ISBN 978-82-575-1009-1
ISSN 1503-1667

Norwegian University of Life Sciences
NO–1432 Ås, Norway
Phone +47 64 96 50 00
www.umb.no, e-mail: postmottak@umb.no

Guro Dørum

Norwegian University of Life Sciences • Universitetet for miljø- og biovitenskap
Department of Chemistry, Biotechnology and Food Science
Philosophiae Doctor (PhD) Thesis 2011:46

Philosophiae Doctor (PhD) Thesis 2011:46

Knowledge-based methods handling complex dependency structures - Applications to gene expression data

Kunnskapsbaserte metoder som håndterer komplekse avhengighetsstrukturer - Anvendelser på genekspresjonsdata

Guro Dørum

Knowledge-based methods handling complex dependency structures - Applications to gene expression data

Kunnskapsbaserte metoder som håndterer komplekse avhengighetsstrukturer - Anvendelser på genekspresjonsdata

Philosophiae Doctor (PhD) Thesis

Guro Dørum

Dept. of Chemistry, Biotechnology and Food Science
Norwegian University of Life Sciences

Ås 2011

Thesis number 2011:46
ISSN 1503-1667
ISBN 978-82-575-1009-1

Acknowledgements

This thesis has been carried out in the period 2007-2011 at the department of Chemistry, Biotechnology and Food Science (IKBM) at the Norwegian University of Life Sciences (UMB) under supervision by Solve Sæbø and Lars Snipen.

First of all I would like to thank Solve for all the time and effort he has put into my PhD thesis, especially during the last hectic weeks - helping me keep my head just above water. Thank you for invaluable guidance and for managing to see things clearly surprisingly fast. Secondly I would like to thank Lars for his philosophical thoughts about a wide range of topics that get one's own thoughts going. I am very grateful for having had such a skilled and inspirational supervisor team.

So many people have left a mark during these four years I have been working with the thesis. I would like to thank Trygve Almøy for always showing interest in whatever I am working with, and the many late afternoon discussions triggered by what was intended to be a quick question from me (some which we learned from and some that turned out to be slightly less useful). Kristian Hovde Liland I would like to thank for being an R guru and for in general having answers to most things that comes to mind. Thanks to Margrete Solheim for giving biological alibis to the papers. Thanks to Trygve and Simen for Thursday/Tuesday morning chocolate and coffee with interesting conversation topics. I would further like to thank my fellow PhD student colleagues for being pleasant officemates and conference/summer school companions, and especially Julia for great creative collaboration on the statistics rap. How fortunate that we didn't have much else to do for those few weeks! Thanks to the members of the Biostatistics group for making it enjoyable to come to work every day, and for after-hours social activities including the annual cabin trip in the mountains (with sometimes more dramatic outcomes such as PhD students crashing the rental van or group leaders getting locked in at shopping centres).

Last, but not least I would like to thank my family and friends for being supportive and encouraging, and for just being there. Thanks to Ida for motivating me to finish in time for the IndoMalay trip, to my sister Siri for being a good role model and to mum and dad for paying for the party! ☺

Ås, June 2011

Guro Dørum

List of papers

I. Dørum, G., Snipen, L., Solheim, M. and Sæbø, S. (2009) *Rotation testing in Gene Set Enrichment Analysis for small direct comparison experiments*, Statistical Applications in Genetics and Molecular Biology **8**(1), Article 34

II. Dørum, G., Snipen, L., Solheim, M. and Sæbø, S. (2011) *Smoothing gene expression data with network information improves consistency of regulated genes*, Statistical Applications in Genetics and Molecular Biology, **10**(1), Article 37

III. Dørum, G., Snipen, L., Solheim, M. and Sæbø, S. *Rotation gene set testing for longitudinal expression data*, Submitted manuscript

IV. Dørum, G. and Sæbø, S. *Improved preprocessing for rotation gene set testing for longitudinal expression data*, Manuscript

Summary

Microarray gene expression data are usually associated with a large number of correlated variables measured on few samples. This type of data typically contain high levels of noise, and the biological signals may be difficult to extract. The classical approach for analysing gene expression data is to test individual genes for differential expression. This basically implies performing tests on possibly thousands of dependent variables while incorrectly assuming statistical independence. The probability of doing false positive discoveries is accordingly high, the results of the analysis may be difficult to reproduce, and the outcome may be a list of biologically unrelated genes that leaves *very much* to the imagination.

An increasing number of publications have therefore started to focus on incorporating prior biological information about gene dependencies in the analysis of gene expression data. Vast amounts of knowledge about relationships between genes based on previous studies are available. The motivation behind analysing the data in light of this information, include increased sensitivity and robustness of the analysis, better reproducibility of the results and easier interpretation. The prior information can for example be groups of genes with a similar function, or gene networks that describe some relationship between genes. With this information in hand, the focus can be turned from identifying important individual genes, to identifying larger groups of important genes that are also related.

The aim of this thesis has been to improve and adapt existing methods to accommodate gene expression data from various types of experimental designs, in addition to developing novel procedures that incorporate prior information. A central part of this work has been concerned with significance testing in data sets with few and dependent samples. Most existing methods in this field use permutation tests to assess significance when the distribution of the test statistics is unknown. This is however problematic in data sets with very small sample sizes and complex experimental designs. In paper I we adopt a popular method for analysing gene sets, and replace the permutation test with a rotation test to accommodate it to small sample sizes. Paper III and IV introduce improvements to the method in paper I by adapting it to data from complex experimental designs and time series data. In paper II we propose a novel method that uses gene networks to improve test statistics for individual genes.

Sammendrag

Genekspresjonsdata fra mikromatriser assosieres ofte med et stort antall korrelerte variabler målt på få observasjoner. Denne typen data inneholder vanligvis mye irrelevant variasjon, og de biologiske signalene kan være vanskelig å skille fra bakgrunnsstøyet. Den vanligste måten å analysere geneekspresjonsdata på, har vært å teste hvert enkelt gen for differensiell ekspresjon. Dette innebærer å utføre tester på potensielt tusenvis av avhengige variabler, samtidig som man antar statistisk uavhengighet. Sannsynligheten for å finne falske positive er tilsvarende høy, resultatene kan være vanskelig å reprodusere, og utfallet av analysen kan være en liste med gener uten biologisk relasjon som overlater *veldig mye* til fantasien.

Et økende antall publikasjoner har derfor begynt å fokusere på inkludering av *a priori* informasjon om genavhengigheter i analyse av genekspresjonsdata. Fra tidligere studier finnes store mengder biologisk kunnskap om relasjoner mellom gener. Ved å analysere dataene i lys av denne informasjonen, ønsker man å oppnå en mer sensitiv og robust analyse med resultater som er enklere å reprodusere og tolke. Forhåndsinformasjonen kan for eksempel bestå av grupper av gener med lignende funksjon eller gennettverk som beskriver relasjoner mellom gener. Med denne informasjonen for hånden, kan fokuset flyttes fra viktige enkeltgener, til grupper av viktige gener som også har noe felles.

Målet med denne avhandlingen har vært å forbedre og tilpasse eksisterende metoder til genekspresjonsdata med forkjellige typer forsøksdesign, samt utvikling av nye metoder som benytter seg av *a priori* informasjon. En sentral del av dette arbeidet har vært knyttet til testing av signifikans i datasett med få og avhengige observasjoner. De fleste eksisterende metoder innenfor dette feltet bruker permutasjonstester for å evaluere signifikans når testobservatoren har en ukjent fordeling. Dette er imidlertid problematisk for datasett med veldig få observasjoner som ikke kan antas uavhengige grunnet forsøksdesignet. I artikkel I tar vi for oss en populær metode for å analysere gengrupper og bytter ut permutasjonstesten med en rotasjonstest for å tilpasse metoden til små utvalgsstørrelser. I artikkel III og IV introduseres forbedringer av metoden i artikkel I ved å tilpasse den til data med komplekse forsøksdesign og tidsseriedata. I artikkel II foreslår vi en ny metode som bruker gennettverk til å forbedre testobservatoren til enkeltgener.

Contents

1 Introduction

1.1 Background

The last couple of decades have seen a revolution in the field of biology with the introduction of "omics" techniques - such as genomics, transcriptomics, proteomics and metabolomics - that can measure the complete set of DNA, RNA, proteins or metabolites in a cell or tissue. Figure 1 shows the hierarchy of the most important omics approaches and which part of the cell they are studying. Genomics at the bottom of the pyramid is a relatively well studied field, while metabolomics at the top is less explored so far. Common for these high-throughput technologies is that they give rapid determination of a large number of variables per sample. However, the number of samples is often limited by financial or practical interests, and the variables may be highly correlated. The generation of such high-dimensional data sets has introduced an increasing need for multivariate statistical methods that can handle these types of data. The combination of many correlated variables and few samples causes problems for the classical statistical methods that rely on having more samples than variables,

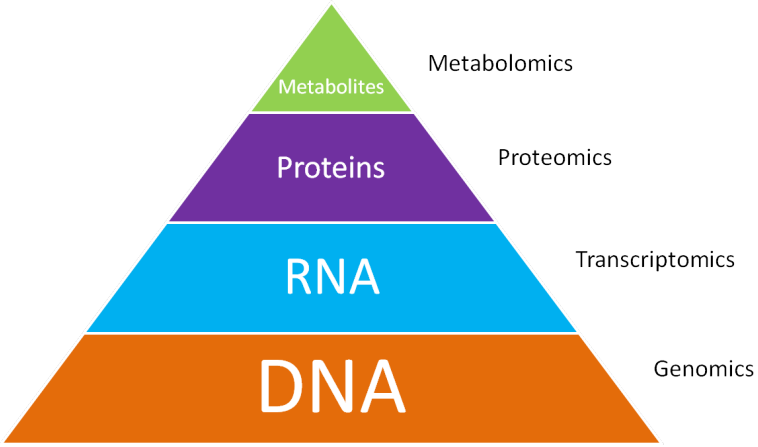


Figure 1: The omics hierarchy.

and that do not take any potentially known variable dependence into account.

A growing number of papers that deal with the analysis of omics data are concerned with bringing prior biological information into the analysis. Large amounts of biological knowledge have been gathered over the years, and analysing the data in light of this information may lead to increased statistical power and improved understanding of the biological processes involved in the condition under study. The focus of this PhD project has been on incorporating biological knowledge into the analysis of transcriptomics data in order to identify genes that are regulated under a certain condition.This has demanded a consideration for complex dependency structures both between variables and samples. Topics that have been treated include testing sets of genes, smoothing data with gene networks and handling small sample sizes and correlated observations. The introduction part of this thesis will give a short preface to the work that is presented in the papers, with a perspective of what have been done before and references to what we have done in our papers. The methods presented in this thesis have been applied to microarray gene expression data, which will be presented in brief in the next section. However, many of the problems faced in microarray data are typical for high-throughput data, and the issues and methods presented here should thus be highly relevant within other omics fields as well. With small adjustments these methods could be applied to different types of omics data.

1.2 Microarray gene expression data

The first paper on microarrays was published by Schena et al. in 1995, and the technology has since become a mainstream tool within the field of molecular biology. Microarrays offer the chance to measure the simultaneous expression of all genes in a cell, quickly and at a relatively low cost per gene. The objective is often to identify genes that are differentially expressed under two or more different conditions or phenotypes.

The number of samples (arrays) in microarray experiments is usually very small compared to the number of variables (genes). There are multiple sources of variation in each step of the experiment which can potentially result in extremely noisy data. Examples of such unwanted effects include differences in dye-intensity, batch effects and array effects. It is important to carefully consider the experimental design in order to

reduce and control the variation so the important effects can be identified. Kerr and Churchill (2001) present some important statistical principles for experimental design of gene expression microarrays.

A microarray experiment can be designed with direct or indirect comparison between the conditions/phenotypes of interest. In direct designs, samples from two phenotypes hybridise to mutual arrays. In indirect designs, samples from two (or more) phenotypes hybridise to different arrays with a common reference. Direct designs give, as the name implies, a direct estimate of the differential expression between two phenotypes. Direct designs require only half as many arrays for the same number of measurements as indirect designs do. The small sample size, which is particularly present in direct designs, is one of the challenges in microarray data that has been treated in this thesis. Indirect design is a natural choice for comparing more than two phenotypes or for comparing results across different experiments. In data from indirect designs, one sample (array) corresponds to one phenotype. Many statistical methods for analysing gene expression data are designed for indirect comparison data, so adapting these methods to direct comparison data is another issue that has been handled in this thesis. Yang and Speed (2002) give a comprehensive review of design considerations for microarray experiments.

