Journal of Biochemistry and Molecular Biology, Vol. 37, No. 6, November 2004, pp. 648-656

JOURNAL OF Biochemistry and Molecular Biology

Composite Dependency-reflecting Model for Core Promoter Recognition in Vertebrate Genomic DNA Sequences

Ki-Bong Kim[†,*] and Seon Hee Park[‡]

[†]Department of Bioinformatics Engineering, Sangmyung University, Cheonan 330-180, Korea

[‡]Electronics and Telecommunications Research Institute, Daejeon 305-350, Korea

Received 10 November 2003, Accepted 2 April 2004

This paper deals with the development of a predictive probabilistic model, a composite dependency-reflecting model (CDRM), which was designed to detect core promoter regions and transcription start sites (TSS) in vertebrate genomic DNA sequences, an issue of some importance for genome annotation. The model actually represents a combination of first-, second-, third- and much higher order or long-range dependencies obtained using the expanded maximal dependency decomposition (EMDD) procedure, which iteratively decomposes data sets into subsets on the basis of dependency degree and patterns inherent in the target promoter region to be modeled. In addition, decomposed subsets are modeled by using a first-order Markov model, allowing the predictive model to reflect dependency between adjacent positions explicitly. In this way, the CDRM allows for potentially complex dependencies between positions in the core promoter region. Such complex dependencies may be closely related to the biological and structural contexts since promoter elements are present in various combinations separated by various distances in the sequence. Thus, CDRM may be appropriate for recognizing core promoter regions and TSSs in vertebrate genomic contig. To demonstrate the effectiveness of our algorithm, we tested it using standardized data and real core promoters, and compared it with some current representative promoter-finding algorithms. The developed algorithm showed better accuracy in terms of specificity and sensitivity than the promoter-finding ones used in performance comparison.

Keywords: CDRM, EMDD procedure, Markov model, Predictive probabilistic model, TSS

*To whom correspondence should be addressed.
Tel: 82-41-550-5377; Fax: 82-41-550-5184
E-mail: kbkim@smu.ac.kr

Introduction

Many other sequencing projects in addition to the Human Genome Project have created a deluge of biological data and ushered us into the functional genomics era. The large and growing amount of genomic DNA sequence data has turned molecular biology into a computationally intensive discipline, and a completely new interdisciplinary field, *bioinformatics* or *computational biology*, has become an integral part of molecular biology research at a genome-wide level. The discipline stems from the storage of large amounts of sequence data in different location, and the cross-linking of this data in an intelligent fashion. However, the application of computers is significantly wider. Since automatic tools are essential for the analysis of large DNA sequences, a plethora of software programs have been designed to solve categories of problems, such as finding protein-coding regions or candidate regulatory elements in genomic DNA. An important part of bioinformatics concerns the interpretation of complete DNA sequences of many organisms. Moreover, without knowing important information, such as, the locations of genes and promoters, the direct impact of a completed genome would be limited to providing the sequence information of a few mapped genes, i.e., genes with known locations. Thus, the automated computer-assisted annotation and analysis of genomes is an important scientific achievement.

A key part of computer - based annotation and analysis concerns regulatory DNA regions - parts of the sequence that influence how and when a gene is activated or expressed. The understanding of the regulation of gene expression is undoubtedly one of the more interesting challenges in molecular biology today (Ko *et al*., 2002). In this context, the problem of identifying promoters in genomic DNA sequences and the significant patterns they harbor using computational methods, has attracted considerable research attention in recent years. One point of view is that the problem is closely related to the fundamental biochemical issues of specifying

the precise sequence determinants of transcription and translation (Hernandez *et al*., 2002). Another is that it may contribute to improvement of gene identification and to prediction of gene expression context. Moreover, finding and decrypting of promoters are interesting in their own right.

The first description of common patterns in eukaryotic promoters, in the form of weight matrices, which are equivalent to linear hidden Markov models, can be found in the ground-breaking publication by Bucher (Bucher, 1990). Depending on their goals, computational approaches to promoters can be divided into two classes: general promoter recognition methods and specific promoter recognition methods. The primary goal of general promoter recognition methods is to identify TSS and/or core promoter elements for all genes in a genome, while specific promoter recognition approaches focus on the analysis of promoter regions to identify harbored regulatory elements (transcription factor binding sites) or to identify specific regulatory elements that are shared by a particular set of transcriptionally related genes. Developed methods (Ohler and Niemann, 2001; Sinha and Tompa, 2002) may be highly specific when searching the whole genome and can provide immediate functional clues as to the identity of the downstream gene. However, because of their broad coverage, general methods are best used for large-scale genome annotation. Two fundamentally different approaches are used by these specific methods: the 'alignment' methods and the 'enumerative' (or 'exhaustive') methods. Alignment methods identify unknown signals using significant local multiple alignments of all sequences. Direct multiple alignment is computationally demanding, so methods of various other strategies. Two important methods for this type are Gibbs sampling (Lawrence *et al*., 1993) and expectation maximization in the MEME system (Bailey and Elkan, 1995). Alternatively, enumerative or exhaustive methods examine all oligomers of a certain length and report those that occur far more often than expected based on the overall promoter sequence composition (Brazma *et al*., 1998; Helden, 2004). This approach has gained in popularity since the elucidation of complete genomes. In common with gene finding approaches, existing methods for general promoter prediction can be classified into two different categories, ab initio and homology based (Pedersen *et al*., 1998; Ohler and Niemann, 2001; Ohler *et al*., 2002). Computational methods that aim at the identification of promoters *ab initio* tackle the promoter-finding task by establishing a model of promoters - and possibly of non-promoters as well -, and then use this model to search for an unknown number of promoters in a contiguous DNA sequence. Depending on how the model captures promoter features, different sub-groups of *ab initio* approaches can be distinguished. For example, one can train Markov chain models of different orders on promoter and non-promoter sequences, and classify a sliding window of fixed length using Bayes' rule. In addition, discriminative counts of oligomers of a certain length are often employed.

