Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Computational Biology and Chemistry

Computational Biology and Chemistry 30 (2006) 58–62

www.elsevier.com/locate/compbiolchem

Brief communication

Comparative analysis of core promoter region: Information content from mono and dinucleotide substitution matrices

D. Ashok Reddy [a], B.V.L.S. Prasad [b], Chanchal K. Mitra [a,*]

[a] *Department of Biochemistry, University of Hyderabad, Hyderabad 500046, India*
[b] *Helix Genomics Pvt. Ltd., Habsiguda, Uppal, Hyderabad 500007, India*

Received 15 August 2005; received in revised form 4 October 2005; accepted 4 October 2005

Abstract

We have studied the core promoter region in five sets of promoter sequences by calculating the average mutual information content *H* (relative entropy). We have used specially constructed substitution matrices to calculate mono and dinucleotide replacements in a given block of aligned sequences. These substitution matrices use log-odds form of scores, which are in bits of information. Here, we constructed and applied nucleotide substitution matrices for the core promoter region to calculate the information content to study the Transcription Start Site (TSS), TATA-box and downstream regions. As expected, the information content decreases with increasing block size. This clearly implies that the TSS region is likely to be 5–10 bases in size (length). We also notice that both in the case of mouse and humans, both TATA-boxes and TSS regions are likely to play important roles in proper transcriptional initiation.

Keywords: Substitution matrices; Core promoter region; Average mutual information content

1. Introduction

Promoter region is a regulatory region of the protein-coding genes and shows variation from species to species. The transcription factors (cell or tissue specific) bind to the promoter region of the DNA that subsequently causes efficient binding of RNA polymerase to initiate mRNA synthesis. Specific DNA sequence elements within the promoter region (like TATA-box, CCAAT-box, Downstream Promoter Element (DPE) and GC-box) exhibit similarities between different promoters of the same DNA as well as between various species. The core promoter region (which can extend ~35 bp upstream and which is a minimal promoter region required to start the pre-initiation complex formation) usually has TATA-box, which is conserved in most of the species (30–50% of promoters) and TSS region, which usually is not conserved. Each nucleotide in the consensus sequence motif (TATA-box, CCAAT-box and GC-box) represents the most frequently occurring nucleotide at that position and does not represent an actual sequence. Reliable identification of the core promoter region by RNA polymerase II prior to transcription initiation is mandatory for the proper initiation and regulation of mRNA synthesis (Smale and Kadonaga, 2003). We use the experimental database of protein-coding promoter sequences (Shahmuradov et al., 2003; Hershberg et al., 2001; Périer et al., 1998) to find out any similarity, consensus sequences, patterns and/or regularities present in a given species as well as between different species. For this purpose, we have used the traditional technique for sequence score computations. However, we also realize that the conventional substitution matrices, which have been developed for the complete genome as a whole may be unsuitable for this purpose. We also believe that it is possible to incorporate the pair-preferences (this is equivalent to a first order Markov dependence) into the substitution matrix. We also assume that the substitution matrices developed from the promoter regions shall lead to poor scores in other regions of the same DNA. Analysis of core promoter region, by using the concept of substitution matrix, is an important method to identify the similarity between promoter elements of different species. These substitution matrices are used to score sequence similarity (Zheng, 2005), database search (like BLAST and FASTA) and also for finding DNA binding sites in protein sequences (Ahmad and Sarai, 2005). The elements of these substitution matrices are explicitly calculated from target frequencies of aligned nucleotides and observed frequencies of the nucleotides.

[*] Corresponding author. Tel.: +91 40 23134668; fax: +91 40 23130120.
E-mail address: c_mitra@yahoo.com (C.K. Mitra).

