Acta Biochimica Polonica

Vol. 54 No. 1/2007, 89–98

on-line at: www.actabp.pl

Regular paper

A common *cis*-element in promoters of protein synthesis and cell cycle genes

Lucjan S. Wyrwicz[1,2]✉, Paweł Gaj[2], Marcin Hoffmann[1], Leszek Rychlewski[1] and Jerzy Ostrowski[2]

[1]*BioInfoBank Institute, Poznań, Poland;* [2]*Department of Gastroenterology, Medical Center for Postgraduate Education and Maria Sklodowska-Curie Memorial Cancer Center and Institute of Oncology, Warszawa, Poland*

Received: 28 November, 2006; revised: 13 February, 2007; accepted: 16 February, 2007
available on-line: 09 March, 2007

Gene promoters contain several classes of functional sequence elements (*cis* elements) recognized by protein agents, e.g. transcription factors and essential components of the transcription machinery. Here we describe a common DNA regulatory element (tandem TCTCGCGAGA motif) of human TATA-less promoters. A combination of bioinformatic and experimental methodology suggests that the element can be critical for expression of genes involved in enhanced protein synthesis and the G1/S transition in the cell cycle. The motif was identified in a substantial fraction of promoters of cell cycle genes, like cyclins (*CCNC, CCNG1*), as well as transcription regulators (*TAF7, TAF13, KLF7, NCOA2*), chromatin structure modulators (*HDAC2, TAF6L*), translation initiation factors (*EIF5, EIF2S1, EIF4G2, EIF3S8, EIF4*) and previously reported 18 ribosomal protein genes. Since the motif can define a subset of promoters with a distinct mechanism of activation involved in regulation of expression of about 5% of human genes, further investigation of this regulatory element is an emerging task.

Keywords: regulation of transcription, gene expression, promoter, bioinformatics, comparative genomics, gel shift assay

INTRODUCTION

The regulation of transcription is the major process modulating expression of genes on both qualitative and quantitative levels. Regulatory elements concentrated in gene promoters include several classes of functional DNA sequence motifs (*cis* elements) recognized by protein agents (*trans* elements), i.e. essential components of the DNA-directed RNA polymerase transcription machinery (GTF, general transcription factors) and complementary transcription factors (TFs). The efficiency of transcription is enhanced by specific interactions between DNA-binding proteins and sequence elements present in promoters (TFBSs, transcription factor binding sites). Apart from the *cis–trans* cooperation other regulating mechanisms include variations in chromatin composition *via* histone modifications (Barrera & Ren, 2006).

The regulation of gene expression is a complex process resulting in enhanced activity of the encoded gene products (proteins). Previously, groups of co-regulated genes (so called gene expression modules) has been identified by comparative measurements of gene expression in various tissues (Segal *et al.*, 2004). For the yeast model gene-expression clusters were effectively translated into regulatory networks defining a molecular background of co-expression (Segal *et al.*, 2003). Specific functions have been assigned to several *cis* elements and the presence (or activation) of the related *trans* agents (TFs) shown to activate specific molecular switches triggering the expression of respective genes (Hughes *et al.*, 2000). Since the regulation of gene expression in higher Eukaryotes

✉To whom correspondence should be addressed: Lucjan S. Wyrwicz, BioInfoBank Institute, ul. Limanowskiego 24A, 60-744 Poznań, Poland; tel: (48 61) 865 3520, fax: (48 61) 864 3350; e-mail: lucjan@bioinfo.pl

Abbreviations: bp, base pair; DDT, dithiothreitol; EMSA, electrophoretic mobility shift assay; EPD, Eukaryotic Promoter Database; GTF, general transcription factor; SAGE, serial analysis of gene expression; TF, transcription factor; TFBS, transcription factor binding site; TSS, transcription start site.

is more complex (Jura *et al.,* 2006), the progress in the field is less advanced. So far, specific regulatory mechanisms has been assigned to a limited number of functional groups of genes (Hardison *et al.,* 1997; Frech *et al.,* 1998; Yoshihama *et al.,* 2002), cellular processes (Kel *et al.,* 2001) and tissue-specific expression patterns (Wasserman & Fickett, 1998).

The interplay between the activity of TFs in a given cell and the presence of TFBS in promoters is one of the most important mechanisms responsible for inducible or tissue-specific transcription. Therefore identifying functional elements of a gene promoter allows prediction of the gene's expression in various tissues and different environmental conditions (Tronche *et al.,* 1997). The description of the full repertoire of transcription factors (*trans*) and their binding specificities (*cis* elements) is one of the most important tasks of bioinformatics in the analysis of gene expression (Pennacchio & Rubin, 2001).

Here we present a common DNA element of human promoters involved in regulation of genes associated with protein translation. Using genome-wide scanning we suggest that the element can take part in regulation of expression of nearly 5% of human genes, mostly those transcribed from TATA-less promoters.

METHODS

Sequence neighborhood. Whole genome human–mouse alignments (genome builds: "hg17", May 2004; "mm5", May 2004) were obtained from the Genome Browser (Kent *et al.,* 2002). Promoter sequences, defined as 1000 base pairs (bp) upstream and 100 bp downstream from transcription start site (TSS), of the 16 749 human genes (non-redundant set from the Reference Sequence project) (Kent *et al.,* 2002) were retrieved from sequence alignments.

The promoter alignments were scanned for occurrences and evolutionary preservation of all k-mers ranging from 6 to 8 bases. A k-mer was recognized as conserved only when it occurred in both genomes at corresponding (homologous) locations with no differences in sequence. The observed conservation ratio (c) of a motif was determined as the proportion of human occurrences (k) that were present in conserved form (non-mutated) in a homologous locus of the mouse genome to all the motif's occurrences in human promoters (n; $c=k/n$).

To analyze the degree of conservation of a k-mer we tested each motif against its "sequence neighborhood" (SN; Table 1) which was defined by all k-mers differing by exactly one nucleotide (e.g.: SN of AAAAA consists of: CAAAA, GAAAA, TAAAA, ACAAA, ..., AAAAT). Such algorithm was introduced to avoid the problem of unequal conservation ratio of motifs of different nucleotide content.

The conservation ratio of each sequence motif was assessed against the average conservation ratio of the other sequences from its sequence neighborhood (C) using a binomial distribution model (probability of k conserved instances out of total n instances for given probability (C) of conservation for any one instance). Z score was calculated according to the binomial approximation of the normal distribution formula (Feller, 1968). Motifs with the Z score of binomial statistics above 4.0 were selected.

