Bibliography

B. Efron, R. Tibshirani, J. Storey and V. Tusher. Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association*, 96, 1151–1160, 2001.

R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry and S. Dudoit, editors. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York, 2005.

D. Scholtens and A. von Heydebreck. Analysis of Differential Gene Expression Studies. In: R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, S. Dudoit, editors. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York, 229–248, 2005.

H. Schwender. Modifying Microarray Analysis Methods for Categorical Data – SAM and PAM for SNPs. In: C. Weihs, W. Gaul, editors. *Classification – The Ubiquitous Challenge*. Springer, Heidelberg, 370–377, 2005.

H. Schwender, A. Krause and K. Ickstadt. Comparison of the Empirical Bayes and the Significance Analysis of Microarrays. *Technical Report*, SFB 475, University of Dortmund, Germany, 2003. URL http://www.sfb475.uni-dortmund.de/berichte/tr44-03.pdf.

J. D. Storey and R. Tibshirani. Statistical Significance for Genomewide Studies. *Proceedings of the National Academy of Sciences*, 100, 9440–9445, 2003.

V. Tusher, R. Tibshirani and G. Chu. Significance Analysis of Microarrays Applied to the Ionizing Radiation Response. *Proceedings of the National Academy of Science*, 98, 5116–5121, 2001.

Holger Schwender, Katja Ickstadt
Department of Statistics, SFB 475
University of Dortmund, Germany
holger.schwender@udo.edu,
ickstadt@statistik.uni-dortmund.de

Andreas Krause
Pharsight Corporation
Mountain View, CA, USA
akrause@pharsight.com

Reverse Engineering Genetic Networks using the GeneNet Package

by Juliane Schäfer, Rainer Opgen-Rhein, and Korbinian Strimmer

GeneNet is a package for analyzing high-dimensional (time series) data obtained from high-throughput functional genomics assays, such as expression microarrays or metabolic profiling. Specifically, **GeneNet** allows to infer large-scale gene association networks. These are graphical Gaussian models (GGMs) that represent multivariate dependencies in biomolecular networks by means of partial correlation. Therefore, the output of an analysis conducted by **GeneNet** is a graph where each gene corresponds to a node and the edges included in the graph portray direct dependencies between them.

GeneNet implements a specific learning algorithm that allows to estimate GGMs from small sample high-dimensional data that is both computationally as well as statistically efficient. This approach relies on analytic shrinkage estimation of covariance and (partial) correlation matrices and on model selection using (local) false discovery rate multiple testing. Hence, **GeneNet** includes a computational algorithm that decides which edges are to be included in the final network, in dependence of the *relative* values of the pairwise partial correlations.

In a recent comparative survey (Werhli et al., 2006) the **GeneNet** procedure was found to recover the topology of gene regulatory networks with similar accuracy as computationally much more demanding methods such as dynamical Bayesian networks (Friedman, 2004).

We note that the approach implemented in **GeneNet** should be regarded as an exploratory approach that may help to identify interesting genes (such as "hubs") or clusters of genes that are functionally related or co-regulated, rather than that it yields the precise network of mechanistic interactions. Therefore, the resulting network topologies need be interpreted and validated in the light of biological background information, ideally accompanied by further integrative analysis employing data from different levels of the cellular system.

Prerequisites

GeneNet is available from the CRAN repository and from the webpage http://strimmerlab.org/software/genenet/. It requires prior installation of four further R packages also found on CRAN: **corpcor**, **longitudinal**, **fdrtool**, and **locfdr** (Efron, 2004).

For installation of the required packages simply enter at the R prompt:

```
> install.packages( c("corpcor",
     "longitudinal", "fdrtool",
     "locfdr", "GeneNet") )
```

Preparation of Input Data

The input data must be arranged in a matrix where columns correspond to genes and where rows correspond to the individual measurements. Note that the data must already be properly preprocessed, i.e. in the case of expression data calibrated and normalized.

