



Promoter prediction in the human genome

Sridhar Hannenhalli and Samuel Levy

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

Received on February 5, 2001; revised and accepted on March 29, 2001

ABSTRACT

Computational prediction of eukaryotic polII promoters has been one of the most elusive problems despite considerable effort devoted to the study. Researchers have looked for various types of signals around the transcriptional start site (TSS), viz. oligo-nucleotide statistics, potential binding sites for core factors, clusters of binding sites, proximity to CpG islands etc.. The proximity of CpG islands to gene starts is now a well established fact, although until recently, it was based on very little genomic data. In this work we explore the possibility of enhancing the promoter prediction accuracy by combining CpG island information with a few other, biologically motivated, seemingly independent signals, that cover most of the known knowledge. We benchmarked the method on a much larger genomic datasets compared to previous studies. We were able to improve slightly upon current prediction accuracy. Furthermore, we observe that CpG islands are the most dominant signals and the other signals do not improve the prediction. This suggests that the computational prediction of promoters for genes with no associated CpG-island (typically having tissue-specific expression) looking only at the immediate neighborhood of the TSS may not even be possible. We suggest some biological experiments and studies to better understand the biology of transcription.

Contact: Sridhar.Hannenhalli@celera.com
Samuel.Levy@celera.com

INTRODUCTION

Computational identification of promoters remains important, not only to detect rarely expressed genes but also for regulatory analysis of genes whose full length transcripts are not known, which is currently true for most of the known genes. A lot of effort has been devoted to this problem in the last few years with moderate success.

Readers are referred to Fickett and Hatzigeorgiou, 1997 for an excellent, although a little dated, review of the field. The conclusion drawn in this review was that the performance of the all the predictors studied was not significantly different and typically, to achieve a sensitivity of 50% they made a prediction every 500bp which is clearly unacceptable due to low specificity. For

a more recent review see Pedersen et al. 1999.

The most notable advancement in the field since then, has been made by Scherf et al. 2000, in their promoter prediction tool, PromoterInspector. They use a slightly generalized notion of oligomers by allowing for variable gaps between fixed oligomers, and upon careful training, report ~40% sensitivity and ~40% specificity on large genomic sequences, making a prediction every 40Kbs which is clearly a great improvement over the previous methods.

CpG islands are stretches of un-methylated DNA with a higher frequency of CpG di-nucleotides when compared to the entire genome (Bird et al. 1987). CpG islands are known to preferentially occur at the transcriptional start of genes, almost 50% of all genes and almost all house-keeping genes have CpG islands at the 5' end of the transcript (Gardiner-Garden and Frommer 1987), (Larsen et al. 1992). In addition, experimental evidence indicates that CpG island methylation is correlated with the level of gene expression (Cross and Bird 1995).

Ioshikhes and Zhang, 2000 (Ioshikhes and Zhang 2000) exploited this by discriminating between CpG islands associated with genes and the ones not associated with “known” genes. They were able to do this with high accuracy by using CpG island length and the CpG score of the island as the discriminating parameters. As a consequence they could predict promoters with 47% sensitivity while making a prediction every 26Kbps. Surprisingly, this is comparable to PromoterInspector considering its simplicity.

The main motivation for this work was to explore the possibility of improving promoter prediction accuracy by using varied evidences from the body of literature in the field.

Most of the work in the past has centered around some notion of word frequencies in the promoter region, eg., Autogene (Kondrakhin et al. 1995), PromFind (Hutchinson 1996), TSSG and TSSW (Solovyev and Salamov 1997), CorePromoter (Zhang 1998) and PromoterInspector (Scherf et al. 2000). Most of the earlier programs were developed before the availability of large sequences and reliable annotations and hence were mostly trained and evaluated against a small region around TSSs. For the first

time they were evaluated in the context of large genomic sequences in (Fickett and Hatzigeorgiou 1997). Features related to DNA structure are believed to play a role in transcription initiation. Pedersen et al. 1998 looked at the DNA structure around the TSS for a set of genes. They looked at three different measures, viz., DNaseI derived bendability, location preference and propeller-twist, all of which were expressed in terms of small oligomers. The feature values were found to differ in the upstream compared to the downstream region of the TSS. These facts motivated us to include the word frequencies upstream and downstream of the TSS as a discriminating factor. In general one would expect the word frequencies to be highly biased by the GC-richness of the promoter region but it's also conceivable that this may reveal words that are discriminating for the non-GC-rich promoters.

