Calibrating Thresholds to Improve the Detection Accuracy of Putative Transcription Factor Binding Sites

Young-Jin Kim[1,3†], Gil-Mi Ryu[1,2†], Chan Park[1], Kyu-Won Kim[2], Bermseok Oh, Young-Youl Kim[1*] and Man Bok Gu[3*]

[1]Center for Genome Science, National Institute of Health, KCDC, Seoul 122-701, Korea, [2]College of Pharmacy, Seoul National University, Seoul 157-742, Korea, [3]School of Life Science & Biotechnology, Korea University, Seoul 136-701, Korea

Abstract

To understand the mechanism of transcriptional regulation, it is essential to detect promoters and regulatory elements. Various kinds of methods have been introduced to improve the prediction accuracy of regulatory elements. Since there are few experimentally validated regulatory elements, previous studies have used criteria based solely on the level of scores over background sequences. However, selecting the detection criteria for different prediction methods is not feasible. Here, we studied the calibration of thresholds to improve regulatory element prediction. We predicted a regulatory element using MATCH, which is a powerful tool for transcription factor binding site (TFBS) detection. To increase the prediction accuracy, we used a regulatory potential (RP) score measuring the similarity of patterns in alignments to those in known regulatory regions. Next, we calibrated the thresholds to find relevant scores, increasing the true positives while decreasing possible false positives. By applying various thresholds, we compared predicted regulatory elements with validated regulatory elements from the Open Regulatory Annotation (ORegAnno) database. The predicted regulators by the selected threshold were validated through enrichment analysis of muscle-specific gene sets from the Tissue-Specific Transcripts and Genes (T-STAG) database. We found 14 known muscle-specific regulators with a less than a 5% false discovery rate (FDR) in a single TFBS analysis, as well as known transcription factor combinations in our combinatorial TFBS analysis.

Contact: inthistime@ngri.go.kr, gm_ryu@ngri.go.kr

Keywords: combinatorial TFBS, regulatory conservation score, regulatory element, transcription factor binding site

Introduction

The discovery of transcription factor binding sites (TFBSs) in promoters is important for understanding transcriptional regulation mechanisms. Over the past few years, numerous tools have been developed for the prediction of TFBSs. However, there has been little information about the selection of thresholds among various prediction tools. A threshold is assumed to be the cutoff for prediction of precise TFBSs.

Recently, Tompa et al. reported a study of 13 previously well-known prediction tools to provide biologists with guidance in their choice of these tools (Tompa, Li, Bailey, Church, De Moor, Eskin, Favorov, Frith, Fu, Kent, Makeev, Mironov, Noble, Pavesi, Pesole, Regnier, Simonis, Sinha, Thijs, van Helden, Vandenbogaert, Weng, Workman, Ye and Zhu 2005). From the results of an analysis of a set of regulatory regions of putatively coregulated genes, almost 80% false positives existed in the prediction results.

We had a various result applying different thresholds and parameters of the prediction tools.In this case, biologists should determine the reliable parameters and validation process of putative regulatory regions by a number of experiments, which is often time-consuming work. Since nothing is assumed, a priori information used to predict regulatory elements in this analysis could significantly improve the detection accuracy with various biological aspects. Here, we report a calibration study to find the relatively relevant detection threshold using incorporated biological information that is summarized in binding profiles (position weight matrix, or PWM), sequence conservation information, and experimentally validated regulatory element data.

Usually,TF-binding sequence information is summarized in a specific data format to facilitate the analysis of possible TFBSs. A PWM is composed of a set of experimentally defined TF-binding sequences and reflects the binding specificity of TFs. TRANSFAC (Wingender, Chen, Fricke, Geffers, Hehl, Liebich, Krull, Matys, Michael, Ohnhauser, Pruss, Schacherer, Thiele and Urbach 2001; Matys, Fricke, Geffers, Gossling, Haubrock, Hehl, Hornischer,

[†]These authors contributed equally to this work.

*Corresponding author: E-mail youngyk@nih.go.kr
Tel +82-2-380-2245, Fax +82-2-354-1063
mbgu@korea.ac.kr
Tel +82-2-3290-3417, Fax +82-2-928-6050

Accepted 1 October 2007

Karas, Kel, Kel-Margoulis, Kloos, Land, Lewicki-Potapov, Michael, Munch, Reuter, Rotert, Saxel, Scheer, Thiele and Wingender 2003) and JASPAR (Vlieghe, Sandelin, De Bleser, Vleminckx, Wasserman, van Roy and Lenhard 2006) are two well-recognized databases that store eukaryotic TFs, descriptions of their respective binding affinities, and PWMs. Here, we used PWM information from TRANSFAC version 10.2.

One of the best strategies for finding functional sequences is to look for sequences that are conserved across species (Margulies and Green 2003; Woolfe, Goodson, Goode, Snell, McEwen, Vavouri, Smith, North, Callaway, Kelly, Walter, Abnizova, Gilks, Edwards, Cooke and Elgar 2005). There has been significant progress in computational approaches to analyze interspecies genomic sequence alignments and to distinguish regulatory regions from neutrally evolving DNA.

