USING NEIGHBORHOOD GRAPHS FOR THE INVESTIGATION OF *E. COLI* GENE CLUSTERS

Theresa Scharl[1,2] *and Friedrich Leisch*[3]

[1]Department of Statistics and Probability Theory, Vienna University of Technology,
Wiedner Hauptstraße 8-10, A-1040 Vienna, Austria
[2]Department of Biotechnology, University of Natural Resources and Applied Life Sciences,
Muthgasse 18, A-1190 Vienna, Austria
[3]Department of Statistics, University of Munich,
Ludwigstraße 33, D-80539 München, Germany
theresa.scharl@ci.tuwien.ac.at, friedrich.leisch@stat.uni-muenchen.de

ABSTRACT

Clustering is commonly used in the analysis of gene expression data to find groups of co–expressed genes. The definition of gene clusters is not very clear as genetic interactions are extremely complex. For this reason the relationship between clusters is very important as co–expressed genes can end up in different clusters. The neighborhood graph is a useful tool to visualize the cluster structure. In this paper the R package `gcExplorer` is presented which is an interactive toolbox for the exploration of gene clusters. Additional information about the gene clusters like the annotation of genes to functional groups (e.g., GO categories) can easily be investigated. The new visualization toolbox is demonstrated on microarray data from *E. coli*.

1. INTRODUCTION

Clusters of co–expressed genes can help to discover potentially co–regulated genes or association to conditions under investigation. Additionally they might suggest pathways or interactions between genes. Cluster analysis is frequently used for the first investigation of a microarray dataset before actually focussing on particular functional subgroups of interest. Gene interactions are extremely complex and the definition of gene clusters is not clear. Further, gene expression data are very noisy and co–expressed genes can easily end up in different clusters. In this context cluster analysis is used as vector quantization as no clear density clusters exist. The data is divided into artificial subsets where the relationship between clusters plays an important role.

The visualization of the cluster structure is important in order to investigate the relationships between clusters. The display of cluster results is very helpful to make cluster analysis useful for practitioners. The Neighborhood graph [1] can be used to display distances between clusters for centroid–based cluster solutions. Microarray data are high–dimensional and complex datasets yielding a high number of clusters. As the linear projection of the data into two dimensions using for example LDA does not scale well in the number of clusters there is a need for new visualization techniques which can handle this situation [2].

In this paper the R package `gcExplorer` is presented which is an interactive toolbox for the exploration of gene clusters. The layout algorithms implemented in the open source graph visualization software Graphviz are used for non–linear arrangement of the clusters. `gcExplorer` contains several possibilities to investigate gene clusters. Further properties of the clusters are included in the neighborhood graph, e.g., cluster size or cluster tightness. Additionally external knowledge from differential expression analysis or functional grouping can be used to investigate the data. `gcExplorer` is currently available at the homepage of the first author (`http://www.ci.tuwien.ac.at/~scharl/Software/`) and will be released as an R package ([3], `http://www.R-project.org`) soon.

The functionality of `gcExplorer` is demonstrated on time–course gene expression data from NCBI Gene Expression Omnibus (GEO, `http://www.ncbi.nlm.nih.gov/geo/`), the *Escherichia Coli* data set with GEO accession number GSE4357-GSE4380 [4]. *E. coli* cells were sampled at several time points (0, 78, 105, 133, 163, 191, 218, 261, 313, 446, 1440 minutes) as

THIS PAPER WAS PUBLISHED AS: THERESA SCHARL AND FRIEDRICH LEISCH. USING NEIGHBORHOOD GRAPHS FOR THE INVESTIGATION OF E. COLI GENE CLUSTERS. IN MIKKA AHDESMÄKI, KORBINIAN STRIMMER, NICOLE RADDE, JÖRG RAHNENFÜHRER, KONSTANTIN KLEMM, HARRI LÄHDESMÄKI AND OLLI YLI-HARJA, EDITORS, PROCEEDINGS OF THE 5TH INTERNATIONAL WORKSHOP ON COMPUTATIONAL SYSTEMS BIOLOGY, WCSB 2008 (JUNE 11-13, 2008, LEIPZIG, GERMANY), PAGES 157–160. PHYSICA VERLAG, TAMPERE UNIVERSITY OF TECHNOLOGY, TAMPERE, FINLAND, 2008.