Since transcription involves binding of various transcription factors (TFs) to the DNA, it makes sense to look for the potential binding sites of these so called core factors to predict promoters. In fact many of the existing programs do that, eg. NNPP (Reese, <http://www-hgc.lbl.gov/inf/nnpp-abstract.html>), PromoterScan (Prestridge 1995), TATA (Bucher 1990) and CorePromoter (Zhang 1998). Typically they have focussed on the TATA signal. In general, the word-statistics based approach will reveal high occurrence of certain binding sites but what the word-statistics does not capture is the positional preference of the binding sites with respect to the TSS. The positional preference of TATA signal is well known (Tsunoda and Takagi, 1999, among others). We have generalized this approach in this work. We computed the likelihood of observing a particular binding site in a particular window with respect to the TSS from the training set and used this likelihood distribution to score a potential TSS in the test set.

Core promoter regions may contain novel binding site sequences that have not yet been discovered as biologically relevant for basal transcriptional control. In this context it is prudent to search the region around the TSS for novel signals using a signal finder such as AnnSpec (Workman and Stormo 2000).

Another distinguishing feature about the promoter region that has been noted is the tendency to find a greater density of TF binding sites in these regions ((Ioshikhes et al. 1999), (Pickert et al. 1998), (States et al. 2000)). In our study we also use the TF site density upstream of a potential TSS as a discriminating parameter.

All the parameters discussed above are seemingly independent and have clear biological motivation. Each of the parameters have been explored individually. Besides generalizing some of them, we have tried to combine these various factors in a unified framework in an effort to improve computational promoter prediction.

SYSTEM AND METHODS

Material

We used three independent data sets for this study. The first set, SET1 consisted of 150 human core promoters from EPD (Bucher and Trifonov 1986) and 77 experimentally determined TSS curated from the published literature that could be confidently mapped to the human genome. We used the Celera human assembly for this purpose (Venter 2001). We used the 100Kb region on either side of the TSS whenever possible. This resulted in a database of 227 sequences with an overall length of ~38Mbs with 227 known TSSs. Our second set, SET2 consisted of publicly available sequence for human chromosome 22. The 35Mb sequence, and the set of 238 TSSs obtained by virtue of aligning known full length mRNAs was obtained from the Sanger Center Web site (<http://www.sanger.ac.uk/HGP/Chr22/>). Our third set, SET3 consisted of 6 Genbank genomic sequences with a total length of 1.38Mb and 35 known TSSs on these sequences that was used for evaluation in Scherf et al. 2000.

Benchmarks

We used three different benchmarks. For the first one, BenchMark1, we divided up randomly SET1 into a training set, SET1_TRAIN of 150 (not the same as EPD derived sequences) sequences and a test set, SET1_TEST of 77 sequences. For the second one, BenchMark2, we used the 227 sequences of SET1 for training and tested it on Chromosome 22, SET2. For the third one, BenchMark3, we used the same training as Benchmark2 and tested it on SET3. A set of 2000 random positions in the genomic sequences corresponding to the training set were used as the negative (or, the background) set of TSSs.

Training

In the following we describe the various computations and training procedure common to all three benchmarks. Every (potential) transcription start site has a few associated parameters. CpG islands are defined in (Larsen et al. 1992) as stretches of 200 bps windows with G+C fraction ≥ 0.5 and the CpG value (ratio of observed versus expected CG di-nucleotides) ≥ 0.6 . Consecutive CpG islands are merged if they are within 200 bps to each other. The main difference in our implementation is that we re-compute the CpG value upon merging and if the CpG value is not over the threshold of 0.6 we output the maximal prefix sub-window whose CpG value is over the threshold. There are three CpG related parameters associated with a TSS, the CpG island length in bps, the CpG value and the distance of the closest CpG island.