The main goal of this work was the development of a predictive probabilistic model that can be employed to computationally recognize core promoters in long, contiguous vertebrate DNA sequences, such as, long contigs. This is of importance for DNA sequence annotation, which should discover as much information as possible. Furthermore, such a computational model could help us to determine which information in a sequence is vital for a reliable recognition, and allow us to accumulate knowledge on the biology of gene expression. The approach of this work belongs to the *general* promoter prediction methodology. Our predictive model was built to reflect the overall structure and biological context of promoter regions. Also, it is designed to predict promoters *ab initio*, i.e. using the DNA sequence of the organisms of interest as information only. We sought to devise a reliable probabilistic model that reflects the underlying biology and biological context, and that is able to cope with diverse variations (i.e., variations in positional nucleotides and transcription elements) and arbitrary interactions (i.e., non-adjacent and adjacent dependencies) inherent in promoter sequences. Non-adjacent dependencies mean the dependencies between positions with various separations.

Materials and Methods

Data set of promoter sequences The total number of promoter sequences used for this work was 2537, which were exclusively taken from the vertebrate data set of EPD (Eukaryotic Promoter Database) release 73 (Perier *et al*., 2002). The database is an annotated non-redundant collection of eukaryotic promotes and their viral pol-II promoters taken from the literature, the transcription start sites of which had been determined experimentally. The sequences have the same size of 600 bp ranging from −500 to +100, and were used as positive samples (or training data set) to build a predictive probabilistic model for vertebrate promoter recognition. The ambiguous nucleotide symbol '*N*' within sequences was randomly converted to *A*, *T*, *G*, or *C*. A portion of this data set was also used to assess predictive accuracy. The use of all vertebrate sequences is justified because of the highly similar transcription machinery in vertebrates and humans. About three-fourths of all the sequences used were of human origin.

Data set of non-promoter sequences We also use the ready-to-use negative background and test data sets containing coding (CDS) and non-coding sequences from the GENIE (Kulp *et al*., 1996) data set, which was compiled for the evaluation and comparison of different gene identification methods for the analysis of human DNA sequences. Complements of coding and non-coding sequences were used as data sets for non-promoters. These sets have already been checked for multiple closely related entries, and those with more than 80% identity by BLAST (Altschul *et al*., 1990) are not contained to be free from data redundancy. The number of coding and non-coding sequences in non-promoter data set amounted to 890 and 4345 respectively, each of 300 bp. These data sets are available at http://www.fruitfly.org/sequence/human-datasets.html. They were used to compute the overall base

composition for the null model in log-odds score computation, and some of them, chosen randomly, were employed as negative samples to evaluate the performance of the predictive model.

Standardized test data set In addition to a high-quality set of sequences suitable for making a predictive model, a large set of promoter annotations in contiguous DNA sequences were collected. This enabled us to assess the results of predictive models in real terms. However, the building of such a set for the evaluation of TSS or, more generally, for promoter recognition, is difficult. Here, one collection of well-mapped promoters was taken to allow relative performance comparisons between the developed predictive model and other promoter prediction programs. This standardized test data set was used by Fickett and Hatzigeorgiou (Fickett and Hatzigeorgiou, 1997) to assess promoter prediction programs and contains 24 promoters in 18 rather short but well studied vertebrate sequences, ranging in length from 565 to 5,663 bp.

Localization of target promoter regions by positional information content Using the promoter sequences collected as described above, we first determined the target promoter regions for our predictive model. Most existing algorithms use promoter regions which are arbitrarily determined on the basis of biological domain knowledge and experimental evidence. In contrast, we used an objective criterion, positional information content, to localize the target region to be modeled. The reason we chose this criterion is that positions with higher information content show considerable nucleotide conservation, which is probably of biological significance. In this context, we employed the definition of Tom Schneider (Schneider and Stephens, 1990) who followed Claude Shannon and defined the uncertainty measure as:

$$H(l) = -\sum_{b=A}^{T} f(b,l)\log_2 f(b,l) \text{ (bits per position)} \qquad (1)$$

where $H(l)$ is the uncertainty at position l, b is one of the bases (A, C, G, or T), and $f(b, l)$ is the frequency of base b at position l. Total information at the position can be represented by a reduction in uncertainty:

$$R_{sequence}(l) = 2 - (H(l) + e(n)) \text{ (bits per position)} \qquad (2)$$

where $R_{sequence}(l)$ is the amount of information present at position l, 2 is the maximum uncertainty at any given position (i.e. $\log_2 M$: in this case, the number of symbols or choices, M, is 4), and $e(n)$ is a correction factor required when the number of sample sequences (n) is small. In this work, we ignored the correction factor, since we have sufficient sample sequences. The entire set of $R_{sequence}(l)$ values forms a curve that represents the importance of various positions in the aligned sequences. In this work, all the promoter sequences were aligned using the transcriptional initiation point (position +1 in Fig. 1) to examine the positional information content, as defined by the formula (2). No gap, i.e. insertion or deletion, was permitted in this alignment. Fig. 1 shows the information content at each position in the original promoter stretch (600 bases ranging from –500 to +100). The importance of a particular position is clearly and consistently provided by positional information content. From Fig. 1, we determined the region ranging from –150 to +100 as a target promoter region relatively rich in information. As expected, it includes the complete core promoter and the downstream part of the proximal promoter, which are believed to be significant as they are recognized by basal transcriptional machinery and in transcriptional regulation control. Now we focus on this target promoter region and use it to construct our predictive model.

