Gene 545 (2014) 102–110

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Slight variations in the SC35 ESE sequence motif among human chromosomes: a computational approach

Olfa Siala [a,*], Ahmed Rebai [b], Faiza Fakhfakh [a]

[a] Laboratoire de Génétique Moléculaire Humaine, Faculté de Médecine de Sfax, Avenue Majida Boulila, 3029 Sfax, Tunisia
[b] Unit of Bioinformatics and Biostatistics, Centre of Biotechnology of Sfax, Sfax 3038, Tunisia

ARTICLE INFO

Article history:
Received 17 February 2014
Received in revised form 29 April 2014
Accepted 30 April 2014
Available online 2 May 2014

Keywords:
Splicing
SR proteins
ESE motif
ESEfinder
Epigenetic modifications

ABSTRACT

Gene expression is initiated by the binding of transcription factors to cis-regulatory modules such as enhancer elements binding to the Serine/Arginine proteins. Recently, we noticed an increased ability to identify the location as well as the motifs of enhancers using genome-wide information on spliceosomal factor occupancy, cofactor recruitment and chromatin modifications. In this study, we have undertaken a large-scale genomic analysis in an attempt to uncover if the exonic splicing enhancer motif binding to the SC35 and the SRp40 SR proteins is conserved among several groups of human genes. For the SRp40, the results showed that the ESE consensus is conserved among human genes. Concerning the SC35 SR protein, results showed an ESE motif conserved among human tissues and between different levels of muscular cell differentiation and within the same chromosome. However, this motif displays subtle discrepancies between genes localized in different chromosomes. These results emphasize the presence of different translational isoforms of the *SFRS2* gene encoding for the SC35, or different post-translational protein maturations in different chromosomes, confirming that chromatin structure is another layer of gene regulation. These links between chromatin pattern and splicing give further mechanistic support to functional interconnections between splicing, transcription and chromatin structure, and raise the intriguing possibility of the existence of a memory for splicing patterns to be inherited through epigenetic modifications.

1. Introduction

Splicing has a profound impact on gene regulatory layers, including mRNA transcription, turnover, transport, and translation. This important step begins with the spliceosome, which is assembled stepwise by the addition of discrete small nuclear ribonucleoprotein particles (snRNPs) and numerous accessory non-snRNP splicing factors (Zhang et al., 2013). The excision of introns followed by the joining of exons depends on the recognition and the usage of the 5′ and the 3′ splice sites (5′ ss and 3′ ss, respectively) by the splicing machinery (Simpson et al., 1999). However, regarding the great number of false splice sites that populate transcripts, exon recognition needs further cis acting elements that lie both within exons and in the adjacent introns: the splicing enhancers and the splicing silencers (Chasin, 2007). Exonic splicing enhancers (ESEs) are discrete (6–8 nt) degenerate sequences that constitute a binding site for the members of the Serine/Arginine (SR)-rich protein family; and it is now estimated that as many as 15–20% of randomly appearing 20-mers contain a splicing enhancer (Blencowe, 2000).

The SR proteins play important regulatory roles in the control of constitutive and alternative splicing and in the export of mature mRNA from the nucleus in a wide range of organisms (Kim et al., 2013). These factors affect splicing through both positive and negative controls of splice site recognition by the pre-spliceosomal factors and are known to promote splice site recognition throughout their Arg-Ser-rich effector domains (Wang et al., 2013). SF2/ASF, SC35, SRp40 and SRp55 are four of the best characterized among the nine human SR proteins identified to date (Zhang and Chasin, 2004). The relationship between sequence-specific binding by SR proteins and the activation of splicing by exonic splicing enhancers is complex and remains incompletely understood. A considerable effort using empirical experiments and computational predictions has been undertaken and identified short, degenerate and sometimes partially overlap motifs characteristic of these splicing regulatory elements (Wang et al., 2005).

In this study, we test by computational and statistical approaches the conservation of the SC35 and the SRp40 ESE motifs in human genes clustered upon different criteria. The SRp40 motif was found to be conserved in the 65 human genes tested here. While the SC35 consensus

Abbreviations: A, adenosine; bp, base pair(s); C, cytidine; EST, expressed sequence tag; G, guanosine; T, thymidine.