Microarray experiments designed as time series have the appealing property of giving a more dynamic picture of gene expression, rather than just doing a ”snapshot” of the genome. Longitudinal studies measure each individual repeatedly over time, while cross-sectional studies measure different individuals at each time point. The advantage of longitudinal studies over cross-sectional studies, is that they can distinguish changes over time within individuals from general differences between individuals. The individuals are usually considered independent, but there will inherently be correlations between measurements from the same individual that must be taken into account. Longitudinal data is a topic in paper III and IV.

The technologies for measuring gene expression is in constant change, and in the future the microarray technology as we know it may be a thing of the past. RNA sequencing (RNA-seq) is gradually taking over in the analysis of transcriptomes (Shendure, 2008, Wang et al., 2009). This technology directly determines the sequence and yields a digital quantification of gene expression, rather than an analog quantification as microarrays yield. The consequence is a dramatic reduction in the level of noise.

Sequencing approaches have traditionally been associated with low throughput and high costs, but with new high-throughput technologies both cost and time are significantly reduced. Although we do not know exactly what is in store for the analysis of transcriptomes, we can be confident that there will always be a need to measure gene expression. The methods presented in this thesis are not exclusive to microarray data, and could easily be adapted to other types of gene expression data. Small sample sizes, one of the issues handled in this thesis, is likely to be a problem also in the future, especially when technologies are new and expensive. As improved technologies produce less technical variation, the biological variation will be even more prominent. Our methods that consider gene dependencies and time dependencies should therefore be highly relevant for gene expression data also from other technologies.

1.3 Including prior knowledge in the analysis

The traditional approach in analysis of gene expression data has been to test each gene for differential expression. Testing thousands of genes simultaneously can potentially lead to a large number of false positives. Correction for multiple hypothesis testing can to a certain degree justify this, but may in many cases give very conservative estimates. In addition, the outcome of the analysis may be a list of significant genes with little or no biological relation, and the results may prove difficult to reproduce in another experiment.

It is known that genes interact on many levels in the cell, and this is likely to be reflected in correlated expression patterns between genes. More and more publications are now focusing on bringing prior biological knowledge into the analysis of expression data. The idea is that genes that are known to be related, should also share a similar expression pattern. The relation may be that the genes take part in the same metabolic pathway, that they have a similar function, or that they are part of the same bacterial operon, just to mention a few. This background information can be brought into the analysis in the form of gene sets or gene networks. What is essential is that the information about gene dependencies is not based on the data at hand, but defined prior to the analysis. Analysing data in light of this prior information can lead to an increased sensitivity by moving the focus from large expression changes in individual genes to more moderate changes in larger groups of related genes. It should also make the

analysis more robust, reduce the number of false positives and give more interpretable results.

1.3.1 Gene set analysis

Gene set analysis methods, or gene set tests as they also will be referred to here, evaluate gene expression data on the basis of a collection of predefined gene sets. Rather than testing individual genes, inferences are made on the gene set level where the goal is to identify differentially expressed sets of genes. Gene set tests may have higher statistical power than individual gene tests because signals from the whole set of genes are accumulated in a gene set score. Significance is assessed for each gene set, usually by computing p-values using a permutation test.

There are numerous ways of defining gene sets. In the papers included in this thesis we have analysed four different categories of gene sets: functional categories, pathways, EC groups and operons. The reasonings for classifying genes into sets may differ, but the fundamental idea is that genes in the same set are expected to have correlated expression patterns. In the functional categories, genes are grouped based on their functional role. Genes in the same pathway take part in successive chemical reactions. EC groups are defined based on the genes' Enzyme Commission number reflecting the biochemical reactions that enzymes catalyse. Operons are clusters of genes in bacterial genomes that are controlled by a common transcription mechanism.

By doing tests on gene sets rather than individual genes, the number of tests, and hence the probability of doing type I errors, is reduced. Gene set tests are also more sensitive for detecting moderate changes in expression that are consistent within the members of a gene set. This can be helpful for seeing gene regulation in a greater context. Larger groups of related genes that show some degree of differential expression, may give more valuable information than a few and possibly unrelated genes with high differential expression. On the other hand, the focus on gene sets may mean that we miss out on some important individual genes. The analysis of gene sets can also be useful for comparing expression patterns in different studies. Subramanian et al. (2005) and Manoli et al. (2006) showed that gene set analysis gave more consistent results than individual gene analysis on different data sets.

A classical method for assessing differential expression in gene sets is the Fisher's exact test. A list of significant genes is compiled, and the density of differentially expressed genes in a given set compared to the remaining genes is tested. The test layout can be presented in a 2×2 table. A number of methods with minor variations of the 2×2 table have been proposed, and an overview is given in Khatri and Draghici (2005). A drawback of these methods is that they still make inferences on the gene level; only the significant genes are included in the computation of a gene set statistic.

Gene set tests that consider the whole set of genes in the computation of a gene set score have become increasingly popular after the introduction of the Gene Set Enrichment Analysis (GSEA) (Mootha et al., 2003, Subramanian et al., 2005). GSEA starts by ranking the genes based on some test statistic, e.g. correlation with a phenotype vector or a t-statistic comparing the differential expression in two groups. The members of the gene sets are located in the ranked list before a Kolmogorov-Smirnov type statistic, a so-called "enrichment score", is calculated for each gene set. Gene sets clustered at the top or bottom of the list tend to get high test scores. The GSEA enrichment score is used as the gene set score in all papers included in this thesis. While GSEA is probably the most popular choice of these types of gene set tests, a string of other tests have been suggested by e.g. Goeman et al. (2004), Tian et al. (2005), Efron and Tibshirani (2007) and Wu et al. (2010). A recent overview can be found in Huang et al. (2009).

The focus of gene set tests so far have been on identifying sets of genes that are differentially expressed at a given time. In longitudinal microarray data, one may also be interested in identifying gene sets that show strong time trends. A group of genes that change expression unanimous over time may give just as important information about gene behaviour as a set of genes that is constantly expressed over time, and is potentially a stronger indicator of correlated genes. In paper III we introduce a gene set test that captures both gene sets with constant differential expression over time and gene sets that show certain trends over time.

1.3.2 Gene networks

Using gene networks is an alternative to gene sets when it comes to including prior information in the analysis. Genes may be arranged into complex networks according

to regulation aspects or location in the DNA. Gene set analysis methods do not take advantage of the explicit structure of gene relationships. With gene networks we can exploit information about distances between genes in a gene set, i.e. how closely related each pair of genes are. The idea is that shorter distances within the network often implies more correlated gene expression patterns. As with gene sets, there are numerous ways of defining gene networks. The most common approach is to use networks based on pathways, but other alternatives include gene regulatory networks or networks derived from bacterial operons. Methods that use gene networks in the analysis of expression data have been presented by i.e. Hanisch et al. (2002), Vert and Kanehisa (2003), Rahnenführer et al. (2004), Rapaport et al. (2007) and Sæbø et al. (2008).

A gene network can be represented as a directed or undirected graph, where each node corresponds to a gene and an edge between two nodes represents some biological relationship. In pathways, an edge between two nodes imply that the genes take part in successive biochemical reactions. In gene regulatory networks, two genes are connected if the transcription of one gene is regulated by the other gene. In operons, two genes are connected because they are located next to each other on the chromosome and are controlled by a common transcription mechanism. Although we have not seen operons being used as networks, they appear to be strong indicators of co-regulation considering that the genes are transcribed simultaneously. Figure 2 shows an example of a gene network derived from pathways in the bacterium *Enterococcus faecalis*. The individual pathways are merged together to one comprehensive network.

The edges in a gene network may also include information about the direction of regulation, that is whether we should expect a positive or negative correlation between each pair of genes. Both in pathways and gene regulatory networks, genes may have either positive or negative effect on each other (positive or negative feedback). Within operons however, we would expect a positive correlation between all genes.

In addition to fully exploiting the structural relationship between genes, gene networks have the appealing feature of not requiring a strict division into gene sets. Since genes may have several functions and take part in several reactions, they do not necessarily fall naturally into one set, and hence there is often considerable overlap between gene sets. In a network however, genes can be connected to several reactions or functions. Groups of genes manifest themselves through so-called community structures in the

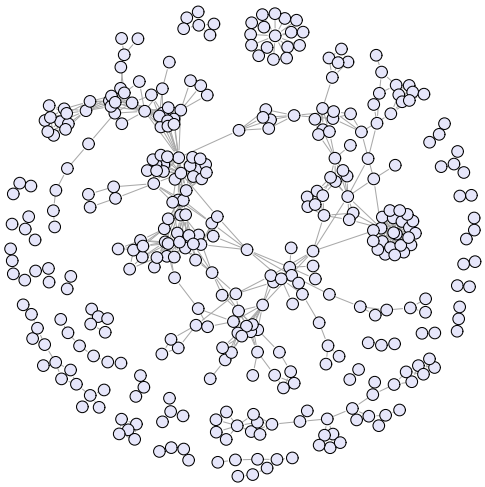


Figure 2: A gene network based on pathways in *Enterococcus faecalis*.

network. Within communities the genes are more tightly connected, and between communities the edges are more sparse. Much effort has been put into finding good ways to divide networks into subnetworks, and examples include using the eigenstructure of the network's Laplacian matrix (Fiedler, 1973, Pothen et al., 1990) or by properties of the network's modularity matrix (Newman and Girvan, 2004). The modularity matrix reflects community structures in the network, and approaches based on this topology will look for natural divisions into communities rather than force a division. The modularity matrix is used in paper IV to divide genes into non-overlapping groups.

In paper II we use a gene network based on pathways to smooth genewise test statistics. The idea is that genes that are closely connected in the network should also have a similar test statistic, and the smoothing should remove false positives and accentuate subnetworks with a high density of important genes. The network is translated into gene dependencies with the use of a graph topology called diffusion (Chung, 1997, Kondor and Lafferty, 2002). Diffusion can be visualised as a liquid travelling through the network, similar to a random walk process. When the diffusion is faster, the liquid will spread faster, and this is equivalent to shortening the distances between genes in the network. Diffusion is also used to model gene dependencies in paper IV, but in this context the gene dependencies are used in the estimation of correlation between

samples. There are a number of other graph topologies that can be used as measures for gene dependencies, for instance the shortest path matrix containing the smallest number of edges separating two nodes, or the already mentioned modularity matrix.

Although gene networks seem to have many advantages over gene sets, a drawback is that there is still limited network information available. Most likely network information for all genes in the data set will not be available. However, in our applications of gene networks it is not required that we have network information for all genes. Our intention is to use whatever information is available to improve the analysis as much as possible.

1.4 Significance testing

The main objective in all papers in this thesis is to identify differentially expressed genes or sets of genes. Whether we are testing individual genes or gene sets, when the distribution of the test statistic is unknown we can use a resampling test to assess significance. Significance testing in gene set analysis is a rather wide-ranging and sometimes confusing subject, so the first section gives an introduction to this topic.

1.4.1 Significance testing in gene set analysis

The general hypothesis tested in all gene set tests is whether a gene set is ”enriched” or not, but the meaning of the word ”enriched” is not always straight forward to interpret. Tian et al. (2005) defined two types of null hypotheses that are used in gene set testing. Hypothesis Q_1 is that the gene set contains no more differentially expressed genes than the remaining gene sets. Hypothesis Q_2 is that the gene set does not contain any differentially expressed genes. In a similar spirit, Goeman and Bühlmann (2007) divided gene set tests into *competitive* and *self-contained* tests depending on how the null hypothesis is defined. Competitive gene set tests aim at identifying gene sets that stand out from a collection of gene sets and are testing the Q_1 hypothesis. Self-contained tests assess each gene set individually and is not affected by the enrichment of other sets, hence they are testing the Q_2 hypothesis. Q_1 and competitive tests are often associated with permutation of genes in the computation of p-values (testing the null hypothesis of the gene set being a random sample of genes from the whole

collection of genes), while Q_2 and self-contained tests are associated with permutation of samples. GSEA however, is rather special. The test statistic is competitive, but sample permutation is used to calculate p-values. The difference in the null hypothesis for the calculation of the test statistics and the p-values may decrease GSEA's power, according to Tian et al. and Goeman and Bühlmann. Permutation of samples compared to permutation of genes is treated in more depth in the next section.

