Published online February 12, 2004

Nucleic Acids Research, 2004, Vol. 32, No. 3 ***949–958***
DOI: 10.1093/nar/gkh246

Statistical analysis of over-represented words in human promoter sequences

Leonardo Mariño-Ramírez, John L. Spouge, Gavin C. Kanga and David Landsman*

Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, MSC 6075 Bethesda, MD 20894-6075, USA

Received November 7, 2003; Revised and Accepted December 30, 2003

ABSTRACT

The identification and characterization of regulatory sequence elements in the proximal promoter region of a gene can be facilitated by knowing the precise location of the transcriptional start site (TSS). Using known TSSs from over 5700 different human full-length cDNAs, this study extracted a set of 4737 distinct putative promoter regions (PPRs) from the human genome. Each PPR consisted of nucleotides from –2000 to +1000 bp, relative to the corresponding TSS. Since many regulatory regions contain short, highly conserved strings of less than 10 nucleotides, we counted eight-letter words within the PPRs, using *z*-scores and other related statistics to evaluate their over- and under-representation. Several over-represented eight-letter words have known biological functions described in the eukaryotic transcription factor database TRANSFAC; however, many did not. Besides calculating a *P*-value with the standard normal approximation associated with *z*-scores, we used two extra statistical controls to evaluate the significance of over-represented words. These controls have important implications for evaluating over- and under-represented words with *z*-scores.

INTRODUCTION

There is presently a major effort to characterize the human transcriptome and to improve genome annotation with full-length cDNA libraries (1,2). A unique opportunity for understanding gene regulation in humans is at hand because of these efforts, and because of the availability of a high-quality human genome sequence (3). To take advantage of the opportunity, this article examines proximal promoters, including gene regulatory sequences located near the transcriptional start site (TSS), which are particularly important in the initiation of transcription (4).

Since proximal promoter sequences are of great importance, many computational methods have been developed for identifying TSSs (5–7). It is possible, however, to identify the TSS by aligning the corresponding full-length cDNA sequence to the human genome. Experimental knowledge of the precise 5′ ends of cDNAs should facilitate the identification and characterization of regulatory sequence elements in proximal promoters (8). As a first experimental step in this direction, Suzuki *et al.* (9) used the oligo-capping method to identify TSSs from cDNA libraries enriched in full-length cDNA sequences, which they have made available at the Database of Transcriptional Start Sites (DBTSS; http://dbtss.hgc.jp/). We have used the DBTSS data set and aligned the full-length cDNAs to the human genome, thereby extracting putative promoter regions (PPRs).

Many regulatory sequence elements in the PPRs are likely to be short, highly conserved strings of less than 10 nucleotides (10). Recently, as many as 1858 exact 8mers were found conserved in orthologous promoters of closely and distantly related yeast species (11). Additionally, conserved 8mers have also been found in regulatory regions of higher eukaryotes using phylogenetic footprinting methods (12). These elements can be identified by enumerative methods, which count all possible DNA words of a certain length in promoter sequences and then use statistics such as *z*-scores to evaluate over-represented words (13–16).

Significant progress in the quality of the human genome sequence has been observed since the publication of the draft sequence in 2001 (3). Many ambiguities in the DNA sequence around the annotated genes have been resolved in the NCBI build 33, allowing a statistical analysis of eight-letter words in PPRs. Here we describe the results of applying enumerative methods to eight-letter words in the human PPRs where the statistical significance of over-represented words was determined using three different methods: (i) analytically derived *z*-scores (the standard method of assigning statistical significances to exact word matches in DNA); (ii) computer simulation of the Markov chain underlying the *z*-scores, to compare the *z*-scores with the actual extreme value distribution that they are supposed to approximate; and (iii) computer simulation of 1000 mock data sets, composed of matched, uniform random DNA sequences from the human genome (which produced *P*-values that were much more conservative than the *z*-scores). Our results show that when applied to our data, the three controls produce very different *P*-values. However, we anticipate that the over-represented words identified in this study should provide seeds for developing position-specific scoring matrices to identify novel transcription factor-binding sites (TFBSs).

*To whom correspondence should be addressed. Tel: +1 301 435 5981; Fax: +1 301 480 2288; Email: landsman@ncbi.nlm.nih.gov

MATERIALS AND METHODS

Identification of PPRs

We extracted 5756 representative full-length human cDNAs from the DBTSS (9). Using the MegaBLAST program (17), we located full-length cDNA clones in a reference human genome sequence, NCBI Build 33. Spidey, a tool that aligns cDNA to genomic DNA, was used to determine the positions of the TSSs within the genomic sequence (http://www.ncbi.nlm.nih.gov/IEB/Research/Ostell/Spidey/). We considered a piece of genomic DNA to correspond to a cDNA, if (i) it contained the TSS; (ii) it covered >90% of the cDNA; and (iii) it had >90% identity in the first exon. A total of 4737 cDNAs met these three criteria. With the TSSs in hand, we defined our 4737 PPRs as the genomic sequences running from –2000 to +1000 bp, relative to each TSS at 0 bp. The PPRs are available at ftp://ftp.ncbi.nlm.nih.gov/pub/marino/published/hs_promoters/fasta/.

CpG islands in the PPRs

A CpG algorithm developed by Takai and Jones (18) and implemented at NCBI identified CpG islands within the PPRs. Note that the algorithm is more restrictive than the original definition due to Gardiner-Garden and Frommer (19). We define a DNA region to be a CpG island, if (i) it is $\geqslant$500 bp in length; (ii) it has a G + C content $\geqslant$50%; and (iii) its ratio of CpG observed/expected is greater than or equal to 0.6 according to the strict NCBI definition (http://www.ncbi.nih.gov/mapview/static/humansearch.html#cpg).

Transcription factor-binding sites in the PPRs

TFBSs were identified from the vertebrate matrices in the TRANSFAC® Professional Suite (Version 7.2) (20), using the TFBS Perl modules for TFBS analysis (21) (http://forkhead.cgb.ki.se/TFBS/). When searching for transcription factors corresponding to TATA, GC and CAAT boxes, we used the same relative cut-off scores as Tsunoda and Takagi (22) to make our results comparable with those of Suzuki *et al.* (23). In all other cases, for simplicity, we used a relative cut-off score of 0.75.