Recently, the phastCons score (Siepel, Bejerano, Pedersen, Hinrichs, Hou, Rosenbloom, Clawson, Spieth, Hillier, Richards, Weinstock, Wilson, Gibbs, Kent, Miller and Haussler 2005) and regulatory potential (RP) score (King, Taylor, Elnitski, Chiaromonte, Miller and Hardison 2005) were introduced to calculate the conservation score for eliminating false putative TFBSs and refine the candidates of highly conserved elements. The phastCons score results from multiple alignments using the phastCons program based on a phylogenetic hidden Markov model (phylo-HMM). RP measurement is a computational score to aid in the identification of putative regulatory sites in the human genome. Unlike tools based on searching for known TFBSs, phastCons and RP scores simulate a comparative genomics method. These scores are computed from genome-wide alignments of the human genome with the genomes of other organisms. This study applied RP scores to putative TFBS information produced by the MATCH program in TRANSFAC.

Moreover, gene expression in higher organisms is regulated by a combinatorial interaction of multiple TFs. A combinatorial element is a functional unit used to identify colocalizing TFBSs in a genome. Multiple TFBSs are clustered together within specific promoter regions. These functionally related binding sites are in a narrow distance range from about 25 to 200 bp (Wasserman and Fickett 1998; Berman, Nibu, Pfeiffer, Tomancak, Celniker, Levine, Rubin and Eisen 2002; Kreiman 2004). Several algorithms have been developed for detecting putative clusters of binding sites and analyzing the combinatorial interaction of multiple TFs (Alkema, Johansson, Lagergren and Wasserman 2004; Kreiman 2004).We searched known combinations of muscle-specific regulatory elements to validate whether the result of our combinatorial TFBS prediction can be used to predict possible combinatorial regulatory units.

Methods

We performed several analysis procedures for enhancing the accuracy of the prediction. In Fig. 1, our analysis scheme is briefly described in a flowchart.

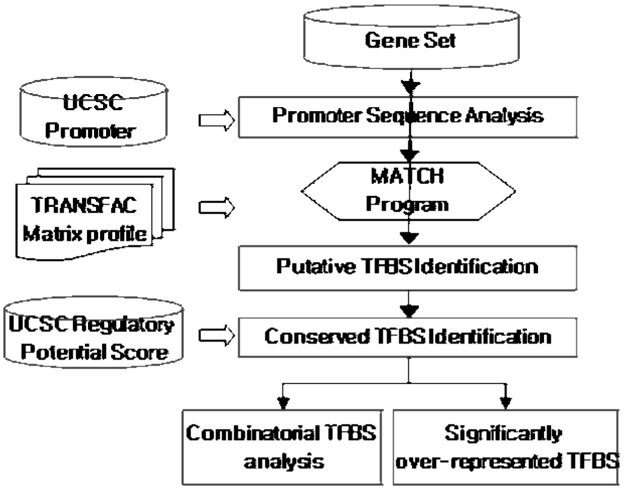
Gene Set
UCSC Promoter
Promoter Sequence Analysis
TRANSFAC Matrix profile
MATCH Program
Putative TFBS Identification
UCSC Regulatory Potential Score
Conserved TFBS Identification
Combinatorial TFBS analysis
Significantly over-represented TFBS

Fig. 1. Analysis flowchart. For prediction of reliable TFBSs, we performed sequential procedures including promoter sequence analysis, putative TFBS identification, conserved TFBS identification, and over-representation of TFBSs.

Data sources and transcription factor binding site prediction

We collected promoter sequences in the 2-kb upstream region of annotated transcription start sites of Refseq genes from the UCSC Genome Browser (Hinrichs, Karolchik, Baertsch, Barber, Bejerano, Clawson, Diekhans, Furey, Harte, Hsu, Hillman-Jackson, Kuhn, Pedersen, Pohl, Raney, Rosenbloom, Siepel, Smith, Sugnet, Sultan-Qurraie, Thomas, Trumbower, Weber, Weirauch, Zweig, Haussler and Kent 2006). Using upstream sequences, we predicted TFBSs using the MATCH program in TRANSFAC professional version 10.2, which has the largest collections of TFBS profiles (Kim, S.B., Ryu, G.M., Kim, Y.J. et al. 2007).

We predicted TFBSs using 584 vertebrate matrices from TRANSFAC. We attempted to select the best TFBS cutoff, adding 0.05 to the value of matrix similarity starting from 0.7 to 0.9 and to the core similarity value ranging from 0.75 to 0.95. The matrix similarity is a score that describes the quality of a match between a matrix and an arbitrary part of the input sequences. Analogously, the core similarity value denotes the quality of a match between the core sequence of a matrix and a part of the input sequence.

Experimentally confirmed TFBSs were retrieved from ORegAnno (an open access database and system for literature-derived promoters, TFBSs, and regulatory variation) for a true positive reference set (Montgomery, Griffith, Sleumer, Bergman, Bilenky, Pleasance, Prychyna, Zhang and Jones 2006). We used known binding sequences and transcription factor binding information of human genes from the ORegAnno database. All predicted binding sequences were produced using position-weight matrices composing the known positive binding site of human genes. Sequences predicted by the MATCH program were compared with the true set from ORegAnno using ClustalW (Fukami-Kobayashi and Saito 2002) for multiple sequence alignments. To detect a matched result from the comparison between ORegAnno real data and the predicted data, we selected all results that were complete or showed a partial overlap between the datasets in their alignments.

