CLEAN: An R package for Clustering Enrichment Analysis

Johannes Freudenberg, Vineet Joshi, Mario Medvedovic

Department of Environmental Health, University of Cincinnati, 3223 Eden Avenue ML 56, Cincinnati OH 45267, USA

Background and Significance

Cluster analysis is frequently applied to DNA microarray data in order to determine groups of similarly expressed genes. It is assumed that such co-expression is a result of co-regulation, involvment in common pathways, or, more broadly, shared membership in a functional category such as biological process, molecular function, or cellular component. Clustering algorithms are inherently unsupervised and do not provide any indication whether or not the underlying biological patterns were recovered. That is, it is not immediately clear whether a particular gene clustering result is biologically meaningful or relevant. In addition clustering algorithms often require users to specify *a-priori* the cluster number or size. For these reasons, an additional analysis step is needed in order to achieve the following goals.

- Given a gene clustering, determine how well clusters relate to biological categories of interest.
- Support cluster prioritization and the choice of cluster number or size.
- Facilitate discovery and prediction of biological pathways and processes underlying a particular experiment or condition.
- Provide a benchmark to compare different clustering methods.

To accomplish these goals we propose a novel compuational approach for assessing functional coherence of clustering results called Clustering Enrichment Analysis. Given both, a set of biological categories represented by corresponding lists of genes and a gene clustering, each gene cluster is compared against all categories and statistically significantly over-represented categories are recorded. The minimum achieved significance level for each cluster is used as an indicator of its biological 'meaningfulness.'

Methods

Computational procedures were implemented as an R package providing the functionality to perform Clustering Enrichment Analysis and to view the results with the new Java-based tool fTreeView. Following are descriptions of the individual features of the package.

Functional categories. These are either derived from publically accessible databases such as Gene Ontology (GO) [1] or Kyoto Encyclopedia of Genes and Genomes (KEGG) [2], or are user-provided sets of gene lists with a corresponding optional table containing category descriptions and other annotations. Gene identifiers are assumed to correspond to the gene identifiers used in the gene clustering.

Determining all clusters. Currently, the gene clustering is assumed to be hierarchical, that is, genes are organized in a binary tree structure where nodes represent gene clusters with leaf nodes corresponding to individual genes. Hence, a list of all possible gene clusters can be obtained by recursively traversing the tree structure and at each node recording the list of corresponding genes. User-provided parameters limit the size of clusters to a range of interest. The approach can be extended to accommodate other, non-hierarchical clustering methods.

Over-representation of functional categories. Given two lists of genes representing a functional category and a cluster, respectively, as well as a background gene list, Fisher's Exact Test is performed and repeated for each category. Resulting *p*-values are adjusted to control the false discovery rate (FDR). A user-specified cutoff-level defines the statistical significance of over-represention of functional categories. Signifant categories are assigned as functional cluster annotion.

Results

The methods described above were implemented as a software package for the statistical programming environment R [4]. The CLEAN package provides functions to perform Cluster Enrichment Analysis, and to import, export, and display the necessary data and annotation files in fTreeView and other formats.

The following R code demonstrates some of the features of the package. The gene clustering of the sample data set was done using the gimmR [5] package.

```
> data(gimmOut)                          #load a sample data set
> funcClustAnnot(gimmOut)                # the typcical use case
> funcClustAnnot(gimmOut, saveDataObjects = T) # save results for later use
> funcClustAnnot(funcCategories = "KEGG")# use KEGG instead of GO
> call.treeview("cluster.cdt")           #display previous results
> funcClustAnnot("cluster.RData", sigFDR=0.01)# more stringend signif. level
> funcClustAnnot(rclust = "minkowski")   # use a different clustering method
> funcClustAnnot("cluster.cdt",cclust=NA)# read expression data from .cdt
                                         # file and do not cluster samples
```

Figure 1. Displaying Clustering Enrichment Analysis results using the Java tool fTreeView.

Methods (cntd.)

Functional Coherence. To measure the functional coherence of a given gene clustering, the per-gene minimum achieved FDR is assesed over all clusters the gene is a member of. To avoid a 'spill-over' effect, less significant super-sets of a cluster with high significance are pruned. The resulting measure is used to prioritize gene clusters of interest and to compare different clustering methods.

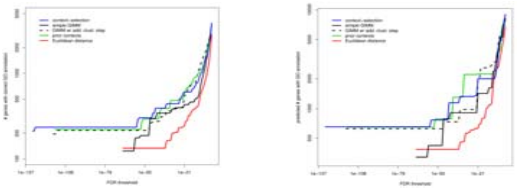
Displaying analysis results. fTreeView is an extension of Java-based Treeview [3]. Clustering Enrichment Analysis results are stored in fTreeView compatible files.

- *.cdt*-file: Contains the gene expression data, gene annotations and information to display the Functional Coherence measure.
- *.gtr*- & *.atr*-file: Contain the gene clustering and sample clustering information, respectively.
- *.fni*-file: Contains functional cluster annotation which is displayed interactively as the user selects a gene cluster of interest.

Results (cntd.)

To demonstrate the features of the CLEAN package we applied our approach to a human breast cancer data set [6]. After pre-processing the raw data we performed cluster analysis and Cluster Enrichment Analysis. **Figure 1** shows a screen shot of fTreeView visualizing the results. The column indicated by **(1)** displays our novel functional coherence measure. The wider the red bar is for a cluster in this column, the higher is the statistical significance of its enrichment for functional categories. Applying this feature the user can select gene clusters of interest by clicking on a subtree indicated by **(2)**. Panel **(3)** then displays the functional annotation obtained for the selected cluster. The CLEAN package can also be employed to compare different gene clustering methods using the per-gene functional coherence measure a benchmark as demonstrated in **Figure 2**. A method is considered superior if it results in more genes in significantly enriched clusters, that is, if the produced clusters are functionally more coherent.

Figure 2. Applying the CLEAN package to compare different gene clustering methods.



Conclusions and Future Directions

Clustering Enrichment Analysis is a computational approach for assessing functional coherence of clustering results. It is readily available as an R software package and can be used for unsupervised functional annotatation of clusters, to facilitate cluster prioritization and discovery of underlying biological patterns, and for comparison of different clustering methods. In its current implementation, the package supports hierarchical methods only but we are planning to generelize our methods to include most other clustering methods as well.

References

[1] The Gene Ontology Consortium (2000) Nature Genet. 25: 25-29. (http://www.geneontology.org)

[2] Kanehisa, M. and Goto, S. (2000) NAR 28, 27-30. (http://www.genome.jp/kegg)

[3] Saldanha AJ. (2004) Bioinformatics 20(17):3246-3248

[4] R Development Core Team (2008). (http://www.R-project.org)

[5] Liu, X. et al. (2006) Bioinformatics 22:1737-44.

[6] Miller LD et al. (2005) Proc Natl Acad Sci U S A 102(38):13550-5.

Acknowledgements

This work is supported by grant 1R01HG003749.