

Identification of highly specific localized sequence motifs in human ribosomal protein gene promoters

Stefan Roepcke^{a,*}, Degui Zhi^b, Martin Vingron^a, Peter F. Arndt^a

^a Max Planck Institute for Molecular Genetics, Ihnestr. 73, 14195 Berlin, Germany

^b Bioinformatics Program, University of California at San Diego, La Jolla, CA 92093-0419, USA

Received 24 March 2005; received in revised form 22 July 2005; accepted 27 September 2005

Available online 15 December 2005

Abstract

For ribosomal protein (RP) genes the start of transcription is rigidly controlled to maintain the 5'-TOP signal on the messenger RNA. The responsible regulatory mechanism is not yet fully understood. Careful comparative analysis of their proximal promoter sequences reveals common characteristics and thus provides clues to the underlying mechanism.

We have extracted the proximal promoters of the 80 human cytosolic ribosomal protein genes together with the orthologous mouse sequences. After annotating the set with transcription factor binding sites based on the available literature, we searched for over-represented sequence motifs. We uncovered a novel motif that is localized at a fixed distance downstream to the transcription start. 31 out of the 80 promoters contain the motif in the same orientation around position +62 (standard deviation 6). A second evolutionary conserved and palindromic motif is found 13 times in the RP promoter set, 9 instances of which are located upstream around position -40. In addition, we see a characteristic profile of the GC-content and of the CpG dinucleotide frequencies.

Our results support a model for the transcription of ribosomal protein genes in which the maintenance of the accurate start of transcription is provided by specific transcription factors. Such a factor binds the target DNA at a fixed location relative to the TSS, and possibly interacts directly with the basal transcription machinery.

© 2005 Elsevier B.V. All rights reserved.

Keywords: Transcription start site; Transcription initiation; Motif search; Transcription factor binding sites

1. Introduction

The first step in the expression of a protein-coding gene is its transcription into messenger RNA (mRNA). This highly orchestrated process is accomplished by the transcriptional apparatus consisting of RNA Polymerase II and several transcription factors (see Reese, 2003, for a review). One usually defines the proximal promoter as a region of a few hundred base pairs around the transcription start site (TSS) and the enhancer and silencer regions, which might be located many thousands of base pairs away. In general, the binding sites of

most transcription factors do not show a strict preference of location or orientation relative to the TSS. Despite of the long history of research on transcriptional regulation (Kadonaga, 2004; Smale and Kadonaga, 2003), the mechanisms on how the machinery is targeted effectively to the TSS and which factors are involved is not well understood.

Although other mechanisms like chromatin remodeling have been shown to be important for transcriptional regulation, the binding of transcription factors to specific sites on the DNA are considered the key events in the initiation of transcription (Kadonaga, 2004). By far the best-characterized sequence motif is the TATA-box. However, as already realized many years ago, only the minority of human promoters contains a TATA-box. Especially the promoters of highly expressed genes are in most cases devoid of a TATA-box. They are in general GC-rich and contain a high number of CpG-dinucleotides (CpG-islands) (Aerts et al., 2004; Gardiner-Garden and Frommer, 1987). A

Abbreviations: RP, ribosomal protein; TSS, transcription start site; TBP, TATA binding protein; PWM, position specific weight matrix; TOP, terminal oligo pyrimidine.

* Corresponding author. Tel.: +49 30 8413 1159; fax: +49 30 8413 1152.

E-mail address: roepcke@molgen.mpg.de (S. Roepcke).

few other motifs have been localized in the very proximity of the TSS such as the TFIIB recognition element at position -37 to -32 , the initiator sequence (Inr) at the TSS and the downstream promoter element (DPE) at $+28$ to $+32$ (Smale and Kadonaga, 2003).

A recently developed technique allows the large-scale identification of the exact 5' end of mRNAs (Suzuki and Sugano, 2003; Suzuki et al., 2004). Mapping these sequences back onto the genome provides us for the first time with genome-scale data sets about observed locations of the TSS. In spite of the fact that many genes have a variable TSS, there are others whose transcripts need a definite and complete 5' end for the proper initiation of the subsequent translational process (Schmid et al., 2004; Suzuki et al., 2004). Comparing a set of promoter sequences with well-annotated TSS allows us to extract positional information that is encoded in the DNA. The goal for this study is the identification of new sequence motifs that are located in the proximity and in a fixed distance to the TSS. We call these motifs localized motifs. As an example, the TATA-box is one such motif. The TATA-box attracts the TATA binding protein (TBP), a part of the general transcription factor TFIID and thus compels the RNA polymerase to start transcription at a fixed location (Reese, 2003).

Human cytosolic ribosomal protein (RP) genes provide an exceptional data set to search for localized motifs. The ribosome is the core complex for protein synthesis and is an essential component in all living cells. The human genome contains 79 ribosomal proteins, encoded by 80 genes (Yoshihama et al., 2002). The RP genes are highly expressed and evolutionarily highly conserved. The exact transcription start site for all human cytosolic RP genes has been determined (Nakao et al., 2004; Yoshihama et al., 2002), which is a necessary requirement for the computational identification of localized motifs. Among

higher organisms we do not know any other promoter sequence set of related genes of similar size and accuracy.

