4278–4284 *Nucleic Acids Research, 2002, Vol. 30 No. 19*

Predicting transcription factor synergism

Sridhar Hannenhalli* and Samuel Levy

Informatics Research, Celera Genomics, 45 West Gude Drive, Rockville, MD 20850, USA

Received May 8, 2002; Revised July 1, 2002; Accepted July 31, 2002

ABSTRACT

Transcriptional regulation is mediated by a battery of transcription factor (TF) proteins, that form complexes involving protein–protein and protein–DNA interactions. Individual TFs bind to their cognate *cis*-elements or transcription factor-binding sites (TFBS). TFBS are organized on the DNA proximal to the gene in groups confined to a few hundred base pair regions. These groups are referred to as modules. Various modules work together to provide the combinatorial regulation of gene transcription in response to various developmental and environmental conditions. The sets of modules constitute a promoter model. Determining the TFs that preferentially work in concert as part of a module is an essential component of understanding transcriptional regulation. The TFs that act synergistically in such a fashion are likely to have their *cis*-elements co-localized on the genome at specific distances apart. We exploit this notion to predict TF pairs that are likely to be part of a transcriptional module on the human genome sequence. The computational method is validated statistically, using known interacting pairs extracted from the literature. There are 251 TFBS pairs up to 50 bp apart and 70 TFBS pairs up to 200 bp apart that score higher than any of the known synergistic pairs. Further investigation of 50 pairs randomly selected from each of these two sets using PubMed queries provided additional supporting evidence from the existing biological literature suggesting TF synergism for these novel pairs.

INTRODUCTION

One of the major challenges in biology is to understand the precise mechanism by which protein expression is regulated. Thus, the information that indicates the spatial, temporal and quantitative contribution of a protein to the organism development is relevant. One of the most important stages at which this regulation occurs is at the level of transcription. Transcriptional regulation is mediated by transcription factor (TF) proteins which bind to specific DNA signals or transcription factor-binding sites (TFBS), varying between 5 and 20 bp in length. The analysis of transcriptional regulation is particularly hard in higher eukaryotes due to large intergenic regions and signals with relatively low information content. Precise regulatory control is achieved by combinatorial and concerted interactions of various transcription factors with their cognate binding sites, with each other and with the transcription initiation complex. In order to enhance our understanding of gene regulation it will be necessary to derive a precise identification of the DNA signals, the proteins that bind to them, the interaction among these proteins to form complexes and the genes these complexes regulate in different tissues under different conditions.

TFBS are organized into modules on the DNA, frequently proximal to the gene, in groups confined to a few hundred base pair regions (1). Various modules interact together (where the ensemble can be referred to as a promoter model) to provide the combinatorial regulation of gene transcription in response to various developmental and environmental conditions. [See Werner (2) for an approach for identifying promoter models using comparative genomics.] Determining the TFs that preferentially work in concert as part of a module is an essential component of understanding transcriptional regulation. Although not all TFs that contribute to regulatory control as part of the same complex bind to neighboring *cis*-elements, many of these TFs that act synergistically in such a fashion are likely to have their *cis*-elements co-localize on the genome. One obvious example of a class of synergistically acting factors is the homo- and heterodimers where the corresponding *cis*-elements are 2–5 bp apart. There have been various attempts to detect pairs of co-localized *cis*-elements, referred to as composite elements, in the literature that deals with detecting patterns in groups of sequences (3,4). There has also been some work in detecting and searching for transcriptional modules (5–8).

We exploit this notion of co-localization of *cis*-elements for synergistically acting TFs to predict TF pairs that are likely to be part of a transcriptional module. The basic idea is to measure the frequency at which the *cis*-elements for two factors co-occur on the genome relative to some appropriate background. We expect that the pair of TFs acting synergistically will have a relatively higher value of this measure, thus suggesting the use of this measure as a predictor of potential synergism.

We used TRANSFAC (9) (http://transfac.gbf.de/TRANSFAC/) and TRANSCompel (10) to validate our methodology. TRANSFAC records for each TF the other TFs that are known to interact with it. Additionally, the TRANSFAC database includes the TRANSCompel database, which presents information on combinatorial gene transcriptional regulation and protein–protein interactions between different TFs bound to their cognate promoter elements. The TRANSFAC database contains curated definitions of TFBS

*To whom correspondence should be addressed. Tel: +1 240 453 3613; Fax: +1 240 453 3324; Email: sridhar.hannenhalli@celera.com

represented as 'positional weight matrices (PWMs)' that provide a quantitative description of the nucleotide base pair variations at each position of the binding sites. These PWMs have been constructed from experimental data taken from different binding studies of transcription factor proteins and their cognate DNA sequence. The PWMs are redundant at two levels. Firstly, there could be many PWMs describing the binding site of the same transcription factor and these matrices are quite similar, if not identical. At a different level, there are potentially distinct TFs that have very similar cognate *cis*-elements resulting in the existence of very similar PWMs. The TFBS described by similar PWMs tend to cluster together artificially, which potentially obscures truly co-localized distinct sites. In order to eliminate these cases we developed a similarity measure between pairs of PWMs in a manner similar to the comparison of amino acid profiles for protein domains to determine relatedness (11). Thus, we have developed a Smith–Waterman style dynamic programming based algorithm to compute a similarity score between a pair of PWMs. Based on this measure we eliminated all the PWM pairs that were similar to each other.