As a result, we selected a matrix similarity of 0.85 and a core similarity of 0.90 to find the real set and minimize false positives. To provide information regarding TFs and TFBSs by selected cutoff, we constructed information databases for genes, TFs, and TFBSs. These databases offered detailed gene and TF information using the selected cutoff. The TFBS information comprised the TFBS name, position, strand, matrix matching score, TFBS sequence, and TF information.

Selection of conserved regulatory elements

To detect conserved regulatory elements, we used the RP score. RP scores were computed from alignments of human (*hg17),* chimpanzee (*panTro1*), mouse (*mm5*), rat (*rn3*), and dog (*canFam1*) genome sequences. RP scores are values obtained by comparing the frequencies of short alignment patterns between known regulatory elements and neutral DNA (King, Taylor, Elnitski, Chiaromonte, Miller and Hardison 2005).

Because these data were the result of the alignment of five entire genomes, the constructed database of RP data had an enormous size of 8.1 gigabytes. To set an appropriate threshold that minimized false positives and maximized true positives, we validated the data using the ORegAnno real dataset by changing the RP threshold. Through the comparison of the true and predicted sets, we measured true positives and possible false positives by changing the RP scores ranging from 0 to 0.1. According to the threshold experiment described above, TFBSs that satisfied the RP cutoff condition among predicted TFBSs were detected as conserved regulatory elements.

Muscle-specific dataset

We used muscle-specific gene lists for significantly overrepresented TFBSs and the combination module. The T-STAG database is a resource for tissue-specific transcripts and genes (Gupta, Vingron and Haas 2005). We extracted 207 muscle-specific genes from the T-STAG database with default parameters. Next, we annotated the genes with the DAVID database (Dennis, Sherman, Hosack, Yang, Gao, Lane and Lempicki 2003), and 207 genes with the Unigene ID were converted and filtered as 91 Refseq mRNAs. We used 91 muscle-specific Refseq mRNAs for further analysis.

Enrichment analysis of muscle-specific gene sets

The enrichment of significant TFBSs and combinatorial TFBS analysis were statistically tested by calculation of *p* values based on Fisher's exact test. Fisher's exact test was used to determine the probability of a nonrandom association between the gene set and significant TFBSs and between the gene set and the TFBS combination of interest. To compute the *p* value for the test, the test compares the proportion of genes containing a particular cis-regulatory module to the proportion of genes having the background regulatory element set derived from whole genome information. One-sided Fisher exact probability was determined using the R statistics package (http://www.r-project.org).

Combinatorial TFBS prediction

Transcriptional regulation in eukaryotes is induced by multiple factors, in contrast with prokaryotes. In higher organisms, more complex signaling machinery plays an important role in the interactions of multiple TFs. For analyzing transcriptional regulation in higher organisms, a combinatorial approach is required for the sophisticated interactions of multiple factors (Halfon, Grad, Church and Michelson 2002; Bluthgen, Kielbasa and Herzel 2005; Kel, Konovalova, Waleev, Cheremushkin, Kel-Margoulis and Wingender 2006). All possible combinations of TFBSs in the entire genome were calculated as TFBS pairs. The distance between motifs in the module sets was less than 100 bp. To calculate possible combinations of TFBSs, we performed a complete search of TFBS clusters in a 100-bp sliding window. In the combination analysis, the distance between TFBSs was considered. Precalculated combinatorial information was built into the database to reduce computing time in the analysis of gene sets.

Improvement of statistical significance

In a previous study, researchers suggested the q value to avoid the immense number of false positives in genome-wide analysis (Storey and Tibshirani 2003). The q value is similar to the well-known *p* value. While the *p* value handles the false positive rate, q value is used to measure the false discovery rate (FDR). With a q value of 5%, one can obtain the significant features with a 5% FDR. This approach is implemented in the R statistics package known as q value. We used the q value package to calculate the significance of overrepresented TFBSs in terms of FDR.

Results and Discussion

Data source for analysis

We used the ORegAnno database, an open database for the curation of known regulatory elements from scientific literature, for choosing the cutoff of predicted TFBSs by TRANSFAC and RP score. Of the 251 TFBSs in this ORegAnno dataset, 210 records were found with known specific binding factor information. Of these, 145 can be found in the alignments of 2 kb upstream from the transcription start site. Finally, 95 TFBSs can also be found in the evolutionary conserved regions, which can be inferred using homologous alignments information of human, mouse, rat, chimpanzee, and dog sequences. Therefore, we used the 95 TFBSs that suited our conditions (Fig. 2).

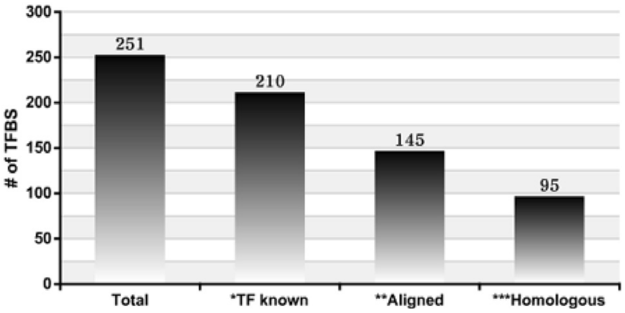

Fig. 2. ORegAnno database contents. Human TFBS data were used for the analysis. A total of 251 data points were filtered through three steps. First, we filtered records with known TF information. Second, we extracted records aligned within 2000 bp upstream from TSS. Finally, we selected gene records with homologous information within 2000 bp upstream. From the last filtering step, 95 records were produced for further analysis.

