Differential Gene Analysis Package Documentation(DGAP)

Choudary L Jagarlamudi

June 14 2005

Contents

1 Introduction

diffGeneAnalysis package (DGAP) is part of the Bioconductor Project. It can be accessed at www.bioconductor.org.The objective of the package is to analyze gene data from micro arrays.The package can be used on data obtained from cDNA arrays as well as affymetrix chips.We intorduce new algorithms for normalizing the microarray data and adjusting the data for bais that can creep in due to various reasons like dye bias, hybridization efficacy,personnel change etc.This package uses the associative T method for performing the differential gene analysis.Details on this method can be obtained form the paper(i) in the refernces.

2 Usage of DGAP

(i) DGAP is used on the raw gene intensities obtained from micro arrays. The data is first subjected to normalization using our normalize function.

Normalization of data is done utilizing information obtained from background fluorescence. Background fluoresce intensity values are used to determine a Gaussian distribution of lowly expressed genes, yielding the background estimates (mean and standard deviation).

(ii) Bias adjustment of the normalized data is performed using our bias Adjust function that takes a normalized dataset and applies a multiplicative scalar derived from the data to help account for expression biases. These expression biases can come from many sources including dye bias, hybridization efficacy, changes in personnel, etc. After bias adjustment the data is ready for differential gene analysis.

(iii) Differential gene analysis of the normalized, bias adjusted data is done using our refGroup function. refGroup performs differential gene analysis in two steps using the associative-T analysis method. The first step is to compute the reference group, which is used as the internal standard for associative analysis and the second and final step is to apply the associative analysis and classify the gene expressions into four groups.

3 Explanation of Results

The results are displayed in a 10 column matrix as follows:

Column Representation.

1 Gene Bank ID

2 Average Signal of the Control Chips/Channels.

3 SD of Control Chips/Channels.

4 Probability that a given gene in the Control Chips/Channels Belongs or doesnt́ belong to background.

5 Average Signal of the Experimental Chips or Channels.

6 SD of Experimental Chips or Channels.

7 Probability that a given gene in the Experimental Chips or Channels belong or dont́ belong to background.

8 P value from a Student T-test.

9 P value from an Associative T-test.

10 Ratio of mean expression values (Control/Experimental).

11 Group Number.

Group Numbers are defined as follows A1 Expressed above background in both sample types but over expressed on the Experimental Chips/Channels

A2 Expressed above background in both sample types but over expressed on the Control Chips/Channels

A3 Expressed above background only on the Experimental Chips/Channel.

A4 Expressed above background only in the Control Chips/Channel.

0 None of the above.

4 Getting Started

(i) Download the package from bioconductor or directly from our website at http:

(ii) Unzip the downloaded file and save it under your R/rw2000/library folders.

(iii) In your R window type load (diffGeneAnalysis) .

(iv) To get help try help(normalize) (or) help(biasAdjust) (or) help(refGroup)

(v) The package can be used in totality or in combination with other tools. The package is divided into three modules namely: normalize biasAdjust and refGroup. One can use other normalization and bias adjustment tools before using associative analysis method (refGroup). The following steps are required to use this package in its entirety.

(vi) Save the micro array data into an Excel sheet and save it in .csv format.

(vii) The first column always consists of the Gene bank Id. Loss of the first chip data may occur if it is missing.

(viii) The first row consists of row headers. Row headers can be avoided by reading the data into R having the option header=FALSE. Loss of gene 1 data can occur if headers are missing while option header=TRUE. It is advised to enter headers for good documentation and readability purposes.

(ix) The control chips are always saved first after Gene ID column (2nd column onwards), followed by experimental chip data in the excel sheet from left to right. Any other combination will not work. No blank columns or blank rows are allowed.

(x) The following command is suggested to read the data from the saved excel sheet(.csv format) into R. rawdata<-read.csv("c:/Book1.csv",sep=",",header=TRUE). To run the examples load the following data provided with the package. data(rawdata) The following sweave style vignette shows how to load a sample dataset(rawdata) and normalize data,bias adjust the normalized data and perform differential gene expression using our associative analysis method.

```
R> library(diffGeneAnalysis) ##load the diffGeneAnalysis package
```

load the sample dataset (rawdata) to run the example.

```
R>data(rawdata)
```

normalize the sample data set. When prompted with a graph of 6 histograms. Select the best mean by observing the right half of the Gaussian fit. Confirm your slection,else reselect the proper histogram fit. Repeat this procedure for as many chips as the dataset consists of.

```
R>normalized<-normalize(rawdata,7,3,4,0.15,0.60)
```

bias adjust the normalized data set.Please note that this function can take upto 20 seconds to run on a large 22400 gene set on a Pentium IV 1.6Ghz, 128MB Ram.

```
R>bAdjusted<-biasAdjust(normalized,7)
```

Perform the associative analysis on the normalized and bias adjusted data set to differentiate gene expressions into groups. The user will be prompted for E and R values. E stand for the increase in fold over background and R stands for the ratio of experimental chips average over control chips average. The higher these values the higher will be the stringency. Example dataset used here was run with an E values of 1 and R value of 1.5 Please note that this function can take upto 40 seconds to run on a 22400 gene set on a pentium IV 1.6Ghz and 128MB RAM.

```
R>results<-refGroup(bAdjusted,7,3,4,0.05)
```

(xi) normalized<-normalize(rawdata,7,3,4,0.15,0.60) normalizing the data. Rawdata is loaded in the package to use as an example. Read data documentation for details.

(xii) bAdjusted<-biasAdjust(normalized,7) bias adjusting the normalized data. normalized is loaded in the package for use as an example.

(xiii) results<-refGroup(bAdjusted,7,3,4,0.05) performing associative analysis after creating the internal standard(refernce group).

(xiv) Write the results into an excel sheet from R for readability and archiving purposes.

(xv) write.table(results, file = "c:/results.csv", sep = ",", col.names = NA)

(xvi) Open results.csv created in the "c:/" directory/ any other directory that u selected and analyze the results.

Note: In order to use the package in its entirety follow steps (x) to (xvi) in the same sequence. If the user decides to use other normalization and bias adjustment methods, save the normalized/ bias adjusted data in an excel file in the .csv format and follow steps (xii) to (xvi) sequentially.

5 References

(i) Dozmorov I,Centola,M. An associative analysis of gene expression array data. Bioinformatics.2003 Jan22;19(2):204-11

(ii) Knowlton N,Dozmorov I, Centola M. Microarray data Analysis Tool box(MDAT): for normalization, adjustment and analysis of gene expression data. Bioonformatics.2004 Dec 12;20(18):3687-90