Construction of the composite dependency-reflecting model The primary goal for building the CDRM was to identify the TSSs and core promoter regions of all protein-coding genes in a vertebrate genome. To achieve such a model, we used two methodologies. The first involved the iterative subclassification of target promoter region sequences into subsets, on the basis of dependency degree and pattern within the target promoter region using our EMDD (expanded maximal dependency decomposition) procedure. This was required to incorporate the biological context and long-range dependencies into the CDRM. The second involved the application of appropriate submodels for subsets generated by the EMDD procedure. As a first step, we needed a way of measuring the amount of dependency between arbitrary positions in the target promoter region. Chi-square statistic was used to measure dependencies between the variables N_i and N_j (which take on the four possible values A, C, G, T), indicating the nucleotides at positions i and j of a sequence, i.e., to ask whether there is an association between the occurrence of particular nucleotide(s) at position i and the occurrence of other nucleotide(s) at position j in the same sequence. In the EMDD procedure, only nucleotide indicator variables N_i and N_j are used to calculate the chi-square statistic $X_{i,j}$ for each pair of positions i, j with $i \neq j$ instead of the consensus indicator K_i and the nucleotide indicator N_j, since sufficient data, with some limitations, are available. The average value of all chi-square statistics in a row of the $X^2_{i,j}$ matrix was used as a yardstick of dependency degree of positions i. Strong dependencies are found between non-adjacent (or long-range) and adjacent positions (as shown in Fig. 2). The EMDD procedure is carried out as follows:

(1) Calculate, for each position i, the average value $A_i = avg(\Sigma_{j \neq i} X^2_{i,j})$ (the average value of a row sum in Table 1), which is a measure of dependency between the variable N_i and the nucleotides at all other positions of the target promoter region, and also considers the dependency pattern of the data set D for a given p-value ($p < 0.001$). Under all null hypotheses of pairwise independences between positions, A_i will have approximately a χ^2 distribution with 9 degrees of freedom.

(2) If the data set D shows a significant dependency pattern between the positions with variable separation, we choose a value i_1 such that A_{i_1} is maximal and significant for the loose p-value ($p < 0.2$) and partition D into two subsets: $+D_{i_1}$, all sequences which have consensus nucleotides(s) at position i_1; and $-D_{i_1}$, all sequences which do not.

(3) The first two steps are now repeated on the subsets $+D_{i_1}$ and $-D_{i_1}$ and on subsets thereof, and so on, yielding a binary "tree" (often called a decision tree) with a maximum theoretical level of $\lambda - 1$ (see Fig. 3).

This subdivision process is carried out successively on each branch of the tree until one of the following four conditions occurs: (i) the $(l-1)th$ level of the tree is reached (so that no further subdivision is possible); (ii) no significant dependencies are detected between positions in a subset (so that further subdivision is