The self-contained null hypothesis is more restrictive than the competitive null hypothesis, and in general self-contained tests will reject more null hypotheses than competitive tests. This will in particular emerge in data sets with many differentially expressed genes. In a competitive test it is harder for a gene set to stand out from the rest when there are many important gene sets. This was shown in a simulation study by Efron and Tibshirani (2007), and is probably also what we observe when we use GSEA in paper II; data sets with many differentially expressed genes get fewer significant gene sets than those with less differentially expressed genes. On the other hand, in such data sets a self-contained test would call almost all gene sets significant, and this may not always be biologically interesting. The choice of null hypothesis and gene set test thus depends on the data at hand and the purpose of the analysis; whether it is to identify all gene sets that are associated with a phenotype or *the most* important gene sets. Alternatively one can test both Q_1 and Q_2, as suggested by Tian et al. and Efron and Tibshirani.

1.4.2 Permutation test

The most common approach in gene set tests is to permute samples to obtain a distribution of the gene set scores under the null hypothesis that none of the genes are differentially expressed (the Q_2 hypothesis and GSEA). In indirect comparison data, each sample (array) represents one phenotype. The permutations are performed by shuffling phenotype labels on the samples. In direct comparison data, a sample represents the ratio of expression for two phenotypes. This is a case of paired data, and permutations can be performed by randomly changing the signs of the samples. Permutation tests require a certain number of samples in order to estimate accurate p-values. Microarray data often have small sample sizes, which restricts the maximum number of permutations. This is especially the case in direct design experiments, which require only half as many arrays as indirect designs to achieve the same number of measurements.

In an experiment with four arrays comparing two phenotypes there are only $\binom{4}{2} = 6$ possible permutations for an indirect design, and $2^4 = 16$ possible permutations for a direct design. The result would be granular null distributions and inaccurate p-value estimates.

Since the number of genes is usually very large, a straightforward solution to the small sample size problem seems to be permutation of genes rather than samples. However, this changes the implicit null hypothesis to being that the genes in the set are drawn at random from the full collection of genes (the Q_1 hypothesis). Fisher's exact test and other 2×2 table methods are equivalent to methods that permute genes. The observed 2×2 table and distribution of significant genes in the gene set is compared to cases where the significant genes are randomly distributed in the 2×2 table. Gene permutation is problematic because it implicitly assumes independent genes, completely contradictory to the whole idea of bringing gene set/network information into the analysis. The genes are grouped or connected because they are believed to have correlated gene expression patterns. The result of resampling correlated genes may be a serious underestimation of the p-values, as shown by Efron and Tibshirani (2007) and as we show in paper I. For this reason, we want to avoid permutation of genes.

A problem with the permutation of samples is the necessary assumption about independent and identically distributed samples, which is often not satisfied in data from complex experiments. Because of effects of fixed design factors the samples are not on the same level, while random design factors introduce correlations between samples. An option could be to permute samples within each level of a factor, but this requires a large number of samples. In paper I and II we approach this problem by fitting an ANOVA model with the uninteresting design variables as fixed factors. The estimated residuals from this model, only containing the interesting effects (e.g. differential expression), are regarded as independent, normalised samples. Similar normalisation techniques have been used by other authors, e.g. Wolfinger et al. (2001). The residuals will never be completely independent, but we argue that they are sufficiently independent for the purposes in these papers. However, in cases where the number of replicates is limited, and the effects of the nuisance factors are small, this type of normalisation may actually increase correlations between samples! In paper I and II we analyse data sets with only two samples from each level of a factor, but by assuming that the effects are independent of gene we can use all genes in the estimation.

Although the problems with the permutation test for small sample sizes and non-exchangeable samples have been discussed in the context of gene set testing, it also applies to single gene testing, which we use in paper II.

1.4.3 Rotation test

As a solution to the problems regarding permutation tests, we have exchanged the permutation test with a rotation test whenever required in all papers in this thesis, both in the context of gene set testing and individual gene testing. The theory behind rotation tests was first published by Langsrud (2005). We were the first to introduce rotation testing for gene set tests in paper I. While a permutation test is restricted to exchange measurement axes, a rotation test can rotate the data in all directions and still preserve covariances between genes. Since there is no limitation to the number of rotations, accurate p-values can be estimated also for small sample sizes. In paper I we compared the power of the rotation test and the permutation test, and found that the rotation test clearly had higher power for very small sample sizes. One advantage the permutation test has over the rotation test, is that it makes no assumptions about distribution except that the samples are identically distributed. The rotation test assumes that the samples come from a multivariate normal distribution, but this appears not to be a critical assumption, as shown by us in paper I and by Wu et al. (2010).

The rotation test can handle data from complex experimental designs by doing rotations in the residual space of a linear model including all factors in the experiment. The data are projected onto a subspace orthogonal to the nuisance factors, a procedure that removes the nuisance effects and obtains independent residuals. This part of the rotation test was first employed in a gene set test context by Wu et al. (2010), and we apply it in paper III and IV where we adapt it to longitudinal microarray data.

1.4.4 Correlated samples

The rotation test can, as mentioned, handle effects of fixed design factors by modelling the data in a linear model and perform rotations in the model's residual space. The rotation test does however assume that the samples are uncorrelated. Wu et al. (2010) dealt with correlation between samples due to random design factors by estimating

the empirical covariance matrix. The estimated covariance can then be included in the model to reduce correlations between samples. The covariance matrix was assumed to be identical for all genes, and genes were assumed to be independent, leaving a large number of samples to base the estimate on. In paper III we attempt to improve this estimate by assuming a structure for the covariance matrix and estimate the various components, rather than estimating each element of the covariance matrix. We use a structure presented in Diggle et al. (1994) for longitudinal data, that assumes that variation between samples is due to random design factors, time and random error. We further assume a common covariance structure for all genes and independence between genes, and estimate the variance components with restricted maximum likelihood. The assumption of independence between genes made by both Wu et al. and us, is however in strong contrast to the assumption about correlation between genes made in the later gene set test. In paper IV we therefore take the estimation of covariance between samples a step further by also including gene dependencies. The genes are divided into non-overlapping groups, and dependencies are assumed only within groups.

1.4.5 False discovery rate

Although testing gene sets rather than individual genes significantly reduces the number of tests performed, some correction for multiple hypothesis testing should be applied. In all papers in this thesis we have used the false discovery rate (FDR) (Benjamini and Hochberg, 1995) as an error rate for controlling the type I error. FDR is defined as the expected proportion of falsely rejected hypotheses among all rejections. An FDR of 5% means that among all rejected hypotheses, on average 5% of these will be false rejections. Storey (2002) introduced the positive false discovery rate (pFDR) which is conditioned on at least one hypothesis being rejected, and the term q-value, the pFDR equivalent of the p-value. The q-values give measures of significance for each hypothesis. It is the lowest significance level at which the hypothesis can be rejected, or the lowest level of pFDR that can be achieved when using the test statistic for the given test as the cut-off.

2 Paper summaries

Paper I – Rotation testing in Gene Set Enrichment Analysis for small direct comparison experiments

The popular Gene Set Enrichment Analysis (GSEA) uses a permutation test to assess significance for gene sets. The permutation test in GSEA is designed for indirect comparison data. To make GSEA applicable also to direct comparison data with few samples, we replace the permutation test in GSEA with a rotation test. The rotation test can, in contrast to the permutation test, calculate accurate p-values also for small sample sizes. We demonstrate in a simulation study how problematic permutation of genes can be when genes are correlated within gene sets. We compare the rotation test with the permutation test on simulated normal and non-normal data, and show that the rotation test outperforms the permutation test on very small sample sizes, and that the rotation test seems to be robust to deviations from the assumption about multinormality. Finally, GSEA with rotation test is applied to a real gene expression data set where the stress responses in the bacterium *E. faecalis* have been investigated.

Paper II – Smoothing gene expression data with network information improves consistency of regulated genes

In this paper we move the focus from gene sets to gene networks. In gene networks we do not only have information about which genes are related, but also how closely related each pair of genes are. As network information we use pathways that are merged into one large gene network. The gene network is used to ”smooth” genewise test statistics in order to reduce the number of false positives, and accentuate parts of the network with high concentrations of important genes. We simulate gene expression data with correlation structures borrowed from both fictional and real networks, and show that the network smoothing improves the power in identifying important genes, but that it also imposes the risk of losing individual genes. We discuss the effect of smoothing on different network structures and the degree of smoothing that should be performed, and propose a criterion for choosing the optimal level of smoothing based on the correlation between the network and the data. The network smoothing is also applied to the *E. faecalis* data set from paper I, and a rotation test is used to calculate a p-value for each gene. The smoothed data are finally analysed with GSEA with

rotation test from paper I to help interpreting the results on a pathway level.

Paper III – Rotation gene set testing for longitudinal expression data

We pick up the thread from paper I and move back to gene sets. In this paper we attempt to improve the method presented in paper I by adapting it to longitudinal data and other complex experimental designs. Longitudinal data introduce intricate correlation structures between samples. In order to reduce these correlations, we assume a structure for the covariance matrix and estimate its components with restricted maximum likelihood. The estimated covariances are included in a preprocessing step. The preprocessed data are represented by a linear model and the rotation test is performed in the residual space of this model, thereby dodging the effects of nuisance factors. This procedure also gives independent residuals from the linear model. The gene set analysis is further improved by allowing testing of several properties simultaneously, so both gene sets that are differentially expressed and gene sets that have interesting time trends can be identified. We show in a simulation study that by taking into account the correlation structure of the samples, we improve the power in identifying important gene sets. Applied to the *E. faecalis* data set from the previous papers, the method is able to identify both gene sets that are constantly differentially expressed over time and gene sets that show strong time trends.

Paper IV – Improved preprocessing for rotation gene set testing for longitudinal expression data

The final paper in this thesis is an extension of paper III. The aim is to further improve the gene set rotation test for longitudinal data, and the focus is now on the estimation of covariances between samples in the preprocessing step. In paper III we made a doubtful assumption about independent genes for the purpose of estimating the covariance matrix. This assumption is in strong contrast to the gene set test's idea about correlation between genes in the same set. We therefore decided to include gene dependencies in the estimation. This paper brings back some topics from paper II concerning gene networks. The genes are divided into non-overlapping groups with the use of a predefined gene network, and dependencies are assumed only between genes in the same group. Gene dependencies are further modeled with distances extracted

from the gene network, and the dependencies are included in the estimation of covariances between samples. In a simulation study we compare the power for the gene set rotation test when using the old and the new preprocessing step. The new preprocessing method seems to emphasize gene sets with strong time trends more than the old method, and attaches less importance to gene sets with constant expression over time. This is further validated by application to the *E. faecalis* data set.

3 Discussion

The focus in this thesis has been on analysing gene expression data with the contribution of prior biological information. The aim has been to extract more of the significant information in the data and filter out the noise, with resulting increased statistical power and improved biological understanding. Incorporation of prior biological information in the analysis is of course dependent on the extent of information available. For some organisms the gene set and gene network information may be limited. In general this seems to be a larger problem when working with networks, since the creation of networks requires more detailed knowledge about gene dependencies. Network information may be available only for sections of the genes to be analysed, and details about the direction of regulation may be unattainable. The applications of networks in this thesis do not require information about all genes, as genes without network information can still be included in the analysis. In gene set tests however, only genes that are members of gene sets are analysed. It can be argued that the exclusion of a small number of genes should not be crucial for the outcome of the analysis when the goal is to identify larger sets of important genes. In addition to the extent of information available, these methods depend upon the accuracy of the biological information. For example are the operons we analysed in paper I, III and IV just predictions. The consequences of incorrect information should be more moderate when working on a gene set level rather than on a gene level. In paper II we used networks to smooth genewise test statistics, where we assumed positive correlations between all genes. Some genes will undoubtedly be negatively correlated, so as a future perspective it would be interesting to explore the robustness towards incorrect network information.