Threshold dependency for TFBS prediction

For TFBS prediction, we used the TRANSFAC MATCH program, which uses a PWM. For human promoter prediction, we used a vertebrate matrix, and we selected the cutoff for minimizing false positives and predicting the exact ORegAnno true set, by applying 0.05 increment to the cutoff ranging from matrix similarity of 0.7 to 0.9 of matrix similarity score and from 0.75 to 0.95 of core similarity score in MATCH. Because little true negative information is available, here we studied the differences between true positives and false negatives. Using those cutoffs, we selected a cutoff of 0.85 and of core similarity of 0.90, having true positives of 80% and minimizing putative false positives (Fig. 3).

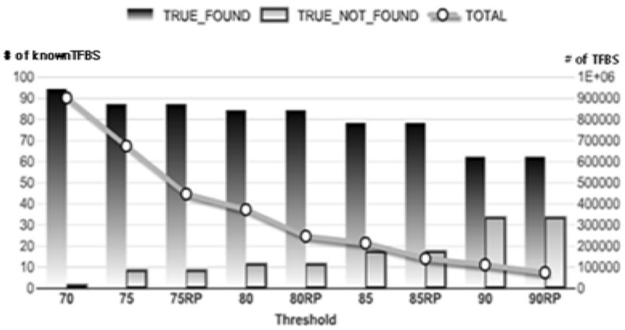

Fig. 3. The performance test with varied thresholds. We performed an accuracy test with 53 genes in 95 records for each threshold by varying 70% to 90% matching thresholds and an RP threshold of zero at the all-match threshold. Gradient bar represents the number of detected true positives with respect to the left y-axis. Outlined gradient bar stands for the undetected numbers that are false negatives. Solid line draws the total number of TFBSs of 53 genes with respect to the right y-axis. While the match threshold increases, accuracy decreases. The total number of TFBSs of 53 genes significantly decreased according to the increased thresholds. However, there was no difference in the performance before and after the conserved sequence information was applied, and only the total number of TFBSs decreased. This explains why the RP measurement is very effective in reducing the putative false positives, while the true positives are retained.

Regulatory potential score dependency for conserved regulatory element prediction

To decrease the false positive rate for predicted TFBSs, we chose a method for detection of conserved regulatory elements. This method calibrates the homology between the known regulatory region and the input promoter sequence of genes and shows the degree of similarity of conserved regulatory motifs by RP score.

We tried a performance test using the ORegAnno human 95 TFBS data with various thresholds. While the match threshold is increasing, the accuracy for motif discovery grows because of the deduction of the false positive prediction rate by high thresholds. As a result of validation using real set data, true positive values slowly change, and false positive values sharply decrease when applying 0.85 matrix similarity and 0.90 core similarity to

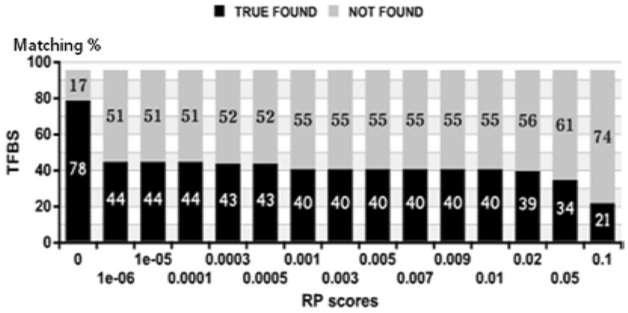

Fig. 4. Calibration test with varied conservation scores. We performed an accuracy test for the data with a threshold of 85% with various conservation scores ranging from 0 to 0.1. The point at which the RP score is zero showed the highest performance, while on the other hand the rest showed poorer performance. At that point, the average number of TFBSs is relatively minimized while accuracy remained above 80%, and 78 were found to be true and 17 were not found.

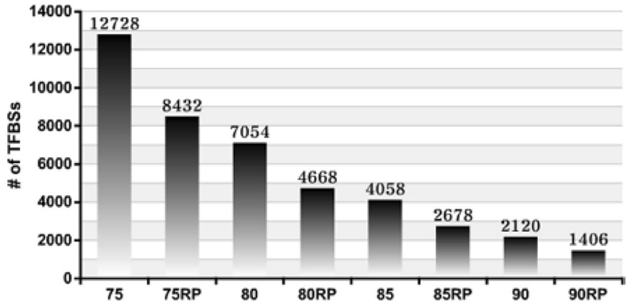

Fig. 5. The average number of TFBSs. We calculated the average number of TFBSs from 53 genes by applying various thresholds. We chose the point of MATCH threshold as 0.85 and an RP threshold of zero for further analysis.

MATCH cutoff and RP scores.

