Institut für Informatik
Praktische Informatik und Bioinformatik
Prof. Dr. Ralf Zimmer
Dr. Gergely Csaba
Markus Gruber

Wintersemester 2018/2019
internal website user: gobiws18
password: rNs3q18

GoBi: Excercise 1

Genome Annotation

proposed deadline: Tuesday, 06.11.2018, 14:00

Save your solution to `/home/proj/biocluster/praktikum/genprakt/${account}/Solution1`. Provide also an executable jar file (containing also the sources) in this directory that allows for reproducing your results. The jar should print a usage info if invoked without parameters. Submit your jarfile also to the `<abgabeserver>` at

`https://services.bio.ifi.lmu.de:1047/abgabeserver/`

to template named `ExonSkipping`.

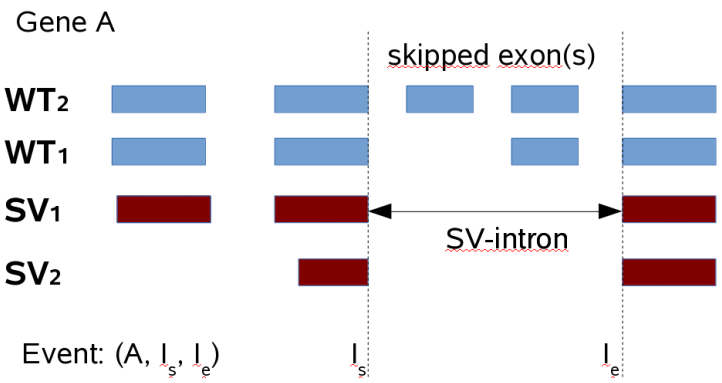
Analyze exon skippings:

One form of alternative splicing is exon skipping. An exon-skipping splicing event is a tuple (gene, intron-start, intron-end) and is defined by (at least) two transcripts: wildtype (WT) and spliced variant (SV) of the same gene, and an intron in SV with start and end corresponding to an exon end and exon start in WT, respectively, and the SV-intron spans at least one exon in WT.

For any exon-skip event there may be several WT-s and several SV-s, and there may be several sets of skipped exons (see figure below; this is **one** exon-skipping event).

Implement a program considering the following parameter:

- `-gtf <GTF file>` : genomic annotation
- `-o <output file path>`: path to the output where the table (defined below) containing all exon skippings will be written.



The program should extract exon-skipping events between **CDS**(!)-es for the gtf file and write the results into a tsv file with the following headers:

- id (gene id)
- symbol (gene symbol)
- chr (chromosome)
- strand (+ or -)
- nprots (number of annotated CDS in the gene)
- ntrans (number of annotated transcripts in the gene)
- SV (the SV intron as start:end)
- WT (the WT introns within the SV intron separated by | as start:end)
- SV_prots (ids of the SV CDS-s, separated by |)
- WT_prots (ids of the WT CDS-s, separated by |)
- min_skipped_exon the minimal number of skipped exons in any WT/SV pair
- max_skipped_exon the maximum number of skipped exons in any WT/SV pair
- min_skipped_bases the minimal number of skipped bases (joint length of skipped exons) in any WT/SV pair
- max_skipped_bases the maximum number of skipped bases (joint length of skipped exons) in any WT/SV pair

An example output for the gene `ENSG00000131018` can be found at `/home/proj/biosoft/praktikum/genprakt/ExonSkipping/ENSG00000131018.exonskippings`.

Apply your program to all GTF file in `/home/proj/biosoft/praktikum/genprakt/gtfs/` to create an overview as specified below. Write two cumulative plots into your output directory named `skipped_exons.jpg` and `skipped_bases.jpg` showing the distributions of the maximum number of skipped exons / skipped bases per event for the different GTF files, and an html file `exon_skipping.html` showing these plots and linking the top 10 genes for both criteria to the current ENSEMBL version.

Use only relative links to the plots in your html-file!