Conservation-based prediction of the transcription regulatory region of the \textit{SCN1A} gene

Yue-Sheng Long, Yi-Wu Shi, Wei-Ping Liao*

\textit{Key Laboratory of Neurogenetics and Channelopathies of Guangdong Province and The Ministry of Education of China, Guangzhou 510260, China Institute of Neuroscience and The Second Affiliated Hospital of Guangzhou Medical College, Guangzhou 510260, China}

Received 27 March 2009; received in revised form 20 May 2009; accepted 21 May 2009

Abstract

A challenge in identifying the transcription regulatory region is that the locations of eukaryotic transcriptional elements are often diverse among different genes. \textit{SCN1A}, a disease-related sodium channel gene, has a complex 5'-untranslated region and diverse mRNA transcripts, which might be driven by different promoters. By cross-species sequence comparison and bioinformatics analysis, human 5'-untranslated exons were found to be conserved within the region of 200 kb upstream of the 5' flanking regions of \textit{SCN1A} in higher mammals, but not in lower mammals and non-mammals. The core promoter elements (INR, DPE, and TATA) were found in the regions flanking different 5'-untranslated exons, suggesting that these sequences (from -45 to +35) might be targeted as core promoters. The nucleotide identity rate of these core promoter sequences are different, and the conservation level of the upstream region of each core promoter varies distinctly, implicating different regulatory mechanisms of the four promoters which exist in the nervous system.

© 2009 National Natural Science Foundation of China and Chinese Academy of Sciences. Published by Elsevier Limited and Science in China Press. All rights reserved.

Keywords: 5'-UTR; Gene expression; \textit{SCN1A}; Promoter

1. Introduction

The transcription regulatory region, including core promoter elements and upstream and downstream regulatory elements, plays a very important role in determining the expression pattern of a gene. Several aspects of gene expression, including the level, tissue specificity, and response to extracellular effectors (e.g., hormones), are under transcriptional control. Regulation of transcription is complex with a multitude of interactions between \textit{cis}-acting regulatory elements and \textit{anti}-acting transcriptional factors that bind these elements [1,2].

A challenge in identifying the \textit{cis}-regulatory elements in higher organisms is that the elements are often diverse among different genes [3]. The most distant element affecting gene expression locates on the site as far as 1 Mb away from the coding region [4]. Computational sequence analysis provides a conventional approach for scanning genomic sequences to identify the regions that may participate in gene regulation [5]. Sequences that mediate gene expression tend to be conserved among different species [6,7]. So, cross-species sequence comparison can be used to identify non-coding sequences that have a reasonable likelihood of having gene regulatory properties. The transcription regulatory regions can then be characterized by analyzing the conserved sequences with the databases of known transcription factor-binding sites [8–11].

\textit{SCN1A}, encoding voltage-gated sodium channel type I (Na,1.1), is an important neuronal sodium channel gene with accurate quantity being essential to normal biological
function. Mutations in the SCN1A coding region are believed to be associated with neurological disorders, including epilepsy, hemiplegic migraine and familial autism [12–14]. Some SCN1A coding-region mutations result in the loss of function of the sodium channel and induce a haploinsufficiency effect [13], suggesting that decreased Na\textsubscript{v}1.1 levels may cause these disorders, which can also be supported by the fact that SCN1A heterozygous knockout mice display seizures [15]. Whereas heterozygous mice for other neuronal sodium channel genes (SCN2A and SCN8A) showed a normal behavioral phenotype [12,16,17], and SCN3A is only expressed in the embryonic period [18], suggesting that the expression level of SCN1A is important to maintain normal biological function. Therefore, it is conceivable that variations of the regulatory region might be associated with human disorders by affecting SCN1A expression.

Comparing other sodium channel genes, SCN1A has a more complex 5'–untranslated region (5′-UTR) with alternatively spliced untranscribed exons [19,20], which is one of the reasons why the SCN1A transcription regulatory region is still poorly known. Spatial differences in the levels of expression of SCN1A and SCN2A 5′-untranslated exons (5UEs) were demonstrated in the mouse brain [19], and the 5UEs were detected to increase gene expression in cultured neuronal cells [20], suggesting that the 5UEs of SCN1A and SCN2A might play important roles in regulating gene expression. Since two SCN2A promoters located upstream of different 5UEs have been identified [21,22], it is possible that multiple SCN1A promoters might independently locate upstream of the 5UEs, which appeared as the first exons in different 5′-UTR transcripts. In order to characterize the SCN1A transcription regulatory region, cross-species comparison was performed in this study to investigate the conservation levels of SCN1A transcription regulatory regions. These results will help to guide further investigation into the biological functions of regulatory elements and the relationship between regulatory element mutation and disease.