* Corresponding author.
E-mail addresses: olfa_siala@yahoo.fr (O. Siala), Ahmed.rebai@cbs.rnrt.tn (A. Rebai), faiza.fakhfakh@yahoo.com (F. Fakhfakh).

http://dx.doi.org/10.1016/j.gene.2014.04.075
0378-1119/

motif remained unchanged in genes specifically expressed in different tissues, in genes involved in different steps of muscular cell differentiation and in genes within the same chromosome, subtle variations concerning the sequence and secondary structure occurred in genes lying on different chromosomes. This result confirms that the splicing regulation process depends on chromatin conformation and remodeling, and supports the hypothesis of the interaction between chromatin and gene transcription.

2. Methods

2.1. Sequence analyses and searching for ESE motifs

Position weighted matrices, defining the consensus motif of four SR proteins (SF2/ASF; SC35; SRp40 and SRp55) were derived by the ESEfinder web-based resource generated in 2003 by Cartegni and colleagues A motif score is significant when it is greater than the threshold value defined by the ESEfinder. These thresholds were defined previously as the median of the highest score for each sequence in a set of randomly chosen 20 nucleotide sequences from the starting pool used for the functional SELEX (Cartegni et al., 2003). In this study, we used the ESEfinder$^{3.0}$ available at: http://rulai.cshl.Edu/tools/ESE/ to analyze the composition of the tested genes on ESEs.

Using the ESEfinder program, we found 3985 ESEs in the constitutive and the alternative coding exons from 65 human genes which have definite annotation in the NCBI Reference Sequence collection available at: http://www.ncbi.nlm.nih.gov. Therefore, and in the aim to avoid the false positive ESEs, we have first select among the four available matrices, only the SC35 and SRp40 ones. Secondly, and for highly significant results, we only considered the ESEs with a score ≥ 4, far above the threshold value (2.38 for SC35 and 2.68 for SRp40). Third, we only consider the coding sequences that were less than 125 nt distant from an exon junction, because it was assumed that ESE motifs were located proximally to exon junctions in order to act as ESEs (Majewski and Ott, 2002; Pettigrew et al., 2005); so that only the first 125 nt and last 125 nt of those exons were analyzed. These filters reduced the number of ESEs to 2780 ones which are potentially functional regulatory sequences.

For the verification of tissue specific expression of the chosen genes, we used the SOURCE dataset available at: http://smd.stanford.edu/cgi-bin/source/sourceSearch and the UniProtKB dataset available at http://www.uniprot.org.

2.2. Statistical procedures for the comparison of ESE motifs

Standard statistical procedures for comparing ESE motifs were performed to claim if they are significantly identical or different. First, and for each gene, we used the MEME (Multiple Em for Motif Elicitation) program version 4.3.0 available at (http://meme.nbcr.net/meme4_3_0/cgibin/meme.cgi) to test the validity of each motif for each gene. Secondly, and to claim if the 4 found motifs showed for the SC35 protein are similar or different, we developed the position weight matrices of the 4 motifs containing 821 ESEs for the G. G. C. C. C. C. T. G motif; 227 ESEs for the G. G. C. T. C. C. T. G motif found in chromosome 1, 250 ESEs for the G. A. C. C. C. C. T. G motif found in chromosome 6, and 224 ESEs for the G. A. C. T. C. C. C. G motif found in chromosome 21. These matrices were compared by a chi2-test with a H0 hypothesis: The motifs have the same distribution in a determined position of the ESE. So, for each position, H0 was rejected when p-value $< 10^{-3}$; and a difference was considered for $-\log10$ (p-value > 3).

2.3. Prediction of the pre mRNA secondary structure

In the aim to compare the pre mRNA secondary structure of the different SC35 ESE motifs, we used the MFOLD program (M. Zuker, Washington University, St. Louis, MO) (Zuker, 2003).

2.4. GC content calculation

The GC content of the tested gene on each chromosome was performed using the ENDMEMO program available at: http://www.endmemo.com/bio/gc.php.

