Mammalian Genome 13, 327–332 (2002).
DOI: 10.1007/s00335-001-2141-8

Mammalian Genome
Incorporating Mouse Genome

Computational analysis of composite regulatory elements

Ping Qiu, Wei Ding, Ying Jiang, Jonathan R. Greene, Luquan Wang

Bioinformatics Group and Human Genomic Research Department at Schering-Plough Research Institute, 2015 Galloping Hill Road, Kenilworth, New Jersey 07033, USA

Received: 5 November 2001 / Accepted: 30 January 2002

Abstract. Combinatorial regulation is a powerful mechanism for generating specificity in gene expression, and it is thought to play a pivotal role in the formation of the complex gene regulatory networks found in higher eukaryotes. The term "Composite Element" (CE) refers to a minimal functional unit where protein–DNA and protein–protein interactions contribute to a highly specific pattern of gene transcriptional regulation. Identification of composite elements will help to better understand gene regulation networks. Experimentally identified CEs are limited in number, and the currently available CE database COMPEL is based on such published information. Here, based on the statistical analysis of over-represented adjacent transcription factor binding sites, we describe a computational method to predict composite regulatory elements in genomic sequences. The algorithm proved to be efficient for extracting composite elements that had been experimentally confirmed and documented in the COMPEL database. Furthermore, putative new composite elements are predicted based on this method, and we have been able to confirm some of our predictions which are not included in the COMPEL database by searching published information.

Eukaryotic gene regulation involves the assembly of an initiation complex at the core promoter region and regulatory complexes at promoter-enhancer regions. The promoter region is usually located just proximal to or overlapping the transcription initiation site and consists of several sequence elements with which transcription factors (TFs) interact in a sequence-specific manner. When recruited, these TFs serve as molecular switches, which turn the transcription of the gene on or off. The combinations of the TF-binding elements in promoters vary depending on the gene, which provides the molecular basis of temporal and spatial gene expression (Mitchell and Tjian 1989; Novina and Roy 1996).

In the last few years, more and more evidence suggests that the complex differential expression of genes in higher organisms is achieved through combinatorial regulation of transcription by a specific combination of transcription factors binding to their target sites in the regulatory regions of these genes. Just a few tissue-specific transcription factors with distinct tissue distributions have the potential to act in different combinations to direct many different patterns of gene expression (Chen 1999; Wolberger 1998). One of the best-studied such examples is that of composite NFAT/AP-1 sites, in which it was demonstrated that these two factors bind cooperatively to activate cytokine gene expression (Jain et al. 1993; Rao 1994; Rao et al. 1997; Northrop et al. 1993; Crabtree 1999; Lee et al 1995; Cockerill et al. 1993, 1995). For genome-wide analysis, microarray data have been used to uncover novel combinatorial functional motif in the promoters of *Saccharomyces cerevisiae* (Pilpel et al. 2001).

Composite Elements (CEs) were first introduced by Diamond et al. (1990) when they studied the interaction between a glucocorticoid receptor binding site and its adjacent AP-1 site in mouse proliferin promoter. The CE model was defined further by Kel-Margoulis et al. (2000) as pairs of closely situated binding sites, corresponding transcription factors, protein–protein interaction between them, and expression patterns provided by this combinatorial regulation. There are two main types of CEs: synergistic and antagonistic. In synergistic CEs, simultaneous interactions of two factors with closely situated target sites result in a high level of transcriptional activation. In an antagonistic CE, two factors interfere with each other, in some cases resulting in mutually exclusive binding. There are other examples where factors can bind to DNA simultaneously, but binding of a repressing factor may mask an activation domain of an activator (Wingender et al. 1997). Computational analysis and prediction of regulatory elements (Scherf et al. 2000; Werner 1999; Frech et al. 1997, 1998; Fickett and Hatzigeorgiou 1997) as well as CEs have been an active research area. Most studies in this direction focused on either target gene identification (Wagner 1999) or on a particular transcription factor (Kel et al. 1999). A recent study utilized a Gibbs sampling strategy to model the cooperativity between two transcription factors and defined position weight matrices for the binding sites (GuhaThakurta and Stormo 2001).

Even with the completed working draft of the human genome sequence, functions of more than half of the human genes are still unknown. It would be beneficial to be able to identify the regulatory regions that confer temporal and spatial expression patterns for the uncharacterized genes. Additionally, it would be advantageous to identify regulatory regions within genes of known expression pattern without performing the costly and time-consuming laboratory studies now required. To achieve these goals, the wealth of case studies performed over the past years will have to be collected. One such ongoing effort is the COMPEL database. Kel-Margoulis et al. developed the COMPEL database (http://compel.bionet.nsc.ru/compel/search.html), in which they have collected published information on composite regulatory elements (Kel et al. 1995, Kel-Margoulis et al. 2000; Wingender et al. 1997). Yet, until now the entries in COMPEL 3.0 are still very limited (178 entries).

In this study, we describe a novel computational approach to detect possible composite elements in genomic sequence. The method is based on the detection of over-represented adjacent transcription binding sites. Such over-represented composite binding sites are very unlikely to occur by chance alone, as opposed to individual sites, which are often abundant in promoter regions as well as in other regions of the genome.