1476-9271/$ – see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2005.10.004

The information in these matrices depends on the quantification approach like evolutionary models, structural properties and chemical properties of aligned sequences (Altschul, 1993; Nicholas et al., 2000; Panchenko and Bryant, 2002; Yu et al., 2003; Yu and Altschul, 2005). The Point Accepted Mutation (PAM; Dayhoff et al., 1978; Schwartz and Dayhoff, 1978), matrices are based on alignments of closely related sequences and by using these PAM matrices one can estimate target frequencies to any desired evolutionary distance by extrapolation. But in case of BLOcks SUbstitution Matrices (BLOSUM) (Henikoff and Henikoff, 1992), the estimation of target frequencies to avoid such extrapolation for different evolutionary distances, it uses the ungapped segments of multiple sequence alignments of protein families. All protein-coding genes have at least one or more TSS regions, which are active under different conditions. There are several attempts to study the TSS with the help of nucleotide frequencies (Majewski and Ott, 2002; Bajic et al., 2002, 2003; Aerts et al., 2004) and the DNA weight matrix methods (Bucher, 1990; Down and Hubbard, 2002) around the TSS but it is poorly understood due to the lack of proper signal in the TSS. The main focus of our study is to find the statistical behavior of the core promoter elements in different species with the help of average mutual information content, which is calculated by using neighbor-independent (4 × 4 matrices) and neighbor-dependent (16 × 16 matrices) nucleotide substitutions (Lunter and Hein, 2004; Arndt and Hwa, 2005).

2. Materials and methods

2.1. Promoter sequence sets

PlantProm DB—a plant promoter database (Shahmuradov et al., 2003), PromEC—*E. coli* promoter database (Hershberg et al., 2001) and EPD—Eukaryotic Promoter Database (Périer et al., 1998) include sequences that are annotated, non-redundant promoter sequences of RNA polymerase II with experimentally determined transcription start sites (Table 1). The EPD sequences included here are “representative sets of not closely related sequences”.

The promoter sequences obtained from the databases are already aligned sequences and can be represented as ungapped blocks with each row a different promoter sequence and each column an aligned base (for each of the five species). From these aligned sequences we extracted set of blocks (TATA-box region, TSS-region and Downstream region) or columns of different sizes (5, 11 and 15 nucleotide wide) for computational purposes. These sequences include both TATA-box containing and TATA-less promoter sequences (Table 2).

Table 1
Databases used in the present study (Taxonomic group or organism with number of sequences used)

S. No.	Database	Taxonomic group/organism	No. of sequences
1	PlantProm DB	All plants	305
2	PromEC	*E. coli*	472
3	EPD	Human	1789
4	EPD	Drosophila	1922
5	EPD	Mouse	118

Table 2
Blocks of nucleotides (positions are with respect to TSS that represents +1)

S. No.	Block size	TATA-box	TSS region	Downstream region
1	5	−30 to −26	−2 to +3	+16 to +20
2	11	−33 to −23	−5 to +6	+13 to +23
3	15	−35 to −21	−7 to +8	+11 to +25

In some of the sequences, there are no TATA-boxes. For such sequences, the table indicates the expected position.

2.2. Construction of substitution matrices and information content

Sequence comparisons are meaningful only if we have some idea of the similarity between different residues/bases. This information about the similarity of the bases must be derived in a contextual fashion. The coding regions and non-coding regions must be compared using a similarity matrix specifically designed for this purpose. For example, a substitution matrix constructed for the coding regions may perform poorly for the non-coding sequences and vice-versa. For this reason, we have constructed a set of substitution matrices and calculated average mutual information content of TSS region, TATA-box and a downstream region that are non-coding regions of the protein-coding genes. First we may see the protocol for single base substitution matrices. These will be 4 × 4 matrix and lack any preferences (this is the standard assumption made in all sequence alignments that the neighboring bases show no preferences). In other words, adjacent bases are considered independent. Next, we see the formulae (they will be very similar except the subscripts will now be pairs) for the base pairs taken together which corresponds to a nearest neighbor preference. As we are considering a pair, there will be 16 × 16 matrix. These matrices include adjacent pair-preferences explicitly.

2.2.1. Neighbor-independent substitution matrices

For each column of the block, we first count the number of matches and mismatches of each type between the first sequence and every other sequence in the block. This procedure is repeated for all columns of all blocks with the summed results are stored in a 4 × 4 matrix. For all sequences in the aligned sequences, the same procedure is followed summing these numbers with those that already in the 4 × 4 matrix. The total number of nucleotide pairs (observed frequency) in a given block is $\frac{ws(s-1)}{2}$ and the total number of nucleotides (expected frequency) in the block is ws, where s is the number of nucleotides in the given position and w is the block width. The resulting matrix (4 × 4 matrix) is used to calculate the odds-ratio between those observed frequencies $q_{i,j}$ and those expected by chance p_i. This odds-ratio ($\frac{q_{i,j}}{p_i p_j}$) is also called a likelihood ratio. Then “log-odds” is calculated (usually logarithm of base 2) from the odds-ratio and is given by $S_{i,j} = \log_2 \frac{q_{i,j}}{p_i p_j}$. Such probabilities (odds-ratios) should be multiplied or log-odds can be added to get the probability of their independent occurrence (Karlin and Altschul, 1990; Altschul, 1991).

