miRNA tap example use

Maciej Pajak, Ian Simpson

October 17, 2016

Contents

1 Introduction  2
2 Installation  2
3 Workflow  3
4 Session Information  5
References  6
1 Introduction

miRNAtap package is designed to facilitate implementation of workflows requiring miRNA prediction. Aggregation of commonly used prediction algorithm outputs in a way that improves on performance of every single one of them on their own when compared against experimentally derived targets. microRNA (miRNA) is a 18-22nt long single strand that binds with RISC (RNA induced silencing complex) and targets mRNAs effectively reducing their translation rates.

Targets are aggregated from 5 most commonly cited prediction algorithms: DIANA (Maragkakis et al., 2011), Miranda (Enright et al., 2003), PicTar (Lall et al., 2006), TargetScan (Friedman et al., 2009), and miRDB (Wong and Wang, 2015).

Programmatic access to sources of data is crucial when streamlining the workflow of our analysis, this way we can run similar analysis for multiple input miRNAs or any other parameters. Not only does it allow us to obtain predictions from multiple sources straight into R but also through aggregation of sources it improves the quality of predictions.

Finally, although direct predictions from all sources are only available for Homo sapiens and Mus musculus, this package includes an algorithm that allows to translate target genes to other speices (currently only Rattus norvegicus) using homology information where direct targets are not available.

2 Installation

This section briefly describes the necessary steps to get miRNAtap running on your system. We assume that the user has the R program (see the R project at http://www.r-project.org) already installed and is familiar with it. You will need to have R 3.2.0 or later to be able to install and run miRNAtap. The miRNAtap package is available from the Bioconductor repository at http://www.bioconductor.org To be able to install the package one needs first to install the core Bioconductor packages. If you have already installed Bioconductor packages on your system then you can skip the two lines below.

```r
> source("http://bioconductor.org/biocLite.R")
> biocLite()
```

Once the core Bioconductor packages are installed, we can install the miRNAtap and accompanying database miRNAtap.db package by

```r
> source("http://bioconductor.org/biocLite.R")
> biocLite("miRNAtap")
> biocLite("miRNAtap.db")
```
3 Workflow

This section explains how miRNAtap package can be integrated in the workflow aimed at predicting which processes can be regulated by a given microRNA. In this example workflow we’ll use miRNAtap as well as another Bioconductor package topGO together with Gene Ontology (GO) annotations. In case we don’t have topGO or GO annotations on our machine we need to install them first:

```r
> source("http://bioconductor.org/biocLite.R")
> biocLite("topGO")
> biocLite("org.Hs.eg.db")
```

Then, let’s load the required libraries

```r
> library(miRNAtap)
> library(topGO)
> library(org.Hs.eg.db)
```

Now we can start the analysis. First, we will obtain predicted targets for human miRNA miR-10b

```r
> mir = 'miR-10b'
> predictions = getPredictedTargets(mir, species = 'hsa',
+ method = 'geom', min_src = 2)
```

Let’s inspect the top of the prediction list.

```r
> head(predictions)
```

<table>
<thead>
<tr>
<th>source_1</th>
<th>source_2</th>
<th>source_3</th>
<th>source_4</th>
<th>source_5</th>
<th>rank_product</th>
<th>rank_final</th>
</tr>
</thead>
<tbody>
<tr>
<td>627</td>
<td>103</td>
<td>10.0</td>
<td>1.0</td>
<td>NA</td>
<td>1.416281</td>
<td>1</td>
</tr>
<tr>
<td>79741</td>
<td>NA</td>
<td>NA</td>
<td>8.0</td>
<td>2</td>
<td>NA</td>
<td>2</td>
</tr>
<tr>
<td>6095</td>
<td>5</td>
<td>2.5</td>
<td>73.5</td>
<td>NA</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>348980</td>
<td>NA</td>
<td>2.5</td>
<td>20.0</td>
<td>NA</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>51365</td>
<td>NA</td>
<td>53.0</td>
<td>3.0</td>
<td>12</td>
<td>3.766392</td>
<td>5</td>
</tr>
<tr>
<td>7022</td>
<td>88</td>
<td>17.5</td>
<td>5.0</td>
<td>149</td>
<td>3</td>
<td>6</td>
</tr>
</tbody>
</table>
```

We are using geometric mean aggregation method as it proves to perform best when tested against experimental data from MirBase (Griffiths-Jones et al., 2008).

We can compare it to the top of the list of the output of minimum method:

