CHIPMODULE: SYSTEMATIC DISCOVERY OF TRANSCRIPTION FACTORS AND THEIR COFACTORS FROM CHIP-SEQ DATA

JUN DING
Department of EECS, University of Central Florida, 4000 central Florida Blvd
Orlando, FL 32816, USA
Email: jding@cs.ucf.edu

XIAOHUI CAI
Shanghai Center for Bioinformation Technology, 100 Qinzhou Rd, Bldg.1, Fl.12
Shanghai, 200235, China
Email: xhcai@scbit.org

YING WANG
Department of EECS, University of Central Florida, 4000 central Florida Blvd
Orlando, FL 32816, USA
Email: ying2010@knights.ucf.edu

HAIYAN HU*
Department of EECS, University of Central Florida, 4000 central Florida Blvd
Orlando, FL 32816, USA
Email: haihu@cs.ucf.edu

XIAOMAN LI*
Burnett School of Biomedical Science, University of Central Florida, 4000 central Florida Blvd
Orlando, FL 32816, USA
Email: xiaoman@mail.ucf.edu

*co-correspondence authors

We have developed a novel approach called ChIPModule to systematically discover transcription factors and their cofactors from ChIP-seq data. Given a ChIP-seq dataset and the binding patterns of a large number of transcription factors, ChIPModule can efficiently identify groups of transcription factors, whose binding sites significantly co-occur in the ChIP-seq peak regions. By testing ChIPModule on simulated data and experimental data, we have shown that ChIPModule identifies known cofactors of transcription factors, and predicts new cofactors that are supported by literature. ChIPModule provides a useful tool for studying gene transcriptional regulation.

1. Introduction

Systematic discovery of transcription factors (TFs) and their cofactors is important for studying gene transcriptional regulation. During gene transcriptional regulation, TFs and their

cofactors bind short DNA segments to activate or repress the expression of genes nearby. In general, a TF can bind to a variety of similar DNA segments, called TF binding sites (TFBSs) of this TF. The common pattern of the TFBSs bound by a TF is termed a motif, often represented as a position weight matrix (PWM) or a consensus sequence. In eukaryotes, multiple TFs often bind their TFBSs in short DNA regions of several hundred base pairs long.[1-3] These short DNA regions are called cis-regulatory modules (CRMs).[1] CRMs are common in high eukaryotes.[3] For instance, more than 110,000 CRMs have been predicted in the human genome and are supported by various sources of functional evidence.[4,5] It is the interaction of multiple TFs and their TFBSs instead of individual TFs that determines the temporal spatial expression patterns of genes.[1,4,5] It is thus critical to identify and study TFs and their cofactors.

The chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-seq) experiments provide an unprecedented opportunity for computational methods to study TFs and their cofactors.[6,7] In a typical ChIP-seq experiment, short DNA segments containing TFBSs of a TF are enriched by the chromatin immunoprecipitation (ChIP) using an antibody specific to the TF. These short DNA segments are then sequenced by the next generation sequencing technologies and mapped to a reference genome.[8-11] Finally, genomic regions in the reference genome enriched with the mapped DNA segments are identified as ChIP-seq peak regions.[12,13] These ChIP-seq peak regions likely contain TFBSs of the TF under consideration.[6,7,14] Compared with potential residing regions of the TFBSs of a TF for a gene,[4,5] which is often several hundred million base pair long, a ChIP-seq peak region is typically no longer than 1000 base pairs. Such short regions thus significantly increase the signal to noise ratio and dramatically help to improve the efficiency of computational identification of TFBSs of a TF and its cofactors.

Available computational methods have already started to provide useful prediction of TFs and their cofactors from ChIP-seq data.[12,14-21] The majority of these computational methods identify motifs of individual TFs at a time.[12,14-17,19-21] The underlying assumption of these methods is that motifs of individual cofactors of a TF are overrepresented in the ChIP-seq peak regions of this TF. However, given the fact that a TF has multiple cofactors and motifs of most cofactors only occur in a small portion of peak regions under a condition, motifs of individual cofactors may be often not overrepresented in the ChIP-seq peak regions of this TF.[18,22] Because TFs and their cofactors often regulate their target genes by binding to CRMs in eukaryotes, one recent study has considered motif co-occurrence of a TF and one of its cofactors.[18] However, a TF may bind regulatory regions together with more than one cofactor to regulate its target genes.[1,4,5]

Here we developed a computational method called ChIPModule to systematically identify TFs and cofactors from ChIP-seq data. ChIPModule considers the co-occurrence of TFBSs of any number of different TFs in ChIP-seq peak regions. In brief, starting from all known TF motifs in public databases,[23,24] ChIPModule scans the ChIP-seq peak regions with these motifs to define putative TFBSs of these TFs. ChIPModule then identifies frequently co-occurring TFBSs of a group of any number of TFs by frequent pattern mining methods.[25,26] Finally, ChIPModule assesses the statistical significance of each group of TFs with frequent co-occurring TFBSs by the Poisson clumping heuristic.[27] The significant groups of TFs are called interacting TF groups. The

TFs in the same interacting TF group with a given TF are designated as the cofactors of this TF. Tested on simulated and experimental data, ChIPModule has been shown to successfully predict known cofactors of TFs. It also predicts new cofactors that were supported by literature. Compared with other methods, ChIPModule shows superior performance in terms of dealing with large datasets and identifying known cofactors. We believe ChIPModule will be useful for future ChIP-seq data analysis and gene transcriptional regulation studies.

