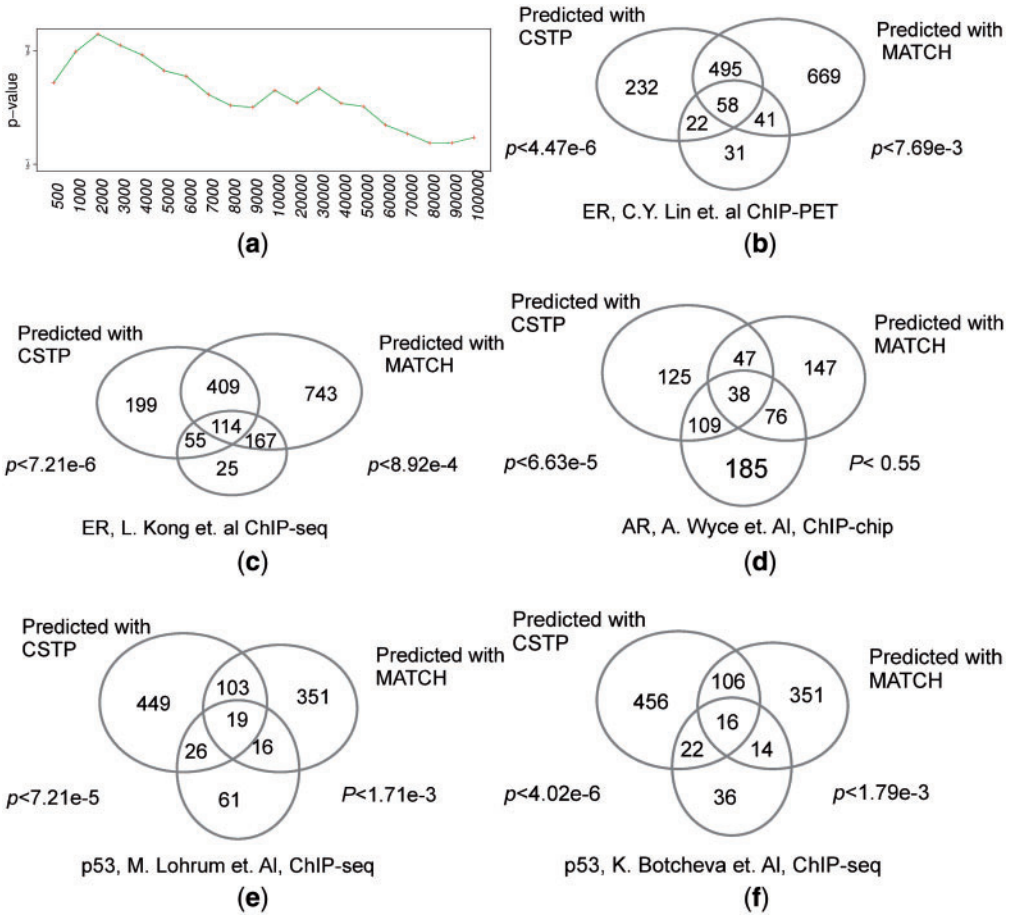

Fig. 3. Comparison of CSTP with MATCH. (**a**) Significance of overlap with true targets varied across different promoter lengths for PWM location. (**b–f**) Venn diagrams comparing CSTP targets, MATCH targets and ChIP–PET/seq targets for different datasets

overlaps between each pair of biological processes are also limited, <10%. This result suggests that CSTP can predict different target genes for the same TFs under different biological conditions.

Then, we investigated whether those predictions reflect true condition-specific regulation. If genes are regulated by ER, their expression is supposed to be correlated (or inversely correlated) to the expression of ER genes. Therefore, we used Pearson's correlation to evaluate the expression similarity between ESR1 and its target genes. Predictions for three independent processes were all checked with the same expression data. Figure 4b presents the smoothed histogram of correlation values that were calculated with expression data from Chin-Yo Lin *et al.*'s (2007) work. We observe that correlation peaks of predicted target genes based on Chin-Yo Lin *et al.*'s (2007) data were closer to +1 or −1 than predictions from the other two biological processes, which indicated more significant expression similarity. The same evaluations were also performed with expression data of the other two processes. Similarly, the more significant expression similarity was observed with predictions associated with the tested expression data (see Fig. 4c and d). These results suggest that CSTP is capable of recovering condition-specific regulation in the studied biological processes.

4 DISCUSSION

In this article, we introduce a computational method for the delineation of regulatory relationships between TFs and target genes, in a manner that is specific for a biological process where gene expression experiments are available. In the public database, numerous expression data from either microarray or RNA-seq experiments are being available for different conditions, which provides data foundation for computational prediction, which promises CSTP to be useful for both biologist and bioinformaticians.