The dataset of regulatory motifs. The collected motifs were grouped before compilation into a database of potential regulatory signals. The rules for clustering were as follows: sequences could differ by a maximum of one nucleotide and could be shifted by a maximum of one position and no gaps were allowed in the alignment. The distribution of clustered motifs was evaluated by the Student's *t*-test for paired data and the clustering was allowed only for motifs of consistent distribution (P<0.05, motifs' occurrences were counted in a 20 bp window along the promoter sequences). The dataset is available for browsing at URL: http://promoter.bioinfo.pl (Wyrwicz L.S., Rychlewski L., Ostrowski J., manuscript in preparation).

The impact of the motif presence on promoter activity was assessed for gene expression profiles

Table 1. The sequence neighborhood of core 8 bp (CTCGCGAG) fragment of TCTCGCGAGA motif

Sequence	*Z* score
CTCGCGAG	21.7013
ATCGCGAG	3.4648
GTCGCGAG	–1.3939
TTCGCGAG	–2.6871
C**A**CGCGAG	–0.4366
C**C**CGCGAG	–0.5648
C**G**CGCGAG	–1.3154
CT**A**GCGAG	–2.0048
CT**G**GCGAG	–3.8648
CT**T**GCGAG	–1.6411
CTC**A**CGAG	–0.7837
CTC**C**CGAG	–2.4787
CTC**T**CGAG	–3.2008
CTCG**A**GAG	–1.9628
CTCG**G**GAG	–2.9035
CTCG**T**GAG	–2.1266
CTCGC**A**AG	–1.342
CTCGC**C**AG	–2.1097
CTCGC**T**AG	–1.1389
CTCGCG**C**G	0.1945
CTCGCG**G**G	–0.4507
CTCGCG**T**G	0.9418
CTCGCGA**A**	–1.6187
CTCGCGA**C**	–0.0373
CTCGCGA**T**	0.8702

obtained from publicly available results of SAGE experiments (Serial Analysis of Gene Expression) (Velculescu *et al.*, 1995) deposited in the GEO database (http://www.ncbi.nlm.nih.gov/geo) (Edgar *et al.*, 2002). The SAGE method was preferred instead of microarrays or other platforms for estimation of gene expression as it has previously been shown to exhibit more precise discrimination between high and low abundance transcripts (van Ruissen *et al.*, 2005). A total of 164 gene expression libraries of 10 bp tags associated with *Nla*III restriction sites representing various tissues and cell lines derived from human normal and cancerous cells were selected (Edgar *et al.*, 2002). The previously published algorithm (Klimek-Tomczak *et al.*, 2004) was used to match the expression data to the set of Reference Sequence project genes. Genes from each SAGE experiment corresponding to tags found in the SAGE library were sorted by the number of tag counts and grouped into: "high expression" (HE; top 40% of expressed genes) and "low expression" (LE; 40% of genes with lowest tag count). Chi-square test was used to compare the number of conserved motif occurrences in both groups of promoters.

Annotation of human promoters. We tested the presence of the motif of interest in human promoters retrieved from the Eukaryotic Promoter Database (EPD) (Schmid *et al.*, 2006) and the UCSC Genome Browser (Kent *et al.*, 2002) databases ("upstream1000" data set) using proprietary scripts written in PERL programming language. The functional annotation of genes was performed with the Gene Ontology (http://geneontology.org) (Harris *et al.*, 2004) and UniProt (http://www.uniprot.org) resources (Bairoch *et al.*, 2005). The scripts, datasets and search results are available as Supplementary materials (URL: http://lucjan.bioinfo.pl/supplemental/cellcycle).

Electrophoretic mobility shift assay. Starved HeLa cells were stimulated with 15% fetal calf serum for: 0, 1, 6 and 24 h. Non-histone nuclear protein extracts were isolated as previously reported (Ostrowski *et al.*, 1991). The sequences of oligonucleotides used in the study were as follows: XPC-single (5′ CTT TCC CGC CTC TCG CGA GAA CAC AAG AGC), COX11 (5′ AGG TCA AAT CTC GCG AGG CGT GCT CCG TCT CGC GAG ATC TGG G), XPC (5′ TCC TCA CGT TTC CGG AGA TTG ACG TTG CTC TTG TGT TCT CGC GAG AGG CGG G), COX11-d1 (5′ AGG TCA AAT CGG CGT GCT CCG TCT CGC GAG ATC TGG G), COX11-d2 (5′ AGG TCA AAT CTC GCG AGG CGT GCT CCG TCG ATC TGG G), COX11-d12 (5′ AGG TCA AAT CGG CGT GCT CCG TCG ATC TGG G), HNRPK (5′ AGT TGT TAG ATC TCG CGA GAG GTT CGC CCC). Double stranded DNA was phosphorylated with [γ-^{32}P]ATP using T4 polynucleotide kinase (Fermentas) according to the manufacturer's protocol. The binding mixture consisted of 3 µl of binding buffer (20% glycerol, 5 mM $MgCl_2$, 2.5 mM EDTA, 2.5 mM DTT, 250 mM NaCl, 50 mM Tris/HCl, pH=7.5), 2.5 µg of nuclear protein extract and 1.4 pmol of oligonucleotide. The reaction was carried out in a total volume of 15 µl (30 min at 25°C). EMSA was performed using 4 µl of the product on an 8% non-denaturating polyacrylamide gel (37.5:1, Promega) for 1 h at 7.5 V/cm. Autoradiograms were obtained using Imaging screen K and Molecular Imager FX (BioRad).

RESULTS AND DISCUSSION

The applied algorithm allowed the identification of a subset of the human genome as potential regulatory motifs. The summary of the motifs' selection is shown in Table 2. Since palindromes constitute one of the most important group of regulatory elements, the dataset was tested for the presence of such motifs. The top scoring palindrome motifs identified are summarized in Table 3.

An uncharacterized palindrome motif TCTCGCGAGA was identified among the most conserved motifs in a genome-wide human–mouse assessment of 6–8 nucleotide segments and is deposited in PromoSignalDB under the accession number H-26.1 (http://promoter.bioinfo.pl/data.pl?acc=H-26.1). The core part of the motif (CTCGCGAG) was conserved in 151 cases out of 283 occurrences in the analyzed human promoters (53%). The conservation ratio increased to 66% for motifs located between base –180 and +40 in relation to TSS.

The motif distribution in human promoters is shown in Fig. 1A. Detailed analysis of the motif distribution within human promoters suggested that the motif tended to be present in more than one copy. In the "upstream1000" dataset the consensus element was present in a duplicated form 12.62 times more often than expected. The motifs within a pair were usually separated by up to 200 nucleotides (Fig. 1B). Selective conservation of the two copies in homologous genomic loci of related species and accumulation of mutations in the spacer sequence were observed (an example motif shown in Fig. 1C).