2. Materials and Methods

2.1. *Framework*

To systematically discover TFs and their cofactors from ChIP-seq data, ChIPModule utilizes the known TF motif information in the TRANSFAC database.[24] Instead of considering one or two TFs at a time, ChIPModule can consider any number of TFs simultaneously. Instead of assuming TFBSs of individual TFs are overrepresented in the ChIP-seq peak regions, ChIPModule assumes that TFBSs of a group of TFs (a TF and its cofactors) are overrepresented in the ChIP-seq peak regions. The framework of ChIPModule consists of the following three steps: prediction of putative TFBSs, identification of frequent co-occurring TF groups, and discovery of TFs and their cofactors. See Figure 1 for the flowchart of ChIPModule. The details are in the following sections.

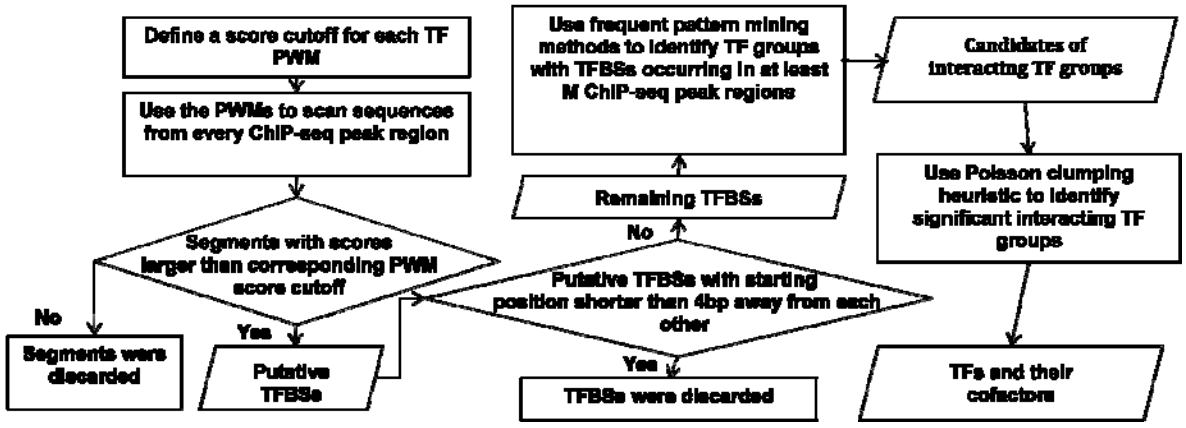

Figure 1. The flowchart of ChIPModule to discover TFs and their cofactors.

2.2. *ChIP-seq Data and Vertebrate Motifs*

We tested ChIPModule on two ChIP-seq datasets and several simulated datasets. The two ChIP-seq datasets are corresponding to the two TFs ESR1 and E2F1, respectively. For ESR1, which is also called estrogen receptor alpha, the ChIP-seq peak regions defined at the p-value cutoff 0.001 were downloaded from the GSM365926 sample in the GEO database.[28] In total, we obtained 3257 peak regions, with the average length of 595 base pairs. For E2F1, the peak region at the p-value cutoff 0.001 were downloaded from the SYDH TFBS track at the UCSC genome

browser.[29] We obtained 10196 peak regions, with the average length of 878 base pairs for E2F1. For each ChIP-seq peak region, we extended it equally on the two sides such that it is at least 800 base pairs long. This extension is to enhance the chance for cofactors to occur in peak regions, as TFBSs of certain cofactors may be not within the originally defined ChIP-seq peak regions. The known TF motifs used in the following study were obtained from the TRANSFAC 9.2 database,[24] where all 522 vertebrate PWMs were extracted. Pseudo counts were introduced to regularize each PWM, as in previous studies.[30,31]

