IDENTIFICATION OF OVER-REPRESENTED COMBINATIONS OF TRANSCRIPTION FACTOR BINDING SITES IN SETS OF CO-EXPRESSED GENES

SHAO-SHAN HUANG[1,2,*], DEBRA L. FULTON[1,2,*], DAVID J. ARENILLAS[1,2,3], PAUL PERCO[4], SHANNAN J. HO SUI[1,2], JAMES R. MORTIMER[5] AND WYETH W. WASSERMAN[1,2,3,#]

[1]*Centre for Molecular Medicine and Therapeutics,*
[2]*Child and Family Research Institute,*
[3]*Department of Medical Genetics,*
University of British Columbia, Vancouver, Canada
[4]*Department of Nephrology, Medical University of Vienna, Vienna, Austria*
[5]*Merck Frosst Centre for Therapeutic Research, Kirkland QC, Canada*

[*]*These authors contributed equally to this work.*
[#]*Corresponding author. E-mail: wyeth@cmmt.ubc.ca*

Transcription regulation is mediated by combinatorial interactions between diverse trans-acting proteins and arrays of cis-regulatory sequences. Revealing this complex interplay between transcription factors and binding sites remains a fundamental problem for understanding the flow of genetic information. The oPOSSUM analysis system facilitates the interpretation of gene expression data through the analysis of transcription factor binding sites shared by sets of co-expressed genes. The system is based on cross-species sequence comparisons for phylogenetic footprinting and motif models for binding site prediction. We introduce a new set of analysis algorithms for the study of the combinatorial properties of transcription factor binding sites shared by sets of co-expressed genes. The new methods circumvent computational challenges through an applied focus on families of transcription factors with similar binding properties. The algorithm accurately identifies combinations of binding sites over-represented in reference collections and clarifies the results obtained by existing methods for the study of isolated binding sites.

1. Introduction

The interaction between transcription factor (TF) proteins and transcription factor binding sites (TFBS) is an important mechanism in regulating gene expression. Each cell in the human body expresses genes in response to its developmental state (e.g., tissue type), external signals from neighboring cells and environmental stimuli (stress/nutrients). Diverse regulatory mechanisms have evolved to facilitate the programming of gene expression, with a primary mechanism for expression modulation targeting the rate of transcript initiation through TF interactions. Given a finite collection of protein structures capable of binding to specific DNA sequences and the diversity of conditions to which cells must respond, it is logical and well-documented that combinatorial interplay between TFs drives much of the observed specificity of gene expression. The arrays of TFBS at which the interactions occur are often termed cis-regulatory modules (CRM)[1].

The sequence specificity of TFs has stimulated development of computational methods for discovery of TFBS on DNA sequences. Well established methods represent aligned collections of binding sites as position weight matrices (PWM). Sequence specificity of individual PWM profiles can be quantified by information content, and scoring a sequence against the PWM of a TF gives a quantitative measure of the sequence's similarity to the binding profile (for review see Wasserman and Sandelin[16]). Searching for high scoring motifs in putative regulatory sequences with a collection of profiles (for instance, JASPAR[9]) can suggest the binding sites in the sequence and the associated TF. However, this methodology is plagued by poor specificity due to the short and variable nature of the binding sites. Phylogenetic footprinting filters have been demonstrated repeatedly to improve specificity[5]. Such filters are justified by the hypothesis that sequences of biological importance are under higher selective pressure and will thus accumulate DNA sequence changes at a slower rate than other sequences. Based on this expectation, the search for potential binding sites can be limited to the most similar non-coding regions of aligned orthologous gene sequences from species of suitable evolutionary distance. A powerful extension incorporates knowledge from experimental data: co-expressed genes are possibly under the control of the same TFs, so over-represented TFBS in the co-expressed genes are likely to be functional. These concepts are implemented by Ho Sui *et al.* in the web service tool oPOSSUM[13], which, when given a set of co-expressed genes, can identify the TFBS motifs that are over-represented with respect to a background set of genes.

The oPOSSUM service has achieved considerable success in finding binding sites known to contribute to the regulation of reference gene sets, but it fails to address the known interplay between TFs at CRMs. Methods for the analysis of over-represented combinations of motifs exist[2,4,11]. We introduce a new approach rooted in the biochemical properties of transcription factors, which allows greater computational efficiency and improved interpretation of results. The resulting method is assessed against diverse reference data to demonstrate its utility for the applied analysis of gene expression data. Supplementary information is available at http://www.cisreg.ca/oPOSSUM2/supplement/.

