Available online at www.sciencedirect.com
SCIENCE DIRECT
Biochemical and Biophysical Research Communications 347 (2006) 166–177
BBRC
www.elsevier.com/locate/ybbrc

A mammalian promoter model links *cis* elements to genetic networks

Junwen Wang, Sridhar Hannenhalli *

Penn Center for Bioinformatics and Department of Genetics, University of Pennsylvania, Philadelphia, PA 19104-6021, USA

Received 8 June 2006
Available online 21 June 2006

Abstract

An accurate identification of gene promoters remains an important challenge. Computational approaches for this problem rely on promoter sequence attributes that are believed to be critical for transcription initiation. Here we report a probabilistic model that captures two important properties of promoters, not used by previous methods, *viz.*, the location preference and co-occurrence of promoter elements. Additionally, we found that many of the position-specific DNA elements are strongly linked with the function of the gene product. For instance, a highly conserved motif CCTTT at −1 position is strongly associated with protein synthesis, cellular and tissue development. Our comparative analysis of promoter classes reveals that the promoters devoid of CpG islands are more conserved and have fewer alternative transcription start sites. The discovered links between promoter elements and gene function allows us to infer genetic networks from promoter elements. The web server for the PSPA promoter predictor is available at http://cagr.pcbi.upenn.edu/PSPA.
© 2006 Elsevier Inc. All rights reserved.

Keywords: Core promoter prediction; CpG island; Genetic networks; Transcription factor binding site (TFBS); Conservation; Position-specific motif; Propensity

Gene transcription relies on proper positioning of RNA polymerase at the start of the gene, which in turn is facilitated by several transcription factors (*TF*) that bind to specific DNA elements in the vicinity of the transcription start site (*TSS*) [1]. The *Core promoter* is the region around the TSS that harbors these required DNA core elements. Only a handful of core elements have been experimentally determined and a majority of known promoters do not contain these elements. An adequate characterization of the promoter sequences presents an important challenge. Computational methods have been developed to identify promoters in the genome by modeling sequence characteristics of the promoter region [2–5], however with limited success.

Transcription requires interactions among several transcription factors and polymerase. This imposes certain location constraints on the corresponding DNA elements. For example, many core elements occur at a specific distance relative to the TSS [6]. Additionally many core elements occur in the same promoter with restricted spacing between them. For example, in the adenovirus 2 *E1B* promoter, increased spacing between the GC-box and the TATA-box diminishes *in vivo* transcription significantly [7]. Similar spacing restrictions have been observed in other cases [8–13]. However, these position-specific constraints have not been fully exploited by current computational approaches. For instance, the methods reported in [3,4] use rather loose positional restriction for the core elements.

Our probabilistic model—*Position-Specific Propensity Analysis* (*PSPA*)—encapsulates the two aspects of promoters, *viz.*, position-specific occurrence of core elements and co-occurrence of these elements. The derived promoter identification tool can predict TSS with higher accuracy than the current methods.

* Corresponding author. Fax: +1 215 573 3111.
E-mail addresses: Junwen@pcbi.upenn.edu (J. Wang), sridharh@pcbi.upenn.edu (S. Hannenhalli).

0006-291X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2006.06.062

Methods

The datasets. We retrieved the full length cDNA sequences from DBTSS database [14] and mapped them to human genome (version hg16). For each of the 12,253 genes, we extracted the ±5 kb sequence flanking the TSS. We removed the promoters with greater than 95% sequence identity in the ±100 bp flanking the TSS, resulting in 10,342 non-redundant promoters. Using the CpG island search tool [15], we identified 7893 CpG-rich promoters (the TSS was contained within a CpG island) and 1277 *CpG-poor* promoters (no CpG-island in the 10 kb sequence). We have modeled these two promoter classes separately due to several known differences between them, e.g., they direct dissimilar expression pattern and there are differences in their gene product function [16,17]. The human-mouse sequence alignments were obtained from UCSC database (genome versions hg16-mm5); the alignment is obtained using the BLASTZ tool [18].

The PSPA model. Our promoter model captures the position-specific propensity (overrepresentation) of DNA elements and their co-occurrence relative to TSS. We define the propensity of a k bases long DNA element (k-*mer*) x at position p as: $\text{propensity}(p_{x,p}) = \frac{\text{observed frequency}(o_{x,p})}{\text{expected frequency}(e_{x,p})}$. Expected frequency is based on the single nucleotide frequencies in the genome. To model the additional synergistic effect of two elements, we calculate the association propensity score between k-mer x_1 at position p_1 and x_2 at position p_2 as: $\text{Association propensity}(A_{x_1,p_1,x_2,p_2}) = \frac{p_{x_1,p_1,x_2,p_2}}{p_{x_1,p_1} \times p_{x_2,p_2}}$, where p_{x_1,p_1,x_2,p_2} is defined analogously to $p_{x,p}$. The propensity and association scores are integrated into the overall score.

Model training. The model training was done separately for CpG-rich and CpG-poor promoters. To train the model, we first compute the propensity for all k-mers (k from 1 to 5) at each position in the ±100 bp relative to TSS. The position of a k-mer is defined as the position of the first base of the k-mer from the transcription start site (negative value refers to upstream and positive value refers to downstream of TSS). We then calculate the association propensity between two non-overlapping k-mers. For the top n k-mers (ranked by propensity) at each position we calculate its association propensity with top n k-mers at other downstream positions ($n = 5$ in current implementation).

Model scoring. PSPA score of a potential TSS is the product of scores for the 200 positions in the ±100 bp relative to the potential TSS. The PSPA score for a position is the product of the positional propensity score of the k-mer at the position and its best association score with a k-mer at another position. We assign an association propensity score of 1 if no such k-mer existed. The positional propensity score is determined by the longest word present in the training data at the position, with a maximum length of 5. The strategy of using the longest representative k-mer is similar to *variable length Markov chain* approach [19]. The difference is that we use propensity score instead of Markov model probabilistic score. To score a sequence with length L, we use a sliding-window frame of 200 bp and obtain L-199 scores. We rank the scores, select the top-scoring predictions (at least 50 bp apart from each other). For chromosome-wide predictions, we only retain the top-scoring predictions that are at least 1 kb apart from each other.

Cross validation. The performance of PSPA was evaluated using 10-fold cross validation. We randomly partition the promoter sequences into 10 parts, train the model on nine parts, and test on the other part. For all other programs, such as FirstEF, DragonPF, and DragonGSF, we used pre-trained model provided by the authors to predict the same set of test sequences. Because these programs do not make predictions on some sequences, and make multiple predictions on other sequence, we compared PSPA with each program separately. For example, when we compare PSPA vs. FirstEF, we consider all predictions made by FirstEF on a given sequence and consider the same number of predictions for PSPA. The sequences, on which FirstEF did not make any prediction, were not included in the evaluation, even though PSPA may make a correct prediction. In this way we ensure that both methods make the same total number of predictions. A prediction is considered correct if any of the predicted TSS is within ±L bp from the true TSS. The prediction accuracy is defined as percentage of correctly predicted sequence over the total number of sequences that have at least one prediction.

