FlipFlop: Fast Lasso-based Isoform Prediction as a Flow Problem

Elsa Bernard Laurent Jacob Julien Mairal Jean-Philippe Vert

September 14, 2014

Abstract

FlipFlop implements a fast method for *de novo* transcript discovery and abundance estimation from RNA-Seq data. It differs from Cufflinks by simultaneously performing the transcript and quantitation tasks using a penalized maximum likelihood approach, which leads to improved precision/recall. Other softwares taking this approach have an exponential complexity in the number of exons in the gene. We use a novel algorithm based on network flow formalism, which gives us a polynomial runtime. In practice, *FlipFlop* was shown to outperform penalized maximum likelihood based softwares in terms of speed and to perform transcript discovery in less than 1/2 second even for large genes.

1 Introduction

Over the past decade, quantitation of mRNA molecules in a cell population has become a popular approach to study the effect of several factors on cellular activity. Typical applications include the detection of genes whose expression varies between two or more populations of samples (differential analysis), classification of samples based on gene expression [van't Veer et al., 2002], and clustering, which consists of identifying a grouping structure in a sample set [Perou et al., 2000]. While probe-based DNA microarray technologies only allow to quantitate mRNA molecules whose sequence is known in advance, the recent development of deep sequencing has removed this restriction. More precisely, RNA-Seq technologies [Mortazavi et al., 2008] allow the sequencing of cDNA molecules obtained by reverse transcription of RNA molecules present in the cell. Consequently, any transcript can be sequenced and therefore quantitated, even though its sequence might not be available a priori for designing a specific probe. In addition to facilitating the study of non-coding parts of known genomes and organisms whose genome has not been sequenced [Mortazavi et al., 2010], RNA-Seq technologies facilitate the quantitation of alternatively spliced genes. Genes in eukaryote cells indeed contain a succession of exon and intron sequences. Transcription results in a pre-mRNA molecule from which most introns

are removed and some exons are retained during a processing step called RNA splicing. It is estimated that more than 95% of multiexonic genes are subject to alternative splicing [Pan et al., 2008]: the set of exons retained during splicing can vary, resulting for the same gene in different versions of the mRNA, referred to as transcripts or isoforms. Identification and quantification of isoforms present in a sample is of outmost interest because different isoforms can later be translated as different proteins. Detection of isoforms whose presence or quantity varies between samples may lead to new biomarkers and highlight novel biological processes invisible at the gene level.

Sequencing technologies are well suited to transcript quantitation as the read density observed along the different exons of a gene provide information on which alternatively spliced mRNAs were expressed in the sample, and in which proportions. Since the read length is typically smaller than the mRNA molecule of a transcript, identifying and quantifying the transcripts is however difficult: an observed read mapping to a particular exon may come from an mRNA molecule of any transcript containing this exon. Some methods consider that the set of expressed isoforms [Jiang and Wong, 2009] or a candidate superset [Huang et al., 2012, Xing et al., 2006] is known in advance, in which case the only problem is to estimate their expression. However little is known in practice about the possible isoforms of genes, and restricting oneself to isoforms that have been described in the literature may lead to missing new ones.

Two main paradigms have been used so far to estimate expression at the transcript level while allowing de novo transcript discovery. On the one hand, the Cufflinks software package [Trapnell et al., 2010] proceeds in two separate steps to identify expressed isoforms and estimate their abundances. It first estimates the list of alternatively spliced transcripts by building a small set of isoforms containing all observed exons and exon junctions. In a second step, the expression of each transcript is quantified by likelihood maximization given the list of transcripts. Identification and quantification are therefore done independently. On the other hand, a second family of methods [Xia et al., 2011, Li et al., 2011b, Bohnert and Rätsch, 2010, Li et al., 2011a, Mezlini et al., 2013] jointly estimates the set of transcripts and their expression using a penalized likelihood approach. These methods model the likelihood of the expression of all possible transcripts, possibly after some preselection, and the penalty encourages sparse solutions that have a few expressed transcripts.

The two-step approach of Cufflinks [Trapnell et al., 2010] is reasonably fast, but does not exploit the observed read density along the gene, which can be a valuable information to identify the set of transcripts. This is indeed a conclusion drawn experimentally using methods from the second paradigm [see Xia et al., 2011, Li et al., 2011b, Bohnert and Rätsch, 2010, Li et al., 2011a, Mezlini et al., 2013].

To summarize, the first paradigm is fast but can be less statistically powerful than the second one in some cases, and the second paradigm should always be powerful but becomes untractable for genes with many exons. The contribution of this paper is to allow methods of the second family to run efficiently without prefiltering the set of isoform candidates, although they solve a non-smooth optimization problem over an exponential

number of variables. To do so, we show that the penalized likelihood maximization can be reformulated as a convex cost network flow problem, which can be solved efficiently [Ahuja et al., 1993, Bertsekas, 1998, Mairal and Yu, 2012].