Intense research over a long period of time revealed basic features common to all RP promoters. Transcription of RP genes starts at a C-residue that is embedded in an oligopyrimidine tract of length 5–25 base pairs, also known as the 5'-TOP signal. The 5'-TOP signal at the very start of the transcript was found to be an essential *cis*-regulatory motif for the translational control of RP gene expression (Levy et al., 1991; Meyuhas, 2000). The proximal promoter is generally GC-rich and is located in a strong CpG-island (see Fig. 1). The proximal promoters of all RPs but RPL7, RPL11 and RPL35A contain a strong CpG island (Karolchik et al., 2003; Yoshihama et al., 2002). The first exons are relatively short, 45 bp on average. For the majority of RP genes the start codon is situated close to the first exon–intron boundary. And in 20 out of the 80 cases the splice donor of the first intron follows exactly after the ATG. A TATA-box or a TATA-like sequence is reported to start around position -30 in 59 RP genes (Yoshihama et al., 2002). Along another line Perry et al. have studied the proximal promoter sequences of many mammalian RP genes in great details and deduce general features of the promoter architecture (Perry, 2003, 2005; Safrany and Perry, 1995). Consensus motifs for the 5'-TOP and for transcription factor binding sites of Sp1, GABP and YY1 have been identified. And some of them are already validated by wet lab experiments.

For ab-initio motif identification we apply MEME (Bailey and Elkan, 1994) on the proximal promoter sequences of these genes. Careful post-processing of the data leads to the identification of two new localized sequence motifs. Based on our analysis we hypothesize that the proper definition of the transcription start for RP genes is mediated by specialized

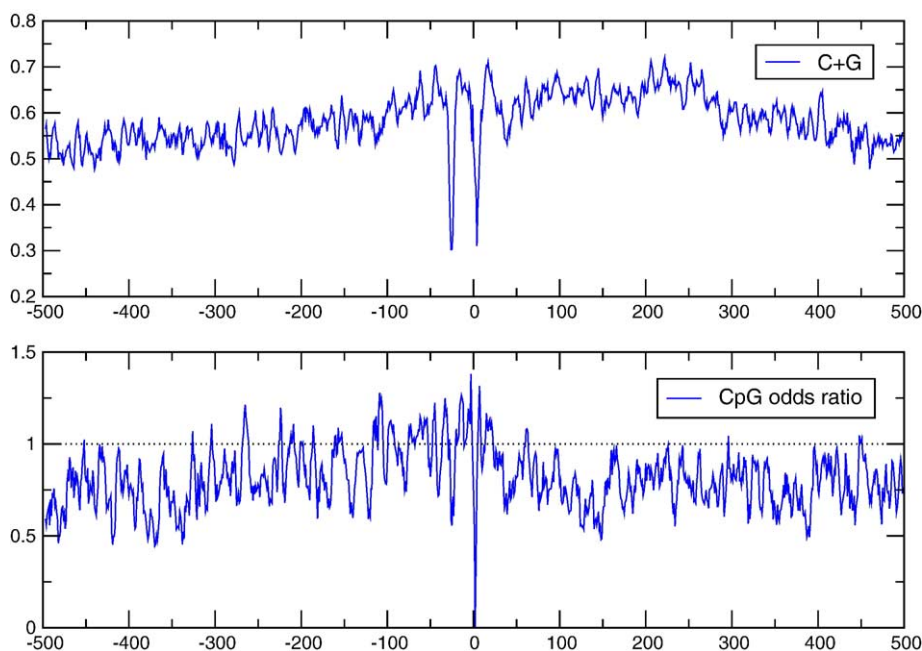


Fig. 1. The GC-content and the CpG odds ratio (i.e. CpG dinucleotide frequency normalized by the C and G frequency) across the 80 human ribosomal protein gene promoters relative to the TSS. The data has been smoothed using running averages over 5 nucleotide positions.

transcription factors that bind the DNA at a fixed distance to the TSS.

2. Material and methods

There are 79 human ribosomal proteins, encoded by 80 genes (Yoshihama et al., 2002). RPS4 is encoded by two different genes, one is on chromosome X and the other on Y. The sequences and the exon–intron structures for the 80 genes are obtained from the Ribosomal Protein Gene database (RPGDB) (Nakao et al., 2004). The accurate TSS given in RPGDB were determined by Yoshihama et al. (2002) and are not yet incorporated into the RefSeq database (Pruitt and Maglott, 2001). 79 mouse homologs are also downloaded from the RPGDB. RPS4Y is missing in current mouse genome sequence. To infer the TSS of mouse RP genes, we first utilized BLAT with standard parameters (Kent, 2002) to identify the human RP genes in the human genomic sequence at the UCSC genome browser (Karolchik et al., 2003). Then we compared the mouse RP gene sequences to the human/mouse/chimp/rat/chicken MultiZ alignment track (Blanchette et al., 2004) and mapped the TSS. Sequence data together with the annotated motifs is available as Supplementary Material from our webpage at http://evogen.molgen.mpg.de/supplementary_material/human_rp.

