COMPARATIVE COMPUTATIONAL ANALYSIS OF 5'-REGION OF CYTOPLASMIC TYROSYL-TRNA SYNTHETASE GENE IN HIGHER EUKARYOTES


Institute of Molecular Biology and Genetics, National Academy of Sciences of Ukraine, Kyiv, Ukraine, e-mail: kornelyuk@imbg.org.ua
*Corresponding author

Key words: orthologous genes, housekeeping genes, transcription factor binding site, promoter model and module, CpG islands, Scaffold/Matrix Attachment Regions

Resume

Motivation: Comparative analysis of 5'-regions of orthological housekeeping genes by computational tools allows estimating quickly their conservative regulatory pattern and revealing the differences that have emerged during evolution.

Results: The structure of 5'-region of TyrRS gene was examined in four eukaryotic organisms in various aspects. In general it shows the typical features of housekeeping genes. Orthological sequences lack TATA box in appropriate position. In the case of human and mouse they are located within CpG islands and include Sp1 sites, what is expected for vertebrates. The approximate boundaries of the promoter zone have been defined for each 5'-region. Conservative promoter framework has been determined for orthological TyrRS genes by computational generating of four promoter models. One common hexameric palindrome has been predicted. One Scaffold/Matrix Attachment Region has been found in 5'-region of mouse gene. This work is a striking instance of testing the bioinformatics tools on the specific biological objects.

Introduction

The comprehensive study of the 5'-region of eukaryotic genes is of a great importance in functional gene analysis. 5'-region of eukaryotic gene encompasses the first exon beginning and regulatory structures located either in transcribed or untranscribed gene sequence. The promoter zone, represented by a number of various transcription factor binding sites (TFS), must be there necessarily. Scaffold/Matrix Attachment Regions (S/MAR), by which chromatin DNA is anchored on nuclear matrix, is also expected in 5'-region, because such physical contact is required for the initiation of transcription. It was demonstrated that a range of crucial nuclear processes, such as replication and recombination, are initiated at S/MAR as well (Travers, 1994; Bode et al., 1992). Virtually, each S/MAR is a complex of various consensuses, such as Ori pattern, topoisomerase II sites, AT-richness, kinked DNA patterns. Palindromic motifs can be also expected in the 5'-region, as they form hairpins or loops that facilitate protein binding, active in the sites of transcription, replication or recombination.

In our study we have performed thorough online computational analysis of the 5'-region of cytoplasmic tyrosyl-tRNA synthetase (TyrRS) gene in four eukaryotes: human (Homo sapiens), mouse (Mus musculus), fruit fly (Drosophila melanogaster) and nematode (Caenorhabditis elegans). The major purpose was to learn, how well its structure corresponds to the expected for eukaryotic housekeeping genes and what regulatory features are the most conservative in orthological sequences.

Methods and Algorithms

Genome sequences used for TyrRS gene analysis. All the sequences were found from the NCBI GenBank database. H. sapiens: AL356459, chromosome 1, contig NT_004511.8, HTGS_PHASE1; M. musculus: AL607123, Chr. 4, HTGS_PHASE1; D. melanogaster: AE003527 genomic scaffold 14200013386050, Chr. 3L; HTG (HTGS_PHASE3). Gene is completely annotated as CG4561; C. elegans: AL132880, cosmid Y105E8E, Chr. 1; HTG (HTGS_PHASE3). Gene is annotated as Y105E8E.v.


Implementation and Results

Nucleotide sequences of human and mouse orthological TyrRS genes were obtained by genome BLAST alignment with the respective TyrRS cDNA. As we hadn’t found any fly and nematode cDNA, annotated as TyrRS cDNA, we had to submit human cDNA for their TyrRS gene search, expecting their high homology. As a result, we found High Throughput Genomic Sequences (HTGS) of the first stage of progress for human and mouse genes, and of the third stage for fly and nematode.

Then we calculated the most probable transcription start site (TSS) for each orthological TyrRS gene. It was found as the average beginning of all expression sequence tags (EST) well aligned. In this study we didn’t trust the longest EST variants, considering them as the products of alternative splicing or background transcription. For the next analysis we tried to cut out the equal genomic regions, so that calculated potential TSS were at the distance from 200 to 250 bp from the end of such fragments and corresponded to the sense strand of orthological TyrRS genes. Consequently, DNA sequences 1351, 1350, 1420 and 1430 bp long were obtained for human, mouse, fly and nematode respectively and then analyzed. Defining the boundaries of the potential promoter, we applied the original approach. Predictions of different search programs were overlapped for the sequence of each organism, and the overlapping of at least three program predictions, adjacent to the potential TSS, was considered as a potential promoter. We always restricted the search by the promoter library of correspondent organism and applied default cutoffs. Calculated promoter zones were positioned relatively to potential TSS. As it is shown in the Table 1, found promoter regions are at the large distances from TSS, which contradicts to the idea of minimal promoter in the close vicinity to TSS. However, we think these regions should significantly overlap with real promoters.