```r
> predictions_min = getPredictedTargets(mir, species = 'hsa',
+ method = 'min', min_src = 2)
> head(predictions_min)
```
<table>
<thead>
<tr>
<th>source_1</th>
<th>source_2</th>
<th>source_3</th>
<th>source_4</th>
<th>source_5</th>
<th>rank_product</th>
<th>rank_final</th>
</tr>
</thead>
<tbody>
<tr>
<td>627</td>
<td>103</td>
<td>10</td>
<td>1.0</td>
<td>NA</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8897</td>
<td>1</td>
<td>183</td>
<td>282.0</td>
<td>NA</td>
<td>NA</td>
<td>1</td>
</tr>
<tr>
<td>79042</td>
<td>NA</td>
<td>107</td>
<td>99.5</td>
<td>1</td>
<td>NA</td>
<td>1</td>
</tr>
<tr>
<td>7182</td>
<td>2</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>106</td>
<td>2</td>
</tr>
<tr>
<td>10739</td>
<td>NA</td>
<td>42</td>
<td>2.0</td>
<td>NA</td>
<td>NA</td>
<td>2</td>
</tr>
<tr>
<td>79741</td>
<td>NA</td>
<td>NA</td>
<td>8.0</td>
<td>2</td>
<td>NA</td>
<td>2</td>
</tr>
</tbody>
</table>

Where predictions for rat genes are not available we can obtain predictions for mouse genes and translate them into rat genes through homology. The operation happens automatically if we specify species as `rno` (for *Rattus norvegicus*)

```r
> predictions_rat = getPredictedTargets(mir, species = 'rno',
+ method = 'geom', min_src = 2)
```

Now we can use the ranked results as input to GO enrichment analysis. For that we will use our initial prediction for human miR-10b

```r
> rankedGenes = predictions[, 'rank_product']
> selection = function(x) TRUE
> # we do not want to impose a cut off, instead we are using rank information
> allGO2genes = annFUN.org(whichOnto='BP', feasibleGenes = NULL,
+   mapping="org.Hs.eg.db", ID = "entrez")
> GOdata = new('topGOdata', ontology = 'BP', allGenes = rankedGenes,
+   annot = annFUN.GO2genes, GO2genes = allGO2genes,
+   geneSel = selection, nodeSize=10)
```

In order to make use of the rank information we will use Kolomogorov Smirnov (K-S) test instead of Fisher exact test which is based only on counts.

```r
> results.ks = runTest(GOdata, algorithm = "classic", statistic = "ks")
> results.ks
```

**Description:**

Ontology: BP

'classic' algorithm with the 'ks' test

599 GO terms scored: 5 terms with p < 0.01

Annotation data:

  Annotated genes: 335
  Significant genes: 335
  Min. no. of genes annotated to a GO: 10
  Nontrivial nodes: 599

We can view the most enriched GO terms (and potentially feed them to further steps in our workflow)

