Databases and ontologies

SpliceAid: a database of experimental RNA target motifs bound by splicing proteins in humans

Francesco Piva*, Matteo Giulietti, Linda Nocchi and Giovanni Principato
Institute of Biology and Genetics, Polytechnic University of Marche, 60131, Ancona, Italy
Received on October 14, 2008; revised on January 29, 2009; accepted on February 27, 2009
Advance Access publication March 4, 2009
Associate Editor: Alfonso Valencia

ABSTRACT
Summary: The correct post-transcriptional RNA processing is finely regulated by RNA-binding proteins. Unfortunately, there is little experimental information on target RNA sequences of RNA-binding proteins and moreover such experimentally derived target sequences are annotated in a compact form by the score matrices that overestimate the number of possible recognized sequences. We carried out an exhaustive hand curated literature search to create a database, SpliceAid, collecting all the experimentally assessed target RNA sequences that are bound by splicing proteins in humans. We built a web resource, database driven, to easy query SpliceAid and give back the results by an accurate and dynamic graphic representation.
Availability: SpliceAid database is freely accessible at http://www.introni.it/splicing.html
Contact: f.piva@univpm.it

1 INTRODUCTION
The pre-mRNA sequences of eukaryotes harbour different languages: not only the one specifying the amino acid sequence, but also those regulating the splicing process, RNA folding, capping, polyadenylation and nuclear export. Particularly, the pre-mRNA splicing is driven by proteins that recognize and bind specific target RNA sequences and by the proteins recruited from the former. DNA mutations can affect the splicing process by destroying or creating target sequences for the proteins thereby changing the complex of proteins bound to the pre-mRNA. Among splicing regulatory proteins, some marking 5′ of proteins bound to the pre-mRNA. Among splicing regulatory proteins, some marking 5′ of proteins bound to the pre-mRNA and 3′ splice sites, polyadenylation site and branch point, others acting as exonic splicing enhancer (ESE), exonic splicing silencer, intronic splicing enhancer (ISE) or intronic splicing silencer (ISS) (Blencowe, 2000; Cartegni et al., 2002).

To explain the mutation effects on the splicing process, it is important to know the pattern of proteins bound to the pre-mRNA and then to take into account all the cited dynamics.

Web tools exist that predict the target RNA sequences of splicing proteins in humans, for example, RESCUE-ESE (Fairbrother et al., 2002) identifies candidate ESE s as hexanucleotides significantly enriched in human exons and/or located with a significant frequency in the vicinity of weak 3′ and 5′ splice sites. It shows the position of the predicted enhancer sequences but it does not show the kind of proteins bound to the found targets and the score. It gives many false positive results because its algorithm does not use experimentally assessed RNA target sequences. ESEfinder 3.0 (Cartegni et al., 2003; Smith et al., 2006) predicts the targets only of five out of about 50 proteins nowadays known to regulate splicing. In a recent work (Skoko et al., 2008), the ESEfinder predictions in human neurofibromin 1 (NF1) exon 37 are experimentally checked. ESEfinder predicted the SF2/ASF, SC35, SRp40 and SRp55 protein binding sites, instead UV cross-linking and immuno-precipitation analyses showed that none of the predicted SR proteins bound the exon. This overestimation is very probably due to the use of the score matrix formalism to store the target sequences in ESEfinder database. Also Splicing Rainbow (Stamm et al., 2006) adopts the score matrix formalism despite its database contains target RNA sequences of many proteins.

The score matrices are simple to use and can increase the sensitivity of a prediction tool although false positives could arise. To understand the creation of a false positive binding site due to the score matrix notation, it can think that the experimental binding sites of a hypothetical experiment gives the sequences AGGT, AGGA and CGGA. Summarizing them by score matrix notation and then extracting the consensus site, we obtain (A/C)GG(A/T) and this information is usually stored in the databases of the prediction tools. Rebuilding sequences contained within the consensus site obtained by multiplying the permutations, a new sequence (CGGT) appears that does not belong to the starting set. So, a prediction tool using this formalism will report a binding in connection with this false site. This spurious sequence is due to the method of generation of the score matrix that takes into account only the nucleotide frequency independently from its position, but it does not take into account the nucleotide correlations among the different positions of the experimentally assessed sequences. In another work (Raponi et al., 2007) the number of predicted ESE and silencer motifs using ESEfinder and RESCUE-ESE tools was compared with human CFTR exon 12 inclusion efficiency resulting from 22 single and 55 multiple synonymous substitutions. Unfortunately, the predictions did not correlate with the splicing efficiency indicating that, in that context, the used in silico programs are not useful to predict the effect of single or multiple mutations. These comparative studies show that up till now the regulatory splicing sequences prediction tools give not very reliable results and therefore they are not very useful to foresee mutations or polymorphisms effects.

In order to improve the reliability of the current prediction programs and to provide the researchers an exhaustive collection of primary data, we have carried out an exhaustive literature

*To whom correspondence should be addressed.

© The Author 2009. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org
We have created a database, SpliceAid that collects all the experimentally assessed target RNA sequences that are bound by splicing proteins in humans.

2 METHODS

Human RNA binding proteins were prepared from the UniProt database (Release 14.2) by entering ‘organism: “Human” [9606]’ and keyword: ‘RNA-binding [694]’. From the obtained items we picked out the pre-mRNA binding proteins certainly involved in the splicing process rejecting the proteins with a notation like ‘May be involved in pre-mRNA splicing’. Some of the remained proteins were rejected because they have helicase activity or help the folding so they bind RNA in a sequence-independent manner. Finally, from a literature review on the filtered proteins, we found the experimentally assessed target RNA sequences of 54 proteins.