We found that the pairs of TFs that are known to act synergistically from experimental analysis indeed have a relatively higher value of the co-localization measure, thus validating the use of this measure as a predictor of potential synergism. A preliminary result based on the PWM hits in the upstream regions of a small set of 500 disease-related genes was published earlier (12). In the current study we extend this to annotated *cis*-elements on the entire human genome. By using PubMed (http://www.ncbi.nlm.nih.gov/), which is the largest freely available online database of biological literature, we further investigated the TF pairs with a high value of this measure but not recorded in the databases of synergistically interacting pairs. We found experimental evidence for synergism for a large fraction of these novel TF pairs.

MATERIALS AND METHODS

Identifying TFBS on the human genome sequence

PWMs representing TFBS known to occur in human were extracted from TRANSFAC version 4.4 (13) and used in our annotation of the human genome. We used only 247 PWMs based on the human and mouse sequence. Identifying the alignment of PWMs on sequences was performed with PWM_SCAN (M.Flanigan, in preparation), which is a faster implementation of the publicly available program PATSER (14).

Acceptable PWM alignments were detected with a measured probability (P) of 2×10^{-4} or lower. This results in approximately 1 100 000 hits per PWM on the human genome sequence of size ~2.8 Gb, considering both strands independently.

Transcription factor co-localization

For each pair of PWMs we compute a measure of frequency of co-localization of the two PWM hits relative to an appropriate background model. For a pair of PWMs i and j, let N_{ij} be the number of times the two PWMs co-localize within w bp in S. We randomly swap the PWM identity for all the hits to generate the background model. For example, for two randomly selected hits for PWMs i and j, we swap the hits, i.e. what was the hit for PWM i now becomes the hit for PWM j and vice versa. There are other ways of generating the background that we have examined (12). This approach preserves the total number of hits for each PWM as well as preserves the density of hits within various sequence regions. Let R_{ij} be the number of times the two PWMs co-localize within w bp in S after the aforementioned perturbation. The preferential co-localization of a pair of PWMs i and j is measured in terms of a co-localization index (CI), defined as:

$$CI = N_{ij}/R_{ij}$$

Note that CI is symmetric for a pair of factors, resulting in $(n^2 - n)/2$ values, where n is the total number of PWMs from TRANSFAC. Values for w of 50 and 200 bp were used.

Transcription factor with known synergism

Two sources of data were used to extract the PWM pairs that are known in the literature to act synergistically. In TRANSFAC the entry for transcription factors included an interaction field, which lists other factors known in the biological literature to interact with the factor in a synergistic fashion. Additionally, each PWM has a factor associated with it. We took a pair of PWMs to be interacting if the corresponding factors were recorded as interacting in TRANSFAC. This yielded 295 pairs from TRANSFAC 5.3.

The TRANSCompel database originates from the COMPEL database (10) and collects information about composite regulatory elements (CEs): pairs of closely situated sites and transcription factors binding to them. Composite elements are defined as a minimal functional unit within which both protein–DNA and protein–protein interactions contribute to a highly specific pattern of gene transcriptional regulation (15). As before, we associate a pair of PWMs as a composite pair via their corresponding factors. This yielded 156 unique pairs of PWMs.

Combining the two sources yielded 444 unique pairs of PWMs that have been identified as either acting synergistically towards transcription regulation or are known to interact in the context of transcriptional regulation.

Similarity score of PWMs

A Smith–Waterman (16) style alignment algorithm was implemented to align a pair of PWMs. A PWM can be viewed as a sequence of probability distributions where each position in the PWM has a corresponding probability distribution representing the base pair preferences at that position.

The alignment score of a position i of PWM I to the position j of PWM J is the relative entropy (RE) (17) of the two probability distributions corresponding to the two positions. For probability distributions $X = (x_A x_C x_G x_T)$ and $Y = (y_A y_C y_G y_T)$,

$$RE(X,Y) = \sum_{i \in \{A,C,G,T\}} x_i \times \ln(x_i/y_i)$$

Since RE is an asymmetric measure, we take the minimum of $RE(X,Y)$ and $RE(Y,X)$. The resulting score of a pair of positions is calibrated appropriately to make the expected score non-positive, which is a requirement for a Smith–Waterman algorithm. A stringent gap penalty was used based

on empirical distribution of *RE* values. Finally, the resulting alignment score for a PWM pair is normalized using the length of the longer PWM. As a result of our choice of various parameters the alignment score has a value between 0 and 0.4, where 0.4 is achieved for identical PWMs.

Constructing a PubMed query to verify novel predictions

For each PWM the list of protein name synonyms was extracted from TRANSFAC via corresponding transcription factors. We will describe the query to retrieve literature relating to factor pairs by illustrating an example. The synonyms for PWM M00005 (TRANSFAC identifier) are AP-4 and AP4. The synonyms for PWM M00141 are LyF-1, Lyf-1, Ikaros and lymphoid transcription factor. To gather evidence from PubMed for the interaction of the two matrices we did the following PubMed query: (''AP-4'' OR ''AP4'') AND (''LyF-1'' OR ''Lyf-1'' OR ''Ikaros'' OR ''lymphoid transcription factor'') AND (complex OR composite OR interact OR interacts OR interaction) AND (transcription OR transcriptional).

One could think of additional ways to make this query more specific and sensitive but this query does extract useful information and is sufficient for our current purposes. The resulting abstracts are manually scrutinized to gather the final evidence.