3. Results

An extensive list of 65 human genes was chosen according to the criteria quoted in the above section (Tables 1 and 2). In addition, each gene must provide at least 10 significant ESEs to allow statistical comparisons. The ESEfinder program leads to the identification of 1522 ESEs binding the SC35 protein and 1258 ESEs specific to the SRp40 protein with a score ≥ 4.

3.1. Identification of SC35 and SRp40 ESE consensus motif

We first aimed to define a specific consensus for the SC35 and SRp40 ESE sequence for each studied gene. For example, sequence analyses of the *MAPT* gene (17q21.1), coding for the human microtubule associated-protein expressed in brain revealed the presence of 40 specific ESEs binding to SC35 with a score above 4 (Fig. 1A). Sequence alignment and position weighted matrices were derived, and according to the frequency of each nucleotide at each octamer position (Fig. 1B), a G. G. C. C. C. C. T. G motif (Fig. 1C) was found. The same strategy was used for the 65 studied genes in search for the ESE motif (Table 1). For the SRp40 protein, *MAPT* gene analyses revealed the presence of 45 specific ESEs with a score above 4 (Fig. 2A). Sequence alignment and position weighted matrices were derived, and the frequency of each nucleotide at each heptamer position (Fig. 2B) revealed a Y. C. A. C. A. G. S motif (Fig. 2C and Table 2). The same strategy was used for the 65 studied genes in search for the ESE motif (Table 2).

3.1.1. The SC35 ESE motif is conserved between tissues

In order to test if the G. G. C. C. C. C. T. G SC35 ESE motif is conserved between genes specifically expressed in different tissues, we examined genes expressed in skeletal muscle (*LAMA2*, *DYSF*, *DES*, *DMD*, *DAG1* and *SSPN*) (Fig. 3A), liver (*TF1*, *CES1*, *HNF*, *TFPI*, *TFR2* and *APOA1* genes) (Fig. 3E), pancreas (*CCKBR*, *SLC30A*, *PDX1*, *PCBD1*, *FOXA2* and *DYRK1B*) (Fig. 3F), and brain (*KCNQ2*, *SNAP91*, *KCNQ3*, *GPI19*, *CTTNBP2*, *MAPT*, *CALM3* and *SYP* genes) (Fig. 3D) (Table 2). The results showed that the same G. G. C. C. C. C. T. G motif was retrieved for all genes supporting that this motif is conserved between tissues.

3.1.2. The SC35 ESE motif is conserved among genes involved in different stages of muscle cell differentiation

Therefore, we tested the conservation of the ESE motif specific to the SC35 protein between genes specifically in smooth and skeletal muscles and also between genes in differentiated muscular cells versus more specifically genes expressed in myoblasts. We analyzed several genes expressed in smooth muscle: thyroid (*TG*, *TPO*, *NKX2*, *PAX8*, *TGOLN2* and *TSHR* genes), skeletal muscles (*LAMA2*, *DYSF*, *DES*, *DMD*, *DAG1* and *SSPN*) and myoblasts (*PEDF*, *NEF3*, *PAX7*, *MYOD1* and *PAX3*). The results showed that the SC35 ESE motif was found to be conserved between skeletal and smooth muscle genes (Fig. 3A and B), and also in genes expressed in myoblasts (Fig. 3C) (Table 1).

3.1.3. The SC35 ESE motif is conserved within the same chromosome but not between different chromosomes

The conservation of a G. G. C. C. C. C. T. G SC35 enhancer motif between the tested genes listed above led us to explore other genes expressed in the p and q arms of the same chromosome, and in different chromosomes. The same computational strategy was used to analyze several genes on chromosome 1 (*ARID1A*, *TTL7*, *PRG4*, *DNM3*, *LPHN2*, *POLZF1*, *EXTL1*, *POGZ*, *TDRD5*, *KAZ1*, *SEPN* and *SLC5A9* genes), chromosome 6 (*LAMA2*, *EYA4*, *PTPRK*, *LRRC1*, *RALBP1*, *RREB1*, *NEU1*, *PRRT* and