Physiol Genomics 31: 374–384, 2007.
First published September 11, 2007; doi:10.1152/physiolgenomics.00085.2006.

CALL FOR PAPERS | *Comparative Genomics*

Identifying *cis*-regulatory elements by statistical analysis and phylogenetic footprinting and analyzing their coexistence and related gene ontology

Wei Shi,[1] Wanlei Zhou,[2] and Dakang Xu[3]

[1]*Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, Melbourne;* [2]*School of Engineering and Information Technology, Deakin University, Melbourne;* [3]*Centre for Cancer Research, Monash Institute of Medical Research, Monash University, Melbourne, Australia*

Submitted 10 May 2006; accepted in final form 4 September 2007

Shi W, Zhou W, Xu D. Identifying *cis*-regulatory elements by statistical analysis and phylogenetic footprinting and analyzing their coexistence and related gene ontology. *Physiol Genomics* 31: 374–384, 2007. First published September 11, 2007; doi:10.1152/physiolgenomics.00085.2006.—Discovery of *cis*-regulatory elements in gene promoters is a highly challenging research issue in computational molecular biology. This paper presents a novel approach to searching putative *cis*-regulatory elements in human promoters by first finding 8-mer sequences of high statistical significance from gene promoters of humans, mice, and *Drosophila melanogaster,* respectively, and then identifying the most conserved ones across the three species (phylogenetic footprinting). In this study, a conservation analysis on both closely related species (humans and mice) and distantly related species (humans/mice and Drosophila) is conducted not only to examine more candidates but also to improve the prediction accuracy. We have found 124 putative *cis*-regulatory elements and grouped these into 20 clusters. The investigation on the coexistence of these clusters in human gene promoters reveals that SP1, EGR, and NRF-1 are the dominant clusters appearing in the combinatorial combination of up to five clusters. Gene Ontology (GO) analysis also shows that many GO categories of transcription factors binding to these *cis*-regulatory elements match the GO categories of genes whose promoters contain these elements. Compared with previous research, the contribution of this study lies not only in the finding of new *cis*-regulatory elements, but also in its pioneering exploration on the coexistence of discovered elements and the GO relationship between transcription factors and regulated genes. This exploration verifies the putative *cis*-regulatory elements that have been found from this study and also gives new insight on the regulation mechanisms of gene expression.

transcription factor

BINDING OF TRANSCRIPTION FACTORS to *cis*-regulatory elements or transcription factor binding sites (TFBSs) are involved in transcriptional regulation of gene expression. Discovering *cis*-regulatory elements has been an important research challenge for some years (9, 30). The traditional approach for the search of *cis*-regulatory elements has been via the wet-lab experiment. This approach, however, can be expensive and is impractical on large numbers of genes. With the availability of genome sequences, the computational approach provides an alternate low-cost method to find *cis*-regulatory elements that can effectively deal with a large number of genes (4, 7, 9, 15, 24, 30).

In recent years, many computational approaches have been proposed to find putative *cis*-regulatory elements using diverse algorithms. To evaluate the accuracy of the algorithms, computationally discovered *cis*-regulatory elements are compared with the known TFBSs from public or propriety databases or published literatures. This type of evaluation, however, does not validate or associate the putative *cis*-regulatory elements with biological functions. In this study we take advantage of gene ontology (GO) annotations to validate the putative *cis*-regulatory elements. GO categories of transcription factors are compared with the GO categories of genes whose promoters contain *cis*-regulatory elements putatively bound by these transcription factors. Transcription factors and their regulated genes are assumed to share some common GO categories. Another disadvantage of the other studies is that they did not investigate the relationships between the putative *cis*-regulatory elements. The expression of a gene is usually not regulated by a single transcription factor, but by clusters of transcription factors that might bind to different *cis*-regulatory elements. Therefore by exploring the combinatorial regulation of gene expression, one can obtain a better understanding of the complex gene regulation machinery. This study will endeavor to determine the coexistence of putative *cis*-regulatory elements and will also take advantage of phylogenetic footprinting to improve the prediction accuracy in the search of *cis*-regulatory elements.

Phylogenetic footprinting has been demonstrated to be a very useful tool in the discovery of evolutionarily conserved *cis*-regulatory elements (13, 22, 29, 30). Three species, humans, mice, and Drosophila, will be compared to examine the conservation of putative *cis*-regulatory elements. In addition, this study will limit the search for putative *cis*-regulatory elements to the promoter regions immediate 5′ of transcription start sites (TSSs) because most known *cis*-regulatory elements are located in the proximal promoter region of genes (2, 25).

Due to the combinatorial nature of transcription regulation, the elucidation of coexistence of *cis*-regulatory elements is important to the understanding of the mechanisms regulating gene expression (3, 20, 28). This will shed light on how transcription factors are combined to regulate gene expression and will also contribute to the reconstruction of gene regulatory networks. In the past, there have been few studies conducted on the identification of coexisting *cis*-regulatory elements using computational approaches (20). This is partly due to the com-

Article published online before print. See web site for date of publication (http://physiolgenomics.physiology.org).

Address for reprint requests and other correspondence: *W. Shi, Bioinformatics Div., Walter & Eliza Hall Inst. of Medical Research, Melbourne, Australia (e-mail: shi@wehi.edu.au).

plexity of this problem and the lack of enough data. In this study we take advantage of the UCSC genome database to analyze a large set of gene promoters to extract the correlation information among clusters of putative *cis*-regulatory elements.

Past studies using GO have focused on the GO annotation for individual genes (7, 29, 31). Many genes have been annotated with gene ontology categories in the GO database (10). However, biological processes are usually implemented by a set of genes that interact with each other. Annotation to individual genes, therefore, cannot be used to interpret the complex interactions among them correctly. This study correlates the GO categories of transcription factors with the GO categories of putatively regulated genes by taking advantage of gene annotation in GO database. To our knowledge, this is the first study to report the matching of the biological functions of transcription factors with the biological functions of putatively regulated genes. This also serves as a validation of the discovered putative *cis*-regulatory elements because transcription factors are correlated to the putatively regulated genes through these elements in this study.

METHODS

Data acquisition. Gene promoters of length 1,000 upstream from the TSSs of human genes, mouse genes, and *Drosophila melanogaster* genes were downloaded from the UCSC genome database (http://genome.ucsc.edu/; human version, May 2004; mouse version, September 2005; *Drosophila melanogaster* version, April 2004). The original data sets contained 20,647, 16,052, and 15,464 gene promoters of humans, mice, and Drosophila separately. After deleting redundant (the redundancy was brought about by multiple RefSeq mRNAs corresponding to a same gene) and incomplete promoters (<1,000 bp long), we were left with 17,407 human promoters, 15,644 mouse promoters, and 12,841 Drosophila promoters.

Calculation of cis-regulatory element factor. For each species, all gene promoters were aligned to their TSSs. These aligned promoters were then divided into 50 bins, each of which contained 20 bp from each promoter. The search for the putative *cis*-regulatory elements was allowed to cross bin boundaries.

All possible 8-mer sequences (65,536 sequences) were investigated for the gene promoter set from each species because only single strand DNA were available for each gene in the dataset. To calculate the statistical significance of an 8-mer sequence s, we first calculated its number of occurrences in each bin. This number was calculated as the sum of each occurrence number of s appearing in each promoter at that bin. For a bin b in a promoter p, the occurring of s in b meant s was a substring of p and the first letter of s was located in bin b. So if s fell on a bin boundary, the position of s's first letter would determine which bin it belonged to. A promoter was counted multiple times if s appeared multiple times in that promoter. The *Cis*-regulatory element factor (CRE_f) of s was then calculated as:

$$\mathrm{CRE}_f = \frac{x_{\max} - \bar{x}}{x_3 - x_1} \quad (1)$$

$x_{\max}$ was the maximal occurrence number among all occurrence numbers of s in all bins. $\bar{x}$ was the median value of all occurrence numbers of s in all bins. x_3 and x_1 were the third quartile and first quartile of occurrence numbers of s in all bins, respectively.