Two 5-mer dictionaries were built, one using the 600 bps region upstream and another using the 600 bps

downstream region of the known TSSs in the training set against the background of randomly picked 2000, 600 bps regions from the training set genomic sequences. The score associated with a 5-mer is log of the ratio of its frequencies in the positive set to its frequency in the background. A word-based score is associated with each (potential) TSS as the sum of score of all overlapping 5-mers in the upstream (downstream) region using the upstream (downstream) dictionary.

The promoter regions proximal to a CpG island have GC-rich words with high scores, as expected. To be able to identify promoter regions not related to CpG islands we need to train separately for GC-poor TSSs. To do this, we extracted from the known TSSs in the training set, the TSSs which were not contained inside a CpG island. These were used as described in the previous paragraph to compute a GC-poor word dictionary.

A transcription factor binding site distribution was built in the $\{-600:+600\}$ region around the TSS as described below. We only considered the well known 7 core factors for this purpose, namely, Sp1, ATF, NF-kappaB, Oct-1, TATA, Inr, CCAAT (Lewin 1997). The corresponding positional weight matrices (PWMs), M00008, M00017, M00054, M00138, M00252, M00253, M00254, were extracted from TRANSFAC (Heinemeyer et al. 1998) and searched against the 1200 bps sequence regions using Patser, which is part of the Consensus suite of programs (Hertz and Stormo 1999). The $\{-600:+600\}$ region was divided into 24 non-overlapping 50 bps windows, or bins. For a particular PWM and particular bin, a score was computed as the log of the ratio of the frequency of the PWM hits in the bin for all training sequences to its frequency in the background region, in a similar fashion as for the 5-mer dictionary. A TFdistribution-based score is associated with each (potential) TSS as the sum of score of all PWM hits in the $\{-600:+600\}$ region around it.

As mentioned before, we used AnnSpec (Workman and Stormo 2000) to detect novel signals in the promoter region. AnnSpec uses a combination of neural networks and Gibbs sampling to amplify signals as compared to background sequences. Identified signals can be described as PWMs and treated as TRANSFAC matrices and subjected to the same processing as described in the previous paragraph. A similar score, TFdistributionANNSPEC was derived using the ANNSPEC derived matrices.

A TF site cluster based parameter was associated with each (potential) TSS. This is defined as the number of maximum hits in any 100 bps window in the 600 bps region upstream of the TSS. Only the core factors were considered for this purpose.

To summarize, with each (potential) TSS we associated 8 different parameters, namely, length, value and distance of the closest CpG island as described above, the word-based score, GC-poor word based score, TFdistribution-

based score, TFdistributionANNSPEC-based score, and the TFSite cluster-based score. Each known TSS (positive) and the set of 2000 random positions on the genome corresponding to the training set, were treated like points in a 8-dimensional space and linear discriminant analysis (LDA) (Afifi and Azen 1979) was used to find the most discriminating hyperplane in this space. For another application of this technique to a computational biology problem, see (Solovyev et al. 1994). The score of a data point is $w \cdot p$ where w is the vector of weights for the parameters and p is the vector representing the data point. The vector w that maximizes the ratio of inter-class variation of score to intra-class variation of score is computed as

$$w = S^{-1} * (\mu_1 - \mu_2)$$

where S is the pooled covariance matrix of the parameters, μ_1 and μ_2 are the sample mean vectors of parameters for the positive and negative data respectively. After we compute the optimum weight vector w and apply it to the data points in the test set, we vary the score threshold to study the sensitivity and specificity tradeoffs.

Application

To apply the method, we scan the genomic sequence, sampling at 10 bps intervals, and at each position, compute the 8 parameters and the TSS score as the dot product of the vectors of parameters and the vector of coefficients obtained from training. We retained only local maxima and among those, if two maxima were closer than 2 Kb apart, we only retained the higher scoring maxima. A 1200 bps region centered around these retained maxima were reported as the promoter prediction. We consider a predicted promoter correct, if it overlaps the 200 bps upstream and 100 bps downstream region around a TSS as per Fickett and Hatzigeorgiou, 1997.

