CABIOS

Vol. 12 no. 5 1996
Pages 391–398

The prediction of vertebrate promoter regions using differential hexamer frequency analysis

G.B.Hutchinson[1]

Abstract

Motivation: *To develop an algorithm utilizing differential hexamer frequency analysis to discriminate promoter from non-promoter regions in vertebrate DNA sequence, without relying upon an extensive database of known transcriptional elements.*

Results: *By determining hexamer frequencies derived from known promoter regions, coding regions and non-coding regions in vertebrates' DNA sequence, and a formula first applied by Claverie and Bougueleret (1986), a discriminant measure was created that compares promoter regions with coding (D1) and non-coding (D2) sequence. The algorithm is able to identify correctly the promoter regions in 18 of 29 loci (62.1%) from an independent test data set. With program options set to identify only one promoter region in the forward strand, there are 11 false-positive predictions in 208 714 nucleotides (one false positive in 18 974 single-stranded bp). With options set to analyze sequence in discrete segments, there is no appreciable improvement in sensitivity, whereas the specificity falls off predictably. It is of particular interest than a search for a peak score (independent of an absolute threshold) is more accurate that a search based upon a fixed scoring threshold. This suggests that the selection of promoter sites may be influenced by the global properties of an entire sequence domain, rather than exclusively upon local characteristics.*

Availability: *A binary-executable, MS-DOS version of PromFind is available free of charge by anonymous ftp, address: iubio.bio.indiana.edu, directory: molbio/ibmpc.*

Contact: *E-mail: hutch@netshop.bc.ca*

Introduction

The Human Genome Project is now entering the Megabase Sequence Era, with several viral and prokaryotic genomes already completely sequenced and very long stretches of contiguous DNA sequence appearing in databases (Anon., 1995). It is desirable that genes and their regulatory regions be identified through computational means prior to detailed experimentation. An eventual goal is the automated annotation of a DNA sequence, to include a complete description of coding regions, promoters, repetitive sequence and other features (Hutchinson, 1995). In addition to the recognition of promoter regions, the identification of specific transcriptional elements and their context might assist in the functional classification of a gene, with a predicted pattern of expression.

Eukaryotic polymerase II promoter regions are characterized by short sequences, termed transcriptional elements (TE), spread out over a region upstream of the transcriptional start site of genes (Lewin, 1990). The computational identification of promoter regions is rendered difficult by the short, degenerate nature of these elements, such that even relatively sophisticated matrix methods lead to the false prediction of many sites. For example, Prestridge and Burks (1993) found, on average, a TATA box predicted for every 120 bp in a non-promoter data set. Evidence suggests that the functionality of promoter regions derives from the interaction of transcription factors that bind to the TEs and subsequently to each other and the polymerase complex (Latchman, 1995). The presence of individual TEs may be a necessary, but not sufficient condition for the promotion of efficient transcription. For transcriptional elements, context is at least as important as the degree of match to a consensus.

Algorithms on promoter recognition published to date rely upon the initial recognition of transcriptional elements through consensus, matrix methods or artificial intelligence techniques that rely upon an extensive database of known transcriptional elements (Bucher, 1990; Claverie *et al.*, 1990; Ghosh, 1990, 1991; Prestridge, 1991; Frech *et al.*, 1993; Prestridge and Stormo, 1993; Fickett, 1996; Wingender *et al.*, 1996). Regions of DNA containing a higher density of putative TEs relative to non-promoter sequences are more likely to be true promoter regions (Prestridge and Burks, 1993). This technique could be considered a complex adaptation of 'search by signal' methodology for identifying protein-coding regions (Staden, 1990; Stormo, 1990).

The current study began with the hypothesis that 'search by content', another methodology used in the prediction of protein-coding regions, might also prove useful in the identification of promoter regions. Of the 20 content measures identified and reviewed by Fickett and Tung (1992), hexamer measures were found to be among the most useful. Claverie and Bougueleret (1986) first described a

Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada

[1]Correspondence address: c/o RabbitHutch Biotechnology Corporation, PO Box 506, 108 Mile Ranch, British Columbia, V0K 2Z0, Canada

method of expressing the discriminant index of a word of length k in two different sequence subsets. In the current study, the discriminant index $k = 6$, here termed the differential hexamer measure, was applied to sequence from promoter regions in order to discriminate them from both coding and non-coding transcribed sequence.