Calculation of Z-score. For each species, 1,000 random sets of gene promoters were generated to evaluate the likelihood that an 8-mer sequence obtained its statistical significance (CRE_f) by chance. We used the seventh-order Markov model to generate random data sets (9). Each random set for humans contained 17,407 gene promoters of length 1,000. Each promoter started with a random 7-mer sequence, and the next base was determined randomly under the condition that frequencies of 8-mer sequences in the real human data set were maintained. This process was repeated until the entire promoter sequence was completed. The same procedure was applied to gene promoters from mice and Drosophila to generate their background sets of promoters. A CRE_f value of an 8-mer sequence was calculated on each random set of gene promoters. The Z-score of an 8-mer sequence, representing the level of confidence that the statistical significance of this 8-mer sequence was not obtained randomly, was then calculated based on its CRE_f value on the real set of promoters and CRE_f values on 1,000 random sets of promoters using the following equation:

$$Z(s) = \frac{\mathrm{CRE}_f(s)_{\mathrm{real}} - \mathrm{CRE}_f(s)_{\mathrm{median}}}{\mathrm{CRE}_f(s)_3 - \mathrm{CRE}_f(s)_1} \quad (2)$$

$\mathrm{CRE}_f(s)_{\mathrm{real}}$ denoted the CRE_f value at s in the real set of promoters. $\mathrm{CRE}_f(s)_{\mathrm{median}}$ denoted the median of CRE_f values at s in the 1,000 background sets. $\mathrm{CRE}_f(s)_3$ and $\mathrm{CRE}_f(s)_1$ denoted the third quartile and the first quartile of CRE_f values at s in the 1,000 background sets, respectively.

Calculation of P value. To test the statistical significance of the number of common GO categories found between GO categories for transcription factors and GO categories for putatively regulated genes in Table 3, we calculated the probability of having m members in the intersection of two subsets taken from the entire set of GO categories.

For two subsets containing s_1 and s_2 members, respectively, the number of possible combinations in which they have m members in common was

$$C^{\mathrm{n}}_{\mathrm{s}_1} C^{\mathrm{s}_1}_{\mathrm{m}} C^{\mathrm{n}-\mathrm{s}_1}_{\mathrm{s}_2-\mathrm{m}}$$

where n was the total number of GO categories in the GO database (n was equal to 20,458 for the Feb. 1, 2006 version of the database that was used in this study). The total number of combinations was

$$C^{\mathrm{n}}_{\mathrm{s}_1} C^{\mathrm{n}}_{\mathrm{s}_2}$$

The probability of having m members in common between the two subsets was

$$P = \frac{C^{\mathrm{n}}_{\mathrm{s}_1} C^{\mathrm{s}_1}_{\mathrm{m}} C^{\mathrm{n}-\mathrm{s}_1}_{\mathrm{s}_2-\mathrm{m}}}{C^{\mathrm{n}}_{\mathrm{s}_1} C^{\mathrm{n}}_{\mathrm{s}_2}} = \frac{C^{\mathrm{s}_1}_{\mathrm{m}} C^{\mathrm{n}-\mathrm{s}_1}_{\mathrm{s}_2-\mathrm{m}}}{C^{\mathrm{n}}_{\mathrm{s}_2}} \quad (3)$$

RESULTS

Identifying putative cis-regulatory elements and grouping them into clusters. Putative promoter regions 1,000 bp of 5′ genomic sequence immediately upstream of TSS of genes from humans, mice, and Drosophila species were analyzed to extract 8-mer sequences that were not only statistically significant in each species but commonly conserved across these species. Using closely related species (humans and mice) and distantly related species (humans/mice and Drosophila), we found more candidate *cis*-regulatory elements and generated fewer false positives than by using either closely related species or distantly related species alone. In total, 17,407 human promoters, 15,644 mouse promoters, and 12,841 Drosophila promoters were analyzed.

Figure 1 outlined the process of identifying *cis*-regulatory elements using a combination of statistical analysis and phylogenetic footprinting techniques. A simple statistical algorithm was first applied to the set of promoters for each species. The CRE_f and Z-score, which represented the statistical significance of an 8-mer sequence and the likelihood of attaining that significance by chance, were calculated for each 8-mer se-

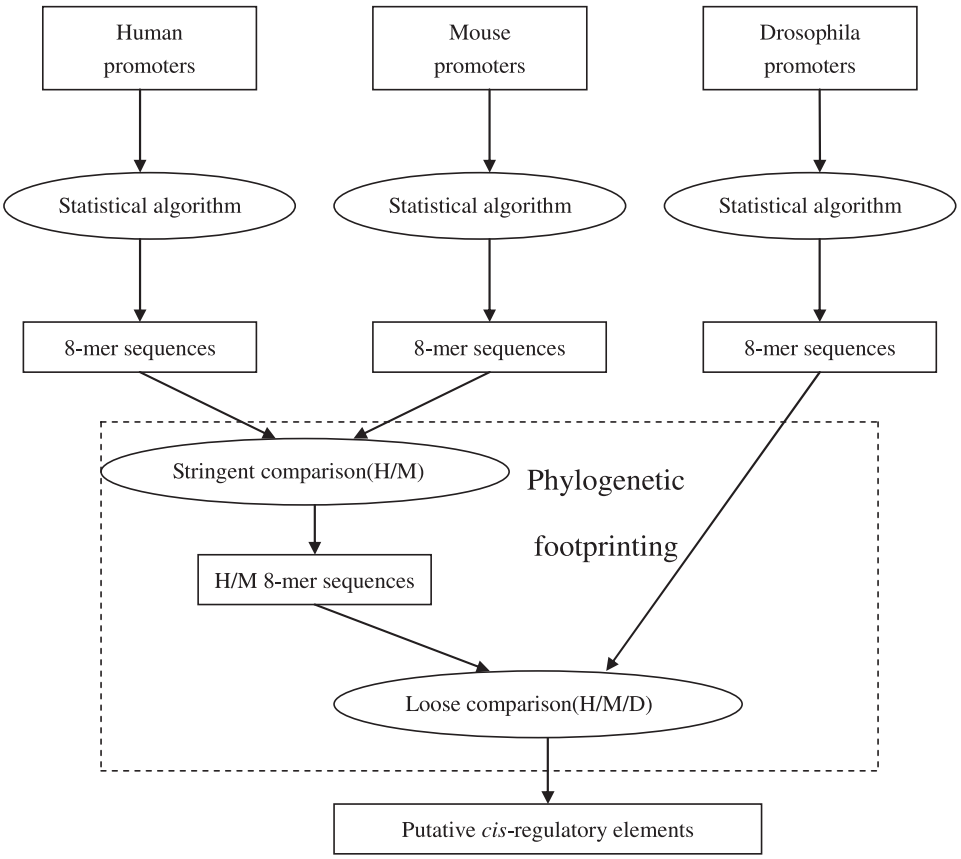

Fig. 1. Flowchart of *cis*-regulatory element discovery process.

quence for each species (see METHODS for details). The 8-mer sequences meeting the selection criteria for each species were then selected for the next step of processing.