For more detail about the statistical model and method, see Bernard et al. [2013] and references therein.

2 Software features

FlipFlop takes aligned reads in `sam` format and offers the following functionalities:

Transcript discovery *FlipFlop* estimates the set of transcripts which are most likely to be expressed according to the model described in Bernard et al. [2013].

Abundance estimation The implemented method simultaneously estimates the abundance of the expressed transcripts in FPKM.

3 Case studies

We now show on a simple one gene example how *FlipFlop* can be used to estimate which transcripts are expressed in an RNA-Seq experiment, and what are the transcript abundances.

3.1 Loading the library and the data

We load the *FlipFlop* package by typing or pasting the following codes in R command line:

```
> library(flipflop)
```

A `.sam` data file can be loaded by the following command:

```
> data.file <- system.file(file.path('extdata', 'vignette-sam.txt'), package='flipflop')
```

These toy data correspond to the alignments of 1000 single-end reads of 125 base-pair long against the hg19 reference genome, available on the UCSC genome browser [1]. The reads have been simulated with the RNASeqReadSimulator [2] from two annotated human transcripts (see reference ID uc001alm.1 and uc001aln.3 in the UCSC genome browser).

In a general context `data.file` should simply be the path to the `.sam` alignement file.

FlipFlop pre-processing of the reads (exctracting exon boundaries, junctions and associated counts), is based on the `processsam` function from the `isolasso` software. More information about the `isolasso` software and the `processsam` options can be found at the following link: `http://alumni.cs.ucr.edu/~liw/isolasso.html`.

[1] http://genome.ucsc.edu

[2] http://alumni.cs.ucr.edu/ liw/rnaseqreadsimulator.html

Note that the `.sam` file has to be sorted according to chromosome name and starting position. In Unix or Mac systems, it can be done with the command `sort -k 3,3 -k 4,4n in.sam > in.sorted.sam`.

3.2 Estimation

In order to estimate the set of expressed isoforms and their abundances, we run the `flipflop` function on the `.sam` file. By default the reads are considered as single-end and no annotation file is necessary. If you have paired-end reads you can use the option `paired=TRUE` and give the mean fragment size (option `frag`) and standard deviation (option `std`) of your RNA-seq library.

3.2.1 without annotation

```
> # The minimum number of clustered reads
> # to consider a cluster of reads as a gene (default 40):
> min.read <- 50
> # The maximum number of isoforms given
> # during regularization path (default 10):
> max.iso <- 7
> ff.res <- flipflop(data.file=data.file,
+                    out.file='FlipFlop_output.gtf',
+                    minReadNum=min.read,
+                    max_isoforms=max.iso)

> names(ff.res)

[1] "transcripts"     "abundancesFPKM"  "expected.counts" "timer"
```

The `flipflop` function outputs a list whose important features are lists `transcripts`, `abundancesFPKM` and `expected.counts`. Each element of these lists corresponds to a different gene in the `sam` file.

The `transcripts` list is a `GRangesList` object from the *GenomicRanges* package [Aboyoun et al.]. More information concerning manipulations of this object can be found in [Aboyoun et al.]. Each element of the list is a `GRanges` object that describes the structure of the transcripts that are found to be expressed. Rows of the object correspond to exons. On the left hand side each exon is described by the gene name, the chromosome, its genomic position on the chromosome and the strand. Transcripts are described on the right hand side. Every transcript is a binary vector where an exon is labelled by 1 if it is included in the transcript. Elements of `abundancesFPKM` are vector whose length is the number of isoforms listed in the `transcripts` object. Each element of the vector is the estimated abundance in FPKM of the corresponding transcript. `expected.counts` has the

same structure whereas it corresponds to the expected fragment counts for each transcript (ie the expected number of mapped fragments by transcript).

```
> transcripts <- ff.res$transcripts[[1]]
> abundancesFPKM <- ff.res$abundancesFPKM[[1]]
> expected.counts <- ff.res$expected.counts[[1]]
> print(transcripts)

GRanges with 7 ranges and 3 metadata columns:
        seqnames             ranges strand | read.count transcript.V1
           <Rle>          <IRanges>  <Rle> |  <numeric>     <numeric>
  Inst1     chr1 [4715106, 4715515]      + |        147             1
  Inst1     chr1 [4771960, 4772760]      + |        382             1
  Inst1     chr1 [4829913, 4830001]      + |         85             1
  Inst1     chr1 [4832340, 4832586]      + |        137             1
  Inst1     chr1 [4834487, 4834619]      + |         88             1
  Inst1     chr1 [4837461, 4837845]      + |         41             0
  Inst1     chr1 [4842605, 4843848]      + |        339             1
        transcript.V2
           <numeric>
  Inst1             1
  Inst1             1
  Inst1             1
  Inst1             1
  Inst1             1
  Inst1             1
  Inst1             0
  ---
  seqlengths:
   chr1
     NA

> print(abundancesFPKM)

[1] 278210.4 114046.5

> print(expected.counts)

[1] 777.3198 220.6800
```

Our example `sam` file contains a gene with 7 exons. Two transcripts were found to be expressed, with respective abundances 278210.2 and 114046.3 FPKM. The first of the

expressed isoforms contains all exons except the exon 6, the second isoform does not contain exon 7.