Evaluation of genome-wide predictions. We first selected all the predictions with $\log_2(\text{PSPA score}) \geqslant 100$. We then grouped predictions within 1000 bp, and retained only the highest-scoring prediction in the group. A prediction was deemed correct if it was within ±L bp of the TSS of the gene (L is the stringency cutoff; see Fig. 1 for illustration). We used sensitivity *Se*, positive predictive value (*ppv*, also called specificity), and true positive cost (*TPC*) to evaluate chromosome-wide prediction accuracy. *Se* = (correctly predicted promoters)/(total number of promoters), ppv = (correct predictions)/(valid predictions), and TPC = (incorrect predictions)/(correct predictions). Valid prediction is a prediction that is either within 2000 bp upstream of TSS or within the gene. Because we do not know whether the predictions outside of this region are true, we only include the valid prediction for evaluation, adopted from [4,20,21]. Since multiple predictions can be made for one promoter, the number of correctly predicted promoter is not always the same as that of correct prediction. We did the training and testing on independent datasets, e.g., we trained the model on all DBTSS promoters other than chr4 when testing promoters on chr4 (same for chr21, chr22). When we tested on mouse genomic region, we used a model trained on all human promoters.

The gene ontology (GO) analysis. High-Throughput GoMiner provided by the National Cancer Institute was used to determine the GO biological function categories for the genes that share position specific element in their core promoter region. This annotation tool not only provides us p-value (Fisher test), but also the 'false discovery rate' (FDR, or q-value) to evaluate multiple comparisons. Genes with p-value <0.05 and FDR < 0.20 were subjected to Ingenuity pathway analysis. The use of FDR addresses the issue of multiple testing.

Pathway analysis on gene targets of core elements. Given a set of genes that are enriched in GO analysis, we use them as seed genes (called *focus genes*) for generating biological networks through Ingenuity Pathway Analysis (Ingenuity Systems, http://www.ingenuity.com) tool. The Ingenuity score indicates the likelihood of the focus genes in a network being found at random. A score of 2 indicates that there is a 1 in 100 chance that the focus genes are together due to random chance. Biological functions

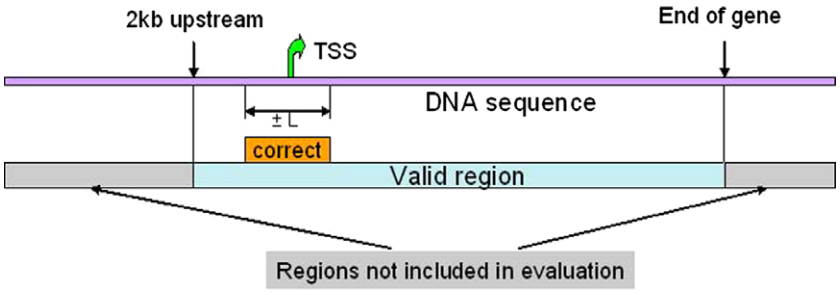

Fig. 1. An illustration of our evaluation scheme for genome-wide prediction. A prediction is deemed *correct* if it is within *cutoff* distance from the true TSS. *Valid prediction* is the number of predictions within 2000 bp upstream of the TSS and the end of the gene. Predictions outside the "valid" region are not included in the evaluation, following [20].

associated with genes within the newly formed networks are also provided by Ingenuity Pathway Knowledge Base.

Multiple sequence alignment. For the genes with significant GO annotation enrichment and Ingenuity score, we extract the sequences in the core promoter region. The sequences were aligned by CLUSTW under default setting.

CpG-rich and CpG-poor promoter comparison. (i) *Human mouse conservation*: for both CpG-rich and CpG-poor datasets, we calculated sequence identities at each position, which are the number of promoter conserved at the position divided by the total number of promoters. We then used Wilcoxon ranked sum paired one-way test to compare these two datasets for conservation. (ii) *Alternative transcription start sites (ATSS)*: we obtained all the ATSS from DBTSS website [14] and mapped them to UCSC human genome 16. We then calculate the occurrences of ATSS at each position; the occurrences were divided by total number of promoters. Wilcoxon ranked sum paired one-way test was used to test the abundances of ATSS in CpG-rich and CpG-poor promoters.

Significance of k-mer conservation. For an element x that occurs in n promoters at a specific position, let c be the fraction of bases among all n instances of the element that are conserved between human and mouse. We randomly pool n promoters and compute the conservation at the same position (5 consecutive bases). Out of 1000 trials the fraction of times the conservation of the pooled elements exceeds c provides the conservation p-value of element x.

Results

Overview of PSPA

Given a large collection of 10,342 experimentally determined transcription start sites, the ±100 bp regions flanking the TSS are used to train the PSPA model. For each position relative to the TSS, and for each *cis* element (words up to 5 bases long) occurring at the position in any of the promoters, we measure the position-specific *propensity* of the *cis* element. Also, for each pair of positions and each pair of *cis* elements at these positions, we measure their position-specific *association propensity*. The *propensity* is defined as the ratio of the observed and the expected frequencies. The comprehensive collection on propensities and association propensities from the training set forms the PSPA model. This model is then used to score a novel TSS, by taking the product of propensities of the *cis* elements and association propensities of the *cis* element pairs in the ±100 bp flanking the TSS. Our results are based on 7893 CpG-rich promoters and 1277 CpG-poor promoters. We have modeled these two promoter classes separately because of known differences in their expression pattern and their gene product function [16,17].

Human promoter prediction accuracy via cross-validation

We first evaluate our model on limited genomic context. The objective here is: given a 10 kb region containing exactly one experimentally verified TSS, how well a program can identify this TSS. We compared PSPA with previously described methods—FirstEF [3], DragonPF, and DragonGSF [4,20]. DragonGSF (gene start finder) is an improvement on DragonPF (promoter finder) [21] by the same authors and is recommended over DragonPF. Under default setting suggested by the authors, FirstEF, DragonGSF, and DragonPF do not always make a prediction on each sequence, and make multiple predictions on some sequences. FirstEF made a prediction on 94.2% and 11.4% of the genes for CPG-rich and CpG-poor promoters, respectively. The corresponding numbers for DragonGSF are 76.4% and 0.0%, and for DragonPF, 98.4% and 70.4%. For the sequences for which a prediction was made, FirstEF makes an average of 2.3 predictions per sequence, with a maximum of 15 predictions. DragonPF makes 4.2 predictions per sequence, with a maximum of 23 predictions. DragonGSF is the most stringent, making about 1 prediction per sequence.

Because of the reasons mentioned above, we compared PSPA separately with Dragon and FirstEF. For each sequence, if the other program (Dragon or FirstEF) makes k predictions, we considered the k best predictions made by PSPA to compare the prediction accuracies. If the program does not make a prediction, we exclude this sequence from the assessment. We considered a prediction correct if any of the k predictions is within L bases from the true TSS. A relatively large value of $L = 2$ kb has been used previously [3,4]. We tested the programs at various values of L ranging from 10 to 500 bp. The overall prediction accuracy of program on a set of TSS-containing sequences is simply the fraction sequences correctly predicted. Fig. 2a and b compare PSPA prediction accuracy with that of DragonGSF and FirstEF on CpG-rich promoters. Fig. 2c and d compare PSPA with DragonPF and FirstEF on CpG-poor promoters (DragonGSF did not make any prediction on CpG-poor promoters). In general, every tool performed poorly on CpG-poor promoters relative to CpG-rich promoters. In all cases PSPA out-performs the other two programs. The only exception is Fig. 1d. In this case, however, FirstEF does not make any prediction on ~89% of the CpG-poor sequences, and thus, those sequences were not included in the evaluation. The performance improvement by PSPA compared with other methods is especially acute for stringent cutoffs. This underscores the value of PSPA in precise prediction of TSS.