2.3. *Identification of Putative TFBSs in ChIP-seq Peak Regions*

To identify putative TFBSs of a TF in ChIP-seq peak regions, we scan the non-repetitive sequences in each peak region and calculate the score of each segment in a peak region by using the above regularized PWM of this TF. A slide window is used to define segments. That is, given a TF motif of length k and a peak region of length L, we consider all L – k+1 distinct segments. We calculate the score of a segment by the following formula: $score(a\ segment\ x_1x_2\cdots x_k) = \sum_{i=1}^{k} \log\left(\frac{f(x_i,i)}{fb(x_i)}\right)$. Here $fb(x_i)$ is the average frequency of the nucleotide x_i in the human reference genome, $f(x_i, i)$ is the frequency of the nucleotide x_i at the i-th position of the motif PWM, and k is the width of the motif. If the score is larger than a predefined cutoff for this TF, this segment will be claimed as a putative TFBS of this TF. In this study, the predefined cutoff for each TF is defined as the 99.99% quartile of the score distribution of DNA segments of length k, when using the PWM of this TF to scan 100 kb long random sequences. The random sequences were generated by permuting input sequences from ChIP-seq peak regions. Note that motifs of certain TFs may have the tendency to occur together, merely due to the similarity of their PWMs. To deal with it, we sort the putative TFBSs by their start positions and discard the overlapped TFBSs with the lower score, when the start positions of two putative TFBSs are smaller than 4 base pairs. We use 4 base pairs here to remove overlapping TFBSs as in previous studies.[31,32]

2.4. *Identification of Groups of TFs with Frequently Co-occurring Motifs in Peak Regions*

We aim to identify groups of TFs whose putative TFBSs co-occur in more than a specified number of peak regions, say *M* peak regions. The rationale is that the chance that multiple TFs with their TFBSs co-occurring in a ChIP-seq region is much smaller than that of individual TFs. That is, if we observe a group of TFs with their TFBSs co-occurring in a large number of peak regions, it is likely their co-occurrence is not by chance and thus this group of TFs likely work together to regulate genes. To discover such a group of TFs, we use a tree to represent the above identified TFBSs and identify all groups of TFs with their TFBSs co-occurring in at least *M* peak regions (Figure 2). In brief, first, we count the number of peak regions containing TFBSs of each TF and sort these TFs according to the corresponding number, from the largest to the smallest. Second, we sort the TFBSs in each peak region, such that TFBSs of the TFs occurring in more peak regions rank at the beginning. Third, starting from the first peak region until the last peak

region, we build a tree to store the TFs whose TFBSs occurring in a peak region (Figure 2). At the beginning, a tree with only a root node is built. Next, the nodes for TFs in the first peak region are added in order. Finally, nodes for TFs in other peak regions are added, if there is no branch in the current tree matching the order of TFs in the peak regions under consideration (Figure 2). With the built tree, we will identify all groups of TFs with TFBSs occurring in at least *M* peak regions. In brief, starting from the TF that occurs in at least M peak regions and occurs in the smallest number of peak regions, we will obtain all the branches in the built tree that contains this TF. For instance, when M=2, we will start from the TF M1 or M6 in Figure 2. Assume we will start from the TF M1. In this case, we obtain two branches, M4:3-M3:3-M1:1 and M7:1-M1:1. We will then construct a tree using the obtain branches for this specific TF, by assuming each branch represent motifs in a peak region. In this case, we will have a tree with the above two branches. It is clear that no group of TFs that includes TF M1 and occurs at least M times. Next, we will obtain all branches and construct a tree for the TF that occurs in the second smallest number of peak regions. Since we already consider the TF M1, this time TF M6 occurs in the second smallest number of peak regions. This time we have only one branch that containing M6, which is M4:3-M3:3-M7:2-M6:2. In this case, it is evident that the group of TFs (M4,M3,M7,M6) co-occur twice in the peak regions considered in Figure 2. We will keep considering a TF each time until we find the groups of TFs that co-occur at least M times for the TF occurring in the most peak regions.