The output is also stored in a standart `gtf` format file. For more details about the GTF format visit `http://mblab.wustl.edu/GTF2.html`. In the so-called attributes column, `FPKM` corresponds to abundances in FPKM unit while `EXP-COUNT` corresponds to the expected fragment counts.

3.2.2 with annotation

The `flipflop` function allows as well the use of an annotated transcript file in `bed` format to settle a priori the exon boundaries. More precisely, the `bed` file must be a `bed12` file with 12 columns. It also has to be sorted according to chromosome name and starting position of isoforms.

A `.bed` annotation file can be loaded by the following command:

```
> annot.file <- system.file(file.path('extdata', 'vignette-annot.bed.txt'),
+                           package='flipflop')

> ff.res.annot <- flipflop(data.file=data.file,
+                     out.file='FlipFlop_output.gtf',
+                     annot.file=annot.file)

> transcripts.annot <- ff.res.annot$transcripts[[1]]
> print(transcripts.annot)

GRanges with 13 ranges and 3 metadata columns:
        seqnames             ranges strand   | read.count transcript.V1
           <Rle>          <IRanges>  <Rle>   |  <numeric>     <numeric>
  Inst1     chr1 [4715105, 4715515]      +   |        147             1
  Inst1     chr1 [4715515, 4771960]      +   |          0             0
  Inst1     chr1 [4771960, 4772760]      +   |        382             1
  Inst1     chr1 [4772760, 4829913]      +   |          0             0
  Inst1     chr1 [4829913, 4830001]      +   |         85             1
    ...      ...                 ...    ... ...        ...           ...
  Inst1     chr1 [4834487, 4834619]      +   |         88             1
  Inst1     chr1 [4834619, 4837461]      +   |          0             0
  Inst1     chr1 [4837461, 4837855]      +   |         42             0
  Inst1     chr1 [4837855, 4842605]      +   |          0             0
  Inst1     chr1 [4842605, 4843851]      +   |        339             1
        transcript.V2
           <numeric>
  Inst1             1
```

```
  Inst1             0
  Inst1             1
  Inst1             0
  Inst1             1
    ...           ...
  Inst1             1
  Inst1             0
  Inst1             1
  Inst1             0
  Inst1             0
  ---
  seqlengths:
   chr1
     NA
```

Two transcripts are again found to be expressed. The number of exons and the positions on the genome are not the same as in the previous example because boundaries now include exons and introns from the annotation.

3.3 Read the output GTF file

The transcripts which are found to be expressed are stored in a `gtf` format file. The transcript information from the GTF file can be easily extracted using the `makeTranscriptDbFromGFF` function from the *GenomicFeatures* package [Carlson et al.].

The following example shows (when the *GenomicFeatures* is installed) how to create a `TxDb` object from the GTF file, and several accessor functions allow to manipulate it. More information can be found in [Carlson et al.]. For instance the `exonsBy` function extracts the list of exons for each gene or transcript:

```
> if(require(GenomicFeatures)){
+     txdb <- makeTranscriptDbFromGFF(file='FlipFlop_output.gtf',
+                                     format='gtf',
+                                     exonRankAttributeName='exon_number')
+     # List of exons for each transcript:
+     exonsBy(txdb, by='tx')
+ }

GRangesList of length 2:
$1
GRanges with 6 ranges and 3 metadata columns:
      seqnames             ranges strand |   exon_id   exon_name exon_rank
         <Rle>          <IRanges>  <Rle> | <integer> <character> <integer>
```

```
  [1]     chr1 [4715105, 4715514]      + |         1        <NA>         1
  [2]     chr1 [4771960, 4772759]      + |         2        <NA>         2
  [3]     chr1 [4829913, 4830000]      + |         3        <NA>         3
  [4]     chr1 [4832340, 4832585]      + |         4        <NA>         4
  [5]     chr1 [4834487, 4834618]      + |         5        <NA>         5
  [6]     chr1 [4837461, 4837854]      + |         6        <NA>         6

$2
GRanges with 6 ranges and 3 metadata columns:
      seqnames             ranges strand | exon_id exon_name exon_rank
  [1]     chr1 [4715105, 4715514]      + |       1      <NA>         1
  [2]     chr1 [4771960, 4772759]      + |       2      <NA>         2
  [3]     chr1 [4829913, 4830000]      + |       3      <NA>         3
  [4]     chr1 [4832340, 4832585]      + |       4      <NA>         4
  [5]     chr1 [4834487, 4834618]      + |       5      <NA>         5
  [6]     chr1 [4842605, 4843851]      + |       7      <NA>         6

---
seqlengths:
 chr1
   NA
```