2. Methods

2.1. *Background: the oPOSSUM database*

Ho Sui *et al.*[13] describe the creation of the oPOSSUM database which stores predicted, evolutionarily conserved TFBS to support over-representation analysis of single TFBS. Here is a brief summary of the process: first, human-mouse orthologs are retrieved from Ensembl. The JASPAR TFBS profiles are used to identify putative TFBS within the conserved non-coding regions from 5000 base pairs (bp) upstream to 5000 bp downstream of the annotated transcription start site (TSS) on both strands. A minimum matrix match score of 65% is required for a position to be reported as a putative binding site. The oPOSSUM database stores the start and end positions and the score of each site. oPOSSUM II uses these data in searching for over-represented binding site combinations (see algorithm described below).

2.2. *Overview and rationale of oPOSSUM II algorithm*

Finding over-represented combinations of TFBS presents several new issues that are not encountered in single site analysis. We address two of the main challenges: computational complexity and TFBS class redundancy.

First, the number of possible combinations of size n from m TFBS ($n \leq m$) increases combinatorially with respect to both m and n, which greatly impacts computing time. Secondly, some TFs in the databases have similar binding properties, thus subsets of profiles may be effectively redundant. Consequently, an exhaustive search is not an efficient approach to find over-represented combinations of patterns. Instead, we should seek to reduce profile redundancy.

To address both problems, we use a novel method to group the profiles into classes. Rather than using protein sequence similarity, a hierarchical clustering procedure is applied to group the profiles into classes according to their quantitative similarity. One representative member is selected from each class for further analysis. We then search for the occurrences of class combinations in both co-regulated genes (foreground) and a set of background genes. To further reduce the chance co-occurrence of TFBS due to profile similarity, the binding sites within a combination are not allowed to overlap. In addition, since many co-operative TFBS are found to occur in clusters but without strict ordering constraints[1], the TFBS in any combination must satisfy a given inter-binding site distance and they are allowed to occur in any order. We also consider the possibility that multiple occurrences of one combination may be relevant. A scoring scheme is adopted from the Fisher exact test to compare the degree of over-representation of the class combinations. The highly over-represented class combinations are re-assessed using all possible profile combinations within the indicated classes.

The overall scheme of oPOSSUM II analysis is shown in Fig. 1. The sections below describe the details of each step.

2.3. *TFBS in foreground gene set*

When presented with a set of co-expressed genes S, oPOSSUM II queries the oPOSSUM database for all putative TFBS T present in S within a maximum of 5000 bp upstream and 5000 bp downstream from the TSS on each gene. The analysis may also include only the TFs from certain taxonomic subgroups (currently plant, vertebrate and insect), or TFs whose profiles exceed a minimum information content.

2.4. *Classification of TFBS profiles*

Binding profiles for T are retrieved from the JASPAR database. A profile comparison method, CompareACE [3] or matrix aligner[10], calculates the pairwise similarity scores of all the profiles using profile alignment methods. The similarity score $s(t_i, t_j)$ between profiles t_i and t_j is converted to distance $d(t_i, t_j)$ by $d(t_i, t_j) = 1 - s(t_i, t_j)$. A distance matrix M is formed from these pairwise distances. From M, an agglomerative clustering procedure creates a hierarchy of clusters (subsets) of T. The complete linkage method is used since it tends to find cohesive classes. Cutting the cluster tree at a specified height thr_H partitions T into classes.

