FESTA:
*F*LEXIBLE *E*XON-BASED *S*PLICING AND *T*RANSCRIPTION *A*NNOTATION

Rago, Alfredo

April 11, 2018

Abstract

I introduce FESTA, an R based algorithm that allows detection of alternative splicing based on experiment-specific exon expression data. FESTA disentangles alternative splicing signal from whole-gene transcription, facilitating the discovery and characterization of novel regulatory events even in the absence of transcript annotations or paired-end data. I also include customization options to increase its applicability on different platforms and experimental designs as well as a tool for the conversion from transcript expression to inclusion ratios.

1 Background

Alternative splicing is a widespread feature of eukaryotic gene regulation which can be represented as a two step process. Transcription generates the total amount pre-mRNA per gene locus, whereas splicing determines the proportions of each alternative transcript that is produced. Based on this model I can distinguish between constitutive exons, which are present across all isoforms and facultative exons, which are present in only a subset of alternative transcripts.

Commonly used methods discard information contained in constitutive exons or average it to match the proportions provided by transcript-specific exons [?], effectively conflating transcription and splicing dependent signal. Furthermore if reads are mapped to pre-annotated transcripts novel transcriptional events might be missed entirely. Dataset-specific estimation of constitutive and transcript-specific exons is therefore advisable for the discovery of novel alternative splicing events relevant to the design of interest [?, ?].

Correlation based exon clustering is a simple implementation of such a method [?]. Since strong correlations among exons arise from their coexpression as part of a single transcript, every cluster represents either an alternative transcript or the subset of exons that are present across all isoforms (constitutive exons, [?]). Constitutive exon clusters will be present in all isoforms and can therefore be identified by having an absolute expression value either higher or equal to any other exon group. Despite being intuitive and effective, correlation based hierarchical clustering is limited by its choice of an a priori threshold.

In this chapter I define a simple algorithm that solves this issue by setting gene-specific thresholds based on highly customizable biological expectations. I also provide a function to calculate inclusion ratios of alternative exon groups in order to allow analysis of transcription independent effects of splicing.

2 Implementation and Usage

2.1 Data input and filtering

FESTA requires two input files: an exon by sample expression table and an exon to gene assignment table. In order to avoid spurious grouping resulting from correlations in the noise component I advise thresholding raw expression data, removing all values that score below minimal signal and excluding all exons that lack expression in a sufficient number of biological replicates for at least one of the dataset's conditions.

2.2 Isoform detection

Figure 1 shows an outline of the FESTA algorithm, which is applied to iteratively to each gene. If a gene has more than one expressed exon, FESTA calculates a clustering tree based on the correlation matrix of expressed exons. FESTA then cuts the tree at the lowest level (one exon per cluster) and ranks each group's expression in each sample (figure 1, A).