RESULTS

Correlation between *CI* and TF synergism

We wanted to determine whether synergistic or interacting TF pairs from the experimental literature had a preferentially higher *CI* value. Therefore we computed *CI* for every pair of PWMs as described in Materials and Methods by using the TFBS data from the human genomic sequence. From this set of values we extracted the *CI* values corresponding to the subset of 444 pairs with known synergism. This latter set provides us with a foreground and the whole set of *CI* values provides a background. Figure 1 shows these two distributions for $w = 50$. The *CI* distribution for the synergistic PWM pairs (the foreground) is biased towards higher values of *CI.*

This bias in the *CI* immediately suggests that this measure could be used to predict synergistic PWMs and hence TFs which could potentially be tested experimentally.

Correlation between *CI* and similarity score

Upon further investigation of PWM pairs with high *CI* not included in the TRANSFAC database, we realized that a large proportion of these PWM pairs either corresponded to the same transcription factor or to different factors which nevertheless had very similar PWMs. To test the contribution of this effect to the bias shown in Figure 1, we investigated the correlation of the PWM similarity score (computed as described in Materials and Methods) and the *CI*. Figure 2 shows the plot of these two quantities for all pairs of PWMs. The correlation coefficient of these two quantities is 0.5. It is clear from this high correlation between the two quantities that a majority of the TF pairs with high values of *CI* in fact have very similar PWMs and hence are not significant. The correlation shown in Figure 2 is bimodal, where the pairs with perfect alignment score (PWM similarity score = 0.4)

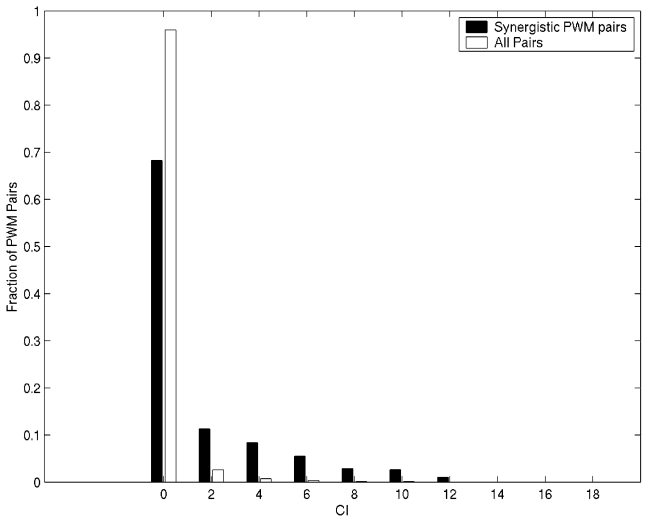

Figure 1. Distribution of *CI* for 444 synergistic pairs of PWMs versus all pairs, for $w = 50$, based on a set of 247 TRANSFAC PWM hits on the entirety of the human genome sequence at a stringency of $P \leqslant 0.0002$. The synergistic PWM pairs were extracted from the TRANSFAC and TRANSCompel databases as described in Materials and Methods.

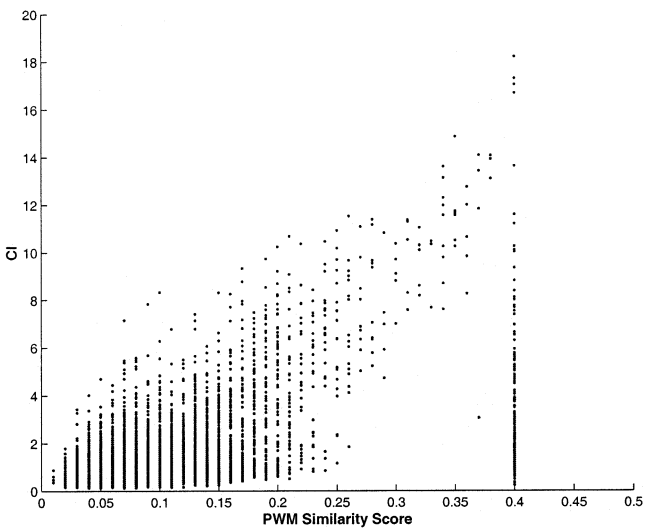

Figure 2. PWM similarity score versus *CI*. The correlation coefficient is 0.5.

form one mode. Upon further inspection it is clear that these correspond to scores of matrices aligned against themselves. The range of *CI* values for these reflect either: (i) a tendency to form homodimers, (ii) a tendency to occur in clusters or (iii) a repeat nature of the binding sites. Further investigation, although important, will not be dealt with in this work and these cases will be eliminated from further analysis. Figure 2 also suggests that this correlation between *CI* and similarity score drops off below 0.25. Upon further inspection of the similarity scores we decided upon a stringent cut-off threshold of 0.2, below which the PWMs are quite dissimilar.