The authors of MEME observed that the addition of homologous sequences enhances the signal to noise ratio (Bailey et al., 1997). In order to find motifs that are conserved between species, 80 human RP promoters and 79 murine RP promoters (–300 to +200) are supplied to MEME (Bailey and Elkan, 1994). We use the MEME command “meme seq.fasta-

mod tcm-dna-minsites 4-maxsize 100000-nmotifs 20-minw 5-maxw 15-revcomp” to obtain top 20 motifs. As background model we use the position independent letter distribution of the input sequences, which corresponds to the default in MEME. The complete MEME results together with the set of promoter sequences can be found in the Supplementary Material. Corresponding sequence logos have been created using WebLogo (Crooks et al., 2004). The MEME motifs are also matched against position specific weight matrices (PWM) available in the Transfac Professional database (Matys et al., 2003) and Jaspar (Sandelin et al., 2004). Two matrices are compared by shifting the shorter one along the longer while computing for each shift the average divergence per matching position. Partial hits where at least half of the shorter matrix matches with at least three positions are also considered. The minimum value is taken as measure of the dissimilarity between the matrices.

We have scanned the promoter sequences on both strands for these matrices using the GENEREG package, which is an in-house implementation of the method described in Rahmann et al. (2003). For each single motif the balanced threshold was calculated. The balanced threshold t denotes the score value at which the type I error and the type II error are equal. In our context the type I error for a motif and a threshold t is defined as the probability to observe a score greater than t on a random background sequence of length 500. The type II error denotes the probability to observe a score less than t if the sequence is actually generated by the motif (Rahmann et al., 2003). To gain sensitivity we lowered the balanced threshold by 4 for the promoter sequence annotation. Additionally, the sequences were scanned with

Table 1
The identified sequence motifs in proximal promoters of human ribosomal proteins

Motif	Width	E-value	Consensus	Hits	Location relative to TSS	Comment
M1	15	2.00E–66	AATCCGCCGCCATCC	32/8 (31/0)	60 bp downstream	
M2	15	2.50E–59	CTTCCTTTTCTTTT	113/65 (62/5)	at TSS	CT
M3	15	7.90E–34	CAACATGGTGAGTGT	15/4 (13/1)	downstream	ATG+splicing signal
M4	15	6.00E–32	AGTCTCGCGAGATCT	17/18 (11/11)	40 bp upstream	
M5	12	8.20E–27	CTCTTCTTTTC	162/111 (80/8)	at TSS	CT
M6	12	1.60E–16	CGGAAGTGACGC	36/84 (11/57)	upstream	GABP
M7	14	3.20E–14	CACTGCCCATGCCG	42/37 (17/15)	40 bp upstream	palindrome
M8	12	1.50E–17	CCGCCATCTTGG	40/25 (23/6)	15 bp downstream	YY1 related
M9	12	2.50E–05	AGCCATGGTAAG	24/5 (15/1)	downstream	ATG+splicing signal
M10	12	9.00E–08	AAAGAAAAA	125/149 (4/22)	>100 bp upstream	A-rich
M11	15	5.40E–04	AACTACATTCCCAG	16/9 (3/1)	upstream, few hits	
M12	15	1.50E–03	TGACCCGGAAGTTAT	7/3 (6/2)	50 bp upstream, few hits	GABP
M13	12	1.40E–10	TTTCCGTTCCC	42/32 (37/8)	at TSS	CT-stretch
M14	15	7.80E–09	GAGGGGGCGGGCCA	46/45 (15/14)	>50 upstream	G-rich motif
M15	15	1.50E+01	TTTTAAATATTTT	55/54 (0/6)	>100 upstream	AT-rich
M16	15	2.10E+03	GCAGCCATGAGGTAA	3/1 (3/0)	15 bp downstream	
M17	9	3.0E+04	TCTCTCTT	24/7 (16/0)	at TSS	CT-stretch
M18	9	8.80E+05	GTGAGTGTT	8/1 (8/0)	20 bp downstream	GT slicing signal
M19	15	1.30E+06	AAGTGCTTAGCTTTT	6/7 (1/0)	only a few hits	
M20	14	5.00E+05	CGTGCTATATAAGC	2/0 (2/0)	only a few hits	
TATA	14		NTATAAANNNNNN	325/319 (50/42)	26 bp upstream	TATA motif by Transfac
YY1	8		GCCATNT	1319/1167 (306/202)	Not localized	YY1 motif by Transfac

The width, E-value, and consensus are cited from the meme output. In the column ‘hits’ we report the numbers of identified motifs (using the GENEREG package) in the direct/reverse strand in the whole proximal promoters and (in brackets) in a window from 80 bp upstream to 120 bp downstream of the TSS. Sequence logos for the first twelve motifs are shown in Fig. 2.

