CABIOS

Vol. 12 no. 5 1996
Pages 431-439

The statistical significance of nucleotide position–weight matrix matches

Jean-Michel Claverie and Stéphane Audic

Abstract

Motivation: *To improve the detection of nucleotide sequence signals (e.g. promoter elements) by position–weight matrices (PWM) using the concept of statistically significant matches.*
Results: *The Mksite program was originally developed for analyzing protein sequences. We report NMksite, a new version adapted to the processing of nucleotide sequences. NMksite creates PWM from nucleotide sequence block alignments or occurrence tables using three weight computation schemes. An original feature of NMksite is the numerical computation of the statistical significance of PWM matches. The utility of this concept is demonstrated in the context of the prediction of splice sites and promoter regions.*
Availability: *Mksite and other components of the MODEST (Motif DEsign and Search Tool) package (written in C/Unix) are available at http://igs-server.cnrs-mrs.fr*
Contact: *E-mail: jmc@igs.cnrs-mrs.fr*

Introduction

Position–weight matrices (PWM) are traditionally used to represent the subtle sequence patterns stored in the local multi-alignment of functionally related molecular sequences. Besides the definition of protein motifs [Gribskov *et al.* (1990) and references therein], PWM have also found successful applications in the representation of nucleotide sequence signals (Staden, 1984, 1989; Stormo, 1990; Fickett, 1996a) such as promoter elements (Bucher, 1990; Fickett, 1996b; Wingender *et al.*, 1996) and splice sites (Senapathy *et al.*, 1990). Several methods have been proposed to derive weights from the propensity of a given symbol to be found at a given position in a block alignment (Figure 1A) or an occurrence table (Figure 1B). Three of them have been implemented in NMksite, the nucleotide version of the Mksite program (Claverie, 1994). Nucleotide PWM are simple mathematical objects and can lead to fast sequence searches. Yet their use has been limited to somewhat heuristic implementations due to a lack of reliable methods to assess the statistical significance of the matches. Here, we introduce the numerical computation of the statistical significance of nucleotide PWM matches, as implemented in NMksite. Sequence patterns tend to have a lower information content in nucleic acids than in proteins. The concept of statistical significance is thus expected to be particularly useful for the interpretation of nucleotide PWM matches. We take examples from two difficult problems, the prediction of splice sites and the identification of promoter regions, to illustrate this claim.

Theory

Different methods have been proposed to store the information contained in a block alignment or an occurrence table (Figure 1A and B) in a set of weights representing the propensity of a given symbol to occur at a given position. For nucleic acids, it is customary to attribute to each nucleotide (e.g. A, C, G, T, hence $i = 1$–4) at each position p a score s_{ip}, such as:

$$s_{ip} = \log \frac{q_{ip}}{f_i} \qquad p = 1 \text{ to } w \tag{1}$$

where q_{ip} is the observed frequency of nucleotide i at position p in a block alignment of N sequences and width w (Figure 1A), and f_i is its frequency in a random sequence of reference (most often $f_i = 25\%$). The local goodness of fit between a target sequence and the 'signal' represented by the matrix is measured by an aggregate score:

$$S = \sum_{p=1,w} s*_p \tag{2}$$

where s^*_p denotes the weight corresponding to the nucleotide found in the target sequence. According to equation (1), the aggregate score S can be interpreted as a lod score between two competing hypotheses: the current target sequence corresponds to an occurrence of the signal, or the current target sequence is random.

The problem of unobserved nucleotides

PWM are defined from a finite number of sequences exhibiting a limited variability. Hence, every nucleotide is not always observed at least once at each position, an obvious problem with equation (1). To solve this problem, we need to be able to encode the information from a block alignment of an arbitrary small number ($N \geq 1$) of sequences. For this, we

Structural and Genetic Information Laboratory, CNRS–E.P. 91, Institute of Structural Biology and Microbiology, 31 Chemin Joseph Aiguier, Marseille 13402, France

A

```
GTATAAAAAAGCGG
CTATAAAAGGCCC
GTATAAAGGGGCG
GTATATAAGCGCG   N=6
CTATAAAGGGGCC
GTATAAAGGCGGG
<-- W=13  -->
```

B

Nmksite -o ali.tata > occ.tata

```
1:C  2.00 G  4.00
2:T  6.00
3:A  6.00
4:T  6.00
5:A  6.00
6:A  5.00 T  1.00
7:A  6.00
8:A  3.00 G  3.00
9:A  1.00 G  5.00
10:C  2.00 G  4.00
11:C  2.00 G  4.00
12:C  4.00 G  2.00
13:C  2.00 G  4.00
```

