2005

Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites

Naum I. Gershenzon
Ohio State University - Main Campus
Gary D. Stormo
Washington University School of Medicine in St. Louis
Illya P. Ioshikhes
Ohio State University - Main Campus

2290–2301 *Nucleic Acids Research, 2005, Vol. 33, No. 7*
doi:10.1093/nar/gki519

Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites

Naum I. Gershenzon*, Gary D. Stormo[1] and Ilya P. Ioshikhes

Department of Biomedical Informatics, The Ohio State University, 3184 Graves Hall, 333 W. 10th Avenue, Columbus, OH 43210, USA and [1]Department of Genetics, Washington University Medical School 660 S. Euclid Avenue, Box 8232 St Louis, MO 63110, USA

Received December 9, 2004; Revised February 24, 2005; Accepted March 29, 2005

ABSTRACT

Position-weight matrices (PWMs) are broadly used to locate transcription factor binding sites in DNA sequences. The majority of existing PWMs provide a low level of both sensitivity and specificity. We present a new computational algorithm, a modification of the Staden–Bucher approach, that improves the PWM. We applied the proposed technique on the PWM of the GC-box, binding site for Sp1. The comparison of old and new PWMs shows that the latter increase both sensitivity and specificity. The statistical parameters of GC-box distribution in promoter regions and in the human genome, as well as in each chromosome, are presented. The majority of commonly used PWMs are the 4-row mononucleotide matrices, although 16-row dinucleotide matrices are known to be more informative. The algorithm efficiently determines the 16-row matrices and preliminary results show that such matrices provide better results than 4-row matrices.

INTRODUCTION

Regulation of gene expression involves the participation of many regulatory transcription factors (TFs). Understanding a regulatory system requires detailed knowledge of both the *trans*-acting TFs and the respective promoter *cis*-elements (binding sites). TFs bind to specific DNA sites among a vast excess of structurally similar non-specific sites. These specific sites share common features, consensus base pairs that almost always appear at the same position in every site (1). The consensus pattern can help to identify unknown sites (2). However, DNA–protein binding sites (signals) are often highly degenerate, so it is impractical to use just a consensus to evaluate the presence or absence of a signal in a DNA sequence (3). The position-weight matrix (PWM) technique has been developed to find a signal (4–14). The advantages of using a PWM in the search for transcription factor binding sites (TFBS) compared with motif consensus have been demonstrated (3,4).

PWMs have been used for the prediction of the binding affinity for numerous bacterial (15) and eukaryotic TFs (16–19). Currently PWMs are routinely used, e.g. in the TRANSFAC database (http://www.gene-regulation.com/pub/databases.html#transfac) and in some of the accompanying software (http://www.gene-regulation.com/pub/programs.html). Despite the obvious advantages of PWMs, the majority of existing PWMs provide a low level of both sensitivity and specificity (17).

There are several approaches to building PWMs. The most widely used methods are similar to the one proposed by Staden (7), and are as follows. One takes a collection of aligned TFBS and builds a base frequency table, i.e. counts the number of times each base occurs at each position. The base frequency table has four rows (one row for each letter: A, C, G and T) and the number of columns are equal to the motif length. The weight matrix has the same number of rows and columns with the value at each position being the natural logarithm of the value from the frequency table divided by the number of sequences in the original collection, i.e. the weight matrix contains the estimates of the log-probabilities of each base occurring at each position in true binding sites, based on the sample of known sites. A score for a particular sequence is the sum of the weights that correspond to the sequence which, under some simplifying assumptions, should be equal to the log-probability of seeing that sequence given that it is a binding site. A common alternative is to use log-odds weights which are the logarithm of the ratio of the probability of observing the sequence among a collection of sites compared to observing the sequence in the genome as a whole (3). The resulting matrix allows searching of DNA sequences in order

*To whom correspondence should be addressed. Tel: +1 614 688 3236; Fax: +1 614 688 6600; Email: gershenzon-1@medctr.osu.edu

to find sites similar to the original set of known sites, typically using a cutoff value, or score threshold, to predict new binding sites.

The basic methods do not show how to find an optimal cutoff value or how to pick the representative set of sites with which to build PWM; this can be important, since it is not uncommon to have false positives among the 'known' sites (3). Bucher presented a method to optimize the cutoff value for a given PWM (20). He applied his method to build PWMs for the most widely used promoter elements such as TATA box, Initiator, and GC-box and these matrices are still in use (see TRANSFAC). Tsunoda and Takagi (21) further refined Bucher's method and calculated the optimal cutoff values for the 205 vertebrate TFs from the TRANSFAC database. Here we describe further improvements to this method.

Besides the biological significance of recognized signals, the statistical significance of PWM matches (22,23) should also be verified. In the absence of experimental validation, the statistical significance allows estimating the expected rate of false-positives. The relationship between sensitivity, specificity and statistical significance of PWM matches were studied by Claverie and Audic (23). Based on several examples, they demonstrated the importance of calculating the statistical significance and showed how to estimate the rate of false-positive matches for a particular PWM. Hence, the statistical significance of a PWM with a given cutoff value is a necessary parameter of the PWM.

The main principles of our new technique

The new computational algorithm is a further development of the Staden–Bucher approach in determining a PWM (7,20) that improves the quality of the existing PWM (or in the worst case leaves it unmodified) based on the information in the promoter database. The promoter database, a set of promoter sequences aligned by transcription start site (TSS), is utilized in addition to the experimentally determined sites as a reservoir of sequences expected to be enriched in the binding sites being modeled. We believe that evolution has preserved the sites necessary for promoter regulation and, therefore, their occurrence frequencies in the promoter area are far from random. Since there are always a limited number of sites known for sure (from experiment) as functional, and since the true functional sites are usually highly degenerate, the only way to get more functional sites (without extensive experimentation) is to find them in the promoter sequences. Indeed, the spatial distributions of the sites in the proximal promoter area are very different from the non-promoter area or from randomly generated sequences (24,25).

Based on the preliminary knowledge about the TFBS being modeled, such as a list of experimentally verified sites, an existing PWM, or even just a consensus, the set of putative sites is extracted from the promoter sequences and used to build a new matrix. Some of the extracted sites are potentially non-functional (false positives), and we may miss some functional ones (true positives). But we expect that these putative sites are enriched in functional sites and by an iterative procedure of finding optimal matrices we will converge on an improved PWM for the sites of interest. In some respects this procedure is similar to the Gibbs sampling algorithm for discovering a common motif in a set of functionally related sequences (11), but by starting with a good estimate of the pattern we expect it will converge to the global optimum more frequently. We also optimize a different objective function—the correlation coefficient (CC) (26) that takes into account both the sensitivity and the specificity of the PWM. We assess the improvement on independent datasets.

The majority of practically used PWMs are the 4-row mononucleotide matrices based on the 'additivity hypothesis', which considers the contributions from each position of the binding site as independent and additive (1). Some experimental evidence (27,28) and theoretical considerations (29) show that a dinucleotide approach (counting of dependence between adjacent nucleotides of TFBS) could be, in some cases, the more appropriate approximation. Recently developed approaches of modeling dependencies in protein–DNA binding sites (30–32) also conclude that counting of interdependence between nucleotides (not necessarily adjacent) gives better results than regular PWM in many cases. Using the same methodology, we built the 16-row dinucleotide matrices. We demonstrate the applicability and advantage of our algorithm by building new 4-row and 16-row matrices for the GC-box element, TFBS for Sp1.

DATA AND ALGORITHM

Data

To build a new PWM for a particular TFBS we need (i) an existing motif consensus or a PWM defined, (ii) a database expected to be enriched in the TFBS of interest and (iii) a control set of experimentally defined binding sites. As an initial (original) matrix we use the PWM for GC-box TFBS obtained by Bucher (20). As a control set of sites, we use 122 non-redundant experimentally defined Sp1 binding sites from human genes from the TRANSFAC database (see list of these sites in Table 1 of Supplementary Material 1). For our database enriched in sites we use the Eukaryotic Promoter Database (EPD) release 75 (33) (http://www.epd.isb-sib.ch/), which contains a total of 1871 non-redundant, experimentally verified, human promoter sequences, each 600 bp long (−499 to +100 bp around the TSS). We evaluate the performance of the new PWM using the Database of Transcriptional Start Sites (DBTSS) (34) (http://dbtss.hgc.jp/index.html) that contain 8793 promoter sequences. We also search the entire human genome (http://genome.ucsc.edu/).