In addition to evaluating the overall performance, we have also quantified the relative contributions made by various aspects of the model. The strict positional restriction on the motifs imposed by our model is the major contributor to the overall performance improvement, with additional contribution made by the use of infrequent elements and co-occurrence of elements. To measure the effect of looser positional restriction, to score a k-mer at position i in the test sequence, we selected the best propensity of the k-mer in positions $i \pm 5$ in the training sequences. When we allow this positional 'variability' of 5 bases on CpG-poor promoters, the cross-validation accuracy goes down by 15%, and this value goes down by 20% for a 20 bases 'variability'. The CpG-rich promoters are less sensitive to positional restriction with a 2.8% decrease when a variability of 5 bases is allowed. With regard to the contribution of co-occurrence of elements, when we use only the position-specific propensity and

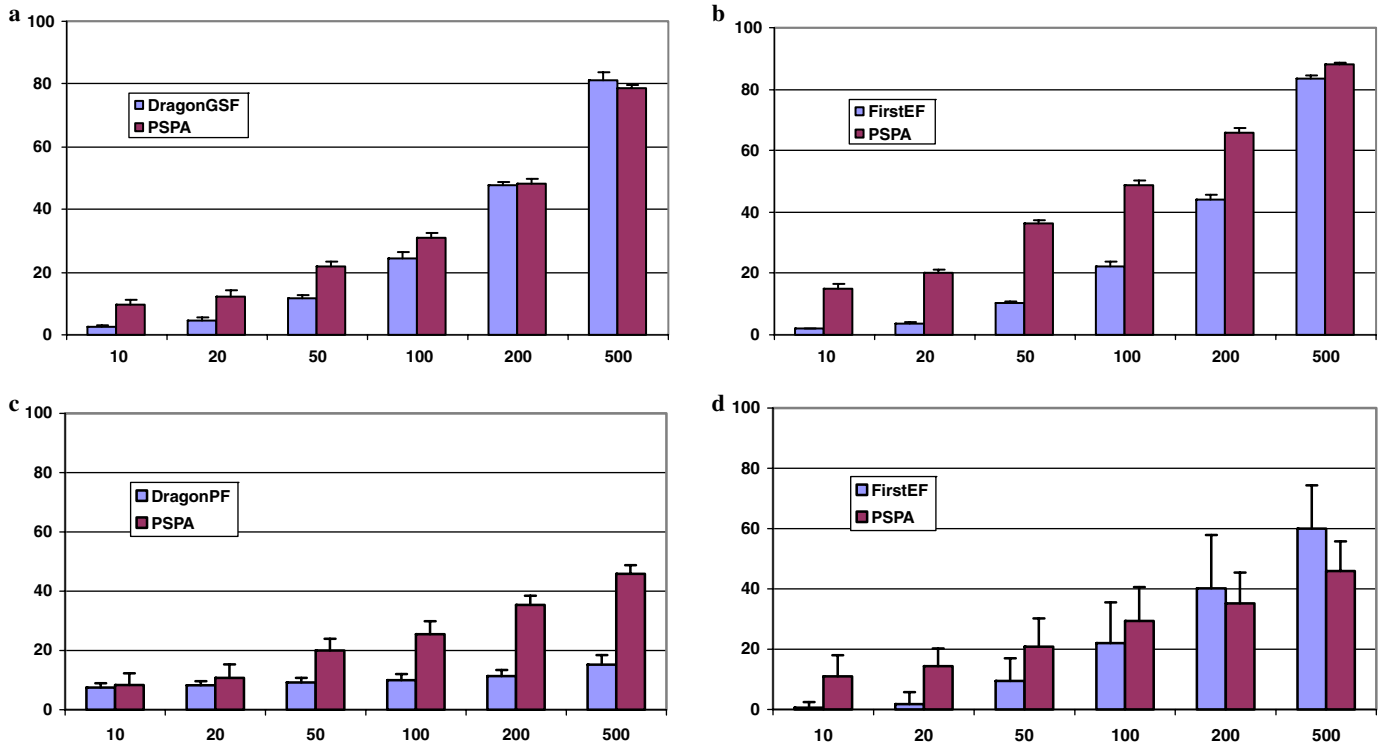

Fig. 2. Prediction evaluations on 10 kb human promoter sequences. While forcing PSPA to make the same number of predictions as that by Dragon or FirstEF, the figure shows the prediction accuracies of PSPA against (a) DragonGSF on CpG-rich promoters (b) FirstEF on CpG-rich promoters, (c) DragonPF on CpG-poor promoters, and (d) FirstEF on CpG-poor promoters. The *x*-axis represents the allowed distance between a predicted and true TSS to be considered correct, and *y*-axis represents the fraction of true TSS correctly predicted.

not the co-association score, the accuracy on CpG-poor promoter is reduced by a modest 3.2%. We currently use a maximum word length of 5, which achieves the best performance with our current training data size. A larger dataset will allow a use of larger word size, which may improve the performance.

Table 1
Comparison of PSPA, DragonGSF, and FirstEF on human chromosomes 4, 21, and 22, based on 747 non-redundant promoters from DBTSS

Cutoff (±L bp)	Program	Total predictions	% of valid predictions	Sensitivity (%)	PPV (%)	True positive cost (TPC)
10	PSPA	2831	19.71	5.49	7.35	12.61
	DragonGSF	2597	23.49	1.87	2.30	42.57
	FirstEF	7863	15.92	2.41	1.52	64.89
20	PSPA	2831	19.71	7.10	9.50	9.53
	DragonGSF	2597	23.49	3.21	3.93	24.42
	FirstEF	7863	15.92	4.42	2.72	35.82
50	PSPA	2831	19.71	12.58	16.85	4.94
	DragonGSF	2597	23.49	6.96	8.36	10.96
	FirstEF	7863	15.96	10.04	5.98	15.73
100	PSPA	2831	19.71	18.34	24.55	3.07
	DragonGSF	2597	23.49	13.92	16.89	4.92
	FirstEF	7863	15.96	14.99	9.08	10.01
200	PSPA	2831	19.71	28.65	38.35	1.61
	DragonGSF	2597	23.49	26.91	32.79	2.05
	FirstEF	7863	15.96	24.77	14.82	5.75
500	PSPA	2831	19.71	45.92	61.29	0.63
	DragonGSF	2597	23.49	51.41	62.79	0.59
	FirstEF	7863	15.96	37.48	22.87	3.37

A prediction is deemed *correct* if it is within *cutoff* distance from the true TSS. Sensitivity = (correctly predicted promoters)/(total number of promoters), ppv = (correct clusters)/(valid clusters), and TPC = (incorrect clusters)/(correct clusters). *Total predictions* measure the number of predictions (after clustering), *Correct promoter* is the number of TSS we predict correctly, *Correct prediction* is the number of predictions that are correct. *Valid prediction* is the number of predictions within 2000 bp upstream of the TSS and the end of the gene. *Percentage of valid clusters* = (valid clusters)/(total clusters predicted). Due to the overlap of some TSSs, the number of correct promoters is sometimes larger than the number of correct clusters. See Methods for definition of the terms.