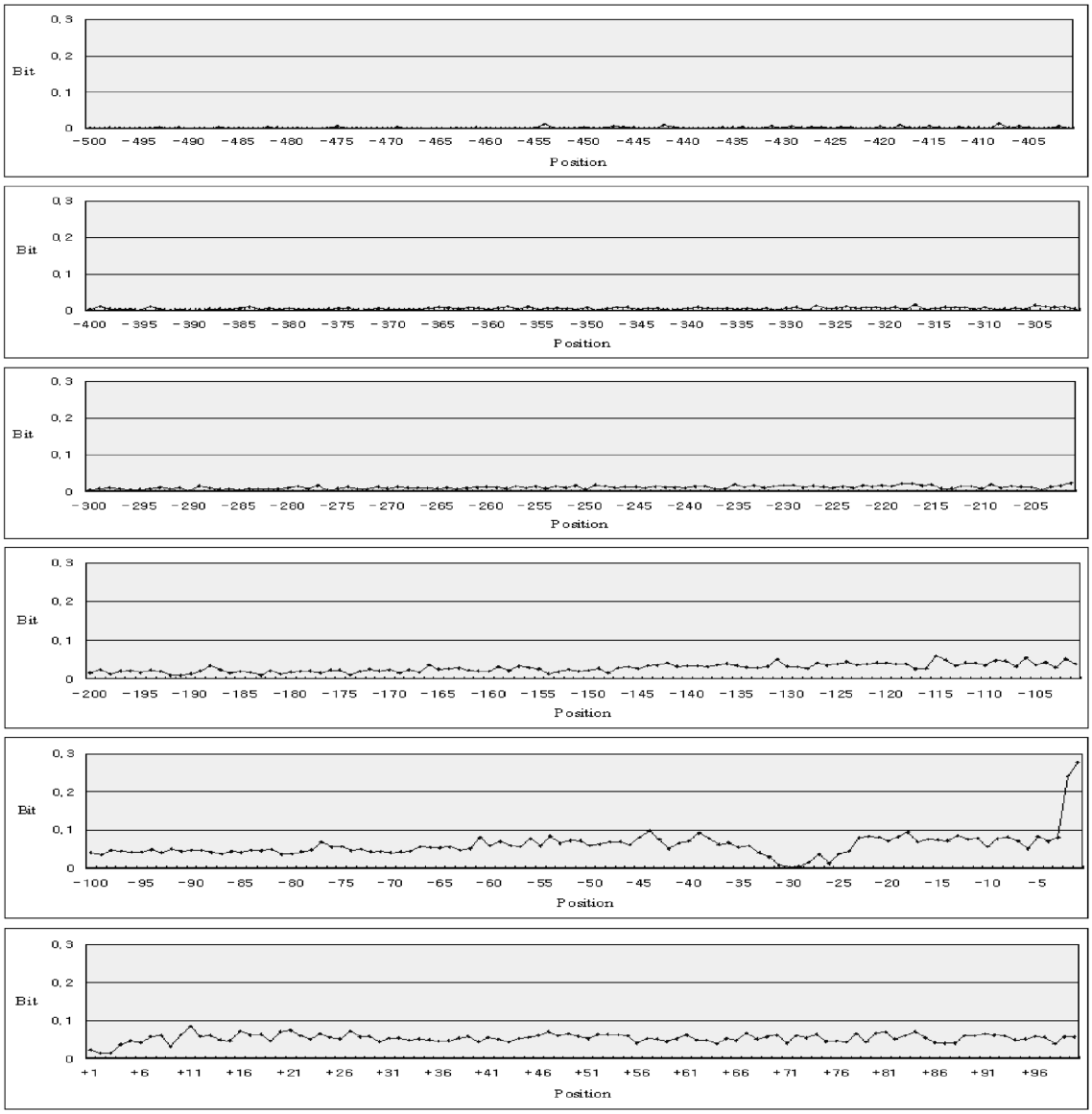

Fig. 1. Positional information content of training promoter sequences. To compute positional information content, the promoter sequences were aligned by TSS (the position +1) without any gap (i.e. insertion/deletion) permitted. The height of the curve signifies the information content of the sequences at that position and the curve itself displays both significant positions and subtle sequence patterns.

not indicated; (iii) if no significant dependency pattern with variable separation in a subset is detected (so that further subdivision is of no use); or the number of sequences in a resulting subset falls below a preset minimum value $M = 110$, as reliable conditional probabilities cannot be determined by further subdivision.

After the EMDD procedure, dependency patterns on each subset generated were examined to determine a proper submodel. If significant dependencies are detected, but are exclusively or predominantly between adjacent positions, then a first-order Markov model may be appropriate (see the Fig. 4). So we applied a first-order Markov model for each subset. Under a first-order Markov model, the probability of generating the sequence $X = x_1, x_2, ..., x_\lambda$ is given by

$$P_{WAM}(X) = p^{(1)}(x_1)p^{(2)}(x_2|x_1)p^{(3)}(x_3|x_2)\ldots p^{(\lambda)}(x_\lambda|x_{\lambda-1})$$

$$= p^{(1)}(x_1)\prod_{i=2}^{\lambda} p^{(i)}(x_i|x_{i-1}) \quad (3)$$

where $p^{(i)}(z|y)$ is the conditional probability of generating nucleotide z at position i given nucleotide y at position $i-1$, which is estimated from the corresponding conditional frequency in this work.

Scoring strategy Probability itself is not the most convenient number to use as a score, so we used the log odds score, which is the logarithm of the probability of a sequence divided by the