Definitions

The mononucleotide matrix is a table with 4 rows (one row for each letter: A, C, G and T) and a number of columns of motif length l. The dinucleotide matrix is a table with 16 rows (one row for each dinucleotide) and a number of columns of motif length minus one. The formula for weight at the ith position of the motif for 4-row matrices was taken from article (20) with slight changes:

$$w_{bi} = \ln\left(\frac{n_{bi}}{e_{bi}} + s_i\right) + c_i, \qquad \mathbf{1}$$

where b is one of 4 nt, n_{bi} is the number of times base b occurs at the ith position of the motif, c_i is a constant providing column maximum value to be zero, s_i is a 'smoothing' parameter

preventing the logarithm of zero (or too small a value), and $e_{bi} = (\sum_{i=1}^{L} n_{bi})/L$ is the expected frequency of base b at position i, and L is the length of the sequence (in our case $L = 600$, the length of promoter area). The choice of s_i is important, since when n_{bi} equals zero, the value of ln (s_i) defines the relative importance of different nucleotides. Different approaches were proposed to define the s_i value [see references and discussion in (23)]. Here, we use the following value for parameter s_i: $s_i = 0$ if the first term under logarithm in Equation 1 is larger than $0.01 \times n/(4 \times e_{bi})$ and $s_i = 0.01 \times n/(4 \times e_{bi})$ otherwise, where $n = \sum_{b=1}^{4} n_b$. To calculate the weight score (S) for a specific sequence we use the formula (20):

$$S = \left(\sum_{i=1}^{l} \min_b(w_{bi}) - \sum_{i=1}^{l} w_{bi} \right) \Big/ \sum_{i=1}^{l} \min(w_{bi}) \qquad \mathbf{2}$$

Equation 1 will also be used for finding the weights for the 16-row matrices. In this case b is one of 16 dinucleotides, n_{bi} is the number of times dinucleotide b occurs at the ith position of the motif, and e_{bi} is the expected frequency of dinucleotide b at position i; c_i and s_i have the same meaning; $s_i = 0$ if the first term under logarithm in Equation 1 is larger than $0.01 \times n/(16 \times e_{bi})$ and $s_i = 0.01 \times n/(16 \times e_{bi})$ otherwise, $n = \sum_{b=1}^{16} n_b$. To calculate the weight score for the 16-row matrices we will use Equation 2 with $l - 1$ instead of l.

To consider averaged positional distribution of elements along aligned promoter sequences we use the occurence frequency (OF) of the element $OF_i = n_i/N_s$, where n_i is the number of promoters containing considered element centered at position i and N_s is the number of sequences. We will use the term 'functional window' to designate the positions relative to TSS, where the occurrence frequency of considered element (OF_{real}) is much larger than expected, namely where $OF_{real} - OF_{random} \geqslant 3 \times SD$; here OF_{random} is the occurrence frequency of the respective elements in the randomly generated DNA sequences with the same proportion of 4 nt as in the training set of promoters and SD is the standard deviation. Therefore, we suppose that sites appearing in that window are likely to have a functional (biological) meaning and some of these putative sites could be used to build a new matrix. Note that we will need the functional window only to define approximately the initial range of positions where we will extract the respective sites; the final optimal matrices will not be affected by the parameters of the functional window.

Several different approaches were applied to optimize the cutoff value and motif length. Bucher used a local over-representation parameter [see equation 5 in his article (20)]. We found that the CC is the most appropriate parameter for optimization:

$$CC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}, \qquad \mathbf{3}$$

where TP, FP, TN and FN are the number of true positives, false positives, true negatives and false negatives, respectively. The definition of these parameters is not straightforward and depends on the specific task. We define them in the following way (see also Figure 1 for clarification): TP is the number of sites positively identified by the new matrix at expected positions (in the given window) among the sites identified there by the original matrix; FP is the difference between the number (N_{new}) of sites positively identified by the new matrix at expected positions at all considered promoter sequences in the same window and TP; FN is the difference between the number (N_{orig}) of positively identified sites by the original matrix at expected positions and TP; TN is total number of all sites in the given functional window with length l_w in all considered promoter sequences (N_s) minus TP, FP and FN; so that:

$$FP = N_{new} - TP,$$

$$FN = N_{orig} - TP,$$

$$TN = l_w \times N_s - TP - FP - FN.$$

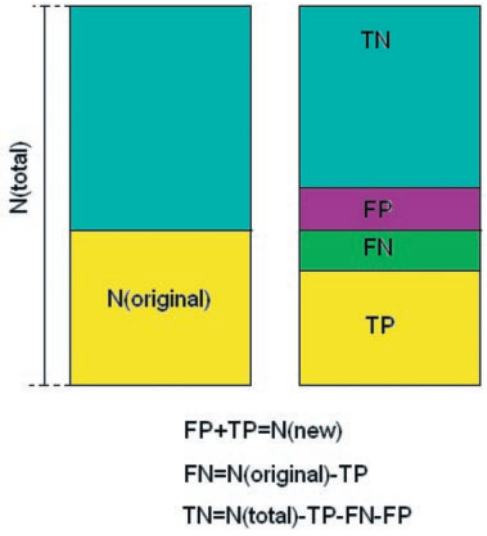

Figure 1. The schematic presentation of TP, FP, TN and FN. The number N_{orig} in the left rectangle represents the amount of sites recognized by the original matrix as respective TFBS among all considered sites $N_{total} \times l_w$ in the given window l_w in all promoter sequences N_s from the training dataset. The number TP is the amount of sites recognized by the new matrix among N_{orig} so the rest of N_{orig} is FN. The number of sites recognized by the new matrix but not included in N_{orig} is FP. Finally, the number TN is the total number of considered sites N_{total} minus TP, FP and FN.

Defining CC in this way we suppose for a moment that the initial matrix is ideal, since the initial matrix is the only information about a particular TFBS we rely on at the beginning. But the sites extracted from the promoter sequences based on the original matrix carry additional data. The new matrix built on these new sites will be different from the original. Maximizing CC parameter we try to be as close as possible to the original matrix, but gradually picking up the additional information (new putative sites) concealed in the promoter sequences. Note that parameters TP, FP, TN and FN defined here to calculate CC could not be used to evaluate the sensitivity and specificity of the final matrix.

The question of how to calculate and compare the sensitivity and specificity of the new and original matrices is crucial. We will define the sensitivity (Sn) simply as a percentage of experimentally confirmed sites recognized by the respective matrix. To compare the specificity of two matrices we will suppose that the majority of sites found by these matrices in

the randomly generated DNA sequences are false positives. If this is true, the ratio of the occurrence frequencies found by the new and original matrices is inversely proportional to the ratio of their specificities. Therefore, we will consider the averaged occurrence frequency of sites in the randomly generated sequences as a parameter describing the specificity (Sp) of the PWM.

To construct the random sequence, we calculate the average percentage of each of 4 nucleotides in the DNA sequences of interest (for example, in all sequences of the training set of promoter database or in respective chromosomes of the human genome) and then we generate a four-letter random sequence with the probability of each letter being proportional to its average percentage in those sequences.

Algorithm description

The flowchart of the algorithm is depicted in Figure 2. The input parameters are the initial PWM with a given cutoff value, a set of experimentally defined sites, and the given sensitivity (Sn_0)—the minimal portion of experimental sites recognized by the matrix. Note that the set of experimentally defined sites was not used for the building of the initial matrix and will not be included in the set of putative sites used to construct new matrices. The first step is the extraction of the dataset of putative binding site sequences for the considered element from the promoter database based on an existing PWM. To distinguish a functional window and to estimate the percentage of the noisy sites we use the randomly generated DNA sequences in addition to the sequences from the promoter database. The procedure to define a functional window includes the following steps: (i) calculate the average percentage of each of 4 nt in the sequences of the training promoter database; (ii) generate a random DNA sequence with the percentage of nucleotides defined in the previous step; (iii) calculate the averaged number and SD of respective TFBS in this randomly generated sequence using the initial matrix; (iv) calculate the positional distributions of TFBS averaged on all sequences from training promoter database. The results from the last two steps allow estimation of the noise level and accurate definition of the functional window.