The GSEA enrichment score is used as the gene set score in all papers included in this thesis, but could easily have been replaced with other types of statistics. The choice of gene set score is heavily debated in the literature, and new gene set scores that appear

to outperform existing scores in given scenarios, are constantly introduced (Kim and Volsky, 2005, Efron and Tibshirani, 2007, Dinu et al., 2007). Different scores capture different sorts of correlation structures between genes, and the optimal choice of score depends on the questions being asked, the gene sets being tested and the researcher's biological knowledge. Huang et al. (2009) give some guidelines for choosing the most appropriate gene set test. The gene set score has however not been a major topic in this thesis, so we chose to retain the GSEA approach for measuring enrichment in gene sets.

We first introduced rotation testing as an alternative to permutation testing to assess significance for gene sets in paper I. Rotation testing was used also in papers II-IV, both for testing gene sets and individual genes. In paper I we investigated properties of the rotation test, and showed that it had superior power over the permutation test when the sample size was very small, and that it controlled the type I error satisfactorily. The rotation test does make an assumption about the samples being multinormally distributed, but in paper I we showed that the rotation test appears to be robust to deviations from normality. This has also later been confirmed by Wu et al. (2010). We did however note that the rotation test had slightly lower power than the permutation test in a data set with even stronger deviations from normality, so it would be interesting to carry out a more thorough survey on the rotation test's properties. In paper III and IV we adapted gene set analysis with rotation test to data with complex correlation structures between samples, in particular longitudinal data. By estimating covariances between samples and taking these into account, we showed on simulated data that we were able to increase the power in identifying important gene sets. We also noted that the estimated type I error was not controlled at a proper level when we ignored these covariances. By assuming independence between samples when they are in fact correlated, we believe that we have more observations than we actually have, and as a result we underestimate the variance. In paper III we assumed independence between genes during the estimation, an assumption that is very doubtful. In paper IV we therefore tried to improve the estimation of the covariance matrix by also taking gene dependencies into account. By dividing genes into non-overlapping gene groups based on network topology, most of the gene-gene correlations within these groups are hopefully taken into account. Including gene dependencies in the covariance matrix helped in the identification of gene sets with strong time trends, but made it more difficult to identify gene sets with differential expression and no time trends. The question

of whether to include gene dependencies in the covariance matrix or not depends on the magnitude of the correlations. If the correlations are large, then the type I error will be more efficiently controlled when including gene dependencies. If however the correlations are small, we risk overfitting. The division into meaningful groups is therefore important to accommodate the assumption of correlations within groups and independence between groups as much as possible.

In all papers included in this thesis we have used simulated data to demonstrate our methods. Microarray data contain large amounts of noise from various sources, so a severe simplification of the real world is necessary when simulating data. Correlations both between genes and samples will undoubtedly be much more complex than what is assumed here, and one may experience that the method is behaving completely different when applied to real data. However, in order to understand the new methods' effect on certain factors, it may be necessary to neglect other less important factors. By applying the method to simulated data in a controlled environment, we hope to get an increased understanding of why the method acts like it does on real data.

The increasing amounts of biological knowledge being accumulated through different studies offer good prospects for including prior knowledge in the analysis. It seems only natural to take advantage of this additional information when analysing new data. New biological associations have been identified with methods that include prior information, and all in all the result may be an increased insight into the behaviour of genes. The information will also become more reliable and more detailed as it is verified in more studies. There are certainly numerous ways to incorporate this information that have not been thought of yet, and future challenges may include combining different sources of information to obtain even better estimates.

4 References

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

F. Chung. Spectral graph theory. *No. 92 in Regional Conference Series in Mathematics. American Mathematical Society*, 1997.

P. Diggle, K.-Y. Liang, and S. Zeger. *Analysis of Longitudinal Data*. Oxford University Press, 1994.

I. Dinu, J. Potter, T. Mueller, Q. Liu, A. Adewale, G. Jhangri, G. Einecke, K. Famulski, P. Halloran, and Y. Yasui. Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics*, 8, 2007. ISSN 1471-2105.

B. Efron and R. Tibshirani. On testing the significance of sets of genes. *Annals of Applied Statistics*, 1(1):107–129, 2007. ISSN 1932-6157.

M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(2):298–305, 1973. ISSN 0011-4642.

J. Goeman and P. Bühlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, 2007. ISSN 1367-4803.

J. Goeman, S. van de Geer, F. de Kort, and H. van Houwelingen. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1): 93–99, 2004. ISSN 1367-4803.

D. Hanisch, A. Zien, R. Zimmer, and T. Lengauer. Co-clustering of biological networks and gene expression data. *Bioinformatics*, 18:145–154, 2002.

D. Huang, B. Sherman, and R. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13, 2009. ISSN 0305-1048.

M. Kerr and G. Churchill. Experimental design for gene expression microarrays. *Biostatistics*, 2(2):183–201, 2001.

P. Khatri and S. Draghici. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21(18):3587–3595, 2005. ISSN 1367-4803.

S. Kim and D. Volsky. PAGE: Parametric analysis of gene set enrichment. *BMC Bioinformatics*, 6, 2005. ISSN 1471-2105.

R. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete input spaces. In *Proceedings of the Nineteenth International Conference on Machine Learning*, 2002. ISBN 1-55860-873-7.

O. Langsrud. Rotation tests. *Statistics and Computing*, 15(1):53–60, 2005. ISSN 0960-3174.

T. Manoli, N. Gretz, H.-J. Grone, M. Kenzelmann, R. Eils, and B. Brors. Group testing for pathway analysis improves comparability of different microarray datasets. *Bioinformatics*, 22(20):2500–2506, 2006. ISSN 1367-4803.

V. Mootha, C. Lindgren, K. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstrale, E. Laurila, N. Houstis, M. Daly, N. Patterson, J. Mesirov, T. Golub, P. Tamayo, B. Spiegelman, E. Lander, J. Hirschhorn, D. Altshuler, and L. Groop. PGC-1 alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34(3):267–273, 2003. ISSN 1061-4036.

M. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2, Part 2), 2004. ISSN 1063-651X.

A. Pothen, H. Simon, and K. Liou. Partitioning sparse matrices with eigenvectors of graphs. *Siam Journal on Matrix Analysis and Applications*, 11(3):430–452, 1990. ISSN 0895-4798.

J. Rahnenführer, F. Domingues, J. Maydt, and T. Lengauer. Calculating the statistical significance of changes in pathway activity from gene expression data. *Statistical Applications in Genetics and Molecular Biology*, 3, 2004.

F. Rapaport, A. Zinovyev, M. Dutreix, E. Barillot, and J. Vert. Classification of microarray data using gene networks. *BMC Bioinformatics*, 8, 2007. ISSN 1471-2105.

S. Sæbø, T. Almøy, A. Flatberg, A. Aastveit, and H. Martens. LPLS-regression: a method for prediction and classification under the influence of background information on predictor variables. *Chemometrics and Intelligent Laboratory Systems*, 91 (2):121–132, 2008. ISSN 0169-7439.

M. Schena, D. Shalon, R. Davis, and P. Brown. Quantitative monitoring of gene-expression patterns with a complementary-DNA microarray. *Science*, 270(5235): 467–470, 1995. ISSN 0036-8075.

J. Shendure. The beginning of the end for microarrays? *Nature Methods*, 5:585–587, 2008.

J. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B - Statistical Methodology*, 64(Part 3):479–498, 2002. ISSN 1369-7412.

A. Subramanian, P. Tamayo, V. Mootha, S. Mukherjee, B. Ebert, M. Gillette, A. Paulovich, S. Pomeroy, T. Golub, E. Lander, and J. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 102(43):15545–15550, 2005. ISSN 0027-8424.

L. Tian, S. Greenberg, S. Kong, J. Altschuler, I. Kohane, and P. Park. Discovering statistically significant pathways in expression profiling studies. *PNAS*, 102(38): 13544–13549, 2005. ISSN 0027-8424.

J. Vert and M. Kanehisa. Extracting active pathways from gene expression data. *Bioinformatics*, 19(Suppl. 2):II238–II244, 2003. ISSN 1367-4803.

Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009. ISSN 1471-0056.

R. Wolfinger, G. Gibson, E. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari, and R. Paules. Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology*, 8(6):625–637, 2001. ISSN 1066-5277.

D. Wu, E. Lim, F. Vaillant, M.-L. Asselin-Labat, J. Visvader, and G. Smyth. ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics*, 26(17): 2176–2182, 2010. ISSN 1367-4803.

Y. Yang and T. Speed. Design issues for cDNA microarray experiments. *Nature Reviews Genetics*, 3(8):579–588, 2002. ISSN 1471-0056.

Paper I

Statistical Applications in Genetics and Molecular Biology

Volume 8, *Issue* 1 2009 *Article* 34

Rotation Testing in Gene Set Enrichment Analysis for Small Direct Comparison Experiments

Guro Dørum* Lars Snipen†
Margrete Solheim‡ Solve Sæbø**

*Norwegian University of Life Sciences, guro.dorum@umb.no
†Norwegian University of Life Sciences, lars.snipen@umb.no
‡Norwegian University of Life Sciences, margrete.solheim@umb.no
**Norwegian University of Life Sciences, solve.sabo@umb.no

Rotation Testing in Gene Set Enrichment Analysis for Small Direct Comparison Experiments*

Guro Dørum, Lars Snipen, Margrete Solheim, Solve Sæbø

Abstract

Gene Set Enrichment Analysis (GSEA) is a method for analysing gene expression data with a focus on *a priori* defined gene sets. The permutation test generally used in GSEA for testing the significance of gene set enrichment involves permutation of a phenotype vector and is developed for data from an indirect comparison design, i.e. unpaired data. In some studies the samples representing two phenotypes are paired, e.g. samples taken from a patient before and after treatment, or if samples representing two phenotypes are hybridised to the same two-channel array (direct comparison design). In this paper we will focus on data from direct comparison experiments, but the methods can be applied to paired data in general. For these types of data, a standard permutation test for paired data that randomly re-signs samples can be used. However, if the sample size is very small, which is often the case for a direct comparison design, a permutation test will give very imprecise estimates of the *p*-values. Here we propose using a rotation test rather than a permutation test for estimation of significance in GSEA of direct comparison data with a limited number of samples. Our proposed rotation test makes GSEA applicable to direct comparison data with few samples, by depending on rotations of the data instead of permutations. The rotation test is a generalisation of the permutation test, and can in addition be used on indirect comparison data and for testing significance of other types of test statistics outside the GSEA framework.

KEYWORDS: gene set analysis, gene expression, microarray data analysis

*We would like to thank Ågot Aakra for her advice in the early stages of the preparation of this manuscript, and Øyvind Langsrud for his valuable comments on rotation testing in general. We would also like to thank two anonymous referees for their help in improving this manuscript.

1 Introduction

The most widespread use of the microarray technology is for identification of differential expression of genes between samples from two or more conditions/ treatments/populations. The traditional approach for analysing expression data involves testing single genes for differential expression and assembling lists of post hoc interesting genes, but the results can often be difficult to interpret into a biological context. In many cases we are more interested in identifying groups of genes rather than single genes. A growing number of methods are focusing on the identification of *a priori* defined sets of genes that are in some way affected by the experimental conditions. A set of genes that is known to be functionally related is likely to share a similar expression pattern, so the use of prior biological knowledge can make the analysis more robust and give more meaningful results. Examples of such *a priori* defined gene sets are pathways, functional categories and Gene Ontology categories (Ashburner *et al.*, 2000).

Most approaches for identifying affected gene sets start by ranking each gene by its differential expression, and some cut-off is set to compile a list of differentially expressed genes. Groups of genes that are overrepresented in this list are classified as potentially interesting. Khatri and Drăghici (2005) give an overview of many of these approaches.

In the following we adopt the Gene Set Enrichment Analysis (GSEA) approach, first introduced by Mootha *et al.* (2003) and later modified by Subramanian *et al.* (2005), for identifying differentially expressed gene sets. One interesting aspect of GSEA is that no inferences are made on the level of single genes, but instead all members of a gene set are included in the calculation of a set score. The idea is that GSEA is more sensitive for detecting sets of genes with a moderate, but consistent effect, compared to methods that use a strict cut-off value. GSEA starts by ranking all genes based on their association with a phenotype vector. The rank positions of all members of a set are identified, and an enrichment score, which is essentially a weighted Kolmogorov-Smirnov statistic, is calculated for each gene set. The enrichment score is normalised to account for gene set size. The significance of a gene set is in Mootha *et al.* and Subramanian *et al.* estimated by permuting the class labels of the phenotype vector and recalculating the normalised enrichment score for each permutation. The p-value is calculated as the proportion of this distribution at least as extreme as the observed normalised enrichment score. See Subramanian *et al.* for further details about the GSEA procedure.