Correlation between *CI* and TF synergism for non-similar PWM pairs

It is clear that the correlation of high *CI* scores with similar PWMs is a strong contributor to the bias shown in Figure 1,

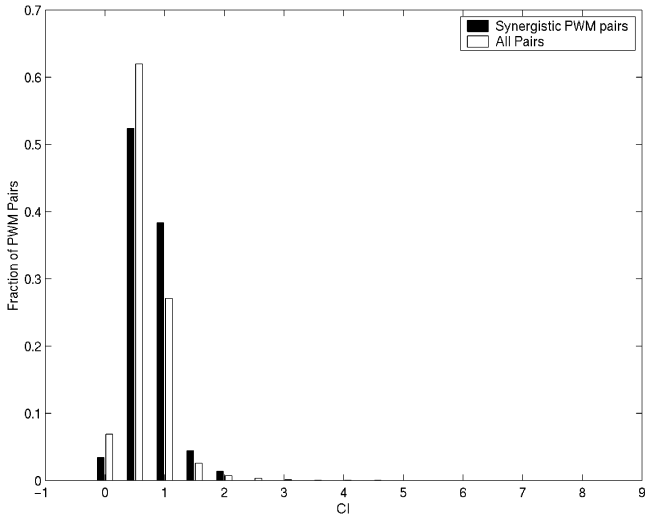

Figure 3. Distribution of *CI* for synergistic pairs of PWMs versus all pairs, similar to Figure 1, after eliminating similar PWM pairs, for $w = 50$. The maximum *CI* for synergistic pairs is 2.31. The maximum for the all pairs is 8.10 and there are 251 pairs with scores >2.31. Even though the distributions are seemingly not as discriminating, from a different perspective there are 30% of the PWM pairs with $CI > 1$ whereas there are 43% of PWM pairs with known synergism with $CI > 1$.

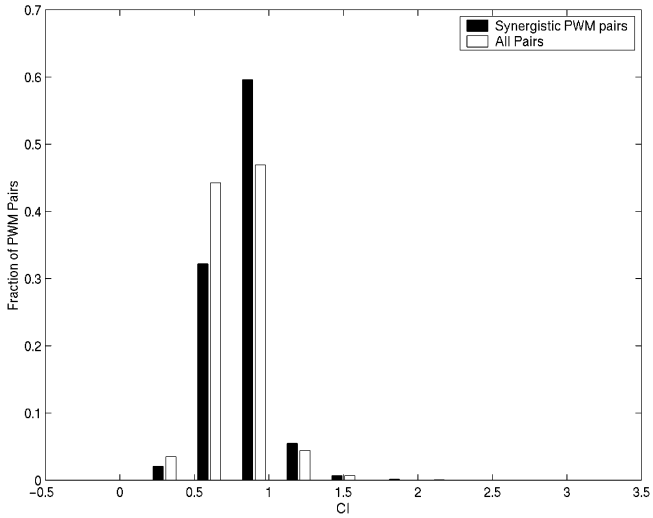

Figure 4. Distribution of *CI* for synergistic pairs of PWMs versus all pairs, similar to Figure 1, after eliminating similar PWM pairs, for $w = 200$. The maximum *CI* for synergistic pairs is 1.77. The maximum for the all pairs is 2.96 and there are 70 pairs with scores >1.77. Even though the distributions are seemingly not as discriminating, from a different perspective there are 25% of the PWM pairs with $CI > 1$ whereas there are 38% of PWM pairs with known synergism with $CI > 1$.

therefore, to remove this bias from our analysis, we made two changes: (i) we consider PWM hits on the genome within w base pairs as neighboring only if the sites identified do not overlap by >4 bp; (ii) we eliminate all pairs of PWMs from our analysis that either have the same factor name or their similarity score is at least 0.2. Both of these criteria try to address the same issue. Removing similar PWMs results in the retention of 292 PWM pairs with known synergism, corresponding to 105 unique PWMs. This was a reduction from 444 pairs corresponding to 139 unique PWMs. This reduction of ~25% is due to our stringent threshold and may not reflect what is truly redundant in TRANSFAC, which probably needs closer manual inspection.

We repeated the analysis presented in Figure 1 after making these changes. Figures 3 and 4 show the corresponding plots for $w = 50$ and $w = 200$. Even though the distributions are seemingly not as discriminating, they are in fact discriminating looking from a different perspective. For instance, for $w = 50$ there are 30% of the PWM pairs with $CI > 1$, whereas there are 43% of PWM pairs with known synergism with $CI > 1$. Similarly, for $w = 200$ there are 25% of the PWM pairs with $CI > 1$, whereas there are 38% of PWM pairs with known synergism with $CI > 1$. These facts indicate that experimentally determined synergistic TF pairs still have a preference for higher *CI* values.

Analysis of novel predictions

All the PWM pairs with high *CI* not reported in TRANSFAC are potentially novel synergisms. It is also possible that some of them are in fact known in the biological literature but have not been assigned by TRANSFAC as yet. To test the validity of our approach, we extracted at random 50 PWM pairs whose *CI* values were greater than those of the known synergistic pairs, both for $w = 50$ and $w = 200$. For $w = 50$, the maximum *CI* among the pairs with known synergism is 2.31 and there are 251 PWM pairs with $CI > 2.31$. For $w = 200$, the maximum *CI* among the pairs with known synergism is 1.77 and there are 70 PWM pairs with $CI > 1.77$. There are only seven PWM pairs in common between the two random sets of 50 pairs. For each PWM pair analyzed we queried PubMed for documents supporting synergistic interactions of the PWMs. The type of query used is described in Materials and Methods.

In 26 out of 50 cases there are PWM pairs that have experimental literature supporting synergistic interactions involved in transcriptional control using $w = 50$. Out of these 26, nine of the pairs have evidence for physical interaction, either by direct protein–protein interactions or indirectly by virtue of being in the transcriptional complex. Out of these 26 pairs, two of the pairs correspond to the same factor pair, namely CREB:SREBP-1, and three of the pairs correspond to CREB:NF-κB. If we discount these redundancies this leaves 23 pairs with evidence, of which eight pairs have evidence of physical interaction. Table 1 lists these PWM pairs and the supporting literature reference obtained from PubMed.