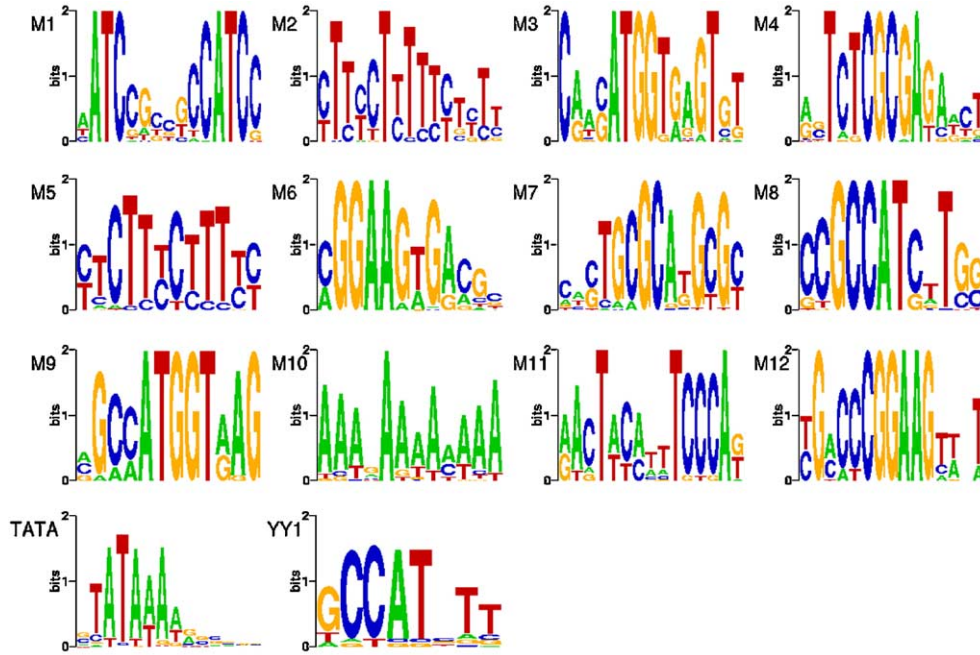


Fig. 2. Sequence logos for the first 12 motifs identified by MEME and the TATA and YY1 motifs from Transfac.

the Transfac motifs V\$TATA_01 and V\$YY1_Q6, using the same method.

For the GelPlot and the table we counted the occurrences of the motifs on the direct and reverse strand relative to the TSS. Partially overlapping hits of the same motif along the DNA were counted as one instance. For the GelPlot we further counted an overlapping hit on the direct and reverse strand as one, whereas in Table 1 we report both numbers separately.

3. Results

We compiled a set of sequences around the TSS of the 80 human cytosolic ribosomal protein (RP) genes taken from the RPG database (Nakao et al., 2004). We restricted our analysis to the region from –300 to +200 bp relative to the annotated TSS. Extending the sequences to the region from –500 to +500 gave essentially the same results (data not shown). To improve the

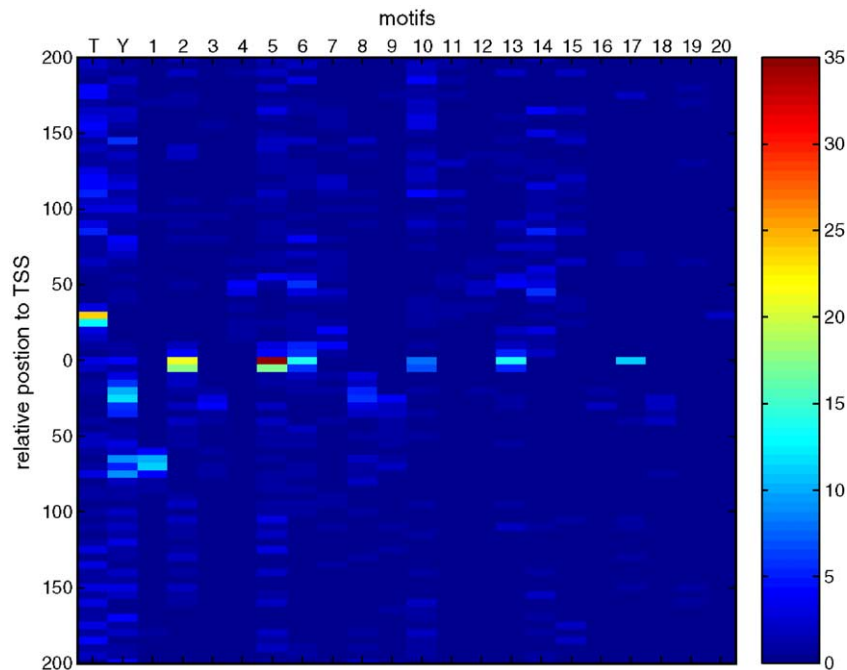


Fig. 3. Histogram of found motifs along the proximal promoters of ribosomal protein genes. The transcription start site is at 0. The first two tracks represent the TATA motif (T) and the YY1 motif (Y) taken from Transfac. Motifs 1–20 have been found by MEME.