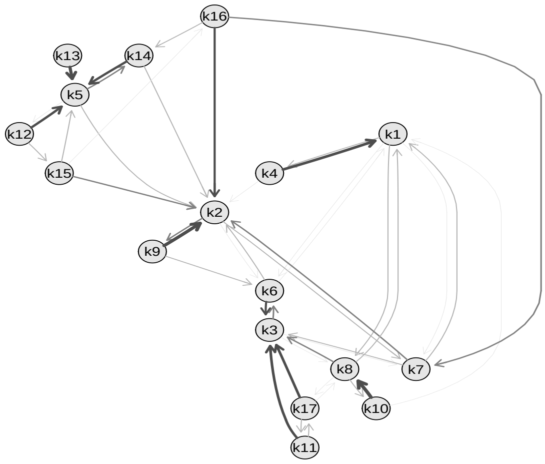


Figure 1. A neighborhood graph of a QT–Clust cluster solution for the *E. coli* data.

they recover from stationary phase versus the Bonner-Vogel medium OD 0.5. After filtering out incomplete and constant observations over time the data set consists of 1672 genes at 11 time points.

2. NEIGHBORHOOD GRAPHS

Neighborhood graphs [1] can be used to visualize cluster solutions of centroid–based cluster algorithms like K–means and PAM or others where clusters can be represented by centroids (e.g., QT–Clust, [5]). For a given data set $X_N = \{x_1, \dots, x_N\}$ the distance between points x and y is given by $d(x, y)$, e.g., the Euclidean or absolute distance. $C_K = \{c_1, \dots, c_K\}$ is a set of centroids and the centroid closest to x is denoted by

$$c(x) = \operatorname*{argmin}_{c \in C_K} d(x, c).$$

Minimizing the average distance between each data point and its closest centroid

$$D(X_n, C_K) = \frac{1}{N} \sum_{n=1}^{N} d(x_n, c(x_n)) \rightarrow \min_{C_K}$$

is the task of most cluster algorithms.

Neighborhood graphs use the mean relative distances between points as edge weights in order to measure how separated pairs of clusters are. Hence they display the distance between clusters. In the graph each node corresponds to a cluster centroid and two nodes are connected by an edge if there exists at least one point that has these two as closest and second–closest centroid.

As described above the centroid closest to x is denoted by $c(x)$ and the second closest centroid to x is denoted by

$$\tilde{c}(x) = \operatorname*{argmin}_{c \in C_K \setminus \{c(x)\}} d(x, c).$$

The set of all points where c_k is the closest centroid is given by

$$A_k = \{x_n | c(x_n) = c_k\}.$$

Now the set of all points where c_i is the closest centroid and c_j is second–closest is given by

$$A_{ij} = \{x_n | c(x_n) = c_i, \tilde{c}(x_n) = c_j\}.$$

For each observation x $s(x)$ is defined as

$$s(x) = \frac{2d(x, c(x))}{d(x, c(x)) + d(x, \tilde{c}(x))}.$$

$s(x)$ is small if x is close to its cluster centroid and close to 1 if it is almost equidistant between the two cluster centroids. The average s–value of all points where cluster i is closest and cluster j is second closest can be used as a proximity measure between clusters and as edge weight in the graph.

$$s_{ij} = \begin{cases} |A_i|^{-1} \sum_{x \in |A_{ij}|} s(x), & A_{ij} \neq \emptyset \\ 0, & A_{ij} = \emptyset \end{cases}$$

$|A_i|$ is used in the denominator instead of $|A_{ij}|$ to make sure that a small set A_{ij} consisting only of badly clustered points with large shadow values does not induce large cluster similarity.