A microarray experiment comparing two different classes (conditions/ treatments/populations) can be performed by either direct or indirect com-

parison of data. Indirect comparison for two-colour arrays, which has been the common data type in GSEA applications insofar, involves hybridising samples from different classes to different arrays with a common reference. Alternatively, gene expression analysis may be conducted by letting samples from two classes hybridise to the same array, and by that obtain a direct estimate of the genes' differential expression between the classes. Direct comparison requires only half as many arrays as indirect comparison, and the inevitable variation between two arrays is avoided. On the other hand, indirect comparison is a natural choice for experiments where you want to compare more than two classes, or if you want to compare the results with other experiments in which the same reference has been used.

In direct comparisons, data from a single array does not represent only one phenotype or class, and hence it is impossible to use a permutation test that shuffles class labels. A number of modifications of GSEA has been suggested, i.a. by Kim and Volsky (2005), Jiang and Gentleman (2007) and Efron and Tibshirani (2007). None of these directly addresses the case of direct comparison data, but e.g. Kim and Volsky avoid permutation tests by assuming normal distribution for the data. A Z-score is calculated for each group and the p-value is estimated by comparing the score to the normal distribution.

Direct comparison data is a case of paired data, for which the standard permutation test involves randomly changing the signs of samples, yielding zero expectation under the null hypothesis (the procedure is outlined in i.a. Box *et al.*, 1978). By re-signing whole arrays, the correlation structure between genes is preserved. A problem with permutation tests arises when the number of replicate arrays is small, which is often the case for direct comparison data (there can be as little as two arrays, see e.g. Kerr *et al.*, 2000). In this paper we analyse a data set with four arrays, meaning there are only 2^4=16 possible re-signings, and the smallest p-value that can be obtained is 1/17 (when the observed value is included).

A simple solution for situations with small sample sizes would be to permute genes rather than samples, which here will be referred to as randomisations (following the nomenclature of Efron and Tibshirani, 2007). As discussed briefly by Efron and Tibshirani and also demonstrated here, a problem with the randomisation approach is that it assumes independence between genes. However, the genes in a set are grouped because they are believed to be functionally related, so presumably there is a considerable correlation between the members' expression values. By shuffling genes, new artificial sets are created and the original correlation structure between the genes is lost. Since the null distribution is based on gene sets with no correlation, the p-values are likely to be severely underestimated. See also Nam and Kim (2008)

DOI: 10.2202/1544-6115.1418

and references therein for discussions concerning randomisation of genes.

As a solution to the problem raised by the often small number of samples in the direct comparison situation, we propose a non-parametric approach for estimating p-values in GSEA based on rotations instead of permutations. The rotation test handles correlation within sets in a similar way to permutations by conditioning rotations on the covariance matrix (Langsrud, 2005).

In the following we will use the notation GSEAperm, GSEArand and GSEArot for GSEA with permutation test, GSEA with randomisation test and GSEA with rotation test, respectively.

In an initial simulation study we illustrate how the randomisation test tends to have increased type I error levels when genes are correlated within gene sets, whereas the rotation test does not show this deficiency. We further use simulated data to compare the power and type I error of the rotation test and the permutation test, and to check the robustness of the rotation test to deviations from its assumptions. We apply GSEArot to data where the genome-wide effect of bile stress on the bacteria *Enterococcus faecalis* V583 has been studied in a direct comparison experiment using two-colour DNA microarrays. We also apply GSEArot to the p53 data set (Olivier *et al.*, 2002) used by Subramanian *et al.* to illustrate that a rotation test can be seen as a generalisation of the permutation test, and is applicable also to indirect data.

2 Materials and Methods

2.1 Rotation test

Due to the fact that a permutation test gives very imprecise estimates of the p-values when the number of samples is small, we here introduce rotation testing as an alternative to permutation testing in GSEA. The observed values for each gene may be considered as a vector in the n-dimensional sample space, $\mathcal{R}^n$. Random rotations of these p gene vectors are used to simulate new data matrices $\mathbf{X}^\star$. The rotations are conditioned on the covariance matrix, i.e. the correlation between genes are maintained also after the rotation.

Consider a $n \times p$ data matrix $\mathbf{X}$ of log-ratios, where n is the number of samples and p is the number of genes. The rotation test assumes that the rows of $\mathbf{X}$ are multinormal and independent, i.e. that each array $\mathbf{x_i} \sim N_p(\mu, \mathbf{\Sigma_x})$ and that the arrays are independent. By a random rotation of $\mathbf{x_i}$ we get $\mathbf{x_i^\star} \sim N_p(\mathbf{0}, \mathbf{\Sigma_x})$. The rotated genes have expectation 0, but the covariance matrix is maintained. The test statistic for GSEA, the enrichment score (ES), is computed for each of a number of rotations in order to construct a null

distribution for ES under the complete null hypothesis that all gene sets consist of only non-differentially expressed genes. This is the same null hypothesis as is tested with the permutation test, though not clearly stated by Subramanian *et al.* (2005).

Note that since our data are log-ratios, assuming each gene to have expectation zero is equivalent to assuming that the expected gene expression for the two phenotypes are identical. This is also the assumption made when permuting phenotypes for indirect comparison data.

In order to describe the procedure for performing random rotations, we adopt a similar notation as Langsrud (2005). The data matrix $\mathbf{X}$ can be decomposed into a random configuration matrix $\mathbf{X_R}$ and a random orientation matrix $\mathbf{X_Q}$ by QR decomposition:

$$\mathbf{X} = \mathbf{X_Q}\mathbf{X_R} \tag{1}$$

Here $\mathbf{X_Q}$ is an orthonormal matrix of size $n \times r$, where r is the rank of $\mathbf{X}$, and $\mathbf{X_R}$ is an upper triangular $r \times p$ matrix with positive diagonal elements. The configuration matrix $\mathbf{X_R}$ is a sufficient statistic for the covariance matrix $\mathbf{\Sigma}$. In order to rotate $\mathbf{X}$, we want a new rotation matrix $\mathbf{X^\star_Q}$ while keeping the structure $\mathbf{X_R}$:

$$\mathbf{X^\star} = \mathbf{X^\star_Q}\mathbf{X_R} \tag{2}$$

A random rotation matrix multiplied by another rotation matrix is still a random rotation matrix. A random rotation matrix $\mathbf{X^\star_Q}$ can therefore be simulated by

$$\mathbf{X^\star_Q} = \mathbf{Q}\mathbf{X_Q} \tag{3}$$

where $\mathbf{Q}$ is a simulated random rotation matrix. This means that $\mathbf{X^\star}$ can be simulated as

$$\mathbf{X^\star} = \mathbf{Q}\mathbf{X_Q}\mathbf{X_R} = \mathbf{Q}\mathbf{X} \tag{4}$$

It can be shown that a random rotation matrix can be generated as follows. A $n \times n$ matrix $\mathbf{W}$ is comprised by elements drawn at random from a standard normal distribution. A QR decomposition of $\mathbf{W}$ then gives $\mathbf{W} = \mathbf{W_Q}\mathbf{W_R}$, where $\mathbf{W_Q}$ is a random $n \times n$ rotation matrix. The rotated data set $\mathbf{X^\star}$ was generated as

$$\mathbf{X^\star} = \mathbf{W_Q}\mathbf{X} \tag{5}$$

Note that some implementations of the QR decomposition algorithm gives a matrix $\mathbf{W_Q}$ where the column-vectors are reverted to give, as far as possible, positive diagonal elements of the corresponding matrix $\mathbf{W_R}$. This will not give full rotational freedom of $\mathbf{X}$. We implemented our own QR-function in R based on Householder reflections to avoid this.

DOI: 10.2202/1544-6115.1418

Whereas the rotations in the rotation test are allowed to vary freely in $\mathcal{R}^n$, permutations can be seen as rotations with restrictions. Permutations of the rows of a matrix $\mathbf{X}$ can be achieved by pre-multiplying by a permutation matrix $\mathbf{P}$, which can be constructed by permuting the columns of an identity matrix. The permutation matrix can be considered a restricted rotation matrix for which rotation is equivalent to exchanging measurement axes. On the other hand, the rotation test assumes multinormality for the rows in the data matrix, which is not a necessary assumption for the permutation test.

Although the motivation for rotation testing in this paper is for performing tests on direct comparison data with small sample sizes, the method may also be used for indirect comparison data as an alternative to the commonly used permutation test. This is illustrated below by reanalysing the p53 data set used by Subramanian *et al.* with GSEArot. Since for indirect comparison data the assumption of null-expectation under the null hypothesis is no longer reasonable, a slight modification of the rotation test must be done to give rotated data where both mean and covariance structure are preserved (as is the case for the permutation method). This means that the data rotation must be done in the (n-1) dimensional space orthogonal to the constant vector $\mathbf{1}$. A procedure for achieving such subspace rotations is described by Wedderburn (1975).

2.2 Simulations

2.2.1 Comparison of rotation test and randomisation test

By permuting genes rather than samples, the randomisation test breaks down the correlation structure within gene sets, resulting in an underestimation of the p-values. To illustrate the effect correlation has on the randomisation test versus the rotation test, we simulated data with increasing correlation within gene sets.

We assume that each gene is member of one and only one gene set, and that all sets have equal internal correlation. The genes were simulated under the complete null hypothesis that all gene sets contain only non-differentially expressed genes. Let x_{kj} be the expression value of the jth gene in the kth set, where $k = 1, ..., K$ and $j = 1, ..., J$. We generated values imitating log-ratios with the model

$$x_{kj} = a_k + \epsilon_{kj} \tag{6}$$

where $a_k \sim N(0, \sigma_a^2)$ is a random gene set effect and $\epsilon_{kj} \sim N(0, \sigma_\epsilon^2)$ is a random gene effect. We further assumed that $x_{kj} \sim N(0, 1)$, such that a gene's total variance is $\sigma_a^2 + \sigma_\epsilon^2 = 1$. For a given correlation ρ, we generated each x_{kj}

with gene set variance $\sigma_a^2 = \rho$ and gene variance $\sigma_\epsilon^2 = 1 - \rho$. Data were generated for $K = 50$ gene sets of size $J = 20$, a total of 1000 genes. The number of samples (arrays) was set to 8. The correlation levels chosen were $\rho = \{0, 0.1, 0.2, ..., 0.9\}$. The simulation was repeated 100 times for each ρ.

2.2.2 Rotation test and permutation test on normal data

Control of the type I error and the sensitivity of the rotation test versus the permutation test were tested in a simulation study. Standard normally distributed data for 1000 genes distributed over 50 gene sets were generated as in section 2.2.1, and the within gene set correlation was set to $\rho = 0.4$. A gene effect γ was added to all genes in the first gene set, where $\gamma = \{0, 0.4, 0.55, 0.7\}$. The number of samples chosen were $n = \{4, 8\}$. The simulation was repeated 100 times for each combination of γ and n.

2.2.3 Rotation and permutation test on non-normal data

The rotation test's assumption about multinormal distribution for each array is what separates it from the permutation test. To test the robustness of the rotation test to deviations from normality, we simulated random data from a log-normal distribution with mean 0 and variance 1. First a matrix $\mathbf{X}$ of normally distributed data for 1000 genes in 50 gene sets were generated with the model in (6), with a gene's total variance corresponding to variance 1 on log-normal scale, and an internal gene set correlation of $\rho = 0.4$. The number of samples chosen were $n = \{4, 8\}$. Log-normal data were then generated as $y_{ij} = e^{x_{ij}}$, and the expected mean was subtracted to obtain expression values with mean 0. A gene effect γ was then added to all genes in the first gene set, with $\gamma = \{0, 0.1, 0.15, 0.2\}$. The γ values for these log-normal data correspond to the percentiles for the γ's used in the normal data in section 2.2.2. The simulation was repeated 100 times for each combination of γ and n.

2.3 Real data

2.3.1 Stress response in *E. faecalis*

Our initial motivation for introducing the rotation test for use in GSEA, was to be able to apply this method to direct comparison data with few samples. The data set described here is an example of such data, with a sample size of only four arrays.