Three criteria were then used to filter 8-mer sequences in this statistical analysis step: *1*) $CRE_f \geq 3$, *2*) $Z \geq 2$, and *3*) MaxBin ≥ 20; MaxBin is the maximal occurrence number of any 8-mer sequence in all bins in the real promoter set. The CRE_f value describes the strength of a peak in the frequency distribution of an 8-mer sequence (see METHODS). If an 8-mer sequence has a high abundance at a particular bin (e.g., bin 48 for TATA box) and low abundance at other bins, there would be a strong peak shown at that particular bin. It was our assumption that an exceptional peak at a particular location implied some biological functions. This assumption was confirmed by the known sites, an example of which is the binding of TATA box binding protein (TBP) to the core promoter regions of TATA-containing genes. Here, the binding site is located at -25 to -30 bp upstream from TSSs. Strong peaks were detected in bin 48 (-20 to -40 bp upstream from TSSs) for TATA box elements in this study (see Table 1). A *Z*-score was used to calculate the likelihood of obtaining a CRE_f value by chance for an 8-mer sequence. The calculation of *Z*-scores was based on 1,000 randomly generated promoter sets for each species (see METHODS). Any 8-mer sequences that had very low abundance in all bins were filtered out with MaxBin values. This is because these sequences were found in a very small number of genes, and it would be hard to tell if such 8-mer sequences were statistically or biologically significant in such a high-throughput analysis.

After filtering with these three criteria, there remained 926, 1,176 and 558 8-mer sequences for humans, mice, and Drosophila, respectively. These 8-mer sequences were then fed into a phylogenetic footprinting analysis module for comparison with their homologs in other species. A two-stage comparison was conducted: first, 8-mer sequences selected from humans were stringently compared with 8-mer sequences selected from mice. Only identical 8-mer sequences from both humans and mice were selected for further analysis. Secondly, these 8-mer sequences were loosely compared with 8-mer sequences selected from Drosophila. This loose comparison allowed insertion, deletion, and substitution of up to one base to simulate the mutation of *cis*-regulatory elements in the course of evolution (see METHODS for details). After these analyses, 124 putative *cis*-regulatory elements were found and grouped into 20 clusters as shown in Table 1. The last five clusters (clus16 to clus20) were new and were predicted by this study. No known transcription factors bind to the elements in these clusters. The elements in each new cluster were grouped together according to their similarities.

These elements only accounted for ~0.2% of all studied 8-mer sequences; however, they appear in the majority of gene promoters in each species (89.4% in humans, 86.6% in mice, and 82.1% in Drosophila). The calculation of the statistical significance of these 20 clusters in the area of 1,000–2,000 bp upstream from TSSs in each species revealed that up to 80% of 124 elements did not show statistical significance at all, and the remaining 20% showed much lower significance than in the area of 1,000 bp from TSSs (data not shown).

Coexistence of clusters of cis-regulatory elements. Investigations into the coexistence of *cis*-regulatory elements in gene promoters will shed light on the understanding of the combi-

Table 1. *124 putative cis-regulatory elements grouped in 20 clusters*

ID	Cluster	8-mer	*Z*	CRE_f	MaxBin	Pos	Total	ID	Cluster	8-mer	*Z*	CRE_f	MaxBin	Pos	Total
1	TATA	tataaaag*†	22.9	19	64	48	444		NRF-1 (con't)	tgcgcatg*†‡	4.4	5.4	42	46	381
		tataaagg*†	11.1	11.3	49	48	307			gcgcgtgc	3.3	4.1	45	48	601
		cctataaa*†	10.4	10.7	36	48	242			ctgcgcat	2.8	4.3	25	46	262
		ataaaagg*†	8.5	8.2	49	48	438			gtgcgcag	2.7	4.1	34	47	392
		tatataag*	8.4	9	30	48	211			ggctgcgg	2.7	3.3	76	48	1,234
		ggtataaa†	8.3	7.5	26	47	229			gcgcagtg	2.5	3.7	31	48	515
		ctataaag*†	7.2	8	36	48	267			cggctgcg	2.5	3.6	54	48	663
		atataagg†‡	6.4	7.3	24	48	168			gctgcggc†	2.5	3.4	58	48	919
		ctataaaa†	6.4	6.6	41	48	444			tggctgcg	2.3	3.5	32	47	523
		ataaaagc†	6.4	6.6	40	48	421			gcacgcgc†	2.2	3.4	43	47	630
		gtataaaa†	5.1	6.3	30	48	310			gcaggcgc†	2.2	3.3	94	48	1,177
		gtatataa†	4.4	5.7	21	47	233			agctgcgc	2	3.3	27	48	391
		taaaaggc*†	4.3	5.1	26	48	328	7	CRE	tgacgtca*†‡	11.9	10.1	87	47	641
		ataaaggg†	4.2	5.5	27	48	300			gacgtcag†‡	9.6	10	34	47	297
		ctatataa†	4.2	5.5	26	48	221			tgtgacgt*†	6.8	7.7	25	47	182
		gctataaa*†	4.2	5	25	47	276			ctgacgtc*†	5.6	6	27	48	282
		taaaaggg†	2.3	3.8	26	48	387	8	PAX-3	gacgtcac†	4.5	5.2	51	47	402
2	CCAAT	gccaatcg*	10.2	9.3	30	46	242			cgtgacgt*†	4	6	25	47	182
		ccattggc†‡	7.2	7.2	49	46	501			acgtcacg†	3.4	4.8	25	47	172
		gattggcc†‡	7.0	7.1	54	46	442	9	C(A/T)CCC	acgcccac	2.9	4.4	27	46	306
		cgattggc†‡	5.7	6.8	29	46	249		box	ctcccttc†	2.8	3.6	43	38	1,101
		ggccaatg*	5.7	6.1	36	45	406			tcctccct‡	2.5	3.3	66	47	1,437
		ctattggc†	4.8	6	21	45	205			cctccctc*†	2.2	3.0	113	48	2,545
		cggccaat*†	4.7	5.9	33	46	313			gctcctcc	2.1	3.1	69	45	1,289
		ttggctgc†	4.5	5.2	34	47	487	10	Y box	gctgattg†	9.2	8.8	40	46	336
		cattggct†‡	4.4	5.5	38	45	408	11	LSF	ggtggagc†‡	4	4.8	34	47	592
		tggccaat†	4.3	5	27	44	419			ggtggcgc†‡	2.6	3.6	31	46	696
		tctgattg†	4.2	5.4	32	46	390			ctggcgct‡	2.3	3.5	21	46	363
		cattggcc†‡	4.1	5.3	32	46	393	12	EGR	gcgcggcg	3.7	4.1	128	48	1,489
		ctcattgg†	3.8	5	31	45	354			tggcgcgg†	3.3	4.3	32	45	435
		caatcaga*†	3.4	4.4	28	46	403			cgcgagag†	3.2	4.7	27	47	283
		attggcca†	2.2	3.5	25	46	437			ctggcgcg	2.9	4	27	47	447
3	SP1	ggaggcgg†	6.2	5.4	180	47	3,132			gcgggcac	2.9	4	24	46	452
		gaggcggg	5.8	5.1	146	46	2,529			ggcgcggc†‡	2.8	3.3	106	48	1,431
		cggcggcg	5.7	5.0	290	48	3,072			gcggcgcg	2.6	3.3	120	48	1,534
		ggcggcgg†	4.5	4.5	332	48	3,935			cggcgcgc	2.4	3.4	78	47	1,086
		gcggcggc†	4.4	4.4	350	48	3,771			cggcgcgg†	2.4	3.3	99	48	1,342
		gcaggcgg	4.2	4.5	67	47	1,039			gacgcgcg	2.3	3.9	28	46	352
		cggccgcg	4.2	3.8	111	48	1,170			aggcgcgc	2.3	3.4	43	46	730
		gctgctgc	3.7	4.2	62	48	1,101			gcggcagc†	2.1	3.2	62	48	937
		gcgcgccg	3.3	4.0	71	47	1,014			ggcgcgcg	2.1	3.0	116	47	1,650
		cgaggcgg†	3.3	4.0	46	48	985			gggcgcgg†‡	2.1	3.0	113	45	2,041
		cggcgcag	3	3.9	47	48	575			ggcgcgcc†	2.1	3.0	53	44	858
		gaggcggt	2.9	4	30	47	400	13	HSF	gtttccgg†	10.3	8.8	29	48	247
		ggcggcgc†	2.9	3.7	133	48	1,645			tttccggc†	7.6	8.3	28	48	243
		ggaggagc†	2.8	3.5	77	46	1,341			ttccggtt†	7.6	8.3	28	48	200
		gaggagcc†	2.6	3.2	47	46	997	14	MIF	gctgcgcg†	4.7	5.1	58	48	617
		gcggtgcc†	2.5	3.8	37	46	488			tgcgcggc†	2.9	3.9	37	48	461
		gcgcgctg	2.5	3.6	54	48	627			ctgcgcgg†	2.3	3.5	49	47	570
		ggcagggc†	2.5	3.2	63	46	1,352			tgcgcgcg†	2.2	3.3	52	48	690
		gccgccgc†	2.4	3.2	227	48	3194	15	TCF11/ MAFG	gaagcggc†	3.1	4.4	33	48	376
		cgccgccg†	2.3	3.2	194	48	2,589			gaggcggc†	2.6	3.3	85	48	1,133
		ggcaggcg	2.2	3.3	43	47	916			gcagcggc†	2.4	3.4	68	49	1,129
4	USF	cacgtgac*†‡	3.1	4.2	46	47	439	16	Clus16	tgtctctc†	2.5	3.7	23	26	584
5	ETS	cggaagcg*†‡	9.8	9.2	60	48	424	17	Clus17	gcgctgct	4.4	5.5	34	48	397
		gaagcgga†	5.9	6.8	32	48	361			gcggctgc	3.9	4.6	80	48	1,028
		ggaagcgg†	4.9	5.3	53	48	698			gcggtgac†	3.2	4.3	21	48	275
6	NRF-1	gcatgcgc†	8.2	7.6	61	47	749	18	Clus18	cgctgatt	6.4	7.3	24	46	167
		gcgcatgc*†‡	7.1	6.8	76	48	756			agctgctg†	4.0	4.6	33	48	739
		gcgcctgc*	6.3	5.8	91	48	1,159			tgctgctg	3.9	4.4	38	48	856
		cgcctgcg*	6.2	5.6	95	48	983	19	Clus19	gtgggcag†	2.4	3.4	32	35	796
		gcctgcgc*	5.5	5.4	105	48	1,137	20	Clus20	ctctcgcg	2.7	3.8	22	45	273

