

Recent advances in genomic studies

Recenti sviluppi negli studi di genomica

Chiara Romualdi

Abstract Current demand for understanding the behaviour of groups of related genes, combined with the greater availability of data, has led to an increased focus on statistical methods in gene set analysis. In this talk, we will review the most recent advances in the field, focussing in particular on the methodology that exploits the potential of the graphical models theory to incorporate information contained into pathways signalling networks, with the aim to develop statistically sound procedures for gene set analysis.

Abstract *La crescente necessità di comprensione del comportamento di gruppi di geni, unita alla ampia disponibilità di dati ha accresciuto l'attenzione verso i metodi statistici per la gene-set analysis. In questo intervento verranno passati in rassegna i principali contributi in tema di analisi statistica di insieme di geni, soffermandosi in particolare sulle metodologie basate sui modelli grafici.*

Key words: gene set analysis, graphical models, pathway, structured covariance.

1 Background

Recently, much attention has been directed towards the study of gene sets in the context of microarray data analysis (hereafter GSA). A microarray experiment typically provides a list of differentially expressed genes (DEGs) that represent the starting point of a highly challenging process of results interpretation. The grouping of genes into functionally related entities is of great help for interpreting the results. In this context, statistical methods for the identification of groups of functionally related genes with moderate, but coordinated, expression changes are fundamental to help biologists in the process of results comprehension.

Chiara Romualdi, Department of Biology, University of Padova, Italy, e-mail: chiara.romualdi@unipd.it

Several GSA tests, both univariate and multivariate, have been recently developed. GSA methods can be divided into two broad categories: (i) methods based on enrichment analysis performed on a list of genes selected through a gene-level test; (ii) methods based on global and multivariate approaches that define a model on the whole gene set (Ackermann and Strimmer, 2009). In general these two approaches are based on two fundamentally different null hypotheses: the first type hypothesizes the same level of association of a gene set with the given phenotype as the complement of the gene set. The second type only considers the genes within a gene set and hypothesizes that there is no gene in the gene set associated with the phenotype (Tian *et al.*, 2005). Goeman and Buhlmann (2007) termed these approaches *competitive* and *self-contained*, respectively.

In general, the *a priori* definition of gene sets is obtained from Gene Ontology (GO) information or from biological pathways; while genes belonging to a GO category do not have any explicit connections among them (apart from being involved in the same function), genes in the same pathway are structured in a network with explicit biological interactions. Almost all of the *self-contained* approaches, when applied to biological pathways, use merely the list of genes belonging to a pathway, and therefore, although effective, miss the relevant topological information contained.

Although pathway topology is a valuable well-defined experimentally validated biological information describing the relationships among genes, in the last years little effort has been done (Draghici *et al.*, 2007; Tarca *et al.*, 2009; Laurent *et al.*, 2012) to consider the topological information within the *self-contained* GSA methods. This information could greatly improve the analysis and help result interpretation. In this contribution, we will review some of the most recent advances, based on Gaussian graphical models theory, that explicitly take such form of *a priori* information into account.

2 From pathway to graph

Biological pathways can be described as sets of linked biological components interacting with each other over time to generate a single biological effect. They comprise a myriad of interactions, reactions, and regulations, which are often identified piecemeal over extended periods and by a variety of researchers. Figure 1 represents an example of such pathways, the Prostate cancer pathway taken from the Kyoto Encyclopedia of Genes and Genomes (KEGG).

A vast variety of databases containing information such as that shown in Figure 1 exist. Among the widely used databases, we find Biocarta, KEGG (Ogata *et al.*, 1999), NCI/Nature Pathway Interaction Database (Schaefer *et al.*, 2009) and Reactome, (Matthews *et al.*, 2009).