Therefore, we applied the RP score to our system (Fig. 3). We also conducted a test, increasing the regulatory score cutoff from 0 to 0.1. As a result, 78 true positives were detected, and there were 17 undetected TFBSs (Fig. 4). Detection accuracy increased to 80% when applying an RP score of 0. Therefore, the prediction method for conserved regulatory elements by applying RP score is an efficient method to minimize false positives.

We predicted an average of 2678 TFBSs in a promoter (Fig. 5). This is about 1.3 regulatory elements per base in the promoter region. Despite the high accuracy obtained with experimentally validated evidence, it is possible that a lot of noise exists in the processed data. These putative false positives may lead to confusing results and false discoveries. However, regardless of possibly false information, one can find the target TFBS of coexpressed genes using statistical methods. This statistical approach can tolerate up to 50% of overrepresented TFBSs of coexpressed genes (Ho Sui, Mortimer, Arenillas, Brumm, Walsh, Kennedy and Wasserman 2005). We applied a similar method to determine whether our data could produce robust detection of overrepresented TFBS.

Significantly overexpressed TFBSs in muscle tissue

We used muscle-specific genes derived from the T-STAG database to analyze tissue- or tumor-specific expression patterns in human and mouse transcriptomes. The statistical method was used to analyze 91 records. We applied Fisher's exact test and q value to the submitted gene set. Through statistical analysis, we obtained overrepresented TFBSs (Table 1).

There are several well-known muscle-specific TFs. MyoD and mef-2 are the most prominent TFs related to muscle-specific gene expression (Fickett 1996). Additionally, SRF, TEF-1, and SP-1 were frequently mentioned as muscle-specific regulators in the previous studies (Wasserman and Fickett 1998).

Table 1 shows a similar result to previous studies; 16 TFBSs were selected as overrepresented in muscle tissue with a q value of less than 0.05. We found that 14 of the 16 TFBSs were known as muscle-specific regulators in previous studies. Well-known MyoD binding sites were ranked as high as second and fourth. The MYF binding profile, E2A_Q6, was also found at fifth place. MEF-2, SP-1, and TEF-1 were found at the 27th, 49th, and 55th places, with a *p* value less than 0.05 and relatively high q value ranging from 0.11 to 0.34 (data not shown). However, the SRF binding profile turned out to be insignificant in this

Table 1. Significantly overrepresented TFBSs in muscle tissue. Binding sites are selected with a less than a 5% false discovery rate. Of 16 overrepresented TFBSs, 14 are known muscle-specific regulators.

TFBS	*p* value	q value
LRF_Q2	5.32E−11	2.11E−08
MYOD_Q6	1.26E−07	1.23E−05
E12_Q6	1.27E−07	1.23E−05
MYOD_Q6_01	1.44E−07	1.23E−05
E2A_Q6	1.55E−07	1.23E−05
HEB_Q6	6.76E−07	4.46E−05
AP4_Q6	9.42E−07	5.04E−05
AP4_Q6_01	1.02E−06	5.04E−05
LBP1_Q6	2.89E−06	0.000127
MZF1_01	4.72E−06	0.000187
E47_01	3.56E−05	0.001283
VDR_Q3	4.95E−05	0.001632
LMO2COM_01	0.000191	0.005821
AP4_Q5	0.000466	0.013175
E2A_Q2	0.000519	0.013709
AP4_01	0.001449	0.035856

analysis because of the high *p* value.

E12 and E2A binding profiles are representative of E12 and E47 target sequences. The splice variants of the E2A gene product, E12 and E47, are the binding partners of MyoD, which is a well-known muscle-specific regulator (Lingbeck, Trausch-Azar, Ciechanover and Schwartz 2005). There have been reports that the biochemical interaction of pRb with a tal-1-E2A-Lmo2-Ldb1 tetramer complex is found in human adult proerythroblasts and erythroblasts (Vitelli, Condorelli, Lulli, Hoang, Luchetti, Croce and Peschle 2000). The Lmo2 target sequence profile LMO2COM_01 is ranked 13th with q value 0.0058. Also, HEB has a functional role of regulating gene expression in the development of skeletal muscle (Conway, Pin, Kiernan and Merrifield 2004).

In a previous study, AP4 was reported to have specific function in human myocardial tissue (Westhoff, Jankowski, Schmidt, Luo, Giebing, Schluter, Tepel, Zidek and van der Giet 2003). Moreover, a polymorphism of VDR affects muscle function (Pfeifer, Begerow and Minne 2002), and Pbx-Meis1/Prep1 binds DNA with heterodimers of E2A and MyoD, myogenin, and Mrf-4 or Myf-5 (Knoepfler, Bergstrom, Uetsuki, Dac-Korytko, Sun, Wright, Tapscott and Kamps 1999). Despite the possible false discoveries in our TFBS prediction, our calibrated data showed good performance, finding 14 of 16 overrepresented TFBSs with a less than 5% FDR.

Detecting TFBS combinations

In higher organisms, multiple TFs are involved in transcriptional regulation. In the analysis above, we found significantly overrepresented TFBSs in muscle tissue. Using a combinatorial approach, we tested our data for possible TFBS combinations. We performed a x^2 test instead of Fisher's exact test on the muscle-specific gene set, owing to the extreme values of the sample size of total TFBS combinations found in the genome. However, the x^2 test might not be very accurate if the margin is very uneven or if there is a small value in one of the cells in the contingency table. Therefore, the statistical data from this analysis may be confusing because of the possible errors in the approximation.

