Institut für Informatik
Praktische Informatik und Bioinformatik
Prof. Dr. Ralf Zimmer
Armin Hadziahmetovic
Samuel Klein

Wintersemester 2025/2026

GoBI: Excercise 1

Information Extraction and Pattern Detection in Strings
Genome Annotation: Gene Structure Definition

Deadline: Monday, 03.11.2025, 12:00

Background:

Genomes are represented as sets of long strings (the chromosomes) over the alphabet $\Sigma = A, C, G, T$ (the nucleotides). A human genome consists of 46 chromosome strings with approximately 50 to 250 million letters. Certain sets of substrings (exons, introns) of the chromosomes form the genes. Gene transcripts are built from combinations of exons (and sometimes also introns) of a gene. Naturally, there is an exponential number of such combinations. The known and predicted combinations are defined in a format called a GTF file. The Gene Transfer Format (GTF) is a text-based, tab-delimited file format used to hold information about gene structure. Specifications for the GTF format, tools, validators, etc., can be found on various official websites, such as the Ensembl webpage `https://www.ensembl.org/info/website/upload/gff.html`.

Aim:

This programming task involves parsing a large GTF file (e.g., ∼50MB gzipped, ∼1.5GB uncompressed) to extract the genome annotation and gene structures. The program should then identify certain transcript structure patterns from the given annotation, known as exon skipping splicing events. The output of the program consists of summaries of the identified splice events, supplemented and displayed with appropriate statistics and plots.

The goal of this task is to understand a complex input format and implement the defined pattern identification. Next, appropriate data structures and algorithms must be designed to efficiently identify all pattern instances in large inputs and strings. A detailed program specification, with several parameters, must be implemented correctly to meet all requirements exactly as described in the task, along with the accompanying command line interface.

Problem description: Analyze exon skippings for a given annotation (GTF file)

One form of alternative splicing is exon skipping. We define *Exon Skipping Splicing Event* (**ES-SE**) as a tuple (gene g, intron-start I_s, intron-end I_e). An ES-SE requires (at least) two

transcripts: a so-called *wildtype* (**WT**) and a *splice variant* (**SV**) of the same gene, involving a specific intron. The intron corresponds to a defined exon end and a downstream exon start, and spans at least one exon in the WT.

For any ES-SE (g, I_s, I_e), there may exist several WTs and several SVs. The SVs of an ES-SE contain exactly the specified intron $[I_s, I_e)$; all other transcripts are WTs for this ES-SE. For the WTs, there may be several sets of skipped exons in the specified intronic region $[I_s, I_e)$. WTs and SVs can have various structures to the left and right of the intronic region. The figure illustrates **one** ES-SE (for the indicated SV-intron $[I_s, I_e)$).