In the following we describe an example for inferring the gene association network among 102 genes from a microarray data set on the microorganism *Escherichia coli* with observations at 9 time points (Schmidt-Heck et al., 2004). These example data are part of **GeneNet**:

```
> library("GeneNet")
> data(ecoli)
> dim(ecoli)
[1]   9 102
```

Shrinkage Estimators of Covariance and (Partial) Correlation

The first step in the inference of a graphical Gaussian model is the reliable estimation of the partial correlation matrix:

```
> inferred.pcor <- ggm.estimate.pcor(ecoli)
> dim(inferred.pcor)
[1] 102 102
```

For this purpose, the function `ggm.estimate.pcor` offers an interface to a shrinkage estimator of partial correlation implemented in the **corpcor** package that is statistically efficient and can be used for analyzing small sample data. By default, the option `method="static"` is selected, which employs the function `pcor.shrink`. Standard graphical modeling theory (e.g. Whittaker, 1990) shows that the matrix of partial correlations $\tilde{\mathbf{P}} = (\tilde{\rho}_{ij})$ is related to the inverse of the covariance matrix $\mathbf{\Sigma}$. This relationship leads to the straightforward estimator

$$\tilde{r}_{ij} = -\hat{\omega}_{ij} / \sqrt{\hat{\omega}_{ii}\hat{\omega}_{jj}}, \tag{1}$$

where

$$\hat{\mathbf{\Omega}} = (\hat{\omega}_{ij}) = \hat{\mathbf{\Sigma}}^{-1}. \tag{2}$$

In Equation 2, it is absolutely crucial that the covariance is estimated accurately, and that $\hat{\mathbf{\Sigma}}$ is well conditioned – otherwise the above formulae will result in a rather poor estimate of partial correlation (cf. Schäfer and Strimmer, 2005a). For this purpose, the `pcor.shrink` function uses an analytic shrinkage estimator of the correlation matrix developed in Schäfer and Strimmer (2005b). This linearly combines the unrestricted sample correlation with a suitable correlation target in a weighted average. Selecting this target requires some diligence: specifically, we choose to shrink the empirical correlations $\mathbf{R} = (r_{ij})$ towards the identity matrix, while empirical variances are left intact. In this case the analytically determined shrinkage intensity is

$$\lambda^{\star} = \frac{\sum_{i \neq j} \mathrm{var}(r_{ij})}{\sum_{i \neq j} r_{ij}^2}. \tag{3}$$

The resulting shrinkage estimate exhibits a number of favorable properties. For instance, it is much more efficient, always positive definite, and well conditioned. It is inexpensive to compute and does not require any tuning parameters, as the analytically derived optimal shrinkage intensity is estimated from the data. Moreover, there are no assumptions about the underlying distributions of the individual estimates, except for the existence of the first two moments. These properties carry over to derived quantities, such as partial correlations. Furthermore, the resulting estimates are in a form that allows for fast computation of their inverse using the Woodbury matrix identity.

Note that the function `ggm.estimate.pcor` also allows the specification of a `protect` argument, with default value `protect=0`. This imposes limited translation (Efron and Morris, 1972) onto the specified fraction of entries of the estimated shrinkage correlation matrix, thereby protecting those components against overshrinkage (see also Opgen-Rhein and Strimmer, 2006c).

Taking Time Series Aspects Into Account

Standard Gaussian graphical models assume i.i.d. data whereas in practice many expression data sets result from time course experiments. One possibility to generalize the above procedure correspondingly is to employ dynamic (partial) correlation (Opgen-Rhein and Strimmer, 2006a). This is available in the function `ggm.estimate.pcor` by specifying the option `method="dynamic"`, which in turn relies on the **longitudinal** package for computation.