A microarray experiment was conducted to test the genome-wide responses in the bacterium *Enterococcus faecalis* V583 to bile stress (sublethal concentra-

DOI: 10.2202/1544-6115.1418

tions). The experiment was designed as a direct comparison study: labelled cDNA from both bacteria treated with bile and untreated bacteria was hybridised to mutual slides. In this experiment one wanted to investigate the gene expression response to various treatment durations, and bile-treatment durations were chosen to be 10, 20, 60 or 120 minutes. In the following, these are treated as four separate experiments. There were data from four arrays in each experiment (2 $\times$ dye-swap), where bacteria had been sampled from two different batches. For further details on the labelling, hybridisations and data pre-processing, see Solheim *et al.* (2007).

Initially 3287 genes were spotted on each array. A minimum requirement for a gene to be included in the analysis, was to be present on at least 3 out of the 4 arrays from a time point. To be able to compare the results from all time points, only genes with sufficient observations in all four experiments (time 10, 20, 60 and 120) were included in the analysis. The number of genes meeting this requirement was 2350. Finally, missing data were imputed by k-nearest neighbours imputation (Troyanskaya *et al.*, 2001) implemented in R.

The differential expression between treated and untreated bacteria was measured as log_2(signal treated) - log_2(signal untreated). Loess normalisation implemented in the LIMMA package for R (Smyth and Speed, 2003) was used to correct for intensity dependent trends in the data. To correct for effect of batch and dye, an ANOVA model with the main effects of these two factors was fitted to the data. The residuals from the model represent the normalised log-ratios and were used in the following analysis. A similar normalisation was done by Wolfinger *et al.* (2001). We are aware that this normalisation will not remove all dependencies between arrays, which is one of the assumptions behind the rotation test. However, the normalisation part is not the main focus of this paper, so we will treat the arrays as independent samples after this normalisation.

Four different types of gene sets were tested in GSEA: 1) functional categories defined by The J. Craig Venter Institute (JVCI) (http://www.tigr.org), 2) pathways from Kyoto Encyclopedia of Genes and Genomes (KEGG) (http://www.genome.jp/kegg), 3) genes classified by first EC (Enzyme Commission) number downloaded from JVCI, and 4) operon predictions by the Virtual Institute for Microbial Stress and Survival (VIMSS) (http://www.microbesonline.org). Operons are sets of genes located adjacently in a bacterial genome and controlled by a common regulatory sequence. If transcription of an operon is induced, usually all genes in the operon are transcribed. Hence, high correlations are expected between the expression values of genes belonging to the same operon.

Gene sets were required to have at least 5 members corresponding to genes

on the microarray to be included in the analysis, which resulted in a total of 132 gene sets: 19 functional categories, 59 pathways, 6 EC groups and 48 operons.

In GSEA the genes are typically ranked by their individual association with a phenotype vector (correlation/two-sample t-statistic/signal-to-noise ratio etc.). However, since direct comparison data do not have a phenotype vector, we chose to rank genes by their t-statistic for testing the expected expression log-ratio to be different from zero. Because of the small number of arrays, the estimated variance of each gene was stabilised by adding the 90th percentile of the estimated variances for the p genes (Efron *et al.*, 2001). The t-statistic for gene j was estimated as

$$t_j = \frac{\bar{x}_j}{\sqrt{\frac{v_j + \tilde{v}}{2n}}} \tag{7}$$

where $\bar{x}_j$ is the average log-ratio for gene j over all arrays, v_j is the estimated variance of the gene, and $\tilde{v}$ is the 90th percentile variance estimate.

2.3.2 p53 status in cancer cell lines

Small sample sizes can be a problem also in indirect comparison data, thus the rotation test can be highly relevant for this kind of data as well. To illustrate that the rotation test is an alternative to the permutation test for GSEA on indirect comparison data, we applied GSEArot and GSEAperm to the p53 data set (Olivier *et al.*, 2002) used by Subramanian *et al.* (2005). Although this is a data set with large sample size (50 arrays), we were able to compare the results of the permutation test and the rotation test on a benchmark data set.

The aim of this analysis was to identify targets of the transcription factor p53 in expression patterns from the NCI-60 collection of cancer cell lines. The data set contains expression profiles from 50 cell lines, of which 17 were classified as wild type of the p53 gene and 33 were classified as mutant. GSEArot and GSEAperm were applied to a catalogue of 308 functional gene sets (see Subramanian *et al.* for details about the data).

To make the comparison of the rotation test and the permutation test as accurate as possible, we implemented our own version of GSEAperm. Due to this, there may be some differences in our procedure and the procedure of Subramanian *et al.*, and hence also in the results. Genes were ordered by their signal-to-noise ratio calculated as

$$\mathrm{SNR} = \frac{\bar{x}_1 - \bar{x}_2}{s_1 + s_2} \tag{8}$$

DOI: 10.2202/1544-6115.1418

where $\bar{x}_1$ and s_1 denote the gene's sample mean and sample standard deviation, respectively, for mutant, while $\bar{x}_2$ and s_2 denote the gene's sample mean and sample standard deviation for wild type. For these indirect comparison data we used a slightly different method for rotation, that in addition to maintaining the covariance matrix also maintains the mean vector, i.e. it allows non-zero means. See Langsrud (2005) for details.

3 Results

3.1 Simulations

3.1.1 Comparison of randomisation test and rotation test

The simulated data sets from section 2.2.1 were analysed with both GSEArand and GSEArot. For each level of the gene set correlation ρ, the probability of making a type I error was estimated as the average proportion of significant gene sets in the 100 simulated data sets. Figure 1 shows the estimated prob-

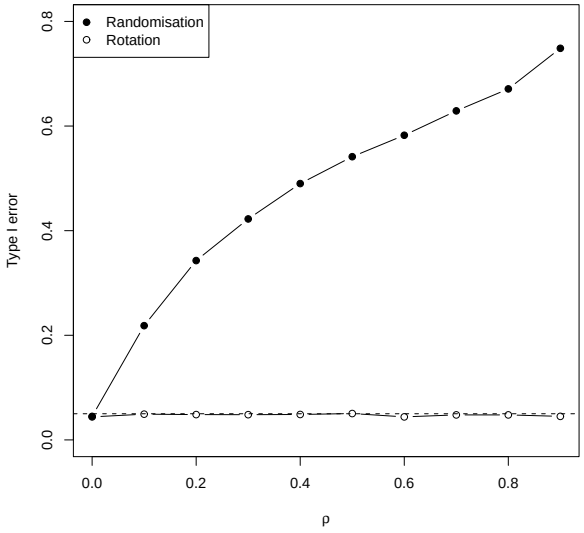


Figure 1: Estimated type I error for GSEA with randomisation test and GSEA with rotation test for different levels of correlation ρ within gene sets. The dashed line indicates the expected type I error of 0.05. The type I error of the randomisation test increases with correlation, while the rotation test controls the type I error for all levels of correlation.

ability of type I error for the randomisation test and the rotation test, for different levels of gene set correlation.

Considering that the genes were generated under the null hypothesis that all gene sets contain only non-differentially expressed genes, and a nominal test level of 5 % was used, we would expect a type I error of approximately 0.05 (indicated by the dashed line). As suspected though, the type I error for the randomisation test increases rapidly as the gene set correlation increases. For gene sets with correlation 0.1 the type I error is over 0.2, while a within gene set correlation of 0.9 gives an estimated type I error of more than 0.7. The rotation test, however, controls the type I error at a level of 0.05 for all correlations. This shows that gene randomisation has severe weaknesses when it comes to testing significance for a correlated set of genes, and hence the rotation test is to be preferred.

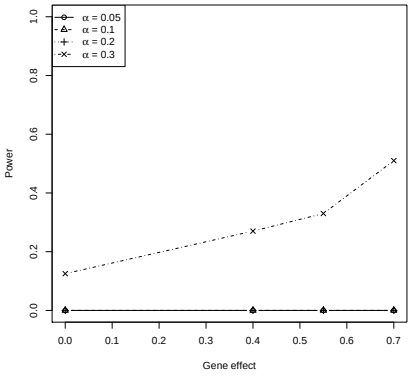
3.1.2 Rotation test and permutation test on normal data

The simulated normal data described in section 2.2.2 were analysed with GSEAperm and GSEArot. The power for each of the two methods was calculated as the proportion of the 100 simulations in which the gene set with the added gene effect was found to be significant. The probability of type I error was calculated from the simulated data set with an added gene effect of 0, as the average proportion of the 50 gene sets that was found to be significant over all simulations. The results of the power study and the type I error for different levels of the significance level α can be seen in Figure 2. The permutation test clearly has lower power than the rotation test for the data set with 4 samples. When the number of samples increases to 8, the power is more or less equal for both tests. The type I error rate is very low for the permutation test on 4 samples due to the low power, while the rotation test has the expected type I error rate for both sample sizes.

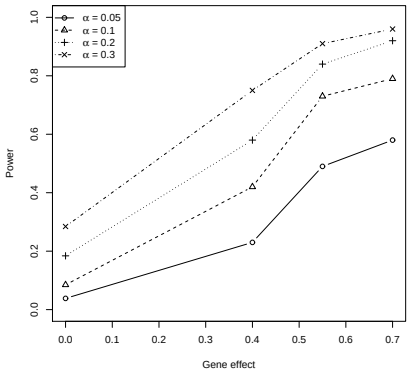
3.1.3 Rotation test and permutation test on non-normal data

The simulated log-normal data from section 2.2.3 were analysed with GSEAperm and GSEArot, and the power and probability of type I error were estimated as for the normal data in section 3.1.2. Figure 3 shows the estimated power and type I error for varying sample size, gene effect and significance level α.

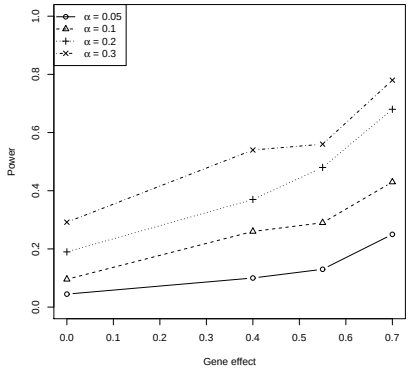
DOI: 10.2202/1544-6115.1418



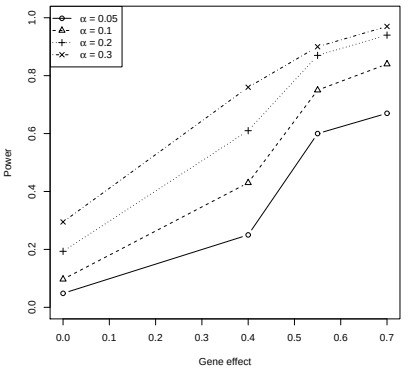
(a) Permutation test on 4 samples



(b) Permutation test on 8 samples

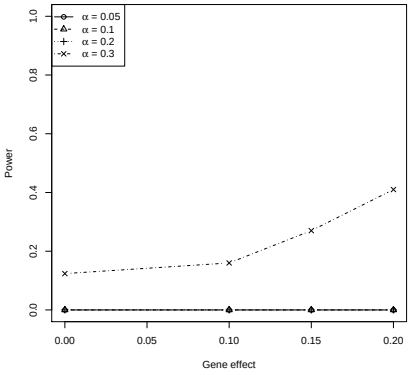


(c) Rotation test on 4 samples

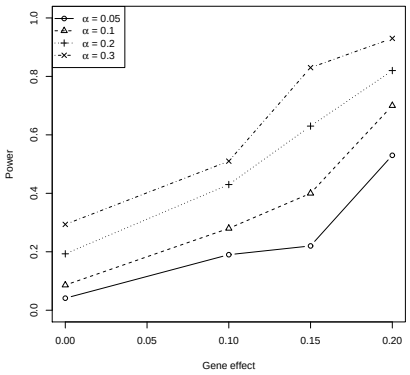


(d) Rotation test on 8 samples

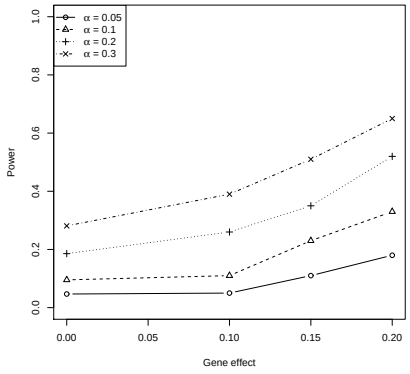
Figure 2: Power study for permutation test and rotation test on simulated normal data with added gene effect of 0, 0.4, 0.55 or 0.7 for the first gene set. The chosen significance level is indicated by α. The value for gene effect 0 is the estimated probability of type I error.