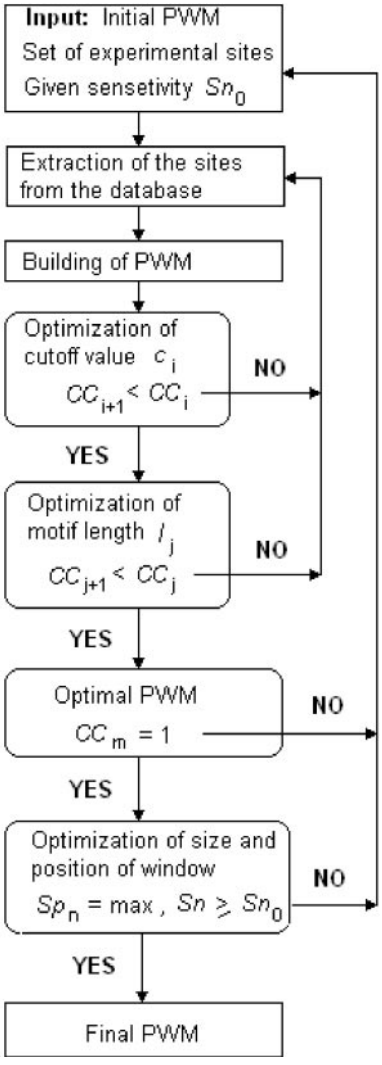

Figure 2. The flowchart of optimization process.

The next step is the alignment of extracted sequences and construction of PWM using Equation 1. The alignment is straightforward since the length of all sites is the same. The new matrix absorbs information from a larger set of sequences containing putative sites. Presumably, this dataset is more representative than the one used to build the previous version of the PWM. The goal of the following steps is a minimization of the percentage of the noisy sites (to reach the maximal specificity) and maximization of the percentage of recognized experimental sites (to reach the maximal sensitivity). There are two levels of optimization at the beginning: cutoff value and motif length (see flowchart in Figure 2). The CC (Equation 3) is used as the optimization criterion. In the given range of parameters (reasonably chosen) we apply a new matrix to the promoter database every time changing one of the parameters and recalculating CC. First, we find the optimal cutoff value for the given motif length and for the given position and size of the window where we pick up the putative motives. To do this we calculate the CC parameter for every cut-off value in the range, for example, from 0.700 to 0.950 with step 0.001. The cutoff value is considered optimal if the CC value for this cutoff is maximal. Second, we change the length of the motif and find the optimal cutoff value and respective value of CC for each length in the given range for the given window. The maximal value of CC defined the optimal length. The final result of this procedure is a PWM with optimal length l, and cutoff value c. Now the new optimal matrix is utilized as an initial matrix for the next optimization cycle, which repeats all the aforementioned steps. This refinement process is continued up to m cycles, until $CC_m = 1$ (usually it takes from 6 to 12 cycles). Each cycle brings a portion of new sites typical for this particular window and excludes some not typical sites increasing the influence of sites from that window. This influence is strongly limited by the requirement to be as close as possible to the previous matrix expressed by the definition of CC. All aforementioned steps should be repeated for each window from the functional window. As a result we will have a set of optimal matrices, one matrix for each considered window. Each matrix has its own sensitivity and specificity. Now we should choose the best matrix among all optimal matrices using the input set of experimentally defined sites and given sensitivity Sn_0.

Accordingly, we choose a matrix with sensitivity Sn $\geqslant$ Sn_0 having the lowest score of occurrence frequency in the randomly generated DNA sequence, thus Sp = max(Sp_n) (see Figure 2). This final matrix is built on the set of sites extracted from a window which is part of the functional window. We will use the term 'optimal window' to designate this window.

To implement the described algorithm and to perform the statistical analysis a set of C++ Window-based programs was created. We also used the software package, Promoter Classifier (35), available at site http://bmi.osu.edu/~ilya/promoter_classifier/.

RESULTS

GC-box presence in the promoter area

Figure 3 depicts the spatial distribution of OF of GC-box sites defined by scanning promoter sequences from EPD and DBTSS databases by the original PWM (20). From this picture, one can see that the OF values (dark blue and green curves) are larger than in randomly generated DNA sequences (horizontal line) across the entire promoter region (from −1000 to +200 bp) with an exception in the immediate downstream area. The difference between OF in promoter area and OF in random sequence exceeds 3 SD in that area. The ratio (OF in promoter area)/(OF in random sequence) is ∼10 in the proximal upstream area, indicating the essential overrepresentation of the GC-box motif there. Surprisingly, the positions and values of absolute maxima on both OF curves, for the EPD and DBTSS databases, practically coincide and the curves are virtually identical. These features point to the non-random nature of GC-box presence even more than the large OF values. Note that both strands of promoter DNA sequence are overrepresented with GC-box (see magenta and red curves at Figure 3). Most of the human promoters contain CpG islands (36). Since the GC-box motif is rich with C and G nucleotides, the presence of those islands could be an explanation of GC-box overrepresentation. However, the occurrence frequency of GC-box in CpG-less and CpG-containing subsets of promoters exhibit similar behavior (see Figure 1 in Supplementary Material 2) excluding 'CpG island' explanation. Thus, the statistical analysis indicates that proximal upstream regions of promoter sequences contain GC-box like motif. It is likely that the majority of those sites have functional (biochemical) meaning as binding sites for Sp1 and, therefore, they could be used for building the PWM for Sp1 TFBS.

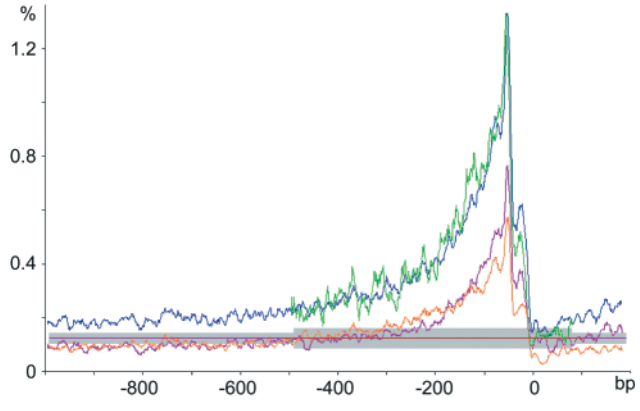

Figure 3. The occurrence frequency (the percentage of sequences having a considered motif centered at particular position) distribution of the GC-box sites found by the original matrix. The distribution is based on scanning of DBTSS (magenta, positive strand; red, negative strand; dark blue, both strands) and EPD (green, both stands) sequences. The value at each position is an 11 point sliding average. The TSS is placed at position +1. The straight horizontal line depicts the average amount of GC-box sites found in both strands of the randomly generated sequence with the same percentage of each of 4 nt as in the training set of promoter sequences, namely 20.6% for A and T, and 29.4% for C and G. The shadow rectangles indicate SD calculated based on 1871 random sequences (short rectangle) and on 8973 random sequences (long rectangle), respectively.