Materials and methods

Resources for databases and computer programs. Genebank release 120 was downloaded from *ftp://ncbi.nlm.nih.gov*. TRANSFAC (Wingender et al. 1996, 2001) and Matinspector (Quandt et al. 1995) were licensed from Biobase. TRANSFAC is a database on transcription factors, their

Correspondence to: P. Qiu; E-mail: ping.qiu@spcorp.com

genomic binding sites, and DNA-binding site sequence profiles (http://transfac.gbf.de/TRANSFAC/). One of the most important parts of TRANSFAC is the MATRIX entries, which represent DNA binding site sequence profiles for individuals or groups of transcription factors. Matinspector is a computer program that can detect potential sequence matches by automatic searches with a library of pre-compiled matrices. Sequence alignment software for transcript mapping AAT (Huang et al. 1997) was licensed from Michigan Technological University. AAT is a local alignment software that extended the BLAST algorithm by assigning fixed penalty to long gaps. All non-commercial software used in this study was written in PERL 5.0.

Transcript mapping and construction of reference promoter database. A collection of human mRNA was first extracted from the primate division of GenBank flat file (Release 120). To ensure that the 5′ end of an mRNA is close to the transcription start site, only mRNAs that encode the N-terminus of the protein were used for transcript mapping, and only sequence in the Genbank Refseq database is used to reduce gene redundancy. Transcript mapping was done based on the October 2000 Freeze of the University of California at Santa Cruz's Working Draft Sequence (*http://genome.ucsc.edu*), which presents a tentative assembly of the finished and draft human genomic sequence based on the Washington University-Saint Louis clone map (*http://genome.wustl.edu/gsc*). For alignment of the 5′ end of the cDNA with the genome sequence, we used a local alignment software package AAT (Huang et al. 1997). To reduce the number of undesirable matches due to interspersed repeats, the DNA sequence is screened for interspersed repeats by using the RepeatMasker program (Smit, AFA and Green, P at http://ftp.genome.washington.edu/RM/RepeatMasker.html). Promoter regions were defined as the sequences extending from 2000 bp upstream of the first exon, but not beyond the gaps of unfinished genomic BAC sequence if such a gap existed. The validation of this promoter reference database by comparing with GenBank annotated promoters has been described in a previous published paper (Wang et al. 2001).

TF site analysis and statistical analysis. Promoter sequences are fetched by taking 2000 bp upstream of the first exon based on the transcript mapping of each sequence. The promoter sequences are then checked for the transcription factor binding site by running Matinspector against TRANSFAC TF binding site matrix library. The output file from Matinspector was parsed and stored in Sybase relational database table. Matrix similarity scores (MSS) of 0.8 and 0.9 were used as cutoff scores in separate analyses. Matrix similarity score is between 0.0 and 1.0, and 0.8 is considered to be a significant high score. If two TF sites can occupy any position in a sequence of n-bp, then the total number of the combinations is n*n. If two TF sites maintain an inter-distance less than m-bp in a sequence of n-bp, then the number of combinations can be calculated as following:

$n + 2\,[(n-1) + (n-1) + \ldots + (n-m)] = n + 2\,[nm - m(m+1)/2] = n + 2nm - m*m - m = (2n-m)(m+1) - n$

Therefore, the chance of two TF sites to exist within m-bp distance in a n-bp long sequence can be defined by the following:

$$F(f1,f2) = \frac{F(f1)F(f2)((2n-m)(m+1)-n)}{n*n}$$

Where $F(f1)$ is the frequency of TF site1 to appear in one n-bp long sequence in our reference promoter database with size of N, $F(f2)$ is the frequency of TF site2 to appear in one n-bp long sequence in our reference promoter database, n = 2000-bp, m = 20-bp (and 50-bp), and N = 1370 promoter sequences in our case.

The expected frequency of any pair of two TF sites to appear within 20-bp (or 50-bp) in our promoter sequence database is calculated by:

$$\text{expected} = N * F(f1, f2), \text{ where } N = 1370$$

The observed frequency of any pair of two TF sites to appear within 20-bp (or 50-bp) in our promoter sequence database is obtained by querying the database constructed from Matinspector output. As the discrepancies between the observed and expected values increase, the value of the statistical variable chi-square (χ^2) becomes larger and the resulting P value becomes smaller, which describes the probability of randomly selected subjects having this large a discrepancy between observed and expected values. With the degree of freedom = 1 in our case, to exclude the false positives with a simple Bonferroni correction, a reasonable significance level would be $P = 0.005/1370 = 3.65 \times 10\text{E-}6$, which correspond to $\chi^2 = 21$. Chi-square value is calculated by:

$$\chi^2 = \sum \frac{(|\text{observed} - \text{expected}| - 0.5)^2}{\text{expected}}$$

Results

Composite elements prediction. To understand the mechanism of transcriptional regulation for a given gene, it is very important to identify and characterize its promoter. Despite the important roles of the promoters, the number of genes whose promoters have been identified is limited. In the Eukaryotic Promoter Database (EPD; *http://www.epd.isb-sib.ch*) (Perier et al. 2000), which collected previously experimentally characterized promoter sequences, only a small amount of human promoters have been registered. To circumvent this problem, a computational transcript mapping approach was used to locate promoter sequences for human genes within their genomic organization, as described in Materials and methods. The promoter reference database was validated by comparing with GenBank annotated promoters. We sampled 150 promoters annotated in GenBank; 133 (88%) were perfectly predicted by the transcript mapping, suggesting that the transcript mapping procedure could properly predict most promoters. This result has been described in a previously published paper (Wang et al. 2001).