C

Nmksite -occ=occ.tata

```
# pattern source= occ.tata
# **** This is an OCCURRENCE file ****
# Pseudo Score mode= 1, threshold= 0, from 6 to 6 seqs, width= 13
# [PWM] max score: 1.65  , min score: -1.79
# [SumScore] perfect matches range:[17.5 - 13.4]
# [SumScore] Average: -9.1 , Variance= 24.4,  Std. Deviation= 4.94
# ------------------------------------------
#     For an average (1000 nt) Sequence
# ------------------------------------------
# Median   of best   score= 7.55
# 1/10     significant score= 10.1
# 1/100    significant score= 12.9
# 1/1000   significant score= 14.9
# 1/10000  significant score= 16.7
# 1/100000 significant score= 17.5
1:A -1.79 C  0.31 G  1.13 T -1.79 N -0.53
2:A -1.79 C -1.79 G -1.79 T  1.65 N -0.93
3:A  1.65 C -1.79 G -1.79 T -1.79 N -0.93
4:A -1.79 C -1.79 G -1.79 T  1.65 N -0.93
5:A  1.65 C -1.79 G -1.79 T -1.79 N -0.93
6:A  1.41 C -1.79 G -1.79 T -0.39 N -0.64
7:A  1.65 C -1.79 G -1.79 T -1.79 N -0.93
8:A  0.77 C -1.79 G  0.77 T -1.79 N -0.51
9:A -0.39 C -1.79 G  1.41 T -1.79 N -0.64
10:A -1.79 C  0.31 G  1.13 T -1.79 N -0.53
11:A -1.79 C  0.31 G  1.13 T -1.79 N -0.53
12:A -1.79 C  1.13 G  0.31 T -1.79 N -0.53
13:A -1.79 C  0.31 G  1.13 T -1.79 N -0.53
```

Fig. 1. (**A**) A nucleotide sequence signal represented as a block alignment. (**B**) The occurrence table corresponding to the same alignment. (**C**) A lod score PWM derived from (**A**) or (**B**).

will now compute the weight matrix as:

$$s_{ip} = \log \frac{Q_{ip}}{f_i} \qquad (3)$$

and require Q_{ip} to follow:

$$Q_{ip} \rightarrow q_{ip} = \frac{ob_{ip}}{N} \text{ for large } N\text{, and}$$

$$Q_{ip} > 0 \text{ for } ob_{ip} = 0 \qquad (4)$$

thus retaining a finite value for unobserved nucleotides. ob_{ip} denotes the number of nucleotides of type i observed at position p.

We will require two additional properties to retain a simple biological explanation of the s_{ip} weights in term of preference ($s_{ip} > 0$) or avoidance ($s_{ip} < 0$) of a given nucleotide. First, a nucleotide i not yet observed at position p should not correspond to a positive weight, hence:

$$s_{ip}(ob_{ip} = 0) \leq 0 \Rightarrow Q_{ip}(ob_{ip} = 0) \leq f_i \qquad (5)$$

On the other hand, a nucleotide occurring more often than its random frequency f_i should correspond to a non-negative score, hence:

$$s_{ip}(ob_{ip} \geq Nf_i + 1) \geq 0 \Rightarrow Q_{ip}(ob_{ip} \geq Nf_i + 1) \geq f_i \qquad (6)$$

For each position p, a suitable expression for Q_{ip} is:

$$Q_{ip} = \frac{ob_{ip} + \epsilon_i}{N + \epsilon} \qquad (7)$$

where ϵ_i is a 'pseudo-count' for nucleotide i and

$$\epsilon = \sum_i \epsilon_i$$

As previously discussed in the context of protein sequence PWM (Claverie, 1994), there are three main ways to compute ϵ_i.

'Constant mode': ϵ_i independent of i. With ϵ_i identical for each nucleotide (Berg and von Hippel, 1987), equation (7) becomes:

$$Q_{ip} = \frac{ob_{ip} + \frac{\epsilon}{4}}{N + \epsilon} \qquad (8)$$

Given $f_i \approx 1/4$, equations (5) and (6) are verified for any positive N and ϵ values, and the weight attributed to an unobserved nucleotides:

$$s_i^0 = \log \frac{\epsilon}{4(N + \epsilon)f_i} \qquad (9)$$

'Proportional mode': ϵ_i as a function of the priori frequency f_i. The pseudo-counts for each nucleotide can also be distributed according to their background frequency f_i (Lawrence *et al.*, 1993). Equation (7) then becomes:

$$Q_{ip} = \frac{ob_{ip} + \epsilon f_i}{N + \epsilon} \qquad (10)$$

The weight attributed to an unobserved nucleotide now becomes independent of its nature:

$$s_i^0 = \log \frac{\epsilon}{N + \epsilon} \qquad (11)$$

again, equations (5) and (6) are verified for any positive value of N and ϵ.

'Matrix mode': ϵ_i as a function of nucleotide substitution frequencies. Finally, pseudo-counts (i.e. an anticipation of what could be observed next) might be inferred from the information already gathered on the nucleotide signal we are trying to capture with the PWM. Given the type of nucleotides that have been observed at a given position, we can express the corrected observed frequency as:

$$Q_{ip} = \frac{ob_{ip} + \epsilon \sum_{j=1}^{4} q_{jp} T_{ij}}{N + \epsilon} \qquad (12)$$

where the pseudo-counts for unobserved nucleotides ($ob_{ip} = q_{ip} = 0$) at position p are now proportional to the off-diagonal T_{ij} elements of a transition matrix, i.e. a set of substitution rates from nucleotide i to j. Equations (5) and (6) are then valid when

$$Max(T_{ij}) \leq \frac{N + \epsilon}{\epsilon} f_i, \qquad \forall_i \text{ and } j \neq i \qquad (13)$$

and

$$\frac{1}{\epsilon} \geq f_i - Min(T_{ij}) \qquad (14)$$

The use of the transition matrix of Gojobori *et al.* (1982), adapted by Li *et al.* (1984), allows:

$$\epsilon \leq \sqrt{N} \text{ and } \epsilon \leq 4 \qquad (15)$$

as a natural choice for ϵ.