In 16 out of 50 cases there are PWM pairs that have experimental literature supporting synergistic interaction involved in transcriptional control using $w = 200$. Out of these 16, seven of the pairs have evidence for physical interaction, either directly or indirectly by virtue of being in the transcriptional complex. Table 2 lists these PWM pairs and the supporting literature evidence.

DISCUSSION

From a previous study on the TFBS identified in the 5 kb upstream region of 500 human disease genes (12), we have been able to show that the measure of *CI*, the tendency for TFBS to co-localize on the genome, is shifted toward higher

Table 1. PubMed evidence for synergistic transcriptional regulation for the 23 PWM pairs among the randomly chosen 50 pairs with *CI* values >2.31, for $w = 50$

ID, factor name		Reference	Complex (Y/N)
PWM1	PWM2		
M00008, Sp1	M00141, Lyf-1	(19)	N
M00008, Sp1	M00189, AP-2	(20)	N
M00008, Sp1	M00253, cap signal	(21)	N
M00033, p300	M00039, CREB	(22)	Y
M00033, p300	M00041, CRE-BP1/c-Jun	(23)	Y
M00033, p300	M00141, Lyf-1	(24)	N
M00033, p300	M00220, SREBP-1	(25)	Y
M00033, p300	M00233, MEF-2	(26)	Y
M00039, CREB	M00053, c-Rel	(27)	Y
M00039, CREB	M00054, NF-κB	(28)	Y
M00039, CREB	M00077, GATA-3	(29)	N
M00039, CREB	M00141, Lyf-1	(30,31)	N
M00039, CREB	M00220, SREBP-1	(25)	Y
M00039, CREB	M00233, MEF-2	(32)	N
M00039, CREB	M00243, Egr-1	(33)	N
M00039, CREB	M00253, cap signal	(34,35)	N
M00041, CRE-BP1/c-Jun	M00054, NF-κB	(36)	N
M00041, CRE-BP1/c-Jun	M00077, GATA-3	(37)	Y
M00041, CRE-BP1/c-Jun	M00141, Lyf-1	(38)	N
M00041, CRE-BP1/c-Jun	M00233, MEF-2	(39)	N
M00054, NF-κB	M00077, GATA-3	(40)	N
M00054, NF-κB	M00119, Max	(41)	N
M00054, NF-κB	M00121, USF	(42)	N

Eight out of the 23 pairs in fact interact physically via direct interaction or as part of same complex. A Y in the last column indicates evidence of direct or indirect physical interaction among the factors. An N indicates an absence of such evidence.

Table 2. PubMed evidence for synergistic transcriptional regulation for the 16 PWM pairs among the randomly chosen 50 pairs with *CI* values >1.77, for $w = 200$

ID, factor name		Reference	Complex (Y/N)
PWM1	PWM2		
M00039, CREB	M00139, AhR	(43)	Y
M00039, CREB	M00141, Lyf-1	(30,31)	N
M00039, CREB	M00191, ER	(44,45)	Y
M00039, CREB	M00220, SREBP-1	(25)	Y
M00039, CREB	M00233, MEF-2	(32)	Y
M00039, CREB	M00249, CHOP-C/EBPα	(46)	N
M00039, CREB	M00253, cap signal	(34,35)	N
M00041, CRE-BP1/c-Jun	M00191, ER	(47)	Y
M00087, Ikaros 2	M00141, Lyf-1	(19,48)	N
M00087, Ikaros 2	M00253, cap signal	(49)	N
M00114, Tax/CREB	M00141, Lyf-1	(50)	N
M00139, AhR	M00180, E2F	(51)	N
M00139, AhR	M00191, ER	(52,53)	Y
M00139, AhR	M00253, cap signal	(54)	N
M00155, ARP-1	M00191, ER	(53,55)	Y
M00155, ARP-1	M00253, cap signal	(56)	N

Seven out of the 16 pairs in fact interact physically via direct interaction or as part of same complex. A Y in the last column indicates evidence of direct or indirect physical interaction among the factors. An N indicates an absence of such evidence.

values. Upon closer manual scrutiny of PWM pairs with high *CI* values we observed a significant number of PWM pairs with the same factor name but distinct PWM identifiers. There was an obvious need to remove spurious, multiply identified PWM pairs since this redundancy could obscure any potential value to the *CI* scoring scheme. This led to the elimination of PWM pairs with a Smith–Waterman similarity score of at least 0.2 and otherwise to eliminate any TFBS annotated pair that overlaps by >4 bp. The second condition will address cases where a Smith–Waterman alignment does not indicate sufficient similarity, but where different PWMs identify the same region of DNA. After removing this redundancy in our TFBS identification process it was possible to find the same trend of shift towards high *CI* upon recomputation of the remaining PWM pairs. This is consistent with the notion that TFBS co-localize in clusters or modules and that they convey the necessary information for the transcriptional control of the gene under different conditions (1). For example, in the case of

the ENDO-16 gene there are as many as eight distinct structural modules, collectively containing up to 35 TFBS, that control the expression of the gene during late embryonic and larval development in the sea urchin. Other examples of the modular design, this time in higher eukaryotes, have been described for the control of human muscle-specific genes (18), where five TFBS are described in what is likely to be a minimal set of control elements in muscle-specific gene regulation.