To eliminate the possible redundancy in our reference promoter database, we used only the mRNA sequences from Genbank Refseq section for promoter region extraction in our analysis. For each gene, the genomic sequence 2000-bp upstream of the 5′ end of the mRNA was retrieved as a promoter region. This resulted in a set of 1370 promoter regions to be used in this analysis. These promoter regions were scanned for potential binding sites by using the Matinspector program and the TRANSFAC transcription binding site scoring matrix library, as described in Materials and methods. Two separate sets of potential TF binding sites were gathered by using different stringency of matrix similarity score (MSS) with cutoff values of 0.8 and 0.9. Matrix similarity score 0.8 is the default similarity value for Matinspector.

In most composite regulatory elements, the two TF binding sites exist within a short distance. We analyzed all the entries documented in the COMPEL database; about 65% of the CEs exist within a 20-bp distance, and about 87% of CEs are within a 50-bp distance. In our study, we used 20-bp and 50-bp as distance cutoffs to predict composite elements. Therefore, our analysis resulted in four determinations of composite elements by using the MSS cutoffs of 0.8 and 0.9 in conjunction with the 20-bp and 50-bp distance cutoffs, hereafter referred to as DIS = 20 and DIS = 50 respectively. The random frequency of two TF binding sites existing within 20-bp or 50-bp over a 2000-bp promoter sequence was calculated as described in Materials and methods. The discrepancy between the actual frequency of the composite elements and the random frequency was evaluated by determining the statistically variable chi-square (χ^2). The p value was further derived from chi-square. The higher the χ^2 or the lower the p value, the more unlikely it is that the composite elements exist within 20-bp (or 50-bp) randomly, which means the more likely it is that their close-by co-existence is biologically significant.

Table 1 lists all the CEs computed to have a $\chi^2 \geq 21$ (MSS = 0.9 and DIS = 20, denoted as MSS = 0.9/DIS = 20). 163 human TF binding site matrices from TRANSFAC 4.4.2 were used for this analysis. Out of the 13,203 possible combinations of any two TF matrices, 236 pairs co-exist within a 20-bp (DIS = 20) distance, are over-represented in the reference promoter database, and have a χ^2 value of 21 or above (MSS = 0.9), which accounts for 1.8% of the total possible combination. Given the fact that for some TFs more than one matrix was generated in the TRANSFAC matrix library, only the ones that have the highest χ^2 values are

Table 1. List of potential composite elements predicted by the in silico method. Matrix similarity score cutoff MSS = 0.9 and composite element distance cutoff DIS = 20-bp.