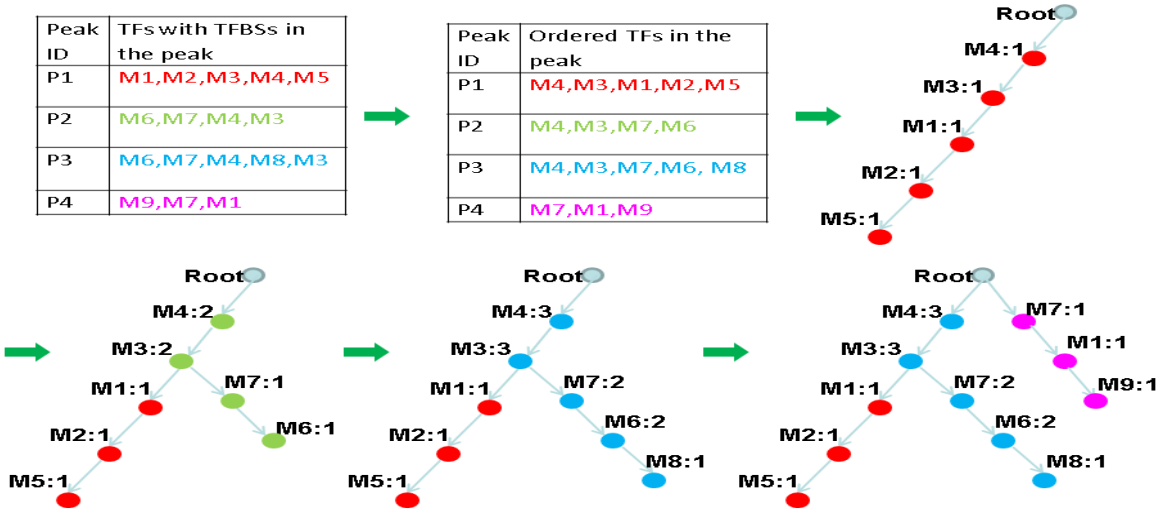

Figure 2. The procedure to construct a tree to represent TF co-occurrence.

2.5. *Identification of TFs and Their Cofactors*

With groups of TFs identified above, we want to assess their statistical significance to obtain interacting TFs. As mentioned above, a TF in a group of interacting TFs is a cofactor of all other TFs in the same group and vice versa. We use the Poisson clumping heuristic[27] to compute the statistical significance of a group of TFs with the assumption that each TF bind a ChIP-seq peak region independently according to a Poisson process. In brief, assume there are N_1 ChIP-seq peak regions, and the average length of a peak region is L, the total number of known motifs is N_2, and λ_k is the rate parameter of the Poisson process for the k-th TF. For a group of TFs composed of

the $m_1, m_2, \ldots, m_n$-th TFs identified above, the probability that TFBSs of the n TFs occur in a peak region of length L is $P_1 = \prod_{i=1}^{n}(1 - e^{-L\lambda_{m_i}})$. The probability that this group of TFs with TFBSs co-occurring in at least K peak regions is $P_2 = 1 - \sum_{k=0}^{K-1} \mathrm{C}_{\mathrm{N}_1}^{\mathrm{k}} \mathrm{P}_1^{\mathrm{k}} (1 - \mathrm{P}_1)^{(\mathrm{N}_1 - \mathrm{k})}$. Since N_2 TFs can produce $\mathrm{C}_{\mathrm{N}_2}^{\mathrm{n}}$ different group of TFs by chance, we require $P_2 < 0.05/C_{N_2}^{n}$ to claim a group of TFs as a group interacting TFs. With the groups of interacting TFs, we then treat each TF and all other TFs from the same group of interacting TFs as TFs and cofactors. The genes closest to the peak regions containing TFBSs of a group of interacting TFs are defined as the target genes of this group of interacting TFs. Similarly, genes closest to the peak regions containing TFBSs of a TF are defined as the target genes of this TF.

3. Results

3.1. *ChIPModule Identified Implanted TFs and Their Cofactors in Simulated Data*

We tested ChIPModule on three simulated datasets with five different parameter setups (Table 1). In each simulated dataset, we generated 2000 to 8000 random sequences, with the length distribution of these sequences the same as those in the E2F1 ChIP-seq dataset. We then randomly inserted TFBSs of 20 groups of TFs, using known TF PWMs in the TRANSFAC database.[24] The number of TFs in a group varied from 2 to 13, the largest number of TFs in a TF group from a previous study.[5] For each group of TFs, we inserted their TFBSs in only 10% randomly chosen sequences. We then applied ChIPModule to these simulated datasets with M as the 10% of the number of sequences. Recall that M is the minimal number of sequences (peak regions) required to contain TFBSs of each TF in a TF group. From Table 1, it is clear that ChIPModule identified as many as 17 of the inserted TF groups, which represents a sensitivity of 85% (the percent of inserted TF groups predicted). We also calculated the specificity of ChIPModule by checking how many percent of predicted TF groups are similar to the inserted TF groups. Note that we could not require the predicted TF groups are exactly as the inserted TF groups, since different TFs may bind similar motifs. A group of predicted TFs is claimed to be similar to a group of inserted TFs, if for each TF in one group, there is one TF in the other group that share a similar motif with the TF under consideration. A pair of TFs shares similar motifs if the STAMP p-value of the similarity of the two motifs is less than 1E-5, as in previous studies.[33-35]

We also noticed that several inserted TF groups were not identified. We hypothesized that these TF groups were missed by ChIPModule because not all TFBSs of the TFs in these TF groups satisfied the required putative TFBS cutoff used in Section 2.3, or TFBSs of different TFs may overlap and some of them were thus discarded. If this hypothesis was true, ChIPModule could correctly identify even more inserted TF groups if one used a smaller M. We thus further tested two of the smaller datasets using a smaller M. From the last two rows in Table 1, it is clear that using a smaller M indeed improved the accuracy of ChIPModule. For instance, ChIPModule successfully predicted all inserted TF groups when we used M as 7.5% of the number of sequences. This demonstrates that the developed tool, ChIPModule, can systematically identify TFs and their cofactors in ChIP-seq datasets.