In order to estimate the relative contribution of the parameters, we repeated the training and testing using only the CpG related parameters, that is, the length, the value and the distance of the closest CpG island from the potential TSS. When we use only these CpG related parameters, clearly all the maxima will be within CpG islands. So, instead of scanning the genomic sequence as before, we scored each CpG using distance as 0 and applied the length and value coefficients from the training. The 1200 bps region centered at the middle position of the CpG island was reported as a promoter prediction.

Evaluation

Traditionally, sensitivity (S_n) of a predictive method, is measured in terms of fraction of the positive data points correctly predicted by the method and specificity (S_p) of the method is measured in terms of fraction of all the predictions made by the method that are

correct. However, this measure of specificity is only appropriate when we have a definitive set of true and false data. In the context of promoter prediction in the genome, a more reasonable measure of specificity is the frequency of predictions made, expressed in bps. Clearly the higher the frequency of predictions, the higher the sensitivity achieved. The average distance between consecutive predictions is the ratio of the length of the genomic sequence processed and total number of predictions made. We define *specificity* (Sp) as the average distance between consecutive predictions. We present the accuracy of the method as a plot showing the relation between the specificity of the predictions made, to the sensitivity with respect to the known TSS. Since, PromoterInspector is clearly the most accurate promoter predictor currently available, we compare our results only to that of PromoterInspector. The comparison was not done on BenchMark1 due to proprietary restrictions.

RESULTS

Figure 1 shows the distribution of the 7 core factors around the 227 TSSs from SET1. The y-axis represents the log-likelihood as described in the previous section. Each bar corresponds to a 50 bps bin centered around TSS. Clearly Sp1 has high frequency of occurrence in the promoter region and surprisingly Oct-1 is extremely under-represented in this region. The sites that show the most positional preference around TSS are TATA, CCATT and ATF.

In the following we report the accuracy of our predictions on various Benchmarks.

BenchMark1

Out of 150 TSSs in the training set, 90 have an associated CpG island and out of 77 TSSs in the test set 47 have an associated CpG island. There were 2975 promoter predictions made and there were a total of 1908 CpG islands in 12.8 Mb of test sequences at a frequency of 6700 bps per CpG. Figure 2. shows the result of applying the method on BenchMark1. For instance, to achieve a sensitivity of 50% we need to make a prediction every 22 Kbps, as indicated by the box on the plot. The solid plot represents the accuracy using only CpG related parameters for training and testing.

BenchMark2

Out of 227 TSSs in the training set 137 have an associated CpG island and out of 238 TSSs in the test set (Chromosome 22), 162 have an associated CpG island. There were a total of 7950 CpG islands predicted in 35 Mb of sequence at a frequency of 4402 bps per CpG. The higher frequency of CpG islands compared to the BenchMark1 is consistent with GC-richness and

gene-richness of Chromosome 22, compared to the entire genome. There were a total of 8934 promoter predictions made on chromosome 22. Figure 3. shows the result on BenchMark2. We extracted the promoter predictions made by PromoterInspector from Genomatix (http://genomatix.gsf.de/chr22/P_predictions.html). They make 421 predictions, corresponding to a frequency of 82000 bps per prediction and get 92/238 correct, corresponding to a sensitivity of 39%. This coordinate is indicated by a box in the plot. To make an absolute comparison, when we make 421 predictions (by choosing the top 421 scoring predictions), we make 114 (48%) correct predictions. We notice that the PromoterInspector predictions on chromosome 22 have an average length of 592 bps, where our predictions are 1200 bps long. So even though we make the same number of predictions, we