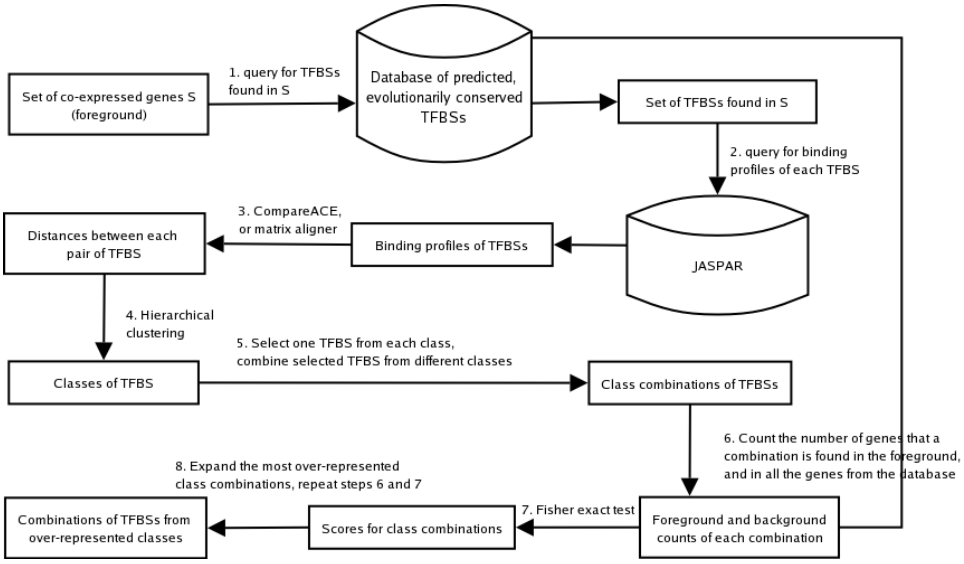

Figure 1. Overview of the oPOSSUM II analysis algorithm. Steps are numbered in the order executed. The database of predicted TFBS is identical to that of the oPOSSUM analysis system (Ho Sui *et al.*[13]).

2.5. *Selection of TFBS and enumeration of combinations*

For each class C, we select the profile that is the most similar to other profiles in C as the class representative. We chose this approach instead of using a consensus matrix, as we could not identify an adequate procedure that would generate a profile with comparable specificity to the matrices within the class. First, we calculate the sum of pairwise similarity score σ_i between a profile t_i and other profiles in C, i.e., $\sigma_i = \sum_{t_i,t_j \in C} s(t_i, t_j)$. The profile with the maximum sum of similarity score is chosen. From the selected TFBS, unordered combinations of specified cardinality are created. oPOSSUM II then searches the foreground gene set (the co-expressed genes) and the background gene set (default is all the genes in the database) for occurrences of these combinations. Let max_d be the maximum inter-binding site distance. For each gene, find the occurrences of the combinations in a sliding window of width max_d within the required search region. The result is the number of genes where a combination is found in the foreground and background gene sets.

2.6. *Scoring of combinations*

The Fisher exact test detects non-random association between two categorical variables. We adopt the Fisher P-values to rank the significance of non-random association between the occurrence of a combination and the foreground gene set, i.e., over-representation of the combination in the foreground compared to background. For each combination, a two-dimensional contingency table is constructed from the foreground and background count distributions:

	Number of genes with a given combination	Number of genes without a given combination
Foreground	a_{11}	a_{12}
Background	a_{21}	a_{22}

For $i, j = 1, 2$, row sum $R_i = a_{i1} + a_{i2}$ and column sum $C_j = a_{1j} + a_{2j}$, and the total count $N = \sum_i R_i = \sum_j C_j$. From the hypergeometric probability function, the conditional probability P_{cutoff} given the row and column sums is

$$P_{\text{cutoff}} = \frac{(C_1!C_2!)(R_1!R_2!)}{N! \prod_{i,j=1,2} a_{ij}}.$$

Calculate the P-values for all other possible contingency tables with row sums equal to R_i and column sums equal to C_j. The Fisher P-value is the sum of all the P-values less than or equal to P_{cutoff}, which are from tables representing equal or greater deviation from independence than the observed table.

Caution must be taken when interpreting these Fisher P-values. First, the foreground and background genes are allowed to overlap, which is a violation of an assumption for the statistical test. Secondly, the Fisher exact test model may not precisely characterize the data sets being analyzed. As a result, the Fisher P-values should be used purely as a measure to compare the degree of over-representation between different combinations. We will hereafter refer to the P-values as "scores". Although the scores do not describe the probabilistic nature of the over-representation, the ranking they provide was shown to be useful[13].

2.7. *Finding significant TFs from over-represented class combinations*

Let thr_C be the maximum score for a combination to be considered significant. Our empirical studies of reference collections suggest that a default value of 0.01 detects relevant TF combinations. Let x_i be any class combination with a score less than or equal to thr_C, and $X = \{x_i | score(x_i) \leq thr_C\}$. Recall that combinations in X are combinations of distinct classes. For each x_i, let $C_1, \ldots, C_h$ be the classes represented, where h is the combination cardinality. Compute the Cartesian product $\mathbb{C}_p$ of $C_1, \ldots, C_h$. This is called "expanding the classes". Repeat the enumeration and ranking procedures for the h-tuples in $\mathbb{C}_p$.