System and methods

The algorithm was implemented on a 90 MHz Pentium microcomputer with 16 Mbytes of random access memory running the Windows 95 operating system. The executable program, PromFind, is designed to run in a 16-bit MS-DOS environment on any microcomputer with a 386 or higher microprocessor. The program is written in C++ and compiled using the Borland C++ compiler, Version 4.51, and is available as PROFINnn.EXE (where nn represents the version number) via ftp from iubio.bio.indiana.edu/molbio/ibmpc, ftp.bchs.uh.edu/gene-server/dos or from cgat.bch.u-montreal.ca. The author may be contacted by E-mail for further information.

Algorithm

The Eukaryotic Promoter Database (EPD), Release 43 (Bucher, 1995) contains 1250 entries for promoter regions recognized by eukaryotic RNA polymerase II. To be included in the database, an entry must be experimentally defined and active in a higher eukaryote, biologically functional and distinct from other promoters in the database. Of these, there are 659 entries for vertebrate promoters for chromosomal genes, and 429 are independent (non-homologous with other promoters in the database). These independent entries were selected as the starting point for the creation of the experimental training and testing data set in order to avoid bias by multiples of closely related sequences. The purpose of the data set was to match known promoter regions derived from the EPD with the GenBank or EMBL entry for the transcribed portion of the gene (associated locus), so that a count could be made of hexamers in three distinct regions of each gene: the promoter, known coding sequence and known non-coding sequence. It was possible to match the promoter region in the EPD database with EMBL or GenBank database entries for 197 genes from primate (77), rodent (68), mammal (14), bird (29) and frog (9) sources. Other sequences were discarded primarily because of insufficient non-coding sequence or ambiguity in the feature table that precluded the clear assignment of coding or non-coding status. Since the 5' flanking regions of genes are often sequenced after the transcribed portions have been entered into the database, it was common to find that the entry for the associated locus did not contain the promoter. These sequences were still used to determine hexamer counts for coding and non-coding sequence downstream of the start of translation, although they could not be used for testing the algorithm. The EPD database is conveniently arranged into sections with genes of similar function. The major functional headings for vertebrate chromosomal genes are small nuclear RNAs, structural proteins, storage and transport proteins, enzymes, hormones, growth factors, regulatory proteins, proteins related to stress or pathogen defense and unclassified genes. Genes within the same functional group are expected to share similar transcription elements, and it is therefore desirable that training and testing data sets contain representatives of each functional group. The original data set of 197 promoters and associated genes, ordered as in the EPD by function, was therefore subdivided by choosing approximately every sixth gene for inclusion in the test set, with the remaining sequences remaining in the training set. This resulted in a training set of 168 promoter/associated locus combinations from which the discriminant measures were derived. In 126 cases, the promoter region was included in the associated locus, and could therefore be used as a subset of the training set suitable for initial testing of the scoring algorithm.

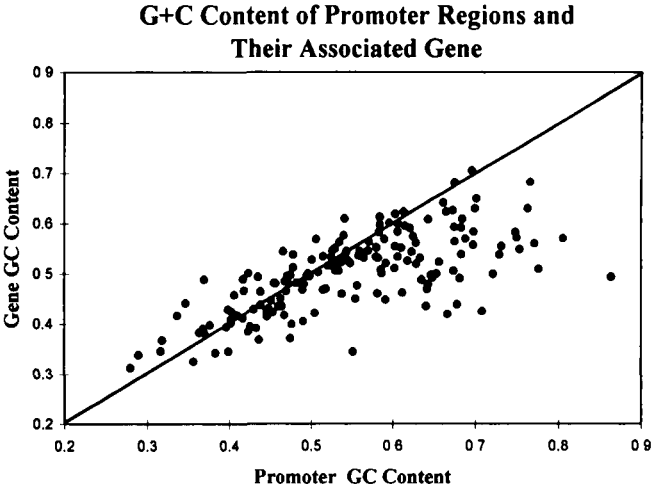

Fig. 1. G + C content for promoter regions of the 126 training loci is plotted against the G + C content of the associated locus. Most promoter regions have a higher G + C content than their associated locus, but a positive correlation is readily apparent (correlation coefficient 0.707).

A comparison of the G + C content of the EPD promoter entry and the associated locus for the 126 gene data set is displayed in Figure 1. It can be seen that, in general, promoter regions have a higher G + C content than their associated locus, but the G + C content of some promoter regions is nevertheless low. The G + C content of the promoter is positively correlated (correlation coefficient 0.707) with that of the gene as a whole.

