bioDist Introduction

October 13, 2014

Introduction

The *bioDist* package contains some distance functions that have been shown to be useful in a number of different biological or bioinformatic problems. The return values are typically instances of the S3 class *dist*.

1 Data

We will use the `sample.ExpressionSet` object from the *Biobase* package as our data. The *bioDist* functions are in some ways extensions of the distance functions available via the `dist` function in R, and hence they compute pairwise distances between the rows of the input. For an expression matrix, this will correspond to the genes or features on the array. Since we are generally more interested in distances between samples, we will transpose the data in this demonstration.

```
> library("bioDist")
> data(sample.ExpressionSet)
> exData = t(exprs(sample.ExpressionSet))
```

2 Distance Measures

The two most used distance measures in the *bioDist* package are MI and KLD. These measures focus on very different distributional aspects of the data. MI is large when the joint distribution is quite different from the product of the marginals, while KLD measures how much the shape of one distribution resembles that of the other.

MI can be considered as a multivariate measure of association, and if the transformation

$$\delta^* = [1 - \exp(-2MI)]^{1/2} \tag{1}$$

is used, then δ^* takes values in the interval $[0, 1]$ and can be interpreted as a a generalization of the correlation. We will make the further transformation to $1 - \delta^*$ so that this measure has the same interpretation as other correlation-based distance measures.

There are two functions for computing mutual information distance measures: `mutualInfo` that computes the distance from independence and `MIdist` that computes the transformation in Equation (1). We note that the computations are not terribly fast, and computing these distances on very large data sets is time consuming.

```
> s1 = MIdist(exData)
> s2 = as.matrix(s1)
> dim(s2)

[1] 26 26

> r1 = mutualInfo(exData)
```

For KL distances, there is one implementation that uses binning, `KLdist.matrix`, and one that uses density estimation followed by numerical integration, `KLD.matrix`.

```
> kl1 = KLdist.matrix(exData)
> kl2 = KLD.matrix(exData, method="density", supp=range(exData))
```

The *bioDist* package also provides implementations of distances based on two other measures of correlation: Kendall's tau and Pearson's rho. In the examples below we will measure distance between genes, not between samples as was done in the first few examples. We will also restrict our analysis to the last 100 genes in the sample data in order to keep computing times low.

```
> eS = sample.ExpressionSet[401:500,]
> tauD = tau.dist(eS, sample=FALSE)
> sp = spearman.dist(eS, sample=FALSE)
```

To find a specified number of nearest neighbors, we will use a simple helper function called `closest.top`.

```
> f1 = featureNames(eS)[1]
> closest.top(f1, sp, 3)

[1] "31699_at" "31710_at" "31696_at"
```

3 Session Information

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

```
R version 3.1.1 Patched (2014-09-25 r66681)
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] bioDist_1.38.0      KernSmooth_2.23-13  Biobase_2.26.0
[4] BiocGenerics_0.12.0

loaded via a namespace (and not attached):
[1] tools_3.1.1
```