Promoter prediction on human genome

Next, we evaluated how well these programs were able to identify experimentally determined TSS on human chromosomes 4, 21 and 22. We chose the accuracy measures—*sensitivity, specificity (ppv), and true positive cost (TPC)*—used in [20] (see Methods). As summarized in Table 1, PSPA is better at stringent cutoff values. At a cutoff of 50 bp, PSPA achieves a sensitivity and specificity (*ppv*) of 12.58% and 16.85% as compared to 6.96% and 8.36% for DragonGSF, and 10.04% and 5.98% for FirstEF. Furthermore, this is achieved at a much lower TPC. At a stringent cutoff of $L = 10$, PSPA predictions are 2- to 4-fold more accurate than the other two programs. At a cutoff of ±500 bp, because all three programs capture CpG islands, which are usually 500–1000 bp in length (average of 764 bp in human genome), PSPA is marginally worse than DragonGSF, but still far better than FirstEF. The true positive cost or TPcost for PSPA is better than those for FirstEF and DragonGSF for stringent cutoffs below 200 bp. We identified the overlap among the correctly predicted genes by the three programs at cutoff $L = 50$ (Fig. 3). Of the 94 correctly predicted genes by PSPA, 70 (74.5%) are unique, the numbers for FirstEF and DragonGSF are 55/75 (73.3%) and 36/52 (69.2%), respectively. Only one gene was predicted correctly by all three programs. This result indicates that three programs are complementary to each other, and suggests potential improvements in prediction accuracy by integrating all three predictors.

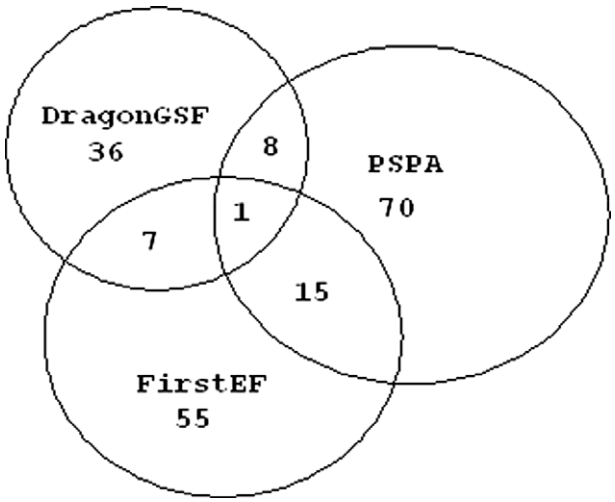

Fig. 3. Venn diagram shows the intersection of correctly predicted genes by three different methods in human chromosomes 4, 21, and 22 using a stringency cutoff of 50 bp. PSPA, DragonGSF, and FirstEF correctly predicted 94, 52, and 75 genes. Among them there are 70, 55, and 36 unique genes for PSPA, DragonGSF, and FirstEF, respectively. Only one gene is correctly predicted by all three methods. This implies three programs are supplementary to each other in promoter prediction.

Promoter prediction on mouse genome

To evaluate the PSPA algorithm on other vertebrate genomes, we applied the model (trained on all non-redundant human promoters) to predict the TSS in mouse genome Chromosome 5, within the range of 3 Mb and 70 Mb. We chose this region as it is being intensely studied as part of expanding the functional genomics approaches using model organisms [22,23]. We used RefSeq genes from UCSC genome browser to evaluate our prediction. There are 398 genes in this region, 256 (64.3%) of them are CpG-rich and 40 (10.1%) are CpG-poor. Table 2 shows that PSPA

Table 2
Comparison of PSPA, DragonGSF, and FirstEF for genome wide promoter prediction on mouse chromosome 5, region 3–70 Mb, which contains 398 genes

Cutoff (±L bp)	Program	Total predictions	% of valid predictions	Sensitivity (%)	PPV (%)	True positive cost (TPC)
10	PSPA	443	41.53	5.53	9.78	9.22
	DragonGSF	529	39.51	1.01	1.91	51.25
	FirstEF	2569	25.26	1.76	0.92	107.17
20	PSPA	443	41.53	7.04	12.50	7.00
	DragonGSF	529	39.51	2.51	4.78	19.90
	FirstEF	2569	25.26	4.27	2.47	39.56
50	PSPA	443	41.53	13.32	23.91	3.18
	DragonGSF	529	39.51	5.78	9.09	10.00
	FirstEF	2569	25.26	10.55	6.01	15.64
100	PSPA	443	41.53	18.59	34.78	1.87
	DragonGSF	529	39.51	12.06	19.14	4.22
	FirstEF	2569	25.26	16.83	9.09	10.00
200	PSPA	443	41.53	34.92	52.17	0.92
	DragonGSF	529	39.51	25.38	42.58	1.35
	FirstEF	2569	25.26	30.40	13.71	6.29
500	PSPA	443	41.53	48.74	82.07	0.22
	DragonGSF	529	39.51	47.49	72.25	0.38
	FirstEF	2569	25.26	43.22	20.96	3.77

See Table 1 legend and method section for further details of column definitions.

performs better than both DragonGSF and FirstEF in all the evaluation ranges in terms of ppv. Compared with the performance in the human genome, there is substantial decrease in the sensitivity for DragonGSF predictions in mouse at stringent cutoff range, and for FirstEF there is slight decrease in ppv and increase in sensitivity. However,