Promoter sequences in the EPD run from base −499 to +100 with respect to the transcription start site. It has previously been shown that the density of transcription elements increases from position −500, with the greatest increase seen from about −200 to the transcription start site

(Prestridge and Burks, 1993). The sampling algorithm was therefore set to scan the region from −299 to 0 in each entry, counting hexamers as it went. The associated locus was then scanned with reference to the 'CDS' line of the feature table, counting hexamers arising from coding and non-coding transcribed regions downstream of the start of translation. The use of sequence downstream of the coding start avoided the re-sampling of the promoter region. Hexamer counting for coding and non-coding sequence for each associated locus was limited to a maximum of five times the promoter sequence available in order to prevent bias from long sequence entries. The resulting counts for each of the 4096 possible hexamers from each of the three regions were then entered into a spreadsheet program.

The frequency, $F(h)$ of each hexamer h was determined for each of the three regions. The frequency differential, $D_n(h)$ for each hexamer was then calculated according to Claverie and Bougueleret (1986) as follows:

$$D_1(h) = \frac{F_{promoter}(h)}{F_{promoter}(h) + F_{non\text{-}coding}(h)}$$

$$D_2(h) = \frac{F_{promoter}(h)}{F_{promoter}(h) + F_{coding}(h)}$$

$D_1(h)$ thus represents the hexamer frequency differential between promoter and non-coding regions, while $D_2(h)$ represents the differential between promoter and coding sequence.

Scoring algorithm

The D1 and D2 measures were applied to an unknown sequence by averaging the $D_1(h)$ or $D_2(h)$, respectively, for each hexamer in a 300 bp window. It was noted during early attempts at developing a scoring algorithm that the D1 measure provided the best contrast between promoter and non-promoter sequence, except for many false-positive peaks within coding regions. It thus appeared that the D1 measure, having been derived only from promoter and non-coding sequence, misinterpreted coding sequence as a possible promoter region. To help reduce these false-positive identifications, threshold scores for the D1 and D2 measure were empirically determined that filtered out many of the erroneous results. During the early experiments on this two-step approach, in which fixed thresholds for D1 and D2 were used, an unacceptable trade-off between sensitivity and specificity occurred. For example, it was found that a D1 threshold score of 0.47 was necessary to permit the correct identification of the promoter region for the human interleukin-2 (IL-2) gene (GenBank and EMBL Accession X00695) without false positives. However, the same threshold led to the identification of the true promoter plus seven false-positive regions in the mouse LDH-A gene for lactate dehydrogenase A (Accession Y00309) and similar multiple false-positive results in other loci. The scoring algorithm was therefore modified to identify only the single highest-scoring region in a sequence. It scans in steps of 10 bp and identifies the 300 bp window within a sequence or segment that scores highest using the D1 measure and also exceeds an empirically determined threshold D2 value.

Table I. The 20 most frequent hexamers found within the promoter region from base −299 to 0 with respect to the transcriptional start site. Although AAAAAA (and TTTTTT, ranked 22nd) are both common, this is due to long tracts of poly(A) and poly(T) in a few loci. The canonical TATAAA element is ranked ninth, and most of the remaining 18 hexamers show similarity to the Sp1 binding site

Rank	Hexamer	Frequency ×1000	Rank	Hexamer	Frequency ×1000
1	AAAAAA	2.46	11	CCTCCC	1.35
2	GGGGGG	2.05	12	GCGGGG	1.31
3	GGGCGG	2.03	13	CCCCTC	1.26
4	GGCGGG	1.73	14	CTCCCC	1.26
5	CCGCCC	1.51	15	CTCCTC	1.22
6	GGGGCG	1.49	16	GGAGGG	1.22
7	CCCGCC	1.42	17	GGGGGC	1.22
8	CCCCGC	1.37	18	CAGCCC	1.19
9	TATAAA	1.37	19	CCCACC	1.19
10	CCCTCC	1.37	20	GGCGGC	1.17

Implementation

Sampling algorithm and hexamer statistics

Counts were collected for all 4096 hexamers over the three regions in the training set of 168 genes, for total counts of 44 399 promoter, 105 201 coding and 232 644 non-coding hexamers, and hexamer frequencies were calculated. Table I shows the 20 most frequent hexamers encountered in the promoter region and their frequencies. The highest-ranking hexamer, AAAAAA, or its reverse complement, TTTTTT (ranked 22nd in order of frequency) was found to be present in only 23.2% of promoters, but with sufficiently long tracts of poly(A) or poly(T) to account for the high rankings. The ninth ranked hexamer, TATAAA , matches the least variable six nucleotides of the canonical TATA box motif. The other 18 of the 20 most frequent hexamers are notable for being G + C rich, including several hexamers that show similarity to Sp1 binding sites.