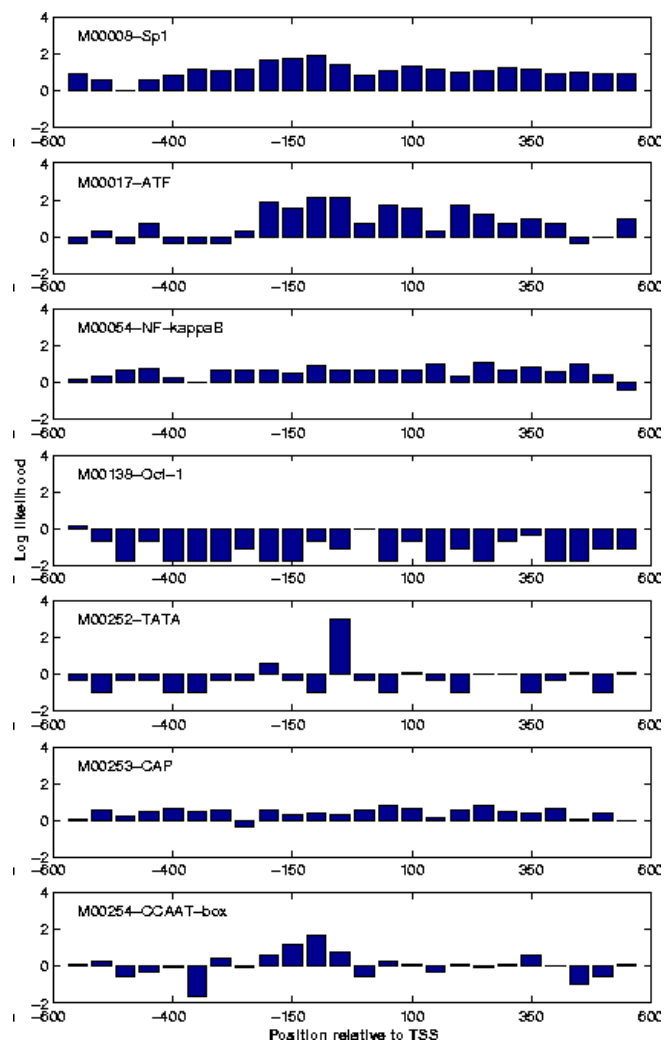


Fig. 1. Frequency of TF sites around TSS, expressed as a log likelihood.

are in fact making many more predictions in terms of base pairs. To make a direct comparison, if we only output 592 bps long region centered around the predicted TSSs as the predicted promoter, we make 102 (43%) correct predictions.

Benchmark3

Out of 35 TSSs in SET3, 26 had an associated CpG island. There were a total of 362 predictions made and a total

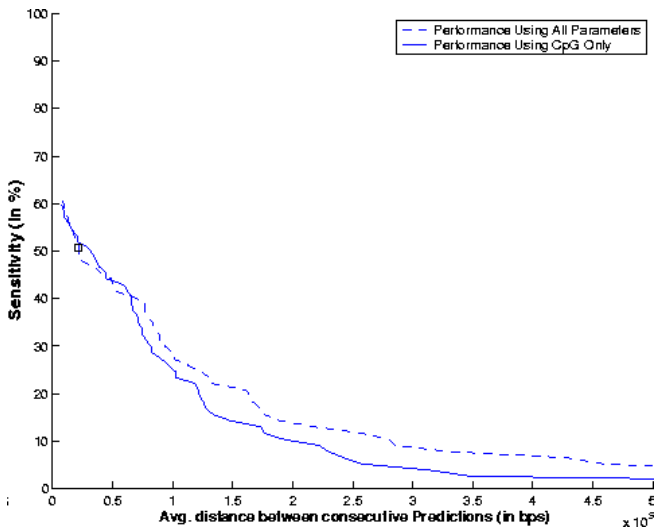


Fig. 2. Prediction accuracy on Benchmark1. To achieve a sensitivity of 50% we make a prediction every 22Kbp, indicated by the box.

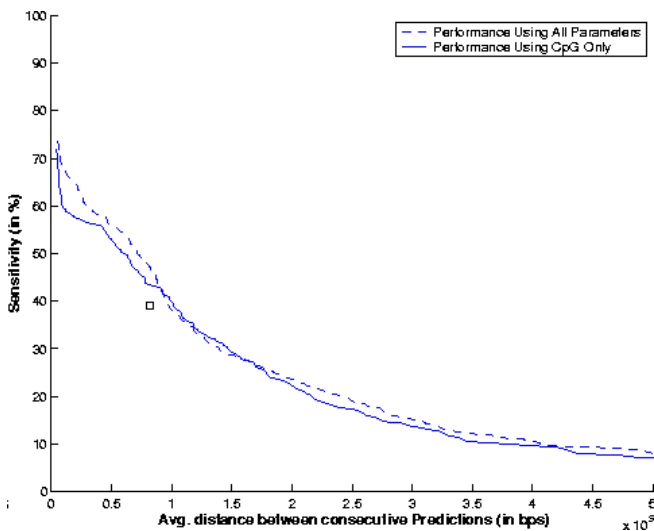


Fig. 3. Prediction accuracy on Benchmark2. The box represents the performance of PromoterInspector, corresponding to 39% sensitivity and a prediction every 82Kbp on average.