The expected distribution of nucleotide PWM scores

Once computed from a block alignment or an occurrence table, PWM are applied to new (anonymous) sequences eventually to locate novel instances of the signal they represent. For example, PWM are widely used for the identification of transcription factor binding sites (Bucher, 1990; Fickett, 1996b; Wingender *et al.*, 1996). For a PWM to be useful, biologically significant signals must correspond to the highest scoring matches, with rare exceptions of false-positive identifications. Optimal score thresholds are often computed using a test set of 'positive' versus 'negative' sequences. An optimal threshold tries to realize the best compromise between the sensitivity of signal detection (e.g. the fraction of biologically significant signal actually retrieved) and its specificity (e.g. the fraction of actual signals among all detected). Besides its biological significance, the validity of a score threshold is also directly related to its statistical significance in the context of a given experiment. Obviously, we cannot expect to identify biologically relevant signals with a given PWM if the best-matching scores have a high probability of occurring by chance in a random nucleotide sequence. This probability of random occurrence will depend on the target sequence length and its nucleotide composition, and affect the result of a PWM search in a way we would like to predict. Until now, nucleotide PWM have mostly been used without reference to the statistical significance of their matches. In the following sections, we describe the general properties of the probability distribution of nucleotide PWM matching scores, and apply the concept of statistical significance to some classical problems of nucleotide signal detection.

Numerical computation of the random score distribution

A nucleotide PWM of width w (Figure 1) can be matched in 4^w distinct ways corresponding to the same number of distinct pathways to the final aggregate scores. From the nucleotide composition of the target sequence, one can, in principle, compute the individual probabilities associated with every pathway and score. In practice, this becomes a cumbersome computation as w increases. Fortunately, the number of practically distinct aggregate scores is much less than 4^w as many different matches produce the same final score. To a suitable approximation (e.g. by using 100 or 1000 values to span the min–max score range), the various s_{ip} (hence the aggregate scores S) can be scaled to positive integers. The various integer approximations of s_{ip} are then used to index an array where the associated probabilities (obtained by summing the contributions of the pathways leading to the same aggregate score) are stored iteratively from $p = 1$ to w. Using this approach, the probability of every distinct (to a given approximation) aggregate score can be rapidly computed for any realistic width w (i.e. up 100 positions). Our random model assumes that all sequence positions are independent. We have previously shown (Claverie, 1996) that, despite its simplicity, this model provides good estimates for the frequencies of sequence patterns found in natural sequences (with the exception of low-entropy and CG-rich motifs). As anticipated, the probability density $p(S)$ of observing an individual PWM match associated with score S is close to Gaussian (Staden, 1989). From $p(S)$, we can compute the cumulative distribution $c(S)$, i.e. the probability for an individual match to score $\leq S$. Figure 2 exhibits $p(S)$ (bottom, 'single') and $c(S)$ (top, 'single') for Bucher's TATA box PWM (Bucher, 1990) used on a random target sequence with an even nucleotide composition (A:T:G:C). The cumulative distribution $c(S)$ is not yet the proper one to assess the statistical significance of a given match in a target sequence of arbitrary length L. For this, we need to compute the cumulative probability $P(S)$ of all matches scoring $\leq S$ when trying the PWM on all sequence windows of width w within the target sequence:

$$P(S) = c(S)^{L-w+1} \qquad (16)$$

and its complement, the probability for at least one (best) match to score $> S$, i.e. the statistical significance of such a score:

$$1 - P(S) \qquad (17)$$

Again, equation (16) assumes that all sequence positions are independent and that overlapping PWM matches can occur. Figure 2 (top) shows $P(S)$ for various lengths of the target sequence ($L = 250$, 1000 and 10 000) as well as (bottom) the density $\frac{dP}{dS}$ for $L = 250$. The distribution $c(S)$ (individual match scores) is very different in range and shape from the distribution $P(S)$ (best-match scores). While the average expected individual score is about -21 (SD = 4.9), it is not unlikely that the best PWM match within a sequence of length ≥ 250 could exceed -7 (thus > 3 SD away from the mean of expected individual scores). Similarly, statistically significant (best) matches in the context of a 250-nucleotide-long target sequence may become much less so in the context of a larger target sequence. For instance, a match scoring -2.5 is highly significant within a 250-nucleotide sequence, but quite likely to occur by chance in a 10 000-nucleotide sequence.