After calculating the $D_1(h)$ and $D_2(h)$ measures for each hexamer, the hexamers were ranked in order of decreasing $D_1(h)$. Inspection of the highest-ranking hexamers revealed many containing CpG dinucleotides. CpG islands are often found at the 5' end of genes (Gardiner-Garden and Frommer, 1987). To investigate whether an excess of CpG dinucleotides might account for differences in hexamer frequency between promoter regions and other parts of genes, the 200 highest-ranked hexamers with respect to both frequency and the $D_1(h)$ measure were examined. There are 1208 hexamers that

Table II. The 20 hexamers that rank highest in *D1(h)* score and that do not contain a CpG dinucleotide. Predominant among these hexamers are patterns similar to the TATA or CCAAT binding site consensus

Rank	Hexamer	$D_1(h)$ Score	Rank	Hexamer	$D_1(h)$ Score
1	CCAATC	0.814	11	CCAATG	0.739
2	TATAAA	0.796	12	GATAAG	0.735
3	GCCAAT	0.792	13	GGCTAT	0.734
4	CAATCA	0.763	14	GGGGGG	0.732
5	CTATAA	0.763	15	GCTATA	0.731
6	GGTTCC	0.752	16	TATATA	0.730
7	TATAAG	0.750	17	CATAAA	0.724
8	ATATAA	0.741	18	TTAGTC	0.724
9	GACCAA	0.741	19	TGATTG	0.715
10	GTAGGC	0.739	20	GATTGG	0.713

contain at least one CpG. If CpG-containing hexamers are randomly distributed among the top-ranking hexamers, one would expect ~63 to be in the top 200. Ordered in descending frequency rank, 45 of the top 200 contained at least one CpG (91 contained at least one GpC). Compared to a random distribution, hexamers containing CpG dinucleotides are therefore somewhat under-represented. However, when ordered with respect to descending $D_1(h)$ values, 196 of the highest-ranked 200 hexamers contained a CpG (135 contained a GpC). Non-coding, non-promoter regions are therefore relatively more depleted in CpG dinucleotides when compared with promoter regions. Similar findings were encountered when ranking hexamers with respect to the $D_2(h)$ values (data not shown), showing that coding regions also have a lower frequency of CpG dinucleotides when compared with promoters.

An analysis was then undertaken to determine the highest-ranked hexamers with respect to the D1 measure that do not contain a CpG dinucleotide (Table II). The table does not include any hexamers showing similarity to Sp1 sites, probably because most of these would contain a CpG. However, several of the highest-ranked hexamers match part of the TATA-box or CCAAT-box motifs (Bucher, 1995).

Determination of program accuracy

For each test sequence, the program reports the centre of the 300 bp window within a sequence segment that obtains the highest score. The size of the segment can be defined by the user to be a fixed value or the entire sequence. For the purposes of accuracy determination, the 'true' promoter region is defined as the 300 nucleotides directly upstream of the transcription start site. A prediction is considered to be correct if the centre of the predicted promoter region is within 200 bp of the centre of the true promoter region. This guarantees an overlap of the predicted and true promoter region of at least 100 bp. However, with particularly short test sequences, there will exist some probability that the promoter could be correctly predicted by a random guess. Although it would be ideal to choose only test sequences that exceed a length of (for example) 5000 bp, it was necessary in this study to include many shorter sequences to obtain sufficient data. For this reason, results are given for all sequences as well as for sequences that exceed 5000 and 10 000 bp. The probability that a promoter region could be predicted randomly within a sequence greater than 5000 and 10 000 bp in length is less than 0.08 and 0.04, respectively. The program operates in two modes. In the first mode, the program makes one prediction for the forward strand of each entire sequence, producing either a match to the true promoter, or a single false-positive prediction (i.e. the entire sequence is chosen as the size of a segment). The user may optionally analyze the reverse strand. In the second mode, the program is supplied by the user with an integer describing a window size. One prediction is subsequently made for each segment along the length of the sequence.