Random genomic DNA data sets

A total of 1000 random genomic data sets, each composed of 4737 randomly selected human genomic DNA sequences of length 3001 bp, were generated as follows. The 545 contiguous sequences (contigs) from NCBI build 33 (2 866 452 029 total base pairs) were used as a source of human genomic DNA. Each contig was sampled roughly in proportion to its length, as follows. All the contigs were ordered end-to-end in chromosomal order and assigned a virtual start and end position. Then, uniformly from 1 to 2 866 452 029, the starting position of a mock PPR was selected at random. If it was in the last 3000 bases of a particular contig, the mock PPR was rejected to avoid running over the end of the contig. Otherwise, the mock PPR ran from the starting position, to its end 3000 bases downstream in the contig. The two possible DNA orientations for each contig were also sampled, each with probability 0.5. Sampling continued until 4737 sampled mock PPR sequences were accepted, forming a single random genomic data set. The sampling was repeated to form 1000 random genomic data sets.

Evaluation of over- and under-representation of eight-letter words in the PPRs

Using overlapping windows, we counted all possible two-, three- and eight-letter words in the 4737 PPRs, considering only words consisting of the four unambiguous nucleotides acgt. Then, we performed the same counts for the 4737 mock PPRs from the 1000 random genomic data sets. Somewhat arbitrarily, the numbers 2 and 8 were selected before the data analysis, for the following reasons. A protein interacting with DNA often attaches to several binding elements of contiguous nucleotides, with gaps between the elements. A binding element usually does not include both sides of the DNA double helix. Somewhat arbitrarily, therefore, the number 2 represents an approximate lower bound for the number of nucleotides on the non-binding side. With each turn of the DNA helix being 10.5 nucleotides, the number 8 then approximates the remaining number of nucleotides per turn. Words of length 3 = 2 + 1 were counted, because some of the statistics below require this count, in addition to the two- and eight-letter word counts.

Several statistics were used to evaluate over- or under-representation of words in the PPRs relative to the human genome. In previous notations (24), all statistics were standardized with the equation

$$z(W) = \frac{N(W) - \hat{N}(W)}{\sqrt{\hat{V}(W)}} \qquad (\mathbf{1})$$

In equation **1**, W represents the word being evaluated, and $N(W)$ is its count in the PPRs. The quantities $\hat{N}(W)$ and $\hat{V}(W)$ are various estimates (given below) of the mean and variance of $N(W)$.

Calculation of the S-score (standardized score)

The S-score is $z(W)$ in equation **1**, with $\hat{N}(W)$ and $\hat{V}(W)$ being the sample mean and sample variance of $N(W)$ over all mock PPRs from the 1000 random genomic data sets. Thus, if $N_i(W)$ was the count for the word W in the i-th random genomic data set, then

$$\hat{N}(W) = \frac{1}{1000}\sum_{i=1}^{1000} N_i(W) \qquad (\mathbf{2})$$

and

$$\hat{V}(W) = \frac{1}{1000-1}\sum_{i=1}^{1000}\{N_i(W) - \hat{N}(W)\}^2 \qquad (\mathbf{3})$$

Calculation of the z-score

The z-score is $z(W)$ in equation **1**, with $N(W)$ and $V(W)$ being the (conditional) mean and variance of $N(W)$ derived from a second-order Markov model. The (transitional) word frequencies in the Markov model were determined from the empirical