2.8. *Random sampling simulations of foreground genes*

oPOSSUM II needs to accommodate input gene sets of different cardinalities, so we wish to investigate the relationship between gene set size and the false positive rate. 100 random samples of r genes are selected from the background and given to oPOSSUM II as foreground genes. For each sample, oPOSSUM II reports the scores for all the class combinations. As these random samples of genes are not expected to be co-regulated, any combination is considered false positive. Let $(0, max_s]$ be the interval over which false positives are accumulated. We report the number of false positive class combinations for a range of max_s when $r = 20, 40, 60, 80, 100$.

Table 1. The parameter values used in validation studies with three human reference gene sets described in the Methods section. Abbreviations: bp - base pairs; TSS - transcription start site.

Parameter	Value
phylum	vertebrate
search region	5000 bp upstream, 5000 bp downstream from TSS
minimum matrix match score	75%
conservation level	1
maximum inter-binding site distance (max_d)	100 bp
profile comparison method	CompareACE
class distance threshold (thr_H)	0.45
combination cardinality	2

2.9. *Validation*

Three reference sets of human genes were used as inputs to oPOSSUM II to assess the performance of the algorithm. Two independent sets of skeletal muscle genes were tested. The first set (muscle set 1) was compiled from the reference collection identified by Wasserman and Fickett[15], as updated by a review of recent literature. A second set (muscle set 2) combines the results of microarray studies of Moran *et al.*[7] and Tomczak *et al.*[14] The third set contains smooth muscle-specific genes experimentally verified by Nelander *et al.*[8] Table 1 summarizes the parameter values for the analyses.

As a further comparison to the methods in Kreiman[4], which were validated in part against the yeast CLB2 gene cluster[12], the yeast CLB2 cluster was analyzed using the yeast oPOSSUM database (Ho Sui, unpublished).

3. Results

3.1. *TFBS classification*

Since the three reference gene sets are restricted to vertebrates, the first step in oPOSSUM II analysis is to cluster the available vertebrate TFBS. Cutting the hierarchical cluster tree at a height of 0.45 ($thr_H = 0.45$) creates groupings that correlate well with the structural families defined in JASPAR (partial cluster tree in Fig. 2, complete tree in web supplement). Most notably, binding profiles from FORKHEAD, HMG and ETS families are grouped correctly. In contrast, the zinc finger family is composed of divergent profiles, so it is appropriate for them to be dispersed into distinct classes.

The 68 vertebrate TFBS in JASPAR are divided into 32 classes. For pairs of binding sites, this step reduces the search space by a factor of four.

3.2. *Validation with reference data sets*

3.2.1. *Yeast CLB2 cluster*

The yeast CLB2 gene cluster contains genes whose transcription peaks at late G2/early M phase of the cell cycle. Transcription of these genes is regulated by the TF FKH, which is a component of the TF SFF, and which interacts with the TF MCM1. The top ten scoring class combinations found by oPOSSUM II all contain the binding sites of the ECB class, of which MCM1 is a member. The highest ranked combination is {ECB, FKH1}, which is consistent with the literature and the results of Kreiman[4]. The complete results are available

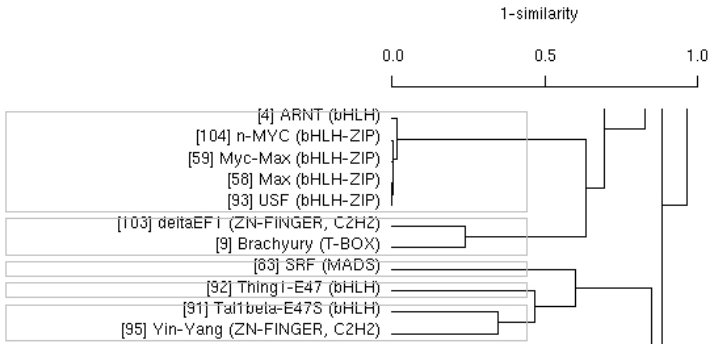

Figure 2. Partial hierarchical clustering tree of vertebrate TFBS profiles. The leaves of the tree are the TF names and the structural family of the TF is in parenthesis. The rectangular boxes are drawn at a height of 0.45.

on the supplementary web site.

