Description of MergeMaid

Xiaogang Zhong, Leslie Cope, Elizabeth Garrett-Mayer, Giovanni Parmigiani

April 30, 2008

1 Introduction

MergeMaid is designed to facilitate multi-study analysis. The merging function generates objects that can efficiently support a variety of joint analyses. Visualization tools allow for exploration of the data without requiring normalization across platforms. We have updated the package by adding a quick approximate calculation of the integrative correlation.

Version 2.1.6 of MergeMaid includes the following primary functions, with corresponding data classes

mergeExprs	Merge Datasets into an object of class **mergeExpressionSet**.
intCor	Compute integrative correlation coefficients, returns an object of class **mergeCor**.
modelOutcome	Fit various models to the data, models currently available include linear and logistic regression, and Cox hazards, returns an object of class **mergeCoeff**.

In addition, there are a number of functions for the manipulation, retrieval and visualization of data. These functions depend on the data class for which they are defined and will be discussed below.

The mergeExprs function and the mergeExpressionSet class The primary data class in the MergeMaid package is the `mergeExpressionSet`, based on the ExpressionSet class defined in Bioconductor. 'mergeExprs' returns an object of class `mergeExpressionSet`, required for all analytic functions included in the package. A `mergeExpressionSet` object contains the following slots

data	a list of `ExpressionSet` objects, one per study
geneStudy	incidence matrix indicating which genes are measured in each study.
notes	

The standard way to build a `mergeExpressionSet` object is with the function `mergeExprs`. This function accepts expression data in a variety of formats, including `ExpressionSet` objects, simple matrices of expression values and other `mergeExpressionSets`. Any combination of these is acceptable. Merging is based on user-supplied gene ids (e.g. Genbank, Unigene, or LocusLink ID's). These IDs should make up the rownames for each expression data matrix. Frequently an expression array will include multiple probesets for some genes, and these may be assigned the same geneid. This presents a special problem for the merging of data across platforms, becoming important when carrying out an analysis on the merged data, (e.g. regression or survival analysis) for which genes need to be unambiguously matched. In general, appropriate measures are left up to the user at ID assignment. To prevent potential problems, replicates within a dataset which still share the same ID are averaged during the merging process.

There are a number of functions to access and manipulate the data in a `mergeExpressionSet`.

exprs	returns the contents of the `data` slot
geneStudy	returns the contents of the `data` slot
notes	returns the contents of the `data` slot
names	returns study names
geneNames	returns the entire list of gene IDs
phenoData	returns a list containing the phenodata (if any) included for each study
[	returns a `mergeExpressionSet` object containing only the indicated studies
intersection	returns a single `ExpressionSet` containing all studies and all common genes
notes<-	replaces the contents of the `data` slot
names<-	replaces the study names
geneNames<-	replaces gene IDs.
plot	Draw scatterplots to compare integrative correlations for genes.

The two main analytic functions in the package are defined for `mergeExpressionSet` objects as well, but are discussed in separate sections, as each has an associated class.

The intCor function and the mergeCor class When working with data from different sources is important to identify those genes which are measured in similar ways in the various datasets, and can be used in joint analyses.

MergeMaid includes a gene reproducibility index called the **integrative correlation coefficient** and calculated using the function `intCor`. Within each study, and for each pair of genes, we calculate the correlation coefficient of expression values across subjects. By examining whether, for a specific gene, these correlations agree across studies we can quantify the reproducibility of results without relying on direct comparison of expression across platforms. The integrative correlations provides a reproducibility score for each gene. This analysis is unsupervised in that consistency is measured without using information about sample phenotypes.

The output from the `intCor` function is an object of class `mergeCor`, containing integrative correlation coefficients for a single `mergeExpressionSet` object. Such an object contains the following slots

pairwise.cors	matrix containing the integrative correlation for each pair of studies.
max.cors	vector representing maximal canonical correlation (pairwise canonical correlations) for each pair of studies.

If n is the number of studies then for $i < j \leq n$, the pairwise correlation of correlations for studies i and j is stored in column $(i-1)*(n-1)-(i-2)*(i-1)/2+j-i$ of the pairwise.cors slot.

The *total integrative correlation* for each gene is obtained by averaging the $n(n-1)/2$ pairwise integrative correlations.

The methods available for this class are:

pairwise.cors	Accessor function for the pairwise.cors slot
max.cors	Accessor function for the maximal canonical correlation (pairwise canonical correlations) for each pair of studies.
integrative.cors	Accessor function, returns total integrative correlation for each gene.

In addition, there is a function called `intcorDens`, which plots a smooth density curve for the true distribution of integrative correlation coefficients as well as the null distribution density curve obtained by permuting expression values. These plots can be used to help identify a useful threshold of reproducibility. Since the permutation required the original expression data, this function is defined for mergeExpressionSet objects rather than for mergeCor objects, but in spirit belongs here.

The modelOutcome function and the mergeCoeff class The function `modelOutcome` calculates gene/study specific coefficients for a variety of models. The output from the `modelOutcome` function is an object of class `mergeCoeff` Such an object contains the following slots

coeffs	a matrix of coefficients, rows=genes, columns=studies
coeff.std	matrix of standardized coefficients
zscore	matrix of zscores for the coefficients

Only 3 models are implemented in the first version of MergeMaid: linear regression, logistic regression and cox hazard rate.

Methods for this class include:

coeff	Accessor function for the coeff slot.
coeffstd	Accessor function for the coeff.std slot.
zscore	Accessor function for the zscore slot.
plot	Draw scatterplots to compare coefficients from different studies.

The plot function is actually defined for the matrix class, rather than for the mergeCoeff class. The usual syntax is `plot(coeff(mergeCoeff))` so that the coefficients are selected.