With CSTP, only the core genes are predicted for their transcriptional regulators. Proper selection of core genes is essential for accuracy of prediction. First, core genes should be regulated at transcriptional level in the studied biological process. One obvious feature of those genes is their differential expression in microarray or RNA-seq experiments. Otherwise, the expression vector of the core gene lacks the variance to describe the changes under different conditions, which makes its co-expressed genes less likely to be co-regulated. Such genes might, e.g. be housekeeping genes that are stably expressed under different conditions or in different tissues (Eisenberg and Levanon, 2003). Second, a core gene should have enough co-expressed genes as

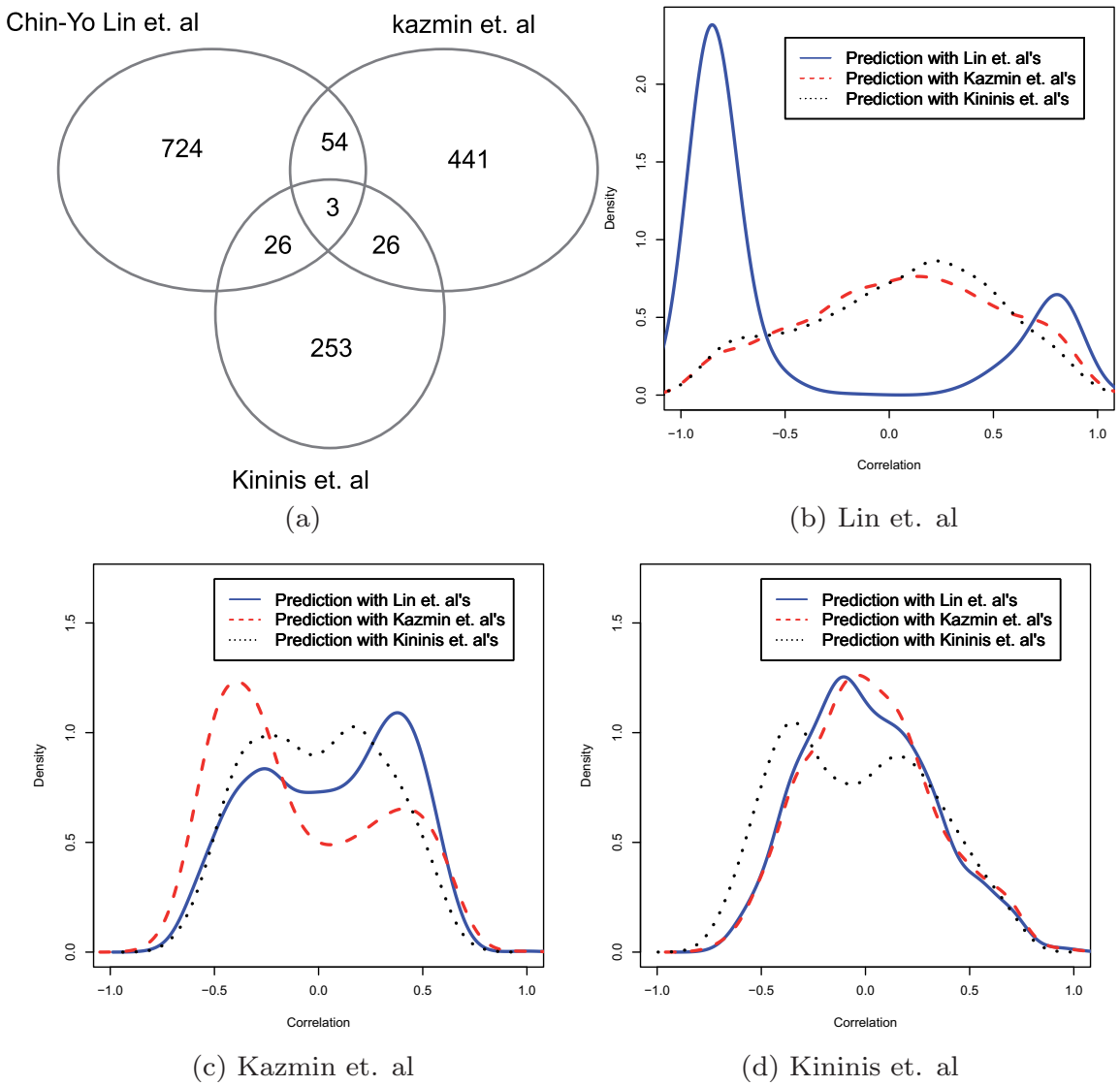

Fig. 4. (**a**) Venn diagram for three CSTP predictions computed from three different condition-specific expression datasets. (**b–d**) Condition-specific distributions of correlations among different prediction

its neighborhood. Our method recovers the regulators of each core gene by recovering the regulators of its co-expressed gene group. Those core genes with few co-expressed genes will be removed for further analysis. Taking Chin-Yo Lin's data, we found 51 co-expressed gene groups without enough co-expressed genes for overrepresentation analysis. These considerations led us to define the core genes as differentially expressed genes between conditions. It would also be feasible to define core genes for time-series gene expression experiments by selecting ones with most expression variances among different time points.

