How to find genes whose expression profile is similar to that of specified genes

April 24, 2017

Introduction

In some cases you have certain genes of interest and you would like to find other genes that are close to the genes of interest. This can be done using the `genefinder` function.

You need to specify either the index position of the genes you want (which row of the expression array the gene is in) or the name (consistent with the `featureNames` of the ExpressionSet).

A vector of names can be specified and matches for all will be computed. The number of matches and the distance measure used can all be specified. The examples will be carried out using the artificial data set, `sample.ExpressionSet`.

Two other options for `genefinder` are `scale` and `method`. The `scale` option controls the scaling of the rows (this is often desirable) while the `method` option controls the distance measure used between genes. The possible values and their meanings are listed at the end of this document.

```r
> library("Biobase")
> library("genefilter")
> data(sample.ExpressionSet)
> igenes<- c(300,333,355,419) ##the interesting genes
> closeg <- genefinder(sample.ExpressionSet, igenes, 10,
+    method="euc", scale="none")
> names(closeg)
[1] "31539_r_at" "31572_at" "31594_at" "31658_at"

The Affymetrix identifiers (since these were originally Affymetrix data) are 31539_r_at, 31572_at, 31594_at and 31658_at. We can find the nearest genes (by index) for any of these by simply accessing the relevant component of `closeg`.

```r
> closeg$"31539_r_at"
$indices
 [1] 220 425 457 131 372 137 380 231 161 38

$dist
 [1] 70.3 70.9 71.7 71.7 73.6 74.8 77.4 77.9 83.6

> Nms1 <- featureNames(sample.ExpressionSet)[closeg$"31539_r_at"$indices]
> Nms1
[1] "31459_i_at" "31664_at" "31696_at" "31370_at"
[5] "31611_s_at" "31376_at" "31619_at" "31470_at"
[9] "31400_at" "AFFX-TrpnX-3_at"
```
You could then take these names (from Nms1) and the annotate package and explore them further. See the various HOWTO's in annotate to see how to further explore your data. Examples include finding and searching all PubMed abstracts associated with these data. Finding and downloading associated sequence information. The data can also be visualized using the geneplotter package (again there are a number of HOWTO documents there).

**Parameter Settings**

The `scale` parameter can take the following values:

- **none** No scaling is done.
- **range** Scaling is done by \((x_i - x_{(1)})/(x_{(n)} - x_{(1)})\).
- **zscore** Scaling is done by \((x_i - \bar{x})/s_x\). Where \(s_x\) is the standard deviation.

The `method` parameter can take the following values:

- **euclidean** Euclidean distance is used.
- **maximum** Maximum distance between any two elements of \(x\) and \(y\) (supremum norm).
- **manhattan** Absolute distance between the two vectors (1 norm).
- **canberra** The \(\sum(|x_i - y_i|/|x_i + y_i|)\). Terms with zero numerator and denominator are omitted from the sum and treated as if the values were missing.
- **binary** (aka asymmetric binary): The vectors are regarded as binary bits, so non-zero elements are on and zero elements are off. The distance is the proportion of bits in which only one is on amongst those in which at least one is on.

**Session Information**

The version number of R and packages loaded for generating the vignette were:

- R version 3.4.0 (2017-04-21), 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
- Running under: Ubuntu 16.04.2 LTS
- Matrix products: default
- BLAS: /home/biocbuild/bbs-3.5-bioc/R/lib/libRblas.so
- LAPACK: /home/biocbuild/bbs-3.5-bioc/R/lib/libRlapack.so
- Base packages: base, datasets, grDevices, graphics, methods, parallel, stats, utils
- Other packages: ALL 1.17.0, Biocbase 2.36.0, BiocGenerics 0.22.0, DESeq 1.28.0, RColorBrewer 1.1-2, class 7.3-14, genefilter 1.58.0, knitr 1.15.1, lattice 0.20-35, locfit 1.5-9.1, pasilla 1.3.1
• Loaded via a namespace (and not attached): AnnotationDbi 1.38.0, BiocStyle 2.4.0, DBI 0.6-1, IRanges 2.10.0, Matrix 1.2-9, RCurl 1.95-4.8, RSQLite 1.1-2, Rcpp 0.12.10, S4Vectors 0.14.0, XML 3.98-1.6, annotate 1.54.0, backports 1.0.5, bitops 1.0-6, codetools 0.2-15, compiler 3.4.0, digest 0.6.12, evaluate 0.10, geneplotter 1.54.0, grid 3.4.0, highr 0.6, htmltools 0.3.5, magrittr 1.5, memoise 1.1.0, rmarkdown 1.4, rprojroot 1.2, splines 3.4.0, stats4 3.4.0, stringi 1.1.5, stringr 1.2.0, survival 2.41-3, tools 3.4.0, xtable 1.8-2, yaml 2.1.14