Application of the scoring algorithm

Of the 168 associated sequence files in the training set, 126 contained sequence that included the promoter. The scoring algorithm was first applied to this training subset. The entire list of tested loci is not included here due to space considerations, but is available from the author on request. The average sequence length included in each locus was 4546 bp, with a minimum of 739 bp and a maximum of 25 759 bp. Application of the algorithm to all 126 testable loci of the training set (operating in mode 1 and making only one prediction for each sequence) resulted in the correct identification of 88 of the 126 promoters (69.8%), with 38 false-positive predictions over the 572 827 nucleotides of the data set (see Table IIIa). The G+C content of the genes tested ranged from 0.31 to 0.71, with a median G + C content of 0.51. Of those genes with a G+C content less than the median, 42/63 promoters were correctly predicted (66.7%), while the promoters of genes with a G + C content higher than the median were predicted correctly in 46/63 (73.0%), suggesting only a minor effect of G + C content on prediction ability. Of the 88 promoters correctly predicted, the predicted centre of the promoter region differed from the actual centre by an average of 76 bp. In 12 predictions considered to be false positive by the above criteria, there was still some overlap of the predicted promoter region with the actual promoter.

Although operating the program as above produces a low false-positive rate of one false prediction per 15 074 single-stranded bp, this is not directly comparable to statistics quoted by other programs since the program (operating in the first mode) only makes one prediction per sequence. Other measures of sensitivity and specificity may be obtained by restricting the analysis to two subsets: sequences of length

Table III. Performance of PromFind on various training and test data sets. The 'All sequences' line refers to testing the program on entire sequences, with only one prediction per sequence. This has the effect of making the false-positive rate appear very low. Extending this one prediction per sequence methodology becomes problematical when looking at large sequences that are likely to contain more than one promoter. In the 'Length > 10 000' line, the program was tested on segments of 10 000 bp only on genes exceeding 10 000 bp in length, whereas the 'Length > 5000' line shows results for testing all gene entries >5000 bp in length, in segments of 5000 bp. Analyzing sequences in segments of fixed length rather than as entire loci minimally improves sensitivity while greatly increasing the false-positive rate

Data-set	Number of sequences	True positive	Sensitivity	False positive	False-positive rate per ss bp	Maximum probability
(a) Training set						
All sequences	126	88	0.698	38	1:15074	0.541
Length >10 000	12	7	0.583	19	1:9111	0.040
Length >5000	36	24	0.667	66	1:5223	0.076
(b) Test set						
All sequences	29	18	0.621	11	1:18974	0.516
Length >10 000	7	5	0.714	11	1:9636	0.039
Length >5000	17	10	0.588	36	1:4781	0.078

>5000 bp (making one prediction for each 5000 bp) and sequences of length >10 000 (with one prediction for each 10 000 bp). The results of these analyses are also shown in Table IIIa. There were 36 loci in the training set that exceeded 5000 bp in length. In 24 of the 36 (66.7%), the promoter was correctly predicted, with one false prediction per 5223 bp in the forward strand. Of the 12 loci in the data set with a sequence length of >10 000 bp, the promoter was correctly predicted in seven (58.3%) with one false prediction per 9111 bp in the forward strand.

Table IV. Performance of the scoring algorithm on 29 loci from the independent test set. Loci are presented according to their order in the Eukaryotic Promoter Database. The actual promoter centre is defined as a point 150 bp upstream of the transcription start site. Predictions are considered correct if the predicted and actual promoter centres differ by <200 bp. The algorithm correctly predicted 18 of the 29 promoters, with 11 false-positive predictions in the forward strand