Pathway annotations is often too rich for the conversion to a graph. In particular, challenges are posed by i) the presence of chemical compounds mediating inter-

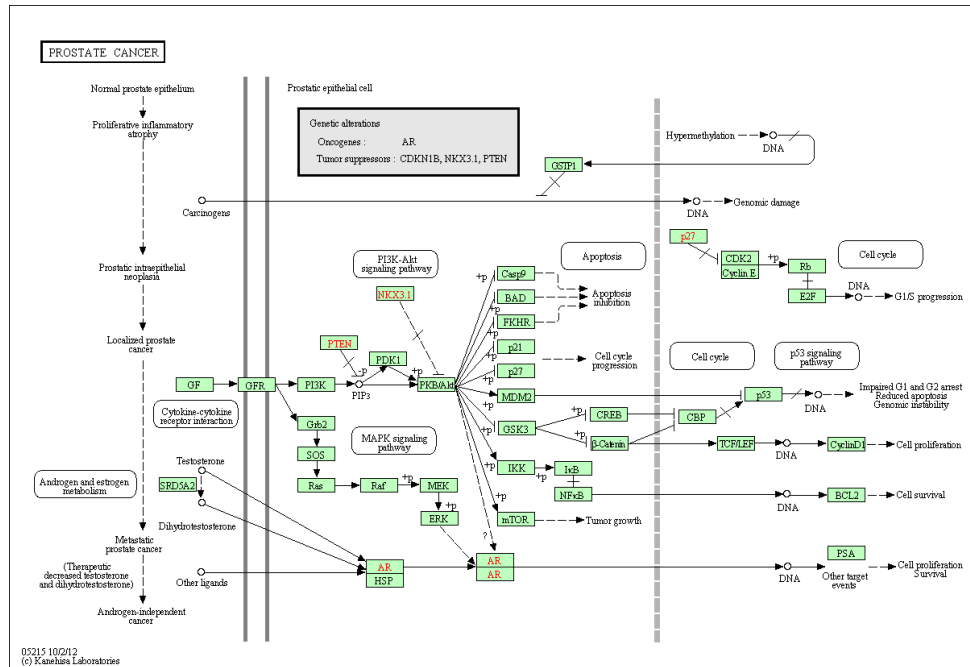


Fig. 1 Prostate cancer: pathway taken from the Kyoto Encyclopedia of Genes and Genomes (KEGG).

actions (interactions for which a compound acts as a bridge between two elements) and by ii) different types of gene groups that are usually represented as single nodes.

Sales *et al.* (2012) recently developed *graphite* (GRAPH Interaction from pathway Topological Environment), a Bioconductor package that provides graphs from the pathways of the main four databases. *graphite* expands the protein complexes into a clique (all proteins connected to the others) and gene families into a module without connections among the element of the group. Moreover, it propagates the signal through compounds considering compound compartment.

The final output of the pathway conversion is a graph-based structure with as many nodes as the number of genes and as many edges as the number of biological reactions among them. There may be different type of reactions and they can be directed or undirected.

3 Graph Inference using pathway topology

Within the graphical models context, data are considered as coming from Gaussian multivariate distributions with a structured concentration matrix (inverse of the covariance matrix), which reflects dependencies among variables. In this context

Massa *et al.* (2010) recently introduced two statistical tests for comparing pathways under different experimental conditions, which naturally stem from the adopted theoretical framework. The first one addresses the question of testing whether the strength of the connections among genes is altered in different experimental conditions. It is likely to figure that a pathological condition does not change the structure of a pathway, but, rather, can influence the strength of the biochemical reactions. The second test is more traditionally designed for testing for differential expression. In doing the test, the procedure specifically employs the information about the behaviour of the partial correlations among genes and about their possible heteroschedasticity in different experimental conditions. We stress that the two tests can be performed independently one from each other. Of course, if one performs both of them, there is a suggested natural order, but the particular research question will define if they are both necessary or if, in a preliminary phase, only one of them is requested. This approach is based on a very different null hypothesis with respect to the method proposed by Tarca *et al.* (2009) called SPIA. First of all, SPIA uses only differentially expressed genes and their average log fold-changes. Then, SPIA combines two independent p -values. The first one captures the significance of the given pathway as provided by an over-representation analysis of the number of differentially expressed genes (hypergeometric test). The second one is derived from a topological analysis called Impact Analysis. The impact analysis calculates a score giving more weight to the differentially expressed genes lying near the entry point of the pathway, following the hypothesis that closer the deregulation to the upstream region of the pathway, more impacted will be the pathway itself. The significance level of the impact score is calculated using a bootstrap approach.