*8-Mer sequences that were also found by FitzGerald et al. (9); †8-mer sequences that matched motifs discovered by Xie et al. (30); ‡8-mer sequences that matched TRANSFAC motifs inferred by Xie et al. (30). An 8-mer sequence matches a motif sequence if it is a subsequence of an instance of the motif or if an instance of the motif is a subsequence of the 8-mer sequence, and the matching part should not have more than 2 consecutive *n*'s (*n* is the degenerate letter representing a, t, g, or c).

natorial regulation mechanism of gene expression. Table 2 shows the statistical results of coexistence of up to five clusters of *cis*-regulatory elements in human gene promoters. For each cluster (e.g., TATA), we have shown combinations of clusters (e.g., TATA, SP1) appearing in >20% of gene promoters that contain one or more putative *cis*-regulatory elements from that seeding cluster (TATA). Results were obtained by first fixing a seeding cluster and then looking for the clusters that could be combined with it. For example, *cluster C* would be combined with the seeding *cluster S* if at least 20% of human gene promoters that contained one or more elements from *S* also contained one or more elements from *C*. A third cluster, *T,* would be combined with *C* and *S*, if at least 20% of human gene promoters containing one or more elements from *S* contained not only one or more elements from *C* but also one or more elements from *T*.

It was observed that coexistence of two or three clusters accounted for the majority of cluster coexistence. SP1, EGR, and NRF-1 were the dominant clusters resulting from the combinatorial combination of up to five clusters. The two bases, "G" and "C", were the dominant bases in *cis*-regulatory elements from these three clusters, and consecutive "GC" was found in many of these elements. This implies that the CpG island is an important part of human gene promoters. It was also observed that each of the 20 clusters coexisted with at least one other cluster, indicating that coexistence of *cis*-regulatory elements in human promoters is a very universal phenomenon. The bias of combined clusters of *cis*-regulatory elements discovered in this study may contribute to the construction of the gene regulatory network.

GO analysis. We examined the relationship between GO categories of transcription factors binding to the putative *cis*-regulatory elements from a cluster and GO categories of genes (regulated genes) whose promoters contained at least one of these elements from the same cluster. We hypothesized that if found *cis*-regulatory elements and the clusters they were grouped into were correct, the GO categories of transcription factors binding to the cluster of elements should match the GO categories of corresponding regulated genes (7, 10, 29, 31).

For each cluster of elements, we first determined RefSeq names of human genes whose promoters contained one or more elements from that cluster, and then converted RefSeq names to HUGO symbols using MatchMiner (5). HUGO symbols of genes were then submitted to High-Throughput GOMiner to mine the GO categories (32). The false discovery rate threshold was set to be 0.1. Two files were analyzed by High-Throughput GOMiner: one was the total genes file including the list of all the genes, the other was the changed genes file, which was a zip file including 20 lists of genes for all the clusters. Names of genes in these two files were HUGO symbols. The result from High-Throughput GOMiner was then clustered by Genesis using hierarchical clustering and Pearson correlation (23). The result of clustering is shown in Fig. 2. The number of GO categories for the regulated genes for each cluster is shown in Table 3. Details of these GO categories can be seen in Supplemental Table S1 (Additional File 1).[1]

Names (HUGO symbols) of genes encoding transcription factors in each cluster were submitted to GOMiner to search GO categories for transcription factors (32). The number of GO categories found for the transcription factors in each cluster is shown in Table 3 as well. Details can be seen in Supplemental Table S2 (Additional File 2).

Results showed matches in 12 out of 15 clusters that have known transcription factors. No matches were found in three clusters with known transcription factors: TATA, Y box, and LSF. For TATA, TBP has the main function of initiating gene expression, which in fact bears no relevance with the functions of regulated genes. For Y box and LSF, only a small number of GO categories resulted for either transcription factors or regulated genes. This led to no correlation between them. This was probably brought about by the inaccuracy of the software (High-Throughput GOMiner and Genesis) or by the lack of knowledge on the function of transcription factors or regulated genes.

Five new clusters of putative *cis*-regulatory elements appeared in a large number of human gene promoters; however, only a few GO categories were able to be assigned to them. Also no GO categories were found for clus16, clus18, and clus19. This suggests that these elements, which have high statistical significance, could have some biological functions that have not yet been recognized.

In summary, common GO categories between transcription factors and regulated genes have been shown in the majority of clusters with known transcription factors (see Table 3). This demonstrates the putative *cis*-regulatory elements discovered in this study and their groupings are biologically meaningful. Details of the common GO categories can be seen in Supplemental Table S3 (Additional File 3).

Comparison with other lists of putative cis-regulatory elements or motifs. Of the 124 putative *cis*-regulatory elements found in this study, only 24 sequences (19.4%) were also identified by FitzGerald et al. (9). Our statistical algorithm is similar to FitzGerald's algorithm from the point of view of splitting promoters into bins and calculating the statistical significance of 8-mer sequences according to their occurrence frequencies in different bins. However, three key differences led to very different results. The first is the median value used in this method rather than the mean value used in the research by FitzGerald et al. Median value is believed to be better than mean value in finding wild values in a sequence of numbers in statistics. Thus putative *cis*-regulatory elements discovered in this study were potentially more specific; The second is a larger set of human promoters used in this study (17,407 human promoters in this study and only 13,010 human promoters in FitzGerald et al.'s study). For obvious reasons, prediction accuracy is improved by investigating more genes. The last and the most important difference is the phylogenetic footprinting technique used in this study. This technique helped to eliminate 8-mer sequences that had high statistical significance in human promoters but were not conserved across the three species of humans, mice, and Drosophila. Conservation is an inherent feature of *cis*-regulatory elements, and we believe that conservation analysis could greatly reduce false positives in the discovery of putative *cis*-regulatory elements.