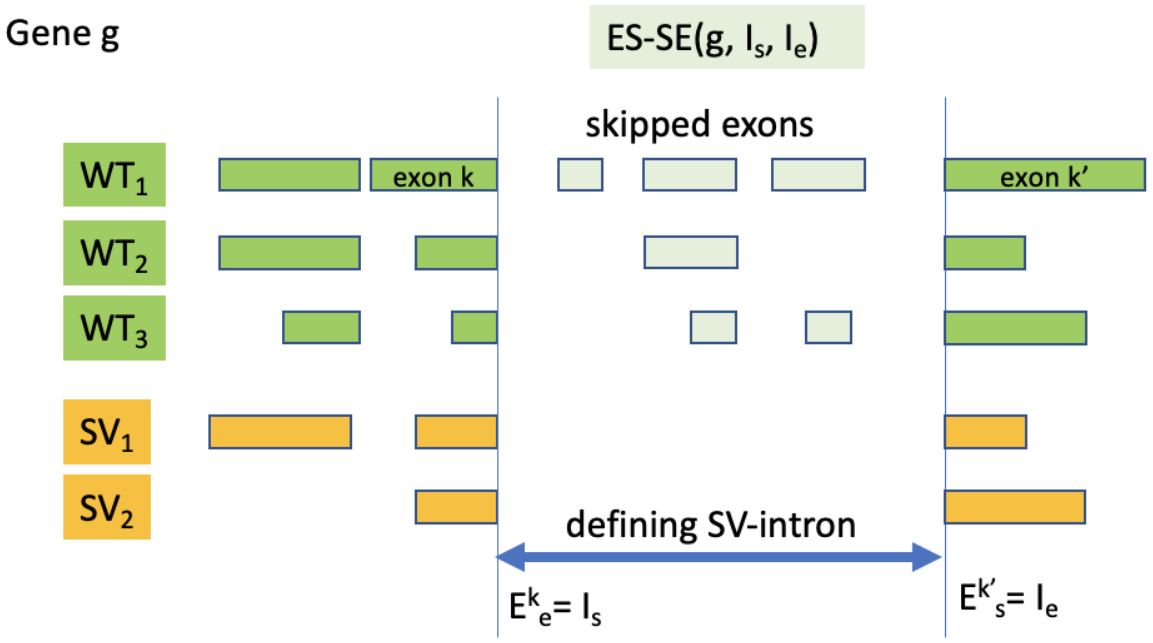

Figure 1: An example of an *Exon skipping splicing event (ES-SE)*. Exons are defined by start and end positions and span the interval $[E_s, E_e)$. The event is described by a consistent exon end E_e^k and a consistent exon start $E_s^{k'}$ for a $k' > k$. Consistent means it is the same for all WT for the respective ES-SE. The defining SV-intron of the ES-SE spans from I_s to I_e, i.e., $[I_s, I_e)$, where $I_s = E_e^k$ (i.e., the `<end of the k-th exon in the GTF>` with +1 optionally added, depending on whether the start/end is defined as inclusive or exclusive) and $I_e = E_s^{k'}$. Note that WT can contain any combination of exons in the intronic regions whereas, by definition, all SV do not have any (non skipped) exon in the defining intronic region.

Program specification:

Implement a program with the following parameters:

- `-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 jar file should print a usage info if invoked without parameters.

The program should extract all ES-SE defined only (!) by coding sequences (**CDS**) listed in the input GTF file. The identified ES-SE are written into a tab separated text (.tsv) file with the following headers/columns:

- 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 (all the WT introns within the SV intron separated by | as start:end)
- SV_prots (ids of the SV CDS, separated by |)
- WT_prots (ids of the WT CDS, 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 in:
`/mnt/biosoft/praktikum/genprakt/ExonSkipping/ENSG00000131018.exonskippings`.

Apply your program to all GTFs in `/mnt/biosoft/praktikum/genprakt/gtfs/` and create a (scientific[1]) report of the exon skippings for all the GTFs. The report should contain at least two cumulative plots (see definition on the internet or an R tutorial) showing the

[1]Guideline: a bunch of plots without a clear explanation of what is shown, without description of what can be observed, or without a statement what can be concluded is not a sufficient report and will be graded accordingly.

distributions of the maximum number of skipped exons and skipped bases, respectively, per ES-SE for every GTF file. The plots must be saved to your output directory named `skipped_exons.jpg` and `skipped_bases.jpg`.

In addition to the plots please also list top 10 genes, for both criteria, with associated links to the current ENSEMBL version. As a supplement to your report you should briefly describe your approach for solving the exon skipping problem. This includes some analysis of the correctness, time complexity, and actual time usage of your method.

Create your personal directory if it doesn't exist yet and create adequate sub-directories for each assignment, e.g. for the first solution:
`/mnt/biocluster/praktikum/genprakt/${user}/Solution1`

Guidelines:

The report must be written in LaTeX and submitted as a PDF (`exon_skipping.pdf`) file and as a source archive (`exon_skipping.zip` containing the LaTeX project: `.tex`, plots, etc.). It must include your plots and a description of the observations and the conclusions you can draw from them.

Upload your solution to your personal directory. This must include your `.jar` file, reports, plots, and R/Python scripts/notebooks, or any other resources used to generate the plots.

Provide the jar as an executable file (including the sources) to allow for the reproduction of your results. Please also upload the Report and associated Supplement **before the deadline** to your directory.

In order to evaluate your program, submit your jar file also to the `<abgabeserver>`
(task `exon_skipping` at `https://services.bio.ifi.lmu.de:1047/abgabeserver/`)

Every participant has to make individual submissions (jar, report, and supplement) to the `<abgabeserver>` and their directory. We will check for plagiarism of the source code. It is allowed, and maybe instructive, to discuss with your fellow GoBI participants, in order to understand the specifications, to debug the implementation, to analyse the data, and to write the report. Again, individual reports must be written and submitted for review and grading.

A rough weighting of the various deliverables is given as follows (total 200 out of 1000 points):

- up to 120 pts - `<abgabeserver>` submission
- up to 45 pts - Report
- up to 35 pts - Report Supplement

Note that 100% on the `<abgabeserver>` does not imply the correctness of your submission; it is, however, a necessary condition for correctness. Therefore, you should further demonstrate your correctness in your report supplement.