Table 2. Summary of the motif identification procedure

Sequence length (nt)	Total number of motifs (4^n)	Observed motifs	Potential regulatory motifs (fraction of observed motifs)
6	4096	4096 (100%)	361 (0.0888)
7	16384	16375 (99.95%)	578 (0.0353)
8	65536	63730 (97.24%)	685 (0.0108)

Table 3. Summary of top scoring palindromes of 6 and 8 nucleotides

		Sequence	*Z* score	Fraction of conserved occurrences	Description
n=6 nt	1.	CACGTG	31.70	0.51591	cMyc
	2.	TCGCGA	27.00	0.69642	Unknown
	3.	CCATGG	26.38	0.37763	CCAAT
	4.	TGCGCA	22.23	0.37326	Unknown
	5.	GACGTC	21.00	0.53813	AP1
	6.	GCGCGC	17.19	0.31609	GC rich motif
	7.	GCATGC	17.00	0.35995	Unknown
	8.	CAGCTG	12.09	0.28532	AP1
	9.	TAATTA	8.76	0.29905	TATA-box
	10.	CTGCAG	8.42	0.23389	Unknown
n=8 nt	1.	CTCGCGAG	21.701	0.53356	Unknown
	2.	TGACGTCA	15.535	0.58663	CREB / AP1
	3.	GCCATGGC	11.843	0.41535	Unknown
	4.	CTGCGCAG	10.600	0.31567	Unknown
	5.	TCACGTGA	10.568	0.56554	cMyc
	6.	CGCATGCG	8.526	0.50967	Unknown
	7.	CCACGTGG	6.858	0.3913	cMyc
	8.	TCAGCTGA	6.060	0.31578	AP1
	9.	GCGCGCGC	5.495	0.28672	GC rich motif
	10.	TTCCGGAA	4.809	0.33139	Ets

The preference of the motif to occur in more than one copy is unusual. To assess if the motif can be recognized in the single or double configuration an experimental study of electrophoretic mobility shift assay (EMSA) of an oligonucleotide containing the TCTCGCGAGA motif was performed. For testing we selected native oligonucleotides containing a motif nearly identical to the consensus sequence. The sequences were obtained from proximal promoters of *XPC* (xeroderma pigmentosum, complementation group C) and *COX11* (cytochrome *c* oxidase assembly protein 11). In the selected promoters two copies of the element were present in close proximity, spaced by 18 and 9 nucleotides, respectively. Both elements were conserved in homologous loci of different species of vertebrates (mouse, rat, dog).

The binding of nuclear proteins to the double stranded oligonucleotides (dsDNA) was assessed by EMSA. To induce transcription, HeLa cells were first starved for 48 h, then stimulated with fetal calf serum for 0, 1, 6 or 24 h. We observed a specific mobility shift for dsDNA probes containing two copies of the motif (Fig. 2B, C), while no shift was observed for an oligonucleotide containing a single copy of the element (XPC-single; Fig. 2A) or for a single nearly identical motif retrieved from promoter of *HNRPK* (not shown). The specific shift was present only when nuclear protein extract from induced cells was used and the amount of shifted probe increased with extension of time of serum induction.

To investigate the function of the presence of two copies of the motif, deletion mutants of *COX11* native element were assayed (Fig. 3). No shift was observed for probes with the central six nucleotides deleted in either one (lanes 3, 4) or both copies (lane 5).

The described regulatory motif was previously identified in other genome-wide studies but no details on its activity were provided. FitzGerald and coworkers (2004) identified this element as a com-

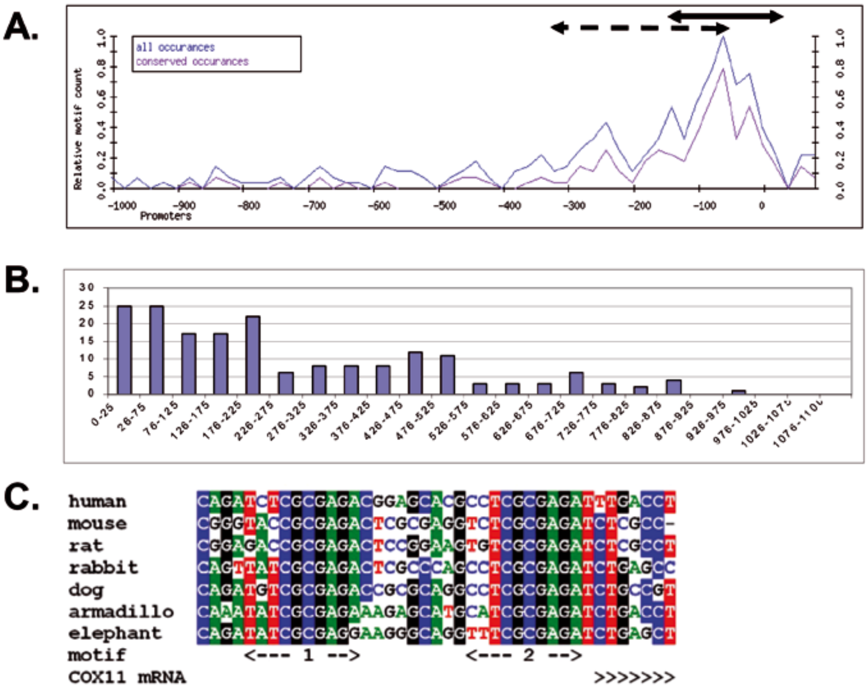


Figure 1. Characteristics of TCTCGCGAGA motif.
A. The distribution of the motif in human promoters in relation to TSS. Notice the high relative ratio of conserved *vs.* all occurrences. Bimodal distribution of the motif is a result of its tandem occurrences. Ranges of proximal and distal motifs are marked with solid and dashed arrows, respectively. **B**. The distribution of spacer length in EPD dataset. **C**. Evolutionary conservation of the tandem element among various species of vertebrates (*COX11* promoter).