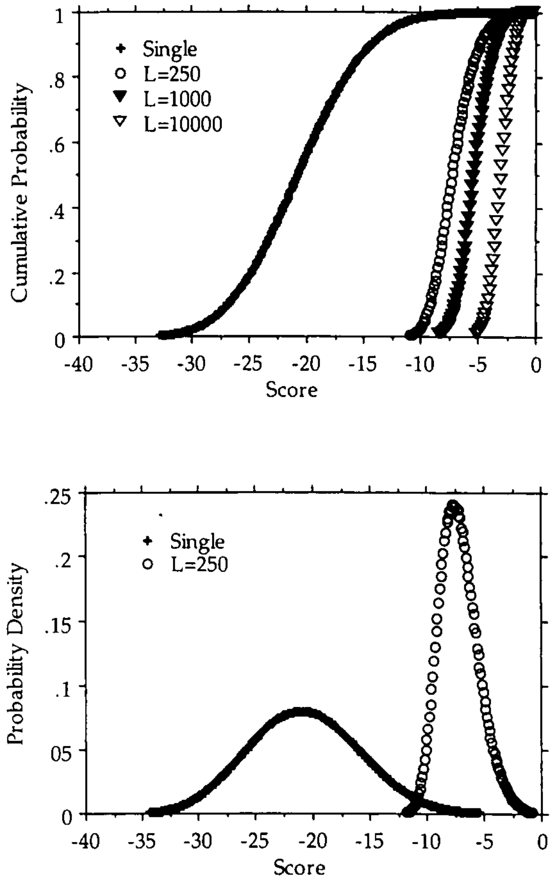


Fig. 2. Probability of best random score versus target sequence length. Bucher's TATA box PWM is used. **Top**: Cumulative probability that all matches score less than a given value within random target sequences of various lengths: $L = 15$ (single, the width of the TATA box PWM), $L = 250$ (a typical size for a promoter region), $L = 1000$ and $L = 10\,000$ (a typical span for a small human gene). **Bottom**: corresponding probability density for $L = 15$ and $L = 250$. As the target sequence length increases, the best-score probability density shifts from a Gaussian to an extreme value distribution.

Implementation

A standard C/unix program, Mksite (in two separate versions for nucleotide or protein PWM), implements two functions: (i) the computation of a lod score PWM from a block alignment (Figure 1A) or an occurrence table (Figure 1B), using the 'constant', 'proportional' or 'matrix' pseudo-count calculation modes; (ii) for the resulting PWM (Figure 1C), the computation of the score thresholds corresponding to various statistical significance levels for a target sequence of a given length and composition.

Figure 3 shows the command line syntax of NMksite (as printed when typing 'NMksite' with no other argument. The protein and nucleotide versions of Mksite are part of the MODEST package (MOtif DEsign and Search Tool). This package includes the Dbsite program for scanning a set of fasta-formatted sequences with a PWM, a program to generate directly block alignments from BLAST (Altschul *et al.*, 1990) outputs, as well as other tools for manipulating block alignments.

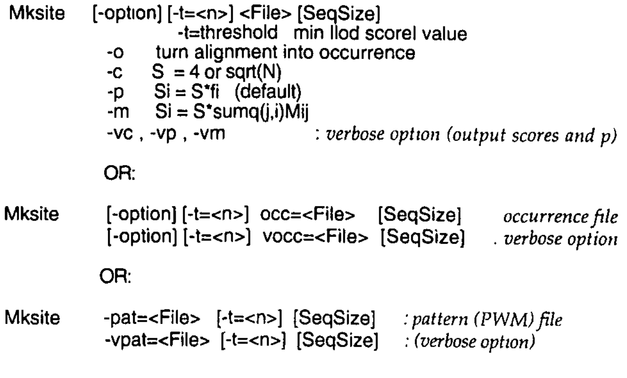
Results

The expected best-score distribution tends to the Gumbel distribution

Figure 4 shows $\frac{dP}{dS}$ for four PWM representing four different transcription-related signals. Those signals are defined on different widths: $w = 8$ for the cap site (Bucher, 1990), $w = 12$ for the CCAAT box (Bucher, 1990), $w = 14$ for the GC box (Bucher, 1990) and $w = 30$ for the Hox1.3 binding motif [Odenwald *et al.*, 1989; accession number T00377 in the TRANSFAC database (Wingender *et al.*, 1996)]. A reduced (Z) score is used to allow a direct comparison of the different PWM. As the width increases, the probability density becomes more regular and resembles the Gumbel (1958) distribution:

$$G(Z) = \frac{1}{\beta} e^{-(Z-\alpha)/\beta} \exp\left[-e^{-(Z-\alpha)/\beta}\right] \tag{18}$$

with $\alpha \cong -0.45$ and $\beta \cong 0.78$, for $Z_{\text{mean}} = 0$ and $\sigma = 1$.



```
Mksite   [-option] [-t=<n>] <File> [SeqSize]
                  -t=threshold   min llod scorel value
           -o    turn alignment into occurrence
           -c    S  = 4 or sqrt(N)
           -p    Si = S*fi   (default)
           -m    Si = S*sumq(j,i)Mij
           -vc , -vp , -vm              : verbose option (output scores and p)

           OR:

Mksite     [-option] [-t=<n>]  occ=<File>   [SeqSize]     occurrence file
           [-option] [-t=<n>]  vocc=<File>  [SeqSize]    . verbose option

          OR:

Mksite     -pat=<File>   [-t=<n>]  [SeqSize]    : pattern (PWM) file
           -vpat=<File>  [-t=<n>]  [SeqSize]    : (verbose option)
```

Fig. 3. Command line syntax of the NMksite program A first set of options {-c, -p, -m} governs the pseudo-count mode, and is only relevant when working with a block alignment or an occurrence table as an input. File names corresponding to an occurrence table must be indicated by the 'occ=' prefix. NMksite can also take an already computed PWM as input, in which case the corresponding file names must be indicated by '-pat=' (for 'pattern'). A verbose option is available in all cases and produces the expected random distribution of best scores [equation (17)]. The weights constituting the PWM can be filtered using a threshold option (thus, negative weights or weights with small absolute values can be ignored). In all cases, the final argument is the sequence target size. By default, the input file is expected to contain a block alignment, the pseudo-count mode is 'constant', and the size of the target sequence is 1000. The format of the PWM generated by NMksite (not using a verbose option) is shown in (C) and is directly usable by the Dbsite program.