Table 1. Summary of the result at single data point for each Benchmark. The result is represented by the sensitivity at a certain frequency of prediction. 1. Using 1200 bps long predictions. 2. Using 592 bps long predictions. 3. The average length of PromoterInspector prediction was not available.

	Promoter-Inspector	CpG+
Benchmark1		50% @ 22Kb
Benchmark2		48% @ 82Kb ¹
	39% @ 82Kb	43% @ 82Kb ²
Benchmark3	43% @ 39Kb	47% @ 39Kb ³

of 226 CpG islands predicted in 1.38 Mb of sequence at a frequency of 6 Kbps per CpG. Figure 4. shows the result on Benchmark3. PromoterInspector (Scherf et al. 2000) makes a total of 35 predictions on these sequences at the rate of one prediction every 39Kb on an average, where 15 correct predictions are made, corresponding to a sensitivity of 43%. This coordinate is indicated by a box in the plot. To make a direct comparison, when we make 35 predictions (by choosing the top 35 scoring predictions), we make 17 (49%) correct predictions. Again, it's likely that the average PromoterInspector prediction is smaller than ours and that would explain the slightly better performance we get. We didn't have access to this number to make a better comparison. Table 1. summarizes the results from the three plots at single points, for ease of comparison. The method developed here is referred to as CpG+.

In all the above benchmarks, the parameters other than CpG related ones hardly make any additional contribu-

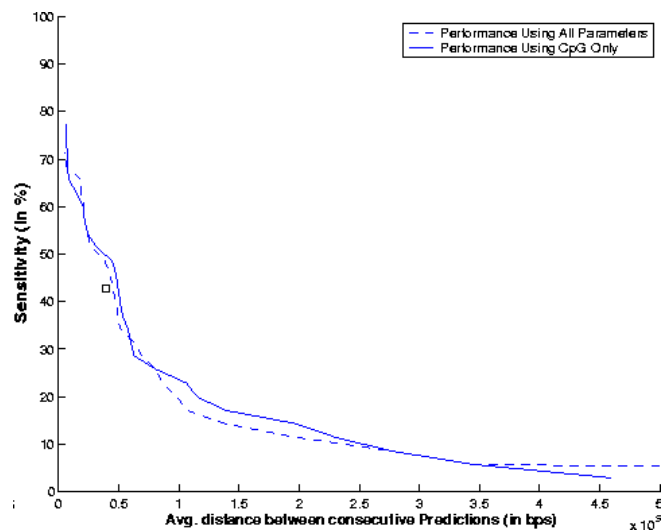


Fig. 4. Prediction accuracy on Benchmark3. The box represents the performance of PromoterInspector, corresponding to 43% sensitivity and a prediction every 39Kbp on average.

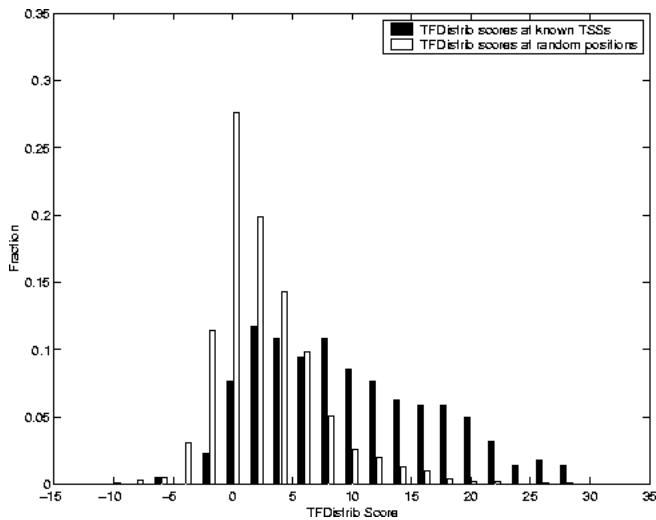


Fig. 5. TFdistribution score distribution for positive and random data points.