2.2.2. Neighbor-dependent substitution matrices

With the incorporation of the pair-preferences into the substitution matrix that gives neighbor-dependent substitution matrices. These are very similar to neighbor-independent substitution matrices except the subscripts will be pairs of nucleotides in a given block. While calculating matches and mismatches the sliding window of one nucleotide along the sequence is used to count of all possible pairs in the given block. The total number of dinucleotide pairs (observed frequency) in a given block is $\frac{(w-1)s(s-1)}{2}$ and the total number of dinucleotides (expected frequency) is given by $(w-1)s$, where s is the number of sequences and w is the block width. The resulting matrix (16×16 matrix) is used to calculate the odds-ratio between those observed frequencies $q_{ij,kl}$ and those expected by chance p_{ij}. This odds-ratio $\frac{q_{ij,kl}}{p_{ij}p_{kl}}$ (likelihood ratio) is then used to calculate the "log-odds" and is given by $S_{ij,kl} = \log_2 \frac{q_{ij,kl}}{p_{ij}p_{kl}}$.

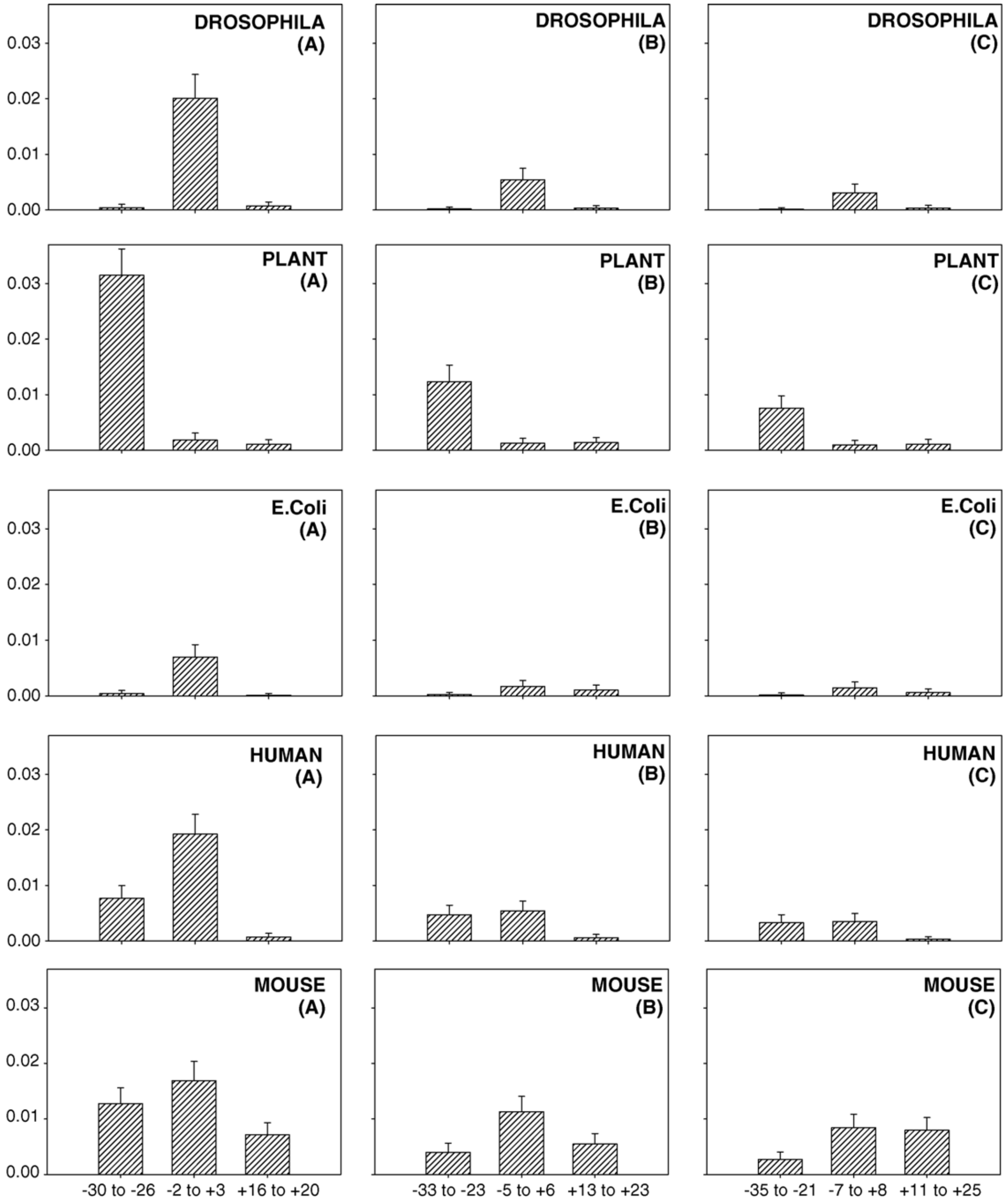

Fig. 1. The average mutual information content H, (in bits) of core promoter elements (calculated by neighbor-independent nucleotide substitutions) from different datasets. In all the figures 'A', 'B' and 'C' represents block size 5, 11 and 15, respectively. Each graph has three bars representing TATA-box region, TSS region and downstream region. The bars on top of the histograms represent the standard errors of the 16 H_{ij} values.

2.2.3. Average mutual information content (H)

The comparison of these non-coding regions can be performed either by scores in the substitution matrices themselves or by the information content of these substitution matrices. In information theoretic terms average mutual information content (H), is the relative entropy of the target and background pair frequencies and can be thought of as a measure of the average amount of information (in bits) available per nucleotide pair. In neighbor-independent substitution matrices, the log-odds of each nucleotide pair s_{ij} (in the units of $\log_2$, called bits) multiplied by the probability of occurrence of that pair q_{ij} will give the weighted score and is then summed overall for the nucleotide pairs to produce a score that represents the ability of the average nucleotide pair in the matrix to discriminate actual from chance alignments. The average mutual information content is given by $H = \sum_{ij} q_{ij} s_{ij} = \sum_{ij} q_{ij} \log_2 \frac{q_{ij}}{p_i p_j}$. The higher the value of the relative entropy of target and background distributions, the more easily they are distinguished (Altschul, 1991). The same procedure is applied for calculating the average mutual information content in the case of