3.2.2. *Three human reference gene sets*

Tables 2 to 4 list the top five over-represented class combinations for each of the three human reference gene sets. The score values for these combinations were less than 2.0E-3. Also listed are the five most over-represented TFBS classes in the total 32 classes created, as reported by oPOSSUM single site analysis. Enclosed in parenthesis is the name of TFs within the class known to mediate transcription in the assessed tissue.

Table 2. The top five over-represented pairwise pairs of TFBS classes in muscle set 1. Over-represented single TFBS classes are displayed in the second column.

TF combination (reported by oPOSSUM II)	TF (reported by oPOSSUM)
class 8 (Bsap); class 20 (MEF2)	class 1 (Myf)
class 8 (Bsap); class 29 (SRF)	class 8 (Bsap)
class 1 (Myf); class 31 (Yin-Yang)	class 29 (SRF)
class 20 (MEF2); class 28 (SP1)	class 26 (RREB-1)
class 20 (MEF2); class 29 (SRF)	class 28 (SP1)

Table 3. The top five over-represented pairs of TFBS classes in muscle set 2. Over-represented single TFBS classes are displayed in the second column.

TF combination (reported by oPOSSUM II)	TF (reported by oPOSSUM)
class 20 (MEF2); class 28 (SP1)	class 20 (MEF2)
class 20 (MEF2); class 32 (Thing1-E47)	class 25 (Androgen)
class 20 (MEF2); class 21 (MZF_5-13)	class 29 (SRF)
class 8 (Bsap); class 20 (MEF2)	class 1 (Myf)
class 1 (Myf); class 20 (MEF2)	class 7 (Spz1)

Prior studies of muscle set 1[15] show the occurrences of clusters of muscle regulatory sites including MEF2, SRF, Myf/MyoD, SP1 and TEF. MEF2 and SP1 containing classes dominate the top combinations in both muscle sets (Tables 2 and 3). Yin-Yang modulates SRF-dependent, skeletal muscle expression. Thing1-E47 is a bHLH TF localized to gut

Table 4. The top five over-represented pairs of TFBS classes in smooth muscle genes. Over-represented single TFBS classes are displayed in the second column.

TF combination (reported by oPOSSUM II)	TF (reported by oPOSSUM)
class 28 (SP1); class 29 (SRF)	class 29 (SRF)
class 21 (MZF_5-13); class 29 (SRF)	class 26 (RREB-1)
class 29 (SRF); class 31 (Yin-Yang)	class 20 (MEF2)
class 29 (SRF); class 7 (Spz1)	class 7 (Spz1)
class 29 (SRF); class 32 (Thing1-E47)	class 1 (Myf)

smooth muscle in adult mice, so the presence of class 32 in the list may be linked to other myogenic factors in the bHLH superfamily (such as Myf). Bsap and MZF are not muscle specific. The Bsap motif is long (20 bp) and exhibits an unusual pattern of low information content distributed across the entire motif, suggesting that it may behave differently than other binding profiles. The inclusion of this profile in the JASPAR database is under review (B. Lenhard, personal communication).

For the smooth muscle genes, the SRF class appears in all the top five combinations, consistent with established knowledge[6]. The top combination, {SP1, SRF}, is required for the expression of smooth muscle myosin heavy chain in rat. Yin-Yang can stimulate smooth muscle growth. Spz1 acts in spermatogenesis, and has no known role in muscle expression.

For all three reference sets, the top scoring combinations cover some different classes as compared to single site analysis. In all cases, there were relevant TFBS only identified by the combination analysis.

3.3. *Effect of set size on false positive*

The result of random sampling simulation of foreground genes is shown in Figure 3, which plots the rate of false positive predictions for a range of gene set sizes as a function of max_s. The data suggests no dependency of the false prediction rate on set size. We also note that at low score values, the proportion of false positives is low.

3.4. *Web interface*

oPOSSUM II web service is available at http://www.cisreg.ca/oPOSSUM2/opossum2.php. A user enters a set of putatively co-expressed genes and specifies parameter values of the algorithm (see example in Table 1). Since the analysis is very computationally intensive (each reference gene set takes about 10 minutes to analyze), the web server queues the request and notifies the user via e-mail when the analysis is complete.