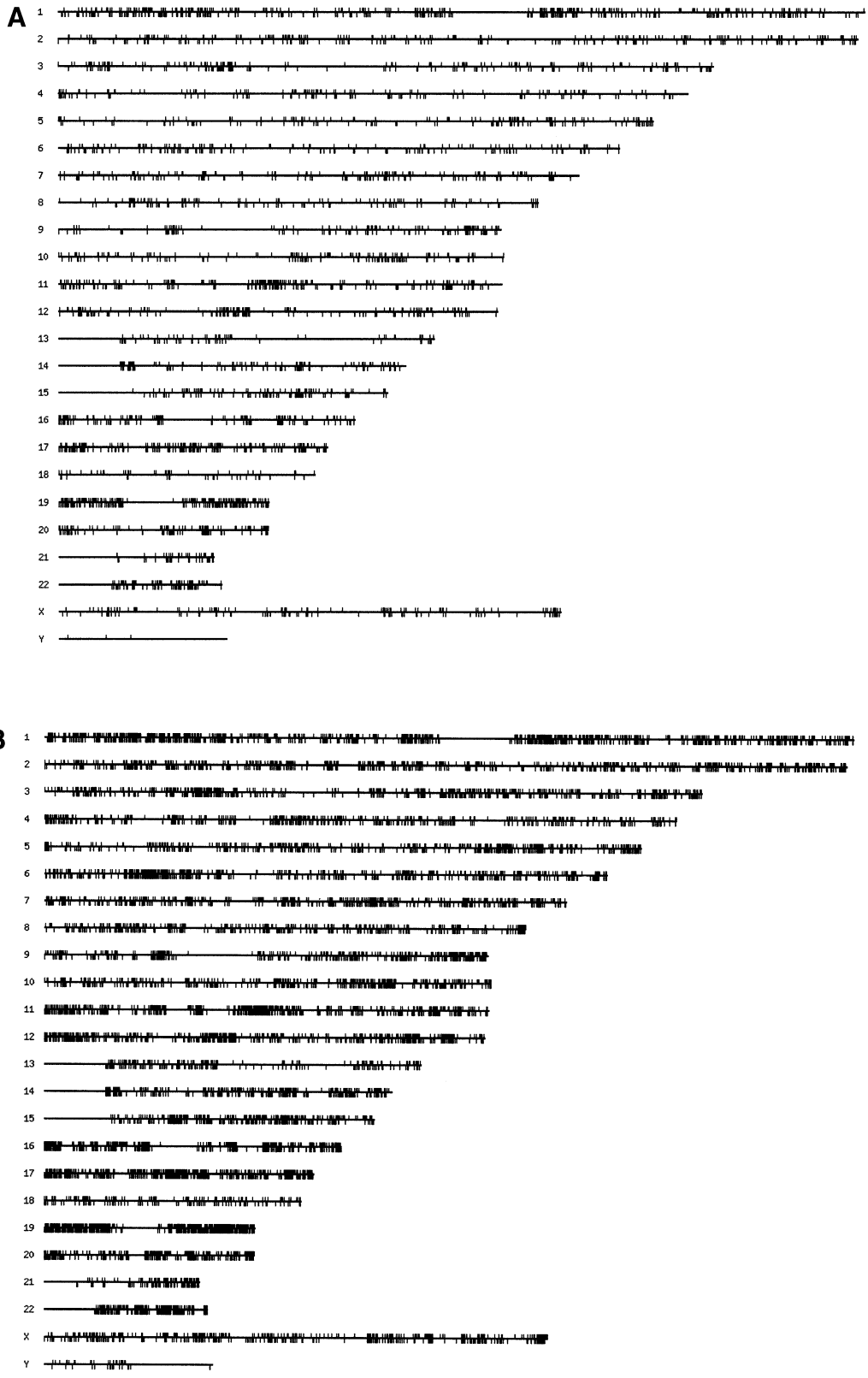


Figure 1. Distribution of the PPRs in the human genome. (**A**) The positions of the TSS from the 4737 PPRs on each chromosome (horizontal lines) are represented as vertical bars. The vertical lines above the chromosome lines correspond to TSSs mapped in the positive chromosome orientations; and those below, the negative chromosome orientations. (**B**) The location of the 5′ end of 18 372 reviewed RefSeq transcripts present in NCBI build 33 and obtained from ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/RNA/rna.fa.gz.

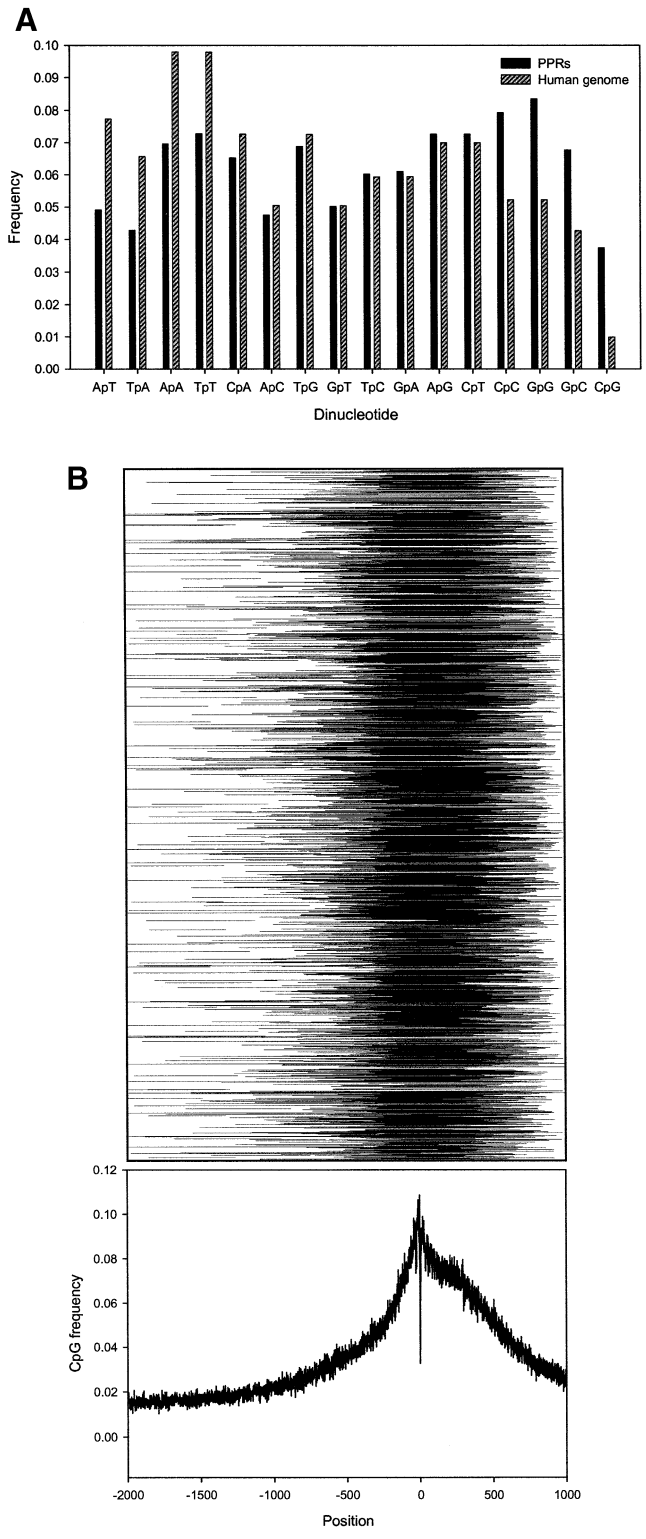


Figure 2. Dinucleotide frequencies in the PPRs and the human genome (NCBI build 33). (**A**) The dinucleotides are ranked left to right according to their S-score from most under-represented to most over-represented. (**B**) CpG Islands in the PPRs (upper panel). The CpG islands are represented as black lines in the PPRs where other regions are represented as white spaces. Positional CpG frequency in the PPRs (lower panel). Total CpG dinucleotide counts were obtained in overlapping 2 bp windows for all PPRs and their frequency estimated relative to the number of sequences (4737). The nucleotide positions are relative to the TSS.

Table 1. Dinucleotide frequencies in the PPRs

Dinucleotide	Counts in PPRs	%	Counts in random DNA	%	S-score
AT	696 056	4.90	1 098 285	7.73	–90.90
TA	607 901	4.28	933 555	6.57	–72.89
AA	989 468	6.96	1 392 168	9.80	–60.39
TT	1 034 667	7.28	1 390 939	9.79	–50.83
CA	927 872	6.53	1 031 830	7.26	–46.57
AC	674 321	4.75	716 326	5.04	–24.35
TG	977 909	6.88	1 031 298	7.26	–24.30
GT	712 115	5.01	716 059	5.04	–2.00
TC	854 573	6.01	843 167	5.93	5.33
GA	866 219	6.10	843 311	5.93	9.68
AG	1 031 731	7.26	994 087	7.00	15.60
CT	1 032 230	7.26	993 679	6.99	15.70
CC	1 126 357	7.93	739 821	5.21	88.94
GG	1 186 183	8.35	739911	5.21	100.73
GC	962 288	6.77	605 789	4.26	120.38
CG	531 105	3.74	139 773	0.98	229.03