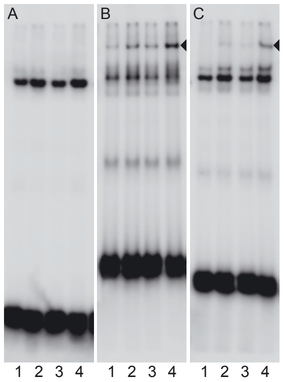


Figure 2. Electrophoretic mobility shift assay of *in vitro* binding of novel *cis* element TCTCGCGAGA.
Starved HeLa cells were stimulated with 15% fetal calf serum for: 0 (lane 1); 1 h (lane 2), 6 h (lane 3) and 24 h (lane 4). **A**, *XPC*-single; **B**, *COX11*; **C**, *XPC*. The locations of the delayed probe likely corresponding to a specific interaction of nuclear extract protein(s) with the tandem motif oligonucleotide are marked with arrowheads.

mon motif clustering in the human genome in close proximity to transcription start sites. Xie *et al*. (2005) identified the element as a conserved motif in several mammalian genomes. Haun and coworkers (1993) investigated the role of the TCTCGCGAGA element in promoter of *ARF3* and concluded that mutation of a single copy of the element diminished the transcriptional activity of the *ARF3* promoter *in vivo*. Notably — the *ARF3* promoter also contains a second imperfect copy of the motif 23 nucleotides apart, which was present in the analyzed gene construct, but not reported by Haun and coworkers (TCT CGC GAG AAC TGC CGC TAG CTA CCG CGC AGC TCT CGC GCG A). The effect of mutation or deletion of the latter site was not investigated.

The presence of similar motifs was postulated by Roepcke and coworkers (2006) (motif M4; AGTCTCGCGAGATCT) and Perry (2005) in their studies on sequence elements overrepresented in promoters of human ribosomal genes. None of the presented studies suggested the tandem composition of the active element.

We performed a search of human promoters containing the composite tandem motif in the Eukaryotic Promoter Database. Other genes containing the motif in their promoters were functionally related to enhanced protein synthesis and included translation initiation factors (*EIF5, EIF2S1, EIF4G2, EIF3S8, EIF4*), cell cycle genes active in G1/S phase (*CDK8, CDC25A, CUL1*), cyclins (*CCNC, CCNG1*), genes linking gene expression and cell cycle regulation (*TAF7*), transcription regulators (*TAF13, PROX1, KLF7, NCOA2*) and chromatin structure modulators (*HDAC2, TAF6L*). The motifs identified in promoters of the mentioned genes are shown in Table 4.

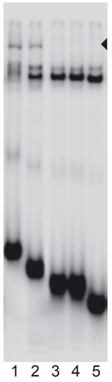


Figure 3. Electrophoretic mobility shift assay of *in vitro* binding of native elements and mutants.
Protein extracts were obtained from starved HeLa cells stimulated with 15% fetal calf serum for 24 h. **1**, *XPC*; **2**, *COX11*; **3**, *COX11*-d1; **4**, *COX11*-d2; **5,** *COX11*-d12. The location of the delayed probe corresponding to a specific interaction is marked with the arrowhead.

Since the role of the motif in gene expression had not been investigated before, we tested whether the motif occurs in promoters of tissue-specific genes. Analysis of gene expression profiles obtained with the SAGE method revealed that the motif was overrepresented in promoters of highly expressed genes when compared with the low expression subset in 121 of 164 tested tissues and cell lines. Similar results (association in >50% of tested gene expression profiles) were achieved for the general or very common regulatory elements which do not exhibit tissue selectivity, i.e. TATA-box, CAAT enhancer, Oct1 and Ets motifs, as well as the Kozak sequence (a motif associated with highly efficient translation) (Kozak, 1987). The results of the analysis are available as Supplementary materials (URL: http://lucjan.bioinfo.pl/supplemental/cellcycle).

Although the consensus motif has been determined, analysis of reference human promoters (EPD) and comparative genomics analysis suggests that a certain degree of variation is accepted within the site, as can be shown for the tandem motif of *COX11* promoter (Fig. 1C, human: TCTCGCGAGA N_9 CCTCGCGAGA, mouse: TACCGCGAGA N_9 TCTCGCGAGA). Moreover, in several promoters of ribosomal protein genes we identified two imperfect copies of the motif located in close range, where neither the proximal nor the distal copy matched 10 bp consensus (Table 4).

An analysis of human promoters retrieved from the UCSC Genome Browser dataset matching the tandem element was performed and ribosomal protein genes were observed among the most abundant class of genes (Table 5). Since the bioinformatic identification of the full repertoire of genes associated with a motif relies on an assumption of a minimal degree of similarity to its consensus sequence

Table 4. List of selected genes involved in protein translation and cell cycle regulation with the duplicated motif within their promoters.

Promoter sequences retrieved from genome assembly (not deposited in EPD) are marked with asterix.

Process	Gene	Motif 1	Spacer	Motif 2	Relation to TSS
Cell cycle regulation	*CCNC*	TCTCGCGAGA	10	CGCCGCGAGC	–172
	CCNG1	GATCGCGGAG	47	TCTCGCGAGA	–44
	CDC10	TCTCGCGAGA	199	GGTCGCGGAG	–21
	CDC5L	TCTCGCGAGA	136	TCTCGCGAGA	–37
	CDK8 *	GTTCGCGAGT	100	GCGCGCGAGA	–29
	CUL1 *	GCTCGCGAGG	141	CCACGCGAGC	–278
Transcription regulators	*TAF7*	TCTCGCGAAA	107	CGTCGCGACG	–5
	TAF13 *	GATCGCGCCA	317	ACTCGCGACC	–192
	KLF7 *	GCTCGCGCTG	381	GATCGCGAGA	–108
	NCOA2 *	AGTCGCGCCT	666	GCTCGCGAGC	–129
	HDAC2	GCTCGCGGCA	21	CTTCGCGACA	–361
	TAF6L *	AGTCGCGAAA	16	ACTCGCGAGC	–309
Translation initiation factors	*EIF5*	AGTCGCCATG	120	TCTCGCGAGA	–260
	EIF4B *	GCTCGCGTGT	23	TTTCGCGACA	–180
	EIF2S1	ATGCGCGAAG	160	ATTCGCGAGG	–686
	EIF4G2 *	GCACGCGAGG	291	CTTCGCGAAT	–194
	EIF3S8	AATCCCGACA	15	ATTCGCGACG	–667
Ribosomal proteins	*RPS15*	TCTCGCGATA	55	TGTTGCGATT	+63
	RPS19	TCTCGCGAGC	9	TCTCGCGAGA	–40
	RPS9	TCTCGCGAGA	11	CTCCGCGAGG	–74
	RPS7	CCTCGCGCTG	39	TCTCGCGAGA	–8
	RPS6	TCTCGCGAGA	98	GGTGGCGAGT	+65
	RPS5	GCTCGGGATC	4	GCTCGCGAGC	–544
	RPS28	GTCCGCGATA	60	TATCGCGAGA	–57
	RPS19	TCTCGCGAGC	9	TCTCGCGAGA	–40
	RPS11	TCTCGCGATA	73	TCTCGCGAGA	–187
	RPLP0	CAACGCGAGG	10	CTTCGCGACC	–180
	RPL9	GATCGCGAGG	6	GTTCGAGACC	+29
	RPL37A	AGTCGCGAGC	130	TGTCGCTAGC	–467
	RPL35	CATGGCGAAA	134	GATCGCGACA	–618
	RPL27A	TCCCGCGAGA	73	CCTCGCGAGA	–17
	RPL24	CCTCGCGATG	139	TGTCGCCATG	+8
	RPL12	CATGGCGACA	35	TCTCGCGATA	–47
	RPL10A	GGGCGCGAAT	165	TATCGCGAGA	–41
	RPL10	TCTCGCGGTC	122	TCTCGCGACC	–334