4. Discussion

The analysis of over-represented combinations of TFBS in the promoters of co-expressed genes was motivated by biochemical and genetic studies which revealed the functional importance of cis-regulatory modules. As opposed to previously described methods for the identification of single over-represented motifs, the analysis of combinations must solve or circumvent the consequent combinatorial explosion. Such an explosion can be prohibitive in terms of computing time. To reduce the search space, oPOSSUM II restricts the search to binding site combinations that satisfy some biologically justified criteria.

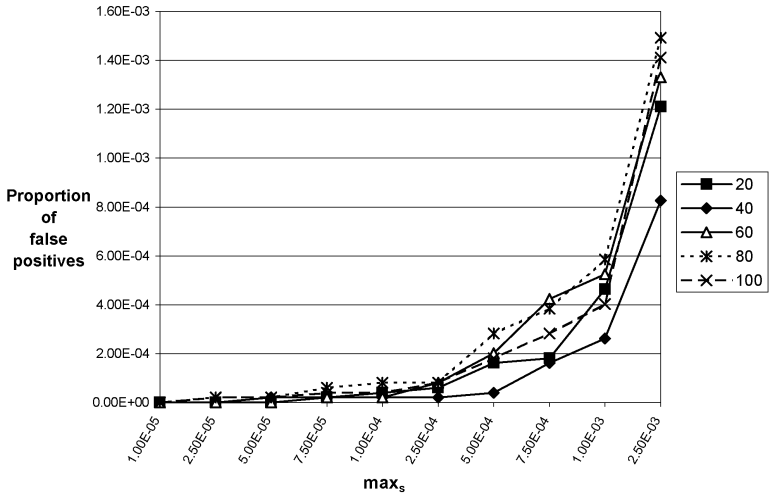

Figure 3. Effect of gene set size on false positive rate observed from pairwise TFBS combinations in randomly generated foreground gene sets.

The results indicate two valuable contributions to the interpretation of existing TFBS over-representation methods. First, in each reference gene set, there is one TF class that appears in multiple combinations, an observation that is not immediately obvious in single site analysis. Second, the algorithm finds functional TFBS that are not prominent in single site analysis. For instance with the yeast CLB2 gene cluster, members of the top scoring combination, ECB and FKH1, are ranked the first and eleventh in single site analysis. In the smooth muscle reference set, the SRF and SP1 combination is the most significant, but they are ranked the first and fourteenth in single site analysis. These demonstrate the benefits of considering combinations of binding sites.

Analysis of the microarray-based skeletal muscle reference set correctly implicates the combination of MEF2 and SP1 containing TF classes in myogenesis. This success confirms the utility of high-quality microarray data for regulatory sequence analysis.

While our result for the yeast CLB2 cluster is comparable to that reported by Kreiman[4], there are significant differences between the methods to consider. Kreiman first uses a motif discovery procedure to select motifs from the set of submitted genes, and subsequently looks for over-represented combinations of motifs using both the newly derived patterns and a TFBS profile database. In our interpretation, there is circular logic in performing motif discovery on a set and then identifying the derived over-represented pattern as being over-represented in combinations. For the CLB2 cluster, the profiles were from the existing database and our results are comparable. For the first skeletal muscle collection, Kreiman reported the top scoring combination consisted of SP1, SRF, TEF and a putative motif.

Although this paper only presents the results from pairs of TFBS, oPOSSUM II implementation is able to handle combinations of higher cardinality. However, validation of larger combinations is seriously limited by the lack of good reference data sets that are

known to be regulated by multiple binding sites.

A few issues remain to be addressed by future research. First, the interpretation of analysis results is confounded by intra-class binding similarity. While this property facilitates the oPOSSUM II algorithm, users must be prepared to consider which proteins in a family are most likely to act within the tissue or under the condition studied. For instance, the fact that an E-box motif is over-represented in the skeletal muscle data does not directly lead the researcher to the MyoD protein; instead the user must consider the entire range of bHLH-domain TFs. Second, inter-class similarity can influence the results. Although oPOSSUM II does not allow overlap between TFBS in the analysis of a given combination, TFBS from different combinations can overlap. Thus two G-rich motifs may be reported as over-represented in different combinations (for instance, the SP1 and MZF motifs from Table 3) but highlight the same candidate TFBS within the sequences analyzed. A related issue is the compositional sequence bias in tissue specific genes[17], which would motivate selection of a more refined background gene set. Finally, the computing time required is prohibitively long for a synchronous web service. Parallelization of the enumeration algorithm is a natural way to improve the running time, which may provide better response times for users.