Table 1. Correctly Predicted TF groups by ChIPModule on simulated datasets.

#total sequences	#sequences with inserted TFBSs of a group of TFs	M	#Correctly predicted TF groups	Sensitivity	specificity
2000	200	200	17	85%	88.8%
4000	400	400	15	75%	93.6%
8000	800	800	12	60%	95.8%
2000	200	150	20	100%	79.9%
4000	400	350	16	80%	74.2%

3.2. *ChIPModule Identified TFs and Their Cofactors in Experimental Data*

We further tested ChIPModule on the two ChIP-seq datasets mentioned above. These two datasets were used because the two TFs, ESR1 and E2F1, are well studied. In addition, several cofactors are known for each TF. Similar to the simulated studies, we used *M* as 10% of the number of ChIP-seq peak regions we obtained for the two TFs, respectively, when applying ChIPModule to the two ChIP-seq datasets.

In total, we identified 1334 and 6428 groups of interacting TFs in the ESR1 dataset and the E2F1 dataset, respectively. The number of TFs in these interacting TF groups is from 2 to 5 for the ESR1 dataset (average 2.16), and from 2 to 7 for the E2F1 dataset (average 4.8). To see whether ChIPModule predicted these interacting TF groups by chance, we permuted the input sequences from the ChIP-seq peak regions in each dataset and applied ChIPModule to these random sequences generated by permutation for each TF. We found that ChIPModule predicted 0 and 87 interacting TF groups in the two permuted random datasets, respectively. The much lower number of predicted interacting TF groups demonstrates that ChIPModule has a low false positive prediction rate (87/6428=1.35%), which confirms a high specificity of ChIPModule and implies the functionality of the majority of the predicted interacting TF groups.

Table 2. Several identified cofactors and their literature support.

Dataset	Known cofactors	Supported new cofactors
ESR1	FOXA[36], OCT1[36], C/EBP[36], AP-1[36]	p300[37], VDR[38]
E2F1	SP1[39], MYC[40]	NF-kappaB,[41] YY1[42]

We next checked whether ChIPModule identified known cofactors of the two TFs. For the TF ESR1, we found a few known cofactors, such as FOXA, OCT1, C/EBP and AP-1 (Table 2).[36] For the TF E2F1, we also found several known cofactors, such as SP1[37] and MYC[38] (Table 2). The de

novo discovery of the known cofactors of the two TFs supports the fact that ChIPModule can identify cofactors of TFs from ChIP-seq data.

Besides known cofactors, ChIPModule also identified new cofactors for the TF ESR1 (Table 2). ChIPModule predicted that P300 and VDR are also cofactors of ESR1, which are supported by literature.[37,38] For instance, ChIPModule identified a group of interacting TFs composed of three TFs, ESR1, VDR, and COUPTF. The TF VDR was reported to interact with ESR1.[38] It is also known that ESR1 is regulated by COUPTF, through both direct DNA binding competition and protein-protein interactions.[43] Moreover, it is suggested that COUPTF plays a master role in regulating the transactivation by VDR.[44] Based on these studies,[38,43,44] it is highly likely that TFs in the this predicted interacting TF group interact with each other, which supports the functionality of this group of TFs. We further investigated the function of the target genes of this group of TFs by the gene ontology (GO) enrichment analysis. The GO enrichment analysis is a common approach to test whether a group of gene significantly share functions based on their annotated GO terms.[45] We found that the target genes of this group of TFs significantly share a function, in utero embryonic development (GO:0001701, corrected p-value= 9.66E-05).[45] The sharing of functions by target genes suggests that these target genes are likely co-regulated, which further supports the functionality of this predicted interacting TF group.

ChIPModule identified new cofactors for the TF E2F1 as well (Table 2). Several of the predicted cofactors of E2F1 are supported by literature, such as NF-kappaB and YY1.[41,42] For instance, ChIPModule predicted a group of interacting TFs consisting of four TFs. These four TFs are YY1, E2F1, SP1, and BSAP. The TFs YY1 and SP1 are reported to be interacted with E2F1.[39,42] It is also known that YY1 interacts with the TFs SP1 and BSAP.[46,47] These studies suggest that the other three TFs in this group interact with E2F1 directly or indirectly, which supports the functionality of this group of interacting TFs. The GO enrichment analysis shown that the target genes of this group of TFs significantly share a function, positive regulation of transcription factor activity (GO: 0051091, corrected p-value=2.6E-4). Thus, the four TFs in this group of interacting TFs likely coordinately regulate their common target genes.