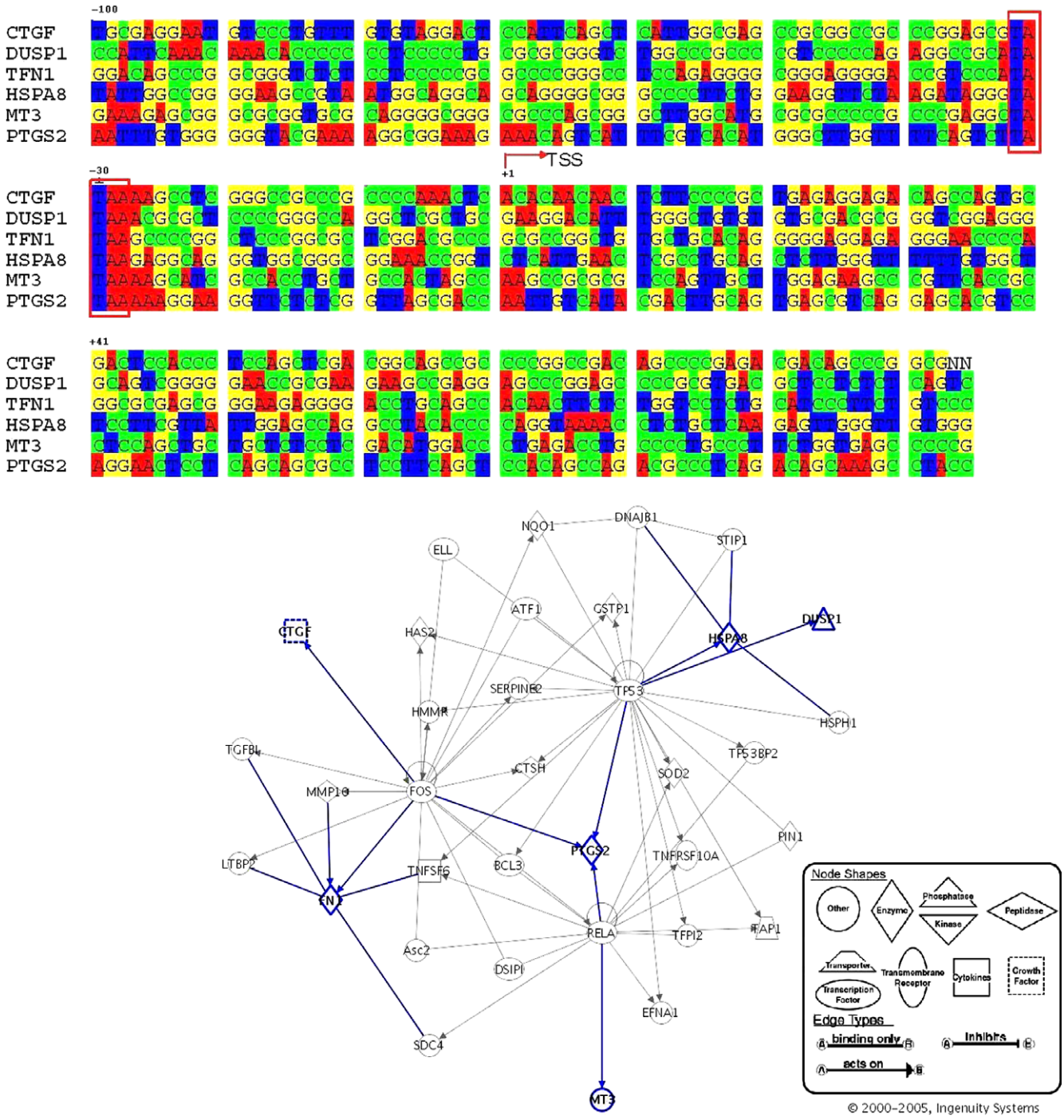

Fig. 4. The genes containing TATA-box-related motif TATAA at position −32 are correlated with stress response network. The top figure shows the six gene promoter sequences. The lower figure shows the genes in the stress response network indicating the positions of the six genes in shaded boxes. Promoter element TATAA has the highest propensity at position −32, with a total CpG-rich sequence count of 32. Among 32 genes, six genes—CTGF, DUSP1, TFN1, HSPA8, MT3, and PTGS2—are associated with response to stress (with $\log_{10}(p\text{-value}) = -2.43$). The six genes have very little sequence similarity in the core promoter region except for the TATAA motif at position −32 (multiple alignment using CLUSTALW with default setting misaligns the TATA box). Ingenuity analysis reveals that all six genes are in the same stress response pathway. The Ingenuity tool also finds significant links of the input genes with various biological functions. All of the six genes are associated with cell death (p-value = 4.7E−10), five (except MT3) of them are associated with apoptosis of tumor cell lines (6E−8), four (FN1, HSPA8, MT3, and PTGS) are related to cancer (4E−7) and proliferation of cells (4E−5) and other cancers (see Supplementary Table S2a).

Table 3
Evolutionarily conserved elements in CpG-rich promoters with significant association with a GO functional class and a gene network

General information		Propensity analysis		GO analysis			Ingenuity analysis data				
Position to TSS	k-mer	$\log_2$ (prop en.)	Propen. rank	$\log_{10}$ (p-value)	FDR (<)	GO term	Genes	Total genes in network	Ingenuity score	Genes in network	Top functions
−19	GGCGG	2.33	1	−2.44	0.178	GO:0051244 regulation of cellular physiological process	8	5	10	CAPG, CASP3, CUL1, PRKRA, UBE2C	Cancer, cell death, reproductive system disease
−18	TTCCG	1.68	13	−3.18	0.115	GO:0008152 metabolism	17	7	15	GTF2A2, GTF2H1, POLR3D, PTPN1, RUVBL2, TTF1, UQCRH	Gene expression, cellular growth and proliferation, hepatic system development and function
−5	CCGCC	2.99	1	−3.88	0.009	GO:0000375 RNA splicing via transesterification reactions	8	5	11	DDX1, DHX15, SFRS3, SNRPD3, SYNCRIP	RNA post-transcriptional modification, cancer, cell death
−4	CGCCA	4.12	1	−5.93	0.001	GO:0008380 RNA splicing	14	6	12	DDX1, DHX15, PPP2CA, SFRS3, SNRPD3, SYNCRIP	RNA post-transcriptional modification, cancer, cell death
−4	CGCCA	4.12	1	−5.93	0.001	GO:0008380 RNA splicing	14	10	10	DDX1, DHX15, PPP2CA, SF3A1, SF3B1, SFRS1, SFRS3, SFRS6, SNRPD3, SYNCRIP	RNA post-transcriptional modification, viral function, cancer
−3	GCCAT	4.43	1	−8.30	0.001	GO:0008380 RNA splicing	16	9	18	DHX15, HNRPC, PPP2CA, SFRS1, SFRS3, SNRPD3, SYNCRIP	RNA post-transcriptional modification, cancer, cellular growth and proliferation
−3	GCCAT	4.43	1	−3.36	0.015	GO:0045449 regulation of transcription	31	14	28	ATF4, CHD4, CRK, HNRPD, ILF3, MDM4, NFYA, PPP2CA, RPS6KA4, SAFB, TERF1, UBTF, ZNF197, ZNF8	Gene expression, cellular growth and proliferation, connective tissue development and function
−2	CCATT	4.62	1	−9.39	0.001	GO:0006397 mRNA processing	18	10	21	HNRPR, PRPF8, SFRS1, SFRS3, SFRS6, SNRPD3, SNRPF, SYNCRIP	RNA post-transcriptional modification, DNA replication, recombination, and repair, cancer
−2	CCATT	4.62	1	−4.50	0.001	GO:0045449 regulation of transcription	32	16	33	ATF4, CHD4, CRK, E2F2, HNRPD, ILF3, MDM4, NFYA, PKNOX1, PPP2CA, SAFB, SIN3A, TERF1, UBTF, ZNF193, ZNF8	Gene expression, cellular growth and proliferation, connective tissue development and function

−1	CATTT	5.06	1	−8.76	0.001	GO:0006397 mRNA processing	17	9	18	DHX15, HNRPC, HNRPR, KHDRBS1, SFRS1, SFRS3, SNRPB, SNRPF, SYNCRIP	RNA post-transcriptional modification, gene expression, molecular transport
−1	CATTT	5.06	1	−3.90	0.001	GO:0043283 biopolymer metabolism	32	12	22	DHX15, HNRPC, HNRPD, HNRPR, KHDRBS1, MDM4, PSMB4, SFRS1, SFRS3, SNRPB, SNRPF, SYNCRIP	RNA post-transcriptional modification, skeletal and muscular system development and function, tissue development
−1	CCTTT	3.34	12	−19.25	0.001	GO:0006412 protein biosynthesis	31	16	38	EEF1G, RPL13, RPL13A, RPL18A, RPL22, RPL23, RPL23A, RPL27A, RPL31, RPL37A, RPL4, RPLP2, RPS16, RPS19, RPS27, RPS3	Protein synthesis, cellular development, connective tissue development and function
1	ATTTT	4.54	1	−6.95	0.001	GO:0006396 RNA processing	13	3	6	HNRPC, HNRPD, SFRS1	RNA post-transcriptional modification, cancer, cell death
1	ATTTT	4.54	1	−3.96	0.003	GO:0043283 biopolymer metabolism	20	10	21	CHD4, CUGBP1, DDX17, HNRPC, HNRPD, MSH2, PSMB4, SAFB, SFRS1, YME1L1	Cellular growth and proliferation, RNA post-transcriptional modification, cancer
2	TTTTG	3.64	2	−6.20	0.001	GO:0006396 RNA processing	11	3	8	E1B-AP5, HNRPC, SFRS1	RNA post-transcriptional modification, post-translational modification, DNA replication, recombination, and repair
2	TTTTT	3.19	5	−3.40	0.015	GO:0006412 protein biosynthesis	7	6	14	EEF1A1, RPL23, RPL27A, RPL38, RPS11, RPS20	Protein synthesis, cancer, immunological disease
24	ATGGC	2.42	3	−2.80	0.089	GO:0044237 cellular metabolism	18	8	16	CAPN2, COX5B, PPP1R8, PSMA2, PSMB8, RPS11, RPS6KB1, SF3B4	Protein synthesis, cancer, cell death