3. SOFTWARE

R package `flexclust` [1] is a flexible toolbox for clustering and contains extensible implementations of the K–centroids and QT–Clust algorithm. The plotting method for cluster solutions in `flexclust` is the neighborhood graph using for example LDA for a linear projection of the data into two dimensions. In `gcExplorer` the neighborhood graph is displayed using non-linear arrangement of the nodes (see for example Figure 1). Bioconductor ([6], `http://www.bioconductor.org`) packages `graph` and `Rgraphviz` [7] provide tools for creating, manipulating, and visualizing graphs in R as well as several non–linear layout algorithms.

3.1. Using gcExplorer

Now the functionality of the interactive software toolbox `gcExplorer` is demonstrated on publicly available *E. coli* time–course gene expression data. The dataset is clustered using the QT–Clust algorithm by the following R commands

```
> library("gcExplorer")
> library("flexclust")
> data("GSE4363")
> cl1 = qtclust(GSE4363, radius = 3, simple = FALSE)
> gcExplorer(cl1, filt = 0.1)
```

The resulting cluster object consists of 17 clusters and the corresponding neighborhood graph is plotted using function `gcExplorer` (see Figure 1). The graph is simplified by using the argument filt. In this case edges between nodes are only drawn if the similarity of a cluster to another cluster is at least 10%. The number of edges pointing from one node to other nodes indicates how distinct the expression profiles are within the corresponding cluster as well as between clusters.

Now there are several possibilities to explore this cluster result. Function gcExplorer is an interactive function if interactive is set equal to TRUE so the clusters can be investigated by clicking on the nodes of the graph. Argument dev offers the possibility to choose if each cluster should be opened in a new window or not. The display method for single clusters is given by the argument panel.function. In the case of expression profiles over time function gcProfile is used as the plotting function. However, any kind of plotting method can be used instead as well as the display of a cluster in form of an html table with links for each gene to databases like NCBI Entrez Gene (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene). The interactive plotting method can be obtained by the following R commands

```
> gcExplorer(cl1, interactive = TRUE, dev = "many",
+     panel.function = gcProfile)
```

There are several possibilities how to include additional information about the clusters in the representation of nodes. The most simple method is to use color coding, e.g., to color nodes by cluster size or cluster tightness.

3.2. Functional Grouping

The annotation of genes to categories or classes is a very important aspect in the analysis of gene expression data. The genes can for example be mapped to functional groups like Gene Ontology (GO, [8]) classifications or to protein complexes. Gene functions are very complex, therefore genes are usually mapped to multiple classes. In any case the mapping is known a priori and does not depend on the experimental data.

External information about the annotation of genes to functional groups can easily be included in the neighborhood graph, e.g., the accumulation of GO classifications in certain gene clusters can be highlighted in the node representation. In the implementation several functional groupings are included, i.e., GO classifications about Biological Process, Molecular Function and Cellular Component, the GenProtEC ([9], http://genprotec.mbl.edu/) classification system for cellular and physiological roles of *E. coli* gene products and some information about operons and regulons from the RegulonDB ([10], http://regulondb.ccg.unam.mx/).

The information of interest can be included in the node representation using the corresponding node.function. Function node.go is used to highlight clusters with accumulation of certain gene functions. The functional group of interest is passed to node.go by the argument node.args. In this example genes assigned to the GO Biological Process group (”gobp”) Metabolism (GO number 8152) are highlighted. This is obtained by the following R commands

```
> gcExplorer(cl1, interactive = TRUE, dev = "many",
+     panel.function = gcProfile,
+     node.function = node.go,
+     node.args = list(gonr = "8152",
+                      source = "gobp"))
```

In Figure 2 a screenshot of an analysis of the *E. coli* data using gcExplorer is given. Nodes of clusters containing genes involved in metabolism are highlighted. Clusters 2, 5, 12 and 13 contain a large number of genes related to metabolism. Clusters 1, 3, 9, 14 and 16 contain a few genes related to metabolism. In the top right of the screenshot an html table of cluster 5 is shown containing links to NCBI Entrez Gene. Additionally the expression profiles of several clusters involved in metabolism are shown. The expression profiles are given with the 11 time points on the x–axis and gene expression on the y–axis. A legend containing the corresponding gene symbols is added to each plot.