Thus, at the limit of large w, the Gumbel distribution (also called the 'extreme value distribution') may govern all types of ungapped pairwise alignments: the maximal segment pairs, as found by BLAST (Altschul *et al.*, 1990; Karlin and Altschul, 1990), the maximal PWM matching scores for protein (Claverie, 1994) and nucleotide sequences and, as a special case of PWM (see below), the best score for matching fixed-window segments. The fit between the numerically computed $\frac{dP}{dS}$ and the Gumbel distribution is best in the high-score region, i.e. the region of interest to assess the statistical significance.

Validation by comparison with an analytical tractable case

The PWM formalism can be used to represent a specific nucleotide sequence query. For this we use source (s_{ip}) values of one for the nucleotide at position p, and zero for the others (e.g. ACGT: $s_{a1} = 1$, $s_{c2} = 1$, $S_{g3} = 1$, $s_{t4} = 1$, all others $= 0$). Then, the aggregate score equals the number of identically matched nucleotides between the PWM query and the target sequence. An analytical formula can be derived to compute $P(S)$ in this a simple case. Let us consider any 20mer (e.g. $[\text{ACGT}]_5$) and a target sequence with an even nucleotide composition. By direct application of the binomial distribution, we have:

$$c(S) = \sum_{s=0}^{S} \frac{20!}{s!(20-s)!} \left(\frac{1}{4}\right)^{s} \left(\frac{3}{4}\right)^{(20-s)} \tag{19}$$

and

$$P(S) = c(S)^{L-20+1} \tag{20}$$

Figure 5 compares the above analytical computation [equation (20), with $L = 250$] and the numerical computation performed by the NMksite program for a PWM corresponding to $[\text{ACGT}]_5$. The cumulative probability (top) and the probability density (bottom) computed by both methods are identical. This simple example already illustrates the fact that the best score statistics are not intuitive. Figure 5 shows that it is very unlikely that the best match will involve < 10 identical positions. Yet, each individual random match only involves five identities on average.

Relationship between sensitivity, specificity and statistical significance

In the absence of experimental validation, computing the statistical significance is a great help in estimating the expected rate of false positives corresponding to a given PWM match at a given stringency (i.e. score). This is shown in Figure 6, using the identification of 5′ splice sites as an

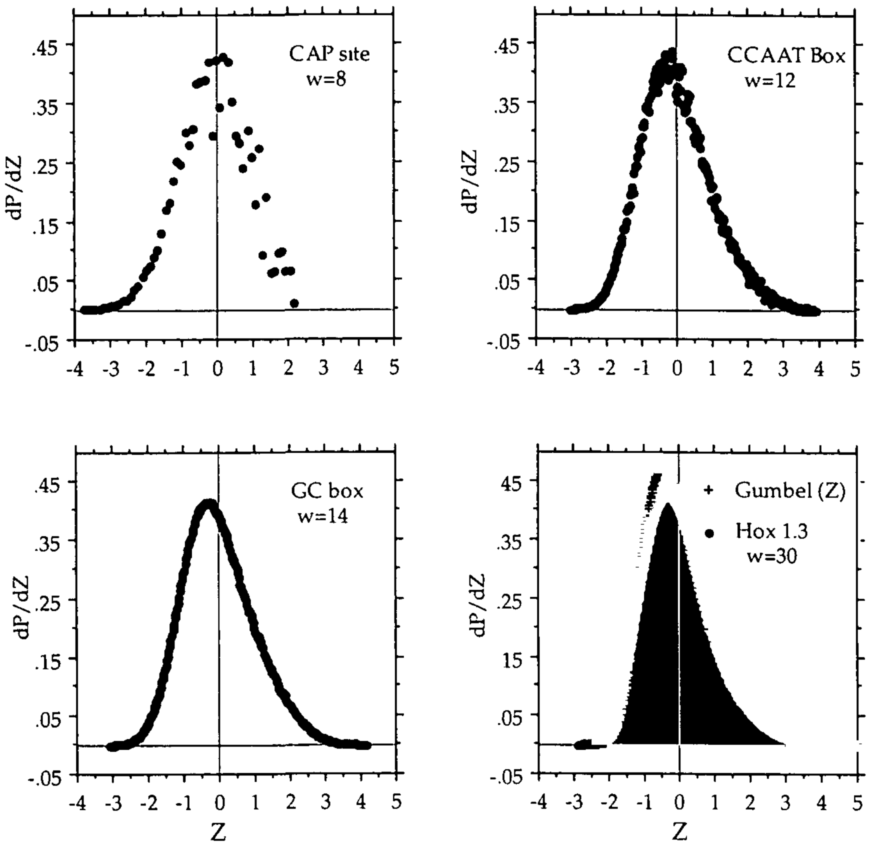


Fig. 4. Expected random best-score distribution for four PWM describing four different promoter elements of width w = 8 to w = 30. All computations apply to a 250-nucleotide target sequence.

example. First, we defined a simple 5′ site PWM from the consensus pattern YYYYYYYYYN$\frac{C}{t}$AG and weighed it according to the variability at each position (Senapathy *et al.*, 1990):

	Position											
	1	2	3	4	5	6	7	8	9	10	11	12
A	0	0	0	0	0	0	0	0	0	0	10	0
C	1	1	1	1	1	1	1	1	0	2	0	0
G	0	0	0	0	0	0	0	0	0	0	0	10
T	1	1	1	1	1	1	1	1	0	1	0	0