Factor 1	Factor 2	χ^2	Factor 1	Factor 2	χ^2	Factor 1	Factor 2	χ^2	Factor 1	Factor 2	χ^2
HFH3	SRY	22116.5	MYOD	SREBP1	178.9	CREB	SREBP1	61.5	HLF	Oct-1	29.3
FREAC7	HFH3	20094.2	RORA1	TCF11	178.2	CEBP	GATA	59.2	BRN2	MEF2	29.1
CEBPA	E4BP4	4762.6	AP1FJ	CREB	176.6	E47	TST1	58.5	BRN2	TCF11	28.9
AP2	SP1	1708.7	ER	PAX3	173.1	MZF1	RREB1	58.4	MYCMAX	WHN	28.9
COUP	HNF4	1698.5	NFY	PBX1	170.4	NFKAPPAB50	SP1	57.6	CEBP	GATA1	28.9
AHRARNT	MYCMAX	1424.3	HSF2	NFAT	170.3	BRN2	HNF1	56.6	ISRE	MYB	28.7
AP4	E47	1334.6	CDP	NFY	167.6	ER	TCF11	56.2	AP4	NFKB	28.4
AML1	CEBP	1263.9	TFC11	TFC11MAFG	154.3	CDPCR1	GATA1	56.1	CEBP	SRF	28.3
CDPCR3HD	PBX1	1177.3	FREAC7	GATA1	153.0	CDP	GATA1	52.0	P300	SREBP1	27.9
MZF1	SP1	1089.2	CREBP1	XBP1	146.8	GATA3	NFY	51.9	AML1	E47	27.3
BRN2	Oct-1	919.4	CREL	NFAT	144.7	GR	TAL1BETAITF2	51.8	STAT	TST1	27.3
CREB	WHN	828.6	FREAC7	Oct-1	139.3	CREB	TCF11	51.4	E4BP4	FREAC7	27.0
FREAC7	TATA	796.3	CEBP	HFH3	136.5	CEBPB	Oct-1	50.7	LMO2COM	Oct-1	26.1
FREAC3	HFH3	708.4	HSF2	Oct-1	135.1	SRY	TATA	50.3	AHRARNT	CREBP1CJUN	26.0
AP4	MYOD	703.3	Oct-1	YY1	131.1	BRN2	HFH3	50.3	AP1FJ	RORA2	25.9
AP1	TCF11	633.5	Oct-1	TST1	130.1	AHRARNT	WHN	49.4	AHRARNT	SP1	25.7
ATF	XBP1	631.2	AP1	PAX2	129.2	GDPCR3HD	YY1	49.3	FREAC2	HFH3	25.1
GATA1	Oct-1	615.6	CREBP1CJUN	XBP1	125.7	CEBP	CREB	49.0	GATA1	TAL1BETAE37	25.0
Oct-1	PBX1	573.9	IRF1	NFAT	125.3	CEBP	FREAC2	46.5	CEBP	ETS2	25.0
CEBP	CHOP	559.1	CREB	PAX3	125.3	GATA1	PBX1	46.3	AP2	MYCMAX	24.8
ATF	WHN	546.3	CEBPA	HLF	124.0	MZF1	NFKAPPAB	46.0	CREL	MZF1	24.7
NF1	NFY	490.5	Oct-1	TCF11	123.3	MEF2	TATA	45.9	ELK1	SP1	24.6
CREB	XBP1	454.6	SRF	YY1	123.0	MZF1	NFKB	45.6	CDP	PBX1	24.3
CEBPB	E4BP4	449.0	HFKAPPAB65	Oct-1	118.4	NFKB	USF	44.9	CREBP1	Oct-1	24.1
NFAT	NFKAPPAB65	431.3	NFAT	YY1	111.4	MYB	RFX1	43.7	CEBP	RORA2	23.8
PAX2	TCF11	426.5	COUP	RORA1	104.0	SRF	TATA	43.7	BRN2	CDPCR3HD	23.7
MZF1	NFKAPPAB50	389.8	NFAT	Oct-1	100.7	CREB	GRE	43.2	GATA1	YY1	23.6
CEBP	Oct-1	374.7	AP2	MZF1	99.8	ATF	PAX3	43.0	PAX2	USF	23.6
AP1	PBX1	359.6	RSRFC4	TATA	95.2	ATF	TCF11	42.3	E47	TAL1BETAE47	23.5
CEBPA	Oct-1	356.2	CREB	XBP1	95.0	CDPCR3	ETS2	40.2	AP1	CREB	23.2
FREAC2	SRY	329.7	AHRARNT	P53	92.1	ISRE	NFAT	40.1	EGR2	USF	22.9
CREL	ELK1	316.7	E47	TAL1ALPHAE47	91.3	NFY	Oct-1	38.1	AHRARNT	LMO2COM	22.7
MYOD	TAL1ALPHAE47	313.3	CEBPB	YY1	88.8	ER	P300	38.1	AP4	RFX1	22.6
NFAT	STAT	307.9	AML1	LMO2COM	88.4	PAX3	WHN	37.4	ELK1	HSF2	22.4
ATF	XBP1	306.2	CREB	HNF1	87.4	ETS2	IRF1	36.9	GATA2	MZF1	22.3
BRN2	FREAC7	303.1	FREAC7	GATA	85.5	ATF	USF	36.9	NFAT	SRY	22.0
FREAC7	SRY	302.5	BRN2	PBX1	82.0	P53	XBP1	36.4	CREL	SRF	22.0
CREL	GABP	276.3	IRF1	SRY	81.9	CREB	USF	35.0	TATA	YY1	21.6
CREBP1	CREBP1	265.6	RORA2	TCF11	80.9	GR	HNF4	34.9	AHRARNT	ATF	21.6
BRN2	TATA	257.1	HNF4	RORA1	80.4	CREB	SRF	34.4	CHOP	NF1	21.4
USF	XBP1	256.3	EGR3	SP1	75.1	HNF1	TATA	33.9	HNF1	Oct-1	21.2
MZF1	P300	242.6	CDPCR3HD	TCF11	73.0	CREL	STAT	33.3			
Oct-1	TATA	230.5	MEF2	TATA	70.6	AP2	EGR3	33.1			
CEBP	SRY	224.2	CREL	ETS1	67.0	AP1	RORA2	33.1			
FREAC3	FREAC7	220.8	PBX1	TATA	66.3	SP1	USF	33.0			
ARNT	GRE	220.0	ATF	P53	63.4	MZF1	SP1	32.8			
E4BP4	Oct-1	209.3	MYCMAX	SP1	63.1	FREAC4	ISRE	32.8			
NFKAPPAB	P53	193.8	CEBPB	HLF	63.0	CREBP1CJUN	TCF11	31.5			
AP1FJ	CREB	188.9	HSF2	STAT	62.8	HNF1	SRF	31.0			
CEBP	NFAT	184.3	PAX2	PAX3	61.8	CDPCR3HD	SRF	30.6			

listed in the table, and therefore the number of unique CEs was reduced to 191.