Out of 124 elements, 83 (66.9%) matched the motifs found by Xie et al. (30). In their study, Xie et al. searched motifs of diverse lengths. A *cis*-regulatory element identified in our study would be deemed as a match with a motif identified by Xie et al. if the *cis*-regulatory element is one instance of the

[1] The online version of this article contains supplemental material.

Table 2. *Coexistence of clusters of cis-regulatory elements*

		Num = 2			Num = 3			Num = 4			Num = 5		
	Total (peak)	Clusters	#	%	Clusters	#	%	Clusters	#	%	Clusters	#	%
TATA	837	(TATA, TATA)	392	46.8									
		(TATA, SP1)	175	20.9									
CCAAT	1,680	(CCAAT, CCAAT)	695	41.4	(CCAAT, SP1, SP1)	353	21.0						
		(CCAAT, SP1)	607	36.1									
		(CCAAT, NRF-1)	354	21.1									
		(CCAAT, EGR)	407	24.2									
SP1	5,813	(SP1, SP1)	3,249	55.9	(SP1, SP1, SP1)	1,841	31.7	(SP1, SP1, SP1, SP1)	1,221	21.0			
		(SP1, NRF-1)	1,661	28.6	(SP1, SP1, EGR)	1,454	25.0						
		(SP1, EGR)	2,220	38.2	(SP1, EGR, EGR)	1,286	22.1						
USF	252	(USF, SP1)	119	47.2	(USF, SP1, SP1)	68	27.0						
		(USF, NRF-1)	84	33.3	(USF, SP1, NRF-1)	51	20.2						
		(USF, EGR)	73	29.0									
ETS	469	(ETS, SP1)	216	46.1	(ETS, SP1, SP1)	127	27.1						
		(ETS, ETS)	185	39.4	(ETS, SP1, ETS)	94	20.0						
		(ETS, NRF-1)	127	27.1	(ETS, SP1, EGR)	95	20.3						
		(ETS, EGR)	138	29.4									
		(ETS, TCF11)	115	24.5									
NRF-1	3,314	(NRF-1, SP1)	1,661	50.1	(NRF-1, SP1, SP1)	958	28.9						
		(NRF-1, NRF-1)	1,517	45.8	(NRF-1, NRF-1, SP1)	754	22.8						
		(NRF-1, EGR)	1,152	34.8	(NRF-1, SP1, EGR)	761	23.0						
					(NRF-1, NRF-1, NRF-1)	757	22.8						
CRE	540	(CRE, SP1)	225	41.7	(CRE, SP1, SP1)	126	23.3						
		(CRE, NRF-1)	135	25.0									
		(CRE, CRE)	228	42.2									
		(CRE, PAX-3)	222	41.1									
		(CRE, EGR)	138	25.6									
PAX-3	346	(PAX-3, SP1)	150	43.4	(PAX-3, SP1, SP1)	82	23.7						
		(PAX-3, NRF-1)	80	23.1	(PAX-3, SP1, CRE)	101	29.2						
		(PAX-3, CRE)	222	64.2	(PAX-3, CRE, CRE)	98	28.3						
		(PAX-3, PAX-3)	107	30.9	(PAX-3, PAX-3 CRE,)	95	27.5						
		(PAX-3, EGR)	90	26.0									
C(A/T)CCC	1,464	(C(A/T)CCC, SP1)	499	34.1									
		[C(A/T)CCC, C(A/T)CCC]	407	27.8									
Y box	150	(Y box, CCAAT)	57	38.0	(Y box, CCAAT, SP1)	33	22.0						
		(Y box, SP1)	67	44.7	(Y box, CCAAT, Clus18)	33	22.0						
		(Y box, NRF-1)	38	25.3	(Y box, SP1, SP1)	34	22.7						
		(Y box, EGR)	39	26	(Y box, SP1, Clus18)	49	32.7						
		(Y box, Clus18)	77	51.3	(Y box, EGR, Clus18)	31	20.7						
LSF	528	(LSF, SP1)	227	43.0	(LSF, SP1, SP1)	122	23.1						
		(LSF, NRF-1)	134	25.4									
		(LSF, EGR)	160	30.3									
EGR	3,485	(EGR, SP1)	2,220	63.7	(EGR, SP1, SP1)	1,454	41.7	(EGR, SP1, SP1, SP1)	953	27.3			
		(EGR, NRF-1)	1,152	33.1	(EGR, SP1, NRF-1)	761	21.8	(EGR, EGR, SP1, SP1)	890	25.5			
		(EGR, EGR)	1,819	52.2	(EGR, EGR, SP1)	1,286	36.9	(EGR, EGR, EGR, SP1)	717	20.6			
					(EGR, EGR, EGR)	932	26.7						
HSF	310	(HSF, SP1)	99	31.9									
		(HSF, NRF-1)	74	23.9									
		(HSF, EGR)	62	20.0									

Continued

Table 2.—*Continued*

		Num = 2			Num = 3			Num = 4			Num = 5		
	Total (peak)	Clusters	#	%	Clusters	#	%	Clusters	#	%	Clusters	#	%
MIF	852	(MIF, SP1)	472	55.4	(MIF, SP1, SP1)	293	34.4	(MIF, SP1, SP1, SP1)	175	20.5			
		(MIF, NRF-1)	462	54.2	(MIF, SP1, NRF-1)	268	31.5	(MIF, SP1, SP1, EGR)	171	20.1			
		(MIF, EGR)	379	44.5	(MIF, SP1, EGR)	246	28.9						
		(MIF, MIF)	253	29.7	(MIF, NRF-1, NRF-1)	223	26.2						
					(MIF, NRF-1, EGR)	216	25.4						
					(MIF, EGR, EGR)	199	23.4						
TCF11/MAFG	1,004	(TCF11, SP1)	771	76.8	(TCF11, SP1, SP1)	564	56.2	(TCF11, SP1, SP1, SP1)	418	41.6	(TCF11, SP1, SP1, SP1, SP1)	328	32.7
		(TCF11, NRF-1)	288	28.7	(TCF11, SP1, NRF-1)	224	22.3	(TCF11, SP1, SP1, EGR)	324	32.3			
		(TCF11, EGR)	469	46.7	(TCF11, SP1, EGR)	403	40.1	(TCF11, SP1, EGR, EGR)	242	24.1	(TCF11, SP1, SP1, SP1, EGR)	252	25.1
					(TCF11, EGR, EGR)	272	27.1						
											(TCF11, SP1, SP1, EGR, EGR)	203	20.2
Clus16	106	(Clus16, SP1)	31	29.2									
		(Clus16, NRF-1)	24	22.6									
Clus17	733	(Clus17, SP1)	443	60.4	(Clus17, SP1, SP1)	293	40.0	(Clus17, SP1, SP1, SP1)	200	27.3			
		(Clus17, NRF-1)	346	47.2	(Clus17, SP1, NRF-1)	227	31.0	(Clus17, SP1, SP1, NRF-1)	158	21.6			
		(Clus17, EGR)	286	39.0	(Clus17, SP1, EGR)	214	29.2	(Clus17, SP1, SP1, EGR)	149	20.3			
					(Clus17, NRF-1, NRF-1)	198	27.0						
					(Clus17, EGR, EGR)	165	22.5						
Clus18	436	(Clus18, SP1)	268	61.5	(Clus18, SP1, SP1)	122	28.0						
		(Clus18, NRF-1)	93	21.3									
Clus19	173	(Clus19, SP1)	67	38.7	(Clus19, SP1, SP1)	36	20.8						
		(Clus19, NRF-1)	37	21.4	(Clus19,SP1, EGR)	30	17.3						
		(Clus19, EGR)	43	24.9									
Clus20	128	(Clus20, SP1)	52	40.6	(Clus20, SP1, SP1)	27	21.1						
		(Clus20, NRF-1)	32	25									
		(Clus20, EGR)	41	32.0									