3.1 The case of $n > p$

The structure of the pathway is converted into a simple directed acyclic graph (DAG) D (see Cowell *et al.*, 1999, pag. 46). Starting from D , its moral graph G is derived (see Cowell *et al.*, 1999, pag. 48). In detail, the moral graph is obtained by adding edges between the parents of each node (if not already present), and, then, by removing the directionality of the original edges. It is important to note that the moral graph can differ from the corresponding DAG, in the sense that it usually has more edges and that they are all undirected, but the choice of working with a moral graph does not affect the purpose of this study.

For the sake of exemplification, in the following we will assume to have two experimental conditions. Let $y_k = (y_k^j)$, $j = 1, \dots, n_k$, be p -variate observations for each condition $k, k = 1, 2$. Such data are modelled as graphical Gaussian models (see Lauritzen, 1996, Chap. 5) with the same undirected graph G , i.e., the Gaussian models

$$\begin{aligned}\mathcal{M}_1(G) &= \{Y_1 \sim N_p(\mu_1, \Sigma_1), \Sigma_1^{-1} \in S^+(G)\} \quad \text{and} \\ \mathcal{M}_2(G) &= \{Y_2 \sim N_p(\mu_2, \Sigma_2), \Sigma_2^{-1} \in S^+(G)\}\end{aligned}$$

are assumed. Here, p is the number of genes (vertices of the graph) and $S^+(G)$ is the set of symmetric positive definite matrices with null elements corresponding to the missing edges of G . Note that the assumption of normality of the data is motivated by the well known fact that relative or absolute gene expression measurements are approximately normal on the log scale.

To evaluate the hypothesis of equality of the strength of the links between genes in two experimental conditions the hypothesis of interest is

$$H_0 : \Sigma_1^{-1} = \Sigma_2^{-1} \quad \text{vs} \quad \Sigma_1^{-1} \neq \Sigma_2^{-1}. \quad (1)$$

The hypothesis of equal expression of the pathway can be explored by testing

$$H_0 : \mu_1 = \mu_2 \quad \text{vs} \quad H_1 : \mu_1 \neq \mu_2 \quad \text{subject to} \quad \Sigma_1^{-1} = \Sigma_2^{-1}, \quad (2)$$

or

$$H_0 : \mu_1 = \mu_2 \quad \text{vs} \quad H_1 : \mu_1 \neq \mu_2 \quad \text{subject to} \quad \Sigma_1^{-1} \neq \Sigma_2^{-1}. \quad (3)$$

Tests for the hypotheses (1), (2) and (3) can be derived by generalizing the standard results in the case of a complete graph, i.e., a graph in which all nodes are connected, to the case of a non complete graph G , i.e., a graph with r edges, say, and $r < p(p+1)/2$. The tests are to be opportunely adjusted by introducing a constrained estimation of the covariance matrices such that their inverse matrices have zeroes corresponding to the missing edges of G . This is the usual setting for Gaussian graphical models when the graph G is known. In this context, the estimates of the covariance matrices, $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ can be obtained by running the Iterative Proportional Scaling algorithm (IPS, see Lauritzen, 1996, pag. 134) on S_1 and S_2 , respectively. The IPS algorithm guarantees that the estimated matrices are positive definite and their inverse have null elements corresponding to the missing edges of the graph.

Under H_0 in (1), let $\hat{S}$ be the fitted covariance matrix obtained via the IPS algorithm using estimated pooled covariance matrix S as starting point of the algorithm and let $\hat{K} = \hat{S}^{-1}$. Under H_1 , let $\hat{K}_1 = \hat{\Sigma}_1^{-1}$ and $\hat{K}_2 = \hat{\Sigma}_2^{-1}$ be the inverse matrices of the fitted covariance matrices via IPS (using as starting points S_1 and S_2 , respectively). The test statistic, denoted as Λ_G is the equivalent of the usual likelihood ratio test statistic (see Anderson, 2003), but takes into account the structure of the graph G . It is given by

$$\Lambda_G = n_1 \log \left(\frac{\det \hat{K}_1}{\det \hat{K}} \right) + n_2 \log \left(\frac{\det \hat{K}_2}{\det \hat{K}} \right),$$

and its asymptotic null distribution is χ^2_{r+p} (see Massa *et al.*, 2010, for more details).