```r
> allRes = GenTable(GOdata, KS = results.ks, orderBy = "KS", topNodes = 20)
> allRes[,c('GO.ID', 'Term', 'KS')]
```
<table>
<thead>
<tr>
<th>GO.ID</th>
<th>Term</th>
<th>KS</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0042692</td>
<td>muscle cell differentiation</td>
<td>0.0016</td>
</tr>
<tr>
<td>GO:0050789</td>
<td>regulation of biological process</td>
<td>0.0023</td>
</tr>
<tr>
<td>GO:0048518</td>
<td>positive regulation of biological process</td>
<td>0.0052</td>
</tr>
<tr>
<td>GO:0050794</td>
<td>regulation of cellular process</td>
<td>0.0079</td>
</tr>
<tr>
<td>GO:0065007</td>
<td>biological regulation</td>
<td>0.0083</td>
</tr>
<tr>
<td>GO:0044087</td>
<td>regulation of cellular component biogenesis</td>
<td>0.0224</td>
</tr>
<tr>
<td>GO:0006352</td>
<td>DNA-templated transcription, initiation</td>
<td>0.0247</td>
</tr>
<tr>
<td>GO:0006367</td>
<td>transcription initiation from RNA polymer</td>
<td>0.0247</td>
</tr>
<tr>
<td>GO:0045944</td>
<td>positive regulation of transcription from RNA polymer</td>
<td>0.0252</td>
</tr>
<tr>
<td>GO:0048522</td>
<td>positive regulation of cellular process</td>
<td>0.0263</td>
</tr>
<tr>
<td>GO:0061061</td>
<td>muscle structure development</td>
<td>0.0264</td>
</tr>
<tr>
<td>GO:0014070</td>
<td>response to organic cyclic compound</td>
<td>0.0316</td>
</tr>
<tr>
<td>GO:0010604</td>
<td>positive regulation of macromolecule metabolism</td>
<td>0.0360</td>
</tr>
<tr>
<td>GO:0031325</td>
<td>positive regulation of cellular metabolism</td>
<td>0.0374</td>
</tr>
<tr>
<td>GO:0080090</td>
<td>regulation of primary metabolic process</td>
<td>0.0401</td>
</tr>
<tr>
<td>GO:0043254</td>
<td>regulation of protein complex assembly</td>
<td>0.0470</td>
</tr>
<tr>
<td>GO:0045893</td>
<td>positive regulation of transcription, DNA</td>
<td>0.0483</td>
</tr>
<tr>
<td>GO:1902680</td>
<td>positive regulation of RNA biosynthetic</td>
<td>0.0483</td>
</tr>
<tr>
<td>GO:1903508</td>
<td>positive regulation of nucleic acid-templ</td>
<td>0.0483</td>
</tr>
<tr>
<td>GO:0010629</td>
<td>negative regulation of gene expression</td>
<td>0.0515</td>
</tr>
</tbody>
</table>

For more details about GO analysis refer to topGO package vignette (Alexa and Rahnenfuhrer, 2010).

Finally, we can use our predictions in a similar way for pathway enrichment analysis based on KEGG (Kanehisa and Goto, 2000), for example using Bioconductor’s KEGGprofile (Zhao, 2012).

4 Session Information

- R version 3.3.1 (2016-06-21), x86_64-pc-linux-gnu
- Locale: LC_CTYPE=en_US.UTF-8, LC_NUMERIC=C, LC_TIME=en_US.UTF-8, LC_COLLATE=C, LC_MONETARY=en_US.UTF-8, LC_MESSAGES=en_US.UTF-8, LC_PAPER=en_US.UTF-8, LC_NAME=C, LC_ADDRESS=C, LC_TELEPHONE=C, LC_MEASUREMENT=en_US.UTF-8, LC_IDENTIFICATION=C
- Base packages: base, datasets, grDevices, graphics, methods, parallel, stats, stats4, utils
- Other packages: AnnotationDbi 1.36.0, Biobase 2.34.0, BiocGenerics 0.20.0, GO.db 3.4.0, IRanges 2.8.0, S4Vectors 0.12.0, SparseM 1.72, graph 1.52.0, miRNAtap 1.8.0, miRNAtap.db 0.99.10, org.Hs.eg.db 3.4.0, topGO 2.26.0
- Loaded via a namespace (and not attached): DBI 0.5-1, RSQLite 1.0.0, Rcpp 0.12.7, chron 2.3-47, grid 3.3.1, gsubfn 0.6-6, lattice 0.20-34,
References