The key difference between dynamical and i.i.d. correlation is that the former takes into account the time that has elapsed between two subsequent measurements. In particular, dynamical correlation allows for unequally spaced time points as often encountered in genomic studies. All small sample learning procedures (shrinkage) developed for i.i.d.

correlation also carry over to dynamical correlation (Opgen-Rhein and Strimmer, 2006b).

Network Search and Model Selection

The second crucial part of gene association network inference is model selection, i.e. assigning statistical significance to the edges in the GGM network:

```
> test.results <-
             ggm.test.edges(inferred.pcor)
> dim(test.results)
[1] 5151    6
```

For this purpose a mixture model,

$$f(\tilde{r}) = \eta_0 f_0(\tilde{r};\kappa) + (1-\eta_0) f_A(\tilde{r})\,, \qquad (4)$$

is fitted to the observed partial correlation coefficients $\tilde{r}$ using the subroutine `cor.fit.mixture`. f_0 is the distribution under the null hypothesis of vanishing partial correlation, η_0 is the (unknown) proportion of "null edges", and f_A the distribution of observed partial correlations assigned to actually existing edges. The latter is assumed to be an arbitrary nonparametric distribution that vanishes for values near zero. This allows for κ, η_0, and even f_A to be determined from the data – see Efron (2004) for an algorithm.

Subsequently, two-sided p-values corresponding to the null hypothesis of zero partial correlation are computed for each potential edge using the function `cor0.test`. Large-scale simultaneous testing is then conducted by obtaining q-values via the function `fdr.control` with the specified value of η_0 taken into account. `fdr.control` uses the algorithms described in Benjamini and Hochberg (1995) and Storey (2002). An alternative to the q-value approach is to use the empirical Bayes local false discovery rate (fdr) statistic (Efron, 2004). This fits naturally with the above mixture model setup, and in addition takes account of the dependencies among the estimated partial correlation coefficients. The posterior probability that a specific edge exists given $\tilde{r}$ equals

$$\mathbb{P}(\text{non-null edge}|\tilde{r}) = 1 - \text{fdr}(\tilde{r}) = 1 - \frac{\eta_0 f_0(\tilde{r};\kappa)}{f(\tilde{r})}. \qquad (5)$$

Following Efron (2005), we typically consider an edge to be "significant" if its local fdr is smaller than 0.2, or equivalently, if the probability of an edge to be "present" is larger that 0.8:

```
> signif <- test.results$prob > 0.80
> sum(signif)
[1] 66
> test.results[signif,]
```

Network Visualization

The network plotting functions in **GeneNet** rely extensively on the infrastructure offered by the **graph** and **Rgraphviz** packages (cf. contribution of Seth Falcon in this R News issue).

First, a `graph` object must be generated containing all significant edges:

```
> node.labels <- colnames(ecoli)
> gr <- ggm.make.graph(
     test.results[signif,],
     node.labels)
> gr
A graphNEL graph with undirected edges
Number of Nodes = 102
Number of Edges = 66
```

Subsequently, the resulting object can be inspected by running the command

```
> show.edge.weights(gr)
```

Finally, the gene network topology of the graphical Gaussian model can be visualized using the function `ggm.plot.graph`:

```
> ggm.plot.graph(gr,
      show.edge.labels=FALSE,
      layoutType="fdp")
```

The plot resulting from the analysis of the `ecoli` data is shown in Figure 1. For `show.edge.labels=TRUE` the partial correlation coefficients will be printed as edge labels. Note that on some platforms (e.g. Windows) the default `layoutType="fdp"` may not yet be available. In this case an alternative variant such as `"layoutType=neato"` needs to be specified.

