3.3 - Tool SCORER (Splice Site Scoring)

Description

We embedded in the astalavista framework the scoring of splice sites that are observed in the provided transcriptome annotation. By tradition in gene finding strategies [1,2,3], these are log-likelihood scores computed by the information stored in Hidden Markov Models (HMMs) which have been trained by splice sites in the corresponding species [1,4]. Pre-computed splice site models for different species can be obtained from the geneid homepage, if no explicit model is provided the astalavista scorer employs a default human HMM.

High scores represent common/well working splice sites whereas low scores reflect sites that are rare and thus probably thermodynamically also less efficient for the splicing process. By the logarithmic nature of the scores, values can drop below 0, for instance the default matrix yields values as low as (-10) for very rare splice sites. However, scores of (-9999) and lower express splice sites that have not ever been observed in the training set, which probably are non-functional.

References

1. R. Guigó, "Assembling genes from predicted exons in linear time with dynamic programming", *Journal of Computational Biology*, 5:681-702 (1998).
2. R. Guigó, S. Knudsen, N. Drake, and T. F. Smith, "Prediction of gene structure", *Journal of Molecular Biology*, 226:141-157 (1992).
3. E. Blanco, G. Parra and R. Guigó, "Using geneid to Identify Genes.", In *Current Protocols in Bioinformatics.* Unit 4.3. (A. Baxevanis, editor) John Wiley & Sons Inc., New York (2002)
4. G. Parra, E. Blanco, and R. Guigó, "Geneid in *Drosophila*", *Genome Research* 10(4):511-515 (2000).

Examples

Scoring Splice Sites of a given annotation employing the default model (i.e., human)

```
astalavista -t scorer -i <annotation.gtf> -c <genome-folder>
```

where <annotation.gtf> is the transcriptome annotation (see GTF format) which contains the splice sites to be scored, and <genome-folder> is the path to the directory containing genomic sequences, one FASTA file per reference sequence (i.e., chromosome, contig, scaffold, etc.).

Scoring Splice Sites with polymorphisms

```
astalavista -t scorer -i <annotation.gtf> -c <genome-folder> --vcf <vcf-file>
```

with <vcf-file> being a file in the VCF (Variant Call Format), describing polymorphisms of the genomic sequence.

Scoring Splice Sites with a custom geneid profile

```
astalavista -t scorer -i <annotation.gtf> --gid <profile>
```

where <profile> is a geneid profile for the HMM model.

The default program output is in VCL format written to a file "<annotation>_sites.vcf" in the same directory as the provided transcriptome annotation <annotation.gtf>. The output file can be changed by the command line flag -f.

Requirements

Hardware

For time efficiency, the positions of all genetic variants are loaded into the computer's memory (RAM), so it is to be ensured that enough memory is provided to the Java Virtual Machine. As an orientation, the variants from the 1000 Genomes project phase 1 and phase 2 just for chr22 occupy 6.4 Gb of disk and require ~1.5Gb for running splice site scoring.

Data Files

1. **Gene Models** (annotation in GTF format), **REQUIRED**: if missing, the program responds with an error like

```
Hey, you forgot to specify a valid input file!
This is a bit important, I cannot work without an input annotation. I want a GTF file with transcript
annotations (exon features, with a mandatory optional attribute named 'transcript_id') IN THE SAME
COLUMN (i.e., if the transcript identifier of the 1st line is in column #10, it has to be in all lines
of the file in column #10. The rest of the file should comply with the standard as specified at
http://mblab.wustl.edu/GTF2.html.
```

2. **Chromosome sequences** (FASTA files, one per chromosome), **REQUIRED**: if missing, the program responds with an error like

```
[ERROR] Splice site scoring requires the genomic sequence, provide a value for parameter 'CHR_SEQ' in
the parameter file, or via the command line flags -c or --chr!
```

The chromosome sequences currently have to be provided as separate files, one per chromosome. All of these files have to be in the same folder (e.g., genomes/H.sapiens/hg19) with a filename prefix that corresponds to the tags in column $1 of the GTF filel provided and a suffix ".fa" or ".fasta"; e.g., if chromosomes are named "chr1", "chr2", etc. then the program expects files named "chr1.fa", "chr2.fa", ...

The first line of every

3. **Genetic variants** (as a pseudo VCF file)

Modified bases can be provided in a 5-column file format that resembles the variant-calling-format (VCF): column 1 is chromosome ID (number or letter), column 2 is position within the chromosome (integer), column 3 is variant identifier, column 4 is reference nucleotide and column 5 is the variant nucleotide.

```
1       948921       rnaedit_1_948921       T       C
1       982994       rnaedit_1_982994       T       C
1       990773       rnaedit_1_990773       C       T
1       1158631       rnaedit_1_1158631       A       G
1       1247494       rnaedit_1_1247494       T       C
1       1309405       rnaedit_1_1309405       T       C
1       1336006       rnaedit_1_1336006       C       T
1       1336626       rnaedit_1_1336626       G       A
1       1342394       rnaedit_1_1342394       G       A
1       1684472       rnaedit_1_1684472       C       T
```

Characters in the first column of the VCF file have to correspond to the suffixes of the chromosome names of the (1) gene annotation and (2) chromosome files, removing the prefix "chr".