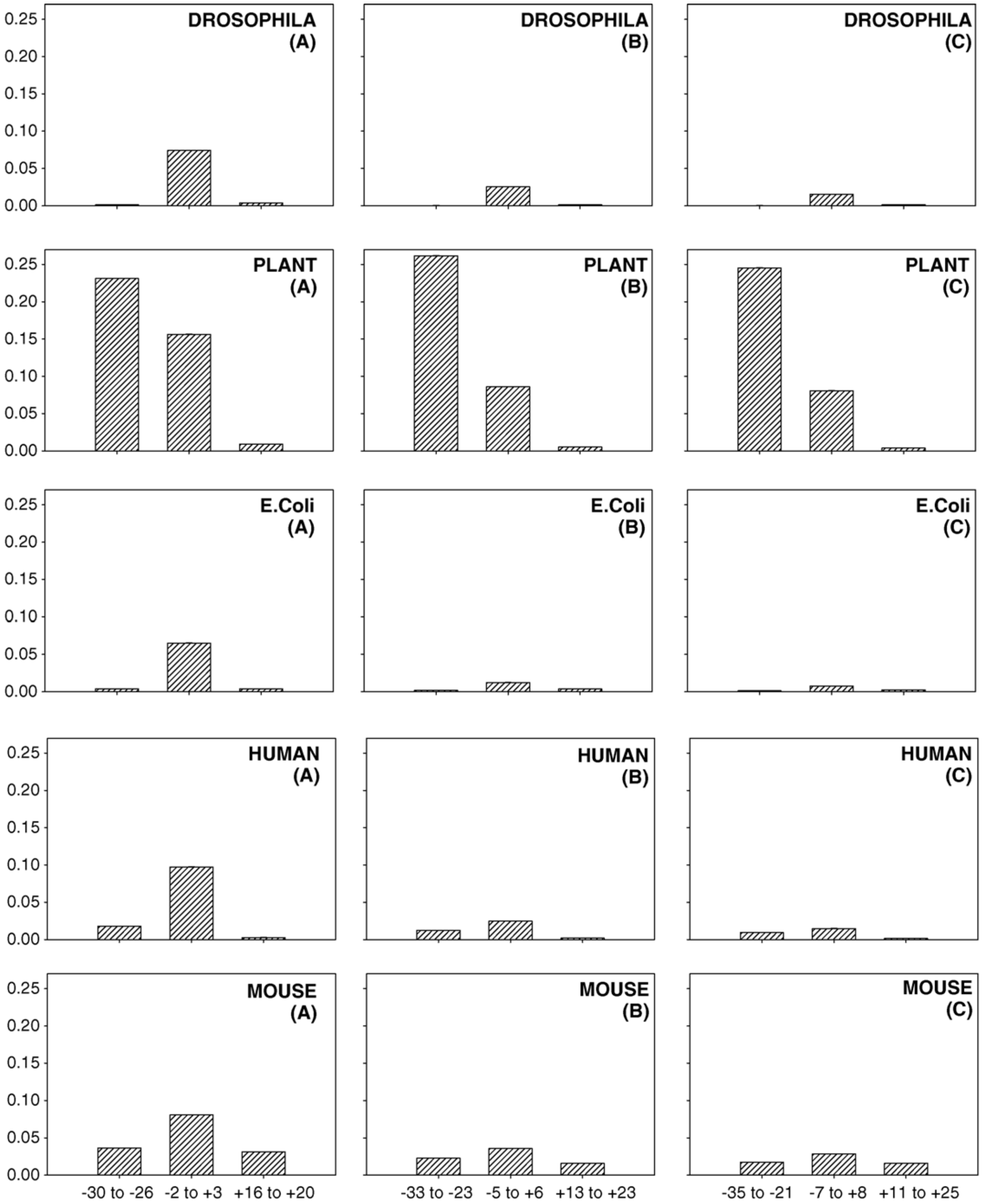

Fig. 2. The average mutual information content, H (in bits) of core promoter elements (calculated by neighbor-dependent nucleotide substitutions) from different datasets. In all the figures ‘A’, ‘B’ and ‘C’ represents block size 5, 11 and 15, respectively. Each graph has three bars representing TATA-box region, TSS region and a downstream region. The standard errors have been actually plotted but cannot be seen, as they are too small.

neighbor-dependent substitution matrices. The average information content in neighbor-dependent substitution matrices is given by $H = \sum_{ij,kl} q_{ij,kl} s_{ij,kl} = \sum_{ij,kl} q_{ij,kl} \log_2 \frac{q_{ij,kl}}{p_{ij} p_{kl}}$.

2.2.4. *Error analysis*

To further assess the reliability of our computations, we have performed a simple error analysis of the results. We consider the matrix elements H_{ij} of the information matrix $s_{ij} * q_{ij}$ as the elements of our data and compute the standard error of the 16 (or 256 in case of the pair-preferences) elements using standard techniques. The standard errors are plotted in the graph along with the histograms.

3. Results

In this study, we constructed the substitution matrices for mono and dinucleotide substitutions and calculated the information content (in bits) of core promoter elements from these substitution matrices. This information content of core promoter elements is represented in bar graphs shown in Fig. 1 (neighbor-independent) and Fig. 2 (neighbor-dependent). In each sub graph ‘A’, ‘B’ and ‘C’ represents block sizes of 5, 11 and 15, respectively.

4. Discussion

As expected, we notice that the information content decreases with increasing block size. This clearly implies that the TSS region is likely to be 5–10 bases in size. This pattern is seen in all the species (even in plants where TATA-boxes evidently play more important roles). We also notice that both in the case of mouse and humans both TATA-boxes and TSS region are likely to play important roles (probably both are involved in binding). We note that TATA-boxes and the TSS are two regions that are physically close together and we do not expect to see a case in which both are relatively less important. The error studies show clearly that the standard errors are sufficiently small that the overall conclusions are not affected. It is important to note that the different species represent different patterns of binding and it may be futile to look for any consensus sequences that are valid in all the cases. However, it may be still possible to locate some patterns in a very closely related group. In this study, we note that mice and men come close together.