(Hoh *et al.,* 2002), the list of genes presented here is only an approximation.

The structure of human ribosomal genes has previously been studied and distinct features of their promoters were identified, including: oligopyrimidine tract around TSS, GC-rich promoters with TATA-like sequences, but usually lacking a typical TATA-box (Yoshihama *et al.,* 2002). We assessed whether the mentioned features are present in promoters containing the motif. Since the exact position of the transcription start site for the "upstream1000" dataset remains uncertain (Makalowski, 2001), we were not able to analyze the neighborhood of TSS in this data. However, for the promoters retrieved from the EPD database we confirmed the presence of the mentioned features characteristic for ribosomal promoters and selected other genes (example entries are shown in Table 6). Since we observed the similar characteristics of the gene promoters presented here we can assume that the described tandem motif complements the previous observation of the functional elements of ribosomal protein gene promoters.

Based on our genome-wide promoter analysis, experimental work and previously published studies we conclude that a tandem TCTCGCGCAGA motif is a common regulator interacting with unknown protein(s) induced during enhanced protein synthesis or/and cell proliferation. The motif consists of a palindrome sequence and is active in a tandem arrangement. Due to the close proximity between the studied element and the transcription start site we also suggest that it may play a central role in expression of a significant fraction of human genes transcribed from a distinct class of TATA-less promoters previously described as ribosomal protein gene-specific promoters (Yoshihama *et al.*, 2002). The identification of the motif in the functionally associated sets of genes of translation and the cell cycle suggests the existence of a common process-specific mechanism of gene expression. Since the previously described specific features of ribosomal gene promoters have a low information content (oligopyrimidine tract around TSS, GC-rich promoters with TATA-like sequences, but usually lacking typical TATA-box) their usefulness in the identification of co-regulated genes is limited. The identification of the described motif enables the identification of a full repertoire of genes regulated in this manner.

Table 5. Search for genes with the tandem motif in promoters.

The genes were grouped into functional classes according to assigned Gene Ontology terms.

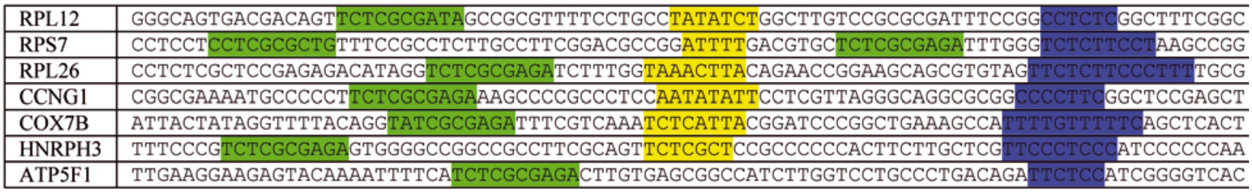
Functional class	Gene Ontology ID	Genes with tandem motif	Gene Ontology term	Gene name (HUGO)
Protein translation	GO:0005842	6	cytosolic large ribosomal subunit (sensu Eukaryota)	*RPLP0, RPL29, RPL35, RPL12, RPL10, RPL6*
	GO:0003743	12	translation initiation factor activity	*EIF2B5, EIF2B4, DENR, EIF5, MRPL49, ITGB4BP, BZW1, EIF2S1, EIF5A2, EIF4G2, EIF3S8, EIF4B*
	GO:0005840	17	ribosome	*RPS11, RPL10A, RPLP0, RPL29, RPL35, RPL12, MRPL2, UBA52, MRPL49, RPL10, RPS2, MRPS7, RPL37A, RPL23A, RPL6, RPS16, MRPS18B*
	GO:0003735	20	structural constituent of ribosome	*RPS11, RPL10A, RPLP0, RPS5, RPL29, RPL35, RPS28, RPL12, MRPL2, UBA52, MRPL49, RPL10, RPS2, MRPS7, RPL37A, MRPS6, RPL23A, RPL6, RPS16, MRPS18B*
	GO:0006412	32	protein biosynthesis	*RPS11, GFM2, RPL10A, RPLP0, RPS5, RPL29, RPL35, EIF5, RPS28, RPL12, PABPC4, MRPL2, UBA52, MRPL49, SCYE1, RPL10, RPS2, MRPS7, ITGB4BP, RPL37A, MRPS6, EIF5A2, QRSL1, RPL23A NACA, EIF3S8, RPL6, RPS16, MRPS18B, KARS, ETF1, EIF4B*
	GO:0016567	6	protein ubiquitination	*TRIM23, FBXO11, UBA52, TRAF7, BARD1, UBC*
	GO:0008565	7	protein transporter activity	*COPB2, AP1S3, KPNA2, SORT1, SNX1, CCT6B, STX16*
	GO:0051082	11	unfolded protein binding	*CALR3, DNAJB1, FUSIP1, HSPA9B, CCT8, PTGES3, CHAF1A, SEC63, MDN1, CCT5, CCT6B*
	GO:0004842	13	ubiquitin–protein ligase activity	*ANAPC4, BRAP, TRIM23, HECTD1, UBE2F FBXO11, UHRF2, UBE2H, DZIP3, TRAF7, BARD1, UBE2D3, UBE2D2*
	GO:0006886	12	intracellular protein transport	*COPB2, AP1S3, KPNA2, SORT1, TLK1, SNX1, NAPB, STX16, TMED2, VPS45A, SSR4, SNX17*
	GO:0006457	20	protein folding	*LOC541473, CALR3, DNAJB1, HSPA9B, GNG10, CCT8, BAG2, PTGES3, HSPA4, HSPA8, NFYC, CHAF1A, SEC63, BAG5, MDN1, CCT5, CABC1, CCT6B, DNAJB6, FKBP6*
	GO:0015031	23	protein transport	*FXC1, VPS33A, RAB33B, GLE1L, SNX5, STX18, ARF1, RAB32, LRSAM1, GDI2, SEC63, RAB6B, ARF3, RAB1A, APBA3, NACA, SCAMP5, LIN7C, POM121, SNX14, AP4E1, ATG4A, EXOC6*
Transcription	GO:0006406	6	mRNA export from nucleus	*FUSIP1, UPF3A, SMG5, KHSRP, UPF3B, POM121*
	GO:0030528	12	transcription regulator activity	*FIGLA, UBA52, FALZ, NEUROG2, BRF1, PROX1, TCF3, ASH2L, HEY2, NCOA2, SMARCA1, UBC*
	GO:0006367	6	transcription initiation from RNA polymerase II promoter	*CRSP7, GTF2A2, GTF2F1, PPARBP, TBPL1, CRSP2*
	GO:0008134	7	transcription factor binding	*SMAD2, DIP2A, FALZ, HDAC2, LMO4, TRAPPC2, RAB1A*
	GO:0006357	15	regulation of transcription from RNA polymerase II promoter	*CRSP7, KLF7, SAP18, TEAD3, RBBP8, HCFC2, TARBP2, POU4F1, HIRA, HTATSF1, NFYC, POU2F3, FOXO1A, TADA3L, TAF6L*
	GO:0003702	11	RNA polymerase II transcription factor activity	*CRKRS, GTF2A2, TBPL1, TEAD3, HCFC2, HTATSF1, BRF1, NFYC, CUTL1, TAF6L, MEF2C*

