Software for Heatmaps for Visualizing Phylogenetic Congruence

Edward Susko

Department of Mathematics and Statistics, Dalhousie University

Installation

The software provides R functions for producing heat map plots of p-values for visualizing phylogenetic congruence like those in Susko et al. (2006).

Functions are provided that produce heat maps with two-way hierarchical clustering with (and without) dendrograms indicating the nature of the clustering. Most of these functions are very similar to the `heatmap`, `levelplot` and `image` functions that are part of the R package. Functions are also provided to reduce a matrix of p-values (genes x topologies) to one that only corresponds to "plausible" vertical descent topologies, that are supported by a majority of genes. Functions for cluster size determination are provided as a separate package

`http://www.mathstat.dal.ca/~tsusko`

in `clust_soft.tar.gz`; see `clust_soft.pdf` for additional information.

To make the functions available, with a running R session, issue the command

```
> source("phylcon_fn.q")
```

Examples of use are contained in `phylcon_example.q`. Full output can be obtained at the R command line with

```
> pdf("phylcon.pdf", paper = "letter")
> source("phylcon_example.q")
> dev.off()
```

This will re-direct output to the screen and graphical output to `phylcon.pdf`, which will contain all plots produced. Alternatively, you can cut and paste lines from `phylcon_example.q` to the R command prompt.

R functions

The function `heatmap.d` is a modfication of the `heatmap` function in R and computes heat maps with dendrograms. The function `heatmap.nod` does not add dendrograms. It can be useful when the number of genes or topologies is too large for easy representation of dendrograms. The function `top.subclust` converts a data frame of p-values to one that contains only the p-values for the topolgies that are "supported" by a majority of genes.

Heatmaps with dendrograms

`heatmap.d(x, nclust.row, nclust.col, ...)`

description: Plots a heat map with dendrograms and lines indicating where the clusters are.

arguments:

`x`: p-value matrix. Each colum gives the p-values for each of the topologies for a given gene.

`nclust.row`: The number of clusters of rows (genes). If `nclust.row` = 3, two vertical lines will be drawn indicating the the boundaries of the three clusters of genes.

`nclust.col`: The number of clusters of columns (topologies).

See `phylcon_example.q` and documentation for the R function `heatmap` for information about other arguments.

examples:

```
heatmap.d(t(pmat), nclust.row = 1, nclust.col = 1,
          scale = "none", ylab = "Topology", xlab = "Gene",
          labRow = rep("",dim(t(pmat))[1]), # no row labels
          labCol = rep("",dim(t(pmat))[2]), # no col labels
          col = terrain.colors(20), main = "Example 1 Data",
          breaks = brks)
```

Heatmaps without dendrograms

```
heatmap.nod(x, nclust.row, nclust.col, method = "average", divisive = F,
            ColSideColors, ColSideColors2, ttle, ...)
```

description: Plots a heat map with no dendrogram and lines indicating where the clusters are.

arguments:

`x`: p-value matrix. Each colum gives the p-values for each of the topologies for a given gene.

`nclust.row`: The number of clusters of rows (genes). If `nclust.row = 3`, two vertical lines will be drawn indicating the the boundaries of the three clusters of genes.

`nclust.col`: The number of clusters of columns (topologies).

`method`: The method of hierarchical clustering used. The default is average clustering which in which the distance between two clusters is the average of the distances for all pairs of items in the two clusters; UPGMA is an example. See the documentation of the R function hclust for non-default options.

`divisive`: If `FALSE` clustering proceeds in an agglomerative fashion by grouping together clusters that are similar. If `TRUE` clusters that differ are split apart.

`ColSideColors, ColSideColors2`: On input should give colours for groups of genes required as input by heatmap* functions so that locations of genes after two way clustering can be tracked. See 'col.groupgenes' and 'rainbow' functions. ColSideColors2 is included to allow two sets of groupings to be included.

`ttle`: The title string.

See `phylcon_example.q` and the documentation of the R function `image` for information about additional arguments.

example:

```
rc <- col.groupgenes(dim(pmat)[1], 205) # The first 205 genes were true genes.
heatmap.nod(t(pmat), nclust.row = 1, nclust.col = 1, method = "average",
            xlab = "Gene", ylab = "Topology", ttl = "Example Data",
            ColSideColors = rc, # the colours indicating true genes
            col = terrain.colors(20), breaks = brks, axes = T)
```

Obtaining Topologies which a majoirty of genes cannot reject

```
plaus.top(x, alpha = 0.05)
```

description: Returns the column indices of the topologies that a majority of genes could not reject at level alpha.If there is no such topology, the first entry of the returned vector is -1.

arguments:

`x`: p-value matrix. Each column gives the p-values for each of the genes for a given topology.

`alpha`: A column of `x` for which a majority of rows have entry larger than alpha will be included in idx.

value:

`idx`: The indices of the topologies that a majority of genes could not reject at level alpha.

examples:

```
idxsub <- plaus.top(pmat[1:205,], alpha = 0.05)
pmat.plaus <- pmat[,idxsub]
```

References

Susko, E., Leigh, J., Doolittle, W.F. and Bapteste, E. (2006). Visualizing and assessing phylogenetic congruence of core gene sets: a case study of the γ-proteobacteria. *Molecular Biology and Evolution.* **23**:1019–1030.