Stable RNA secondary structure of human donor splice sites

Dmitri Pervouchine†*, Joel Graber†, Simon Kasif†

Secondary structure of RNA plays a key role in many biological processes. For instance, initiation and termination, at the level of transcription, and attenuation, at the level of translation, both involve formation of important stem-loop regions, which are crucial for the further RNA processing. The question of whether the secondary structure in pre-mRNA also plays a role in splicing is debatable. In several cases, experimental evidence for facilitation or inhibition of splicing has been documented. However, recent studies indicate that, in general, splice sites have no universal conserved pattern in secondary structure. It can be explained by the weakness of the splicing signals and the high variance of secondary structure prediction.

Using decision tree and support vector machine classifiers, it was shown recently that splice sites in human preferentially exhibit short helices at the acceptor site [1]. In the present work, we attempt to view this problem in a different light by considering not all, but only stable secondary structures. To evaluate the stability, we introduce the following stability measure

$$\varepsilon_{ij} = \frac{E_{ij}}{|j - i|}$$

for a region of RNA between bases $i$ and $j$ with optimal fold energy $E_{ij}$ predicted by Vienna package [2]. This quantity has physical units of the elasticity constant and is referred to as the stability factor or elasticity. With a suitable scaling factor, $\varepsilon_{ij}$ is a number between 0 and 1. Setting different thresholds for elasticity, i.e. discarding regions having elasticity less than a given value, allows us to view the secondary structure at different levels of stability.

The hypothesis we tested in this paper was whether stable elements of secondary structure (namely, stem-loops) occur more often at splice sites compared to the rest of the sequence. The clean dataset of EST-confirmed splice sites from human was obtained from [3]. The positive (+) set consisted of 400 donor splice sites not known to be involved in alternative splicing events.

†Bioinformatics Program, Boston University, 44 Cummington street, Boston MA 02215.
*To whom correspondence should be addressed. E-mail: dp@bu.edu
The negative (-) set consisted of 400 sequences that mimic donor splice sites (pseudo splice sites). It was constructed by taking together internal parts of exons and introns. All sequences had length 280 bp and were centered at the starting GU dinucleotide of an intron. Three subsets of approximately 100 sequences having donor consensus GU, GURA, GUNA respectively were sampled from these sets. The frequencies of a stem-loop structure being observed within 12 bp from the splice site for these three subsets can be summarized here in the following table:

<table>
<thead>
<tr>
<th>Ellasticity threshold</th>
<th>0.51</th>
<th>0.66</th>
<th>0.68</th>
<th>0.71</th>
<th>0.75</th>
</tr>
</thead>
<tbody>
<tr>
<td>Donor site</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>GU</td>
<td>0.97</td>
<td>0.97</td>
<td>0.86</td>
<td>0.82</td>
<td>0.38</td>
</tr>
<tr>
<td>GUNA</td>
<td>0.96</td>
<td>0.88</td>
<td>0.73</td>
<td>0.62</td>
<td>0.28</td>
</tr>
<tr>
<td>GURA</td>
<td>0.99</td>
<td>0.93</td>
<td>0.79</td>
<td>0.58</td>
<td><strong>0.27</strong></td>
</tr>
</tbody>
</table>

The most effective discrimination between actual and pseudo splice sites was achieved at the threshold value of 0.68. This difference is even more striking if we use both secondary structure and sequence information. Although stem-loops, which have $\varepsilon \geq 0.68$, were observed in vicinities of actual donor sites twice as often as in pseudo donor sites, the overall frequency of their occurrence was low and, as a consequence of that, the sensitivity of the method was also low.

The method described in the poster gives another perspective on the interpretation of secondary structure predictions and can be potentially used in gene finding to reduce the rate of false positive predictions.

References

