MantelCorr Package for Bioconductor

Brian Steinmeyer, MS, and William Shannon, PhD
Department of Internal Medicine
Division of General Medical Sciences
Washington University in St. Louis
School of Medicine
email: steinmeb@ilya.wustl.edu, wshannon@wustl.edu

October 3, 2006

Contents

1 Description 1

2 Example R session with the Golub training data 2

3 Reminder 2

4 GetClusters() function 2

5 DistMatrices() function 2

6 MantelCorrs() function 3

7 PermutationTest() function 3

8 ClusterList() function 3

9 ClusterGeneList() function 3

1 Description

The MantelCorr package is based on the methodology developed in Shannon et al. [1], for which six functions are used to locate and identify important gene clusters from standard microarray expression data with \( p \) genes (rows) and \( n \) samples (columns). Mantel statistics have been applied with success to correlate gene expression levels with clinical covariates [3]. We also include a real microarray dataset with the package to help illustrate its functionality. Specifically, the package makes use of the \texttt{k-means()} function in R (with arbitrary \( k \), say \( k \in [5, \frac{p}{2}] \)) to essentially over-partition the gene space into \( k \) non-overlapping clusters. Next, two types of dissimilarity matrices are computed, one based on the original data \( D_{full} \), and one for each resultant cluster, \( D_{subset(k)} \).

Mantel [2] cluster correlations are then found by correlating each \( D_{subset(k)} \) with \( D_{full} \), resulting in \( k \) Mantel correlations. In order to destroy the distance dependent nature of \( D_{full} \) and to obtain an empirical null distribution of distance independence, a permutation test is done, where the
number of permutations and \( \alpha \) significance level parameters can be chosen by the user. Specifically, the significance level provides the criterion value (\( p\)-value) at which a given cluster is considered significant or non-significant. Both significant and non-significant cluster lists can be viewed with the \( \text{ClusterList} \) function. In addition, a summary list of genes within these clusters can also be seen with the \( \text{ClusterGeneList} \) function.

We next introduce a simple application of the \( \text{MantelCorr} \) package with gene-expression training data taken from the Golub et al. [4] leukemia study.

## 2 Example R session with the Golub training data

The Golub training data consists of gene-expression values measured for 38 samples from Affymetrix Hgu6800 chips on 7,129 genes. There are 27 acute lymphoblastic leukemia (ALL) and 11 acute myeloid leukemia (AML) samples. To load the \( \text{MantelCorr} \) package, simply type `library(MantelCorr)`. The data can be loaded by typing `data(GolubTrain)` and a description provided with `?GolubTrain`.

```r
> library(MantelCorr)
> data(GolubTrain)
> dim(GolubTrain)
[1] 7129 38
> data <- GolubTrain
```

## 3 Reminder

Help on any of the following \( \text{MantelCorr} \) package functions can be viewed by `?FunctionName`, which provides a complete description and overview of the function’s purpose and syntax. In addition, all input ‘data’ values are assumed to be interval-scale (e.g., numeric data), with gene and sample labels assigned from the \( \text{dimnames()} \) function.

## 4 GetClusters() function

The \( \text{GetClusters()} \) function over-partitions the gene-space as described in the package description. We select \( k = 500 \) clusters and store the result in an object called "kmeans.result".

```r
> kmeans.result <- GetClusters(data, 500, 100)
```

## 5 DistMatrices() function

A function used to compute distance matrices \( D_{\text{full}} \) and \( D_{\text{subset}(k)} \) from the \( k \) non-overlapping clusters stored in "kmeans.result". The result is assigned to "DistMatrices.result".

```r
> DistMatrices.result <- DistMatrices(data, kmeans.result$clusters)
```
6 MantelCorrs() function

The MantelCorrs() function uses Dfull and Dsubset(k) to compute a Mantel correlation for each kth cluster by correlating these two dissimilarity matrices. The result is saved in "MantelCorrs.result".

> MantelCorrs.result <- MantelCorrs(DistMatrices.result$Dfull, 
+ DistMatrices.result$Dsubsets)

7 PermutationTest() function

PermutationTest() permutes Dfull to obtain an empirical null distribution for which cluster significance is determined. We have selected 100 permutations in order to conserve CPU time, and chosen an \( \alpha \)-value of 0.05 for the 38 Golub leukemia samples. The result is stored in an object called "permuted.pval". NOTE: we recommend using at least 1000 permutations for a thorough analysis.

> permuted.pval <- PermutationTest(DistMatrices.result$Dfull, DistMatrices.result$Dsubsets, 
+ 100, 38, 0.05)

8 ClusterList() function

A function used to generate a complete list of both significant and non-significant clusters found by the permutation test and associated level of significance. Cluster size and correlation are provided with each type of cluster. We assign the result to the R object "ClusterLists" as follows:

> ClusterLists <- ClusterList(permuted.pval, kmeans.result$cluster.sizes, 
+ MantelCorrs.result)

9 ClusterGeneList() function

A final function that uses information from the "ClusterList" function, coupled with the dimnames function to generate a composite list of the genes found in both cluster types (significant and non-significant). We store the result in R object "ClusterGenes".

> ClusterGenes <- ClusterGeneList(kmeans.result$clusters, ClusterLists$SignificantClusters, 
+ data)

References