This PWM is not a lod score matrix, but has the advantage of generating aggregate scores that can be easily related to the degree of consensus matching (for instance, a match scoring < 20 does not obey the minimal AG rule). A maximal score of 30 indicates a perfect consensus 5′ splice site, while a score of 26 would correspond to a mediocre signal (four mismatches in the consensus pyrimidine track). From a standard data set of human genes prepared by Kulp and Reese (1995), we isolated 1486 windows of 250 nucleotides each, centered on 1486 proven 5′ splice sites. Using the above PWM with the scanning program Dbsite, we recorded all matches with a score greater than or equal to a given threshold. Because we knew the location of the biologically significant sites, we could compute the sensitivity (fraction of actual splice sites located) and the specificity (1–fraction of false identification) of the PWM detection. This is shown in Figure 6 (top). As the specificity increases for more stringent PWM matches, the sensitivity decreases in the same proportion, making the decision about an optimal score threshold quite arbitrary. For instance, the score ($S = 23$) needed to achieve 100% sensitivity (i.e. locating all the actual sites) corresponds to a specificity of 10% (i.e. nine out of 10 identified sites are not real). On the other hand, the highest stringency (maximal score = 30) only achieved a specificity of 70% at the price of a low 20% sensitivity (i.e. only one out of five sites are detected). The best compromise is found for scores equal to or greater than 29, for which approximately one out of two of real sites are detected, and one out of two detected sites is not real.

A plot of the statistical significance (computed by NMksite from the above 5′ splice site PWM for a 250-nucleotide target sequence) explains our specificity problem: no score but the highest achievable one ($S = 30$) is in fact statistically significant ($p = 1.5\%$). There is a perfect agreement between the theoretical statistical significance and the range of practically useful scores for this PWM. In the specific context of mRNA splicing, this indicates that the window of 250

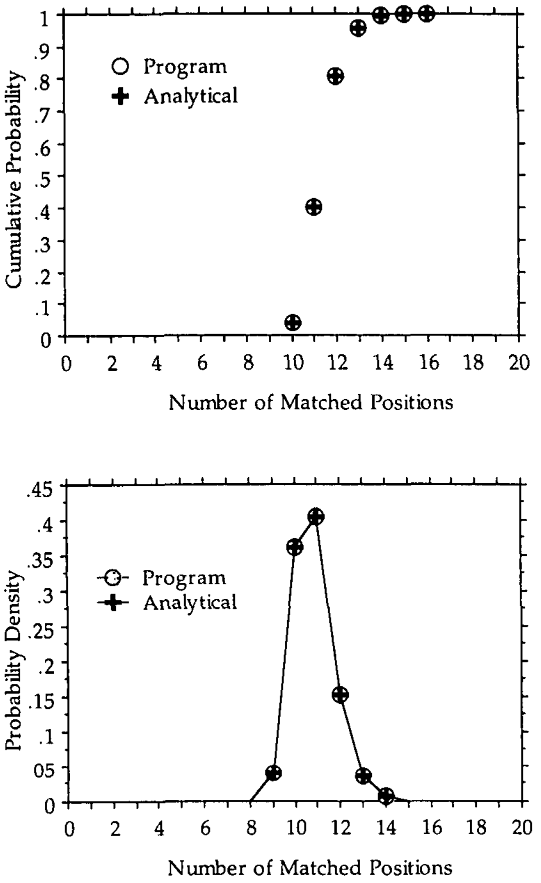


Fig. 5. Matching a 20-nucleotide query to a 250-nucleotide random target sequence. Analytical and numerical computations of the probability of the best match involving less than N identical positions.

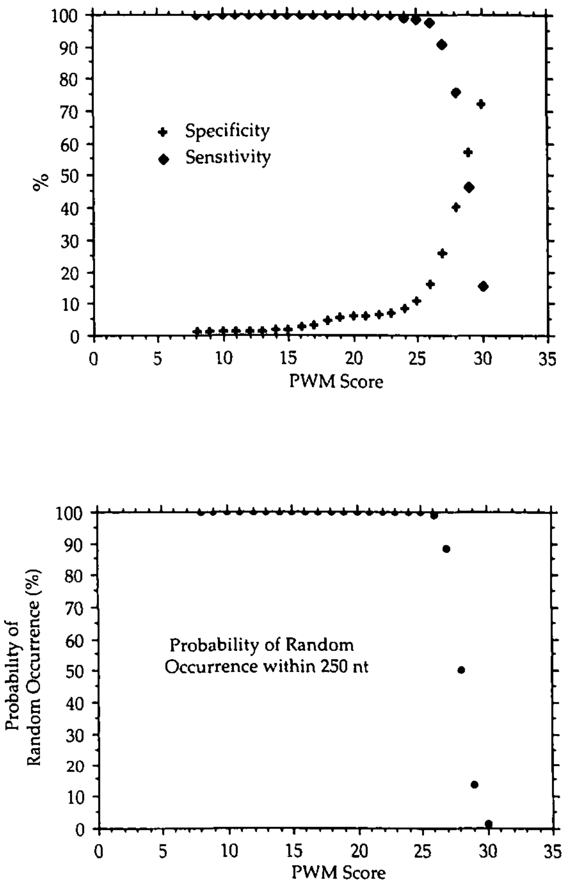


Fig. 6. 5′ mRNA splice site detection: specificity, sensitivity and statistical significance.

nucleotides centered on an actual splice signal does not exhibit a peculiar statistical signature: the frequency of 'random' signal (as defined by our PWM) is not diminished in the vicinity of the actual site. The specificity of the detection of splices is inherently limited by their low level of statistical significance, and even sophisticated methods cannot overcome this fundamental handicap (Brunak *et al.*, 1991).