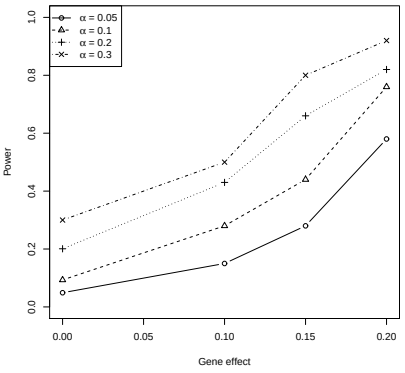
(a) Permutation test on 4 samples



(b) Permutation test on 8 samples



(c) Rotation test on 4 samples



(d) Rotation test on 8 samples

Figure 3: Power study for GSEA with rotation test on simulated log-normal data with added gene effect of 0, 0.1, 0.15 or 0.2 for the first gene set. The chosen significance level is indicated by α. The value for gene effect 0 is the estimated probability of type I error.

DOI: 10.2202/1544-6115.1418

3.2 Real data

3.2.1 Stress response in *E. faecalis*

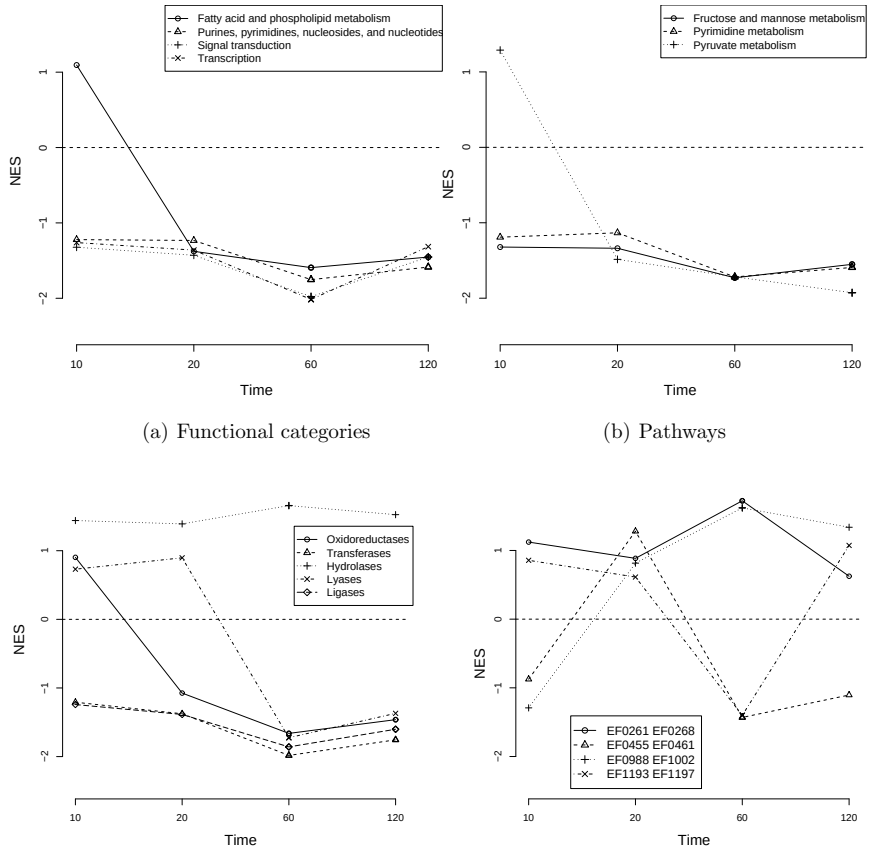
The bacteria data described in section 2.3.1 were analysed with GSEArot. Data from each of the four time points were analysed separately. The significant gene sets are presented in Table 1, along with q-values (the FDR analogue of the p-value, Storey, 2002), p-values, normalised enrichment scores (NES) and size of the gene sets. With a significance level of 0.2 for the q-value, no significant gene sets are found at times 10 and 20. At time 60 we find the largest share of significant sets, where 23 out of the 132 sets are significantly enriched. At time 120 there are 12 significant gene sets.

Although the enrichment scores are computed and all testing is done for each time point separately, it is valuable to study the time-course of the enrichment score. If a gene set is (significantly) enriched at several time points and there is a clear trend in the scores, this gives strong support to the conclusion that the genes in the set are affected by bile treatment. False positives for single time points may correspondingly be discovered through the absence of a trend over time. In Figure 4 the normalised enrichment score (NES) is plotted against time for all gene sets that were found to be significantly enriched at minimum one time point. A bold point indicates that the gene set is significant at the given time point. Figure 4(a) shows the time trend for significantly enriched functional categories. For most categories, the general trend seem to be repression in bacteria treated with bile, indicated by a negative NES at all time points. However, one functional category shows a different trend. Fatty acid and phospholipid metabolism has a decreasing trend, going from induced to repressed. In Figure 4(b) the expression pattern of the significant pathways can be studied. To reduce the number of pathways in the figure, only pathways that are significant at minimum two time points were plotted. Pyruvate metabolism show a decreasing trend, while the other two pathways are repressed at all times, similar to the general trend among the functional categories. Figure 4(c) shows the differential expression of the EC groups over time. Hydrolases has a positive NES for all time points, indicating that it is induced in bile-treated bacteria. Transferases and Ligases are repressed at all times. Oxidoreductases show a decreasing trend over time, while for Lyases it is difficult to point out a trend. Figure 4(d) shows that the operon EF0261 EF0268 has a positive NES at all time points, while EF0988 EF1002 show an increasing trend over time. The two operons EF0455 EF0461 and EF1193 EF1197 do not have any apparent trends.

Table 1: Significant gene sets ($q \leq 0.2$) from GSEArot on *E. faecalis* data set, with their estimated q-value, p-value and normalised enrichment score (NES). Size gives the number of genes in the gene set (notice that this is the number of genes in the set that correspond to genes in the expression data).

Gene set	q	p	NES	Size
Time 60				
Functional categories				
Signal transduction	0.000	0.002	-1.979	76
Transcription	0.000	0.002	-2.016	32
Purines, pyrimidines, nucleosides, and nucleotides	0.005	0.008	-1.753	56
Fatty acid and phospholipid metabolism	0.029	0.002	-1.593	33
Pathways				
Phosphotransferase system (PTS)	0.000	0.002	-1.930	34
Pyruvate metabolism	0.006	0.002	-1.717	22
Pyrimidine metabolism	0.006	0.006	-1.718	41
Fructose and mannose metabolism	0.007	0.002	-1.729	28
Glycerolipid metabolism	0.010	0.008	-1.672	11
Lysine biosynthesis	0.027	0.002	1.776	13
Citrate cycle (TCA cycle)	0.116	0.006	-1.468	6
Tyrosine metabolism	0.157	0.036	-1.428	9
Two-component system - General	0.163	0.012	-1.418	25
Glycolysis/Gluconeogenesis	0.200	0.018	-1.381	25
EC				
Transferases	0.000	0.002	-1.983	131
Ligases	0.001	0.002	-1.859	52
Lyases	0.006	0.004	-1.725	38
Oxidoreductases	0.011	0.004	-1.662	42
Hydrolases	0.049	0.004	1.658	84
Operons				
EF0261 EF0268	0.028	0.010	1.723	8
EF0988 EF1002	0.065	0.004	1.620	15
EF0455 EF0461	0.162	0.033	-1.431	7
EF1193 EF1197	0.185	0.062	-1.400	5
Time 120				
Functional categories				
Purines, pyrimidines, nucleosides, and nucleotides	0.075	0.002	-1.586	56
Signal transduction	0.170	0.070	-1.453	76
Fatty acid and phospholipid metabolism	0.185	0.002	-1.453	33
Pathways				
Pyruvate metabolism	0.003	0.002	-1.927	22
Alanine and aspartate metabolism	0.018	0.002	-1.777	17
Pyrimidine metabolism	0.083	0.002	-1.593	41
Fructose and mannose metabolism	0.102	0.002	-1.551	28
Glutamate metabolism	0.128	0.002	-1.520	17
Propanoate metabolism	0.141	0.002	-1.498	10
EC				
Transferases	0.019	0.002	-1.756	131
Ligases	0.095	0.002	-1.602	52
Oxidoreductases	0.188	0.002	-1.462	42

DOI: 10.2202/1544-6115.1418



(a) Functional categories

(b) Pathways

(c) EC

(d) Operons

Figure 4: Time series plots, showing the change in normalised enrichment score (NES) over time, for gene sets significantly enriched ($q \leq 0.2$) at minimum one time point (minimum two time points for pathways) in *E. faecalis* treated with bile. Gene sets that are significant at a given time point are marked in bold.

3.2.2 p53 status in cancer cell lines

GSEArot and GSEAperm were applied to the p53 data set, described in section 2.3.2, to illustrate that the rotation test is an alternative to the permutation test also for indirect comparison data. The comparison wild type>mutant identified five significant gene sets (q-value ≤ 0.25) with GSEAperm, and four significant gene sets with GSEArot. The comparison wild type<mutant gave one significant gene set with both methods. Both GSEArot and GSEAperm gave the same top six gene sets, although GSEArot gave slightly higher q-values. The results are presented in Table 2.

Table 2: Top six gene sets, with their corresponding q-value and p-value, from GSEAperm and GSEArot on the p53 data set.

	GSEAperm		GSEArot	
Gene set	q	p	q	p
Enriched in p53 mutant				
Ras signaling pathway	0.238	0.004	0.485	0.004
Enriched in p53 wild type				
Stress induction of HSP regulation	0.001	0.002	0.001	0.002
Hypoxia and p53 in the Cardiovascular system	0.001	0.002	0.002	0.002
p53 signaling pathway	0.002	0.002	0.008	0.002
p53 upregulated genes	0.003	0.004	0.012	0.002
Radiation sensitivity genes	0.051	0.007	0.131	0.006

No preprocessing of the data had been done, apart from gene probe reduction (see supporting information in Subramanian *et al.*, 2005). Since one of the assumptions of the rotation test is multinormally distributed arrays, we log-transformed the data to make them more normal and reanalysed them with GSEArot and GSEAperm. The results of the analysis can be viewed in Table 3. The comparison wild type>mutant gave no significant gene sets, while the comparison wild type<mutant gave three significant gene sets with both GSEArot and GSEAperm. Interestingly, two of the gene sets that were now identified, ngf pathway and igf1 pathway, fell just short of the significance threshold before the log-transformation. These gene sets were examined further by Subramanian *et al.*. Their leading edge subsets (i.e. the core members of the set) were found to contain four genes that they both share with the Ras signaling pathwway, which was significant both before and after log-transformation.

DOI: 10.2202/1544-6115.1418

Table 3: Top three gene sets, with their corresponding q-value and p-value, from GSEAperm and GSEArot on the log-transformed p53 data set.

	GSEAperm		GSEArot	
Gene set	q	p	q	p
Enriched in p53 mutant				
Ras signaling pathway	0.015	0.004	0.031	0.002
ngf pathway	0.099	0.006	0.143	0.006
igf1 pathway	0.234	0.006	0.215	0.008

4 Discussion

In this paper we have presented the rotation test as an alternative to the permutation test generally used for testing significance in Gene Set Enrichment Analysis. Since the permutation test requires a certain number of samples to generate accurate estimates of the p-values, applying GSEA to data with small sample sizes would be of little value. Especially for microarray experiments where direct comparison design has been used, the number of samples tends to be small.

Although our initial motivation was to find a significance test that would make GSEA applicable to direct comparison data with few samples, we would like to emphasise the versatility of the proposed rotation test. The rotation test can be applied to both direct (paired) and indirect (unpaired) comparison data. We would also like to point out that the test procedure is not only for use within the GSEA framework, but can also be used for other types of statistics testing significance for sets of genes.

Through a simulation study we demonstrated problems occurring when trying to permute genes rather than samples (randomisation), as an alternative to the permutation test. Gene sets with varying internal correlation levels were simulated, and analysis with GSEArand showed that the type I error for the randomisation test increased rapidly when correlation within gene sets increased. GSEArot, on the other hand, controlled the type I error for all levels of correlation. The simulation study comparing the power of the permutation test and the rotation test on normally distributed data, showed that the rotation test had higher power than the permutation test on 4 samples, while the power was approximately the same for 8 samples. The same results could be observed for the comparison of the two tests on simulated log-normal data, which indicates that the rotation test is fairly robust against deviations from the assumption about multinormally distributed arrays.