Table 2. Representative eight-letter words[a] with high *z*-scores in the PPRs

Eight-letter word	*z*-score (rank)	*P*-value	Total count	No. of vertebrate TRANSFAC (version 7.2) site entries
gattacag	+311.86 (01)	0	3149	5
aaaaaaaa	+271.10 (02)	0	20828	4
tttttttt	+264.51 (03)	0	22589	3
ggattaca	+244.97 (04)	0	3039	4
gtaatccc	+227.57 (05)	0	2381	3
tgtaatcc	+215.46 (06)	0	2560	3
ctactaaa	+203.21 (07)	0	1447	0
ttagtaga	+197.92 (08)	0	1672	0
agtagctg	+195.68 (09)	0	2048	3
tgtgtgtg	+184.66 (10)	0	3866	5
tcgaactc	+182.06 (11)	0	813	0
gtagctgg	+181.49 (12)	0	1968	4
tagtagag	+178.41 (13)	0	1588	1
attacagg	+174.46 (14)	0	3153	4
ctaatttt	+172.18 (15)	0	2263	1
ttgtattt	+166.35 (16)	0	2187	0
ccgccgcc	+141.04 (31)	0	1975	4
gggcgggg	+108.41 (68)	0	3498	58
ggggcggg	+104.18 (85)	0	3388	43
ggcggggc	+79.13 (177)	0	2459	35
ggtgagtg	+39.52 (640)	0	740	2
cgacgcgg	+34.55 (791)	1.1e–256	171	0
tgacgtca	+31.39 (906)	1.8e–211	245	26
gcatgcgc	+30.20 (954)	1.8e–195	418	3

[a]Web queries available at http://www.ncbi.nlm.nih.gov/CBBresearch/Landsman/HRSE/.

three-letter word counts within the PPRs. Relevant formulas can be found in Schbath (24), Prum *et al*. (25) and Schbath *et al*. (26).

Under the second-order Markov model, each individual *z*-score has an asymptotic standard Gaussian (normal) distribution as $N(W) \rightarrow \infty$. We were mainly interested in the *P*-values of the largest *z*-scores over $4^8 = 65\,536$ eight-letter words, however, so normal approximations are very inaccurate in the present context (see Results and Discussion). We

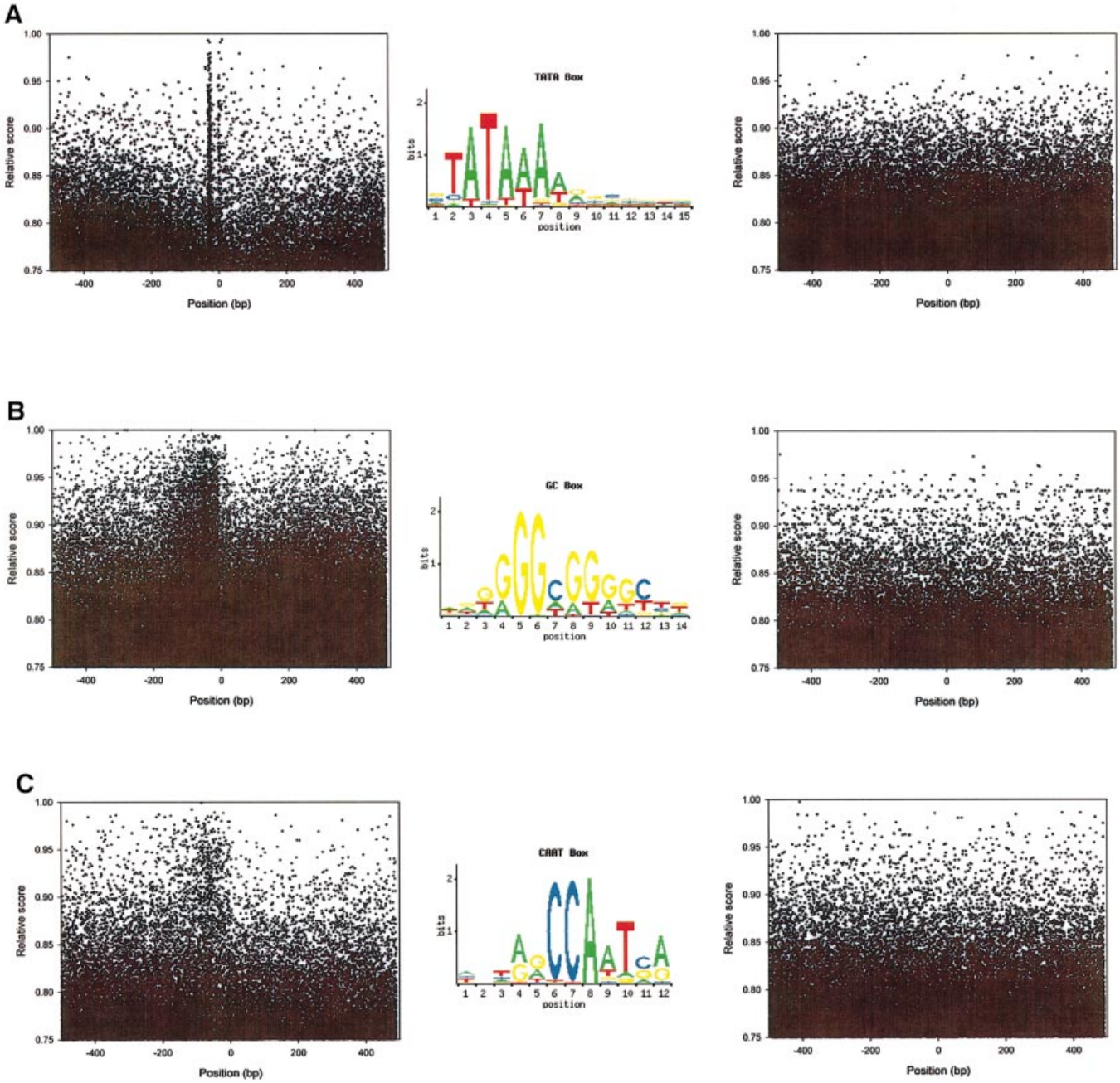
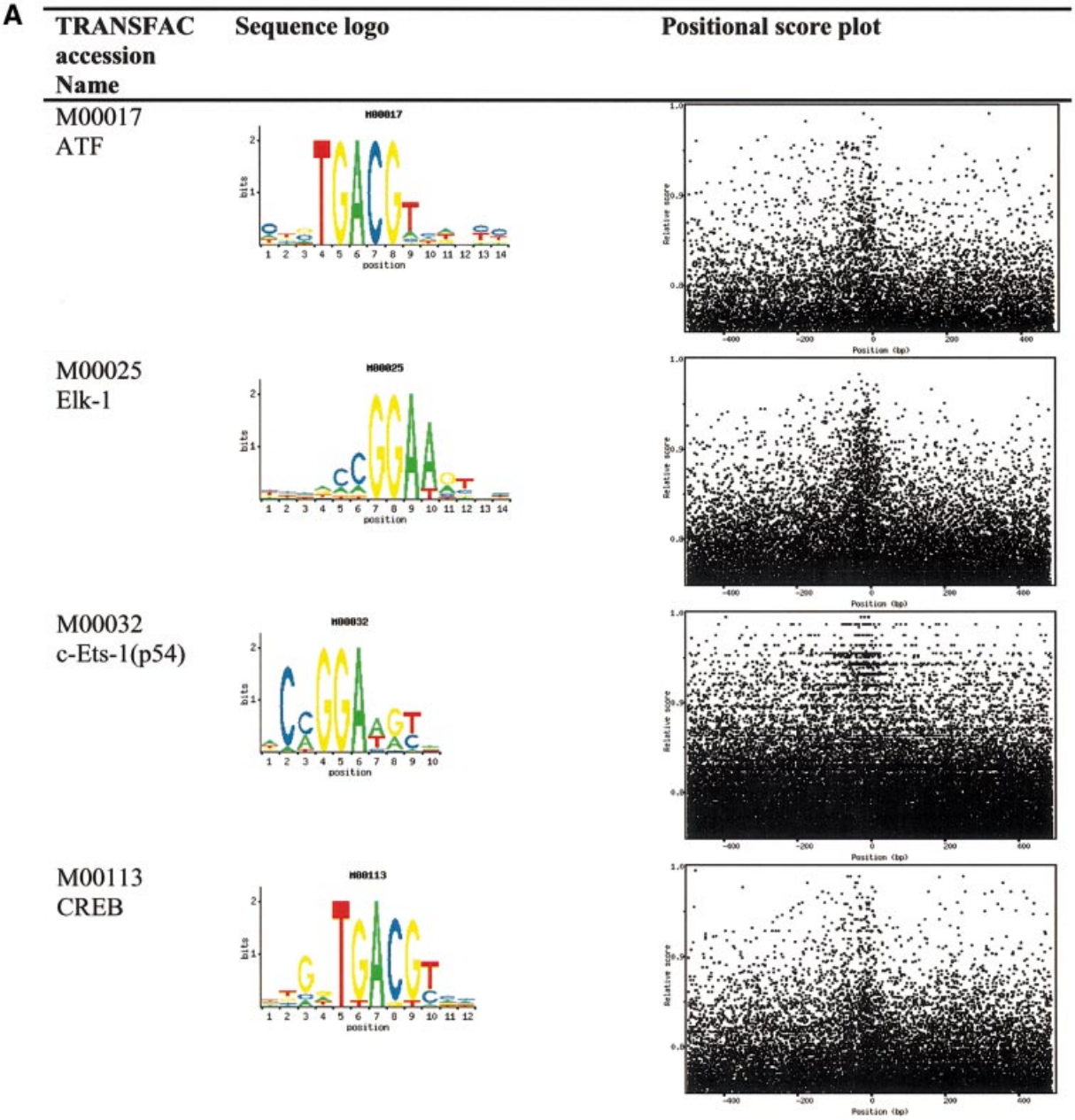