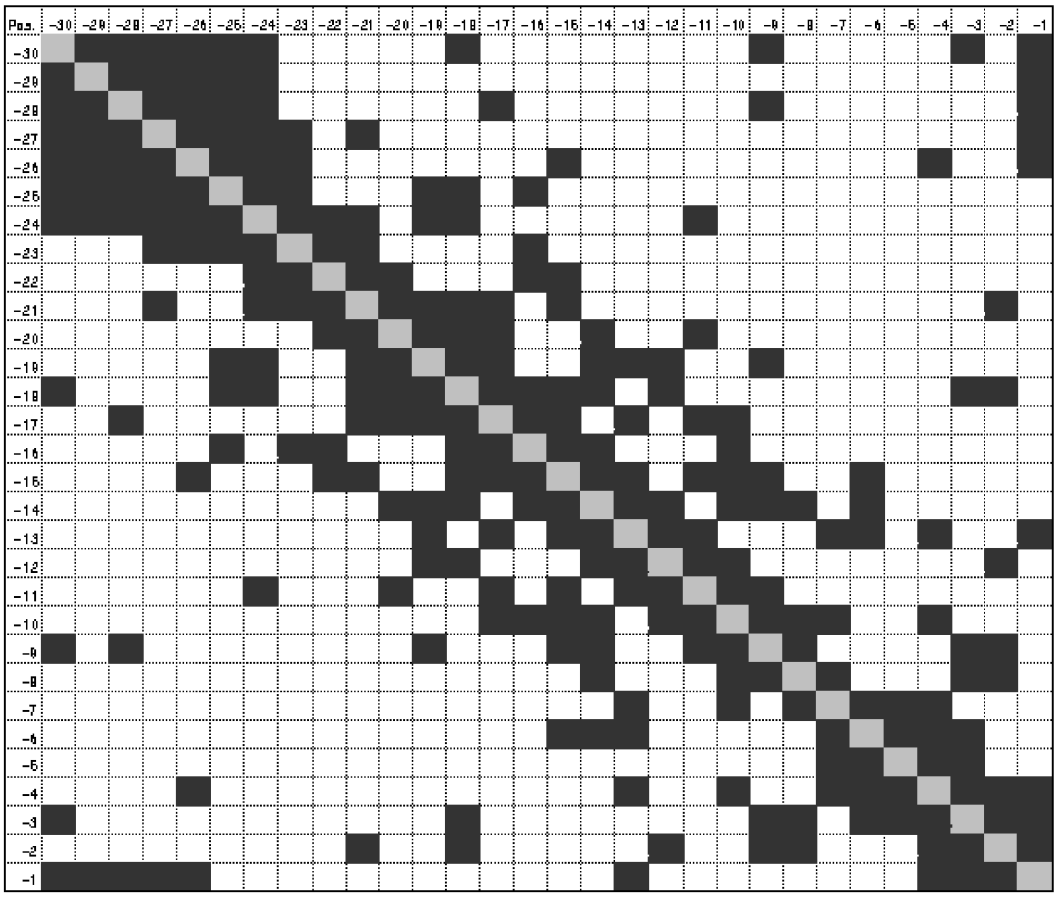

Fig. 2. Statistically significant dependencies between arbitrary positions. For each pair of positions i (y-axis) and j (x-axis) with $i \neq j$, a N_i versus N_j contingency table was constructed and the X^2 values of chi-square statistics were calculated. Black squares indicate pairs with significant chi-square values at the $p < 0.001$ (9 degree of freedom) level; all other squares are white, except those on the diagonal ($i = j$), which are colored light gray for visual reference. Owing to limited space, only pairs in the range [–30, –1] are shown in the figure.

probability according to the null model. Log odds scores are summed to obtain a combined log odds score for a particular position in the promoter sequence. The sums of log odds scores in bits may be converted to odds scores using the odds score formula $= 2^{(\text{log-odds score})}$. These numbers vary from small fractions to large numbers and reflect variations in the likelihood of a target promoter region at each sequence position. As a result these numbers might be difficult to intuitively understand, so a new formula for scoring,

Table 1. Dependencies between positions in target promoter region sequences

Position i	Position j								Average
	-150...-30	-29	-28	-27	-26	-25	-24	-23...+100	
-150... -30	....	....	....	....	....	....	....	....	....
-29	....	-	283.71	383.67	223.69	171.28	28.02	....	26.91
-28	....	283.71	-	356.45	306.63	182.18	62.52	....	28.45
-27	....	383.67	356.45	-	287.74	262.4	77.71	....	29.16*
-26	....	223.69	306.63	287.74	-	189.95	106.13	....	22.65
-25	....	171.28	182.18	262.4	262.4	-	168.02	....	22.31
-24	....	28.02	62.52	77.71	77.71	168.02	-	....	16.29
-23... +100	....	....	....	....	....	....	....	....	....

χ^2 statistic for nucleotide indicators N_i versus N_j. For each pair of positions $\{i, j\}$ with $i \neq j$, a 4×4 contingency table was constructed for nucleotide indicator variables N_i versus N_j, identifying the nucleotide at position i and j respectively. For each contingency table, the χ^2 statistic was calculated and is listed in the table below (not all data are shown because of limited space). The last column in the table lists the average of the values in each row, which is a measure of dependency between N_i and N_j.