3.3. *A Large Number of Predicted Interacting TF Groups do not Contain the TFs Used for the ChIP-seq Experiments*

In the above analysis, we found that a large percentage of predicted interacting TF groups do not contain the TFs used for the ChIP-seq experiments. For instance, in the E2F1 ChIP-seq dataset, 4782 out of the 6248 predicted interacting TF groups do not contain the TF E2F1. We hypothesized that the exclusion of the corresponding TFs in our predictions is most likely due to the indirect binding of the corresponding TFs to the ChIP-seq peak regions through the interaction with cofactors. In other word, there are at least two types of ChIP-seq peak regions, one bound by the corresponding TF directly, the other bound by the cofactor of the corresponding TF that interact with the cofactors. To support this hypothesis, we examined the predicted interacting TF

groups and found that this is the case for several interacting TF groups. We provided two of such supporting examples below.

Example 1. An interacting TF group composed of the TFs GATA1 and SP1 was found in the ESR1 dataset. A previous study has shown that GATA1 interacts with SP1 to regulate their target genes.[48] In addition, we found that the target genes of SP1 shared the function, synaptic vesicle (GO:000802, corrected p-value=9.0E-3). Meanwhile, the target genes of GATA1 shared a similar function, synaptic transmission (GO:0007268, corrected p-value=3.0E-2). Consistently, the target genes of this interacting TF group significantly shared the function, synaptic vesicle (GO:000802, corrected p-value =4.7E-3). The interaction of the two TFs and the consistency of the function of individual TFs and the TF group suggest that this group of interacting TFs is likely functional. In addition, the TF ESR1 was reported to interact with SP1 in breast cancer cells.[49] It is thus likely that ESR1 interacts with this group of interacting TFs, which directly bind the ChIP-seq peak regions.

Example 2. The interacting TF group with two TFs ETS and SP1 was identified from the E2F1 dataset. Although E2F1 was not included in this group, the two TFs in this group were found to interact with E2F1.[39,50] A previous study has shown that E2F1 specifically interacts with ETS-related TFs.[50] The TF SP1 has also been found to interact with E2F1.[39] Moreover, the ETS TF family cooperates with SP1 to activate the human Tenascin-C promoter.[51] In addition, the target genes of this TF group significantly shared a function, RNA splicing (GO:0008380, corrected p-value 5.29E-09). Therefore, E2F1 likely interacts with this group of TFs, which directly bind the ChIP-seq peak regions. These pieces of evidence support the above hypothesis that the corresponding TF indirectly bind the ChIP-seq regions through the interaction with its cofactors.

3.4. *Comparisons with Other Methods*

We attempted to compare ChIPModule with coMOTIF[18] and W-ChIPMotifs[17]. coMOTIF jointly considers two motifs in ChIP-seq peak regions, and W-ChIPMotifs is a web application tool for de novo motif discovery from ChIP-based high throughput data. Under default parameters, coMOTIF took more than a week to run on the ESR1 dataset (3257 peaks, each 595 base pair long on average). We could not make it work on the E2F1 dataset, which may be due to the much larger data size of this dataset (10196 peaks, each 878 base pair long on average). As to W-ChIPMotifs, we were unable to obtain a local version of this tool and the online version of this tool cannot accept more than 3000 sequences. On the contrary, ChIPModule took about 533 seconds on the ESR1 dataset and 1129 seconds to run on the E2F1 dataset on a desktop computer (Intel core 2 Duo CPU, 2.93 GHz, 4G RAM), which make it suitable for gene transcriptional regulation studies based on ChIP-seq experiments. We provide both the command line mode of the ChIPModule that can be run on the DOS, Linux, and OS environments and the GUI mode of the Windows version ChIPModule. Detailed information about ChIPModule is in the readme file on the download package at http://www.cs.ucf.edu/~xiaoman/ChIPModule/ChIPModule.html.

Because it is difficult to run W-ChIPMotifs and coMOTIF on the original datasets, we chose to compare ChIPModule with the two software tools on the top 100 peak regions of the ESR1 and E2F1 datasets. In the top 100 peak regions of ESR1, for the known cofactors FOXA, OCT1,C/EBP,AP-1,p300, and VDR mentioned above, ChIPModule identified two known co-factors VDR and p300, W-ChIPMotifs identified C/EBP, and coMOTIF did not identify any of the above co-factors. In the top 100 peak regions of E2F1, for the known aforementioned co-factors sp1, myc, NF-kappaB, and YY1, ChIPModule identified all four cofactors, W-ChIPMotifs identified sp1, and coMOTIF identified the TF combination E2F1 and sp1.

We also compared ChIPModule with the two tools on simulated data. We inserted TFBSs of 50 groups of TFs into 10 out of 100 random sequences. There are 43 TFs contained in the 50 groups. W-ChIPMotifs identified motifs of 10 out of 43 TFs. coMOTIF correctly predicted two TFs in 11 out of 50 inserted TF groups. ChIPModule discovered 45 out of 50 inserted TF groups. In addition, motifs of 39 out of the 43 inserted TFs have been included in these predictions.