Figure 3. Major TFBSs in the PPRs. Positional scores for (**A**) TATA, (**B**) GC and (**C**) CAAT boxes (TRANSFAC accessions M00252, M00255 and M00254, respectively). The relative scores of the boxes are plotted as a function of their positions in the PPRs on the left. The sequence logos (31) for TRANSFAC count matrices corresponding to the boxes appear on the center. The relative scores of the boxes plotted as a function of their DNA location using a random genomic data set are presented on the right. All computations were performed with the TFBS Perl modules for transcription factor detection and analysis as described in Materials and Methods.

therefore also estimated extreme value distributions for the *z*-score with two Monte Carlo methods.

Extreme value distribution of the *z*-score (Markov chain method)

Using empirical transition probabilities derived from the three-letter word counts in the PPRs, we simulated the second-order Markov model underlying the *z*-scores. Each of 1000 resulting synthetic Markov data sets was composed of 4737 sequences of length 3001 to match our actual PPR data set. We calculated *z*-scores for all possible eight-letter words using the transitional frequencies from the empirical three-letter word counts in the synthetic Markov data set under scrutiny. For each synthetic Markov data set, we recorded the maximum *z*-score over all possible eight-letter words. Finally, we determined the 95th percentile (corresponding to a *P*-value



of 0.05) of the 1000 synthetic maximum z-scores. The 95th percentile from our simulation therefore approximates the maximum z-score that is produced from the second-order Markov model by chance alone 5% of the time. In this Markov model of the PPRs, any z-score in the actual PPR data set that lies above the 95th percentile can be deemed statistically significant at $P \leqslant 0.05$. Similar procedures were also carried out for the minimum z-scores.

Extreme value distribution of the z-score (random genomic method)

For each of the 1000 random genomic data sets, which were composed of 4737 sequences of length 3001 to match our PPR data set, we calculated (just as for the original PPR data set) the z-scores of all possible eight-letter words. Just as for the 1000 synthetic Markov data sets, we calculated the 95th percentile of the maximum z-score for the 1000 random genomic data sets. As above for the Markov model, any z-score in the actual PPR data set that lies above the 95th percentile can be deemed statistically significant at $P \leqslant 0.05$. Similar procedures were also carried out for the minimum z-scores.