2. Materials and methods

2.1. SCN1A 5′ flanking sequence data and nucleotide identity rate analysis of 5UEs

SCN1A 5′ flanking sequences within 200 kb upstream of the first-translated exons from different species were downloaded from the Ensembl Web site (http://www.ensembl.org/index.html). The Ensembl gene IDs of SCN1A of all investigated species are ENSG00000144285 (Homo sapiens, human), ENSECAG0000015359 (Equus caballus, horse), ENSBTAG0000018520 (Bos taurus, cow), ENSCAFG0000005058 (Canis familiaris, dog), ENSMUSG00000064329 (Mus musculus, mouse), ENSRNOG0000005989 (Rattus norvegicus, rat), ENSMLUG00000009631 (Myotis lucifugus, microbat), ENSMODG00000006537 (Monodelphis domestica, opossum), ENSGALG00000010943 (Gallus gallus, chicken), ENSXETG00000073011 (Xenopus tropicalis, tropicalis), ENSDARG0000010783 (Danio rerio, zebrafish). The Ensembl gene IDs of SCN2A are ENSG00000136331 (human) and ENSMUSG0000075318 (mouse). The sequences of humans that frequently used 5UEs (h1a, h1b, h1c, h1u, h2u, and h3u: GenBank accession numbers: DQ993522–DQ993524, EU368817–368819), and rarely used 5UEs (h1d, h1e, h1f, and h1g: GenBank accession numbers: DQ993525–DQ993527) were aligned, respectively, with the SCN1A 5′ flanking sequences of the other species with the Clustal X program of vector NTI (Informax; http://www.informax.com). The target sequences were considered as exon orthologs if the alignment e value was <10\textsuperscript{-6}, identity rate was >80%, and alignment length was >85% of the exon [23].

2.2. 5UEs conservation analysis

The genomic sequences within 3500 bp upstream and downstream of h1a (or h1u), h1b, and h1c (or h2u) each and 500 bp upstream of h3u were aligned with the vector NTI (Informax). The default alignment parameters (gap opening penalty: 15; gap extension penalty: 6.66; gap separation penalty range: 8; score matrix: swgapdnmat) were used for alignment [23]. The conservation levels of the aligned nucleotides were evaluated with the absolute complexity calculated by vector NTI software [23]. The sequences were considered as conserved non-coding regions (CNRs) if the average absolute complexity was ≥8.0 and the alignment length was ≥50 bp.

2.3. Core promoter element prediction and core promoter sequence identity rate analysis

The first nucleotides of the first 5UEs on different mRNA transcripts were designated as potential transcription start sites (TSSs, marked as “+1”). The sequences (ranging from −45 to +35) were used to predict core promoter elements according to the consensus of each element [24,25]. The core promoter sequences were aligned with Vector NTI using the default alignment parameters described earlier, and the nucleotide identity rates of these core promoter sequences between humans and other species were calculated. The difference in the nucleotide identity rate was tested with ANOVA Tukey’s test by Excel 2000.

2.4. Conservation analysis for the upstream regulatory region

The genomic sequences (5UEs plus 5000 bp upstream of 5UEs), according to a previous study [26], were aligned with vector NTI using the default alignment parameters described earlier. The method for conservation-level calculation was described earlier.
3. Results

3.1. Nucleotide identity rates of SCN1A 5'UEs between humans and other species and the organization of SCN1A 5'UE orthologs in mammals

The sequences of human 5'UEs were blasted with the regions of 200 kb upstream of SCN1A translation start sites (ATG) of all the other species, in which the full-length sequences were published. Among the human frequently used 5'UEs, the nucleotide identity rates of h1a (or h1u), h1b, and h3u were significantly high (>85%) in higher mammals (Table 1). The high nucleotide identity rate of another frequently used exon h1c (or h2u) was only found in the dog and cow. Among human rarely used 5'UEs, high nucleotide identity rates of h1d, h1e, and h1f were found in all investigated mammals, and h1g was not found in the dog. All the exon orthologs are lined in the same order on the SCN1A 5' flanking region (Fig. 1). The distances between two frequently used exon orthologs are similar to humans in the horse, dog, and rat, but significantly dissimilar to humans in the cow, mouse, and microbat (Fig. 1). The exon h3u and its orthologs are directly linked to the first-translated exons in all investigated species.

3.2. The correlation between SCN1A 5'UEs and the conserved non-coding regions (CNRs)

On the alignment of the genomic sequences around each exon and their corresponding orthologs, the most highly conserved regions within each alignment were designated as CNR1-7 (average absolute complexity ≥8.0, length ≥50 bp, Fig. 2). The 5'UE h1a (or h1u), h1b, and h1c (or h2u) were found to locate within CNRs, and h3u was found to locate overlapping exactly with CNR7. For human rarely used exons, only h1f locate within a highly conserved region CNR4. Within 3500 bp upstream and downstream of these frequently used exons, two CNRs (CNR2 and CNR5) were found located about 1500 bp downstream of 5'UEs.