The table shows the elements with GO $\log_{10}(p\text{-value}) < -2.0$, FDR < 0.20 and ingenuity score $>= 8$.

for PSPA, since we used a full model in this test (we used cross-validation in other tests), we observed significant improvement in PPV, and moderate improvement in sensitivity, as compared to the predictions in human.

Promoter elements are linked to gene product function

Further investigation of genes that share a position-specific promoter element reveals that these genes are often involved in similar function or genetic networks.

Position-specific overrepresented elements

We selected at each position the elements that were (i) among the five most over-represented elements at the position, and (ii) matched one of the core elements—TATA, GC-box, CAAT-box, INR, and MTE. We investigated whether the genes containing the element at the specific position share a gene ontology (*GO*) category [24,25]. In the CpG-poor promoters, position-specific core elements are linked to biosynthesis, cell adhesion, synaptic transmission, and T-cell activation (Supplementary Table S1b). In CpG-rich promoters (Supplementary Table S1a), TATA box element TATAA is one of the most over-represented at positions −33 to −28. Even though TATAA is associated with development in general, it correlates significantly with stress response at position −32, morphogenesis at position −31, and cell growth and proliferation at position −30. Similarly, the GC-box is linked with localization at position −23, mRNA splicing at position −6, and organic acid metabolism at position −41. At position −32, TATAA is shared by 32 CpG-rich genes, six of which—*CTGF*, *DUSP1*, *TFN1*, *HSPA8*, *MT3*, and *PTGS2*—are associated with the response to stress in GO analysis (with $\log_{10}(p\text{-value}) = -2.43$, and false discovery rate (FDR) of 0.16, shown in Supplementary Table S1a). Using the Ingenuity tool (www.ingenuity.com), we found that all six genes are in the stress response pathway (Fig. 4), which is significantly associated

Table 4
Comparison of CpG-rich and CpG-poor genes, in terms of top 10 GO categories—(a) cellular location, (b) biological process, and (c) molecular function

CpG rich		CpG poor	
(a) GO cellular component			
Gene category	EASE score	Gene category	EASE score
Intracellular	1.92E−142	Extracellular	4.38E−23
Cytoplasm	2.66E−106	Extracellular space	3.08E−17
Mitochondrion	8.26E−45	Cell fraction	9.30E−06
Ribonucleoprotein complex	9.24E−44	Microsome	3.54E−05
Cell	7.73E−39	Vesicular fraction	4.09E−05
Ribosome	2.48E−27	Integral to plasma membrane	1.65E−04
Cytosol	9.26E−21	Integral to membrane	2.94E−04
Cytosolic ribosome (sensu Eukarya)	3.96E−18	Plasma membrane	6.93E−04
Nucleolus	3.12E−16	Membrane fraction	1.37E−03
Nucleus	5.16E−14	Extracellular matrix	1.49E−03
(b) GO biological process			
Gene category	EASE score	Gene category	EASE score
Biosynthesis	2.87E−33	Defense response	5.45E−25
Metabolism	7.86E−32	Response to biotic stimulus	2.89E−24
Protein biosynthesis	2.14E−29	Immune response	1.78E−23
Macromolecule biosynthesis	8.74E−28	Response to external stimulus	3.74E−18
Intracellular transport	1.63E−27	Response to pest/pathogen/parasite	1.19E−12
Intracellular protein transport	1.41E−25	Response to wounding	1.46E−09
RNA metabolism	1.01E−24	Inflammatory response	2.45E−08
RNA processing	6.51E−24	Innate immune response	1.10E−07
Protein transport	9.15E−24	Immune cell activation	1.36E−06
Mitotic cell cycle	7.18E−23	Cell activation	1.36E−06
(c) GO molecular function			
Gene category	EASE score	Gene category	EASE score
RNA binding	8.88E−35	Signal transducer activity	3.61 E−06
Structural constituent of ribosome	3.73E−29	Endopeptidase inhibitor activity	2.67E−04
Catalytic activity	5.24E−23	Receptor activity	2.83E−04
Protein transporter activity	7.23E−20	Protease inhibitor activity	3.00E−04
Purine nucleotide binding	6.83E−14	Chitin binding	4.83E−04
Nucleotide binding	1.67E−13	Receptor signaling protein activity	7.33E−04
Ligase activity	2.82E−12	Peptidase activity	9.14E−04
Hydrogen ion transporter activity	3.07E−11	Chemokine receptor binding	9.37E−04
Primary active transporter activity	5.14E−10	Chemokine activity	9.37E−04
ATP binding	1.31E−09	Elastase activity	1.06E−03

EASE score represents the *p*-value of the overlap.

with apoptosis and cancers (Supplementary Table S2). Links between TATA-containing genes and stress response have been previously reported in yeast [26] and our study suggests a similar association in humans and refines the association with TATAA at −32 position.

Position-specific evolutionarily conserved elements

We have observed similar links for promoter elements that are preferentially conserved between human and mouse (see Supplementary Tables S3a,b for detailed lists of conserved elements from CpG-rich and CpG-poor