tion to the predictive accuracy. However, looking at the distribution of the values of these parameters in the positive and negative sets, there is clearly a difference. See Figure 5 for the distributions of the TF distribution based scores around the known TSSs and the random positions in SET1. There are two possible reasons that these parameters do not make a contribution despite a discriminating distribution between the positive and random positions. Firstly, the distribution is based on a small set (only 2000 random points), which does not translate into acceptable sensitivity/specificity for the genome scale prediction. Secondly, a lot of pairs of these parameters do correlate quite well, by virtue of capturing the same sequence property, and consequently, do not make any additional contribution. For example, word based score, TFdistribution-based score and TFdistributionANNSPEC-based correlate with each other quite well (data not shown).

DISCUSSION

We have explored in this study the possibility of enhancing *de novo* promoter prediction by combining multiple, biologically motivated and seemingly independent features. The RNA polymerase II core promoters can be classified into two main classes, one associated with CpG islands, or in other words, GC-rich, and the other class which is not. The first class is known to correspond to so called house-keeping genes. It is the second class of promoters, typically corresponding to the genes with tissue specific expression, that are the bottleneck in accurate promoter prediction. Even the combination of all the features, trained solely on the second class of promoters was not able to make significant predictions for that class in the test case

(data not shown).

CpG-islands (and GC-richness) of the promoters seems to be the most dominating feature that can be detected at the sequence level around the promoter regions. What we want to emphasize in this paper is not our ability to make accurate prediction for the CpG-associated genes but our inability to make predictions for the non-CpG-associated genes despite using multiple, relevant parameters. This raises some very interesting questions and possibilities. It is likely that chromatin structure plays a critical role in the transcript-ability of genes, by making most of the genome inaccessible. The majority of regulatory DNA in vertebrate genomes is found within islands of highly remodeled chromatin termed “DNase I hypersensitive sites” (Gross and Garrard 1988). Clearly, the role of DNA structure in transcription was never questioned, but the fact that all the attempts on promoter prediction have met with modest success, including ours, suggest the criticality of it. Although computational prediction of DNA structure has not been very successful, it is possible to experimentally determine the accessible regions of chromatin. The accessible regions corresponding to the upstream region of a gene is where one needs to search for clues. It is likely that the genes in the second class, which tend to have tissue specific expression, have their “basal” transcription controlled by non-core TFs binding far from the TSS. Therefore it may not be possible to computationally predict the second class of promoters based solely on the region immediately surrounding the promoters. Correlating the expression level and tissue specificity to the types and strengths of signals found around TSS will be an important step in the direction of understanding transcription.

ACKNOWLEDGEMENTS

Authors would like to thank Natalia Milshina for providing a set of known TSSs curated from the literature. Also authors would like to thank Gene Myers and Vineet Bafna for encouragement and critical comments. We would also like to thank the reviewers for their insightful comments.

REFERENCES

- Affi, A.A. and S.P. Azen. (1979). *Statistical Analysis: computer oriented approach*. Academic Press, New York.
- Bird, A.P., M.H. Taggart, R.D. Nicholls, and D.R. Higgs. (1987). Non-methylated CpG-rich islands at the human alpha-globin locus: implications for evolution of the alpha-globin pseudogene. *Embo J.*, **6**, 999-1004.
- Bucher, P (1990). Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J Mol Biol.*, **212**, 563-78.
- Bucher, P. and E.N. Trifonov (1986). Compilation and analysis of eukaryotic POL II promoter sequences. *Nucleic Acids Res.*, **14**, 10009-26.