4. Discussion

We developed a novel method, ChIPModule, to systematically discover TFs and their cofactors from ChIP-seq data. Tested on simulated datasets, ChIPModule identified the majority of all planted interacting TF groups. Applied to experimental datasets, ChIPModule identified known cofactors and predicted new cofactors, which were supported by literature. ChIPModule thus provides a useful method to study gene transcriptional regulation.

A main assumption in the ChIPModule is that multiple TFs instead of individual TFs regulate their target genes under a given condition. This assumption is supported by the GO enrichment analysis[45] of the target genes of the predicted interacting TF groups and those of individual TFs. We found that target genes of 119 out of 150 top groups of interacting TFs (79.33%) have smaller GO enrichment p-value than those of individual TFs in the same groups for the ESR1 dataset. Meanwhile, target genes of 149 out of 150 top groups of interacting TFs (99.9%) have smaller GO enrichment p-value than those of individual TFs in these groups for the E2F1 dataset. Moreover, the target genes of a group of interacting TFs often share functions while target genes of individual TFs may not share any function. For instance, for the interacting TF group composed of the TFs PAX4 and SP1 in the ESR1 dataset, we could find that its target genes significantly share the function, negative regulation of follicle-stimulating hormone secretion (GO:0046882, corrected p-value=8.18E-005). However, the target genes of PAX4 or SP1 share no similar function.

In the above study, we found that the predicted interacting TF groups often do not contain the corresponding TFs used for the ChIP-seq experiments. We provided concrete examples to support the hypothesis that the corresponding TFs could interact with their cofactors, while the cofactors directly bind the ChIP-seq peak regions. Note that alternative explanation exists. For instance, if we lower the p-value cutoffs used to define putative TFBSs in Section 2.3, or choose a smaller M in Section 2.4, we could find more predicted interacting TF groups containing the corresponding TFs. However, our experience with the two ChIP-seq datasets and other ChIP-seq datasets[5,52]

suggests that the proposed hypothesis is likely the main reason for the exclusion of the corresponding TFs in the predicted interacting TF groups.

Several options in our developed software make ChIPModule a widely applicable tool for studying gene transcription regulation. First, besides using the TF PWMs in public databases,[23,24] users can use self-defined TF PWMs. Second, users can choose different p-value cutoffs to define putative TFBSs in ChIPModule. This is necessary, as one wants to use more stringent p-value cutoffs for large datasets while use looser p-value cutoffs for small datasets. Third, the discovered TFs and their cofactors by ChIPModule are organized in four different formats, which help users to study these interacting TF groups at different scales. We believe ChIPModule will be a useful tool for future gene transcriptional regulation studies.

5. Acknowledgement

This work was supported by a National Science Foundation, Chemical, Bioengineering, Environmental, and Transport Systems grant 1125679 (HH), a National Science Foundation, Faculty Career Development grant 1149955 (HH), and a National Science Foundation, Information & Intelligent Systems grant 1218275 (XL).

Reference

1. M. I. Arnone and E. H. Davidson, *Development (Cambridge, England)* **124** (10), 1851 (1997).
2. C. H. Yuh, H. Bolouri, and E. H. Davidson, *Science (New York, N.Y* **279** (5358), 1896 (1998).
3. L. Li, Q. Zhu, X. He et al., *Genome biology* **8** (6), R101 (2007).
4. M. Blanchette, A. R. Bataille, X. Chen et al., *Genome research* **16** (5), 656 (2006).
5. X. Cai, L. Hou, N. Su et al., *BMC genomics* **11**, 567 (2010).
6. D. S. Johnson, A. Mortazavi, R. M. Myers et al., *Science (New York, N.Y* **316** (5830), 1497 (2007).
7. G. Robertson, M. Hirst, M. Bainbridge et al., *Nature methods* **4** (8), 651 (2007).
8. J. Shendure, G. J. Porreca, N. B. Reppas et al., *Science* **309** (5741), 1728 (2005).
9. M. Margulies, M. Egholm, W. E. Altman et al., *Nature* **437** (7057), 376 (2005).
10. H. Li and N. Homer, *Brief Bioinform* (2010).
11. B. Langmead, C. Trapnell, M. Pop et al., *Genome Biol* **10** (3), R25 (2009).
12. H. Ji, H. Jiang, W. Ma et al., *Nature biotechnology* **26** (11), 1293 (2008).
13. Y. Zhang, T. Liu, C. A. Meyer et al., *Genome biology* **9** (9), R137 (2008).
14. A. Valouev, D. S. Johnson, A. Sundquist et al., *Nature methods* **5** (9), 829 (2008).
15. M. Hu, J. Yu, J. M. Taylor et al., *Nucleic acids research* **38** (7), 2154 (2010).
16. E. Mercier, A. Droit, L. Li et al., *PloS one* **6** (2), e16432 (2011).
17. V. X. Jin, J. Apostolos, N. S. Nagisetty et al., *Bioinformatics (Oxford, England)* **25** (23), 3191 (2009).
18. M. Xu, C. R. Weinberg, D. M. Umbach et al., *Bioinformatics (Oxford, England)* **27** (19), 2625 (2011).
19. M. Thomas-Chollier, E. Darbo, C. Herrmann et al., *Nature protocols* **7** (8), 1551 (2012).
20. I. V. Kulakovskiy, V. A. Boeva, A. V. Favorov et al., *Bioinformatics (Oxford, England)* **26** (20), 2622 (2010).