PWM for GC-box

We applied our algorithm to refine PWM for the GC-box element using as input the original matrix and a control set of experimentally defined Sp1 binding sites (see subsection Data under Data and Algorithm). Since one of the input parameters, given by the user, is the number (or percentage) of recognized sites from the control set of sites there are multiple optimal matrices (one for each given number). Figure 4 depicts the occurrence frequency of GC-box sites in the random sequence versus sensitivity for the original matrix (circle at the left upper corner) and two sets of new 4-row (squares) and 16-row (diamonds) matrices, respectively. Each new matrix was built based on its own set of sites extracted from the respective optimal window. Note that each matrix has its own optimal cutoff value, which is considered a part of the matrix. We already mentioned that the set of the experimental sites was not used to build the new matrices. However, we used this set on the final stage of the algorithm procedure to choose the best matrix among all optimal matrices with the same (given by the user) sensitivity. Thus, indirectly the set of experimental sites is involved in the process of matrix refinement. Since we use the same set of sites to compare the initial and new matrices we should confirm by cross-validation procedure that the different subsets of the experimental sites

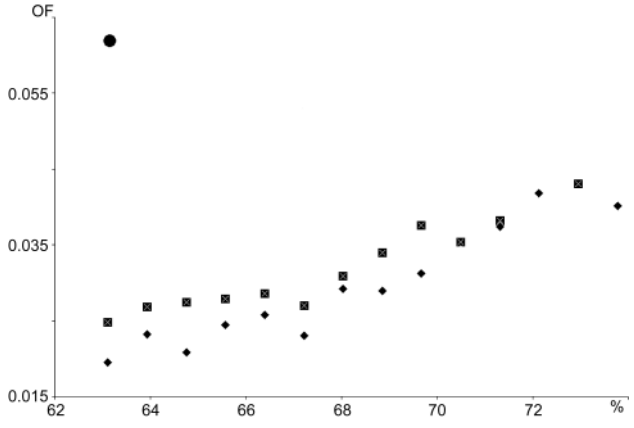

Figure 4. The sensitivity/specificity ratios for the original and new matrices. The averaged occurrence frequency of GC-box sites found by the original matrix (circle at the left upper corner) and two sets of new 4-row (squares) and 16-row (diamonds) matrices in the randomly generated sequence with the same percentage of each of four nucleotides as in the training set of promoter sequences versus sensitivity. The *x*-axis is the percentage of recognized sites from a control set of experimentally defined sites.

lead to the same optimal matrix. To implement it we divided randomly 122 experimental sites on five subsets with ~80% sites in each. We applied our algorithm five times using different subsets of experimental sites and found that indeed every time the comparison of the sensitivity and specificity of all optimal matrices leads us to the same (the best) matrix. For example, the matrix built on the set of putative sites from the window −55 to −47 bp was the matrix with maximal sensitivity among all matrices for each of the 5 subsets of experimental sites. Furthermore, the sensitivity on the remaining 20% of sites is the same, on average, indicating that the procedure does not lead to overfitting to the 80% of sites used to determine the optimal matrices.

Two obvious conclusions follow from Figure 4: (i) there are new matrices with higher levels of both sensitivity and specificity than original matrix; (ii) in all cases, the optimal 16-row matrix has higher specificity than optimal 4-row matrix with the same sensitivity. We consider in more details four new matrices: two (4-row and 16-row) with the highest sensitivity (Tables 1 and 2) and two (4-row and 16-row) with sensitivity equal to the original matrix (Tables 1 and 2 of Supplementary Material 3). The sets of putative sites extracted from the promoter sequences and used for the matrix building can be found in Supplementary Materials 4–7.

The comparison of the new and original PWMs

To compare matrices we applied them to the human genome and two types of randomly generated sequences [type 1 and type 2 are the sequences with the same percentage of each of 4 nt as in the human genome and as in the training set of promoter sequences (EPD database) respectively]. The results are presented in Table 3 and lead to the following conclusions:

(i) The specificities of the new 4-row and 16-row PWMs with the same sensitivity as the original matrix are higher than the original matrix by 10.5 and 18.6 times, respectively (compare lines 1–3 from column 5).
(ii) The specificities of the new 4-row and 16-row PWMs with maximal sensitivity are higher than the specificity of the original matrix by 5.6 and 6.5 times, respectively (compare lines 1, 4 and 5 from column 5). In this case, the sensitivities of new 4-row and 16-row PWMs are 9.9% and 10.7% larger, respectively, than the sensitivity of the original matrix (column 1).
(iii) The specificity of the 16-row matrix with the same sensitivity as the new 4-row matrix has 1.8 times higher specificity than the new 4-row matrix.
(iv) The specificity and sensitivity of the 16-row PWM with maximal sensitivity are higher then specificity and sensitivity of new 4-row PWM with maximal sensitivity.

The common features of the new 4-row matrices are: (i) the positions 5 and 6 contain only nucleotide G; (ii) there is no G at position 7; (iii) there is no C and T at positions 4 and 8; (iv) there is no C at position 10.

The consensus comparison of the original and two new 4-row matrices shows that the main nucleotides at the core positions from 3 to 12 are identical (see Table 1 here and Table 1 in Supplementary Material 3). The major difference between matrices is due to the weight of the second and third nucleotides in those positions. Indeed, the weights of A and T at position 3 in the original matrix are larger than in the new matrices. The same is for the weight of A at positions 4 and 10. Nucleotide T at positions 11 and 12 has higher weight in the original matrix. The edge positions have more differences. Thus, the main nucleotides in the original matrix are A at positions 1 and 2 and T at positions 13 and 14, in contrast to the new matrices having the main nucleotide G at positions 1 and 2, C at position 13 and G at position 14.

The typical features of GC-box motif followed from the 16-row matrices are (i) position 5 contains only dinucleotide GG; (ii) position 4 contains mainly GG and AG; (iii) position 6, 7 and 8 contains mainly 3 among 16 possible dinucleotides (GC, GA and GT at position 6; CG, AG and TG at position 7; GG, GT and AG at position 8); (iv) edge positions never contain dinucleotides AT at position 1, AC and TC at position 2, AA at position 13 and TT at position 14.

It is obvious that the 16-row matrix contains, in general, more information than the 4-row matrix. There are certain positions where some dinucleotides never occur or their occurrence frequencies are much smaller or much higher than expected from the 4-row matrix. The expected and actual numbers of all possible dinucleotides at all positions are presented at frequency table (Table 2). From this table one can see that there are small differences between predicted

Table 1. The GC-box element base frequency table and new 4-row mononucleotide PWM with maximum sensitivity (73.0%)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	51	57	33	36	0	0	47	13	0	41	20	29	26	48
C	66	55	14	0	0	0	199	0	0	0	24	174	116	31
G	125	141	186	234	270	270	0	257	258	218	216	36	77	124
T	28	17	37	0	0	0	24	0	12	11	10	31	51	66
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	−0.377	−0.387	−1.211	−1.354	−5.523	−5.523	−0.949	−2.465	−5.478	−1.152	−1.861	−1.298	−1.002	−0.429
C	−0.613	−0.916	−2.561	−5.874	−6.017	−6.017	0.0	−5.968	−5.971	−5.803	−2.173	0.0	0.0	−1.361
G	0.0	0.0	0.0	0.0	0.0	0.0	−5.762	0.0	0.0	0.0	0.0	−1.601	−0.434	0.0
T	−0.948	−1.569	−1.068	−5.352	−5.495	−5.495	−1.593	−5.445	−2.520	−2.439	−2.526	−1.203	−0.299	−0.083
	g/a	**g/a**	**G**	**G**	**G**	**G**	**C**	**G**	**G**	**G/a**	**G**	**C**	**c/t**	**g/t**

The optimal cutoff value is 0.885. The total number of 270 putative 14 bp long sites were extracted from the EPD promoter sequences at positions from −55 to −47 bp. Here and hereafter the position indicates the 5'-end of the site. The last line of the matrix table contains motif consensus. The consensus of the original matrix is (**a/t-a/g-g/t/a-G/a-G-G-C/t/a-G/a-G/t-g/a/t-g/t-c/t-t/c-t/g**) (20).