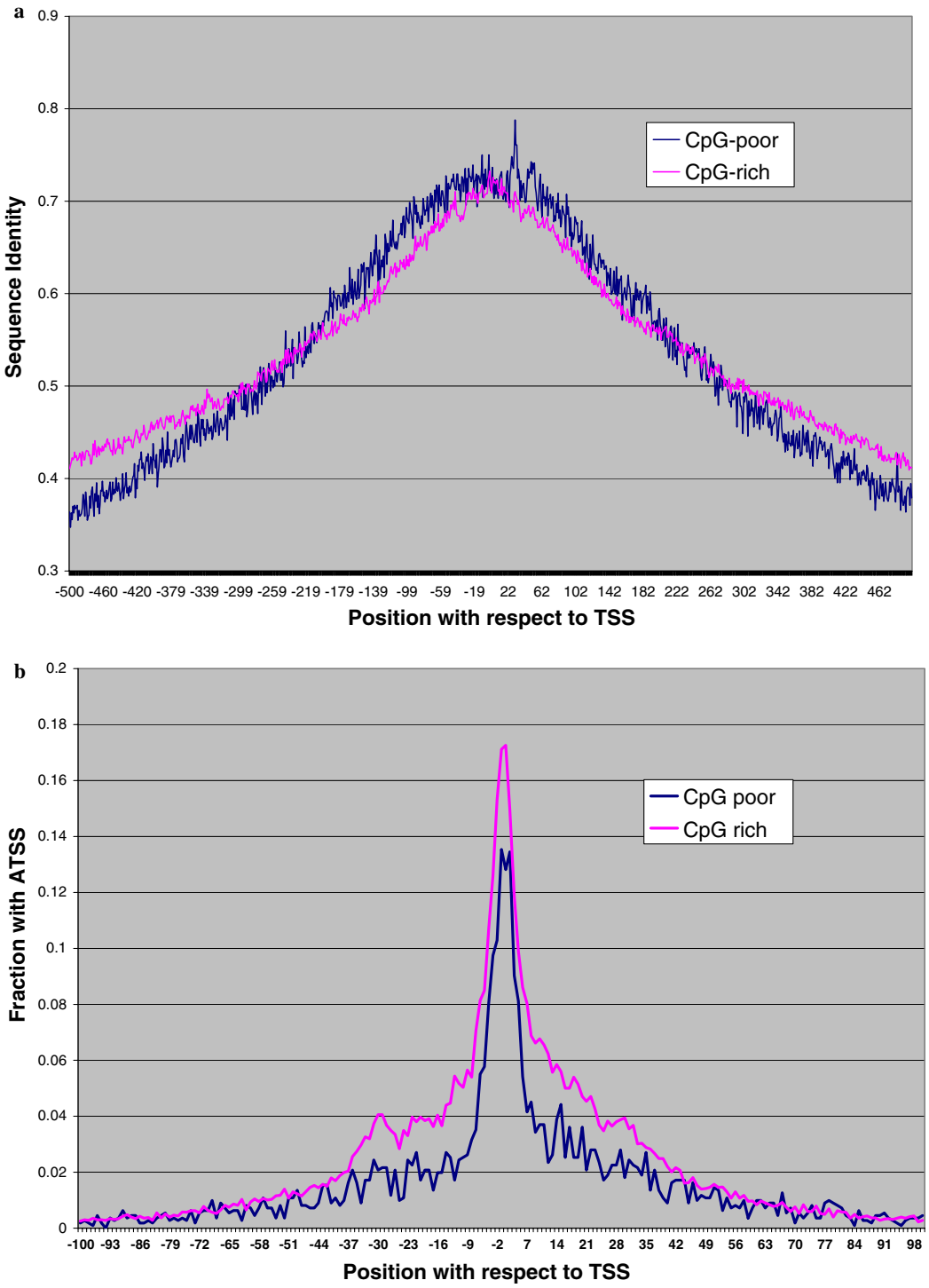

Fig. 5. Comparison of CpG-rich and CpG-poor promoters. (a) For each of the ±500 positions relative to the TSS, the figure shows the fraction of promoters that are conserved between human and mouse. A significantly higher conservation of CpG-poor promoters is observed in ±200 bp ($p < 10^{-16}$). (b) The fraction of genes (y-axis) with an alternative TSS at a position (x-axis) relative to the major TSS. In the immediate vicinity of the most frequent start site, CpG-poor promoters have fewer alternate TSS.

promoters). In CpG-rich promoters, CCGCC at position −5 is frequently conserved and associated with RNA splicing, CGCCA at −4 with biopolymer metabolism, and CCTTT at −1 with protein biosynthesis. A number of these elements have significant association with specific functional pathways based on Ingenuity analysis (Table 3). Most of the highly conserved elements (p-value $\leqslant 0.001$) with significant GO association also have high propensity. A highly conserved element CCTTT at position −1 is strongly associated with the protein synthesis pathway, and it also has high propensity at position −1. There are 66 CpG-rich genes that share this element, 26 of which are involved in protein biosynthesis (GO p-value $= 10^{-19}$). Sixteen of the 26 genes belong to a single ingenuity genetic network for protein synthesis (p-value $= 10^{-38}$). Nine of these—RPL13, RPL18A, RPL22, RPL27A, RPL4, RPLP2, RPS19, RPS27, and RPS3—are ribosomal protein genes (RP) (Supplementary Fig. S1). Studies by Yoshihama et al. have demonstrated that all RP genes' transcription starts at a C residue within a characteristic oligopyrimidine tract [27]. Recent study by Perry refined the transcriptional initiation consensus of RP genes to $(Y)_2C^{+1}TY(T)_2(Y)_3$ [28]. Our finding further identifies a dominant subclass—CCTTT—of this motif.

Comparison of CpG-rich and CpG-poor promoters

We have compared these two promoter classes with respect to three properties—functional classes of their gene products, evolutionary conservation, and the frequency of alternative transcription starts with the promoters. Our GO analysis shows that CpG-poor promoter gene products are mostly secreted and involved in defense response and signal transduction. In contrast CpG-rich gene products are mostly intra-cellular and responsible for biosynthesis, metabolism, and RNA binding, and are structural constituents (Table 4). As shown in Fig. 5a, CpG-poor promoters are significantly more conserved between human and mouse compared with the CpG-rich promoters. The elements correspond to the conservation peaks at positions −2, −10, +30, +51, and +53. Finally, the CpG-poor promoters have significantly fewer alternate TSS than CpG-rich promoters (Fig. 5b). This analysis is based on the experimentally determined alternative TSS provided in DBTSS database (http://dbtss.hgc.jp). Fewer alternative start sites and greater conservation in CpG-poor promoters—suggest a greater regulatory control in CpG-poor genes. This is consistent with tissue-restricted expression of CpG-poor gene promoters.

Discussion

We have described a probabilistic model—PSPA—that by capturing the location preference and co-occurrence of promoter elements can better predict transcription start sites. Accurate promoter models can guide experimental approaches to detect novel genes and alternative transcription start site for known genes. The importance of the latter is highlighted by recent discovery of a second, far upstream promoter of the MODY 1 gene *HNF4α*. Polymorphisms in this promoter are linked to type-II diabetes [29]. Furthermore, our modeling approach can be easily extended to several other genomic site detection problems.

We have found that many position-specific promoter elements are strongly linked with the gene product function. In prokaryotes, a single transcriptional regulator can recruit RNA polymerase via direct contact. In higher organisms, multiple regulators control gene transcription [30]. This might have evolved in response to multiple and often opposing environmental conditions. However, simultaneous activation or repression of a group of genes in response to a stimulus can be effectively accomplished through a common promoter element in these genes. A single transcription factor HNF1α binds to 1.6% of all human genes and likely regulates many of them, thus serving as a master regulator [31]. This master control of multiple genes via a shared promoter element is reminiscent of prokaryotic gene regulation via a single regulator. We have shown that genes sharing a position-specific promoter element have similar function or act within the same genetic networks. By further investigating such shared promoter elements, we will be able to partially unravel the gene networks in the mammalian genome.

Acknowledgments

The authors thank Y. Suzuki for help with DBTSS database and J. Kadonaga, R. Bushman, M. Bucan, S. Date, and Y. Sun for helpful comments. The research was supported by a funding from the Commonwealth of PA.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2006.06.062.