The *CI* score revealed highly co-localized PWM pairs that have not been reported in the curated literature of the TRANSFAC database. Our examination of the PubMed literature for a randomly chosen set of 50 PWM pairs separated by 50 or 200 bp did reveal experimental evidence supporting the proposed hypothesis of synergism for between 32 and 46% of the pairs (Tables 1 and 2). In addition, 39–43% of the reported synergist pairs have evidence for physical interaction between the TF proteins, indicating that the *CI* score is a useful determinant for synergistic and potentially interacting TF proteins. We have been able to identify 251 and 70 novel PWM pairs with *CI* values greater than the maximum *CI* found for PWM pairs reported in TRANSFAC (*CI* = 2.31 and 1.77 for 50 and 200 bp separation, respectively). There are 33 PWM pairs that are in common between these two sets, but for the most part it is clear that certain TFBS pairs have preferred spacing, as indicated by their occurrence in one set of higher *CI* scoring pairs and not the other. This finding is consistent with the notion that the proteins bound to the TFBS will interact and form complexes that themselves are likely to form optimally at given separations through surface interactions.

During the course of this study we noted that some PWMs contained in TRANSFAC are very similar to each other by using a Smith–Waterman based approach to align PWM pairs without gaps. The alignment method resulted in a distance measure between PWM and with a suitably chosen distance threshold we removed similar PWM pairs. This resulted in a corresponding 34% reduction in the 444 PWM pairs known to exist from experimental data. The reduction essentially eliminates multiple PWMs that are either similar or are assigned the same TF name and thus eliminates trivial PWM pair combinations.

The precise control of transcription is probably achieved through combinatorial interactions among TFs. Although higher order combination is relevant in this regard, discovering the genes that are regulated by novel TF protein pairs could provide an initial means by which common regulatory pathways can be identified. Indeed, identifying and clustering all PWM pairs with a *CI* score >2.0 did not result in the discovery of a precise set of higher order interactions. In this situation most PWMs were members of one large cluster (data not shown). Another approach to building higher order interaction models and the genes that may be regulated by proximal TFBS is to start with high confidence co-occurring paired TFBS.

Thus, using the annotation of the known TFBS on the genome sequence it is interesting to compare how many transcripts possess a single TFBS, a TFBS pair as predicted by an elevated *CI* score and finally a TFBS triplet due to the combination of two pairs. This will provide a measure of specificity in looking for TF interactions and the genes so regulated. For example, we were able to search for the occurrence of the PWMs for the AhR, E2F and ER TFBS by explicitly finding the locations of the PWMs M00139, M00180 and M00191, respectively, and the transcripts that contain them in the 5 kb upstream of the start codon. There are 5041 transcripts that have an AhR TFBS in the 5 kb upstream of the start codon, 440 transcripts that contain AhR and E2F within 200 bp of each other and only 21 transcripts that contain the combination of AhR, E2F and ER TFBS within 200 bp. When searching for genes potentially regulated by certain TFs, a search based on higher order combination will be clearly of great help due to the specificity of the search results. In the future, we will extend the notion of *CI* to include multiple PWMs and the information this provides on regulatory modules and the coordinately regulated genes they control.

ACKNOWLEDGEMENTS

We would like to thank the referees and Mike Flanigan for their valuable comments.

REFERENCES

1. Yuh,C.H., Bolouri,H. and Davidson,E.H. (1998) Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science*, **279**, 1896–1902.
2. Werner,T. (2000) Identification and functional modelling of DNA sequence elements of transcription. *Brief Bioinformatics*, **1**, 372–380.
3. GuhaThakurta,D. and Stormo,G.D. (2001) Identifying target sites for cooperatively binding factors. *Bioinformatics*, **17**, 608–621.
4. Kel,A., Kel-Margoulis,O., Babenko,V. and Wingender,E. (1999) Recognition of NFATp/AP-1 composite elements within genes induced upon the activation of immune cells. *J. Mol. Biol.*, **288**, 353–376.
5. Frith,M.C., Hansen,U. and Weng,Z. (2001) Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics*, **17**, 878–889.
6. Klingenhoff,A., Frech,K., Quandt,K. and Werner,T. (1999) Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. *Bioinformatics*, **15**, 180–186.
7. Wagner,A. (1999) Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics*, **15**, 776–784.
8. Pilpel,Y., Sudarsanam,P. and Church,G.M. (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genet.*, **29**, 153–159.
9. Wingender,E., Dietze,P., Karas,H. and Knuppel,R. (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, **24**, 238–241.
10. Kel-Margoulis,O.V., Romashchenko,A.G., Kolchanov,N.A., Wingender,E. and Kel,A.E. (2000) COMPEL: a database on composite regulatory elements providing combinatorial transcriptional regulation. *Nucleic Acids Res.*, **28**, 311–315.
11. Sjolander,K. (1998) Phylogenetic inference of protein superfamilies: analysis of SH2 domains. In *Proceedings of the Sixth International Conference on Intelligent Systems in Molecular Biology*, pp. 165–174.
12. Levy,S., Hannenhalli,S. and Workman,C. (2001) Enrichment of regulatory signals in conserved non-coding genomic sequence. *Bioinformatics*, **17**, 871–877.
13. 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**, 281–283.
14. Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.
15. Kel,O.V., Romaschenko,A.G., Kel,A.E., Wingender,E. and Kolchanov,N.A. (1995) A compilation of composite regulatory elements affecting gene transcription in vertebrates. *Nucleic Acids Res.*, **23**, 4097–4103.

16. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
17. Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.
18. Wasserman,W.W. and Fickett,J.W. (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.*, **278**, 167–181.
19. Supakar,P.C., Fujita,T. and Maruyama,N. (2000) Identification of novel sequence-specific nuclear factors interacting with mouse senescence marker protein-30 gene promoter. *Biochem. Biophys. Res. Commun.*, **272**, 436–440.
20. Pollmann,C., Huang,X., Mall,J., Bech-Otschir,D., Naumann,M. and Dubiel,W. (2001) The constitutive photomorphogenesis 9 signalosome directs vascular endothelial growth factor production in tumor cells. *Cancer Res.*, **61**, 8416–8421.
21. Lewin,B. (1997) *Genes VI*. Oxford University Press, Oxford, UK.
22. Mandolesi,G., Gargano,S., Pennuto,M., Illi,B., Molfetta,R., Soucek,L., Mosca,L., Levi,A., Jucker,R. and Nasi,S. (2002) NGF-dependent and tissue-specific transcription of vgf is regulated by a CREB-p300 and bHLH factor interaction. *FEBS Lett.*, **510**, 50–56.
23. Vries,R.G., Prudenziati,M., Zwartjes,C., Verlaan,M., Kalkhoven,E. and Zantema,A. (2001) A specific lysine in c-Jun is required for transcriptional repression by E1A and is acetylated by p300. *EMBO J.*, **20**, 6095–6103.
24. Miyagishi,M., Fujii,R., Hatta,M., Yoshida,E., Araya,N., Nagafuchi,A., Ishihara,S., Nakajima,T. and Fukamizu,A. (2000) Regulation of Lef-mediated transcription and p53-dependent pathway by associating beta-catenin with CBP/p300. *J. Biol. Chem.*, **275**, 35170–35175.
25. Oliner,J.D., Andresen,J.M., Hansen,S.K., Zhou,S. and Tjian,R. (1996) SREBP transcriptional activity is mediated through an interaction with the CREB-binding protein. *Genes Dev.*, **10**, 2903–2911.
26. Youn,H.D., Chatila,T.A. and Liu,J.O. (2000) Integration of calcineurin and MEF2 signals by the coactivator p300 during T-cell apoptosis. *EMBO J.*, **19**, 4323–4331.
27. Butscher,W.G., Powers,C., Olive,M., Vinson,C. and Gardner,K. (1998) Coordinate transactivation of the interleukin-2 CD28 response element by c-Rel and ATF-1/CREB2. *J. Biol. Chem.*, **273**, 552–560.
28. Furia,B., Deng,L., Wu,K., Baylor,S., Kehn,K., Li,H., Donnelly,R., Coleman,T. and Kashanchi,F. (2002) Enhancement of nuclear factor-kappa B acetylation by coactivator p300 and HIV-1 Tat proteins. *J. Biol. Chem.*, **277**, 4973–4980.
29. Steger,D.J., Hecht,J.H. and Mellon,P.L. (1994) GATA-binding proteins regulate the human gonadotropin alpha-subunit gene in the placenta and pituitary gland. *Mol. Cell. Biol.*, **14**, 5592–5602.
30. Novak,A. and Dedhar,S. (1999) Signaling through beta-catenin and Lef/Tcf. *Cell. Mol. Life Sci.*, **56**, 523–537.
31. Halle,J.P., Haus-Seuffert,P., Woltering,C., Stelzer,G. and Meisterernst,M. (1997) A conserved tissue-specific structure at a human T-cell receptor beta-chain core promoter. *Mol. Cell. Biol.*, **17**, 4220–4229.
32. Sartorelli,V., Huang,J., Hamamori,Y. and Kedes,L. (1997) Molecular mechanisms of myogenic coactivation by p300: direct interaction with the activation domain of MyoD and with the MADS box of MEF2C. *Mol. Cell. Biol.*, **17**, 1010–1026.
33. Tsai,E.Y., Falvo,J.V., Tsytsykova,A.V., Barczak,A.K., Reimold,A.M., Glimcher,L.H., Fenton,M.J., Gordon,D.C., Dunn,I.F. and Goldfeld,A.E. (2000) A lipopolysaccharide-specific enhancer complex involving Ets, Elk-1, Sp1, and CREB binding protein and p300 is recruited to the tumor necrosis factor alpha promoter *in vivo*. *Mol. Cell. Biol.*, **20**, 6084–6094.
34. Chau,N.H., Vanson,C.D. and Kerry,J.A. (1999) Transcriptional regulation of the human cytomegalovirus US11 early gene. *J. Virol.*, **73**, 863–870.
35. Kapatos,G., Stegenga,S.L. and Hirayama,K. (2000) Identification and characterization of basal and cyclic AMP response elements in the promoter of the rat GTP cyclohydrolase I gene. *J. Biol. Chem.*, **275**, 5947–5957.
36. Udalova,I.A. and Kwiatkowski,D. (2001) Interaction of AP-1 with a cluster of NF-kappa B binding elements in the human TNF promoter region. *Biochem. Biophys. Res. Commun.*, **289**, 25–33.
37. Kawana,M., Lee,M.E., Quertermous,E.E. and Quertermous,T. (1995) Cooperative interaction of GATA-2 and AP1 regulates transcription of the endothelin-1 gene. *Mol. Cell. Biol.*, **15**, 4225–4231.