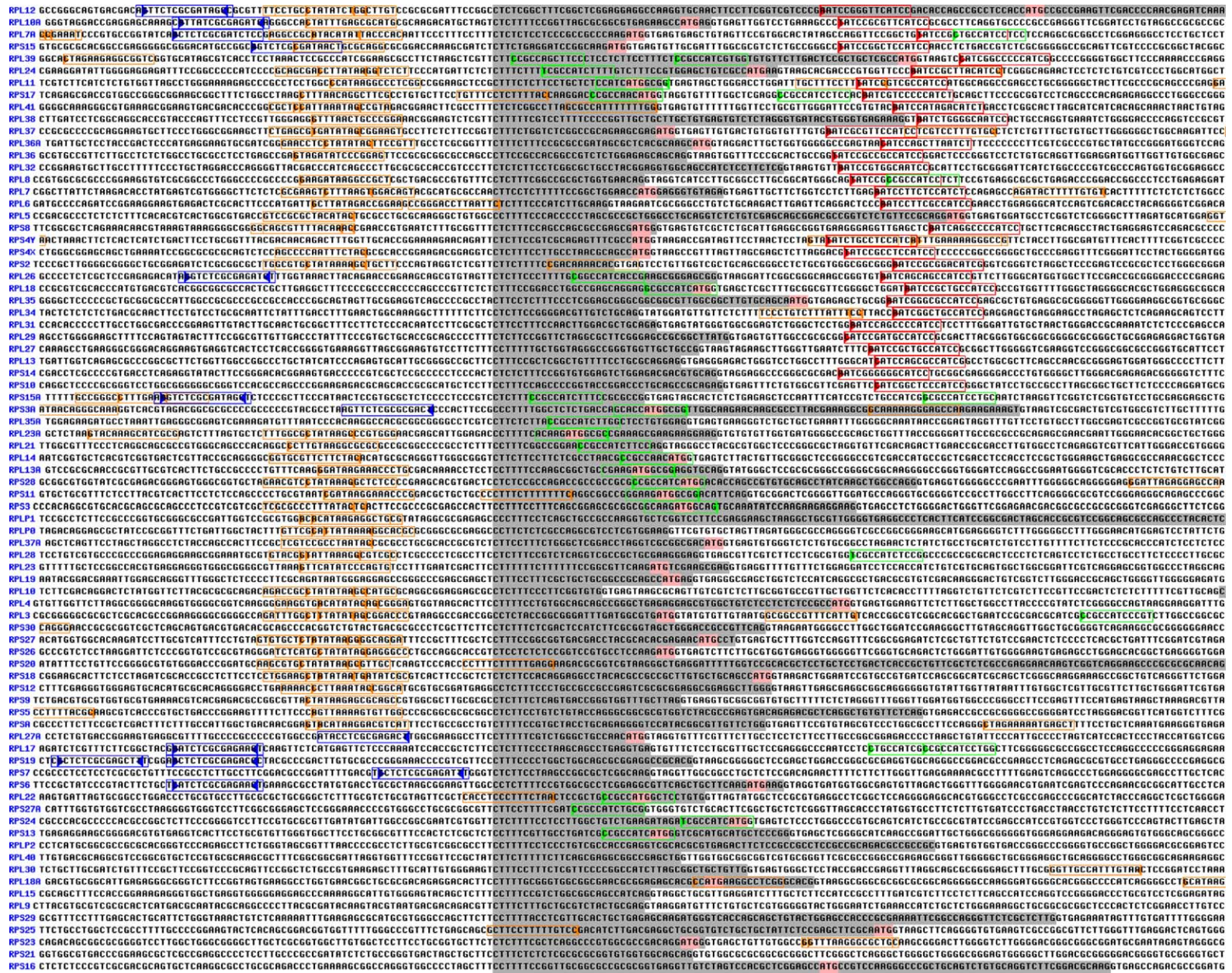


Fig. 4. Annotations of the motifs M1 (red), M4 (blue), M8 (green), and TATA (orange) to the sequences for the human ribosomal proteins. Exons are marked by a grey background and the translation start site by pink background. All sequences are aligned to their TSS. The sequences are sorted according to the occurrence of the different motifs. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

signal-to-noise ratio we augmented our data set by the orthologous mouse sequences (Bailey et al., 1997). We run the program MEME on this promoter set to identify over-represented sequence motifs. In addition to those found by MEME, Transfac (Matys et al., 2003) motifs for the TATA-box and for binding sites of the transcription factor Yin-Yang-1 (YY1) have been included in our analysis because they were already described to be localized in the proximal promoter of RP genes (Perry, 2003; Yoshihama et al., 2002). Details of the sequence preparation and the motif recognition are given in Materials and methods. The top 20 motifs returned by MEME are listed in Table 1. Corresponding sequence logos for the first 12 motifs are presented in Fig. 2. In order to identify localized motifs, we visualized the positions of the identified motifs in a color-coded 2-dimensional histogram (Fig. 3) and annotated a subset of the motifs to the proximal promoter sequences (Fig. 4). The number of hits in our promoter sequence set in fixed distances from the TSS is color-encoded from blue to red (heat map). We dubbed it GelPlot because it resembles a gel electrophoresis image.

The findings are organized as follows. First, we present two new localized motifs in detail. Then we discuss the motifs that either were already known before or significantly overlap with known transcription factor binding sites. Finally, the remaining MEME motifs are described.

3.1. A strong motif downstream of the TSS

The most interesting motif is the first and strongest motif found by MEME, M1 (see the corresponding sequence logo in Fig. 2). This 15 bp motif is structured in the following way. The flanks are composed of the sub-motif ATCC, of which the ATC is very highly conserved. The middle part is made up of a six base pair long, GC-rich sequence with a much less conserved consensus GCCGCC. M1 is found 31 times in the close proximity to the TSS, i.e. from -80 to $+120$ bp (see Table 1). It is clearly over-represented with MEME E-value $2E-66$. In all of these cases it occurs downstream of the TSS and shows a narrow distribution of distances to the TSS with an average of 62 bp (see Figs. 3 and 5). Moreover, M1 is not palindromic and all occurrences at the preferred position have the same orientation. This suggests that the M1-binding factors are directly associated to the transcription initiation complex. In all but one promoter M1 is situated in the first intron (see Fig. 4). It therefore might also represent a splicing signal. However the distribution of distances to the TSS is much narrower compared to the one of distances to the splice site (Fig. 5), which supports the role of motif M1 in the initiation of the transcriptional process. Note that the RPL39 promoter has a M1 site occurring at position $+77$, which is rather far away from the TSS (see Fig. 4). Interestingly, it was reported that RPL39 has an alternative TSS at 59 bp upstream of the M1 site (Yoshihama et al., 2002). This fits much better into the distance distribution of M1 to the TSS with an average of 62 bp. As observed already by the authors of MEME, adding homologous sequences to the training set augments the significance (Bailey et al., 1997). This was especially apparent for motif M1. This suggests that