- Cross, S.H. and A.P. Bird (1995). CpG islands and genes. *Curr Opin Genet Dev.*, **5**,309-14.
- Fickett, J.W. and A.G. Hatzigeorgiou (1997). Eukaryotic promoter recognition. *Genome Res.*, **7**, 861-78.
- Gardiner-Garden, M. and M. Frommer (1987). CpG islands in vertebrate genomes. *J Mol Biol.*, **196**, 261-82.
- Gross, D.S. and W.T. Garrard (1988). Nuclease hypersensitive sites in chromatin. *Annu Rev Biochem.*, **57**, 159-97.
- Heinemeyer, T., E. Wingender, I. Reuter, H. Hermjakob, A.E. Kel, O.V. Kel, E.V. Ignatieva, E.A. Ananko, O.A. Podkolodnaya, F.A. Kolpakov, N.L. Podkolodny, and N.A. Kolchanov (1998). Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. *Nucleic Acids Res.*, **26**, 362-7.
- Hertz, G.Z. and G.D. Stormo (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563-77.
- Hutchinson, G.B (1996). The prediction of vertebrate promoter regions using differential hexamer frequency analysis. *Comput Appl Biosci.*, **12**, 391-8.
- Ioshikhes, I., E.N. Trifonov, and M.Q. Zhang (1999). Periodical distribution of transcription factor sites in promoter regions and connection with chromatin structure. *Proc Natl Acad Sci USA*, **96**, 2891-5.
- Ioshikhes, I.P. and M.Q. Zhang (2000). Large-scale human promoter mapping using CpG islands [In Process Citation]. *Nat Genet.*, **26**,
- Kondrakhin, Y.V., A.E. Kel, N.A. Kolchanov, A.G. Romashchenko, and L. Milanese (1995). Eukaryotic promoter recognition by binding sites for transcription factors. *Comput Appl Biosci.*, **11**, 477-88.
- Larsen, F., G. Gundersen, R. Lopez, and H. Prydz (1992). CpG islands as gene markers in the human genome. *Genomics*, **13**, 1095-107.
- Lewin, B (1997). *Genes*, Oxford University Press, Oxford.
- Pedersen, A.G., P. Baldi, Y. Chauvin, and S. Brunak. (1998). DNA structure in human RNA polymerase II promoters. *J Mol Biol.*, **281**, 663-73.
- Pedersen, A.G., P. Baldi, Y. Chauvin, and S. Brunak (1999). The biology of eukaryotic promoter prediction—a review. *Comput Chem.*, **23**, 191-207.
- Pickert, L., I. Reuter, F. Klawonn, and E. Wingender (1998). Transcription regulatory region analysis using signal detection and fuzzy clustering. *Bioinformatics*, **14**, 244-51.
- Prestridge, D.S. (1995). Predicting Pol II promoter sequences using transcription factor binding sites. *J Mol Biol.*, **249**, 923-32.
- Scherf, M., A. Klingenhoff, and T. Werner (2000). Highly Specific Localization of Promoter Regions in Large Genomic Sequences by PromoterInspector: A Novel Context Analysis Approach. *J Mol Biol.* **297**, 599-606.
- Solovyev, V. and A. Salamov (1997). The Gene-Finder computer tools for analysis of human and model organisms genome sequences. *ISMB* **5** 294-302.
- Solovyev, V.V., A.A. Salamov, and C.B. Lawrence (1994). Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Res.*, **22**, 5156-63.
- States, D.J., T.W. Blackwell, A.W. Lee, and V. Nowotny (2000). Transcription Factor Binding Site Clusters in the Human Genome are Evolutionarily Conserved and Hypomethylated. In S. Salzberg and A.R. Kerlavage,(eds), *4th Annual Conf. on Comp. Genomics*, pp. 11-11, TIGR Genomic Science Series, Baltimore.
- Tsunoda, T. and T. Takagi (1999). Estimating transcription factor bindability on DNA. *Bioinformatics*, **15**,
- Venter, J.C. et al. (2001). The Sequence of the Human Genome. *Science*, **291**, 1304-1351.
- Workman, C.T. and G.D. Stormo (2000). ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac Symp Biocomput.*, 467-78.
- Zhang, M.Q. (1998). Identification of human gene core promoters in silico. *Genome Res.*, **8**, 319-26.