Table 2 lists the combination results sorted by statistical significance. There are an estimated 67,958 TFBS combinations in the muscle gene set. With a q value result less than 0.05, we found that 13,845 of 67,958 TFBS combinations are statistically significant. Because of the large amount of analysis and possible errors in estimation, we empirically selected the top 40 combinations from the results.

Table 2. Significantly overrepresented TFBS combinations in muscle tissue. Because of the huge amount of data, only the top 40 results are displayed. The known muscle−specific combination E2F−SP1 is ranked 35th.

Combination	*p* value	q value
AP2ALPHA_01 E2F_Q2	0	0
PAX2_02 TBP_Q6	3.1E−302	9.6E−298
AP2GAMMA_01 E2F_Q2	6.3E−301	1.3E−296
HNF3_Q6 TBP_Q6	7.1E−239	1.1E−234
HNF3_Q6 PAX2_02	9.5E−237	1.2E−232
FOXD3_01 PAX2_02	3.9E−236	4.1E−232
FOXD3_01 TBP_Q6	1.5E−223	1.4E−219
FOXJ2_01 PAX2_02	1.6E−220	1.3E−216
HFH3_01 TBP_Q6	2.5E−211	1.8E−207
HFH3_01 PAX2_02	2.8E−209	1.7E−205
HNF3_Q6_01 TBP_Q6	3.9E−208	2.2E−204
FOXJ2_01 TBP_Q6	1.9E−206	9.9E−203
FOXD3_01 HNF3_Q6	5.5E−183	2.6E−179
HNF3_Q6 HNF3_Q6_01	3.3E−182	1.5E−178
HNF3_Q6_01 PAX2_02	6.7E−179	2.7E−175
FAC1_01 PAX2_02	4.2E−178	1.6E−174
FAC1_01 TBP_Q6	3.9E−176	1.4E−172
HFH3_01 HNF3_Q6	1.3E−175	4.3E−172
HNF3ALPHA_Q6 TBP_Q6	8.2E−175	2.7E−171
LRF_Q2 LRF_Q2	1.4E−170	4.4E−167
FOXD3_01 HNF3_Q6_01	2.8E−168	8.1E−165
FOXJ2_01 HNF3_Q6	1.9E−160	5.5E−157
FOXD3_01 HFH3_01	1.2E−156	3.3E−153
HFH3_01 HNF3_Q6_01	3.5E−155	9E−152
E2F_Q2 E2F_Q2	4.9E−155	1.2E−151
HNF3ALPHA_Q6 PAX2_02	1.1E−149	2.6E−146
FOXJ2_01 HNF3_Q6_01	1.7E−147	3.8E−144
E2F_Q2 ETF_Q6	1.5E−143	3.3E−140
TBP_Q6 TBP_Q6	2E−143	4.2E−140
FOXD3_01 HNF3ALPHA_Q6	3.4E−143	7E−140
HNF3ALPHA_Q6 HNF3_Q6	7.1E−143	1.4E−139
FOXD3_01 FOXJ2_01	1.8E−140	3.4E−137
FAC1_01 FOXD3_01	1.3E−139	2.4E−136
FOXJ2_01 HFH3_01	1.5E−138	2.7E−135
E2F_Q2 SP1_01	2.7E−137	4.8E−134
FAC1_01 HNF3_Q6	4.3E−134	7.4E−131
FAC1_01 HNF3_Q6_01	8.6E−131	1.4E−127
LBP1_Q6 LRF_Q2	3.1E−130	5E−127
HNF3ALPHA_Q6 HNF3_Q6_01	2.3E−127	3.7E−124
HFH3_01 HNF3ALPHA_Q6	6.8E−127	1E−123

E2F and SP1 are reported to have direct interaction in muscle tissue (Guo, Degnin, Fiddler, Stauffer and Thayer 2003). We found that the predicted E2F and SP1 combination ranked 35th. E2F and GATA6 are reported as important regulators of glomerular mesangial cells (Morrisey 2000). The E2F and GATA6 combination ranked 59th (data not shown). Known transcription factor interactions also followed. E2A and MyoD in 183rd, E12 and LMO2 in 330th, and HEB and MyoD ranked 772nd. Other tissue-specific combinations have been found to have a high rank. The AP-2 and E2F combination is ranked in first and third place. It is reported that the E2F-AP-2 complex may cooperate with c-Myc and α-tocopherol in neuronal cells (Dottori, Gross, Labosky and Goulding 2001). FOXJ2 and PAX2 have an interaction in the eye (Yu, Lin, Zack and Qian 2006), and is ranked 8th.

Web implementation

We partially applied this method for the analysis of TFBS information; the detection of overrepresented TFBSs has been implemented in a user-friendly website (http://www.ngri.go.kr/cmams/CREAT/index.html). Because of the lack of computational power of our web server, we excluded the combinatorial module. Reference ID, gene ID, and gene symbol are allowed as inputs for module analysis. Users may input genes into a text box as part of the analysis program. Through the website, we offer conserved regulatory motif information. The program is implemented in Perl script (http://www.perl.com) and R language (http://www.bioconductor.org), and uses MySQL (http://www.mysql.com) as a database. The program is wrapped by Perl script to maintain a user-friendly web interface.