Table 1. Defining the boundaries of the potential promoter zones.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Number of used programs</th>
<th>Overlapping length, bp</th>
<th>Distance to average TSS, bp</th>
<th>Maximal number of overlappings achieved within the region of three overlappings</th>
</tr>
</thead>
<tbody>
<tr>
<td>H. sapiens</td>
<td>9</td>
<td>328</td>
<td>+97</td>
<td>6</td>
</tr>
<tr>
<td>M. musculus</td>
<td>9</td>
<td>334</td>
<td>+168</td>
<td>5</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>7</td>
<td>267</td>
<td>-88</td>
<td>4</td>
</tr>
<tr>
<td>C. elegans</td>
<td>6</td>
<td>185</td>
<td>+450</td>
<td>4</td>
</tr>
</tbody>
</table>

To reveal conservative regulatory patterns in orthological promoter sequences we applied the strategy of framework search in a set of sequences (Werner, 2000). As a result, four 3-component promoter models were generated computationally. Promoter models are combinations of TFS with definite order and relative distances. In the case of analyzed sequences, these patterns are common for human, mouse, fly and nematode 5'-regions. Specificity of the promoter models (framework score – FS-Score) ranged from 0.44 to 0.88. The model with the highest FS-Score is represented in Table 2.

Table 2. 3-component model with the highest FW-score generated by Genomatix ModelGenerator. Matrix similarity reflects the probability of each element. Distance ranges are caused by different variants-matches of this promoter model.

<table>
<thead>
<tr>
<th>Nr</th>
<th>Element</th>
<th>Strand</th>
<th>Matrix similarity</th>
<th>Distance to next element</th>
<th>FW-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Human and murine ETS1 Factor</td>
<td>±</td>
<td>Optimized (min. 0.87)</td>
<td>25–86 bp</td>
<td>0.80 / 0.80</td>
</tr>
<tr>
<td>2.</td>
<td>E2F-myc activator/cell cycle regulator</td>
<td>±</td>
<td>Optimized (min. 0.78)</td>
<td>11–73 bp</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>Signal transducer and activator of transcript. Factor</td>
<td>±</td>
<td>Optimized (min. 0.73)</td>
<td>–</td>
<td></td>
</tr>
</tbody>
</table>

The set of promoter models was used to scan Eukaryotic Promoter Database with input cutoff 0.95 and resulted by 13 matches. Interestingly, those six matches were annotated as adenovirus promoters. Independently, we compared the analyzed human, mouse, fly and nematode sequences to the database of promoter modules, which are conservative pair combinations of TFS. Although no module, common for all four sequences, was obtained, three modules were common for human and mouse: ETSF_ETS1F_01; SP1F_SP1F_01 and SP1F_CEBP_01. Notably, there were overall 3 copies of Sp1 site in human modules and 7 copies in mouse modules. It is well known that Sp1 is a typical TFS for promoters of housekeeping genes in vertebrates. (Mudge et al., 1998).
Analysis of C+G content resulted by characterization of human and mouse sequences as CpG islands with such properties: C+G content was 67 and 68%; CpG content was 7 and 8% in human and mouse respectively. Along with the copies of Sp1 site this is expected for the promoters of housekeeping genes in vertebrates (Pedersen, 2001).

The orthological 5'-regions were then submitted to TATA box search by specialized program. Besides, the sequences were analyzed for the presence of all TFS. None of these programs demonstrated appropriately positioned (about +30 bp) TATA box, which was expected for the promoters of eukaryotic housekeeping genes. However, two other conservative proximal promoter elements – CCAAT box, GC box – were located in one copy within the region [-200- +200] bp in human and mouse sequences, but in (−)-orientation. While searching the most complex common palindromes we found the single hexameric structure GGANNNTCC with such positions relatively to the calculated TSS: -32 bp in human; -142 bp/mouse; -810 bp/fly; -275 bp/nematode.

At last, we characterized S/MAR consensuses in orthological sequences. Fly and nematode differ by the larger number of potential Ori patterns (15 and 17, respectively) in comparison to human and mouse (3 and 1, respectively), and nematode sequence is AT-enriched (161 matches in contrast to 15 ones in fly and no matches in others). There were 3 matches of topoisomerase II site in human; 1 in mouse; 4 in fly and 6 in nematode. Kinked DNA was found 2 times in human and mouse, 4 in fly and 2 in nematode. Based on S/MAR features the program MAR-Wiz predicted potential S/MAR region, but we succeeded only with mouse sequence. Predicted S/MAR was 201 bp long and positioned at -382 bp relatively to the potential TSS.

Discussion

With the wide range of programs, we have characterized such features of 5'-regions of orthological TyrRS genes as promoters, palindromes, CpG islands and S/MAR patterns in comparative manner and concluded about their functional correlation. One S/MAR was predicted in mouse 5'-region, but all analyzed sequences had a number of other S/MAR features: topoisomerase II sites, Ori pattern, AT-richness, kinked DNA. Orthological promoters have demonstrated typical properties of the promoters of housekeeping genes, which lack TATA box and in the case of vertebrates contain CpG islands and Sp1 sites. However, CCAAT and GC boxes were located close to TSS in human and mouse sequences, what is unusual for eukaryotic housekeeping genes. Despite the evolutionary divergence, these 5'-regions have retained the conservative framework of transcriptional regulation, represented in our study by 4 promoter models. The latter have proved to be close to adenovirus promoters, what may elucidate common patterns of their regulation. Since the analysis was based on the computational tools, our results indicate the potential structures of the 5'-regions and need experimental support. The composition of promoter models also needs deeper analysis.

References