5. Conclusion

oPOSSUM II utilizes putative TFBS identified from comparative genomic analysis, in conjunction with knowledge of co-regulated expression, to search for functional combinations of TFBS that may confer a given gene expression pattern. It uses a novel scheme to group similar binding site profiles. Based on these similarity sets, the oPOSSUM II system circumvents the combinatorial challenge of TFBS studies to identify over-represented combinations of TFBS within the promoters of co-expressed genes. Validation with reference data indicates that analysis of site combinations can give information distinct from analysis of isolated sites.

Acknowledgments

We thank Andrew Kwon for annotation of the muscle reference collections. We acknowledge operating support from the Canadian Institutes for Health Research (CIHR) and Merck Frosst; D.F. was supported by the CIHR/MSFHR Bioinformatics training program and the Merck Frosst Co-op program; WWW is supported as a Michael Smith Foundation for Health Research Scientist and a New Investigator of the CIHR.

References

1. M. I. Arnone and E. H. Davidson. The hardwiring of development: organization and function of genomic regulatory systems. *Development*, 124(10):1851–64, 1997.
2. N. Bluthgen, S. M. Kielbasa, and H. Herzel. Inferring combinatorial regulation of transcription in silico. *Nucleic Acids Res*, 33(1):272–9, 2005.
3. J. D. Hughes, P. W. Estep, S. Tavazoie, and G. M. Church. Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J Mol Biol*, 296(5):1205–14, 2000.
4. G. Kreiman. Identification of sparsely distributed clusters of cis-regulatory elements in sets of co-expressed genes. *Nucleic Acids Res*, 32(9):2889–900, 2004.

5. B. Lenhard, A. Sandelin, L. Mendoza, P. Engstrom, N. Jareborg, and W. W. Wasserman. Identification of conserved regulatory elements by comparative genome analysis. *J Biol*, 2(2):13, 2003.
6. C. S. Madsen, J. C. Hershey, M. B. Hautmann, S. L. White, and G. K. Owens. Expression of the smooth muscle myosin heavy chain gene is regulated by a negative-acting GC-rich element located between two positive-acting serum response factor-binding elements. *J Biol Chem*, 272(10):6332–40, 1997.
7. J. L. Moran, Y. Li, A. A. Hill, W. M. Mounts, and C. P. Miller. Gene expression changes during mouse skeletal myoblast differentiation revealed by transcriptional profiling. *Physiol Genomics*, 10(2):103–11, 2002.
8. S. Nelander, P. Mostad, and P. Lindahl. Prediction of cell type-specific gene modules: identification and initial characterization of a core set of smooth muscle-specific genes. *Genome Res*, 13(8):1838–54, 2003.
9. A. Sandelin, W. Alkema, P. Engstrom, W. W. Wasserman, and B. Lenhard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*, 32(Database issue):D91–4, 2004.
10. A. Sandelin, A. Hoglund, B. Lenhard, and W. W. Wasserman. Integrated analysis of yeast regulatory sequences for biologically linked clusters of genes. *Funct Integr Genomics*, 3(3):125–34, 2003.
11. R. Sharan, A. Ben-Hur, G. G. Loots, and I. Ovcharenko. CREME: Cis-Regulatory Module Explorer for the human genome. *Nucleic Acids Res*, 32(Web Server issue):W253–6, 2004.
12. P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell*, 9(12):3273–97, 1998.
13. S. J. H. Sui, J. R. Mortimer, D. J. Arenillas, J. Brumm, C. J. Walsh, B. P. Kennedy, and W. W. Wasserman. oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes. *Nucleic Acids Res*, 33(10):3154–64, 2005.
14. K. K. Tomczak, V. D. Marinescu, M. F. Ramoni, D. Sanoudou, F. Montanaro, M. Han, L. M. Kunkel, I. S. Kohane, and A. H. Beggs. Expression profiling and identification of novel genes involved in myogenic differentiation. *FASEB J*, 18(2):403–5, 2004.
15. W. W. Wasserman and J. W. Fickett. Identification of regulatory regions which confer muscle-specific gene expression. *J Mol Biol*, 278(1):167–81, 1998.
16. W. W. Wasserman and A. Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet*, 5(4):276–87, 2004.
17. R. Yamashita, Y. Suzuki, S. Sugano, and K. Nakai. Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue specificity. *Gene*, 350(2):129–36, 2005.