4 Session Information

```
R version 3.1.1 (2014-07-10)
Platform: x86_64-unknown-linux-gnu (64-bit)

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] parallel  stats     graphics  grDevices utils     datasets  methods
[8] base

other attached packages:
```

```
[1] GenomicFeatures_1.16.2 AnnotationDbi_1.26.0   Biobase_2.24.0
[4] GenomicRanges_1.16.4   GenomeInfoDb_1.0.2     IRanges_1.22.10
[7] BiocGenerics_0.10.0    flipflop_1.2.2

loaded via a namespace (and not attached):
 [1] BBmisc_1.7               BSgenome_1.32.0         BatchJobs_1.3
 [4] BiocParallel_0.6.1       Biostrings_2.32.1       DBI_0.3.0
 [7] GenomicAlignments_1.0.6 Matrix_1.1-4            RCurl_1.95-4.3
[10] RSQLite_0.11.4          Rsamtools_1.16.1        XML_3.98-1.1
[13] XVector_0.4.0           biomaRt_2.20.0          bitops_1.0-6
[16] brew_1.0-6              checkmate_1.4           codetools_0.2-9
[19] digest_0.6.4            fail_1.2                foreach_1.4.2
[22] grid_3.1.1              iterators_1.0.7         lattice_0.20-29
[25] rtracklayer_1.24.2      sendmailR_1.1-2         stats4_3.1.1
[28] stringr_0.6.2           tools_3.1.1             zlibbioc_1.10.0
```

References

P. Aboyoun, H. Pages, and M. Lawrence. *GenomicRanges: Representation and manipulation of genomic intervals*. R package version 1.12.5.

R.K. Ahuja et al. *Network Flows*. Prentice Hall, 1993.

Elsa Bernard, Laurent Jacob, Julien Mairal, and Jean-Philippe Vert. Efficient rna isoform identification and quantification from rna-seq data with network flows. Technical report, HAL, 2013. hal-00803134.

D.P. Bertsekas. *Network Optimization: Continuous and Discrete Models*. Athena Scientific, 1998.

R. Bohnert and G. Rätsch. rQuant.web: a tool for RNA-Seq-based transcript quantitation. *Nucleic Acids Res*, 38(Web Server issue):W348–W351, 2010.

M. Carlson, H. Pages, P. Aboyoun, S. Falcon, M. Morgan, D. Sarkar, and M. Lawrence. *GenomicFeatures: Tools for making and manipulating transcript centric annotations*. R package version 1.12.4.

Y. Huang et al. A robust method for transcript quantification with RNA-Seq data. In *Proceedings of the 16th Annual international conference on Research in Computational Molecular Biology*, RECOMB'12, pages 127–147, 2012.

H. Jiang and W. H. Wong. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, 25(8):1026–1032, 2009.

J. J. Li et al. Sparse linear modeling of next-generation mRNA sequencing (RNA-Seq) data for isoform discovery and abundance estimation. *P Natl Acad Sci USA*, 108(50): 19867–19872, 2011a.

W. Li et al. IsoLasso: a LASSO regression approach to RNA-Seq based transcriptome assembly. *J Comput Biol*, 18:1693–1707, 2011b.

J. Mairal and B. Yu. Supervised feature selection in graphs with path coding penalties and network flows. *preprint arXiv:1204.4539v1*, 2012. to appear in JMLR.

A. M. Mezlini et al. iReckon: Simultaneous isoform discovery and abundance estimation from RNA-seq data. *Genome res*, 23(3):519–529, 2013.

A. Mortazavi et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, 5(7):621–628, 2008.

A. Mortazavi et al. Scaffolding a caenorhabditis nematode genome with RNA-Seq. *Genome Res*, 20(12):1740–1747, 2010.

Q. Pan et al. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*, 40(12):1413–1415, 2008.

C. M. Perou et al. Molecular portraits of human breast tumours. *Nature*, 406(6797): 747–752, 2000.

C. Trapnell et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, 28(5):511–515, 2010.

L. J. van't Veer et al. Gene expression profiling predicts clinical outcome of breast cancers. *Nature*, 415(6871):530–536, 2002.

Z. Xia et al. NSMAP: a method for spliced isoforms identification and quantification from RNA-Seq. *BMC Bioinformatics*, 12:162, 2011.

Y. Xing et al. An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs. *Nucleic Acids Res*, 34(10):3150–3160, 2006.