The Hotelling test statistic to test hypothesis (2) is given by

$$T_G^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{y}_1 - \bar{y}_2)^T \hat{S}^{-1} (\bar{y}_1 - \bar{y}_2).$$

In the Behrens Fisher setting of hypothesis (3), the test statistic is given by

$$BF_G^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{y}_1 - \bar{y}_2)^T \left(\frac{\hat{S}_1}{n_1} + \frac{\hat{S}_2}{n_2} \right)^{-1} (\bar{y}_1 - \bar{y}_2).$$

In case of paired data, let $y = (y_1^j - y_2^j)$, $j = 1, \dots, n$, be the p -variate differences to be analyzed. Such data are modelled as a graphical Gaussian model

$$\mathcal{M}(G) = \{Y \sim N_p(\mu, \Sigma), \Sigma^{-1} \in S^+(G)\}.$$

The hypothesis to be tested in this case is

$$H_0 : \mu = 0,$$

for which standard theory applies.

For the test statistics described above, a null distribution can be worked out, at least asymptotically. However, following the lines in Janssen (1997), even when the underlying distributions have different variances and with different sample sizes, it is easy to see that permutation tests based on the previous statistics can allow one to construct tests that asymptotically attain the nominal rejection probability when the underlying distributions are different, while maintaining exact rejection probability in finite samples when the underlying distributions are equal. Therefore, to compute p -values for the above given tests, it is possible to rely on permutation strategies.

The tests described above are implemented in an R package called `topologyGSA`. Furthermore, `topologyGSA` can be run automatically on a list of pathways of a specific database from `graphite`.

3.2 The case of $n < p$

As reported above, the IPS guarantees that the estimated matrices belong to $S^+(G)$. In this case, a necessary condition for the existence of the maximum likelihood estimate is that the number of samples is greater than the cardinality of the largest clique (see Lauritzen, 1996, pag. 133), a setting that is easily missed in case of gene expression data.

To estimate the covariance matrix in such circumstances, Martini *et al.* (2013) introduced a shrinking procedure in the estimation of the sample covariance matrices. Apart from increased efficiency, the shrunk estimates have the additional advantage of being always positive definite and well conditioned. The authors use a James-Stein-type shrinkage estimator, as implemented in `corpcor` R package (Schafer and Strimmer, 2005; Opgen-Rhein and Strimmer, 2007).

The shrunk estimates are passed on to the IPS algorithm. The use of a shrinkage estimator, however, precludes the use of the asymptotic distribution of the log-likelihood ratio test. Thus, a permutational approach on the samples is used.

Even if the IPS algorithm implemented in `qpgraph` (Castelo and Roverato, 2006) is one of most computationally efficient, in some cases (very large and com-

plex pathways) it is highly computationally demanding (e.g. for various pathways, the IPS algorithm takes even several days to converge) and sometimes it has problems of convergence. Therefore, the authors have also investigated the possibility of computing the maximum likelihood estimate of the covariance matrices using the approach of Friedman *et al.* (2008), implemented in the R package `glasso`, where they have specified the indices of entries of the inverse covariance matrix to be constrained zero and set the regularization parameter equal to zero. In the examined cases, the estimates of the covariance matrices obtained by `glasso` with no regularization and with the IPS algorithm were the same and no significant improvement in the computational efficiency were found.

4 Looking deeper within the pathway

One of the major advantages of the approach proposed in Massa *et al.* (2010) is the possibility to break down the problem of differences among experimental conditions into testing relevant subgraphs, while accounting for the multiplicity issue arising from the number of subgraphs. In fact, if the graph is decomposable (see Cowell *et al.*, 1999, pag. 50), it is possible to decompose it into its maximal complete subgraphs (cliques) and repeat the previous two tests for each clique. If the graph is not decomposable, it is always possible to add extra edges in order to obtain a triangulation of the graph and therefore a graph that is decomposable (see Cowell *et al.*, 1999, pag. 57). Under decomposability, the density function is uniquely expressible as the product of marginal density functions on cliques and these marginal densities are now the focus of interest. The cliques are, in fact, complete graphs (i.e., all the nodes are connected by an edge), so that standard inference on homogeneity of variances and equality of means can be performed (see previous section).

This decomposition allows to finely identify the portion of the pathway that is really involved in the experimental condition. Following this idea, Martini *et al.* (2013) proposed an empirical approach, called `clipper`, that, exploiting the structure of a junction tree derived from the graph decomposition, identifies signal paths within the entire pathway mostly correlated with the phenotype.