Validation of predicted composite elements using COMPEL. In order to validate the composite elements derived from this computational analysis, one of the most direct ways would be to test the χ^2 value for all the 148 composite elements entries that were identified experimentally and documented in COMPEL release 2.4 database compiled by Kel-Margoulis et al. (Kel-Margoulis et al. 2000). Since only part of the CEs in the COMPEL and their corresponding TFs have binding site matrix entries in TRANSFAC 4.4.2 that were used in this study, the composite elements from COMPEL2.4 whose corresponding TFs have no binding site matrix entries in TRANSFAC are not included in the list for validation. After removal of redundancy and those TFs whose matrices are compiled from less than 10 TF binding sites and collection of only those entries with matrix entries in TRANSFAC, 40 CEs remain that we can test. Out of the 40 CEs, 15 of them were predicted with our method with χ^2 values ≥ 21 (Table 2). Most of them show significantly high χ^2 values. For example, it was shown by electrophoretic mobility shift assays (Zhang et al. 1996) that CCAAT enhancer-binding protein (C/EBP) and AML1 (CBF alpha2) synergistically activate the macrophage colony-stimulating factor receptor promoter. Our analysis shows that the χ^2 value for C/EBP/AML1 is 1263.9, which strongly suggests that these two factors have a very strong tendency to exist as a close pair. For another example, Mietus-Snyder et al. (1992) showed that HNF-4 is an activator of ApoCIII expression; both ARP-1 and COUP-TF are repressors; and Galson et al. (1995) showed antagonism between COUP–TF and HNF-4 in the regulation of tissue-specific and hypoxia-specific erythropoietin gene expression. Again, the χ^2 is 1698.5 for HNF-4/COUP in our analysis (Table 2). The other 25 composite elements from COMPEL that fall below our cutoff ($\chi^2 < 21$) are listed in Table 3 and will be discussed later.

Validation of predicted composite elements that are not in COMPEL database by other published information. It is interesting to know whether our prediction can pinpoint to some real CEs that have not yet been collected by COMPEL. One direct approach would be to take some predicted CEs with extremely high χ^2 values that are not in COMPEL and look for supporting information from the scientific literature.

Table 2. List of composite elements from COMPEL (Release 2.4) that can be predicted from in silico analysis. Highest Chi-square values for different combination of matrix similarity score cutoff (MSS = 0.8 and 0.9) and distance cutoff (DIS = 20-bp and 50-bp) are shown. (Note: only those transcription factors with binding site matrix entries in TRANSFAC 4.42 are shown.)

FACTOR 1	FACTOR 2	MATRIX 1	MATRIX 2	Highest χ^2
HNF-4	COUP	HNF4_01	COUP_01	1698.5
C/EBPalpha	AML1	CEBP_01	AML1_01	1263.9
NF-Y	NF-1	NFY_01	NF1_Q6	490.5
C/EBPalpha	NF-Y	CEBPA_01	NFY_Q6	380.4
HNF-1	Oct-1	HNF1_01	OCT1_03	189.5
YY1	SRF	SRF_Q6	YY1_01	123.0
CREB	HNF-1	CREB_01	HNF1_C	87.4
Sp1	NF-Y	SP1_Q6	NFY_01	67.6
Sp1	E2F-1	SP1_01	E2F_02	61.3
COUP-TF	ER	COUP_01	ER_Q6	49.5
HLH family	Octamer family	USF_C	OCT1_B	29.9
NF-kappaB	Sp1	NFKAPPAB_01	SP1_01	27.9
C/EBPbeta	HNF-1	CEBPB_02	HNF1_01	22.7
CREB/ATF family	NF-Y	CREBP1CJUN_01	NFY_01	21.9
C/EBPalpha	HNF-4	CEBP_01	HNF4_01	21.2

Table 3. List of composite elements from COMPEL (Release 2.4) that can not be predicted from our in silico analysis. Highest Chi-square values for different combination of matrix similarity score cutoff (MSS = 0.8 and 0.9) and distance cutoff (DIS = 20-bp and 50-bp) are shown. (Note: only those transcription factors with binding site matrix entries in TRANSFAC 4.42 are shown. obs<exp means observed frequency of occurrence is smaller than expected value owing to random variation. χ^2 is not calculated in these cases.)

FACTOR 1	FACTOR 2	MATRIX 1	MATRIX 2	Highest χ^2
SP1	c-Ets-1	SP1_Q6	ETS1_B	17.8
RFX	CEBP/ATF family	RFX1_01	ATF_01	8.5
AP-1	NFATp	AP_Q2	NFAT_A6	5.7
Sp1	MyoD	SP1_01	MYOD_Q6	4.4
GR	HNF-1	GR_Q6	HNF1_01	4.2
Elk-1	SRF	ELK1_02	SRF_Q6	4.1
C/EBPbeta	NF-kappaB	CEBPB_01	NFKAPPAB_01	4.0
Sp1	NF-1	SP1_Q6	NF1_Q6	3.2
ATF-3	NF-kappaB	ATF_01	NFKAPPAB65_01	2.4
NF-Atp	c-Fos	NFAT_Q6	AT1FJ_Q2	2.3
GATA-2	c-Jun	GATA1_03	CREBP1CJUN_01	1.9
c-Ets-1	GR	ETS1_B	GR_Q6	1.9
GR	c-Fos	GR_Q6	AP1FJ_Q2	1.7
AML1	c-Myb	AML1_01	MYB_Q6	1.4
GATA-3	CREB	GATA3_01	CREB_02	1.3
ETS family member	SRF-related protein	ETS1_B	SRF_Q6	0.9
RFX	NF-Y	RFX_01	NFY_01	0.8
GR	C/EBPbeta	GR_Q6	CEBPB_01	0.6
AP1	C/EBPbeta	AP1_C	CEBPB_Q2	0.6
YY1	NF-kappaB	YY1_01	NFKAPPAB_01	0.3
c-Jun	c-Ets-1	CREBP1CJUN_01	ETS1_B	0.3
CREB	HNF-4	CREB_02	HNF4_01	0.1
IRF-1	NF-kappaB	IRF1_01	NFKAPPAB65_01	0.1
Sp1	Oct-1	SP1_Q6	OCT1_Q6	obs<exp
Sp1	C/EBPbeta	SP1_Q6	CEBP_01	obs<exp