3.3. Potential core promoter elements present on the regions around 5'UEs

Four human SCN1A 5'UEs (h1u, h1a, h1b, and h1c) have been identified to appear as the first 5'UEs on different mRNA transcripts till now [19,20], and the potential core promoter elements might locate around these 5'UEs. The nucleotide sequences around h1u, h1a, h1b, and h1c (from −45 to +35, named as P1u, P1a, P1b, and P1c, respectively) were predicted using consensus sequences of core promoter elements (Fig. 3a). The core promoter element INR (−5 to

Table 1
Nucleotide identity rates of SCN1A 5'UEs between humans and other species (%).a

<table>
<thead>
<tr>
<th>Species</th>
<th>h1a (h1u) b</th>
<th>h1b</th>
<th>h1c (h2u) c</th>
<th>h3u</th>
<th>h1d</th>
<th>h1e</th>
<th>h1f</th>
<th>h1g</th>
</tr>
</thead>
<tbody>
<tr>
<td>Horse</td>
<td>93.0</td>
<td>93.6</td>
<td>76.1</td>
<td>95.9</td>
<td>88.2</td>
<td>81.4</td>
<td>92.3</td>
<td>76.0</td>
</tr>
<tr>
<td>Cow</td>
<td>93.0</td>
<td>91.2</td>
<td>—</td>
<td>91.8</td>
<td>86.8</td>
<td>74.4</td>
<td>94.9</td>
<td>75.0</td>
</tr>
<tr>
<td>Dog</td>
<td>94.2</td>
<td>86.8</td>
<td>77.9</td>
<td>95.9</td>
<td>86.8</td>
<td>80.2</td>
<td>84.6</td>
<td>—</td>
</tr>
<tr>
<td>Microbat</td>
<td>93.1</td>
<td>91.2</td>
<td>—</td>
<td>91.8</td>
<td>85.3</td>
<td>82.6</td>
<td>94.9</td>
<td>76.2</td>
</tr>
<tr>
<td>Mouse</td>
<td>86.7</td>
<td>83.2</td>
<td>—</td>
<td>87.8</td>
<td>77.9</td>
<td>75.6</td>
<td>92.5</td>
<td>69.3</td>
</tr>
<tr>
<td>Rat</td>
<td>86.7</td>
<td>81.7</td>
<td>—</td>
<td>87.8</td>
<td>74.0</td>
<td>73.3</td>
<td>95.0</td>
<td>67.3</td>
</tr>
<tr>
<td>Opossum</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Chicken</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Tropicalis</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Zebralish</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

a Nucleotide identities lower than 60% are shown as "—".
b h1u coincides with h1a with a 144 bp difference in length at the 5' end.
c h2u coincides with h1c with a 47 bp difference in length at the 5' end. Human 5'UEs are classified into frequently used (with frequencies greater than 5%) and rarely used (with frequencies less than 2%) according to previous studies (Martin et al. [19] and Long et al. [20]).
+6) was found on P1u, P1a, P1b, and P1c, and DPE (+24 to +31) was found on P1u, P1b, and P1c, and a mimic TATA element (−33 to −23) was only found on P1c (Fig. 3b).

3.4. Different nucleotide identity rates of core promoter sequences and different conservation levels of upstream regulatory regions

Core promoter sequences P1u, P1a, P1b, and P1c and their orthologs from other investigated species were aligned, respectively, by Vector NTI software (Fig. 3b), and the nucleotide identity rates between humans and other species were calculated. The average nucleotide identity rates of the core promoter sequences are 93.8 ± 3.2% (P1u, n = 6), 84.4 ± 4.4% (P1b, n = 6), 89.8 ± 2.9% (P1b, n = 6), and 81.4 ± 6.9% (P1c, n = 2, two identity rates between humans and the horse, humans, and the dog), showing that the nucleotide identity rates of the core promoters in high-to-low order were P1u, P1b, P1a, and P1c. The nucleotide identity rate of P1u was significantly high (Fig. 3c) compared to that of P1a (P < 0.01), P1b (P < 0.05), and P1c (P < 0.01). On the 5000 bp regions upstream of all core promoters, the average absolute complexities all decreased gradually with the distance from TSSs increasing (Fig. 4). The average absolute complexity of the P1u (or P1a) upstream region was higher than that of P1b and P1c. The conserved region extended to 4500 bp upstream of P1u (or P1a), 1700 bp upstream of P1b and 700 bp upstream of P1c (Fig. 4, the arrows showed the rapidly decreasing sites in absolute complexity), showing that the conservation levels of the upstream regions of core promoters in high-to-low order were P1u (or P1a), P1b, P1a, and P1c.
4. Discussion