Junction tree construction builds a new hyper-tree having cliques as nodes and satisfying the *running intersection property* according to which, for any cliques C_1 and C_2 in the tree, every clique on the path connecting C_1 and C_2 contains $C_1 \cap C_2$. Using the structure of this hyper-tree, `clipper` identifies *paths* and *sub-paths* ranked according to a score that we called *relevance*. Briefly, the authors define a *path* as the shortest path connecting the root clique with a leaf clique (identified by maximum cardinality search algorithm). For each clique along the paths, the *p*-value of the test on homoscedasticity is considered as weight w of the clique. Then, the *relevance* score is a function of the *p*-values of the cliques belonging to the path. Higher the *relevance* more associated to the phenotype is the *path*. For a given graph, there could be more than one junction tree. Therefore, the root of

the junction tree is forced to be in agreement with the biological structure of the pathway.

5 The time course experimental design without replicates

Time course gene expression experiments are widely used to study the dynamics of biological processes. Usually, the main goal of such experiments is to identify genes modulated along a biological process or after a system perturbation (such as drug treatments or genetic modifications). However, time course data are costly and usually long time series have few or no replicates. Although several statistical models have been proposed to account for differential expression in the context of time series without replicates (Storey *et al.*, 2005; Han *et al.*, 2007; Wu and Wu, 2013), none of them were proposed in the context of pathway analysis.

Martini *et al.* (2014) proposed a two-steps approach based on `clipper`, called `timeClip`, to test pathway time-dependency. Briefly, given that a pathway is composed by multiple genes, to reduce the dimension of a whole or of a portion of a pathway, the authors used principal component analysis. Then, the first principal component is explored for temporal variation.

5.1 Exploring the whole pathway

Let $Y_{n \times t}$ be the normalized log transformed gene expression matrix with genes on the rows and experiments (equal to time points t) on the columns. Let $Y_{p \times t}^P$ the sub-matrix of genes belonging to pathway P . Pathway P has p genes. Then, on the transpose of Y^P , $Y^{P'}$, the authors perform principal component analysis (PCA). They used both the classical (R package `stats`) and the robust (`rrcov` R package) version of PCA. Let $Z_{t \times p}^P$ be the scores matrix and $L_{p \times t}^P$ the loadings matrix. They call $Z_1^P, \dots, Z_p^P$ the p principal components. In this way, the first PCs summarize the temporal variation of the genes in pathway P (if present). Thus, from now on we will indicate Z_i^P as $Z_i^P(t)$.

Then, for irregularly sampled time series the authors assume that the irregularly sampled signal $Z_1^P(t)$ can be decomposed as $Z(t) = p(t) + \varepsilon(t)$, where $p(t)$ is a deterministic function, hereafter called “trend”, and $\varepsilon(t)$ is the realization of a stationary stochastic process with mean zero. Extensive exploratory analysis suggests that a reasonable choice for the trend component is a polynomial of degree 2 in t , i.e.,

$$p(t) = \beta_0 + \beta_1 t + \beta_2 t^2$$

with β_1 capturing existing temporal behaviors of $Z_1^P(t)$ and β_2 correcting for potential non linearities. Moreover, the authors assume that $\varepsilon(t)$ follows a continuous-time Gaussian autoregressive process of order 1. The model is fitted using gen-

eralized least squared (as implemented in `nlme` R package). The representative p -value of pathway P is then taken to be the p -value of the test of nullity of β_1 (obtained by a t-test as implemented in the `glm` function of the `nlme` R package). Bonferroni correction is used to adjust p -values for multiple tests.

5.2 Decomposing the pathway

Pathways declared as time-dependent in step 1 are then moralized, triangulated and decomposed into a junction tree as described in Martini *et al.* (2013).

A clique k of pathway P , noted as C_k^P (with $k = 1, \dots, K$), is composed by a subset of genes in P , c_k^P . Let $Y_{c_k^P}^P$ be the sub-matrix of Y corresponding to the genes of the clique C_k^P . For each clique k of P we apply the same approach as described in step 1: PCA transformation and then a linear model with polynomial trend and autoregressive process of order 1 on the first PCs. The p -value of clique k in pathway P , $p_{C_k^P}$ is given by the p of the β_1 of the polynomial regression. Finally, the best time-dependent paths within a pathway P , hereafter called S_{P_j} , $j = 1, \dots, J$, are identified using the relevance measure as described in previously (Martini *et al.*, 2013).