Our method shows that Pbx and Oct-1 co-exist with a χ^2 value of 573.9. Subramaniam et al. (1998) reported that the ubiquitously expressed POU-homeodomain protein Oct-1, together with a second ubiquitously expressed Pbx protein, is responsible for maximal PRL3 (prolactin) expression. As another example, Metz and Ziff (1991) demonstrated that C/EBP-related factors rNFIL-6 and rE12 bind to the serum response element (SRE) at sites adjacent to the major c-fos regulatory element, the DSE, which is the binding site for serum regulatory factor (SRF); the χ^2 value for SRF and C/EBP in our analysis is 28.3. As another example, Schwenger et al. (1999) reported that the novel combination of YY1 and the nuclear factor of activated T cells (NF-AT) transcription factors bind to a distal hIL-5 promoter element where both factors are involved in down-regulation of hIL-5 gene expression in human T cells; the χ^2 for NF-AT/YY1 in our analysis is 111.4. As yet another example, Belsham and Mellon (2000) showed that Oct-1 and C/EBPβ are both downstream transcriptional regulators involved in the repression of GnRH gene expression by the glutamate/NO/cGMP signal transduction pathway, and χ^2 for Oct-1 and C/EBP is 50.7. Lastly, Fukada and Tonks (2001) demonstrated the reciprocal role of Egr-1 and SP family proteins in the regulation of the PTP1B promoter in response to the p210 Bcr-Abl oncoprotein-tyrosine kinase, and the χ^2 for EGR/SP1 is 75.1.

In our study, we used the fairly conservative chi-square of 21 as the cutoff. By increasing the χ^2 cutoff in our study, the specificity of the prediction can be increased while the sensitivity will be sacrificed. We are also aware of the fact that some false prediction might originate from the non-uniformity of the human DNA composition. Since we cannot validate all the putative composite elements owing to lack of experimental data, it is difficult to evaluate the extent of false-positive predictions in our results. Nevertheless, given the fact that a large percentage of the documented CEs have been predicted by this method, the method is proven to be efficient for predicting and pinpointing real composite elements and suggests many possibilities for further exploration.

Discussion

The accurate identification of regulatory elements within a genomic sequence is a difficult challenge, both experimentally and

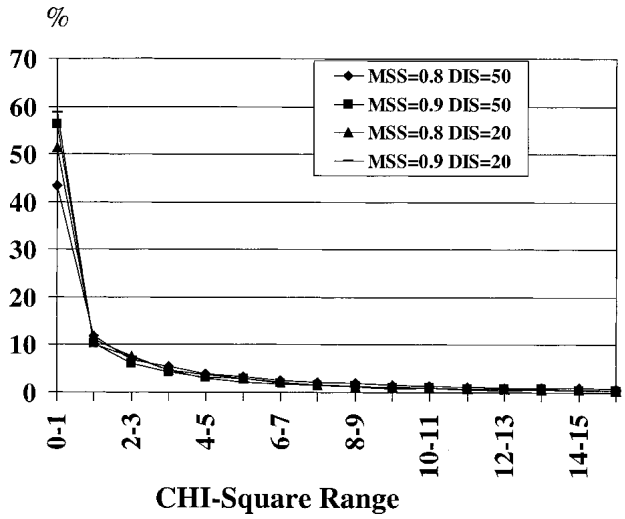

Fig. 1. Plotted distribution of the number of TF binding site pairs with different chi-square range. Four different criteria (matrix similarity score MSS = 0.8 or 0.9 and distance of composite elements DIS = 20-bp or 50-bp) were shown.

computationally. With the available working draft of the human genome (International Human Genome Sequencing Consortium 2001; Venter et al. 2001), the huge amount of uncharacterized genomic sequence will preclude experimental analysis of each gene's regulatory structure, making computational identification of protein *cis*-acting elements valuable. However, given the flexibility of the regulatory mechanisms, one can hardly develop a comprehensive method that could detect all the regulatory signals systematically. By combining profiles of some relatively well-characterized regulatory elements with statistical significance analysis of their close-by co-existence, we have generated an efficient computational means of identifying CEs.