Table 2. The GC-box element dinucleotide frequency table and 16-rows dinucleotide PWM with maximal possible sensitivity (73.8%)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
AA	5	3	0	0	0	0	0	0	0	3	0	4	0	4
	7.3	4.7	3	0	0	0	1.5	0	0	2.1	1.5	1.9	3.1	
AC	7	0	0	0	0	0	0	0	0	0	5	6	2	8
	7.1	2	0	0	0	0	0	0	0	2.5	8.8	8.5	2	
AG	30	29	17	32	0	0	31	9	0	20	0	5	16	12
	18	27	19	25	0	0	30	8.5	0	22	1.8	5.6	8.1	
AT	0	4	0	0	0	0	0	0	0	0	3	0	1	4
	2.2	5.3	0	0	0	0	0	0.4	0	1	1.6	3.7	4.3	
CA	5	5	0	0	0	0	9	0	0	0	0	10	11	1
	9.5	4.6	1.3	0	0	0	6.5	0	0	0	1.8	11	14	
CC	13	3	0	0	0	0	0	0	0	0	11	44	12	8
	9.2	1.9	0	0	0	0	0	0	0	0	11	51	9.1	
CG	21	29	8	3	0	0	132	0	0	6	2	39	29	16
	23	26	8.3	0	0	0	129	0	0	0	2.2	34	36	
CT	2	13	0	0	0	0	0	0	0	0	3	29	23	3
	2.8	5.1	0	0	0	0	0	0	0	0	1.9	22	19	2.8
GA	22	8	29	0	0	31	0	0	23	5	15	5	7	9
	18	12	17	0	0	32	0	0	27	11	16	2.4	9.3	18
GC	26	5	3	0	0	141	0	0	6	16	101	18	6	22
	17	5	0	0	0	136	0	0	0	13	95	10	6	
GG	31	70	101	156	191	0	0	165	144	133	26	2	34	56
	44	66	110	159	184	0	0	167	142	119	20	7	24	
GT	3	15	0	0	0	19	0	17	1	6	19	3	5	7
	5.4	13	0	0	0	16	0	7.8	7.2	5.5	17	4.6	13	
TA	4	1	3	0	0	0	0	0	0	0	0	0	10	1
	1.4	3.4	0	0	0	0.8	0	1.2	0.5	0.7	2	6.2		
TC	4	0	0	0	0	0	0	0	0	0	5	7	8	15
	3.9	0.6	0	0	0	0	0	0	0	0.7	4.4	9.1	4	
TG	16	5	30	0	0	0	19	0	16	2	0	6	15	25
	10	8	22	0	0	0	16	0	6.6	6	0.9	6	16	
TT	2	1	0	0	0	0	0	0	1	0	1	13	12	0
	1.2	1.6	0	0	0	0	0	0	0.3	0.3	0.8	4	8.5	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
AA	−1.338	−2.236	−5.853	−6.288	−6.490	−6.321	−6.575	−6.344	−6.207	−2.877	−5.987	−1.817	−5.100	−1.725
AC	−0.931	−5.415	−5.782	−6.217	−6.419	−6.250	−6.504	−6.273	−6.137	−6.057	−2.155	−1.341	−2.184	−0.960
AG	0.0	−0.420	−1.321	−1.123	−6.943	−6.774	−1.442	−2.448	−6.661	−1.434	−6.441	−2.048	−0.629	−1.080
AT	−4.600	−1.448	−5.353	−5.788	−5.990	−5.821	−6.075	−5.844	−5.707	−5.628	−2.237	−4.856	−2.448	−1.225
CA	−1.596	−1.983	−6.111	−6.545	−6.748	−6.579	−2.483	−6.601	−6.465	−6.386	−6.245	−1.159	−0.808	−3.369
CC	−1.265	−3.117	−6.735	−7.169	−7.372	−7.203	−7.457	−7.226	−7.089	−7.010	−2.319	−0.301	−1.345	−1.913
CG	−0.363	−0.426	−2.081	−3.497	−6.950	−6.781	0.0	−6.804	−6.668	−2.644	−3.602	0.0	−0.040	−0.798
CT	−2.64	−1.163	−6.247	−6.682	−6.884	−6.715	−6.969	−6.738	−6.601	−6.522	−3.131	−0.230	−0.206	−2.406
GA	−0.112	−1.510	−0.589	−6.543	−6.746	−0.990	−6.830	−6.599	−1.176	−2.622	−1.383	−1.850	−1.258	−1.170
GC	−0.469	−2.505	−3.382	−7.068	−7.270	0.0	−7.355	−7.124	−3.044	−1.983	0.0	−1.093	−1.936	−0.799
GG	−0.428	0.0	0.0	0.0	0.0	−7.235	−7.489	0.0	0.0	0.0	−1.491	−3.425	−0.336	0.0
GT	−1.790	−0.566	−5.793	−6.228	−6.431	−1.165	−6.515	−1.299	−3.996	−2.125	−0.831	−2.046	−1.279	−1.106
TA	−0.915	−2.689	−1.957	−5.642	−5.844	−5.675	−5.929	−5.698	−5.562	−5.483	−5.342	−4.710	0.0	−2.466
TC	−1.783	−5.707	−6.074	−6.509	−6.711	−6.542	−6.796	−6.565	−6.429	−6.349	−2.447	−1.479	−1.090	−0.624
TG	−0.405	−1.955	−0.530	−6.518	−6.721	−6.552	−1.709	−6.574	−1.514	−3.514	−6.218	−1.643	−0.470	−0.122
TT	−2.208	−3.288	−5.806	−6.241	−6.443	−6.274	−6.528	−6.297	−4.009	−6.081	−3.789	−0.591	−0.416	−5.216

The optimal cutoff value is 0.863. The total number of 191 putative sites 15 bp long were extracted from the EPD promoter sequences at positions from −57 to −52 bp. The top numbers at each cell in the frequency table are the actual number of dinucleotides at perspective positions; the bottom numbers are the number of dinucleotides predicted based on the base frequency table from Table 1. The gray shadow marks the predicted numbers which are essentially different from the actual numbers.

and actual numbers for the main dinucleotides such as GG at positions 3–5, 8–11 and CG at position 7. At the same time, the dinucleotides with small frequency number could be greatly different from the predicted number. For example, the frequency of dinucleotide AT at positions 12 and 13 are 0 and 1 (see frequency Table 2, first line), although the expected value calculated base on frequency of mononucleotide A at positions taken from frequency Table 1 are 3.7 and 4.3, respectively. Another example is the frequency of GT at positions 8, 9 and 13 are 17, 1 and 5, in contrast to expected values 7.8, 7.2 and 13 (frequency Table 2, line 12).

The distribution of GC-box motif in the promoter area and in the human genome

The curves on Figure 5 depict the spatial distribution of OF for the GC-box motif defined by scanning of EPD and DBTSS sequences with the 16-row matrix with the best sensitivity (Table 2). The respective curves on Figures 3 and 5 qualitatively coincide, but the values are essentially different. The results from Figure 5 are:

(i) The occurrence frequencies on two different promoter databases (as in Figure 3) show close results: the positions

Table 3. The comparison of the original and new matrices

	1 Sensitivity %	2 OF genome	3 OF conserved	4 OF random genome	5 OF random promoter
Original	63.1	1129	974	489	1207
4-Sp^{max}	63.1	388	451	51	505
16-Sp^{max}	63.1	296	335	28	382
4-Sn^{max}	73.0	686	799	90	830
16-Sn^{max}	73.8	637	723	75	777
psOrig	68.8	716	721	494	1211
psNew	59.8	661	733	322	929

Sensitivity (column 1); the occurrence frequency of GC-box sites (the averaged number of sites per one million base pairs in both strands) in human genome (column 2), in conserved between mouse and human sequences of human genome (column 3), in the randomly generated DNA sequences with the same percentage of each of 4 nt as in human genome (column 4), and with the same percentage of each of 4 nt as in the training set of promoter sequences (column 5) obtained by the original matrix (first line), new 4-row matrix with maximal specificity (4-Sp^{max}) and sensitivity equals to original matrix (second line), 16-row matrix with maximal specificity (16-Sp^{max}) and sensitivity equals to the original matrix (third line), new 4-row matrix with maximal sensitivity (4-Sn^{max}) (4th line), 16-row matrix with maximal sensitivity (16-Sn^{max}) (5th line), original pseudo GC-box matrix (psOrig) (6th line), and new pseudo GC-box matrix (psNew) (last line). The average percentages of nucleotides in genome are 0.295 for A and T, and 0.205 for C and G.