Locus	Accession	Sequence length	G + C content	*Dw*1 score	Promoter centre		Correct Prediction?
					Predicted	Actual	
MMNUCLEO	X07699	11478	0.446	0.593	1650	1731	Y
MMRPL3A	K02060	3757	0.513	0.577	470	208	–
HSTUBAG	X01703	4087	0.479	0.568	280	173	Y
GGACTAC	X02212	5463	0.450	0.607	400	150	–
GGMYHE	J02714	31 111	0.419	0.474	12460	2062	–
HSEPKER	J00124	5339	0.544	0.545	400	108	–
HSNFLG	X05608	4542	0.473	0.588	200	11	Y
GGCRYD1	X02222	8160	0.482	0.546	220	72	Y
GGCALB	Y00407	11 068	0.495	0.516	190	117	Y
RNLALB01	X00461	3829	0.466	0.504	1580	1098	–
GGFERH	M16343	7126	0.501	0.635	1210	1125	Y
OCHBAPT	M11113	4031	0.663	0.663	3130	3	–
GGHBBR2	K00824	1955	0.561	0.543	150	49	Y
HSTKRA	M15205	13 500	0.533	0.596	330	309	Y
MUSODCC	J03733	7100	0.482	0.632	510	385	Y
RNHOX	J02722	8377	0.506	0.557	1400	1238	Y
RNPTRY1	J00778	6503	0.405	0.478	3020	3032	Y
MMGUSB01	J02836	16 449	0.488	0.527	2270	2151	Y
HSMHCP42	M12792	5141	0.589	0.539	1800	1520	–
BTHOR01	X00502	1167	0.705	0.617	500	60	–
SSFSHBS	D00621	10 172	0.384	0.493	9420	5514	–
HSIFNG	V00536	5961	0.366	0.463	2730	197	–
HSTNFB	X02911	3037	0.555	0.546	830	668	Y
HSOPS	K02281	6953	0.549	0.511	1150	51	–
HSNRASR	X02751	2436	0.419	0.519	550	463	Y
HSA1ATP	K02212	12 222	0.514	0.519	1940	1807	Y
HSTCBV81	X07192	775	0.471	0.452	170	317	Y
HSTCRT3D	X03934	4186	0.509	0.504	350	193	Y
MMNPGF1	M58691	2789	0.541	0.554	940	846	Y

The algorithm was then applied to the 29 independent loci of the test set (see Tables IIIb and IV). The average sequence length included in each locus was 7197 bp, with a minimum of 775 bp and a maximum of 31 111 bp. The promoter was correctly identified in 18 (62.1%) of the loci, with 11 false-positive predictions for the total of 208 714 nucleotides. In four loci considered false positive, there was still some overlap with the actual promoter. Of the 18 promoters correctly predicted, the predicted centre of the promoter region differed from the actual centre by an average of 111 bp. To test whether higher scoring predictions correlated with accuracy, the test loci were ranked according to the D1 score of the predicted promoter. Eight of the 14 lowest-scoring predictions (57.1%) were correct compared with 10 of the remaining high-scoring 15 (66.7%). The lack of a strong correlation between D1 score and accuracy suggests that the confidence of predictions cannot be determined reliably using the D1 score.

Those test set loci exceeding 5000 and 10 000 bp were analyzed both as entire sequences (mode 1, see Table IV) and in segments (mode 2, Table IIIb). Of the seven loci longer than 10 000 bp in length, the promoter was correctly predicted in five (71.4%) in either mode. Analyzing the loci in segments of 10 000 bp increased the false-positive predictions from 2 to 11. There were 17 loci of length >5000. Of these, 10 promoters were predicted correctly (58.8%) in either mode. A segmental analysis in windows of 5000 bp increased the false-positive predictions from 7 to 36.

To determine the potential increase in false-positive determinations that would be incurred by the analysis of the opposite strand, the algorithm was applied (in mode 1, i.e. one prediction per sequence) to the complementary sequence of the test set members. Of the 29 predicted promoters in the reverse strands, 14 closely overlapped predictions for the forward strand, with a maximum distance between the forward and reverse predicted centres of 38 bp (average 18). Eight of these predicted promoters corresponded to the true promoter, but were made for the opposite strand. The other 15 predictions were all additional false positives.

Discussion

This study demonstrates that a two-step approach using differential hexamer measures derived by contrasting promoter regions with coding and non-coding sequence is capable of identifying the promoter region in 58–71% of vertebrate DNA sequences known to contain a single promoter. The inclusion of several short sequences in both the training and testing sets could potentially bias the results since, in some cases, even a random choice would have a reasonable probability of being found 'correct'. It is, therefore, of importance to note that the sensitivity does not appreciably fall off with sequences of greater length. This accuracy is surprising given the simplicity of the algorithm. Other programs developed for promoter recognition to date have relied upon *a priori* knowledge of binding sites maintained in a transcription factor database. For example, PROMOTER SCAN (Prestridge, 1995) scans a sequence using a promoter recognition profile, scoring each transcriptional element according to its relative density in promoter versus non-promoter sequence. Additional points are given for sequences scoring beyond an empirically derived threshold matrix score for the TATA box. In contrast, the algorithm derived in this study uses no prior knowledge of transcriptional elements and does not specifically recognize the TATA box or any other motif. Owing to the simplicity of the calculations, the program is able to analyze both strands of the sequence from the 126 testable training loci in 83 s (6900 bp/s) using an MS-DOS computer with a 90 MHz Pentium processor.