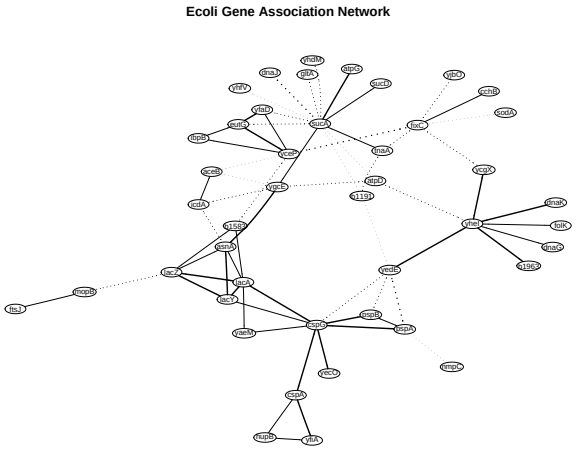

Figure 1: Sparse graphical Gaussian model for 102 genes inferred from an *E. coli* microarray data set with 9 data points. Full and dotted lines indicate positive and negative partial correlation, respectively.

Release History of GeneNet and Example Scripts

The package **GeneNet** emerged from a reorganization of the (now obsolete) package **GeneTS**. This was split into the **GeneNet** part dealing with gene network reconstruction, and the package **GeneCycle** for cell cycle and periodicity analysis (Wichert et al., 2004; Ahdesmäki et al., 2005).

On the home page of **GeneNet** we collect example scripts in order to guide users of **GeneNet** when conducting their own analyses. Currently, this includes the above *E. coli* data but for instance also a network analysis of *A. thaliana* diurnal cycle genes. We welcome further contributions from the biological community.

Bibliography

M. Ahdesmäki, H. Lähdesmäki, R. Pearson, H. Huttunen, and O. Yli-Harja. Robust detection of periodic time series measured from biological systems. *BMC Bioinformatics*, 6:117, 2005.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, 57:289–300, 1995.

B. Efron. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Am. Statist. Assoc.*, 99:96–104, 2004.

B. Efron. Local false discovery rates. Preprint, Dept. of Statistics, Stanford University, 2005.

B. Efron and C. N. Morris. Limiting the risk of Bayes and empirical Bayes estimators – part II: The empirical Bayes case. *J. Am. Statist. Assoc.*, 67:130–139, 1972.

N. Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303:799–805, 2004.

R. Opgen-Rhein and K. Strimmer. Inferring gene dependency networks from genomic longitudinal data: a functional data approach. *REVSTAT*, 4:53–65, 2006a.

R. Opgen-Rhein and K. Strimmer. Using regularized dynamic correlation to infer gene dependency networks from time-series microarray data. In *Proceedings of the 4th International Workshop on Computational Systems Biology (WCSB 2006)*, 12-13 June 2006, Tampere, volume 4, pages 73–76, 2006b.

R. Opgen-Rhein and K. Strimmer. Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. In review.

J. Schäfer and K. Strimmer. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21:754–764, 2005a.

J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statist. Appl. Genet. Mol. Biol.*, 4(1):Article 32, 2005b.

W. Schmidt-Heck, R. Guthke, S. Toepfer, H. Reischer, K. Duerrschmid, and K. Bayer. Reverse engineering of the stress response during expression of a recombinant protein. In *Proceedings of the EUNITE symposium, 10-12 June 2004, Aachen, Germany*, pages 407–412, 2004. Verlag Mainz.

J. D. Storey. A direct approach to false discovery rates. *J. R. Statist. Soc. B*, 64:479–498, 2002.

A. Werhli, M. Grzegorczyk, and D. Husmeier. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks. *Bioinformatics*, 22:2523–2531, 2006.

J. Whittaker. *Graphical Models in Applied Multivariate Statistics*. Wiley, New York, 1990.

S. Wichert, K. Fokianos, and K. Strimmer. Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics*, 20:5–20, 2004.

Juliane Schäfer, ETH Zurich, Switzerland
Rainer Opgen-Rhein, University of Munich, Germany
Korbinian Strimmer, University of Munich, Germany
juliane.schaefer@stat.math.ethz.ch
opgen-rhein@stat.uni-muenchen.de
korbinian.strimmer@lmu.de