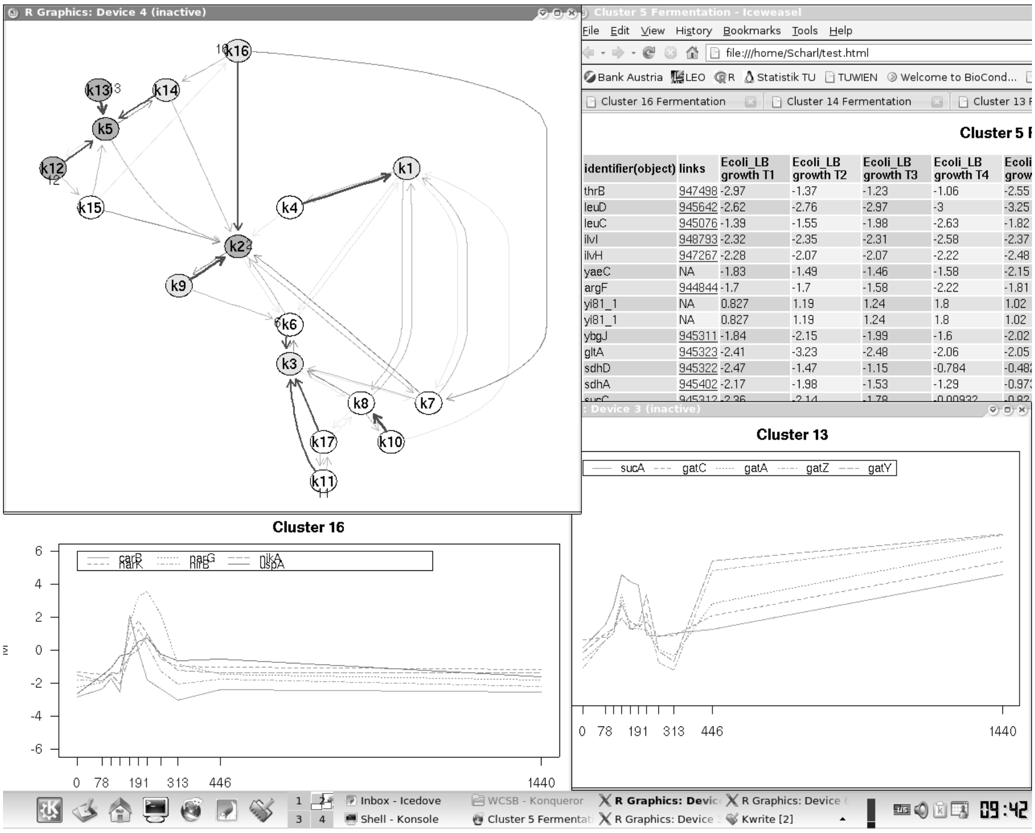
4. SUMMARY

Cluster analysis is commonly used to find groups of co–regulated genes in a microarray dataset without prior knowledge about the gene functions. However, by clustering expression profiles groups of genes with similar biological function are found. For this reason clustering provides a good initial investigation of the data before actually focussing on groups of genes associated to conditions under investigation. As the definition of gene clusters is not very clear and genetic interactions are extremely complex the relationship between clusters is very important as co–expressed genes can end up in different clusters.

In this paper an interactive toolbox for the investigation of gene clusters was presented. Neighborhood graphs were found useful instruments for the investigation of the underlying cluster structure and for gaining insight into the relationships between clusters. gcExplorer is very helpful not only for statisticians but also for practitioners to extract useful information from microarray experiments. It allows not only to visualize the cluster structure, beyond the gene clusters are plotted or shown in html tables with links to databases. Additional properties of the clusters like cluster size or cluster tightness can be highlighted as well as external information like functional grouping. Further extensions of the software are work in progress like the generalization to arbitrary organisms.