6 Future perspectives

The flexibility of the graphical modeling approach, when applied to gene set analysis, allows to tackle key questions in the analysis of the behaviour of groups of genes, as shown by the contributions described in this talk. Clearly, effectiveness of the approach increases by increasing the quality of the a priori information and of the data. When pathways information lacks of trustable evidence, data might help to acknowledge and adjust for the discrepancies in the structure of the pathway that are provided by different sources and to discover novel genetic interactions. Strategies to refine the structure of the pathways become therefore crucial to amplify the future potential of such approaches.

References

- Ackermann, M. and Strimmer, K. (2009). A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, **10**, 47.
- Anderson, T. W. (2003). *An introduction to multivariate statistical analysis*. Wiley, New York.
- Castelo, R. and Roverato, A. (2006). A robust procedure for gaussian graphical model search from microarray data with p larger than n . *Journal of Machine Learning Research*, **7**, 2621–2650.

- Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems*. Springer-Verlag, New York.
- Draghici, S., Khatri, P., Tarca, A. L., Amin, K., Done, A., Voichita, C., Georgescu, C., and Romero, R. (2007). A systems biology approach for pathway level analysis. *Genome Research*, **17**(10), 1537–1545.
- Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**(3), 432–441.
- Goeman, J. J. and Buhlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**(8), 980–987.
- Han, X., Sung, W., and Feng, L. (2007). Identifying differentially expressed genes in time-course microarray experiment without replicate. *J Bioinf Comput Biol*, **5**, 281–296.
- Janssen, A. (1997). Studentized permutation tests for non-i.i.d. hypotheses and the generalized behrens-fisher problem. *Statistics and Probability Letters*, **36**(9), 21.
- Laurent, J., Pierre, N., and Dudoit, S. (2012). Gains in power from structured two-sample tests of means on graphs. *Annals of Applied Statistics*, page In press.
- Lauritzen, S. L. (1996). *Graphical models*. Clarendon Press, Oxford.
- Martini, P., Sales, G., Massa, M. S., Chiogna, M., and Romualdi, C. (2013). Along signal paths: an empirical gene set approach exploiting pathway topology. *Nucleic Acids Research*, **41**(1), e19.
- Martini, P., Sales, G., Calura, E., Cagnin, S., Chiogna, M., and Romualdi, C. (2014). timeclip: pathway analysis for time course data without replicates. *BMC Bioinformatics*, **In Press**.
- Massa, M. S., Chiogna, M., and Romualdi, C. (2010). Gene set analysis exploiting the topology of a pathway. *BMC Systems Biology*, **4**(1), 121.
- Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., Kanapin, A., Lewis, S., Mahajan, S., May, B., Schmidt, E., Vastrik, I., Wu, G., Birney, E., Stein, L., and D'Eustachio, P. (2009). Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.*, **37**, 619–622.
- Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **27**, 29–34.
- Opgen-Rhein, R. and Strimmer, K. (2007). Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Statistical Applications in Genetics and Molecular Biology*, **6**(9).
- Sales, G., Calura, E., Cavalieri, D., and Romualdi, C. (2012). graphite - a bioconductor package to convert pathway topology to gene network. *BMC Bioinformatics*, **13**(1), 20.
- Schaefer, C., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and Buetow, K. (2009). Pid: the pathway interaction database. *Nucleic Acids Res*, **37**(Database issue), D674–9.
- Schafer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, **4**(32).

- Storey, J., Xiao, W., Leek, J., Tompkins, R., and Davis, R. (2005). Significance analysis of time course microarray experiments. *Proc National Acad Sci USA*, **102**(36), 12837–12842.
- Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J.-s., Kim, C. J., Kusanovic, J. P., and Romero, R. (2009). A novel signaling pathway impact analysis. *Bioinformatics*, **25**(1), 75–82.
- Tian, L., Greenberg, S. A., Kong, S. W., Altschuler, J., Kohane, I. S., and Park, P. J. (2005). Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(38), 13544–13549.
- Wu, S. and Wu, H. (2013). More powerful significant testing for time course gene expression data using functional principal component analysis approaches. *BMC Bioinformatics*, **14**(1), 6.