Transcription	GO:0006350	100	transcription	*VDR, CRSP7, PTMA, PRDM14, PRDM8, HNRPUL1, ZNF84, RBPSUH, SMAD2, ZNF174, LEO1, NR4A2, GTF2H1, ZNF582, ZFP91, KLF7, TAF13, GTF2A2, FOXJ1, KLF16, ZNF141, GTF2F1, PPARBP, ZNF694, CDK8, PRDM16, ZNF596, RAP80, SAP18, THAP7, TEAD3, ZNF286, ZNF41, SUPT5H, TBL1XR1, SSX7, MYEF2, ESRRA, ZNF687, MKL2, HDAC2, DIDO1, HLF, KHSRP, MED28, HTATSF1, EGR1, FUBP1, ZF, HNRPK, LMO4, TRAF7, ERN2, ZNF263, NFYC, TCF3, ASH2L, CHAF1A, CCNL2, DEAF1, FIZ1, GRHL1, EGR2, ZFP95, FOXO1A, ZBTB16, TADA3L, CNOT8, RBBP4, TRAPPC2, ZFP30, ASXL1, SOX11, ZNF167, CRSP2, ZNF497, ZBTB7A, DEDD2, POLR3F, NACA, NR6A1, PCGF2, PNRC1, PPP1R10, NRBF2, THRAP2, ZNF398, CCNC, MAFF, TBX15, ZNF3, TAF6L, ZNF260, JMJD2A, SMARCA1, ING2, EAF2, HDAC6, ZNF471, MEF2C*
	GO:0003713	11	transcription coactivator activity	*CRSP7, TAF7, KLF7, GTF2F1, NFKB2, HCFC2, MKL2, NFYC, MNT, NCOA2, MEF2C*
	GO:0006355	121	regulation of transcription, DNA-dependent	*VDR, TAF5L, PRDM14, PRDM8, HNRPUL1 ZNF84, ZNF174, CDC2L1, LEO1, NR4A2, GTF2H1, TAF7, ZNF673, ZNF582, ZFP91, SCAND2, KIAA1718, TAF13, SIX5, FIGLA, GTF2A2, FOXJ1, KLF16, ZNF141, GTF2F1, SIRT5, KIAA1542, PPARBP, TBPL1, ZNF694, CDK8, PRDM16, ZNF596, RAP80, THAP7, NFKB2, TEAD3, UHRF2, ZNF286, ZNF41, SUPT5H, EMX2, TBL1XR1, SSX7, ESRRA, NCOR1, FALZ, ZNF687, MKL2, HDAC2, NEUROG2, DIDO1, MYBL2, HLF, KHSRP, MED28, EGR1, FUBP1, ZF, PHF8, HNRPK, LMO4, BRF1, TRAF7, PROX1, ERN2, ZNF263, TCF3, POU2F3, ASH2L, CHAF1A, CCNL2, SIM1, MNT, DEAF1, FIZ1 GRHL1, EGR2, ZFP95, FOXO1A, CNOT8, HEY2, RBBP4, TRAPPC2, NCOA2, STRA13, ZFP30, ASXL1, RAB1A, SOX11, ZNF167, ZNF497, DEDD2, NR6A1, TLE4, PCGF2, PNRC1, CDC5L, NRBF2, THRAP2, RPL6, ZNF398, LASS3, CCNC, HOXD1, MAFF, TBX15, RBAK, ZNF3, ZNF260, JMJD2A, SATB1, SMARCA1, ING2, EAF2, EMX1, HDAC6, ZNF471, CDC2L2, RERE, MEF2C*
	GO:0016568	9	chromatin modification	*TBL1XR1, NCOR1, HDAC2, TLK1, H2AFY2 TLK2, JMJD2A, ING2, HDAC6*
	GO:0006366	12	transcription from RNA polymerase II promoter	*TAF5L, GTF2H1, NCOR1, MYBL2, HLF, FUBP1, LMO4, ASH2L, MNT, DEAF1, MAFF MEF2C*
	GO:0003677	71	DNA binding	*THAP2, PRDM14, PRDM8, HNRPUL1, ZNF84, HIST1H1B, DMC1, ZNF582, ZFP91 NUCB2, HIST1H2BG, FIGLA, KLF16, ZNF141, GTF2F1, SIRT5, PPARBP, TBPL1 ZNF596, THAP7, PDCD8, UHRF2, NUP153 ZNF286, MYEF2, HELLS, NCOR1, ZNF687, MKL2, NCL, NEUROG2, HIST2H2AA3, SYCP1, KHSRP, KIF15, HNRPK, TMPO, PROX1, KIN, ASH2L, DEAF1, GRHL1, ZBTB16, HEY2, MBD4, ZFP30, RAD51C, SOX11, H2AFY2, ZNF497, ZBTB7A, DEDD2, POLR3F, NACA, LIG4, ZZZ3, CDC5L, PPP1R10, INOC1, ANG, SF3B2, RPL6, ZNF398, NME1, GLI4, TAF6L, ZNF260, SMARCA1, ING2, ZMYM3, ZNF471*
Cell cycle	GO:0051301	19	cell division	*ANAPC4, CDK8, STAG1, LLGL2, CCNG1, CDC2L6, SYCP1, MAD2L1, CDK6, CIT, MIS12, CCDC16, MAD1L1, CDC26, LIG4, CCNC, CCND3, CDC2L2, CDC25A*
	GO:0000074	20	regulation of progression through cell cycle	*PTMA, ANAPC4, CDC2L1, DUSP6, IGF2, CDK8, F2R, MYBL2, NFYC, CDK6, MNT, PDGFD, TADA3L, JAK2, PRKACA, PCTK1, FRAP1, CCNC, CCND3, CDC2L2*
	GO:0007049	36	cell cycle	*ANAPC4, CDC2L1, MN1, DMC1, CDK8, PRC1, STAG1, LLGL2, DBC1, CCNG1, GADD45GIP1, UBA52, SYCP1, LIN9, MAD2L1, CHAF1A, CDK6, CIT, TLK1, MIS12, RBBP4, CCDC16, MAD1L1, LIG4, TLK2, CDC5L, WWOX, CCNC, NME1, CCND3, STAG3, HDAC6, UBC, CDC2L2, CUL1, CDC25A*
	GO:0007067	10	mitosis	*CDC2L1, STAG1, CCNG1, KIF15, MAD2L1, CIT, MIS12, CCDC16, CDC2L2, CDC25A*