We carried out various searches in the PubMed resource by entering each found protein name, including its synonyms, joined with the word ‘splicing’; subsequently the same names were joined with the word ‘binding’. The exhaustive, expert curation, literature review was a crucial part of the work and it was a time consuming and laborious process in which the kinds of experiment to assess the binding were taken into account. The list of the revised articles is available by a link in the SpliceAid web page. Information about experimentally assessed target sequences of the pre-mRNA-binding proteins in humans is stored in the database on our server. A database record includes the target RNA sequence, the protein that binds it, the EntrezGene identifier of the gene encoding the protein, the binding score, the experiments used to assess the binding (EMSA, SELEX, UV cross-link followed from western immunoblot, competition assay, RNAi, ...), the conditions (cell lines, transfection, nuclear extract, ...), the RNA substrate and the references from which all these data were extracted.

Binding quality levels were assigned based on the binding evidence. A positive and negative score was assigned, respectively, to the target sequences that facilitate exon or intron definition. So, ESE and ISS motifs were associated with a positive score. According to the same criteria, exonic splicing silencer (ESS) and ISE motifs were associated with a negative score.

Regarding the extraction of the target RNA sequences of a protein from the literature, where the simple binding to only one sequence was assessed, we attributed a middle score (i.e. 5) to the target because information is unable to evaluate its strength compared with other targets. SELEX experiments can give target RNA sequences with occurrence more than one thus these assays supply enough information to infer that the more a winner sequence occurrence is high the more its affinity is high. We linearly mapped the occurrence of each winner sequence to have a minimum value of 0 and a maximum of 10. Subsequently, the value is set by dividing into bins the occurrence of each winner sequence shown from the experiment for example: if the value of a target is $\geq 0$ and $\leq 1$ we attribute a score of 1, if the value is $> 1$ and $\leq 2$ its score is 2 and so on. We established a score scale with a resolution equal to 1 because generally a SELEX experiment gives no more than some tens of winner sequences so the resolution of the mapped occurrences is not lower than 1, therefore information is lost using a score scale of resolution equal to 1.

After we established the guidelines about literature search and the records annotation, we performed an inter-annotator analysis (Colosimo et al., 2005) carried out by two PhD students in biology and a researcher in order to assure the annotation quality. This check allowed us to refine the annotation guidelines and therefore to improve the database quality.

For storage and query of the data we constructed a MySQL relational database. The web interface uses Perl DBI and CGI to connect with the database, facilitate the visualization and retrieval of the data.

3 RESULTS AND DISCUSSION

We have created a database of strictly experimentally assessed target RNA sequences that are bound by splicing proteins in humans. At present, the total number of target sites stored in SpliceAid is larger than 680 corresponding to 54 proteins: 9G8, CUG-BP1, DAZAP1 (or PRP22), ETR-3, Fox-1, Fox-2, hnRNPA0, hnRNPA1, hnRNPA2/B1, hnRNPC, hnRNPC1, hnRNPC2, hnRNPD, hnRNDD0, hnRNDDL, hnRNPE1, hnRNPF, hnRNPH1, hnRNPH2, hnRNPI (or PTB), hnRNPIJ, hnRNPK, hnRNPL, hnRNPM, hnRNPN, hnRNPU1, HTra2alpha, HTra2betalpha, HuB, HuD, HuR, KSRP, Nova-1, Nova-2, nPTB, PSF, Sam68, SC35, SF1, SF2/ASF, Sm-1, Sm-2, Srp20, Srp30k, Srp38, Srp40, Srp54, Srp55, Srp75, TDP43, TIA1, TIAL1, YB-1. Unfortunately, in the case of Srp38, the authors claimed to have performed SELEX experiment but they did not publish each winner sequence (Shin and Manley, 2002).

User is allowed to query the database by protein name or by submitting a sequence being analyzed with respect to database target RNA sequences. SpliceAid returns the results by histogram in which each target is represented by a colored bar. The bars have variable width and height, respectively, related to the number of nucleotides of the identified target and to its score. A positive score is associated to the target sequences that facilitate exon definition that is ESE and ISS motifs. According to the same criteria, a negative score denotes the target sequences that facilitate intron definition that is ESS and ISE motifs. Moving the mouse over the factor’s name label, a window will appear showing the references that experimentally support the connection of that target RNA sequence with its factor.

In order to interpret the results, researchers have to take into account that the score is deduced based on a specific experimental condition so the same binding efficiency cannot be guaranteed in other contexts. For example, the secondary RNA structure plays an important role in splicing (Hiller et al., 2007; Pagani et al., 2003; Shepard and Hertel, 2008) because most of splicing factors bind only single strand RNA (Auweter et al., 2006; Buk wanna and Darnell, 1997; Skrisovska et al., 2007). In other words, the RNA structure could prevent a signaling binding.

This resource provides primary data useful to bioinformaticians to improve prediction programs because it allows to distinguish the experimental data from the predictions, moreover it is a dataset suitable to train machine-learning algorithms. The biologists can utilize SpliceAid to understand splicing pattern alterations or to explain the effect of the mutations on the target RNA sequences and thus the protein complex that drives the RNA maturity.

We plan a continuous updating of our database also by the collaboration of the users that are invited to report new target RNA sequences or signal annotation slips. To facilitate such cooperation, we have predisposed forms to easy submit the suggestions.

Funding: Polytechnic University of Marche.

Conflict of Interest: none declared.

REFERENCES


Colosimo, M.E. et al. (2005) Data preparation and interannotator agreement: BiCraAteI task 1B. *BMC Bioinformatics*, 6 (Suppl. 1), S12.