Algorithms and implementations for the statistics

All statistical computations were carried out with the Microsoft Visual C++ 6.0 compiler. The computer had a

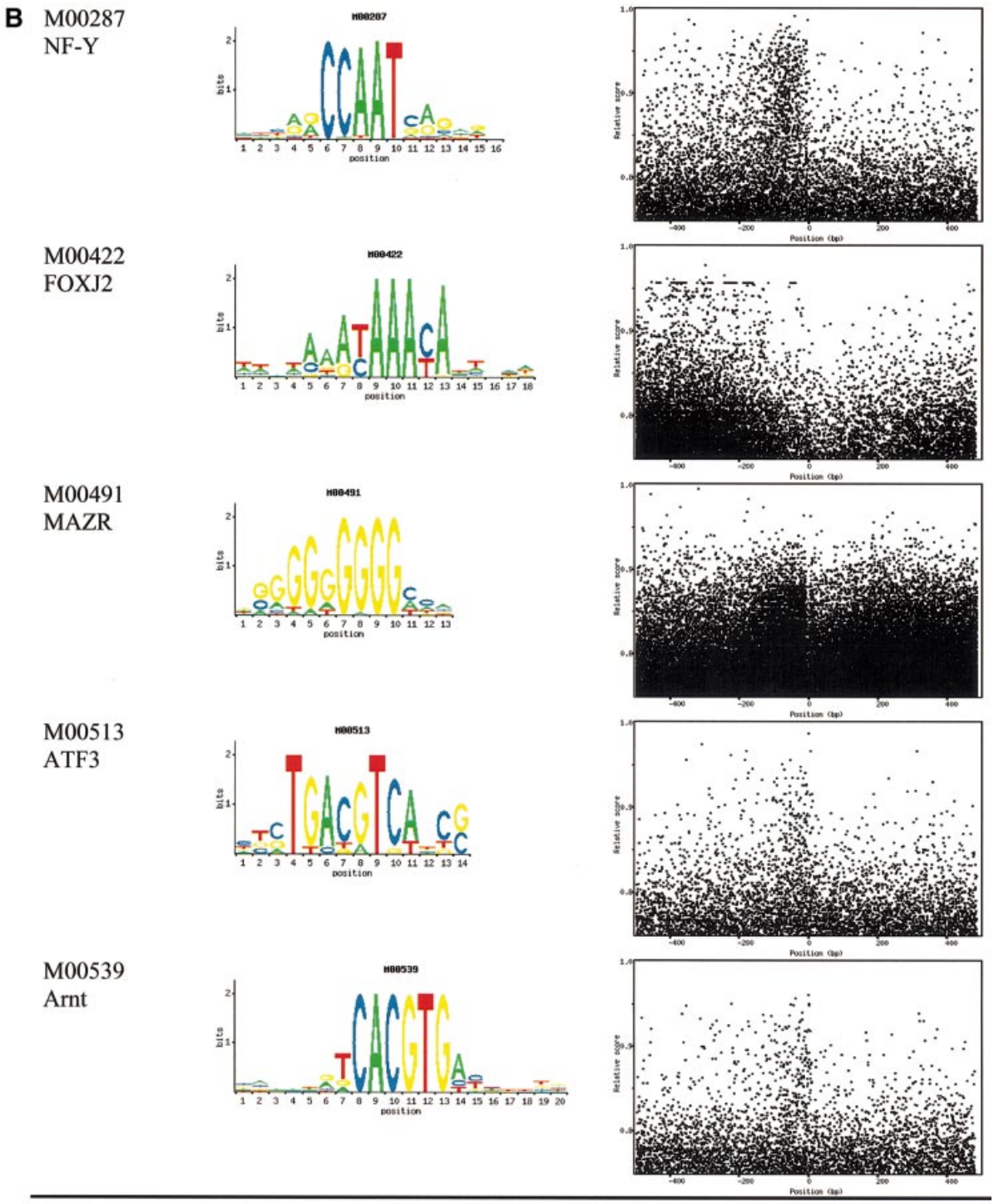


Figure 4. Representative vertebrate TRANSFAC matrices with clusters in the PPRs.

single 2.2 GHz Intel Pentium 4 CPU under the Microsoft Windows 2000 operating system. Because the statistical computations were in the central loop of a Monte Carlo simulation, they had to be efficient.

We used a typical 1–1 hash function for the nucleotide alphabet acgt, one that maps words of length h into a hash space of integers between 0 and $4^h - 1$. We then performed string operations within the hash space. For example, our word counts used a single pass of the relevant sequences. As each new character was read from the sequence, we incrementally updated the hashed integer for the word in the eight-letter window under scrutiny. The two- and three-letter word counts

were calculated from the eight-letter word counts. String operations required by the formulas for the z-score were also performed in the hash space.

RESULTS

Identification and characterization of the putative promoter regions (PPRs)

In all 24 chromosomes of the human genome, the distribution of the 4737 PPRs paralleled that of the 18 474 reviewed (NM-prefixed) RefSeq transcripts (27) in NCBI build 33 (Fig. 1). As a representative sample of human genes, therefore, our PPR data set displays no obvious location bias.

Previous studies estimated that 40–50% of human genes overlap with CpG islands (23,28). We found that 76% of our PPRs (3608 of 4737) overlap with at least one CpG island. To explore further the frequency of CpG and other dinucleotides in the PPRs, we used our S-score to compare frequencies in the PPRs with those in the human genome (Fig. 2A). In the PPRs relative to randomly sampled genomic sequence, the S-score indicated that of all dinucleotides, CpG was the most over-represented (Table 1), as might have been expected. Additionally, dinucleotides containing G and C are over-represented, whereas dinucleotides containing A and T were under-represented.

Figure 2B displays the CpG islands found in the PPRs and the positional preferences of CpG dinucleotides relative to the TSS. The CpG islands appear to cluster near the TSS, but in many cases they extend over longer regions. The CpG frequency generally increases with proximity to the TSS, but it drops dramatically at positions –29 to –24 and –1. Indeed, TATA boxes prefer these positions (see Fig. 3) and, in accord with this preference, the decrease in the CpG frequency was balanced by an increase in the frequency of TpA, ApT and ApA (data not shown).