References

Aerts, S., Thijs, G., Dabrowski, M., Moreau, Y., Moor, B.D., 2004. Comprehensive analysis of the base composition around the transcription start site in Metazoa. BMC Genomics 5, 34.

Ahmad, S., Sarai, A., 2005. PSSM-based prediction of DNA binding sites in proteins. BMC Bioinformatics 6, 33.

Altschul, S.F., 1991. Amino acid substitution matrices from an information theoretic perspective. J. Mol. Biol. 219, 555–565.

Altschul, S.F., 1993. A protein alignment scoring system sensitive at all evolutionary distances. J. Mol. Evol. 36, 290–300.

Arndt, P., Hwa, T., 2005. Identification and measurement of neighbor-dependent nucleotide substitution process. Bioinformatics 21, 2322–2328.

Bajic, V.B., Seah, S.H., Chong, A., Krishnan, S.P.T., Koh, J.L.T., Brusic, V., 2002. Computer model for recognition of functional transcription start sites in RNA Polymerase II promoters of vertebrates. J. Mol. Graph. 21, 323–332.

Bajic, V.B., Choudhary, V., Hock, C.K., 2003. Content analysis of the core promoter region of human genes. In Silico Biol. 4, 0011.

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.

Dayhoff, M.O., Schwartz, R.M., Orcutt, B.C., 1978. In: Dayhoff, M.O. (Ed.), Atlas of Protein Sequence and Structure, vol. 5. Natl. Biomed. Res. Found, Washington, DC, pp. 345–352.

Down, T.A., Hubbard, T.J.P., 2002. Computational detection and location of transcription start sites in mammalian genomic DNA. Genome Res. 12, 458–461.

Henikoff, S., Henikoff, J.G., 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. U.S.A. 89, 10915–10919.

Hershberg, R., Bejerano, G., Santos-Zavaleta, A., Margalit, H., 2001. PromEC: an updated database of *Escherichia coli* mRNA promoters with experimentally identified transcriptional start sites. Nucleic Acids Res. 29, 277.

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

Lunter, G., Hein, J., 2004. A nucleotide substitution model with nearest-neighbor interactions. Bioinformatics 20, i216–i223.

Majewski, J., Ott, J., 2002. Distribution and characterization of regulatory elements in the human genome. Genome Res. 12, 1827–1836.

Nicholas Jr., H.B., Deerfield II, D.W., Ropelewski, A.J., 2000. Overviw: strategies for searching sequence databases. BioTechniques 28, 1174–1191.

Panchenko, A.R., Bryant, S.H., 2002. A comparison of position-specific score matrices based on sequence and structure alignments. Protein Sci. 11, 361–370.

Périer, C.R., Junier, T., Bucher, P., 1998. The Eukaryotic promoter database EPD. Nucleic Acids Res. 26, 353–357.

Schwartz, R.M., Dayhoff, M.O., 1978. In: Dayhoff, M.O. (Ed.), Atlas of Protein Sequence and Structure, vol. 5. Natl. Biomed. Res. Found, Washington, DC, pp. 353–358.

Shahmuradov, I.A., Gammerman, A.J., Hancock, J.M., Bramley, P.M., Solovyev, V.V., 2003. PlantProm: a database of plant promoter sequences. Nucleic Acids Res. 31, 114–117.

Smale, S.T., Kadonaga, J.T., 2003. The RNA polymerase II core promoter. Ann. Rev. Biochem. 72, 449–479.

Yu, Y.-K., Altschul, S.F., 2005. The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions. Bioinformatics 21, 902–911.

Yu, Y.-K., Wootton, J.C., Altschul, S.F., 2003. The compositional adjustment of amino acid substitution matrices. Proc. Natl. Acad. Sci. U.S.A. 100, 15688–15693.

Zheng, W.-M., 2005. Relation between weight matrix and substitution matrix: motif search by similarity. Bioinformatics 21, 938–943.