38. Lennon,A.M., Ottone,C., Rigaud,G., Deaven,L.L., Longmire,J., Fellous,M., Bono,R. and Alcaide-Loridan,C. (1997) Isolation of a B-cell-specific promoter for the human class II transactivator. *Immunogenetics*, **45**, 266–273.
39. Harada,S., Sampath,T.K., Aubin,J.E. and Rodan,G.A. (1997) Osteogenic protein-1 up-regulation of the collagen X promoter activity is mediated by a MEF-2-like sequence and requires an adjacent AP-1 sequence. *Mol. Endocrinol.*, **11**, 1832–1845.
40. Minami,T. and Aird,W.C. (2001) Thrombin stimulation of the vascular cell adhesion molecule-1 promoter in endothelial cells is mediated by tandem nuclear factor-kappa B and GATA motifs. *J. Biol. Chem.*, **276**, 47632–47641.
41. Kirch,H.C., Flaswinkel,S., Rumpf,H., Brockmann,D. and Esche,H. (1999) Expression of human p53 requires synergistic activation of transcription from the p53 promoter by AP-1, NF-kappaB and Myc/Max. *Oncogene*, **18**, 2728–2738.
42. Adams,C.C. and Workman,J.L. (1995) Binding of disparate transcriptional activators to nucleosomal DNA is inherently cooperative. *Mol. Cell. Biol.*, **15**, 1405–1421.
43. Kobayashi,A., Numayama-Tsuruta,K., Sogawa,K. and Fujii-Kuriyama,Y. (1997) CBP/p300 functions as a possible transcriptional coactivator of Ah receptor nuclear translocator (Arnt). *J. Biochem. (Tokyo)*, **122**, 703–710.
44. Lazennec,G., Thomas,J.A. and Katzenellenbogen,B.S. (2001) Involvement of cyclic AMP response element binding protein (CREB) and estrogen receptor phosphorylation in the synergistic activation of the estrogen receptor by estradiol and protein kinase activators. *J. Steroid Biochem. Mol. Biol.*, **77**, 193–203.
45. Castro-Rivera,E., Samudio,I. and Safe,S. (2001) Estrogen regulation of cyclin D1 gene expression in ZR-75 breast cancer cells involves multiple enhancer elements. *J. Biol. Chem.*, **276**, 30853–30861.
46. Dumais,N., Bounou,S., Olivier,M. and Tremblay,M.J. (2002) Prostaglandin E(2)-mediated activation of HIV-1 long terminal repeat transcription in human T cells necessitates CCAAT/enhancer binding protein (C/EBP) binding sites in addition to cooperative interactions between C/EBPbeta and cyclic adenosine 5′-monophosphate response element binding protein. *J. Immunol.*, **168**, 274–282.
47. Jung,D.J., Na,S.Y., Na,D.S. and Lee,J.W. (2002) Molecular cloning and characterization of CAPER, a novel coactivator of activating protein-1 and estrogen receptors. *J. Biol. Chem.*, **277**, 1229–1234.
48. Molnar,A. and Georgopoulos,K. (1994) The Ikaros gene encodes a family of functionally diverse zinc finger DNA-binding proteins. *Mol. Cell. Biol.*, **14**, 8292–8303.
49. Dhulipala,P.D. and Kotlikoff,M.I. (1999) Cloning and characterization of the promoters of the maxiK channel alpha and beta subunits. *Biochim. Biophys. Acta*, **1444**, 254–262.
50. Love,J.J., Li,X., Case,D.A., Giese,K., Grosschedl,R. and Wright,P.E. (1995) Structural basis for DNA bending by the architectural transcription factor LEF-1. *Nature*, **376**, 791–795.
51. Elferink,C.J., Ge,N.L. and Levine,A. (2001) Maximal aryl hydrocarbon receptor activity depends on an interaction with the retinoblastoma protein. *Mol. Pharmacol.*, **59**, 664–673.
52. Klinge,C.M., Jernigan,S.C., Risinger,K.E., Lee,J.E., Tyulmenkov,V.V., Falkner,K.C. and Prough,R.A. (2001) Short heterodimer partner (SHP) orphan nuclear receptor inhibits the transcriptional activity of aryl hydrocarbon receptor (AHR)/AHR nuclear translocator (ARNT). *Arch. Biochem. Biophys.*, **390**, 64–70.
53. Klinge,C.M., Kaur,K. and Swanson,H.I. (2000) The aryl hydrocarbon receptor interacts with estrogen receptor alpha and orphan receptors COUP-TFI and ERRalpha1. *Arch. Biochem. Biophys.*, **373**, 163–174.
54. Gonzalez,F.J., Kimura,S. and Nebert,D.W. (1985) Comparison of the flanking regions and introns of the mouse 2,3,7,8-tetrachlorodibenzo-p-dioxin-inducible cytochrome P1-450 and P3-450 genes. *J. Biol. Chem.*, **260**, 5040–5049.
55. Lazennec,G., Kern,L., Valotaire,Y. and Salbert,G. (1997) The nuclear orphan receptors COUP-TF and ARP-1 positively regulate the trout estrogen receptor gene through enhancing autoregulation. *Mol. Cell. Biol.*, **17**, 5053–5066.
56. Satoh,H., Nagae,Y., Immenschuh,S., Satoh,T. and Muller-Eberhard,U. (1994) Identification of a liver preference enhancer element of the rat hemopexin gene and its interaction with nuclear factors. *J. Biol. Chem.*, **269**, 6851–6858.