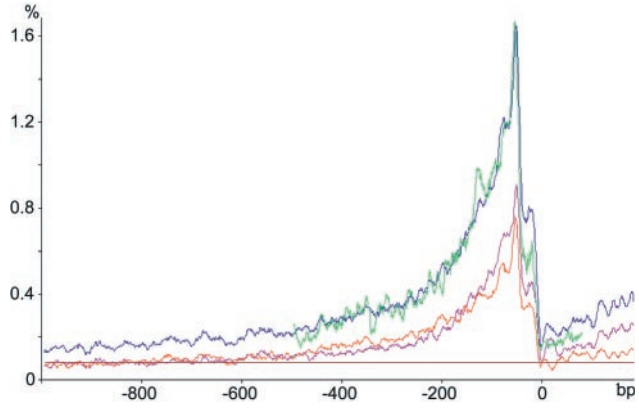

Figure 5. The occurrence frequency distribution of the GC-box sites found by the 16-row PWM with maximal sensitivity. The occurrence frequency distribution of the GC-box sites based on scanning of DBTSS (magenta, positive strand; red, negative strand; dark blue, both strands) and EPD (green, both stands) sequences. The rest is as in Figure 3.

and values of maxima of both curves (dark blue and green) practically coincide and both curves exhibit the same pattern.

(ii) The upstream (up to −1000 bp) and downstream (up to 200 bp) promoter area are overrepresented by GC-box binding sites. The ratio of real occurrence frequency to expected occurrence frequency ranges from 2 at the 5′-end of the promoters up to 20 at positions from −55 to −50 bp. Both DNA strands are overrepresented by GC-boxes (magenta and red curves).

(iii) The majority of promoter sequences (77% in the area from −500 to 100 bp and 85% in the area from −1000 to 200 bp) contain at least one GC-box on at least one strand. The average number of GC-boxes per promoter, excluding the number of potentially noisy sites, is 3.

Table 4. The occurrence frequency of GC-box sites in Human genome

	4-Sn^{max} 1 Number sites	2 Number conserved	3 Random	16-Sn^{max} 4 Number sites	5 Number conserved	6 Random
Chr 1	761	975	106	709	862	81
Chr 2	614	576	81	560	513	62
Chr 3	560	573	72	515	511	56
Chr 4	476	403	62	434	371	47
Chr 5	553	567	73	502	511	56
Chr 6	573	659	72	526	582	57
Chr 7	667	692	84	635	649	69
Chr 8	591	587	77	541	531	61
Chr 9	729	744	95	680	696	80
Chr 10	699	707	101	648	658	87
Chr 11	749	1067	100	675	945	77
Chr 12	678	1009	92	635	872	74
Chr 13	489	369	62	449	348	47
Chr 14	686	713	86	641	646	76
Chr 15	764	836	115	719	749	83
Chr 16	1004	970	158	961	911	133
Chr 17	1161	1652	181	1133	1473	151
Chr 18	557	502	77	510	467	65
Chr 19	1584	2721	255	1579	2539	226
Chr 20	934	1035	146	861	990	120
Chr 21	707	573	95	655	537	69
Chr 22	1340	1589	248	1287	1497	209
Chr X	572	680	75	509	590	57
Chr Y	513	326	81	457	285	60
Averaged	686	799	90	637	723	75

The averaged number of sites per one million base pairs (both strands) in each of 24 chromosomes of human genome (columns 1 and 4) and in conserved between mouse and human sequences (columns 2 and 5) obtained by the new 4-row (4-Sn^{max}) and 16-row (16-Sn^{max}) matrices with maximal sensitivities. The columns 'random' contain averaged OF values in the randomly generated DNA sequences with the same percentage of each of four nucleotides as in respective chromosome. The last line contains the respective values averaged on whole genome.

We scanned with the 4-row and 16-row matrices of maximal sensitivity each chromosome of the human genome in two types of sequences: (i) whole chromosome sequences and (ii) only conserved regions between human and mouse sequences. Table 4 shows the average occurrence frequencies of GC-box sites in each chromosome and whole genome. For comparison, we also present the OF values in randomly generated sequences (columns 3 and 6). We see that the average occurrence frequency of GC boxes ranges widely from 238/million base pair in Chr# 4 to 796/million base pair in Chr# 19 reflecting the different density of genes and/or promoters in chromosomes. Indeed, the occurrence frequency of sites in chromosomes is proportional to the occurrence frequency of known genes (see Figure 6). The average OFs in the randomly generated sequences also vary widely, from 25/million base pair in Chr# 13 to 108/million base pair in Chr# 19. This is because of the difference of nucleotide percentages: Chr# 13 has 30.7% of A and T, in contrast, Chr# 19 has 25.8% of A and T. The average OF value in every chromosome and in the whole Genome are 6- to 8-fold larger than in the randomly generated sequences, but are similar to the OF of random promoters sequences (Table 3), indicating the different composition of promoter regions compared to the whole chromosomes.

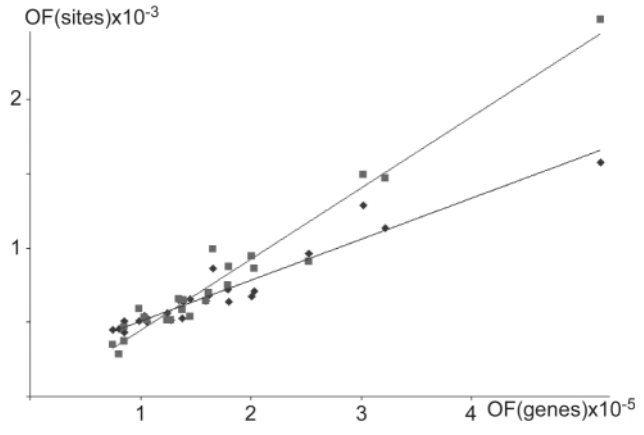

Figure 6. The occurrence frequency of GC-box sites versus occurrence frequency of known genes in chromosomes of human genome. The OF of sites were obtained by 16-row matrix with maximal sensitivity. The diamonds and squares show the averaged OF of each chromosome in whole and conserved sequences, respectively.

Comparison with experimental data

We compare our predictions with experimental results that were obtained by applying high-density oligonucleotide arrays to all nonrepetitive sequences on human chromosomes 21 and 22 to map *in vivo* binding sites, in particular, for Sp1 (37). The method used by Cawley *et al*. (37) allows defining the windows in chromosome sequence (up to a 1000 bp long) containing one or several functional binding sites. They found a total of 353 such windows in both chromosomes. We scanned those 353 sequences by our new matrices (from the Table 1 and 2) and found totally 1158 sites in 256 sequences (75% of total number of sequences) for the 4-row matrix and 861 sites in 231 (65%) sequences for the 16-row matrix. These sites that are predicted within the experimentally identified segments are only a fraction, 1–2%, of the total number of sites predicted across the complete chromosomes 21 and 22. This is due to at least two reasons. The authors used very high stringency in identifying the binding sites, thereby missing many of the true sites (37). In addition, we search the entire chromosomes and identify sites that are not available for binding to the Sp1 due to chromatin structure, although presumably many of them would bind if they were available. But we can still compare the predictions to see if the sites predicted by the new matrices are more enriched in the experimentally validated segments. In comparison with the original matrix, the new 4-row matrix with maximum sensitivity (Table 1) predicts ~40% fewer sites across both chromosomes, consistent with its higher specificity. Yet it actually finds ~6% more sites within the experimentally validated segments, so that it has ~50% higher fraction of its predicted sites within those segments than the original matrix. In this case the 16-row matrix with maximum sensitivity (Table 2) is not quite as good, but still shows ~20% higher enrichment in the validated segments than the original matrix.

