BHC: Bayesian Hierarchical Clustering

Richard S. Savage

April 14, 2011

The BHC method performs bottom-up hierarchical clustering, using a Dirichlet Process (infinite mixture) to model uncertainty in the data and Bayesian model selection to decide at each step which clusters to merge. This avoids several limitations of traditional methods, for example how many clusters there should be and how to choose a principled distance metric. This implementation accepts multinomial (i.e. discrete, with 2+ categories) data.

The paper 'R/BHC: fast Bayesian hierarchical clustering for microarray data' (Savage et al. 2009, BMC Bioinformatics) contains a quantitative comparison of the performance of the BHC algorithm with that of a conventional agglomerative hierarchical clustering method (using an uncentred correlation coefficient as a distance metric and complete linkage). Gene Ontology (GO) annotations are used to show that BHC produces more biologically-relevant gene clustering results.

Here is an example of how to use the BHC package.

```
> require(graphics)
> require(BHC)
> require(affydata)
```

```
     Package    LibPath                                          Item
[1,] "affydata" "/loc/home/biocbuild/bbs-2.8-bioc/R/library" "Dilution"
     Title
[1,] "AffyBatch instance Dilution"
```

```
> require(gcrma)
> data(Dilution)
> ai <- compute.affinities(cdfName(Dilution))
```

```
Computing affinities[1] "Checking to see if your internet connection works..."

The downloaded packages are in
        '/tmp/RtmpEasunU/downloaded_packages'
[1] "Checking to see if your internet connection works..."

The downloaded packages are in
        '/tmp/RtmpEasunU/downloaded_packages'
.Done.
```

```
> Dil.expr <- gcrma(Dilution, affinity.info = ai, type = "affinities")

Adjusting for optical effect....Done.
Adjusting for non-specific binding....Done.
Normalizing
Calculating Expression

> testData <- exprs(Dil.expr)
> keep <- sd(t(testData)) > 0
> testData <- testData[keep, ]
> testData <- testData[1:100, ]
> geneNames <- row.names(testData)
> nGenes <- (dim(testData))[1]
> nFeatures <- (dim(testData))[2]
> nFeatureValues <- 4
> for (i in 1:nFeatures) {
+     newData <- testData[, i]
+     newData <- (newData - mean(newData))/sd(newData)
+     testData[, i] <- newData
+ }
> for (i in 1:nGenes) {
+     newData <- testData[i, ]
+     newData <- rank(newData) - 1
+     testData[i, ] <- newData
+ }
> hc <- bhc(testData, geneNames, nFeatureValues = nFeatureValues)

> plot(hc, axes = FALSE)


> WriteOutClusterLabels(hc, "labels.txt", verbose = FALSE)
```

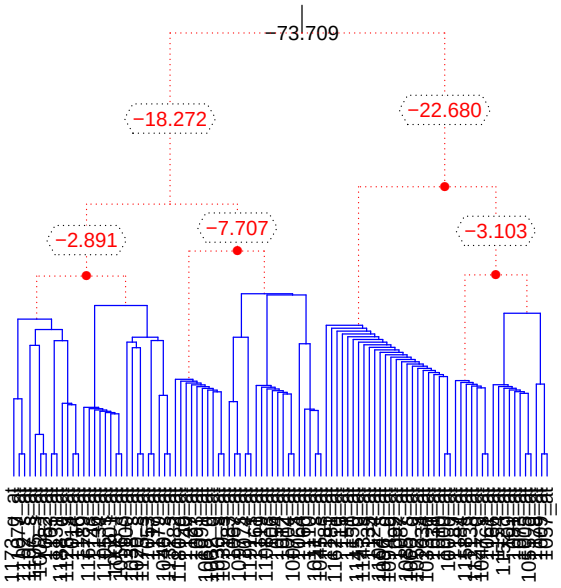

Figure 1: BHC example plot