Being able to assess the *a priori* statistical significance of matching any PWM at any stringency is especially useful when experimental validations are hard to obtain. This is the case when a limited set of proven examples are available, and/or when the discrimination between 'false positive' (i.e. this site is never active) and 'false negative' (i.e. the biological context in which this signal is active has not yet been tested) is difficult. This problem is encountered in the detection of eukaryotic promoter elements.

Random distribution of PWM scores: influence of nucleotide composition

PWM describing a given signal are classically optimized on the basis of a fixed training set of relevant sequences (Bucher, 1990). Optimal cut-off values (minimal aggregate score thresholds) are fixed so as to give the best compromise between sensitivity and specificity in the context of this specific set of sequences. When the PWM is subsequently used to analyze new anonymous sequences, the results are judged on the basis of cut-off values determined once and for all.

The numerical computation of the statistical significance of PWM scores explicitly involves the nucleotide composition of the target sequence within which the signal is searched. This offers a way to examine, for any PWM, the consequence of applying a pre-determined cut-off value to the analysis of target sequences of different compositions. Figure 7 shows the results of such a computation. The expected best score distributions of the TATA box and CCAAT box PWM (Bucher, 1990) were computed for three different nucleotide compositions: $(A+T) = 70\%$ $(A+T) = 50\%$ and $(A+T) = 30\%$. The results are dramatic for the TATA box (Figure 7, top). The cut-off value recommended by Bucher is -8.16. From the top plot, we can now predict that such a score threshold: (i) will be adequate for the analysis of an A + T-poor target sequence; (ii) will gather a sizeable

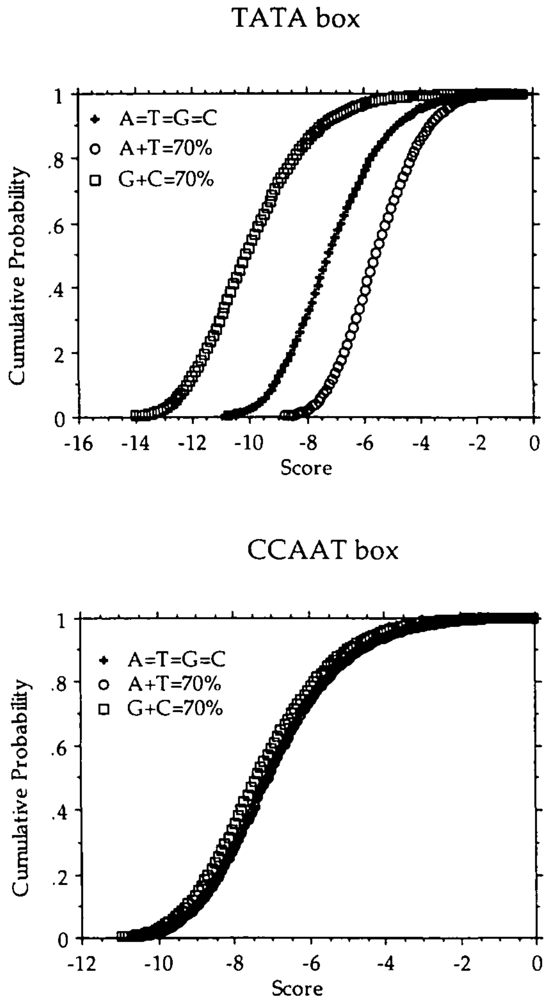


Fig. 7. Effect of the target sequence nucleotide composition on the statistical significance of PWM matches within a random 250-nucleotide target sequence. **Top**: Bucher's TATA box matrix. **Bottom**: Bucher's CCAAT box matrix. The influence is large for the TATA box and negligible for the CCAAT box.

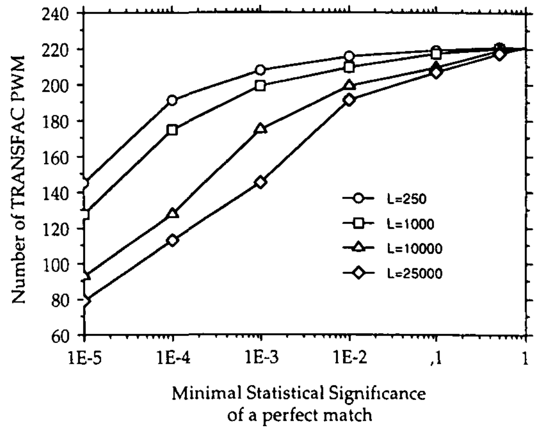


Fig. 8. Number of PWM in the TRANSFAC database for which the maximal score corresponds to a given level of statistical significance within a random segment of 250 ($L = 250$) to 25 000 ($L = 25\,000$) nucleotides.

fraction of false positives for $A + T = 50\%$; (iii) will be totally inadequate for $A + T = 70\%$. In contrast, the detection of the CCAAT box signal (Figure 7, bottom) is predicted to be quite insensitive to changes in nucleotide composition. However, the significance plot predicts that Bucher's recommended cut-off value of −4.5 will gather many false positives in all cases.