Conclusions and Discussion

We used the TRANSFAC MATCH program to predict human TFBSs and used the RP score to identify conserved regulatory elements. For the MATCH program and RP score cutoff decision, we selected cutoff values whose true positive values were not changed and whose false positive values decreased sharply. We developed the cREAT system by applying those cutoffs.

We also validated the running result of the cREAT system using muscle-specific genes of the T-STAG database. As a result, 14 of 16 were predicted TFs by cREAT, previously reported as muscle-specific TFs in the literature.

We also analyzed combinatorial TFBSs commonly affecting gene expression. The combinatorial approach of multiple TFs has been addressed in a considerable number of previous studies. Enriched regulatory module analysis of a set of potentially coregulated genes, combination analysis of evolutionarily conserved regulatory elements, and robust cis-regulatory module analysis using biological information have been introduced and applied to identify regulatory networks and pathways (Cohen, Klingenhoff, Boucherot, Nitsche, Henger, Brunner, Schmid, Merkle, Saleem, Koller, Werner, Grone, Nelson and Kretzler 2006). We were able to find known combinatorial units in the muscle-specific gene sets. The enrichment analysis implied that data filtered by the calibrating thresholds are robust and that true positives are detected while the number of possible false positives is reduced.

Compared with the previously introduced study, this analysis offers researchers relevant threshold and filtered putative TFBS information. First, we adopted a regulatory potential score, which is a specialized measurement of potentially conserved regulatory elements. Second, we calibrated the thresholds in the analysis of TFBS detection and conserved elements with a varied range of thresholds. Finally, we found the relevant thresholds to provide robust detection. We compared the proportion of the difference between the true dataset and the predicted dataset. Next, we validated our result with well-known muscle-specific gene sets from T-STAG. This approach will be very helpful to biologists confronted with finding putative TFBSs and filtering reliable TFBSs with specific thresholds.

Acknowledgments

The authors thank Chang-Bum Hong, Sung-Hoon Lee, and Jeong-Ho Cha for computer system support and helpful comments. This study was supported by the intramural fund of the National Institute of Health, Korea.

References

Alkema, W.B., Johansson, O., Lagergren, J., and Wasserman, W.W. (2004). MSCAN: identification of functional clusters of transcription factor binding sites. *Nucleic Acids Res.* 32(Web Server issue), W195-8.

Berman, B.P., Nibu, Y., Pfeiffer, B.D., Tomancak, P., Celniker, S.E., Levine, M., Rubin, G. M., and Eisen, M.B. (2002). Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc Natl Acad Sci USA* 99(2), 757-62.

Bluthgen, N., Kielbasa, S.M., and Herzel, H. (2005). Inferring combinatorial regulation of transcription in silico. *Nucleic Acids Res.* 33(1), 272-9.

Cohen, C. D., Klingenhoff, A., Boucherot, A., Nitsche, A., Henger, A., Brunner, B., Schmid, H., Merkle, M., Saleem, M.A., Koller, K.P., Werner, T., Grone, H.J., Nelson, P.J., and Kretzler, M. (2006). Comparative promoter analysis allows de novo identification of specialized cell junction-associated proteins. *Proc Natl Acad Sci USA* 103(15), 5682-7.

Conway, K., Pin, C., Kiernan, J.A., and Merrifield, P. (2004). The E protein HEB is preferentially expressed in developing muscle. *Differentiation* 72(7), 327-40.

Dennis, G. Jr., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C., and Lempicki, R.A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* 4(5), P3.

Dottori, M., Gross, M.K., Labosky, P., and Goulding, M. (2001). The winged-helix transcription factor Foxd3 suppresses interneuron differentiation and promotes neural crest cell fate. *Development.* 128(21), 4127-38.

Fickett, J.W. (1996). Quantitative discrimination of MEF2 sites. *Mol Cell Biol.* 16(1), 437-41.

Fukami-Kobayashi, K., and Saito, N. (2002). How to make good use of CLUSTALW. *Tanpakushitsu Kakusan Koso* 47(9), 1237-9.

Guo, C.S., Degnin, C., Fiddler, T.A., Stauffer, D., and Thayer, M.J. (2003). Regulation of MyoD activity and muscle cell differentiation by MDM2, pRb, and Sp1. *J Biol Chem* 278(25), 22615-22.

Gupta, S., Vingron, M., and Haas, S.A. (2005). T-STAG: resource and web-interface for tissue-specific transcripts and genes. *Nucleic Acids Res.* 33(Web Server issue), W654-8.

Halfon, M.S., Grad, Y., Church, G.M., and Michelson, A.M., (2002). Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model. *Genome Res.* 12(7), 1019-28.

Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., Hillman-Jackson, J., Kuhn, R.M., Pedersen, J.S., Pohl, A., Raney, B.J., Rosenbloom, K.R., Siepel, A., Smith, K.E., Sugnet, C.W., Sultan-Qurraie, A., Thomas, D.J., Trumbower, H., Weber, R.J., Weirauch, M., Zweig, A.S., Haussler, D., and Kent, W.J. (2006). The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* 34(Database issue), D590-8.

