BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 7 2012, pages 1031–1032
doi:10.1093/bioinformatics/bts074

Sequence analysis

Advance Access publication February 10, 2012

Spliceman—a computational web server that predicts sequence variations in pre-mRNA splicing

Kian Huat Lim[1] and William Guy Fairbrother[1,2,*]

[1]Department of Molecular Biology, Cellular Biology and Biochemistry, Brown University, Providence, RI 02903 and [2]Center of Computational Molecular Biology, Brown University, Providence, RI 02012, USA

Associate Editor: Ivo Hofacker

ABSTRACT

Summary: It was previously demonstrated that splicing elements are positional dependent. We exploited this relationship between location and function by comparing positional distributions between all possible 4096 hexamers around a database of human splice sites. The distance measure used in this study found point mutations that produced higher distances disrupted splicing, whereas point mutations with smaller distances generally had no effect on splicing. Reasoning the idea that functional splicing elements have signature positional distributions around constitutively spliced exons, we introduce Spliceman—an online tool that predicts how likely distant mutations around annotated splice sites were to disrupt splicing. Spliceman takes a set of DNA sequences with point mutations and returns a ranked list to predict the effects of point mutations on pre-mRNA splicing. The current implementation included the analyses of 11 genomes: human, chimp, rhesus, mouse, rat, dog, cat, chicken, guinea pig, frog and zebrafish.

Availability: Freely available on the web at http://fairbrother.biomed.brown.edu/spliceman/

Contact: fairbrother@brown.edu

Received on June 23, 2011; revised on January 19, 2012; accepted on January 25, 2012

1 INTRODUCTION

Pre-mRNA splicing is an important regulatory step in gene expression pathway: introns are removed and exons are joined to form mRNA. The splicing process is performed by the spliceosome, a macromolecular ribonucleoprotein complex that rivals the ribosome in size and complexity. The intricate assembly of the spliceosome is guided by the consensus splice site sequences (i.e. branch point, polypyrimidine tract, 3′ and 5′ splice sites) and a family of subsidiary elements known as intron and exon splicing enhancers and silencers. Estimates of the fraction of disease mutations that cause aberrant splicing had been reported to range from 15% (Stenson *et al*., 2003) to 62% (Lopez-Bigas *et al*., 2005).

In a previous study (Lim *et al*., 2011), we demonstrated that splicing elements had signature positional distributions around constitutively spliced exons—they were abundant where they functioned positively and rare when they were inhibitory. We captured these positional properties for hexamers with the L1 distance metric (Section 2) and used it to cluster positional distributions of all possible 4096 hexamers around human splice sites. In addition to recognizing consensus splice site sequences, our method successfully identified various classes of intronic and exonic splicing enhancers and silencers. Experimental verifications of the computational results strongly indicated the power of this method to be predictive. Specifically, we found point mutations that produced higher L1 distances disrupted splicing in an *in vivo* minigene system, whereas point mutations with small distances generally had no affect on splicing. To facilitate the analysis of splicing mutations, we present Spliceman—an online tool that predicts how likely a genomic variation is to disrupt splicing. While the effect of mutations found in the consensus splice donor and acceptor sites can often be predicted with high accuracy, Spliceman also excludes splice sites positions in order to predict distant splicing enhancers and silencers.

2 METHODS

2.1 Design and implementation

The computational engine and web interface were developed in Perl and with the use of *Bioperl* toolkit (Stajicn *et al.*, 2002). The tool was designed to accept a set of DNA sequences with mutational data in FASTA format.

2.2 Preparation of exon databases

Exon database of each species was built from Refseq annotations of the following assemblies stored at the UCSC Table Browser (Karolchik *et al*., 2004): human (hg18 and hg19), chimp (panTro3), rhesus (rheMac2), mouse (mm9), rat (rn4), dog (canFam2), cat (felCat4), chicken (galGal3), guinea pig (cavPor3), frog (xenTro2) and zebrafish (danRer7). Duplicated entries were removed, and each sequence was divided into two distinct regions: upstream intron (up to 200 intronic and 100 exonic nucleotides of 3′ss) and downstream intron (up to 200 intronic and 100 exonic nucleotides of 5′ss). Therefore, each sequence in the exon database contained at most 600 nt. In the case where intronic or exonic sequence length was <400 or 200 nt, respectively, the sequences were equally divided and each half was assigned to its nearest splice site.

2.3 Algorithm methodology

Step (1) Selecting word size and generating feature vectors: RNA binding proteins typically contain one to four RNA recognition motif domains so that motifs recovered are expected to be of heterogeneous length. Our analysis of prior SELEX studies indicated that RNA binding proteins recognized motifs between the length of 6–10 nt (Lim *et al.*, 2011). Previous implementations of dictionary methods also illustrated how a smaller word size choice was generally self-correcting (Fairbrother *et al.*, 2002; Zhang and Chasin 2004). For these reasons, as well as computation efficiency, we selected hexamers for the analysis presented here. For each hexamer, the counting algorithm traversed through the exon database and recorded the occurrences of that hexamer at 600 different positions relative to splice sites. Repeated this

*To whom correspondence should be addressed.

procedure for all hexamers generated 4096 feature vectors. Each feature vector highlights the enrichment and depletion characteristics of a hexamer at locations relative to splice sites. Since overlapping occurrences of internally repeated words can occur more frequently than complex words, overlapping occurrences of any words were counted as a single occurrence in a window of 11. For example, a run of 11 A's (i.e. AAAAAAAAAAA) was counted as single occurrence at the position where it was first observed.