Since we expected to find important TFBSs in their known locations, we examined the frequency and positional preferences of TATA, GC and CAAT boxes. We found that 1343 (28.4%) of the PPRs contained TATA boxes; 4197 (88.6%), GC boxes; and 2839 (60%), CAAT boxes, in agreement with previous findings (23). Figure 3 displays the positional preferences of these three motifs with respect to positions –500 to +500 bp relative to the TSS; each motif hit of a particular score at a particular position is represented as a dot. The boxes appear to cluster relative to the TSS, revealing possible regions preferred for transcription factor binding. Additionally, to get an estimate of the background noise for each box, we used a random genomic DNA data set; note that the background noise decreases significantly at relative scores above 0.85. In Figure 4, we present examples of other TRANSFAC matrices with apparent clustering tendencies. The positional relative scores for 466 vertebrate TRANSFAC matrices with respect to positions –500 to +500 bp relative to the TSS, along with their sequence logos, can be found at ftp://ftp.ncbi.nlm.nih.gov/pub/marino/published/hs_promoters/tfbs/. The positional preferences of all words clearly merit a more detailed statistical analysis, and will be presented elsewhere (L. Mariño-Ramírez, J.L. Spouge and D. Landsman, in preparation).

Table 3. Extreme value distributions[a] of the z-score

	Upper z-score $P_{0.05}$ (words above z-score)	Lower z-score $P_{0.05}$ (words below z-score)
Markov chain method	+31.00 (926)	–32.43 (2)
Random genomic method	+357.83 (0)	–54.37 (0)

[a]Web queries available at http://www.ncbi.nlm.nih.gov/CBBresearch/Spouge/Articles/.

Statistical analysis of eight-letter words in the PPRs

Many eight-letter words with high z-scores correspond to known TFBSs in the vertebrate subset of the TRANSFAC database, which in most cases is associated with one or more binding factors and references in the literature (Table 2). The most frequent eight-letter word found in the vertebrate subset of TRANSFAC sites is gggcgggg, which is a DNA sequence recognized by the Sp1 factor and was ranked number 68 according to its z-score. Indeed, other Sp1-related eight-letter words also had high z-scores (e.g. ggggcggg, ggcggggc, cccgcccc). We found TRANSFAC site entries for 12 of the 16 eight-letter words with the highest z-scores, suggesting a biological role in transcriptional regulation for words with high z-scores.

In addition, we found a few eight-letter words that are involved in transcriptional regulation but received poor z-scores. This is the case for atttgcat, which is the target for the ubiquitous octamer-binding factor present in immunoglobulin heavy chain genes (29). This observation can be explained by the fact that our data set contains only a few immunoglobulin genes, making difficult the identification of over-represented words present in those genes.

Unfortunately, the normal approximation implicit in the z-scores does not closely parallel the corresponding extreme value distribution (high or low). For example, in Table 2, the eight-letter word gattacag received the highest z-score, z = 311.86; the word aaaaaaaa was next, with z = 271.10. Under the normal approximation, even after multiplying by 65 536 to correct for multiple testing, 12 615 words received a Bonferroni upper bound less than 0.05, indicating a statistically significant one-sided P-value ($P < 0.05$).

Table 3 presents the $P < 0.05$ levels for z-scores from some related extreme value distributions (extreme value distributions do not require a correction for multiple testing). A simulation of the actual Markov chain underlying the z-scores estimated the 95th percentile of the corresponding extreme value distribution at z = 31.00. Only 926 words had z-scores above z = 31.00. Even more surprisingly, however, the smallest maximum z-score produced by 1000 mock PPR data sets sampled from random genomic DNA was z = 291.10. In the actual PPR data set, only one z-score (z = 311.86) exceeded z = 291.10. Moreover, the actual maximum z-score (z = 311.86) corresponded to the 4.7 percentile of maximum z-scores among the 1000 mock PPR data sets from random genomic DNA, a value that might be considered anomalously low. Both the Markov and random DNA controls indicate that the normal approximation overestimates the significance of a z-score.

We developed a web interface to perform queries for words of length 2–8 and obtain the statistics described here in addition to the PPRs or TRANSFAC sites that contain a particular eight-letter word. The interface is available at the following URL: http://www.ncbi.nlm.nih.gov/CBBresearch/Landsman/HRSE/. The complete list of *z*-scores and their random controls can be found at http://www.ncbi.nlm.nih.gov/CBBresearch/Spouge/Articles/.

DISCUSSION

Using 5756 representative full-length cDNAs from the DBTSS, we identified PPRs for 4737 RefSeq transcripts. The distribution and location of the PPRs in the human genome indicate that they are a representative sample of the reviewed RefSeq transcripts present in NCBI build 33 (Fig. 1).

We found that dinucleotides containing C or G are over-represented in the PPRs, whereas dinucleotides containing T or A are under-represented (Table 1 and Fig. 2). In addition, 76% of the PPRs overlapped at least one CpG island. Despite our stringent definition of a CpG island, the percentage of PPRs present in or containing CpG islands is somewhat higher than the usual 40–50% given in previous studies (23,28), suggesting a tight association between CpG islands and regions surrounding the TSS. The CpG frequency generally increased with proximity to the TSS (Fig. 2), but dropped dramatically at positions –29 to –24 and –1. Probably, the presence of TATA boxes in this region causes the drop, because they increase the frequency of the dinucleotides TpA, ApT and ApA at those positions. Our characterization of dinucleotide compositions near the TSS is in general agreement with previous observations (5).

Figures 3 and 4 show that many TFBSs such as the TATA, GC and CAAT boxes often occur near the TSS (23). We found that most known TFBSs in TRANSFAC have preferred locations between –300 and +50 bp relative to the TSS. This finding suggests that the basal promoter and nearby upstream regulatory elements are found in the region between –300 and +50 bp, in accord with a recent study from the Myers laboratory, where 91% of 152 DNA fragments containing regions –550 to +50 relative to the TSS were active as promoters in at least one of four cell types evaluated (8).

A statistical analysis of over-represented words using *z*-scores yielded a ranked list of eight-letter words with possible transcription factor binding activity (Table 2). Some of these words contain mostly the bases a and t, which pair weakly, with only two hydrogen bonds. These words therefore might be facilitating the transcriptional unzipping of the DNA double helix or might configure the DNA, and thus the chromatin, in a higher order conformation which facilitates transcription regulation. For example, the transcription bubble found at DNA polymerase III transcribed gene initiation sites requires the binding of a complex of several proteins, including RNA polymerase III and transcription initiation factor IIIB, to regions of open chromatin containing promoters, and the opening up of 3–5 bp of DNA before efficient transcription can be achieved (30). Interestingly, the words aaaaaaaa and tttttttt in particular are over-represented relative to the GC-rich background of the PPRs.