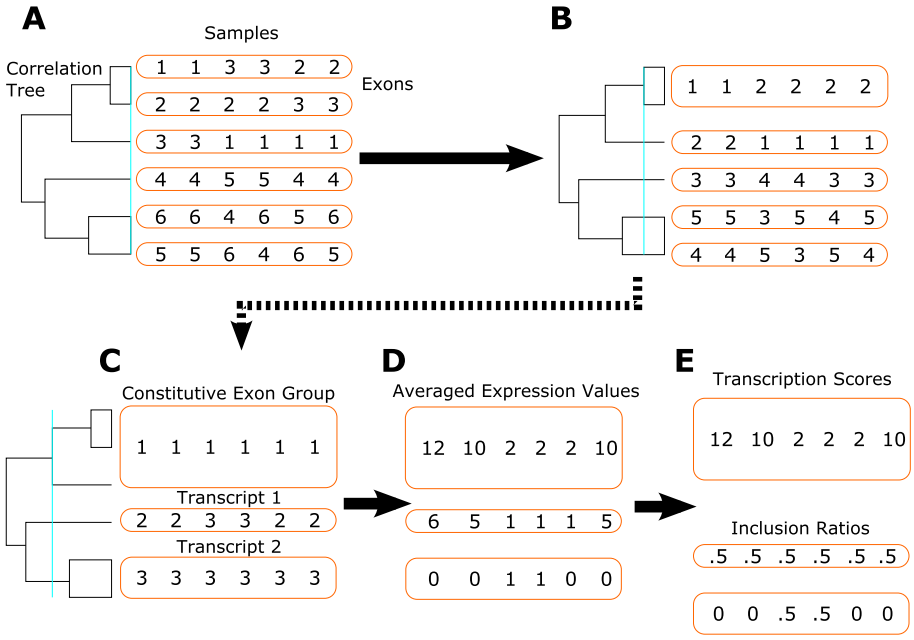

Figure 1: **Outline of the FESTA algorithm**. Steps A-C are handled by the ClusterExons function. Step D and optional step E are handled by the AverageExons function.

If any exon group is ranked first or tied for first across all samples I consider it to be the constitutive part of the gene, record the cluster assignments at the tree cut level and proceed on to the next gene. If no exon group ranks as first or tied across all samples, FESTA moves up a level in the hierarchical clustering tree, averages expression scores in exon groups with more than one exon and re-calculates the exon group rankings across the dataset (figure 1, B).

FESTA iteratively calculates the expression rankings of exon groups at each level until a single exon group shows the highest expression across all samples (figure 1, C). If no exon group meets constitutive exon criteria at the highest level, the algorithm converges on single-group clustering: all exons are annotated as constitutive and the gene is reported as lacking significant splicing events.

FESTA generates a single expression score for each group by averaging the expression scores of all its exons (figure 1, D). These raw expression scores can be directly used for analyses on individual transcript abundance. Alternative splicing events can also be converted to inclusion ratios by dividing them by the transcription score of their gene (figure 1, E). Inclusion ratios range between zero (if the isoform is absent) and one (if all transcripts produced by the gene include those exons) and can be used to analyze the effects of alternative splicing independently of the main gene's overall expression.

2.3 Fine tuning parameters

I include two main parameters can be changed to affect the sensitivity and power of the main clustering algorithm: significant digits and number of exceptions.

Significant digits allows the user to define numerical accuracy of expression measurements. Setting a high number of significant digits will result in less ties between exon groups but might cause over sensitivity to minor fluctuations in expression values between biologically co-expressed exons. Fewer significant digits increase the number of ties in rankings, decreasing the ability to differentiate constitutive exons from highly expressed alternative exons.

Number of exceptions allows to increase the permissiveness of constitutive exon group definitions. If this number is greater than zero, constitutive groups are re-defined as being first or tied with any other group in all samples except the exceptions. For instance, in case the dataset includes 25 samples, exon groups will be identified as constitutive if they are first or tied in at least 24 samples if exception number is set to 1, at least 23 samples if it is set to 2 and so on. This parameter enables setting tree-cut height based on experimental design considerations, with more stringent values resulting in less isoforms and larger constitutive exon groups and more permissive values resulting in more isoforms and smaller constitutive exon groups.

2.4 Caveats

There are three caveats regarding FESTA's current implementation. Firstly, the algorithm depends on the number of biological replicates to generate accurate exon rankings. Secondly, it does not currently make use of paired-end data.

Lastly, as the algorithm attempts to identify isoform-specific exon groups it will not be able to detect isoforms characterized by different combinations of the same exons such as in the case of hypervariable combinatorial genes.

3 Conclusions

I present an intuitive method for the detection of transcription and splicing in transcriptomic data which requires only an exon by sample expression table. FESTA allows the end user to customize sensitivity using easily interpreted parameters which can be tuned to the experimental design and the instrument's sensitivity. FESTA's output is a reduced transcript by sample table, which retains only the splice variants observed in the experiments and can be directly used in downstream analyses. The optional conversion from transcript abundances to splicing ratios allows the investigation of the effects of increasing the proportions of specific isoforms rather than their absolute abundances, allowing for a comparative study of the impact of transcriptional and splicing regulation.

References