RNA biology	GO:0008380	8	RNA splicing	*SMNDC1, SFRS2, SAP130, KHSRP, SFRS4, SF3A2, SNRPG, SF3B2*
	GO:0003723	46	RNA binding	*RBM15, HNRPUL1, SFRS2, MATR3, RPLP0, RPS5, HNRPDL, NOM1, RPL29, AKAP1, FUSIP1, CSTF3, RPL12, RBM4, PABPC4, MYEF2, CPEB4, UPF3A, DCP2, POLDIP3, NCL, KHSRP, HTATSF1, DZIP3, IREB2, HNRPK, HNRPH1, SIAHBP1, LSM3 RPS2, RAVER2, RNH1, CPSF3, EIF2S1, SYNJ2, BARD1, RNMT, UPF3B, SNRPG, EIF4G2, CDC5L, PPP1R10, RPL6, SERBP1, RNPC3, EIF4B*
	GO:0005681	6	spliceosome complex	*SMNDC1, SAP130, SF3A2, SNRPG, CDC5L, SF3B2*
	GO:0006364	6	rRNA processing	*TRIM23, FRG1, ARF1, FTSJ1, ARF3, UTP15*
	GO:0000398	10	nuclear mRNA splicing *via* spliceosome	*SFRS2, SAP130, FUSIP1, RBM4, KHSRP, SFRS4, LSM3, SF3A2, CDC5L, SF3B2*
	GO:0006397	6	mRNA processing	*GLE1L, UPF3A, SLBP, ERN2, UPF3B, POM121*
	GO:0030529	8	ribonucleoprotein complex	*HNRPUL1, MRPS34, MRPL2, HNRPK, SLBP, SIAHBP1, LSM3, SRP19*
Other processes	GO:0006888	10	ER to Golgi vesicle-mediated transport	*COPB2, STX18, ARF1, NAPB, TRAPPC2, RAB6B, ARF3, RAB1A, TMED2, ERGIC2*
	GO:0019992	7	diacylglycerol binding	*CDC42BPA, PRKD1, DGKA, CIT, CDC42BPB, PRKACA, DGKQ*
	GO:0006091	6	generation of precursor metabolites and energy	*ECHS1, PPP1R2, PHKA2, SLC25A4, INSR, ATP5C1*
	GO:0005102	8	receptor binding	*ENSA, F2R, FIZ1, JAK2, GFRA1, GIPC1, ANG, SNX17*
	GO:0006464	9	protein modification	*PHKA2, UBL3, ICMT, LCMT1, TTLL1, UBA52, TMUB1, TMUB2, UBC*
	GO:0004674	33	protein serine/threonine kinase activity	*STK10, TRIO, CRKRS, STK32C, CDC2L1, STK25, GTF2F1, CDK8, CDC42BPA, PRKD1, CDC2L6, WNK3, MASTL, TTBK2, ERN2, MAPK14, CDK6, STK11, MAP3K7, CIT, TLK1, MAP2K2, CDC42BPB, PRKACA PCTK1, COL4A3BP, TLK2, CSNK1D, LIMK1, TAOK1, STK17B, CAMK2G, CDC2L2*
	GO:0006897	6	endocytosis	*AP1S3, ATP6V1H, ANKFY1, SORT1, SNX1 TFRC*
	GO:0005743	6	mitochondrial inner membrane	*ETFDH, SLC25A4, SLC25A11, SLC25A6, SLC25A19, UCP2*
	GO:0016874	16	ligase activity	*BRAP, HECTD1, UBE2F, TTLL1, UHRF2, UBE2H, LRSAM1, DZIP3, TRAF7, PAICS UBE2D3, ADSS, LIG4, UBE2D2, GCLC, KARS*

Table 6. Annotation of transcription start site region of selected human promoters from EPD database containing the tandem motif.

RPL12, RPS7, RPL26, ribosomal protein genes; CCNG1, cyclin G1; HNRPH3, heterogeneous nuclear ribonucleoprotein H3; ATP5F1, ATP synthase; H+ transporting, mitochondrial F0 complex, subunit B1. Atypical TATA-box motifs shown in yellow, polypyrimidine tract around TSS highlighted blue, TCTCGCGAGA motif shown in green.