CSTP does not require clear TF binding motifs in the promoters of predicted target genes. This setting enables CSTP to recover the transcriptional regulation with degenerate motifs or distal binding. One example is ER binding. It is reported that only 5% of its binding sites locate in proximal promoters (Lin *et al*., 2007). Investigation to promoter will ignore those binding sites in enhancers or distal regions. CSTP helps to improve the recovery of those ER regulation. In estrogen simulation process, 70 genes with true ER binding sites were also predicted with ER regulation, whereas 31% of them had no ER binding motif in their promoters. However, non-insistence to TF also complicates the removal of false-positive prediction. One possible problem is indirect regulation. It is especially true in a regulation pathway like A →B →C. If B and C are both core genes and they have similar expression profiles, CSTP may predict both B and C to be regulated by A. One possible example is MYC. In estrogen simulation process, MYC is predicted to repress the expression of ESR1 genes process. This is also confirmed with microarray data (Musgrove *et al*., 2008). However, ChIP-seq/chip results do not support the direct regulation (Hua *et al*., 2008; Raha *et al*., 2010). Therefore, we guess that some other TF might take a mediator role. With some luck, analysis of the network

can help in clarifying the situation: if an intermediate TF gets identified by CSTP, then elimination of transitive edges can lead to better predictions of direct targets.

CSTP aims at infusing condition-specific information into target prediction. There is clearly a wide gap between our ability to computationally predict TF targets and the target sets for a TF that can be observed experimentally (Gerstein *et al*., 2012). The CSTP predictions discussed in Section 3 tend to be equally reasonable predictions as PWM-based predictions, even when our method has no insistence on clear TF binding sites in the promoters. Both sets of predictions have their validity, but the two sets are still largely different. Likewise, the CSTP predictions for different conditions are different, in line with biological expectation. In this work, we did not include ChIP-seq data into the prediction algorithm because we needed the location data for validation purposes. Given that from overrepresentation and gene expression, we could achieve reasonably good results, we would hope that future inclusion of TF location data will get us closer to an understanding of which TF induces which expression changes in a certain condition.

ACKNOWLEDGEMENT

The authors thank Kirsten Kelleher for editing the manuscript. Language errors, if any, were introduced after the manuscript reading by Kirsten Kelleher.

Conflict of Interest: none declared.

REFERENCES

Altintas,D.M. *et al.* (2013) Differentially expressed androgen-regulated genes in androgen-sensitive tissues reveal potential biomarkers of early prostate cancer. *PLoS One*, **8**, e66278.

Badis,G. *et al.* (2009) Diversity and complexity in DNA recognition by transcription factors. *Science*, **324**, 1720–1723.

Botcheva,K. *et al.* (2011) Distinct p53 genomic binding patterns in normal and cancer-derived human cells. *Cell Cycle*, **10**, 4237–4249.

Bryne,J.C. *et al.* (2008) Jaspar, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.*, **36**, D102–D106.

Cappellen,D. (2007) Novel c-myc target genes mediate differential effects on cell proliferation and migration. *EMBO Rep*., **8**, 70–76.

Consortium,T.E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Eisenberg,E. and Levanon,E.Y. (2003) Human housekeeping genes are compact. *Trends Genet*., **19**, 362–365.

Gerstein,M.B. *et al.* (2012) Architecture of the human regulatory network derived from encode data. *Nature*, **489**, 91–100.

Girardini,J.E. *et al.* (2011) A pin1/mutant p53 axis promotes aggressiveness in breast cancer. *Cancer Cell*, **20**, 79–91.

Gonit,M. *et al.* (2011) Hormone depletion-insensitivity of prostate cancer cells is supported by the ar without binding to classical response elements. *Mol. Endocrinol*., **25**, 621–634.

Grant,C.E. *et al.* (2011) Fimo: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.

Grinkevich,V.V. *et al.* (2009) Ablation of key oncogenic pathways by rita-reactivated p53 is required for efficient apoptosis. *Cancer Cell*, **15**, 441–453.

Herbert,B.-S. *et al.* (2010) A molecular signature of normal breast epithelial and stromal cells from li-fraumeni syndrome mutation carriers. *Oncotarget*., **1**, 405–422.