Discussion

In a previous article (Claverie, 1994), we described the Mksite program, a tool to generate lod score PWM from protein block alignments. One of the useful features of Mksite was the computation of the score threshold corresponding to various levels of statistical significance. Here, we have extended this work to the analysis of nucleic acid sequences. Computing the statistical significance of nucleotide PWM matches provides a rigorous framework to understand better why most nucleotide signals are hard to detect. The high scores associated with many nucleotide signals, such as splice sites (Figure 6) or promoter elements (Figure 7), have non-negligible probabilities of occurring by chance in a random sequence. The necessity of combining the detection of multiple elements was recognized early (Claverie and Sauvaget, 1985), as well as the necessity of confining the search within a small window (Bucher, 1990). Those are the two ways by which we can increase the statistical significance of our results. Detecting an unusual concentration of several transcription elements within a small sequence segment (e.g. 250 nucleotides) is still the principle behind current promoter identification programs [Prestridge, (1995) and references therein]. Here, 'unusual' is simply another word for 'statistically significant'. However, the current prediction programs rely on obscure combinations of cut-off values for both PWM scores and consensus string matches, and are still plagued by a high rate of false-positive identification. The explicit use of the probability of random occurrence (statistical significance) associated with the individual scores is both a simple and rigorous way to combine the information from multiple PWM matches. Given the statistical significance $p_t(x)$ of a match with PWM$_t$ at position x in the target sequence, we can locate candidate promoter regions by identifying 250-nucleotide windows $[x_{\min}, x_{\max}]$ obeying the constraints:

$$\prod_t p(x)_t \leq t \qquad \text{with } x_{\min} \leq x \leq x_{\max} \tag{21}$$

where t is a threshold corresponding to the lowest specificity and sensitivity we can tolerate. To assess the feasibility of this approach, we have applied NMksite to the entire TRANSFAC database (Wingender *et al.*, 1996) and computed the statistical significance of the maximal scores for each of the 221 PWM (Figure 8) in the matrix.dat collection. With a few exceptions, most PWM could indeed contribute to the delineation of a promoter region using equation (21) (i.e. most exhibit a finite range of scores with $p_t < 1$). More work has yet to be done to relate the levels of statistical significance to the sensitivity and the specificity of promoter detection. A file of the 221 processed TRANSFAC PWM, and the source code of the various programs mentioned in this article, are available at: http://igs-server.cnrs-mrs.fr

References

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

Berg,O.G. and von Hippel,P.H. (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, **193**, 723–750.

Brunak,S., Engelbrecht,J. and Knudsen,S. (1991) Prediction of human mRNA donor and acceptor sites from the DAN sequence. *J. Mol. Biol.*, **220**, 49–65.

Bucher,P. (1990) Weight matrix description of four eukaryotic RNA polymerase II promoter elements derived from 5023 unrelated promoter sequences. *J. Mol. Biol.*, **212**, 563–578.

Claverie,J.M. (1994) Some statistical properties of position/weight matrix scoring systems. *Comput. Chem.*, **18**, 287–293.

Claverie,J.M. (1996) Exon detection by similarity searches. *Methods Mol. Biol.*, **68**.

Claverie,J.M. and Sauvaget,I. (1985) Assessing the biological significance of primary structure consensus patterns using sequence databanks. I. Heat-shock and glucocorticoid control elements in eukaryotic promoters. *Comput. Applic. Biosci.*, **1**, 95–104.

Fickett,J.W. (1996a) The gene identification problem: an overview for developers. *Comput. Chem.*, **20**, 103–118.

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

Gojobori,T., Li,W.-H. and Graur,D. (1982) Patterns of nucleotide substitution in pseudogenes and functional genes. *J. Mol. Evol.*, **18**, 360–369.

Gribskov,M., Lüthy,R. and Eisenberg,D. (1990) Profile analysis. *Methods Enzymol.*, **183**, 146–159.

Gumbel,E.J. (1958) *Statistics of Extremes*. Columbia University Press, New York.

Karlin,S. and Altschul,S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci.* USA, **87**, 2264–2268.

Kulp,D. and Reese,M. (1995) Standard data set for the prediction of genes from human DNA sequences. Available by ftp from: www-hgc.lbl.gov, in directory: /pub/genesets.

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

Li,W.H., Wu,C.-I. and Luo,C.-C. (1984) Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. *J. Mol. Evol.*, **21**, 58–71.

Odenwald,W.F., Garbern,J., Arnheiter,H., Tournier-Lasserve,E. and Lazzarini,R.A. (1989) The HOX-1.3 homeo box protein is a sequence-specific DNA-binding phosphoprotein. *Genes Dev.*, **3**, 158–172.

Prestridge,D.S. (1995) Predicting Pol II promoter sequences using transcription factor binding sites. *J. Mol. Biol.*, **249**, 923–932.

Quandt,K., Frech,K., Karas,H., Wingender,E. and Werner,T. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.*, **23**, 4878–4884.

Senapathy,P., Shapiro,M.B. and Harris,N.L. (1990) Splice junctions, branch point sites, and exons: sequence statistics identification, and applications to genome project. *Methods Enzymol.*, **183**, 252–278.

Staden,R. (1984) Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res.*, **12**, 505–519.

Staden,R. (1989) Methods for calculating the probabilities of finding patterns in sequences. *Comput. Applic. Biosci.*, **5**, 89–96.

Stormo,G.D. (1990) Consensus patterns in DNA. *Methods Enzymol.*, **183**, 211–221.

Wingender,E., Dietze,P. and Karas,H (1996) 'TRANSFAC': a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, **24**, 283–241.