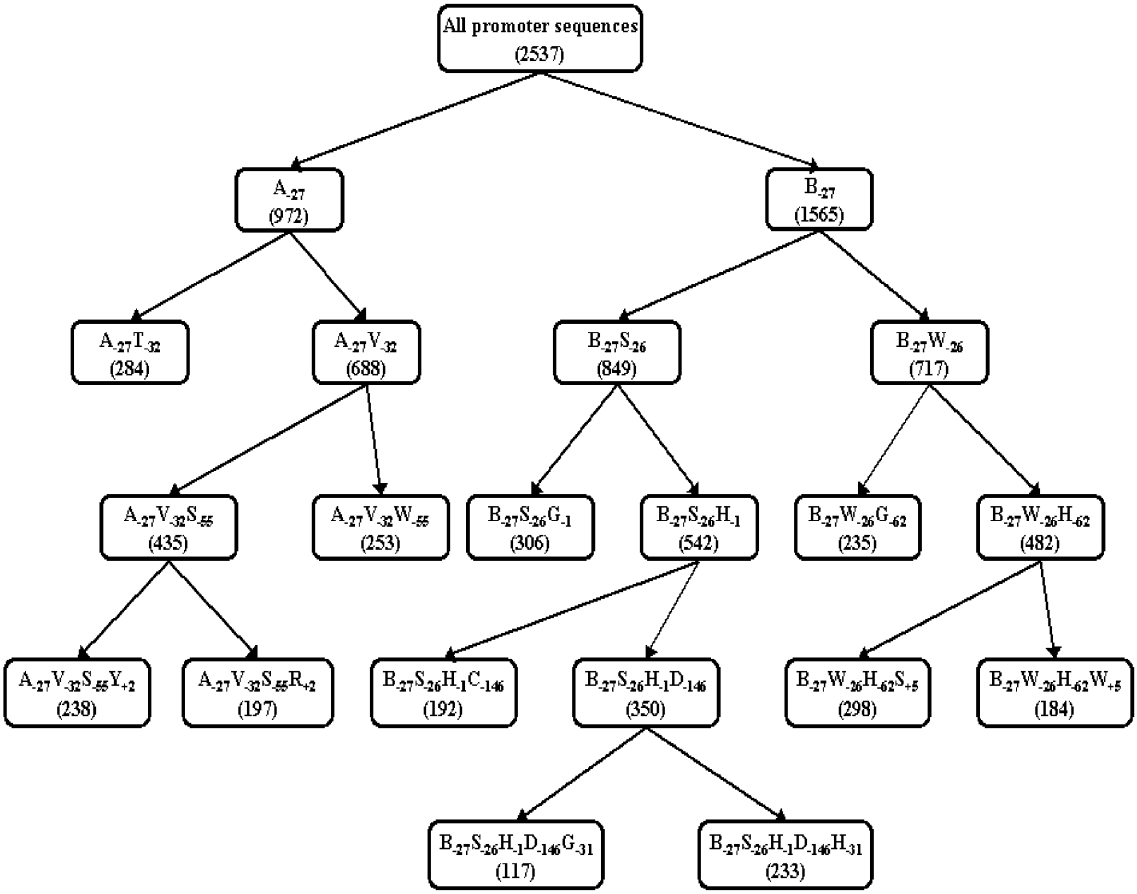

Fig. 3. Subclassification of target promoter sequences using the EMDD procedure. Each rounded box represents a subset of target promoter region sequences corresponding to a pattern of matches/mismatches to the consensus nucleotide(s) or the nucleotide with the highest frequency at a set of positions. For example, $A_{-27}T_{-32}$ is the set of target region sequences with consensus nucleotides *A* and *T* at positions −27 and −32. The number of corresponding target region sequences in each subset is given in parentheses beneath the pattern description. IUB single letter symbols are used to represent groups of nucleotides (for example, V means A or C or G).

based on allocating a score between 0 and 100, is introduced here.

$$Score(w) = 100 \times (O_w - Min)/(Max - Min) \quad (4)$$

where *Min* and *Max* are the minimum and maximum possible totals, *i.e.* the sum of the lowest and highest log-odds scores in each position in the scoring matrix, and O_w is the total log-odds score of the 250 nucleotides in the subsequence being scored.

Strategy used to detect core promoters and TSSs in genomic sequences With the complete CDRM it is straightforward to detect target promoter regions (i.e. core promoter regions and TSSs) in a query sequence. A sliding 250 bp window algorithm is used in this work. It can be summarized as follows:

(A) Set a sliding window of 250 bp at the beginning of the query sequence.

(B) Select the appropriate submodel of the CDRM for a sliding window of 250 bp according to the decision rules provided by the decision tree of Fig. 3. For example, if the nucleotides at positions −27 and −32 in the sliding window are *A* and *T* respectively, the first-order Markov chain model for the $A_{-27}T_{-32}$ (i.e. the submodel for corresponding leaf node in the decision tree in Fig. 3) will be selected.

(C) Calculate the score *S*(*W*) of the sliding window with the corresponding submodel (i.e. one from the separate first-order Markov model for each subset).

(D) If the score *S*(*W*) exceeds the predefined threshold, insert a new node into the double linked list for the hit.

(E) Shift the sliding window 1 bp and repeat steps *B* through *D* until the right end of the sliding window is located at the end of the query sequence.

(F) To obtain all of the most appropriate hits, traverse the double linked list using recursive calls and obtain all hits which meet the following criteria: (i) hits do not overlap with each other; (ii) if hits overlap, the hit with maximal *S*(*W*) is taken.

Results

A variety of quantitative measures have been proposed for characterizing the sensitivity and specificity of prediction methods in the bioinformatics field, especially in the gene prediction realm (Burset and Guigo, 1996). The definition of Burset and Guigo is widely used to assess the accuracy of prediction or classification programs in the bioinformatics