It should be noted that some composite elements in the COMPEL database were not identified by using the parameters of the analysis presented here. Multiple reasons could account for this outcome. 1) Some TRANSFAC TF binding site matrices are outdated in terms of quality and specificity. For example, we have not been able to identify some known p53 target genes using V$P53_01, which is the binding site matrix for p53. 2) Some TRANSFAC matrices are not accurate enough; for example, about 18% of the total matrices (as of TRANSFAC 4.4.2) are built based on less than 10 binding sites, which could cause a substantial sampling error. 3) Another factor would be the matrix-similarity-score cutoff we used for the matrix searching software Matinspector. As mentioned in Materials and methods, we use 0.8 and 0.9 as cutoffs in our study. A lower cutoff score would increase the sensitivity but lower the specificity for some TFs. 4) The distance cutoff between the composite elements we used, 20-bp or 50-bp, might not reflect the actual distance for some composite elements. For example, about 13% of the CEs in COMPEL release 2.4 have a distance greater than 50-bp. Again, increasing the CE distance cutoff in the analysis might increase the sensitivity but decrease the specificity and, therefore, decrease the χ^2 value. For example, AP1 and NF-AT is a pair of well-known synergistic transcription factors, but the χ^2 test for their co-existence failed to pass our cutoff score (MSS = 0.9/DIS = 20, $\chi^2 \geq 21$). If we use MSS = 0.8 and DIS = 50 instead, the χ^2 value is greatly increased from 1.2 to 5.67. In the SP1/ETS-1 example, the χ^2 value is 17.8 in the MSS = 0.8/DIS = 50 combination, while in the MSS = 0.8/DIS = 20 combination the χ^2 is 1.12. Figure 1 shows the distribution of the TF binding site pairs with different χ^2 ranges. It is also important to mention that the candidate composite elements in the MSS = 0.9/DIS = 50 results set are not necessarily in the MSS = 0.8/DIS = 50 results set, since the number of matches might dramatically change if the cutoff for a certain matrix is relaxed. Therefore, the χ^2 calculated might be dramatically decreased accordingly as well. Which cutoff for matrix similarity score and CE distance to use for analysis really depends on the nature of the two factors and the nature of how the matrix is built. 5) Even though we have 1370 promoter sequences in our reference database, this is by far less than the total number of predicted genes with the most conservative recent estimates of human gene numbers, which is ~30,000 (Ewing and Green 2000; Roest Crollius 2000). Some TF binding sites with low frequency of occurrence might never have been represented in our reference promoter database. 6) The promoter quality is another factor that affects the outcome of the prediction. We have taken 2000-bp upstream of the mRNA 5′ end as the promoter region. Since we know that most of the known CEs fall between −250 bp and the transcription start site, 2000-bp might be too long in some cases; thus, the noise level might be increased.

Information about the known CEs and the specific gene regulation achieved through such CEs is going to be extremely useful for promoter prediction, gene function prediction, gene engineering, as well as the gene regulation network and biological pathway modeling. This prediction algorithm might also help to supplement COMPEL or other similar database, since it can efficiently point to high-quality putative composite elements. The performance of the prediction method described here is sufficiently specific to warrant further analysis of predicted composite elements.

Acknowledgments. The authors thank Dr.David Stillman for critical reading and valuable comments on the manuscript.

References

Belsham DD, Mellon PL (2000) Transcription factors Oct-1 and C/EBPβ (CCAAT/Enhancer-Binding Protein-β) are involved in the glutamate/nitric oxide/cyclic-guanosine 5′-monophosphate-mediated repression of gonadotropin-releasing hormone gene expression. Mol Endocrinol 14, 212–228

Chen L (1999) Combinatorial gene regulation by eukaryotic transcription factors. Curr Opin Struct Biol 9, 48–55

Cockerill PN, Shannon MF, Bert AG, Ryan GR, Vadas MA (1993) The granulocyte-macrophage colony-stimulating factor/interleukin 3 locus is regulated by an inducible cyclosporin A sensitive enhancer. Proc Natl Acad Sci USA 90, 2466–2470

Cockerill PN, Bert AG, Jenkins F, Ryan GR, Shannon MF et al. (1995) Human granulocyte-macrophage colony-stimulating factor enhancer function is associated with cooperative interactions between AP-1 and NFATp/c. Mol Cell Biol 15, 2071-2079

Crabtree GR (1999) Generic signals and specific outcomes: signalling through Ca^{2+}, calcineurin, and NF-AT. Cell 96, 611–614

Diamond MI, Miner JN, Yoshinaga SK, Yamamoto KR (1990) Transcription factor interactions: selectors of positive or negative regulation from a single DNA element. Science 249, 1266–1272

Ewing B, Green P (2000) Analysis of expressed sequence tags indicates 35000 human genes. Nat Genet 25, 232–234

Fickett JW, Hatzigeorgiou AG (1997) Eukaryotic promoter recognition. Genome Res 7, 861–878

Frech K, Danescu-Mayer J, Werner T (1997) A novel method to develop hghly specific models for regulatory units detects a new LTR in GenBank which contains a functional promoter. J Mol Biol 270, 674–687

Frech K, Quandt K, Werner T (1998) Muscle actin genes: a first step towards computational classification of tissue specific promoters. In Silico Biol 1, 29–38

Fukada T, Tonks NK (2001) The reciprocal role of Egr-1 and Sp family proteins in regulation of the PTP1B promoter in response to the p210 Bcr-Abl oncoprotein-tyrosine kinase. J Biol Chem 276, 25512–25519

Galson DL, Tsuchiya T, Tendler DS, Huang LE, Ren Y et al. (1995) The orphan receptor hepatic nuclear factor 4 functions as a transcriptional activator for tissue-specific and hypoxia-specific erythropoietin gene expression and is antagonized by EAR3/COUP-TF1. Mol Cell Biol 15, 2135–2144

GuhaThakurta D, Stormo GD (2001) Identifying target sites for cooperatively binding factors. Bioinformatics 17, 608–621