On a methodological note, extreme caution should be exercised when interpreting the significance of a *z*-score for an eight-letter word. Much to our surprise, randomly selected genomic DNA data sets yielded generally higher *z*-scores than the experimentally selected PPR data set. In fact, theoreticians have cautioned that the *z*-score is inaccurate as a normal approximation to the corresponding extreme value distribution (25). Our simulation of the Markov chain underlying the *z*-score confirmed the inaccuracy, showing that the normal approximation was excessively optimistic as an estimate of statistical significance. Random genomic DNA should therefore be used as a statistical control on *z*-scores, whenever appropriate and feasible.

Despite sounding this cautionary note on *z*-scores, we have indeed found many eight-letter words at the top of our ranked list which correspond to TFBSs in the TRANSFAC database. Moreover, many of these words showed positional preferences with respect to the TSS, suggesting a specific role in transcriptional regulation. We are planning to explore the positional preferences of eight-letter words with statistical tests similar to the one described by Wolfsberg and collaborators (13).

ACKNOWLEDGEMENTS

We are grateful to Yutaka Suzuki and Riu Yamashita for providing the full-length cDNAs from DBTSS, and Philip Johnson for providing the code used for CpG island identification. This study utilized the computing facilities at the NCBI and the high-performance computational capabilities of the Biowulf/LoBoS3 cluster at the National Institutes of Health, Bethesda, MD.

REFERENCES

1. Strausberg,R.L., Feingold,E.A., Grouse,L.H., Derge,J.G., Klausner,R.D., Collins,F.S., Wagner,L., Shenmen,C.M., Schuler,G.D., Altschul,S.F. *et al.* (2002) Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl Acad. Sci. USA*, **99**, 16899–16903.
2. Carninci,P., Waki,K., Shiraki,T., Konno,H., Shibata,K., Itoh,M., Aizawa,K., Arakawa,T., Ishii,Y., Sasaki,D. *et al.* (2003) Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. *Genome Res.*, **13**, 1273–1289.
3. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
4. Ohler,U. and Niemann,H. (2001) Identification and analysis of eukaryotic promoters: recent computational approaches. *Trends Genet.*, **17**, 56–60.
5. Down,T.A. and Hubbard,T.J. (2002) Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res.*, **12**, 458–461.
6. Davuluri,R.V., Grosse,I. and Zhang,M.Q. (2001) Computational identification of promoters and first exons in the human genome. *Nature Genet.*, **29**, 412–417.
7. Bajic,V.B. and Seah,S.H. (2003) Dragon gene start finder: an advanced system for finding approximate locations of the start of gene transcriptional units. *Genome Res.*, **13**, 1923–1929.
8. Trinklein,N.D., Aldred,S.J., Saldanha,A.J. and Myers,R.M. (2003) Identification and functional analysis of human transcriptional promoters. *Genome Res.*, **13**, 308–312.
9. Suzuki,Y., Yamashita,R., Nakai,K. and Sugano,S. (2002) DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. *Nucleic Acids Res.*, **30**, 328–331.
10. Wingender,E., Dietze,P., Karas,H. and Knuppel,R. (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, **24**, 238–241.

11. Cliften,P., Sudarsanam,P., Desikan,A., Fulton,L., Fulton,B., Majors,J., Waterston,R., Cohen,B.A. and Johnston,M. (2003) Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science*, **301**, 71–76.
12. Blanchette,M. and Tompa,M. (2002) Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.*, **12**, 739–748.
13. Wolfsberg,T.G., Gabrielian,A.E., Campbell,M.J., Cho,R.J., Spouge,J.L. and Landsman,D. (1999) Candidate regulatory sequence elements for cell cycle-dependent transcription in *Saccharomyces cerevisiae*. *Genome Res.*, **9**, 775–792.
14. vanHelden,J., Aandre,B. and Collado-Vides,J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, **281**, 827–842.
15. Brazma,A., Jonassen,I., Vilo,J. and Ukkonen,E. (1998) Predicting gene regulatory elements in silico on a genomic scale. *Genome Res.*, **8**, 1202–1215.
16. Sinha,S. and Tompa,M. (2002) Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.*, **30**, 5549–5560.
17. Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203–214.
18. Takai,D. and Jones,P.A. (2002) Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl Acad. Sci. USA*, **99**, 3740–3745.
19. Gardiner-Garden,M. and Frommer,M. (1987) CpG islands in vertebrate genomes. *J. Mol. Biol.*, **196**, 261–282.
20. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
21. Lenhard,B. and Wasserman,W.W. (2002) TFBS: computational framework for transcription factor binding site analysis. *Bioinformatics*, **18**, 1135–1136.
22. Tsunoda,T. and Takagi,T. (1999) Estimating transcription factor bindability on DNA. *Bioinformatics*, **15**, 622–630.
23. Suzuki,Y., Tsunoda,T., Sese,J., Taira,H., Mizushima-Sugano,J., Hata,H., Ota,T., Isogai,T., Tanaka,T., Nakamura,Y. *et al.* (2001) Identification and characterization of the potential promoter regions of 1031 kinds of human genes. *Genome Res.*, **11**, 677–684.
24. Schbath,S. (1997) An efficient statistic to detect over- and under-represented words in DNA sequences. *J. Comput. Biol.*, **4**, 189–192.
25. Prum,B., Rodolphe,F. and de Turckheim,E. (1995) Finding words with unexpected frequencies in DNA sequences. *J. R. Stat. Soc. Ser. B, Methodol.*, **57**, 205–220.
26. Schbath,S., Prum,B. and de Turckheim,E. (1995) Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences. *J. Comput. Biol.*, **2**, 417–437.
27. Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.
28. Larsen,F., Gundersen,G., Lopez,R. and Prydz,H. (1992) CpG islands as gene markers in the human genome. *Genomics*, **13**, 1095–1107.
29. Kemler,I., Schreiber,E., Muller,M.M., Matthias,P. and Schaffner,W. (1989) Octamer transcription factors bind to two different sequence motifs of the immunoglobulin heavy chain promoter. *EMBO J.*, **8**, 2001–2008.
30. Kassavetis,G.A., Letts,G.A. and Geiduschek,E.P. (2001) The RNA polymerase III transcription initiation factor TFIIIB participates in two steps of promoter opening. *EMBO J.*, **20**, 2823–2834.
31. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.