References

[1] J.T. Kadonaga, Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors, Cell 116 (2004) 247–257.
[2] S. Hannenhalli, S. Levy, Promoter prediction in the human genome, Bioinformatics 17 (Suppl. 1) (2001) S90–S96.
[3] R.V. Davuluri, I. Grosse, M.Q. Zhang, Computational identification of promoters and first exons in the human genome, Nat. Genet. 29 (2001) 412–417.
[4] V.B. Bajic, S.H. Seah, Dragon gene start finder: an advanced system for finding approximate locations of the start of gene transcriptional units, Genome Res. 13 (2003) 1923–1929.
[5] P. Qiu, Recent advances in computational promoter analysis in understanding the transcriptional regulatory network, Biochem. Biophys. Res. Commun. 309 (2003) 495–501.
[6] C.Y. Lim, B. Santoso, T. Boulay, E. Dong, U. Ohler, J.T. Kadonaga, The MTE, a new core promoter element for transcription by RNA polymerase II, Genes Dev. 18 (2004) 1606–1617.

[7] M.L. Grace, M.B. Chandrasekharan, T.C. Hall, A.J. Crowe, Sequence and spacing of TATA box elements are critical for accurate initiation from the beta-phaseolin promoter, J. Biol. Chem. 279 (2004) 8102–8110.
[8] J.T. Kadonaga, The DPE, a core promoter element for transcription by RNA polymerase II, Exp. Mol. Med. 34 (2002) 259–264.
[9] J.E. Butler, J.T. Kadonaga, Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs, Genes Dev. 15 (2001) 2515–2519.
[10] C.A. Spek, R.M. Bertina, P.H. Reitsma, Unique distance- and DNA-turn-dependent interactions in the human protein C gene promoter confer submaximal transcriptional activity, Biochem. J. 340 (Pt 2) (1999) 513–518.
[11] L. Wu, A. Berk, Constraints on spacing between transcription factor binding sites in a simple adenovirus promoter, Genes Dev. 2 (1988) 403–411.
[12] T. Sugiyama, D.K. Scott, J.C. Wang, D.K. Granner, Structural requirements of the glucocorticoid and retinoic acid response units in the phosphoenolpyruvate carboxykinase gene promoter, Mol. Endocrinol. 12 (1998) 1487–1498.
[13] K. Senger, G.W. Armstrong, W.J. Rowell, J.M. Kwan, M. Markstein, M. Levine, Immunity regulatory DNAs share common organizational features in *Drosophila*, Mol. Cell 13 (2004) 19–32.
[14] Y. Suzuki, R. Yamashita, S. Sugano, K. Nakai, DBTSS, DataBase of Transcriptional Start Sites: progress report 2004, Nucleic Acids Res. 32 (2004) D78–D81.
[15] Y. Wang, F.C. Leung, An evaluation of new criteria for CpG islands in the human genome as gene markers, Bioinformatics 20 (2004) 1170–1177.
[16] F. Antequera, Structure, function and evolution of CpG island promoters, Cell Mol. Life Sci. 60 (2003) 1647–1658.
[17] J. Schug, W.P. Schuller, C. Kappen, J.M. Salbaum, M. Bucan, C.J. Stoeckert Jr., Promoter features related to tissue specificity as measured by Shannon entropy, Genome Biol. 6 (2005) R33.
[18] S. Schwartz, W.J. Kent, A. Smit, Z. Zhang, R. Baertsch, R.C. Hardison, D. Haussler, W. Miller, Human-mouse alignments with BLASTZ, Genome Res. 13 (2003) 103–107.
[19] J. Wang, S. Hannenhalli, Generalizations of Markov model to characterize biological sequences, BMC Bioinformatics 6 (2005) 219.
[20] V.B. Bajic, S.L. Tan, Y. Suzuki, S. Sugano, Promoter prediction analysis on the whole human genome, Nat. Biotechnol. 22 (2004) 1467–1473.
[21] V.B. Bajic, S.H. Seah, A. Chong, G. Zhang, J.L. Koh, V. Brusic, Dragon Promoter Finder: recognition of vertebrate RNA polymerase II promoters, Bioinformatics 18 (2002) 198–199.
[22] J.H. Nadeau, R. Balling, G. Barsh, D. Beier, S.D. Brown, M. Bucan, S. Camper, G. Carlson, N. Copeland, J. Eppig, C. Fletcher, W.N. Frankel, D. Ganten, D. Goldowitz, C. Goodnow, J.L. Guenet, G. Hicks, M. Hrabe de Angelis, I. Jackson, H.J. Jacob, N. Jenkins, D. Johnson, M. Justice, S. Kay, D. Kingsley, H. Lehrach, T. Magnuson, M. Meisler, A. Poustka, E.M. Rinchik, J. Rossant, L.B. Russell, J. Schimenti, T. Shiroishi, W.C. Skarnes, P. Soriano, W. Stanford, J.S. Takahashi, W. Wurst, A. Zimmer, Sequence interpretation. Functional annotation of mouse genome sequences, Science 291 (2001) 1251–1255.
[23] J.C. Schimenti, B.J. Libby, R.A. Bergstrom, L.A. Wilson, D. Naf, L.M. Tarantino, A. Alavizadeh, A. Lengeling, M. Bucan, Interdigitated deletion complexes on mouse chromosome 5 induced by irradiation of embryonic stem cells, Genome Res. 10 (2000) 1043–1050.
[24] D.A. Hosack, G. Dennis Jr., B.T. Sherman, H.C. Lane, R.A. Lempicki, Identifying biological themes within lists of genes with EASE, Genome Biol. 4 (2003) R70.
[25] B.R. Zeeberg, H. Qin, S. Narasimhan, M. Sunshine, H. Cao, D.W. Kane, M. Reimers, R. Stephens, D. Bryant, S.K. Burt, E. Elnekave, D.M. Hari, T.A. Wynn, C. Cunningham-Rundles, D.M. Stewart, D. Nelson, J.N. Weinstein, High-Throughput GoMiner, an 'industrial-strength' integrative Gene Ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency (CVID), BMC Bioinformatics 6 (2005) 168.
[26] A.D. Basehoar, S.J. Zanton, B.F. Pugh, Identification and distinct regulation of yeast TATA box-containing genes, Cell 116 (2004) 699–709.
[27] M. Yoshihama, T. Uechi, S. Asakawa, K. Kawasaki, S. Kato, S. Higa, N. Maeda, S. Minoshima, T. Tanaka, N. Shimizu, N. Kenmochi, The human ribosomal protein genes: sequencing and comparative analysis of 73 genes, Genome Res. 12 (2002) 379–390.
[28] R.P. Perry, The architecture of mammalian ribosomal protein promoters, BMC Evol. Biol. 5 (2005) 15.
[29] L.D. Love-Gregory, J. Wasson, J. Ma, C.H. Jin, B. Glaser, B.K. Suarez, M.A. Permutt, A common polymorphism in the upstream promoter region of the hepatocyte nuclear factor-4 alpha gene on chromosome 20q is associated with type 2 diabetes and appears to contribute to the evidence for linkage in an ashkenazi jewish population, Diabetes 53 (2004) 1134–1140.
[30] B. Lemon, R. Tjian, Orchestrated response: a symphony of transcription factors for gene control, Genes Dev. 14 (2000) 2551–2569.
[31] D.T. Odom, N. Zizlsperger, D.B. Gordon, G.W. Bell, N.J. Rinaldi, H.L. Murray, T.L. Volkert, J. Schreiber, P.A. Rolfe, D.K. Gifford, E. Fraenkel, G.I. Bell, R.A. Young, Control of pancreas and liver gene expression by HNF transcription factors, Science 303 (2004) 1378–1381.