Ho Sui, S.J., Mortimer, J.R., Arenillas, D.J., Brumm, J., Walsh, C.J., Kennedy, B.P., and Wasserman, W.W. (2005). oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. *Nucleic Acids Res.* 33(10), 3154-64.

Kel, A., Konovalova, T., Waleev, T., Cheremushkin, E., Kel-Margoulis, O., and Wingender, E. (2006). Composite Module Analyst: a fitness-based tool for identification of transcription factor binding site combinations. *Bioinformatics*. 22(10), 1190-7.

King, D.C., Taylor, J., Elnitski, L., Chiaromonte, F., Miller, W., and Hardison, R.C. (2005). Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome Res*. 15(8), 1051-60.

Knoepfler, P.S., Bergstrom, D.A., Uetsuki, T., Dac-Korytko, I., Sun, Y.H., Wright, W.E., Tapscott, S.J., and Kamps, M.P. (1999). A conserved motif N-terminal to the DNA-binding domains of myogenic bHLH transcription factors mediates cooperative DNA binding with pbx-Meis1/Prep1. *Nucleic Acids Res*. 27(18), 3752-61.

Kreiman, G. (2004). Identification of sparsely distributed clusters of cis-regulatory elements in sets of co-expressed genes. *Nucleic Acids Res.* 32(9), 2889-900.

Lingbeck, J.M., Trausch-Azar, J.S., Ciechanover, A., and Schwartz, A.L. (2005). E12 and E47 modulate cellular localization and proteasome-mediated degradation of MyoD and Id1. *Oncogene* 24(42), 6376-84.

Margulies, E.H., and Green, E.D. (2003). Detecting highly conserved regions of the human genome by multispecies sequence comparisons. *Cold Spring Harb Symp Quant Biol*. 68, 255-63.

Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., Kloos, D.U., Land, S., Lewicki-Potapov, B., Michael, H., Munch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S., and Wingender, E. (2003). TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* 31(1), 374-8.

Montgomery, S.B., Griffith, O.L., Sleumer, M.C., Bergman, C.M., Bilenky, M., Pleasance, E.D., Prychyna, Y., Zhang, X., and Jones, S.J. (2006). ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation. *Bioinformatics* 22(5), 637-40.

Morrisey, E. E. (2000). GATA-6: the proliferation stops here: cell proliferation in glomerular mesangial and vascular smooth muscle cells. *Circ Res.* 87(8), 638-40.

Pfeifer, M., Begerow, B., and Minne, H.W. (2002). Vitamin D and muscle function. *Osteoporos Int*. 13(3), 187-94.

Kim, S.B., Ryu, G.M., Kim, Y.J., Heo, J.Y., Park, C., Oh, B.S., Kim, H.L., Kimm, K.C., Kim,K.W., and Kim, Y.Y. (2007). FCAnalyzer: A Functional Clustering Analysis Tool for Predicted Transcription Regulatory Elements and Gene Ontology Terms. *Genomics & Informatics* 5(1), 10-18.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., Weinstock, G.M., Wilson, R.K., Gibbs, R.A., Kent, W.J., Miller, W., and Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*. 15(8), 1034-50.

Storey, J. D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* 100(16), 9440-5.

Tompa, M., Li, N., Bailey, T.L., Church, G.M., De Moor, B., Eskin, E., Favorov, A.V., Frith, M.C., Fu, Y., Kent, W.J., Makeev, V.J., Mironov, A.A., Noble, W.S., Pavesi, G., Pesole, G., Regnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C., and Zhu, Z. (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol.* 23(1), 137-44.

Vitelli, L., Condorelli, G., Lulli, V., Hoang, T., Luchetti, L., Croce, C.M., and Peschle, C. (2000). A pentamer transcrip tional complex including tal-1 and retinoblastoma protein downmodulates c-kit expression in normal erythroblasts. *Mol Cell Biol*. 20(14), 5330-42.

Vlieghe, D., Sandelin, A., De Bleser, P.J., Vleminckx, K., Wasserman, W.W., van Roy, F., and Lenhard, B. (2006). A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res*. 34(Database issue), D95-7.

Wasserman, W.W. and Fickett, J.W. (1998). Identification of regulatory regions which confer muscle-specific gene expression. *J Mol Biol*. 278(1), 167-81.

Westhoff, T., Jankowski, J., Schmidt, S., Luo, J., Giebing, G., Schluter, H., Tepel, M., Zidek, W., and van der Giet, M. (2003). Identification and characterization of adenosine 5'-tetraphosphate in human myocardial tissue. *J Biol Chem*. 278(20), 17735-40.

Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Krull, M., Matys, V., Michael, H., Ohnhauser, R., Pruss, M., Schacherer, F., Thiele, S., and Urbach, S. (2001). The TRANSFAC system on gene expression regulation. *Nucleic Acids Res*. 29(1), 281-3.

Woolfe, A., Goodson, M., Goode, D.K., Snell, P., McEwen, G.K., Vavouri, T., Smith, S.F., North, P., Callaway, H., Kelly, K., Walter, K., Abnizova, I., Gilks, W., Edwards, Y.J., Cooke, J.E., and Elgar, G. (2005). Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* 3(1), e7.

Yu, X., Lin, J., Zack, D.J., and Qian, J. (2006). Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues. *Nucleic Acids Res.* 34(17), 4925-36.