Pseudo GC-box matrix

The algorithm proposed here allows the improvement of any initial matrix based on any database of sequences regardless if this initial matrix and database are 'good' (reflects some basic features of a particular TFBS) or 'bad'. To find the difference between good and bad matrices we applied our algorithm to a pseudo GC-box matrix using a fake promoter database. As an initial matrix we used Bucher GC-box matrix where column 7 replaced column 1, the columns from 1 to 6 shifted right one position and the columns 8–14 stayed as before. So the good and bad matrices have identical columns but rearranged order. As a training set of promoter sequences we used the EPD sequences in the interval of positions from 1000 to 1600 from TSS (instead of −500 to +100). We also used the set of experimental sites for the real GC-box where we altered the respective positions as for the pseudo GC-box matrix. Thus, we used a matrix with no biological sense and a training set of sequences where we do not expect an overrepresentation of sites recognized by the pseudo GC-box matrix. Applying our algorithm we obtained a new pseudo GC-box matrix. The results of scanning the human genome, random sequences of type 1 and 2, and promoter sequences are presented in Table 3 (last two lines) and Figures 1 and 2 of Supplemental Material 8, respectively. We see that the new matrix is 'improved' by some criteria: the OF values in random sequences of both types are reduced compared with the initial pseudo GC matrix (Table 3); the OF value found by the new matrix in pseudo-functional window is larger than OF value obtained by the original matrix (Figure 2 of Supplemental Material 8). However, by other criteria the new pseudo matrix is quite different from the new real matrix. First, the reduction in OF between the initial and final matrices for both real genome and random genome sequences is much less for the pseudo matrix than for the real one, indicating a much smaller increase in specificity. Second, the ratios OF(genome)/OF(random) for the new real matrices are larger than the same ratio for the new pseudo matrix (compare column 2 and 5, Table 3), indicating that in the real case a true signal has been identified that distinguishes real promoters from random sequences. And third, the final pseudo matrix loses sensitivity on the pseudo experimental sites: the initial pseudo GC-box matrix recognized 68.8% of pseudo experimental sites and the final optimal matrix with maximal sensitivity recognized only 59.8%. This last result emphasizes the fundamental difference between 'good' data and 'bad' data. The iterative procedure for obtaining a new matrix is done without regard to the experimental data. If the promoter database contains new real examples of the sites specified by the matrix they will be used to improve the matrix by increasing both its specificity, compared with random sequences, and its sensitivity on known binding sites. In the case of the pseudo matrix, for which there is no enrichment in the pseudo promoter database, the iterative procedure does increase specificity of the matrix slightly, but that is matched with a decrease in sensitivity.

DISCUSSION

A new computational algorithm of building improved PWMs is presented. As an input this technique requires an initial PWM including cutoff value (or just motif consensus) and a promoter database as a source of putative sites. New 4-row and 16-row matrices with new cutoff values are built. The main idea underlining this algorithm is that the occurrence frequency

of sites is not random in the functional window of promoter sequences (see Figures 3 and 5) allowing extraction of a set of putative sites with minimal number of random (potentially non-functional) sites.

Our algorithm is an improvement of Staden–Bucher approach (7,20). The main formal difference with Bucher's algorithm is the optimization parameter: we use correlation coefficient (see definition in Data and Algorithm) instead of over-representation parameter (20). Bucher's algorithm is searching for an area (window) inside the set of aligned promoter sequences where the ratio of the number sites inside and outside that area is maximal. It allows locating the potential functional window for the considered element. In this case the number of random sites is not minimized. In contrast, our algorithm minimized the number of random sites by looking for an optimal window inside the functional window (not a functional window itself) where the set of putative sites are as close as possible to the experimentally defined sites and the percentage of the noisy sites is minimal. The proportion of the random sites in the representative dataset is crucial. This can be especially seen for the TFBS with broad spatial distribution of functional positions such as the GC-box. To build a PWM for GC-box, Bucher used the putative sites gathered from promoter sequences at positions from −170 to −5 bp (20). In contrast, we used a much smaller range of positions, which allows us to considerably reduce the portion of the noisy sites. Thus, to build a 16-row matrix with maximal sensitivity we used 190 sites extracted from the promoter sequences at position from −57 to −52 bp. The proportion of the potentially non-functional sites in the window −57 to −52 bp (2.4%) is very much smaller than in the window −170 to −5 bp (8.9%).

Another essential difference with Bucher algorithm is an additional input, a control set of experimentally verified sites. It allows us to control sensitivity of the new matrices and create multiple optimal PWMs with different ratios of specificity/sensitivity (see Figure 4 and Table 3). Finally, we used randomly generated sequences to find the matrix with highest specificity among all matrices with equal sensitivity.

In order to compare the specificities of PWMs, we made the assumption that the majority of sites found in the randomly generated DNA sequences are false positives. How true is this assumption? First of all, it is known that the majority (if not all) of the existing PWMs find in the genome sequence many orders of magnitude more sites than any reasonable estimate of the functional site number. For example, the PWMs for the TATA box and Initiator from (20) found 42 and 272 million potential sites, respectively, in the human genome, although the estimated numbers are 3000–5000 for the TATA box and 15 000 for Initiator [this estimate is made based on the statistics of core promoter elements from (38). One of the intrinsic reasons of 'bad' specificity of PWM is the degeneracy of functional sites. In particular, it means that even if we magically find all real functional sites and build a PWM based on this ideal set of sites, the resulting matrix will still find many more false positive sites in genome. So there is a hope that, in the absence of exhaustive experimental data, the average OF in the random sequences is an appropriate parameter to estimate specificity.

As we see from the Results section, the new matrices (Tables 1 and 2) are better than the original one. First, the percentage of recognized experimental sites by the new matrices, i.e. sensitivity, is higher (Table 3). Second, the proportion of sites in the randomly generated sequences is lower, i.e. the specificity, is also higher (Table 3). Although the main goal of this article is a demonstration of the ability of the proposed algorithm to improve an existing PWM (not necessarily to create an ideal matrix), the new PWMs for the Sp1 binding sites show a reasonably high level of specificity and sensitivity. This conclusion follows from the comparison of the predicted sites and experimental data from chromosomes 21 and 22 (37). The huge difference between the occurrence frequencies of the GC-box sites in the sequences conserved between mouse and human and in the randomly generated sequence with the same composition (see Tables 3 and 4 from the main text) also indirectly supports this statement.

The analysis of the pseudo GC-box matrix demonstrated that if the database of sequences is not enriched in the sites being modeled, the iterative procedure will not return an improved matrix as assessed by the set of experimental sites which is an important component of the training procedure, yet not required to be especially large. There are at least 200 of known TFBS satisfying these requirements (21) and, therefore, their matrices could be improved by the described algorithm. Though in the present article we considered in detail only one example, we already have preliminary results showing improvement of matrices for the EGR-1 and EST-1 TFBSs. Note that there are just a few known experimental sites for the later TFBSs (see TRANSFAC database). The number of experimentally defined sites affects the outcome of the program in two ways. An initial matrix is needed to get the process started, but that could be based on only a few sequences—perhaps just a consensus sequence. The other effect comes at the end of the process when different matrices are ranked by their specificity (frequency of sites in random sequences) and their sensitivity. For that purpose one would like have a reasonable estimate of the sensitivity, but 10 to 20 sequences should be sufficient.

Studying the sensitivity-specificity tradeoff for the information theoretic PWM and determining an optimal cutoff according to some criterion is one possible approach. Recently, Djordjevic *et al*. (39) proposed an alternative solution to the problem. They formulated a maximum likelihood estimation (MLE) method that utilized the physical TF concentration-dependent binding probabilities. This MLE naturally becomes a variant of logistic regression and gives rise to a set of parameter estimation methods that interpolate between the conventional information theoretic weight matrix at one end and a Support Vector Machine at the other end. The Support Vector Machine, called QPMEME (for Quadratic Programming Method for Energy Matrix Estimation) not only provides a PWM but also a threshold. So this method defines the threshold based on intrinsic properties of sites included in the training set of experimentally defined sites, in contrast to our approach, which maximizes the portion of functional sites among a set of putative sites. These two methods do not contradict but compliment each other.

Given that the 'additivity assumption' does not hold in some cases (27,28), we can refine the methods and algorithms employed to build higher order matrices [see, for example, the weight array method (29)] or more generally counting all interdependency between adjusted and non-adjusted

nucleotides (30–32). The limitations of small experimental datasets have persuaded researchers to use less accurate, but fairly reliable 4-rows matrices (40). There is no such limitation in our case since we use a large set of putative sites extracted from the sequences of promoter database. Despite the accuracy of 'additivity assumption', the 16-row matrix, in general, is more statistically informative than the 4-row matrix, and, therefore, should a priori be more precise. Indeed our results (Figure 4 and Tables 3) confirmed this statement. The results obtained by the recently developed approaches to the problem of TFBS prediction (30–32) also confirmed that algorithms counting within-motif dependence in many cases outperform conventional 4-row PWM.