Comparative analysis of mammalian genomes has detected numerous conserved non-coding regions [27–30]. In this study, we searched all the 5UE orthologs within 200 kb upstream of the first-translated exons of SCN1A of the species and found similar organizations of 5UEs in higher mammals, but not in lower mammals and non-mammals. Although there is limited information on SCN1A 5'-UTRs of other species published on genome databases or other publications, it is possible that most of the SUE orthologs might act as 5UEs in other higher mammals according to the human and mouse SCN1A mRNA transcripts [19]. The human frequently used exons located on highly conserved regions, implicating that the exon-used frequencies are consistent with the conservation levels of 5UEs. However, the high identity rate of human frequently used 5UE h1c (or h2u) was only found in the horse and dog, suggesting that h1c might be functional only in these species. It has been reported that h1c was not found in mouse SCN1A mRNA transcripts identified by 5 RACE experiments [19], consistent with our hypothesis. Since those non-coding sequences conserved in a limited number of mammals are frequently functional [31] and SCN1A 5UEs did alter gene expression in cultured neuronal cells [20], it is suggested that the diversity of 5UEs might accommodate different regulatory mechanisms. Further studies are needed to investigate the relationship between 5UEs evolution and the change of the regulating mechanism.

We also found two CNRs (CNR2 and CNR5, Fig. 2) at the sites about 1500 bp downstream of h1a and h1b. These
CNRs may appear as SUEs that are not identified by experiments till now. More attention should be paid to investigate their roles in regulating gene expression in vivo.

The core promoter region can usually be identified according to the location of the TSS [32]. For human SCN1A, four types of 5'-UTR transcripts were identified from different tissues [19,20], and four SUEs (h1u, h1a, h1b and h1c) appeared as the first exon in different transcripts. Distinct expression patterns of SCN1A SUEs in the mouse brain implicated that similar expression patterns might occur on human SCN1A transcripts, raising the possibility of distinct promoters in human SCN1A. Here, we identified core promoter elements around h1u, h1a, h1b, and h1c (Fig. 3a and b), suggesting at least four core promoters (P_{1u}, P_{1a}, P_{1b}, and P_{1c}) of human SCN1A. Among the nucleotide identity rates of these core promoters, P_{1u} is the most conservative and P_{1c} is the least conservative, indicating that these core promoter sequences evolved at different rates. Recent studies found that core promoter sequences evolved at an accelerated rate compared with the neutral controls [33,34], which might accommodate distinct gene expression patterns in the nervous system and partially contribute to the evolution of species.

There is no criterion to determine the length of the upstream regulatory region used for promoter functional nested-deletion analysis, because the locations of upstream regulatory elements vary significantly among eukaryotic genes. Here, we used cross-species sequence comparison to investigate the conservation of SCN1A upstream regulatory regions for each core promoter and found that distinct lengths of conserved sequences vary significantly among these sequences (Fig. 4). A similar difference in conservation level was also found on the upstream regions of SCN2A promoters (unpublished data). Conservation analysis of upstream regulatory sequences could be used to determine the length of the promoter upstream region for nested-deletion analysis, because evolutionary conservation has been considered as an important clue for identifying regulatory elements [5,7,35–38].

Many disorders are associated with mutations in functional regulatory elements [39–42]. It is highly possible that DNA variants on SCN1A regulatory elements might also be associated with neurological disorders since SCN1A haploinsufficiency resulting from coding-region mutations [13] and SCN1A knockout mice [15] induces disorders. The present study demonstrated that SCN1A SUEs, core promoters, and upstream regulatory regions are conserved in higher mammals, which implicates that the DNA variants on these regions from patients might be associated with diseases. It is possible that variants on different promoter regions of SCN1A might result in different types of disorders, because these promoters might drive SCN1A expression in distinct patterns according to the evolution rates of promoter sequences and different expression patterns of SUEs [19]. Further studies are needed to identify the functional elements from the SUEs, core promoters, and upstream regulatory regions and screen the variants on these functional elements from patients to explore the relationship between the regulatory element mutation and disease.
Acknowledgements

We thank Dr. Zao C. Xu and Dr. Zhongmin Yuan for critically proofreading the manuscript and significant suggestions. We are grateful to the He Shanheng Charity Foundation for contributing to the development of our institute. This work was supported by the National Natural Science Foundation of China (Grant No. 30600198), The Guangdong Natural Science Foundation (06301101), and The Science and Technology Projects of Guangzhou Municipal College (08A047).

References

[26] Drews VL, Lieberman AP, Meisler MH. Multiple transcripts of sodium channel SCN5A (Na1.6) with alternative 5′ and 3′-untranslated regions and initial characterization of the SCN5A promoter. Genomics 2005;85(2):245-57.