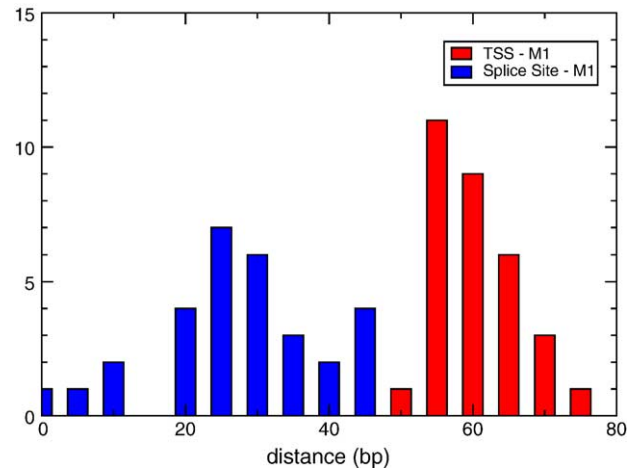


Fig. 5. Distribution of the distance of motif M1 to the TSS and to the splice site. The distribution of the distances to the TSS (mean 62, standard deviation 6) is much more narrow than one of the distances to the splice signal (mean 29, standard deviation 11).

the motif is evolutionarily conserved between human and mouse. And indeed, all instances of M1 in the human RP promoters have a pendant at a similar position in mouse and in some instances M1 is even well conserved from human through chicken (Fig. 6). In order to check whether M1 is a widely occurring motif or a specific one we applied the same annotation scheme to all the 4809 promoter sequences of EPD78 (Eukaryotic Promoter Database, version 78; Schmid et al., 2004). In total there were only 124 hits on the direct and 97 on the reverse strand, suggesting that M1 is RP-specific.

3.2. Another localized motif

Another localized motif is M4 (see Fig. 2 for a sequence logo). It is in itself a statistically strong motif (E-value $6E-32$) and its core part is palindromic. 18 out of the 22 observations in the -80 to $+120$ region represent pairs of overlapping sense–antisense hits ending up in 13 sites (see Fig. 4). All occurrences of M4 are evolutionarily highly conserved. 9 of these 13 sites occur around position -40 . M4 is similar to the BoxA motif, and 11 out of the 13 sites have been reported to contain BoxA recently (Perry, 2005). For M4 we could not find any similar motif in Transfac or Jaspar.

3.3. Other motifs

Motif M3 and M6 of our MEME result correspond to the oligopyrimidine tract at the start of transcription ($5'$ -TOP), the strongest localized signal in the promoters of cytosolic RP genes. $5'$ -TOP is essential for translational control (Meyuhas, 2000). Mutation studies have shown that interruption or displacement of the motif constrains the proper translational regulation of the corresponding mRNA in response to mitogenic and nutritional stimuli (Levy et al., 1991). To our knowledge, it has not been investigated whether the $5'$ -TOP is also important for the accurate start of transcription. The upstream extension of the $5'$ -TOP beyond the TSS is a hint that

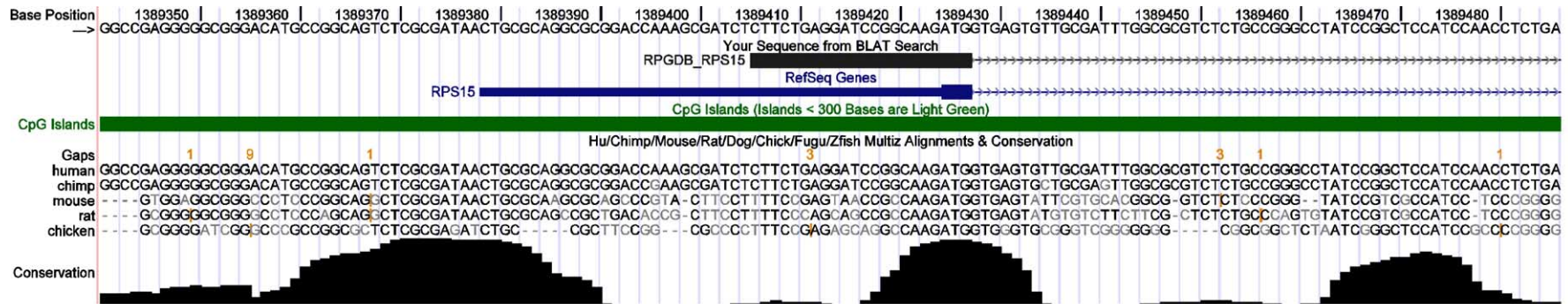


Fig. 6. Cross species conservation at the region surrounding TSS of the human RPS15 gene. The image is generated at the UCSC genome browser with a BLAT search using the mRNA sequence obtained from the RPGDB (Nakao et al., 2004). The beginning of the BLAT hit marks the TSS. The three “bumps” in the multi-species conservation track at the bottom correspond (from left to right) to the TATA-signal, ATGGT (start codon directly followed by a splicing signal), and the M1 site.

the pyrimidine tract is also a transcriptional signal. This extension can also have an alternative explanation, as it may allow for slight flexibility of the TSS (see Fig. 4). Additionally, Safrany and Perry (1995) found that mutation of the 5'-TOP sequence of the murine RPL30 gene impairs transcriptional activity.