5. ACKNOWLEDGMENTS

The project was funded by the Austrian K_{ind}/K_{net} Center of Biopharmaceutical Technology (ACBT).



Cluster 5 F

identifier(object)	links	Ecoli_LB growth T1	Ecoli_LB growth T2	Ecoli_LB growth T3	Ecoli_LB growth T4	Ecoli grow
thrB	947498	-2.97	-1.37	-1.23	-1.06	-2.55
leuD	945642	-2.62	-2.76	-2.97	-3	-3.25
leuC	945076	-1.39	-1.55	-1.98	-2.63	-1.82
ilvI	948793	-2.32	-2.35	-2.31	-2.58	-2.37
ilvH	947267	-2.28	-2.07	-2.07	-2.22	-2.48
yaeC	NA	-1.83	-1.49	-1.46	-1.58	-2.15
argF	944844	-1.7	-1.7	-1.58	-2.22	-1.81
yi81_1	NA	0.827	1.19	1.24	1.8	1.02
yi81_1	NA	0.827	1.19	1.24	1.8	1.02
ybgJ	945311	-1.84	-2.15	-1.99	-1.6	-2.02
gltA	945323	-2.41	-3.23	-2.48	-2.06	-2.05
sdhD	945322	-2.47	-1.47	-1.15	-0.784	-0.48
sdhA	945402	-2.17	-1.98	-1.53	-1.29	-0.97
sucC	945312	-2.36	-2.14	-1.78	-0.00932	-0.82

Figure 2. Screenshot of the functionality of `gcExplorer`.

6. REFERENCES

[1] F. Leisch, “A toolbox for k-centroids cluster analysis,” *Computational Statistics and Data Analysis*, vol. 51, pp. 526–544, 2006.

[2] T. Scharl and F. Leisch, “Visualizing gene clusters using neighborhood graphs in R.,” *Department of Statistics: Technical Reports*, 2008, http://epub.ub.uni-muenchen.de/2110/.

[3] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2007, ISBN 3-900051-07-0.

[4] D. P. Sangurdekar, F. Srienc, and A. B. Khodursky, “A classification based framework for quantitative description of large-scale microarray data,” *Genome Biology*, vol. 7, 2006.

[5] L. J. Heyer, S. Kruglyak, and S. Yooseph, “Exploring expression data: Identification and analysis of coexpressed genes,” *Genome Research*, vol. 9, pp. 1106–1115, 1999.

[6] R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, and S. Dudoit, Eds., *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Statistics for Biology and Health. Springer-Verlag, New York, 2005, ISBN 978-0-387-25146-2.

[7] V. J. Carey, R. Gentleman, W. Huber, and J. Gentry, “Bioconductor software for graphs,” in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, and S. Dudoit, Eds., Statistics for Biology and Health. Springer-Verlag, New York, 2005, ISBN 978-0-387-25146-2.

[8] The Gene Ontology Consortium, “Gene ontology: tool for the unification of biology.,” *Nature Genetics*, vol. 25, pp. 25–29, 2000.

[9] M. Serres, S. Goswami, and M. Riley, “Genprotec: an updated and improved analysis of functions of escherichia coli k-12 proteins.,” *Nucleic Acids Res.*, vol. 32, pp. D300–2, 2004.

[10] H. Salgado, S. Gama-Castro, M. Peralta-Gil, E. Diaz-Peredo, F. Sanchez-Solano, A. Santos-Zavaleta, I. Martinez-Flores, V. Jimenez-Jacinto, C. Bonavides-Martinez, J. Segura-Salazar, A. Martinez-Antonio, and J. Collado-Vides, “Regulondb (version 5.0): Escherichia coli k-12 transcriptional regulatory network, operon organization, and growth conditions.,” *Nucleic Acids Res.*, vol. 34, pp. D394–7, 2006.