21. S. J. van Heeringen and G. J. Veenstra, *Bioinformatics (Oxford, England)* **27** (2), 270 (2011).
22. L. Li, *J Comput Biol* **16** (2), 317 (2009).
23. A. Sandelin, W. Alkema, P. Engstrom et al., *Nucleic acids research* **32** (Database issue), D91 (2004).
24. E. Wingender, P. Dietze, H. Karas et al., *Nucleic acids research* **24** (1), 238 (1996).
25. G. Grahne and J. Zhu, *IEEE transactions on knowledge and data engineering* **17**, 1347 (2005).
26. J. Han, J. Pei, and Y. Yin, in *ACM SIGMOD International Conference on Management of Data* (Dallas, USA, 2000).
27. D. Aldous, *Probability Approximations via the Poisson Clumping Heuristic*. (Springer-Verlag, 1989).
28. R. Edgar, M. Domrachev, and A. E. Lash, *Nucleic acids research* **30** (1), 207 (2002).
29. W. J. Kent, C. W. Sugnet, T. S. Furey et al., *Genome research* **12** (6), 996 (2002).
30. J. M. Claverie and S. Audic, *Comput Appl Biosci* **12** (5), 431 (1996).
31. J. Hu, H. Hu, and X. Li, *Nucleic acids research* **36** (13), 4488 (2008).
32. R. Sharan, I. Ovcharenko, A. Ben-Hur et al., *Bioinformatics (Oxford, England)* **19 Suppl 1**, i283 (2003).
33. F. Fauteux, M. Blanchette, and M. V. Stromvik, *Bioinformatics (Oxford, England)* **24** (20), 2303 (2008).
34. B. D. Reed, A. E. Charos, A. M. Szekely et al., *PLoS genetics* **4** (7), e1000133 (2008).
35. J. Ding, X. Li, and H. Hu, *Plant physiology* (2012).
36. J. S. Carroll, C. A. Meyer, J. Song et al., *Nature genetics* **38** (11), 1289 (2006).
37. B. D. Jeffy, J. K. Hockings, M. Q. Kemp et al., *Neoplasia (New York, N.Y* **7** (9), 873 (2005).
38. E. M. Colin, A. G. Uitterlinden, J. B. Meurs et al., *The Journal of clinical endocrinology and metabolism* **88** (8), 3777 (2003).
39. S. Y. Lin, A. R. Black, D. Kostic et al., *Molecular and cellular biology* **16** (4), 1668 (1996).
40. S. W. Hiebert, M. Lipp, and J. R. Nevins, *Proceedings of the National Academy of Sciences of the United States of America* **86** (10), 3594 (1989).
41. X. Palomer, D. Alvarez-Guardia, M. M. Davidson et al., *PloS one* **6** (5), e19724 (2011).
42. S. Schlisio, T. Halperin, M. Vidal et al., *The EMBO journal* **21** (21), 5775 (2002).
43. C. M. Klinge, B. F. Silver, M. D. Driscoll et al., *The Journal of biological chemistry* **272** (50), 31465 (1997).
44. A. J. Cooney, X. Leng, S. Y. Tsai et al., *The Journal of biological chemistry* **268** (6), 4152 (1993).
45. E. I. Boyle, S. Weng, J. Gollub et al., *Bioinformatics (Oxford, England)* **20** (18), 3710 (2004).
46. K. Calame and M. Atchison, *Genes & development* **21** (10), 1145 (2007).
47. E. Seto, B. Lewis, and T. Shenk, *Nature* **365** (6445), 462 (1993).
48. K. D. Fischer, A. Haese, and J. Nowock, *The Journal of biological chemistry* **268** (32), 23915 (1993).
49. K. Kim, R. Barhoumi, R. Burghardt et al., *Molecular endocrinology (Baltimore, Md* **19** (4), 843 (2005).
50. L. Hauck, R. G. Kaba, M. Lipp et al., *Molecular and cellular biology* **22** (7), 2147 (2002).
51. F. Shirasaki, H. A. Makhluf, C. LeRoy et al., *Oncogene* **18** (54), 7755 (1999).
52. Y. Wang, X. Li, and H. Hu, *Genomics* **98** (6), 445 (2011).