RPL12	GGGCAGTGACGACAGTTCTCGCGATAGCCGCGTTTTCCTGCCTATATCTGGCTTGTCCGCGCGATTTCCGGCCTCTCGGCTTTCGGC
RPS7	CCTCCTCCTCGCGCTGTTTCCGCCTCTTGCCTTCGGACGCCGGATTTTGACGTGCTCTCGCGAGATTTGGGTCTCTTCCTAAGCCGG
RPL26	CCTCTCGCTCCGAGAGACATAGGTCTCGCGAGATCTTTGGTAAACTTACAGAACCGGAAGCAGCGTGTAGTTCTCTTCCCTTTTGCG
CCNG1	CGGCGAAAATGCCCCCTTCTCGCGAGAAAGCCCCGCCCTCCAATATATTCCTCGTTAGGGCAGGCGCGGCCCCTTCGGCTCCGAGCT
COX7B	ATTACTATAGGTTTTACAGGTATCGCGAGATTTCGTCAAATCTCATTACGGATCCCGGCTGAAAGCCATTTTGTTTTTCAGCTCACT
HNRPH3	TTTCCCGTCTCGCGAGAGTGGGGCCGGCCGCCTTCGCAGTTCTCGCTCCGCCCCCCACTTCTTGCTCGTTCCCTCCCATCCCCCCAA
ATP5F1	TTGAAGGAAGAGTACAAAATTTTCATCTCGCGAGACTTGTGAGCGGCCATCTTGGTCCTGCCCTGACAGATTCTCCATCGGGGTCAC

The analysis of gene expression profiles suggests that the motif is rather involved in a general mechanism of regulation of gene expression and is not a tissue-specific *cis* element. Although the detailed mechanism of its action remains undiscovered, we assume that it may play a role of a central GTF, alternative to the TATA-binding protein (TBP) or is a highly active enhancer element recruiting the assembly of the polymerase complex in the neighborhood of TSS and its determination may result in a development of novel therapeutic strategies (Gniazdowski & Czyz, 1999).

Acknowledgements

This work was supported by MNiSW and the European Commission (LSHG-CT-2003-503265 and LSHG-CT-2004-503567) and MNiSW grants (N401

193 32/4033, N401 050 32/1181). L.S.W. is a fellow of the Young Scientist Award from the Foundation for Polish Science.

REFERENCES

Bairoch A, Apweiler R *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res* **33**: D154–159.

Barrera L, Ren B (2006) The transcriptional regulatory code of eukaryotic cells-insights from genome-wide analysis of chromatin organization and transcription factor binding. *Curr Opin Cell Biol* **18**: 291–298.

Edgar R, Domrachev M, Lash A (2002) Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acid Res* **30**: 207–210.

Feller W (1968) An introduction to probability theory and its applications. Vol 1, 3rd edn, Wiley.

FitzGerald P, Shlyakhtenko A, Mir A, Vinson C (2004) Clustering of DNA sequences in human promoters. *Genome Res* **14**: 1562–1574.

Frech K, Quandt K, Werner T (1998) Muscle actin genes: a first step towards computational classification of tissue specific promoters. *In Silico Biol* **1**: 29–38.

Gniazdowski M, Czyz M (1999) Transcription factors as targets of anticancer drugs. *Acta Biochim Polon* **46**: 255–262.

Hardison R, Slightom J, Gumucio D, Goodman M, Stojanovic N, Miller W (1997) Locus control regions of mammalian beta-globin gene clusters: combining phylogenetic analyses and experimental results to gain functional insights. *Gene* **205**: 73–94.

Harris M, Clark J *et al.* (2004) The gene ontology (GO) database and informatics resource. *Nucleic Acids Res* **32**: D258–D261.

Haun R, Moss J, Vaughan M (1993) Characterization of the human ADP-ribosylation factor 3 promoter transcriptional regulation of a TATA-less promoter. *J Biol Chem* **268**: 8793–8800.

Hoh J, Jin S, Parrado T, Edington J, Levine A, Ott J (2002) The p53MH algorithm and its application in detecting p53-responsive genes. *Proc Natl Acad Sci USA* **99**: 8467–8472.

Hughes J, Estep P, Tavazoie S, Church G (2000) Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae. J Mol Biol* **296**: 1205–1214.

Jura J, Wegrzyn P, Jura J, Koj A (2006) Regulatory mechanisms of gene expression: complexity with elements of deterministic chaos. *Acta Biochim Polon* **53**: 1–10.

Kel A, Kel-Margoulis O, Farnham P, Bartley S, Wingender E, Zhang M (2001) Computer-assisted identification of cell cycle-related genes: new targets for E2F transcription factors. *J Mol Biol* **309**: 99–120.

Kent W, Sugnet C, Furey T, Roskin K, Pringle T, Zahler A, Haussler D (2002) The human genome browser at UCSC. *Genome Res* **12**: 996–1006.

Klimek-Tomczak K, Wyrwicz L, Jain S, Bomsztyk K, Ostrowski J (2004) Characterization of hnRNP K protein–RNA interactions. *J Mol Biol* **342**: 1131–1141.

Kozak M (1987) An analysis of 5′-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res* **15**: 8125–8148.

Makalowski W (2001) The human genome structure and organization. *Acta Biochim Polon* **48**: 587–598.

Ostrowski J, Sims J, Sibley C, Valentine M, Dower S, Meier K, Bomsztyk K (1991) A serine/threonine kinase activity is closely associated with a 65-kDa phosphoprotein specifically recognized by the kB enhancer element. *J Biol Chem* **266**: 12722–12733.

Pennacchio L, Rubin E (2001) Genomic strategies to identify mammalian regulatory sequences. *Nat Rev Genet* **2**: 100–109.

Perry R (2005) The architecture of mammalian ribosomal protein promoters. *BMC Evol Biol* **5**: 15.

Roepcke S, Zhi D, Vingron M, Arndt P (2006) Identification of highly specific localized sequence motifs in human ribosomal protein gene promoters. *Gene* **365**: 48–56.

Schmid C, Périer R, Praz V, Bucher P (2006) EPD in its twentieth year: towards complete promoter coverage of selected model organisms. *Nucleic Acids Res* **34**: D82–D85.

Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* **34**: 166–176.

Segal E, Friedman N, Koller D, Regev A (2004) A module map showing conditional activity of expression modules in cancer. *Nat Genet* **36**: 1090–1098.

Tronche F, Ringeisen F, Blumenfeld M, Yaniv M, 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.

van Ruissen F, Ruijter J, Schaaf G, Asgharnegad L, Zwijnenburg D, Kool M, Baas F (2005) Evaluation of the similarity of gene expression data estimated with SAGE and Affymetrix GeneChips. *BMC Genomics* **6**: 91.

Velculescu V, Zhang L, Vogelstein B, Kinzler K (1995) Serial analysis of gene expression. *Science* **270**: 484–487.

Wasserman W, Fickett J (1998) Identification of regulatory regions which confer muscle-specific gene expression. *J Mol Biol* **278**: 167–181.

Xie X, Lu J, Kulbokas E, Golub T, Mootha V, Lindblad-Toh K, Lander E, Kellis M (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434**: 338–345.

Yoshihama M, Uechi T *et al.* (2002) The human ribosomal protein genes: sequencing and comparative analysis of 73 genes. *Genome Res* **12**: 379–390.