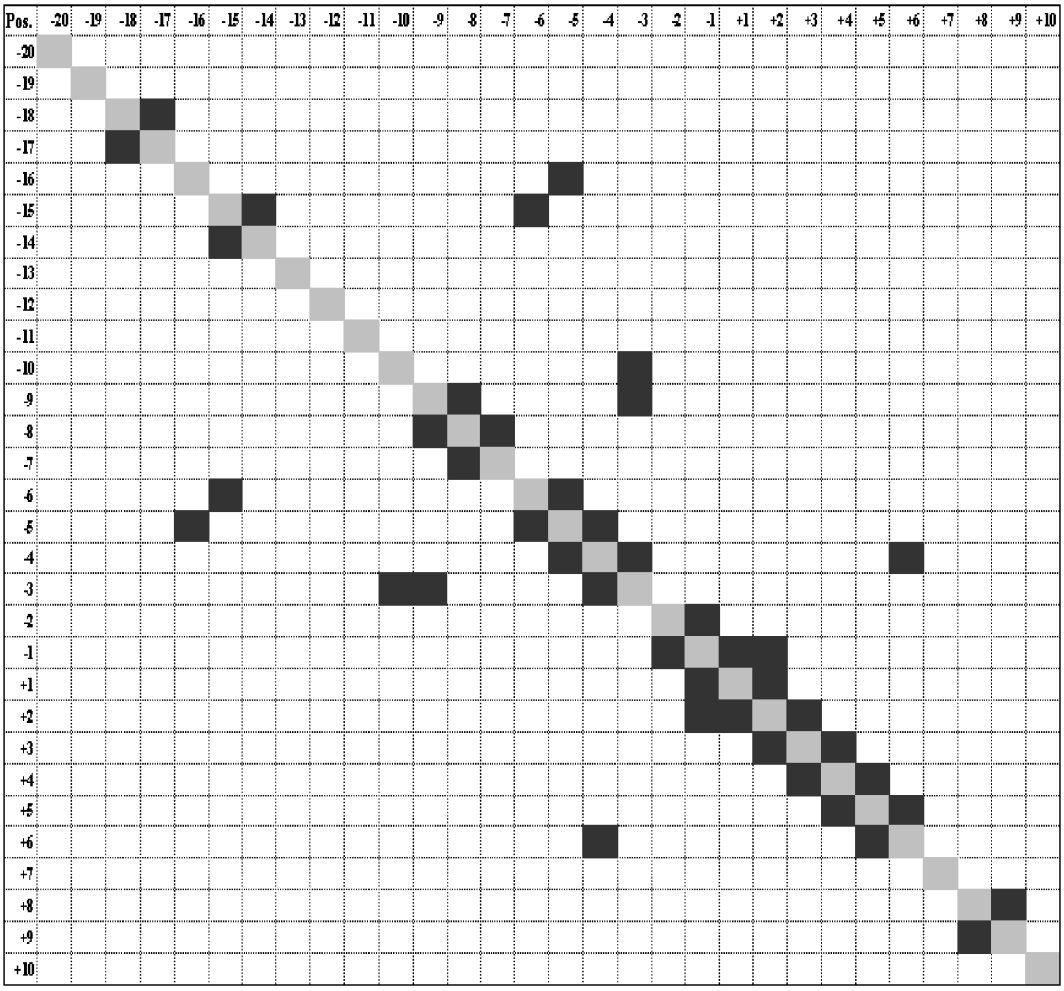

Fig. 4. Dependency pattern of the subset $A_{-27}T_{-32}$. For each pair of positions i (y-axis) and j (x-axis) with $i \neq j$, a N_i versus N_j contingency table was constructed and the X^2 value of the chi-square statistics was calculated. Black squares indicate pairs with significant chi-square values at the $p < 0.001$ (9 degree of freedom) level; all other square are white, except for those on the diagonal ($i = j$), which are colored light gray for reference.

field. Fundamentally, accuracy is related to the degree of concordance between predicted and actual signal sites. The outcome of a classification experiment results in a number of correctly classified samples from both classes, i.e., "true positives (TP)" and "true negative" (TN). In the same sense, the misclassified samples are called "false positives" (FP) and "false negatives" (FN). The correlation coefficient (CC) of an experiment is calculated using these four numbers. It is defined by: -

$$CC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TN + FN) \times (TP + FN) \times (TN + FP)}} \quad (5)$$

Values range between −1 and 1; a CC of one means a perfect prediction, a CC of zero indicates a random prediction, and a CC of −1 shows perfect anti-prediction. The CC value depends on the proportion of negative and positive samples, which means that it cannot be compared across different data sets in general. The sensitivity (Sn) and specificity (Sp) of a classification are defined as: -

$$Sn = \frac{TP}{TP + FN}, \; Sp = \frac{TP}{TP + FP} \quad (6)$$

To assess the predictive accuracy of the CDRM in terms of *Sn*, *Sp* and *CC*, we first constructed a special test data set consisting of 253 positive samples and 251 negative samples. The negative set comprised 140 intron sequences and 111 CDSs, which were extracted randomly from CDSs and intron sequence data sets, as described in the previous section. The positive data set was randomly extracted from all target promoter region sequences (2537), and the remainders were used to construct the CDRM. We then tested predictive accuracy on the test data using a threshold 80, and obtained the result Sn = 0.97, Sp = 0.57, and CC = 0.33 (see the Table 2).

In order to compare the relative performance of the CDRM

Table 2. Result of performance testing (sensitivity and specificity)

		Predicted positive (432)	Predicted negative (72)
Promoter (253)		True positive (246)	False negative (7)
Non-promoter	Intron (140)	False positive (186)	True negative (65)
	CDS (111)		

Numbers in parenthesis indicate the number of corresponding sample sequences.

Table 3. Overall accuracy of programs tested by Fickett and Hatzigeorgiou

	Audic	Autogene	GeneID	NNPP	P'Find	P'Scan	TATA	TSSG	TSSW
Sensitivity	5.24	7/24	10/24	13/24	7/24	3/24	6/24	7/24	10/24
	24%	29%	42%	54%	29%	13%	25%	29%	42%
Specificity	33 fp	51fp	51 fp	72 fp	29 fp	6 fp	47 fp	25 fp	42 fp
	1/1004	1/649	1/649	1/460	1/1142	1/5520	1/705	1/1325	1/789
	bp	bp	bp	bp	bp	bp	bp	bp	bp

For each program, the sensitivity (as a number and as a percentage of promoters correctly detected) and specificity (as the number of false positives and as the number of base pairs per false positive) is given (an excerpt taken from a paper by Fickett and Hatzigeorgiou (1997)).

with other programs, we tested how well it performed on the genomic sequence data used by Fickett and Hatzigeorgiou (Fickett and Hatzigeorgiou, 1997). While it is desirable to compare programs with standardized test data to evaluate relative performance, it should be remembered that not all programs are designed with the same intent, and to some extent any comparison is certain to be unfair. However, such a comparison does serve to illustrate the strengths and weaknesses of programs. From this viewpoint, we tested how well the CDRM would perform on the standardized test data mentioned in previous section and compared its performance result with those of other programs, which Fickett and Hatzigeorgiou had evaluated using the same test data in their famous review paper (Fickett and Hatzigeorgiou, 1997). However, they evaluated only the ability to approximately locate the TSS. If a program produced a promoter prediction but not an explicit TSS, they took the 3' end of any predicted promoter window as the predicted TSS. The predicted TSS, explicit or implicit, was counted as correct if it was within 200 bp 5', or 100 bp 3', of any experimentally mapped TSS. Using these criteria accuracy results are summarized in Table 3. Table 3 shows that the algorithms tested found that 13~54% of the true promoters in the test set and reported, on average, 40 false positives. Given the same criteria, however, our CDRM identified 16 (67%) of the true promoters and reported 39 false positives. There may be room for some debate concerning problems of a limited sample size and the possibly skewed nature of the samples, but this result demonstrates that the CDRM is superior to other programs tested.