Step (2) Quantifying similarities and differences between feature vectors by computing L1 distance metric: this tool used the L1 distance metric to quantify the 'closeness' between two feature vectors. An obvious choice for distance metric is the Euclidean or L2 distance; however, the sharp peaks created by the splice site hexamers themselves dominated the comparison and prevented the detection of more subtle signals. This was remedied by using the Manhattan distance, also referred to as the city block distance or simply L1 distance. L1 distance was calculated as the sum of the absolute differences in feature vectors at each of the 600 positions. The higher the L1 distance between two hexamers (i.e. wild type versus point mutation), the greater the differences are between them, thereby the mutation is predicted to be more likely to alter splicing. In order to facilitate in the analyses of distant splicing elements, Spliceman also calculates L1 distance metric by masking out splice sites positions (e.g. 20 intronic positions upstream and 5 exonic positions downstream of the 3′ splice site; 5 exonic positions upstream and 8 positions downstream of the 5′ splice site).

Step (3) Calculating percentile ranks for L1 distances: this method binned all possible L1 distances into 100 equal intervals and assigned each L1 distance to its corresponding bin. For instance, comparisons between two hexamers that resulted in low L1 distances would be assigned with low percentile ranks.

3 OUTPUT

Spliceman takes a set of DNA sequences with point mutations and computes how likely these single nucleotide variants alter splicing phenotypes. For each mutation given in the input form, the tool reports the L1 distance and percentile rank that correspond to the given mutation. This is the rank that the tool uses to predict how likely a mutation is to disrupt pre-mRNA splicing. The higher the percentile rank, the more likely the point mutation is to disrupt splicing.

4 RECEIVER OPERATING CHARACTERISTIC CURVE STATISTICS

We previously demonstrated the predictive power of the proposed method by clustering hexamers into distinct groups based on positional distributions (Lim *et al.*, 2011). Experimental verifications suggested that mutations with high L1 distances altered splicing, whereas mutations with low L1 distances generally had no affect on splicing. To further analyze the predictive power of this method, we computed receiver operating characteristic (ROC) curve statistics using a binary classifier ('0' corresponds to true positive samples derived from a set of 1987 confirmed splicing mutations found in the Human Gene Mutation Database (HGMD) and '1' corresponds to false positive samples constructed from a set of simulated mutations using equal rates of transversions and transitions). ROC statistics were computed for mutations found in three different regions around annotated human splice sites (upstream 3′ splice site introns, exons and downstream 5′ splice site introns). Since splice site sequences can often be predicted with high accuracy, we removed HGMD mutations that were located in the consensus splice donor and acceptor sites to measure the predictive power of this method on distant splicing enhancers and silencers. The area under curve measurements suggested our proposed method was predictive of distant splicing mutations (Fig. 1).

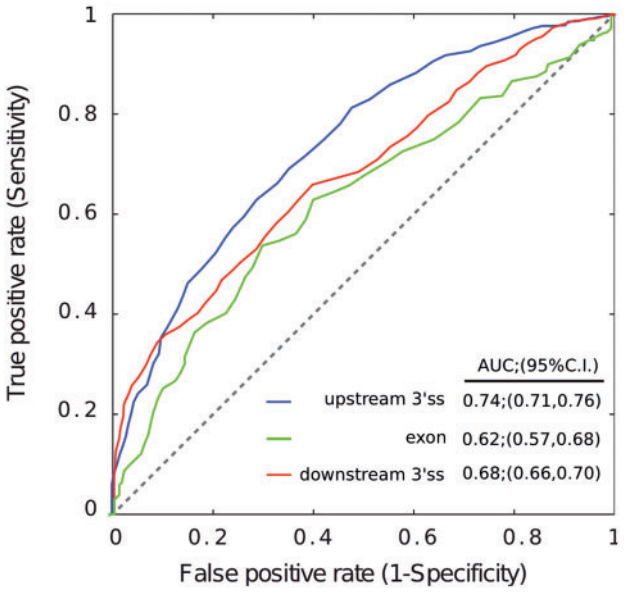

Fig. 1. L1 distance metric is predictive of distant splicing mutations. ROC curve analysis using HGMD splicing mutations to compare L1 distances in three distinct regions around the annotated human splice sites. True positive samples are derived from a total of 1987 HGMD splicing mutants found outside of the donor and acceptor sites, and false positive samples are constructed from simulated mutations using equal rates of transversions and transitions. The exonic region is shown in green; upstream and downstream introns are shown in blue and red, respectively. AUC, area under curve; C.I., confidence interval.

Funding: The lab is funded by the National Institute of Health (1R01GM095612-01).

Conflict of Interest: none declared.

REFERENCES

Fairbrother,W.G. *et al.* (2002) Predictive identification of exonic splicing enhancers in human genes. *Science,* **297**, 1007–1013.

Karolchik,D. *et al.* (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, **1**, D493– D496.

Lim,K.H. *et al.* (2011) Using positional distribution to identify splicing elements and predict mRNA processing defects in human genes. *Proc. Acad. Sci. USA*, **108**, 11093–11098.

Lopez-Bigas,N. *et al.* (2005) Are splicing mutations most frequent cause of hereditary disease? *FEBS Lett.*, **579**, 1900–1903.

Stajicn,J.E. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.

Stenson,P.D. *et al.* (2003) Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.*, **21**, 577–581.

Zhang,X.H. and Chasin,L.A. (2004) Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev.,* **11**, 1241–1250.