The 2nd column is the number of human gene promoters containing any of the putative *cis*-regulatory elements from the corresponding cluster. The 3rd column (Num = 2) provides the information on 2 clusters that coexist in gene promoters accounting for at least 20% of gene promoters that contain 1 or more elements from the cluster shown in the 1st column at the same row (seeding cluster). Names of the 2 clusters, number of gene promoters containing 1 or more elements from each of the 2 clusters, and percentage of these gene promoters in gene promoters that contain 1 or more elements from the cluster shown in the 1st column at the same row, are shown in the 3 subcolumns of the 2nd column, respectively. The remaining columns show the similar information for the combination of 3, 4, and 5 clusters. Note that *cis*-regulatory elements coexisting in gene promoters can overlap with each other.

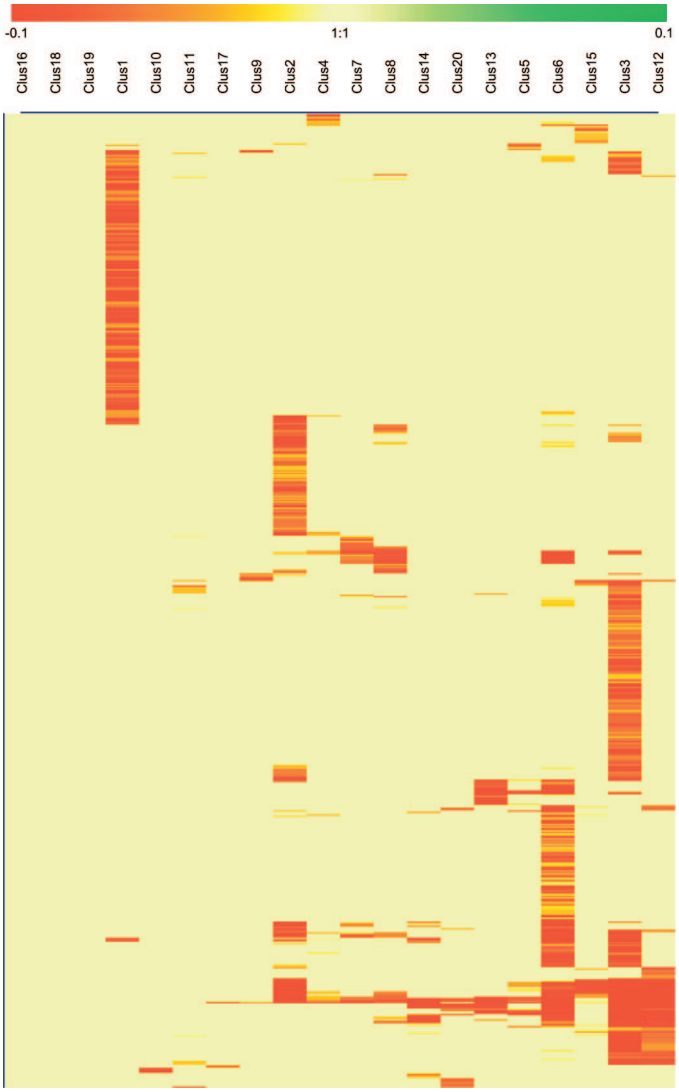

Fig. 2. Hierarchical clustering of Gene Ontology (GO) categories of regulated genes by Genesis. Rows are GO categories. Columns are gene files for 20 clusters. Each file contains the list of genes found for the corresponding cluster. Hierarchical clustering algorithm and Pearson correlation were used. The input to Genesis was the CIM file generated by High-Throughput GOMiner.

motif or an instance of the motif is part of the element. Xie et al. conducted a comparative analysis of the human, mouse, rat, and dog genomes, which are more closely related in the course of evolution compared with the species selected by this study (humans vs. Drosophila). Comparison between more distantly related species could increase the chance of finding *cis*-regulatory elements having more fundamental biological roles and of eliminating false positives. The combination of comparisons between both closely related species and distantly related species will not only provide more candidates for examination but also set a more stringent criterion for eliminating those 8-mer sequences only having the statistical significance, which were difficult to filter out by comparing only closely related species. Thus the putative *cis*-regulatory elements discovered in this study should have more fundamental biological functions and be more accurate than Xie et al.'s finding, even though the number of motifs here was less than that found by Xie et al. (30).

Eighteen out of 124 elements (14.5%) matched the TRANSFAC motifs inferred by Xie et al. (30). Only a small portion of the putative *cis*-regulatory elements we found matched the TRANSFAC motifs. One of the reasons is perhaps the inaccuracy of these TRANSFAC motifs assembled by Xie et al., which might not match well with the TRANSFAC database. Another reason is possibly the stringent criteria for selecting candidate *cis*-regulatory elements in this study, which might filter out some real *cis*-regulatory elements.

Put together, 73.4% of 124 elements discovered in this study were also found by other studies. Four putative *cis*-regulatory elements in the five new clusters matched the motifs identified in Xie et al.'s (30) study, which also found no known transcription factors.

DISCUSSION

This paper focuses on the discovery of putative *cis*-regulatory elements 8 mer in length, which will normally consist of a core motif sequence (e.g., "tata" for TATA box) and several flanking bases. The specificity will be lost if the putative *cis*-regulatory elements under investigation are too short; however it will be computationally intractable if they are too long. FitzGerald et al. (9) have identified 156 putative *cis*-regulatory elements 8 mer in length for human promoters, but it is not clear if these elements are conserved in the course of evolution. Additionally they did not validate these elements by examining the biological functions of transcription factors and regulated genes as undertaken in this study (9).

Table 3. *GO categories for transcription factors and regulated genes in each cluster*

	# of GO Categories for TFs	# of GO Categories for Regulated Genes	# of Common GO Categories	*P*
1	59	211	0	0.54
2	73	149	23	1e-72
3	54	275	24	1e-70
4	56	29	8	1e-32
5	66	28	12	1e-53
6	61	205	31	1e-105
7	138	38	7	1e-19
8	68	50	9	1e-30
9	50	9	1	1e-5
10	32	4	0	0.99
11	5	29	0	0.99
12	87	81	22	1e-78
13	96	34	10	1e-36
14	37	29	3	1e-12
15	46	45	14	1e-62
16	N/A	0	N/A	N/A
17	N/A	3	N/A	N/A
18	N/A	0	N/A	N/A
19	N/A	0	N/A	N/A
20	N/A	20	N/A	N/A

Number of gene ontology (GO) categories for transcription factors (*column 2*) was obtained by directly searching GO database using the transcription factor names. Number of GO categories for regulated genes (*column 3*) was obtained by using High-Throughput GoMiner (30), which searched unique GO categories for each cluster. The Feb. 1, 2006 version of GO database was used in this study, which contained 20,458 GO categories. *P* value (last column) gave the probability of finding common GO categories (*column 4*) between 2 sets of GO categories for each cluster. The calculation of *P* values is described in METHODS.