Motif M8 occurs 65 times in our promoter set and 20 times around position +20. The localization according to the TSS is not very rigid and the orientation does not seem to be as important, since in 6 out of 29 cases it is located on the reverse strand. Motif M8 closely resembles the consensus DNA binding motif for the transcription factor YY1 (compare the logos in Fig. 2). Shrivastava and Calame (1994) gathered binding sites of YY1 and compiled them into weight matrices, corresponding to the rather unspecific Transfac matrix V\$YY1_Q6 (see Table 1 and Supplementary Material). For mouse RPL32 it was shown that the exonic and the intronic delta site bind YY1 (Chung and Perry, 1993). Excision experiments of the two elements suggest that they activate transcription cooperatively. Spacing and orientation does not impair the expression level. Notably, the integrity of the 5'-TOP of the transcripts was not investigated. Delta-binding sites are already described in several RP promoters: L30, L32, L13A, L7, L7A in mouse, S6, L7, and L17 in human, L1 and L14 in clawed frog and in P2 in rat (Antoine and Kiefer, 1998; Chung and Perry, 1993). When scanning our human promoter set with M8 we can identify the reported YY1 binding sites or the corresponding orthologous in L13A, L7, L7A and S17. The sites in S6, L30 and in exon 1 of L32 could not be identified.

4. Discussion

Interestingly, the last 10 bp of the M1 motif roughly matches to the YY1 sites. Indeed we find two overlapping occurrences of M8 and M1 motifs in our set of proximal promoters. We further find three M8 motifs around +62 bp, which is the preferred location of M1 in other proximal promoters. Here we identified two previously characterized YY1 binding sites as M1 motifs, the human intronic sequence that aligns with the intronic delta site of mouse RPL32 and the distal site of S17. The question emerges whether M1 describes YY1 binding sites. We find that motif M1 is much more specific and even quite different from the typical YY1 binding site: The M1 binding site is longer and has a well-conserved part at the beginning and the end, whereas the YY1 binding site lacks the conserved front part. And more importantly for our search of localized motifs, M1 consistently appears further downstream and in a definite orientation. From the above observations we suspect that the exonic and the intronic delta sites serve different functions. To be precise, we do not challenge the experimental finding that YY1 can bind M1 sites, but suggest that the M1 binding site is also target of another factor, which helps to localize the polymerase at the TSS. YY1 appears to fulfill different purposes in different environments. For a number of genes it functions as transcriptional activator and for others as

repressor (Shrivastava and Calame, 1994). In a very recent paper YY1 was shown to account for the accurate positioning of the transcriptional machinery in LINE transposable elements L1 (Athanihar et al., 2004). Mutations of the binding site entail only minor effects on the transcriptional activity but disrupt the proper location of the transcription initiation of L1. It is tempting to consider that YY1 fulfills the same role in RP promoters as well.

In a significant number of 20 RP genes the splice donor GT follows immediately the start codon ATG, resulting naïvely in the significant motif ATGGT. Motifs M3 and M9 contain this pattern. Besides that ATGGT evidently matches a good part to the core consensus of the reported YY1 binding sites. 18 out of the 20 ATGGT motifs overlap with our YY1 annotation (see Supplementary Material for detail). Additionally, M18 contains the motif GT and matches preferentially splice donor sites. Whether translational or splice signals are of relevance for the regulation of transcription initiation is not known to our knowledge.

Motif M6 and M12 resemble binding sites of the ETS family of transcription factors (ETS site). Many potential ETS sites have been recognized before (Yoshihama et al., 2002) and for mouse RPL30 and RPS16 it has been experimentally verified that GABP binds to ETS sites in the promoters and regulates their transcription (Genuario and Perry, 1996). In total we find 121 hits for M6 and 10 hits for M12 in our sequence set. Each of the motifs shows a slight preference for one orientation and the localization is clustered immediately before the TSS and around -50.

To our surprise, there was no motif among the best 20 resembling the TATA-box or a TATA-like sequence although its presence is already described in the literature and well-recognizable in the promoter set (Yoshihama et al., 2002). To annotate the TATA-like sequences in our set we utilized the Transfac matrix V\$TATA_01 (see Fig. 2 for the sequence logo). The hits are localized to a very narrow region around -26 bp upstream to the TSS (Figs. 3 and 4). However, the nucleotide sequences at that position in each single promoter are actually hardly similar to the well-characterized canonical TATA-box, characterized in (Bucher, 1990). Interestingly, its AT-richness shows a sharp contrast to the neighboring high GC profile (Fig. 1).