Ho,S. *et al.* (2007) Opossum: integrated tools for analysis of regulatory motif over-representation. *Nucleic Acids Res*., **35**, W245–W252.

Hua,S. *et al.* (2008) Genomic analysis of estrogen cascade reveals histone variant h2a.z associated with breast cancer progression. *Mol. Syst. Biol*., **4**, 188.

Hua,S., Kittler,R. and White,K.P. (2009) Genomic antagonism between retinoic acid and estrogen signaling in breast cancer. *Cell*, **137**, 1259–1271.

Kazmin,D. *et al.* (2006) Linking ligand-induced alterations in androgen receptor structure to differential gene expression: a first step in the rational design of selective androgen receptor modulators. *Mol. Endocrinol*., **20**, 1201–1217.

Kel,A.E. *et al.* (2003) Match: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res*., **31**, 3576–3579.

Kininis,M. *et al.* (2007) Genomic analyses of transcription factor binding, histone acetylation, and gene expression reveal mechanistically distinct classes of estrogen-regulated promoters. *Mol. Cell. Biol*., **27**, 5090–5104.

Kong,S.L. *et al.* (2011) Cellular reprogramming by the conjoint action of er, foxa1, and gata3 to a ligand-inducible growth state. *Mol. Syst. Biol*., **7**, 526.

Lenhard,B. and Wasserman,W.W. (2002) Tfbs: computational framework for transcription factor binding site analysis. *Bioinformatics*, **18**, 1135–1136.

Levine,M. and Tjian,R. (2003) Transcription regulation and animal diversity. *Nature*, **424**, 147–151.

Lin,C.Y. *et al.* (2007) Whole-genome cartography of estrogen receptor alpha binding sites. *PLoS Genet*., **3**, e87.

Marstrand,T.T. *et al.* (2008) Asap: a framework for over-representation statistics for transcription factor binding sites. *PLoS One*, **3**, e1623.

Meng,G. *et al.* (2010) A computational evaluation of over-representation of regulatory motifs in the promoter regions of differentially expressed genes. *BMC Bioinformatics*, **11**, 267.

Musgrove,E.A. *et al.* (2008) Identification of functional networks of estrogen- and c-myc-responsive genes and their relationship to response to tamoxifen therapy in breast cancer. *PLoS One*, **3**, e2987.

Ong,C.T. and Corces,V.G. (2011) Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat. Rev. Genet*., **12**, 283–293.

Peltonen,K. *et al.* (2010) Identification of novel p53 pathway activating small-molecule compounds reveals unexpected similarities with known therapeutic agents. *PLoS One*, **5**, e12996.

Raha,D. *et al.* (2010) Close association of RNA polymerase II and many transcription factors with pol III genes. *Proc. Natl Acad. Sci. USA*, **107**, 3639–3644.

Roider,H.G. *et al.* (2007) Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, **23**, 134–141.

Roider,H.G. *et al.* (2009) Cpg-depleted promoters harbor tissue-specific transcription factor binding signals–implications for motif overrepresentation analyses. *Nucleic Acids Res*., **37**, 6305–6315.

Saleh,S.A. *et al.* (2011) Estrogen receptor silencing induces epithelial to mesenchymal transition in human breast cancer cells. *PLoS One*, **6**, e20610.

Smeenk,L. *et al.* (2011) Role of p53 serine 46 in p53 target gene regulation. *PLoS One*, **6**, e17574.

Stein,R.A. *et al.* (2008) Estrogen-related receptor alpha is critical for the growth of estrogen receptor-negative breast cancer. *Cancer Res*., **68**, 8805–8812.

Song,L. *et al.* (2011) Open chromatin defined by dnasei and faire identifies regulatory elements that shape cell-type identity. *Genome Res*., **21**, 1757–1767.

Thomas,M.C. and Chiang,C.M. (2006) The general transcription machinery and general cofactors. *Crit. Rev. Biochem. Mol. Biol*., **41**, 105–178.

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

Vellaichamy,A. *et al.* (2009) Proteomic interrogation of androgen action in prostate cancer cells reveals roles of aminoacyl trna synthetases. *PLoS One*, **4**, e7075.

Wyce,A. *et al.* (2010) Research resource: the androgen receptor modulates expression of genes with critical roles in muscle development and function. *Mol. Endocrinol*., **24**, 1665–1674.

Xu,X. *et al.* (2007) A comprehensive chip-chip analysis of e2f1, e2f4, and e2f6 in normal and tumor cells reveals interchangeable roles of e2f family members. *Genome Res*., **17**, 1550–1561.

Zambelli,F. *et al.* (2009) Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes. *Nucleic Acids Res*., **37**, W247–W252.