The putative *cis*-regulatory elements discovered in this study show high abundance at particular sites of human promoters. For example, the overwhelming majority of the elements in the cluster "TATA" show high abundance at bin 48, which is around −20 bp to −40 bp upstream from TSSs. This site is consistent with the assumed location of TATA box in promoters of human genes. The high CRE_f values these elements have obtained indicate they do not show such high abundance at other sites of the 1-kb-long promoters. Rather, they only show strong peaks at particular sites. Supplemental Figure S1 shows frequency distributions of representative putative *cis*-regulatory elements that have the biggest CRE_f values in each cluster. The specificity of these elements is clearly observed in that figure.

When the approach of comparative genomics is used to search DNA sequences that are conserved during the course of evolution, there is always a chance that the identified DNA sequences are in fact general nonspecific sequences that have little or no relevance to the target of the discovery process. However, because the general nonspecific sequences were assumed to be randomly distributed across the genome, they should have little or no statistical significance, and proper statistical algorithms should be able to filter them out. In this study we conducted statistical analysis to filter out those 8-mer sequences that have low statistical significance (CRE_f values) before carrying out the conservation analysis. Comparison between distantly related organisms (humans/mice with Drosophila) will further remove the nonspecific sequences from the resulting list of putative *cis*-regulatory elements. We are confident that the putative *cis*-regulatory elements discovered in this study stand out from a large set of candidate cis-regulatory elements because they are subject to selection pressure in the course of evolution.

In this study, GO classification was applied to validate the putative *cis*-regulatory elements we discovered and the clusters these elements have formed. There are good matches between GO categories of the transcription factors binding to these elements and GO categories of genes whose promoters contain these elements. Some biological evidence has been found that supports these matches. For example, Adnane et al. (1) suggested that GGTI-298-mediated upregulation of $p21^{WAF1/CIP1}$ involved an increase in the amount of DNA-bound Sp1–Sp3 and enhancement of Sp1 transcriptional activity. The dominant negative mutant of the small GTPase RhoA was able to activate $p21^{WAF1/CIP1}$ and constitutively active RhoA repressed $p21^{WAF1/CIP1}$. In this study, the small GTPase RhoA was also found in the list of genes putatively regulated by Sp1, and they share the same GO category "small GTPase mediated signal_transduction" (GO:0007264). Additionally the location of the Sp1 binding site Adnane et al. found is consistent with the location of Sp1 binding sites determined by this study.

The GO category "peripheral nervous system development" is one of the common categories found for EGR and its putatively regulated genes (*cluster 12*). Nerve growth factor (NGF) plays a critical role in the development and survival of neurons in the peripheral nervous system (21). Warner et al. (26) described that stable expression of Egr2 is specifically associated with the onset of myelination in the peripheral nervous system. Egr-2 or Krox-20 has also been observed in the developing mammalian hindbrain (16).

NRF-1 is a nuclear encoded gene product that has been shown to be important for the transcriptional regulation of multiple mitochondrial genes involved in organelle biogenesis and cellular respiration. Deletion or mutation of the sequences containing the NRF-1 site at positions −61 bp to −49 bp upstream from TSSs essentially abolished *CXCR4* promoter activity. CXCR4 is both a chemokine receptor and entry coreceptor for T-cell line-adapted human immunodeficiency virus type 1 (27). In this study, NRF-1 is found to be sharing the GO category "organelle organization and biogenesis" (GO: 0006996)with its putatively regulated genes. Again, the determined NRF-1 binding site is consistent with the binding sites found by Wegner et al. (27).

G/C accounts for 53.4% of promoter regions investigated in this study, which is ~6.8% higher than the A/T content in human promoter regions. However, some G/C-rich putative *cis*-regulatory elements (e.g., some elements in SP1 cluster) are ~10 times more abundant than A/T-rich putative *cis*-regulatory elements (e.g., elements in TATA box cluster). Therefore, the biased G/C content in human promoter regions only makes a very small contribution to the abundance of G/C-rich *cis*-regulatory elements. We believe that the binding requirements by transcription factors (e.g., SP1, EGR, NRF-1, etc.) in the promoter regions are one of the major factors that could explain the number of G/C-rich elements in these regions. Additionally the peaks shown at particular sites by these elements can also support their assumed role in the binding of transcription factors to the promoter regions of genes.

Positions of *cis*-regulatory elements in the promoter region are important for them to be identified and bound by transcription factors to regulate the expression of genes. Table 1 shows the position (bin number) of each discovered putative *cis*-regulatory element where it has the maximum occurrence number in all bins. The TATA box is well known to be located −25 bp to −30 bp upstream from the TSSs. This study shows that the overwhelming majority of regulatory elements in the TATA cluster is located at *bin 48*, which is −20 bp to −40 bp upstream from TSSs. This supports the known TATA box location.

Most of the elements in the CCAAT cluster lie in *bin 45* and *46*, i.e., −100 bp to −60 bp upstream from TSSs. Analysis of 5′-deletion and substitution mutants in HeLa nuclear extracts has shown that the basal activity of the promoter depends primarily on a CCAAT box sequence located at −65 (17). CCAAT box, located between −72 and −77 relative to TSSs, is one of the three regions where mutations would result in a significant decrease in the level of transcription (18). AP-1 (HeLa cell-activating protein 1) sites residing within two promoter elements of the osteocalcin gene bind the Fos-Jun protein complex: the osteocalcin box (OC box; nucleotides −99 to −76), which contains a CCAAT motif as a central element and influences tissue-specific basal levels of osteocalcin gene transcription (19).

SP1 binding sites were found in bin *46, 47,* and *48*, i.e., −80 bp to −20 bp upstream from TSSs. A conserved Sp1 site was found at −43 bp to −38 bp upstream from TSSs, which is associated with maximum reporter gene activity (8). A sequence at −60 bp binds the transcription factor Sp1 in vitro and in vivo and is essential for CD11b promoter activity (6).

Also, the upstream promoter region of the AIDS virus LTR lies between −45 and −77 and contains three tandem, closely spaced SP1 binding sites of variable affinity (14).

NRF-1 binding sites were found to be mainly located in *bin 47* and *48*, which are −60 bp to −20 bp upstream from TSSs. This supports the known NRF-1 binding sites, i.e., −61 bp to −49 bp upstream from TSSs (27).

Elements in the CRE cluster appear in *bin 47* and *48*, i.e., −20 bp to −60 bp upstream from TSSs. Cotransfection experiments showed that the cyclin D1 promoter is inducible by c-Jun and that this induction is mediated predominantly through the protected putative CRE at −52 (12). One of the two identified putative cAMP-response elements appeared at position −38. Functional analysis showed that this element is necessary for complete PKA induction (11).

Conclusions

By combining statistical analysis and the phylogenetic footprinting technique, this study yielded 124 putative *cis*-regulatory elements that not only had high statistical significance but were well conserved across the human, mouse, and Drosophila species. These elements were grouped into 20 clusters, of which 15 clusters had known transcription factors. Examination of the coexistence of these clusters found that SP1, EGR, and NRF-1 were the dominant clusters that appeared most frequently in the combinatorial combination of up to five clusters, implying that the CpG island is an important part of human gene promoters. GO analysis revealed that in most clusters GO categories of transcription factors matched GO categories of regulated genes. However, only a few GO categories have been found for the genes whose promoters contain *cis*-regulatory elements from the new clusters despite their high statistical significance and conservation across the three species. These elements could potentially represent good candidates for further systematic experimental evaluations.

ACKNOWLEDGMENTS

W. Shi conceived the study, conducted the data analysis, and drafted the manuscript. W. Zhou supervised the project. D. Xu suggested the analysis on coexistence of *cis*-regulatory elements and provided comments from the perspective of biology. All authors read and approved the final version.