The improvement of PWM quality is beneficial for studying the variety of transcription regulation scenarios. An alternative way of understanding such scenarios is a strict molecular modeling of the processes of the *cis–trans* interactions (41–43). That approach would utilize the detailed knowledge of the TF structure, in particular those of their DNA-binding domains and of the details of their biochemical interaction with the respective *cis*-elements. An experimental study of these processes may also be used for the refinement of the existing PWMs (44). Even though such a study may be somewhat simplified by taking into account the existence of major types of the TF-binding domains (45), it still would be very laborious and could deal only with a few TFs and TFBS at a time. Our approach allows for significant improvement of PWMs in a short period of time, thereby presenting an alternative to molecular modeling.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank anonymous referees for helpful suggestions. GDS was supported by NIH grant HG00249. Funding to pay the Open Access publication charges for this article was provided by OSU.

Conflict of interest statement. None declared.

REFERENCES

1. Berg,O. and von Hippel,P. (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, **193**, 723–750.
2. Day,W.H. and McMorris,F.R. (1992) Threshold consensus methods for molecular sequences. *J. Theor. Biol.*, **159**, 481–489.
3. Stormo,G. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
4. Stormo,G., Schneider,T., Gold,L. and Ehrenfeucht,A. (1982) Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli*. *Nucleic Acids Res.*, **10**, 2997–3011.
5. Harr,R., Haggstrom,M. and Gustafsson,P. (1983) Search algorithm for pattern match analysis of nucleic acid sequences. *Nucleic Acids Res.*, **11**, 2943–2957.
6. Mulligan,M.E., Hawley,D.K., Entriken,R. and McClure,W.R. (1984) *Escherichia coli* promoter sequences predict *in vitro* RNA polymerase selectivity. *Nucleic Acids Res.*, **12**, 789–800.
7. Staden,R. (1984) Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res.*, **12**, 505–519.
8. Stormo,G.D. and Hartzell,G.,III (1989) Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl Acad. Sci. USA*, **86**, 1183–1187.
9. Goodrich,J.A., Schwartz,M.L. and McClure,W.R. (1990) Searching for and predicting the activity of sites for DNA binding proteins: compilation and analysis of the binding sites for *Escherichia coli* integration host factor (IHF). *Nucleic Acids Res.*, **18**, 4993–5000.
10. Hertz,G.Z., Hartzell,G.W.,III and Stormo,G.D. (1990) Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Appl. Biosci.*, **6**, 81–92.
11. 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.
12. Neuwald,A.F., Liu,J.S. and Lawrence,C.E. (1995) Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci.*, **4**, 1618–1632.
13. Quandt,K., Grote,K. and Werner,T. (1996) GenomeInspector: a new approach to detect correlation patterns of elements on genomic sequences. *Comput. Appl. Biosci.*, **12**, 405–413.
14. Wolfertstetter,F., Frech,K., Herrmann,G. and Werner,T. (1996) Identification of functional elements in unaligned nucleic acid sequences by a novel tuple search algorithm. *Comput. Appl. Biosci.*, **12**, 71–80.
15. Stormo,G.D. (1990) Consensus patterns in DNA. *Meth. Enzymol.*, **183**, 211–221.
16. Fickett,J.W. (1996) Finding genes by computer: the state of the art. *Trends Genet.*, **12**, 316–320.
17. Frech,K., Quandt,K. and Werner,T. (1997) Finding protein-binding sites in DNA sequences: the next generation. *Trends Biochem. Sci.*, **22**, 103–104.
18. Tronche,F., Ringeisen,F., Blumenfeld,M., Yaniv,M. and Pontoglio,M. (1997) Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome. *J. Mol. Biol.*, **266**, 231–245.
19. Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Systematic determination of genetic network architecture. *Nature Genet.*, **22**, 281–285.
20. 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.
21. Tsunoda,T. and Takagi,T. (1999) Estimating transcription factor bindability on DNA. *Bioinformatics*, **15**, 622–630.
22. Staden,R. (1989) Methods for calculating the probabilities of finding patterns in sequences. *Comput. Appl. Biosci.*, **5**, 89–96.
23. Claverie,J.M. and Audic,S. (1996) The statistical significance of nucleotide position-weight matrix matches. *Comput. Appl. Biosci.*, **12**, 431–439.
24. Prestridge,D.S. and Burks,C. (1993) The density of transcriptional elements in promoter and non-promoter sequences. *Hum. Mol. Genet.*, **2**, 1449–1453.
25. Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Systematic determination of genetic network architecture. *Nature Genet.*, **22**, 281–285.
26. Burset,M. and Guigo,R. (1996) Evaluation of gene structure prediction programs. *Genomics*, **34**, 353–367.
27. Man,T.-K. and Stormo,G. (2001) Non-independence of *Mnt* repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res.*, **15**, 2471–2478.
28. Bulyk,M., Johnson,P. and Church,G. (2002) Nucleotides of transcription factor binding sites exert inter-dependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.*, **30**, 1255–1261.
29. Zhang,M. and Marr,T. (1993) A weight array method for splicing signal analysis. *Comput. Appl. Biosci.*, **9**, 499–509.
30. Barash,Y., Elidan,G., Friedman,N. and Kaplan,T. (2003) Modeling dependencies in protein–DNA binding sites. In *Proceedings of the Seventh Annual International Conference on Computational Molecular Biology*, April 2003, Berlin, Germany, ACM Press, NY, pp. 28–37.
31. King,O.D. and Roth,F.P. (2003) A non-parametric model for transcription factor binding sites. *Nucleic Acids Res.*, **31**, e116.
32. Zhou,Q. and Liu,J.S. (2004) Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, **20**, 909–916.

33. Praz,V., Périer,R.C., Bonnard,C. and Bucher,P. (2002) The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data. *Nucleic Acids Res*., **30**, 322–324.
34. Suzuki,Y., Yamashita,R., Nakai,K. and Sugano,S. (2002) DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. *Nucleic Acids Res*., **30**, 328–331.
35. Gershenzon,N. and Ioshikhes,I. (2005) Promoter Classifier: software package for promoter database analysis. *Appl. Bioinformatics*, **4**, in press.
36. Antequera,F. and Bird,A. (1993) Number of CpG islands and genes in human and mouse. *Proc. Natl Acad. Sci*., **90**, 11995–11999.
37. Cawley,S., Bekiranov,S., Ng,H.H., Kapranov,P., Sekinger,E.A., Kampa,D., Piccolboni,A., Sementchenko,V., Cheng,J., Williams,A.J. *et al*. (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell*, **116**, 499–509.
38. Gershenzon,N.I. and Ioshikhes,I.P. (2005) Synergy of human Pol II core promoter elements revealed by statistical sequence analysis. *Bioinformatics*, **21**, 1295–1300.
39. Djordjevic,M., Sengupta,A.M. and Shraiman,B.I. (2003) A biophysical approach to transcription factor binding site discovery. *Genome Res*., **13**, 2381–2390.
40. Benos,P., Bulyk,M. and Stormo,G. (2002) Additivity in protein–DNA interactions: how good an approximation is it? *Nucleic Acids Res*., **30**, 4442–4451.
41. Locker,J., Ghosh,D., Luc,P.V. and Zheng,J. (2002) Definition and prediction of the full range of transcription factor binding sites—the hepatocyte nuclear factor 1 dimeric site. *Nucleic Acids Res*., **30**, 3809–3817.
42. Mandel-Gutfreund,Y., Baron,A. and Margalit,H. (2001) A structure-based approach for prediction of protein binding sites in gene upstream regions. *Pac. Symp. Biocomput*., 139–150.
43. Suzuki,M. and Yagi,N. (1994) DNA recognition code of transcription factors in the helix-turn-helix, probe helix, hormone receptor and zinc finger families. *Proc. Natl Acad. Sci. USA*, **91**, 12357–12361.
44. Kraus,R.J., Murray,E.E., Wiley,S.R., Zink,N.M., Loritz,K., Gelembiuk,G.W. and Mertz,J.E. (1996) Experimentally determined weight matrix definitions of the initiator and TBP binding site elements of promoters. *Nucleic Acids Res*., **24**, 1531–1539.
45. Ponomarenko,J.V., Bourne,P.E. and Shindyalov,I.N. (2002) Building an automated classification of DNA-binding protein domains. *Bioinformatics*, **2**, S192–S201.