The D1 measure appears to be measuring two characteristics that differ between promoter and non-coding sequence: the CpG dinucleotide content and the presence of transcriptional elements. Most highly ranked hexamers with respect to $D_1(h)$ either contain CpG dinucleotides or show similarities to known transcriptional elements such as TATA, Sp1 or CCAAT boxes. It is possible that those high-scoring hexamers that do not belong to either of the above groups will be found to correspond to other, perhaps less recognizable transcriptional elements. As such, the technique shows promise as a means of discovering new binding sites for transcription factors. In analyzing the reverse strand, potential promoters are predicted that overlap the true promoter on the forward strand. In this case, the algorithm may be responding to the mirrored high G + C content of the complementary sequence, and in addition is detecting transcriptional elements (such as Sp1 or CCAAT boxes) that do not show an orientation with respect to the transcribed strand.

Perhaps the most significant finding of this study is the influence of global sequence properties on the correct identification of promoter regions. Early versions of the algorithm set a fixed threshold for the D1 measure, but this resulted in an unacceptable number of false-positive predictions. The discovery that the peak score alone reliably predicted the promoter was unexpected. Moreover, the absolute score did not correlate significantly with accuracy. This finding was initially thought to be biased by the inclusion of several short sequences in the data set, but even the promoter regions of sequences of length greater than 10 kilobases were predicted in over 58% of cases using this method. It is tempting to speculate that this finding speaks to the biology of promoter recognition itself. Perhaps the promoter region is recognized simply as the sequence within a domain that 'looks most like' a promoter, without a set requirement for a minimum number of transcriptional elements or adherence to other criteria.

Operation of the program to make only one prediction per strand per sequence can be thought of as identical to setting a fixed search segment length that exceeds the maximum length of all test sequences. The option of specifying a fixed length was introduced to answer concerns that accuracy determinations (particularly false-positive rates) were misleading if only one prediction was made for a sequence. It was expected that a trade-off would occur, since restricting the search to smaller segments should improve sensitivity. This effect proved to be very small. In the training set, the introduction of a search segment length of 5000 bp led to the identification of only four more promoters out of 36, but false-positive predictions increased dramatically from 15 to 66. In the independent test set, no additional correct predictions were made, but false-positive predictions went from 6 to 36. This finding is again compatible with the hypothesis that the promoter is identified, at least in part, as the sequence within a segment that 'looks most like' a promoter, and that the size of a segment is larger than 5000 or even 10 000 bp.

This author is aware of four other programs in current use that attempt to predict eukaryotic polymerase II promoter regions. However, only one study has been published at the time of writing, on PROMOTER SCAN (Prestridge, 1995). To generate some comparative statistics on accuracy, Version 1.7 of PROMOTER SCAN was tested on the 29 test sequences reported in Table IV. Predictions in either the forward or reverse strands were considered correct if the distance between the centres of the predicted and true promoter regions was <200 bp. Forward and reverse strand predictions within 200 bp of one another were considered to be one prediction (i.e. an incorrect prediction duplicated on the reverse strand was counted as only one false positive). PROMOTER SCAN correctly identified 14 of the 29 promoters (48.3%) with 30 false-positive predictions. This compares with the correct identification of 18 promoters with 26 false-positive predictions in a double-strand search with PromFind using the same criteria of accuracy.

Most of the analysis presented in this study was conducted on sequence from a single (forward) strand. In many cases, such as in sequence derived from primer extension, the researcher will know the orientation and may wish to ignore the reverse strand predictions. Many transcriptional elements, such as CCAAT and Sp1 sites, are bi-directional, and the strandedness of the promoter will remain unknown for sequences in which no prior knowledge of orientation exists. The TATA box, however, is uni-directional, and is present in a majority of genes (Prestridge and Burks, 1993). Future development of the program should include the specific identification of TATA-box motifs in order both to determine the strandedness and localize the exact transcriptional start site with more accuracy.

Programs that predict protein-coding regions can improve accuracy by using the putative identification of the transcriptional start site to identify the first coding exon of a gene, although allowance must be made for the possible presence of an intron between the transcriptional and translational start sites. Conversely, a highly confident prediction for a coding exon should prompt the search for an upstream promoter region. Clearly, the future evolution of gene-finding programs will involve the integration of techniques to identify all recognizable features, with the sharing of information between algorithms.