In conclusion, we have searched the proximal promoter of human ribosomal protein genes for conserved and over represented motifs and focused on those that show a narrow distribution of distances to the TSS. Besides the well-known TATA-like sequence, we have identified two novel localized motifs M1 and M4. They very likely represent binding sites of transcription factors that help to initiate transcription exactly at a definite site, as it is required for RP genes. Thus the positioning of the transcriptional apparatus emerges as a distinguishable regulatory task. We hypothesize that there exists a set of specialized transcription factors that bind the target DNA at a fixed distance from the TSS and interact directly with the basal transcription machinery. We propose that motif M1 sites are excellent candidates to test this model experimentally.

References

- Aerts, S., Thijs, G., Dabrowski, M., Moreau, Y., De Moor, B., 2004. Comprehensive analysis of the base composition around the transcription start site in Metazoa. *BMC Genomics* 5, 34.
- Antoine, M., Kiefer, P., 1998. Functional characterization of transcriptional regulatory elements in the upstream region and intron 1 of the human S6 ribosomal protein gene. *Biochem. J.* 336 (Pt 2), 327–335.
- Athanikar, J.N., Badge, R.M., Moran, J.V., 2004. A YY1-binding site is required for accurate human LINE-1 transcription initiation. *Nucleic Acids Res.* 32, 3846–3855.
- Bailey, T.L., Elkan, C., 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 2, 28–36.
- Bailey, T.L., Baker, M.E., Elkan, C.P., 1997. An artificial intelligence approach to motif discovery in protein sequences: application to steroid dehydrogenases. *J. Steroid Biochem. Mol. Biol.* 62, 29–44.
- Blanchette, M., et al., 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* 14, 708–715.
- 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.
- Chung, S., Perry, R.P., 1993. The importance of downstream delta-factor binding elements for the activity of the rpL32 promoter. *Nucleic Acids Res.* 21, 3301–3308.
- Crooks, G.E., Hon, G., Chandonia, J.M., Brenner, S.E., 2004. WebLogo: a sequence logo generator. *Genome Res.* 14, 1188–1190.
- Gardiner-Garden, M., Frommer, M., 1987. CpG islands in vertebrate genomes. *J. Mol. Biol.* 196, 261–282.
- Genuario, R.R., Perry, R.P., 1996. The GA-binding protein can serve as both an activator and repressor of ribosomal protein gene transcription. *J. Biol. Chem.* 271, 4388–4395.
- Kadonaga, J.T., 2004. Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell* 116, 247–257.
- Karolchik, D., et al., 2003. The UCSC Genome Browser Database. *Nucleic Acids Res.* 31, 51–54.
- Kent, W.J., 2002. BLAT—the BLAST-like alignment tool. *Genome Res.* 12, 656–664.
- Levy, S., Avni, D., Hariharan, N., Perry, R.P., Meyuhas, O., 1991. Oligopyrimidine tract at the 5′ end of mammalian ribosomal protein mRNAs is required for their translational control. *Proc. Natl. Acad. Sci. U. S. A.* 88, 3319–3323.
- Matys, V., et al., 2003. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* 31, 374–378.
- Meyuhas, O., 2000. Synthesis of the translational apparatus is regulated at the translational level. *Eur. J. Biochem.* 267, 6321–6330.
- Nakao, A., Yoshihama, M., Kenmochi, N., 2004. RPB: the Ribosomal Protein Gene database. *Nucleic Acids Res.* 32 (Database issue), D168–D170.
- Perry, R.P., 2003. Architecture of Mammalian Ribosomal Protein Promoters. Fox Chase Cancer Centre.
- Perry, R.P., 2005. The architecture of mammalian ribosomal protein promoters. *BMC Evol. Biol.* 5, 15.
- Pruitt, K.D., Maglott, D.R., 2001. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* 29, 137–140.
- Rahmann, S., Müller, T., Vingron, M., 2003. On the power of profiles for transcription factor binding site detection. *Stat. Appl. Genet. Mol. Biol.* 2.
- Reese, J.C., 2003. Basal transcription factors. *Curr. Opin. Genet. Dev.* 13, 114–118.
- Safrany, G., Perry, R.P., 1995. The relative contributions of various transcription factors to the overall promoter strength of the mouse ribosomal protein L30 gene. *Eur. J. Biochem.* 230, 1066–1072.
- Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W., Lenhard, B., 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* 32 (Database issue), D91–D94.
- Schmid, C.D., Praz, V., Delorenzi, M., Perier, R., Bucher, P., 2004. The Eukaryotic Promoter Database EPD: the impact of in silico primer extension. *Nucleic Acids Res.* 32 (Database issue), D82–D85.
- Shrivastava, A., Calame, K., 1994. An analysis of genes regulated by the multifunctional transcriptional regulator Yin Yang-1. *Nucleic Acids Res.* 22, 5151–5155.
- Smale, S.T., Kadonaga, J.T., 2003. The RNA polymerase II core promoter. *Annu. Rev. Biochem.* 72, 449–479.
- Suzuki, Y., Sugano, S., 2003. Construction of a full-length enriched and a 5′-end enriched cDNA library using the oligo-capping method. *Methods Mol. Biol.* 221, 73–91.
- Suzuki, Y., Yamashita, R., Sugano, S., Nakai, K., 2004. DBTSS, DataBase of Transcriptional Start Sites: progress report 2004. *Nucleic Acids Res.* 32 (Database issue), D78–D81.
- Yoshihama, M., et al., 2002. The human ribosomal protein genes: sequencing and comparative analysis of 73 genes. *Genome Res.* 12, 379–390.