Prediction of Structured Non-Coding RNAs in the Genomes of the Nematodes Caenorhabditis elegans and Caenorhabditis briggsae

Kristin Missal, Xiaopeng Zhu, Dominic Rose, Wei Deng, Geir Skogerbø, Runsheng Chen and Peter F. Stadler

Bioinformatics Group, Department of Computer Science, University of Leipzig, Germany

Email: kristin.dominic@tuda@bioinf.uni-leipzig.de
WWW: http://www.bioinf.uni-leipzig.de

Biological Research Group, Key Laboratory of Biophysics, Chinese Academy of Sciences, China

Biological Research Group, Key Laboratory of Information Processing, China

Chinese National Human Genome Center, China

Department of Theoretical Chemistry, University of Vienna, Austria

The Santa Fe Institute, USA

Motivation

The analysis of animal genomes showed that only a minute part of their DNA codes for proteins. Recent experimental results agree, however, that a large fraction of these genomes is transcribed and hence is probably functional at the RNA level [4]. A computational survey of vertebrate genomes has predicted thousands of previously unknown non-coding RNAs (ncRNAs) with evolutionary conserved secondary structures [7]. An extension of these comparative studies beyond vertebrates is difficult, however, since most non-coding RNAs evolve relatively fast at the sequence level while conserving their characteristic secondary structures. Hence, independent screens in invertebrates are necessary. A first ncRNA prediction approach among urochordates revealed some thousand putative structured RNAs [5]. Here we extend the phylogenetic range of systematic surveys for ncRNAs to the nematodes C. elegans and C. briggsae. Upon the genome release we decided to use those non-coding DNA in C. elegans genome:

Contiguous regions except protein-coding and repetitive elements define putative nc DNA.

We identify conserved non-coding DNA regions between C. elegans and C. briggsae by blast alignments (E < 10−7). Hits with short distance between are combined considering consistency checks:

Global alignments of the resulting regions are computed using clustalo. They are screened with BLAST [8] to detect regions that are also conserved at the secondary structure level. The BLAST algorithm evaluates thermodynamic stability and the evolutionary conservation of secondary structure.

tucre. Evolutionary conserved secondary structure indicates functional significance and a z-score of thermodynamic stability relative to an ensemble of shuffled sequences evaluates if the potentially transcribed RNA is more stable than by chance. For each global alignment, both possible reading directions are considered, because calculating thermodynamic energy is direction dependent.

Upcoming statistical values describe the number of the genomic loci in C. elegans.

Results

We detect 3672 structured RNA motifs, of which only 678 are known ncRNAs or clear homologs of known RNAs. Most of these signals are located in introns or at a distance from known protein-coding genes.

<table>
<thead>
<tr>
<th>Genomic Blast</th>
<th>Number of ncRNA context alignments</th>
<th>length</th>
</tr>
</thead>
<tbody>
<tr>
<td>intronic</td>
<td>597,128</td>
<td>1235</td>
</tr>
<tr>
<td>5'UTR</td>
<td>116,193</td>
<td>119</td>
</tr>
<tr>
<td>3'UTR</td>
<td>128,766</td>
<td>130</td>
</tr>
<tr>
<td>intergenic</td>
<td>810,989</td>
<td>1221</td>
</tr>
<tr>
<td>total</td>
<td>3672</td>
<td>2366</td>
</tr>
</tbody>
</table>

Phylogenetic classification of the nematodes C. elegans and C. briggsae: green numbers represent the amount of predicted ncRNA candidates. The sensitivity of these predictions is based on known ncRNA annotations from the Wormbook [6]. We compare the numbers of predictions with those classed as structured RNAs by RNAz [5].

Comparison of the BIYA results with experimentally validated ncRNAs [5].

[2] Columns have the same meaning as above.

Annotation

Deng et al. identified three putative RNA-specific promoter sequences, denoted by UMI, UMI2 and UMI3. They form bulge-bulge RNAs and are associated with our ncRNAs. UMI1 (96 hits) covers snRNA loci and includes the C. elegans proximal sequence element (PSE), UMI2 (413 hits) was mainly found upstream of snoRNA genes. However, it is similar to the internal RNA promoter and thus comprises RNA loci. UMI3 (7 hits) covers snoRNA, RNase P and 5 functionally unassigned loci.

Furthermore, Deng et al. identified a class of snRNA-like ncRNAs characterized by a recognizable SMN-binding site. We use BLAST to search for the sequence motif AUUUGG followed by a hairpin of rather variable stem and loop length, a common generalization of SMN binding sites in known snRNAs. We require that the pattern coregiously occurs in aligned positions of C. elegans candidate ncRNA sequences. This procedure recovers 122 loci of which more than 60 are plausible snoRNA candidates (among others we count 9 U1, 19 U2, 5 U4, and 12 US loci).

Possible novel microRNA precursors are either identified by manual filtering of the BLAST-based predictions or by running RNAmicro [3] on the input alignments. RNAmicro works in spirit of BIYA, but especially is trained to detect microRNA precursors.

References


Printed by Universitätsexekutivkommission Leipzig