Huang X, Adams MD, Zhou H, Kerlavage AR (1997) A tool for analyzing and annotating genomic sequences. Genomics 46, 37–45

International Human Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. Nature 409, 860–921

Jain J, McCaffrey PG, Miner Z, Kerpola TK, Lambert JN et al. (1993) The T-cell transcription factor NFATp is a substrate for calcineurin and interacts with Fos and Jun. Nature 365, 352–355

Kel A, Kel-Margoulis O, Babenko V, 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

Kel OV, Romaschenko AG, Kel AE, Wingender E, Kolchanov NA (1995) A compilation of composite regulatory elements affecting gene transcription in vertebrates. Nucleic Acids Res 23, 4097–4103

Kel-Margoulis OV, Romashchenko AG, Kolchanov NA, Wingender E, Kel A (2000) COMPEL: a database on composite regulatory elements providing combinatorial transcriptional regulation. Nucleic Acids Res 28, 311–315

Lee HJ, Masuda ES, Arai N, Arai K, Yokota T (1995) Definition of cis-regulatory elements of the mouse interleukin-5 gene promoter. Involvement of nuclear factor of activated T cell-related factors in interleukin-5 expression. J Biol Chem 270, 17541–17550

Metz R, Ziff E (1991) The helix-loop-helix protein rE12 and the C/EBP-related factor rNFIL-6 bind to neighboring sites within the c-fos serum response element. Oncogene 6, 2165–2178

Mietus-Snyder M, Sladek FM, Ginsburg GS, Kuo CF, Ladias JA et al. (1992) Antagonism between apolipoprotein AI regulatory protein 1, Ear3/COUP-TF, and hepatocyte nuclear factor 4 modulates apolipoprotein CIII gene expression in liver and intestinal cells. Mol Cell Biol 12, 1708–1718

Mitchell PJ, Tjian R (1989) Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. Science 245, 371–378

Northrop JP, Ullman KS, Crabtree GR (1993) Characterization of the nuclear and cytoplasmic components of the lymphoid-specific nuclear factor of activated T cells (NFAT) complex. J Biol Chem 268, 2917–2923

Novina CD, Roy AL (1996) Core promoters and transcriptional control. Trends Genet 12, 351–355

Perier RC, Praz V, Junier T, Bonnard C, Bucher P (2000) The eukaryotic promoter database (EPD). Nucleic Acids Res 28, 302–303

Pilpel Y, Sudarsanam P, Church GM (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. Nat Genet 29, 1–7

Quandt K, Frech K, Karas H, Wingender E, Werner T (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. Nucleic Acids Res 23, 4878–4884

Rao A (1994) NF-ATp: a transcription factor required for the coordinate induction of several cytokine genes. Immunol Today 15, 274–281

Rao A, Luo C, Hogan PG (1997) Transcription factors of the NFAT family: regulation and function. Annu Rev Immunol 15, 707–747

Roest Crollius H, Jaillon O, Bernot A, Dasilva C, Bouneau L et al. (2000) Estimate of human gene number provided by genome-wide analysis using DNA Tetraodon nigroviridis DNA sequence. Nat Genet 25, 235–238

Scherf M, Klingenhoff A, Werner T (2000) Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. J Mol Biol 297, 599–606

Schwenger GT, Fournier R, Hall LM, Sanderson CJ, Mordvinov VA (1999) Nuclear factor of activated T cells and YY1 combine to repress IL-5 expression in a human T-cell line. J Allergy Clin Immunol 104, 820–827

Subramaniam N, Cairns W, Okret S (1998) Glucocorticoids repress transcription from a negative glucocorticoid response element recognized by two homeodomain-containing proteins, Pbx and Oct-1. J Biol Chem 273, 23567–23574

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ et al. (2001) The sequence of the human genome. Science 291, 1304–1351

Wagner A (1999) Genes regulated cooperatively by one or more transcription factors and their identification in whole eukariotic genomes. Bioinformatics 15, 776–784

Wang L, Wu Q, Qiu P, Mirza A, McGuirk M et al. (2001) Analyses of P53 target genes in the human genome by bioinformatic and microarray approaches. J Biol Chem 276, 43604–43610

Werner T (1999) Models for prediction and recognition of eukaryotic promoters. Mamm Genome 10, 168–175

Wingender E, Dietze P, Karas H, Knuppel R (1996) TRANSFAC: a database of transcriptional factors and their DNA binding sites. Nucleic Acids Res 24, 238–241

Wingender E, Kel AE, Kel OV, Karas H, Heinemeyer T et al. (1997) TRANSFAC, TRRD and COMPEL: towards a federated database system on transcriptional regulation. Nucleic Acids Res 25, 265–268

Wingender E, Chen X, Fricke E, Geffers R, Hehl R et al. (2001) The TRANSFAC system on gene expression regulation. Nucleic Acids Res 29, 281–283

Wolberger C (1998) Combinatorial transcription factors. Curr Opin Genet Dev 8, 552–559

Zhang DE, Hetherington CJ, Meyers S, Rhoades KL, Larson CJ et al. (1996) CCAAT enhancer-binding protein (C/EBP) and AML1 (CBF alpha2) synergistically activate the macrophage colony-stimulating factor receptor promoter. Mol Cell Biol 16, 1231–1240