Discussion

Promoter recognition has an important bearing on the elucidation of gene regulation, which is one of the most important research topics in molecular biology. It is therefore important to precisely locate regulatory regions and examine them in detail, either computationally or experimentally, to elucidate the mechanisms that control their expressions. An evaluation of which features improve the quality of promoter predictions can also help us to understand how promoters are actually recognized in the cell.

From the annotation point of view, promoter identification can help gene finding algorithms to identify the 5' UTRs (untranslated regions) that can span up to tens of thousands of kilobases, and determine the exact 5' boundary of a gene. In most cases, gene finding algorithms do not determine the exact 5' end of a gene since 5' UTRs show marked length variations and do not show significant probabilistic/statistical properties, which are difficult to model reliably. Furthermore, such algorithms can also help to detect genes, namely those that are rarely expressed and thus are not registered in cDNA libraries, short genes that are easily missed, and non-coding RNA genes with no codon statistics.

Our approach belongs to the *general* promoter prediction type of method, the primary goal of which is to identify the TSSs and core promoter regions of all protein-coding genes in a genome, and not the seeking *specific* regulatory elements. For this purpose we propose CDRM as a predictive model for vertebrate promoters, as it determines the most significant dependencies between positions, and allows for non-adjacent and adjacent dependencies. In order to construct our predictive model, the EMDD procedure was designed, which iteratively decomposes a data set into subsets based on dependency degree and pattern within the target promoter region. Such iterative subclassification leads to a binary tree. First-order Markov models were built on each leaf node of the binary tree, and became submodels of the CDRM. In this

respect, CDRM is well tuned to the nature of target promoter region sequences by reflecting the observed long-range dependencies in the region and by incorporating the adjacent dependencies explicitly. It actually represents a combination of first-, second-, third- and even much higher order or longer-range dependencies through the EMDD procedure.

CDRM improves discrimination between true and false samples by accounting for several potentially important non-adjacent and adjacent interactions, and may also give some insight into how the target promoter region is recognized. It may be of interest in the future to apply this method to other biological signals, e.g. other transcriptional or translational signals in DNA/RNA or perhaps even to protein motifs. If larger sets of sequences have accumulated sufficiently, more complex dependencies can be more reliably measured and modeled. One important future challenge is the development of more flexible and sensitive approaches to the analysis of available sequence data, which may allow the detection of more subtle biological features. In the longer term, it may even be possible to construct realistic models of such complex biological processes as transcription and pre-mRNA splicing *in silico*.

References

Altschul, S., Gish, W., Miller, W., Myers, E. and Lipman, D. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403-410.

Bailey, T. and Elkan, C. (1995) Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning* **21**, 51-83.

Brazma, A., Jonassen, I., Vilo, J. and Ukkonen, E. (1998) Predicting gene regulatory elements in silico on a genomic scale. *Genome Res.* **8**, 1202-1215.

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

Burset, M. and Guigo, R. (1996) Evaluation of gene structure prediction programs. *Genomics* **34**, 353-367.

Fickett, J. and Hatzigeorgiou, A. (1997) Eukaryotic promoter recognition. *Genome Res.* **7**, 861-878.

Helden, J. (2004) Metrics for comparing regulatory sequences on the basis of pattern counts. *Bioinformatics* **20**, 399-406.

Hernandez, E., Johnson, A., Notario, V., Chen, A. and Richert, J. (2002) AUA as a translation initiation site in Vitro for the human transcription factor Sp3. *J. Biochem. Mol. Biol.* **35**, 273-282.

Ko, J., Na, D. S., Lee, Y. H., Shin, S. Y., Kim, J. H., Hwang, B. G., Min, B. I. and Park, D. S. (2002) cDNA microarray analysis of the differential gene expression in the neuropathic pain and electroacupunction treatment models. *J. Biochem. Mol. Biol.* **35**, 420-427.

Kulp, D., Haussler, D., Reese, M. and Eeckman, F. (1996) A generalized Hidden Markov Model for the recognition of human genes in DNA. *ISMB96*, States, D. J., Agarwal, P., Gaasterland, T., Hunter, L., Smith, R. (eds.), pp. 134-142, AAAI/MIT Press, St. Louis, USA.

Lawrence, C., Altschul, S., Boguski, M., Liu, J., Neuwald, A. and Wootton, J. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**, 208-214.

Ohler, U., Liao, G., Niemann, H. and Rubin G. (2002) Computational analysis of core promoters in the Drosophila genome. *Genome Biol.* **3**, 1-17.

Ohler, U. and Niemann, H. (2001) Identification and analysis of eukaryotic promters: recent computational approaches. *Trends Genet.* **17**, 56-60.

Pedersen, A., Baldi, P., Chauvin, Y. and Brunak, S. (1999) The biology of eukaryotic promoter prediction a review. *Comput. Chem.* **23**, 191-207.

Perier, R., Praz, V., Junier, T., Bonnard, C. and Bucher, P. (2000) The Eukaryotic Promoter Database. *Nucleic Acids Res.* **28**, 302-303.

Schneider, T. and Stephens, R. (1990) Sequence Logos: A New Way to Display Consensus Sequences. *Nucleic Acids Res.* **18**, 6096-6100.

Sinha, S. and Tompa, M. (2002) Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.* **30**, 5549-5560.