We thank Barry R. Zeeberg for the help on using MatchMiner, GOMiner, and Genesis; Andrey Shlyakhtenko for the detailed explanation of his TFBS discovery algorithm; Fuchun Huang for the helpful discussion on the statistical algorithm; and Aneta Dowsing, Shamith Samarjiwa, and Lingdi Zhou for improving the English of this manuscript. We are also grateful to the anonymous referees whose comments and suggestions allowed a significant improvement of this work. This study used high-performance computing facilities in APAC and VPAC, Australia.

REFERENCES

1. **Adnane J, Bizouarn FA, Qian Y, Hamilton AD, Sebti SM.** p21$^{WAF1/CIP1}$ is upregulated by the geranylgeranyltransferase I inhibitor GGTI-298 through a transforming growth factor β- and Sp1-responsive element: involvement of the small GTPase RhoA. *Mol Cell Biol* 12: 6962–6970, 1998.
2. **Bortoluzzi S, Coppe1 A, Bisognin A, Pizzi C, Danieli CA.** A multistep bioinformatic approach detects putative regulatory elements in gene promoters. *BMC Bioinformatics* 6: 121, 2005.
3. **Buchler NE, Gerland U, Hwa T.** On schemes of combinatorial transcription logic. *Proc Natl Acad Sci USA* 100: 5136–5141, 2003.
4. **Bussemaker HJ, Li H, Siggia ED.** Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proc Natl Acad Sci USA* 97: 10096–10100, 2000.
5. **Bussey KJ, Kane D, Sunshine M, Narasimhan S, Nishizuka S, Reinhold WC, Zeeberg B, Ajay W, Weinstein JN.** MatchMiner: a tool for batch navigation among gene and gene product identifiers. *Genome Biology* 4: R27, 2003.
6. **Chen HM, Pahl HL, Scheibe RJ, Zhang DE, Tenen DG.** The Sp1 transcription factor binds the CD11b promoter specifically in myeloid cells in vivo and is essential for myeloid-specific promoter activity. *J Biol Chem* 11: 8230–8239, 1993.
7. **Corà D, Herrmann C, Dieterich C, Cunto FD, Provero P, Caselle M.** Ab initio identification of putative human transcription factor binding sites by comparative genomics. *BMC Bioinformatics* 6: 110, 2005.
8. **Croager EJ, Gout AM, Abraham LJ.** Involvement of Sp1 and microsatellite repressor sequences in the transcriptional control of the Human *CD30* gene. *Am J Pathol* 156: 1723–1731, 2000.
9. **FitzGerald PC, Shlyakhtenko A, Mir AA, Vinson C.** Clustering of DNA sequences in human promoters. *Genome Res* 14: 1–13, 2004.
10. **Gene Ontology Consortium.** Creating the gene ontology resource: design and implementation. *Genome Res* 11: 1425–1433, 2001.
11. **Giono LE, Varone CL, Cánepa ET.** 5-Aminolaevulinate synthase gene promoter contains two cAMP-response element (CRE)-like sites that confer positive and negative responsiveness to CRE-binding protein (CREB). *Biochem J* 353: 307–316, 2001.
12. **Herber B, Truss M, Beato M, Muller R.** Inducible regulatory elements in the human cyclin D1 promoter. *Oncogene* 7: 2105–2107, 1994.
13. **Jensen ST, Shen L, Liu JS.** Combining phylogenetic motif discovery and motif clustering to predict co-regulated genes. *Bioinformatics* 20: 3832–3839, 2005.
14. **Jones KA, Kadonaga JT, Luciw PA, Tjian R.** Activation of the AIDS retrovirus promoter by the cellular transcription factor, Sp1. *Science* 4751: 755–759, 1986.
15. **Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC.** Detecting subtle sequence signals: a Gibbs Sampling strategy for multiple alignment. *Science* 262: 208–214, 1993.
16. **Morgan JI, Curran T.** Stimulus-transcription coupling in the nervous system: involvement of the inducible proto-oncogenes fos and jun. *Annu Rev Neurosci* 14: 421–451, 1991.
17. **Morgan WD, Williams GT, Morimoto RI, Greene J, Kingston RE, Tjian R.** Two transcriptional activators, CCAAT-box-binding transcription factor and heat shock transcription factor, interact with a human hsp70 gene promoter. *Mol Cell Biol* 3: 1129–1138, 1987.
18. **Myers RM, Tilly K, Maniatis T.** Fine structure genetic analysis of a beta-globin promoter. *Science* 4750: 613–618, 1986.
19. **Owen TA, Bortell R, Yocum SA, Smock SL, Zhang M, Abate C, Shalhoub V, Aronin N, Wright KL, Wijnen AJ.** Coordinate occupancy of AP-1 sites in the vitamin D-responsive and CCAAT box elements by Fos-Jun in the osteocalcin gene: model for phenotype suppression of transcription. *Proc Natl Acad Sci USA* 24: 9990–9994, 1990.
20. **Pilpel Y, Sydarsanam P, Church GM.** Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat Genet* 29: 153–159, 2001.
21. **Salton SR, Fischberg DJ, Dong KW.** Structure of the gene encoding VGF, a nervous system-specific mRNA that is rapidly and selectively induced by nerve growth factor in PC12 cells. *Mol Cell Biol* 5: 2335–2349, 1991.
22. **Sinha S, Schroeder MD, Unnerstall U, Gaul U, Siggia ED.** Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in Drosophila. *BMC Bioinformatics* 5: 129, 2004.
23. **Sturn A, Quackenbush J, Trajanoski Z.** Genesis: cluster analysis of microarray data. *Bioinformatics* 18: 207–208, 2002.
24. **Tadesse MG, Vannucci M, Lio P.** Identification of DNA regulatory motifs using Bayesian variable selection. *Bioinformatics* 20: 2553–2561, 2004.
25. **Thijs G, Lescot M, Marchal K, Rombauts S, Moor BD, Rouze P, Moreau Y.** A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics* 17: 1113–1122, 2001.
26. **Warner LE, Mancias P, Butler IJ, McDonald CM, Keppen L, Koob KG, Lupsk JR.** Mutations in the early growth response 2 (EGR2) gene are associated with hereditary myelinopathies. *Nat Genet* 18: 382–284, 1998.
27. **Wegner SA, Ehrenberg PK, Chang G, Dayhoff DE, Sleeker AL, Michael NL.** Genomic organization and functional characterization of the chemokine receptor *CXCR4*, a major entry co-receptor for human immunodeficiency virus type 1. *J Biol Chem* 8: 4754–4760, 1998.

28. **Wolberger C.** Combinatorial transcription factors. *Curr Opin Genet Dev* 8: 552–559, 1998.
29. **Wu H, Su Z, Mao F, Olman V, Xu Y.** Prediction of functional modules based on comparative genome analysis and Gene Ontology application. *Nucleic Acids Res* 33: 2822–2837, 2005.
30. **Xie X, Lu J, Kullbokas EJ, Golub T, Mootha V, Lindblad-Toh K, Lander ES, Kellis M.** Systematics discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* 434: 338–345, 2005.
31. **Yoon S, Micheli GD.** Prediction of regulatory modules comprising microRNA and target genes. *Bioinformatics* 21, *Suppl* 2: ii93–ii100, 2005.
32. **Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S, Bussey KJ, Riss J, Barrett JC, Weinstein JN.** GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biology* 4: R28, 2003.
33. **Zeeberg BR, Qin H, Narasimhan S, Sunshine M, Cao H, Kane DW, Reimers M, Stephens R, Bryant D, Burt SK, Elnekave E, Hari DM, Wynn TA, Cunningham-Rundles C, Stewart DM, Nelson D, Weinstein JN.** High-Throughput GoMiner, an 'industrial-strength' integrative Gene Ontology tools for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency(CVID). *BMC Bioinformatics* 6: 168, 2005.