This algorithm was developed using sequences typical of GenBank and EMBL databases at their current level of sophistication. That is, most entries contain at most one gene, and many contain only fragments of genes, with the regulatory sequence entered separately. The program is not designed to determine whether a promoter region exists within an unknown sequence, but rather assumes that there is a promoter and identifies the most probable region containing it. As a default, the program assumes that there is only one promoter and makes one prediction. With the option of specifying a search segment size, the user can examine sequences where multiple promoters might be expected. Within the next few years, an increasing number of large DNA regions will be sequenced without prior knowledge of gene content. It is likely that researchers will rely upon computer programs to annotate the sequence tentatively. PromFind joins related programs by this author, including SorFind (identification of putative protein-coding exons) and RepFind (annotation of common repetitive elements), to form a suite of computer software operating in tandem to assist with this task (Hutchinson and Hayden, 1992).

Acknowledgements

Travel to the International Workshop on Computational Analysis of Eukaryotic Transcriptional Regulatory Elements was assisted by a grant from the Canadian Genome Analysis & Technology Program. Thanks are due to the reviewers and Dr Allen Delaney for their helpful suggestions for revisions.

References

Anonymous (1995) GenBank enters the megabase sequence era. NCBI News, National Center for Biotechnology Information, Sept, pp. 1–2.

Bucher,P. (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.*, **212**, 563–578.

Bucher,P. (1995) *The Eukaryotic Promoter Database EPD*. EMBL Nucleotide Sequence Data Library, European Bioinformatics Institute, Hinxton, Cambridge.

Claverie,J.-M. and Bougueleret,L. (1986) Heuristic informational analysis of sequences. *Nucleic Acids Res.*, **14**, 179–196.

Claverie,J.-M. *et al.* (1990) k-tuple frequency analysis: from intron/exon discrimination to T-cell epitope mapping. *Methods Enzymol.*, **183**, 237–252.

Fickett,J. (1996) Quantitative discrimination of MEF2 sites. *Mol. Cell. Biol.*, **16**, 437–441.

Fickett,J.W. and Tung,C.-S. (1992) Assessment of protein coding measures. *Nucleic Acids Res.*, **20**, 6441–6450.

Frech,K. *et al.* (1993) Computer-assisted prediction, classification, and delimitation of protein binding sites in nucleic acids. *Nucleic Acids Res.*, **21**, 1655–1664.

Gardiner-Garden,M. and Frommer,M. (1987) *J. Mol. Biol.*, **196**, 261–282.
Ghosh,D. (1990) A relational database of transcription factors. *Nucleic Acids Res.*, **18**, 1749–1756.
Ghosh,D. (1991) New developments of a transcription factors database. *Trends Biochem. Sci.*, **16**, 445–447.
Hutchinson,G.B. (1995) *Towards the Automation of Feature Recognition in DNA Sequence.* University of British Columbia, Canada.
Hutchinson,G.B. and Hayden,M.R. (1992) The prediction of exons through an analysis of spliceable open reading frames. *Nucleic Acids Res.*, **20**, 3453–3462.
Latchman,D. (1995) *DNA Sequences and Transcription Factors*, 2nd edn. Academic Press, London.
Lewin,B. (1990) *Genes*, IV edn. Oxford University Press, Oxford.
Prestridge,D.S. (1991) SIGNAL SCAN: a computer program that scans DNA sequences for eukaryotic transcriptional elements. *Comput. Applic. Biosci.*, **7**, 203–206.
Prestridge,D.S. (1995) Predicting pol II promoter sequences using transcription factor binding sites. *J. Mol. Biol.*, **249**, 923–932.
Prestridge,D.S. and Burks,C. (1993) The density of transcriptional elements in promoter and non-promoter sequences. *Hum. Mol. Genet.*, **2**, 1449–1453.
Prestridge,D.S. and Stormo,G.D. (1993) SIGNAL SCAN 3.0: new database and program features. *Comput. Applic. Biosci.*, **9**, 113–115.
Staden,R. (1990) Finding protein coding regions in genomic sequences. *Methods Enzymol.*, **183**, 163–180.
Stormo,G.D. (1990) Consensus patterns in DNA. *Methods Enzymol.*, 211–221.
Wingender,E. *et al.* (1996) TRANSFAC: A database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, **24**, 238–241.