Both GSEAperm and GSEArot are testing a complete null hypothesis, meaning that all null hypotheses are assumed to be true. In this case it translates to assuming that all gene sets consist of only non-differentially expressed genes. The complete null hypothesis is a consequence of permuting or rotating all gene sets simultaneously, and will here lead to loss in power. In theory, we could avoid the complete null hypothesis by permuting/rotating only the gene set to be tested, but a few aspects of the analysis prevent this. First, we have overlapping gene sets, which means that rotating genes in one set would automatically rotate genes in other sets. This could potentially cause problems, and since we have used several different types of gene sets (pathways, functional categories etc.), we would expect a considerable overlap. Second, the choice of enrichment score as the test statistic means that the statistic of one gene set is dependent on the other sets. GSEArand is testing a different type of complete null hypothesis than GSEAperm and GSEArot, namely that all gene sets show a similar expression pattern. The randomisation test would always have to assume a complete null hypothesis, independent of overlapping sets and which test statistic is used, because genes are drawn at random from all sets of genes to create new artificial sets.

GSEArot was applied to a direct comparison data set describing the genome-wide effect of bile stress on *Enterococcus faecalis* V583. Previous studies have suggested that several aspects of the bile response are conserved among gram-positive bacteria (Bron *et al.*, 2006; Solheim *et al.*, 2007). In our analysis of the *E. faecalis* V583 data set, the key role of the membrane architecture and composition in bacterial bile tolerance was reflected in an enrichment of genes that code for proteins with membrane-associated functions and/or locations. Particularly, the functional categories of genes involved in fatty acid and lipid metabolism and signal transduction were strongly affected. The functional category Signal transduction mainly contains genes that code for regulatory two-component systems and phosphotransferase systems (PTS), two of the membrane-associated pathways that showed differential expression. Altogether, these results suggest that bile acids may insert into the membrane and interfere with the dynamics and coordinated function of multienzyme complexes located there.

In addition to emphasising the effects of bile on membrane integrity and structure, studies of bile stress in other gram-positive bacteria have suggested that bile imposes oxidative stress on the cell (reviewed in Begley *et al.*, 2005). By random gene disruption strategies in *E. faecalis*, Le Breton *et al.* (2002) identified a bile-sensitive mutant in which the amino acid sequence of the insertion locus showed homology to a putative oxidoreductase. This EC group (EC1 Oxidoreductases) was also significantly enriched with GSEArot at times

DOI: 10.2202/1544-6115.1418

60 and 120. Another observed response to bile was the downregulation of genes belonging to the functional category Purines, pyrimidines, nucleosides and nucleotides (cf. the pathway Pyrimidine metabolism). Differential expression of genes involved in pyrimidine metabolism have also been observed in response to detergents, antibiotics and NaCl-induced osmotic stress (Aakra *et al.*, 2005; Solheim *et al.*, 2007; Solheim *et al.*, unpublished results), and may be part of a more general stress response in *E. faecalis* V583.

The time trend plots reveal quite clear trends for many of the gene sets. On the other hand, lack of a clear trend may unveil false positives. The EC group Lyases does not show a clear trend, having a positive NES at times 10 and 20 and a negative NES at times 60 and 120, and where only time 60 is significant. This could be an indication that lyases are not affected in bacteria exposed to bile, and that the significant NES at time 60 is a false positive. The same can be hypothesized about the two operons EF0455 EF0461 and EF1193 EF1197, which lack a trend and also have a higher q-value than the two operons that show a time trend.

The rotation test, like the permutation test, is based on the assumption of sample independence. In many experiments there may occur between-array effects like batch effects (e.g. array lots, lab batches), array effects (e.g. for direct comparison data) and effects from other design factors that introduce dependencies between samples. It is quite common in microarray studies to deal with this by normalising the data up-front in an attempt to remove the unwanted and systematic within- and between-array effects. ANOVA-type batch corrections, for instance, are frequently used for microarray data, and for the *E. faecalis* data in this paper, a simple ANOVA correction of dye and batch effects was done. Due to dependence between the residuals, normalisation with ANOVA models will never be able to remove all correlation between arrays. At best it can reduce the correlation levels between arrays, but in some cases where the number of within-batch samples is limited, it may even increase correlations! Hence, the assumption of independence may very often be questionable. We do not intend to follow up this comprehensive, but important issue here. The purpose of applying GSEArot to the *E. faecalis* data in this paper was merely to demonstrate the method on data with very few samples. Even if there are doubts about the independence, the analysis seem to generate meaningful results.

ANOVA corrections with post-analyses on residuals is not the only option for dealing with systematic nuisance factors in relation to GSEA analyses. We may also think of a more integrative approach where both gene-effects and effects from nuisance factors are estimated together using an ANOVA model. The t-statistic in eq. (7) could then be replaced by gene specific t-statistics

from the ANOVA model. The rotations of the GSEA test should then be restricted to take place in a space orthogonal to the space spanned by the nuisance variables. The details on such a procedure need to be subject to further study. A consequence of such an ANOVA approach would be that all genes are assumed to have equal variance (homoscedastic error model), which of course is a stricter assumption than the smoothing of individual variances that is done in eq. (7). An alternative could perhaps be a model along the lines of the empirical Bayes model implemented in the LIMMA package for R (Smyth, 2004), for which random gene specific variances are assumed to follow a common prior distribution.

To illustrate that a rotation test can be seen as a generalisation of the permutation test, GSEArot was applied to an indirect comparison data set used by Subramanian *et al.* (2005). In the p53 data set the comparison of the permutation test and the rotation test showed that both tests gave the same top six gene sets, though the rotation test seemed to generate slightly larger q-values and found one less significant gene set than the permutation test did. Univariate histograms of the data showed quite strong deviations from normality. The rotation test is based on a multinormal distribution assumption, which may give loss in test power if this assumption is false. By the permutation test no explicit assumption of the distribution of the array $\mathbf{x_i}$ is made, other than that the expression values for the two classes come from the same distribution with identical means. The normal assumption of the rotation test may give more correct conclusions regarding gene set enrichment if the assumption is approximately true and the number of samples is limited, which is often the case in microarray studies. As the simulation studies showed, the rotation test had higher power than the permutation test on 4 samples for both normal and log-normal data. On the other hand, when the sample size is very large (50 arrays for these data) the assumption about multinormality can instead lead to some loss in power. There exist tests for multinormality, e.g. Mardia's test of multinormality (Mardia, 1985) and a multivariate Shapiro-Wilk test (Royston, 1982). But to our knowledge, none of these tests are applicable for situations where the number of variables exceeds the number of observations, as is the case for these types of data.

Although the rotation test seems fairly robust against violations of the multinormality assumption, the loss in power will probably be more noticeable for extreme deviations from normality. We also did a comparison of GSEAperm and GSEArot on the ALL-AML data set (Armstrong *et al.*, 2002) used by Subramanian *et al.,* for which univariate histograms showed even stronger deviations from normality than the p53 data. The results are not included in this paper, but the rotation test appeared to have lower power

DOI: 10.2202/1544-6115.1418

than the permutation test, finding only one significant gene set compared to the permutation test's five gene sets.

In an attempt to approach the rotation test's assumption about multinormality, we log-transformed the p53 data before reanalysing them. The analysis of these log-transformed data gave quite different results from before the transformation. GSEArot and GSEAperm found the same three significant gene sets, but only one of these corresponded with the significant gene sets in the untransformed data. A possible explanation for this could be the weighting of the enrichment score (see Subramanian *et al.* for details); the log-transformation will tone down the highest expression values, such that these genes now will have a lower influence on the enrichment score. This raises the question of what should and what does contribute to a gene set getting a high enrichment score, a discussion we will not go any deeper into here.

To summarise, the rotation test seem to be tolerably robust against deviations from the assumption about multinormally distributed arrays, but the power will be somewhat lower for extreme deviations from normality. For very small sample sizes the rotation test has proved to have higher power than the permutation test, both for normal and non-normal data, which was our initial aim with rotation testing in GSEA.

References

Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., den Boer, M. L., Minden, M. D., Sallan, S. E., Lander, E. S., Golub, T. R. and Korsmeyer, S. J. (2002) MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia, *Nature genetics*, **30**, 41-47.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nature genetics*, **25**, 25-29.

Begley, M., Gahan, C. G. and Hill, C. (2005) The interaction between bacteria and bile, *FEMS Microbiol Rev*, **29**, 625-651.

Box, G. E. P, Hunter, W. G. and Hunter, J. S. (1978) Statistics for experimenters, *John Wiley & Sons*.

Breton, Y. L., Maze, A., Hartke, A., Lemarinier, S., Auffray, Y. and Rince, A. (2002) Isolation and characterization of bile salts-sensitive mutants of Enterococcus faecalis, *Curr Microbiol*, **45**, 434-439.

Bron, P. A., Molenaar, D., Vos, W. M. and Kleerebezem, M. (2006) DNA micro-array-based identification of bile-responsive genes in Lactobacillus plantarum, *Journal of Applied Microbiology*, **100**, 728-738.

Efron, B. and Tibshirani, R. (2007) On testing the significance of sets of genes, *The Annals of Applied Statistics*, **1**, 107-129.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001) Empirical Bayes analysis of a microarray experiment, *Journal of the American Statistical Association*, **96**, 1151-1160.

Jiang, Z. and Gentleman, R. (2007) Extensions to gene set enrichment, *Bioinformatics*, **23**, 306-313.

Kerr, M. K., Martin, M., and Churchill, G. A. (2000) Analysis of variance for gene expression microarray data, *Journal of computational biology*, **7**, 819-837.

Khatri, P. and Drăghici, S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems, *Bioinformatics*, **21**, 3587-3595.

Kim, S. and Volsky, D. J. (2005) PAGE: Parametric Analysis of Gene Set Enrichment, *Bioinformatics*, **6**, 144.

Langsrud, Ø. (2005) Rotation tests, *Statistics and Computing*, **15**, 53-60.

Mardia, K. V. (1985). "Mardia's test of multinormality", in S. Kotz and N.L. Johnson, eds., *Encyclopedia of Statistical Sciences*, vol. 5 (NY: Wiley), 217-221.

DOI: 10.2202/1544-6115.1418

Mootha, V. K., Lindgren, C. M., Eriksson, K., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P, Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D. and Groop, L. C. (2003) PGC-1-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes, *Nature Genetics*, **34**, 267-273.

Nam, D. and Kim, S.-Y. (2008) Gene-set approach for expression pattern analysis, *Briefings in Bioinformatics*, **9**, 189-197.

Olivier, M., Eeles, R., Hollstein, M., Khan, M. A., Harris, C. C. and Hainaut, P. (2002) The IARC TP53 database: New online mutation analysis and recommendations to users, *Human Mutation*, **19**, 607-614.

Royston, J. P. (1982) An extension of Shapiro and Wilk's W test for normality to large samples, *Applied Statistics*, **31**, 115-124.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, **3**, No. 1, Article 3.

Smyth, G. K. and Speed, T. P. (2003). Normalization of cDNA microarray data. *Methods*, **31**, 265-273.

Solheim, M., Aakra, Å., Vebø, H., Snipen, L. and Nes, I. (2007) Transcriptional responses of *Enterococcus faecalis* V583 to bovine bile and sodium dodecyl sulfate, *Applied and Environmental Microbiology*, **73**, 5767-5774.

Storey, J. D. (2002) A direct approach to false discovery rates, *J. R. Statist. Soc. B*, **64**, Part 3, 479-498.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. and Mesirov, J. P (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles, *PNAS*, **102**, 15545-15550.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R. B., (2001) Missing value estimation methods for DNA microarrays, *Bioinformatics*, **17**, 520-525.

Wedderburn, R. W. M. (1975) Random rotations and multivariate normal simulation, Research Report, Rothamsted Experimental Station.

Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C. and Paules, R. S., (2001) Assessing gene significance from cDNA microarrays expression data via mixed models, *Journal of Computational Biology*, **8**, 625-637.

Aakra, A., Vebo, H., Snipen, L., Hirt, H., Aastveit, A., Kapur, V., Dunny, G., Murray, B. E. and Nes, I. (2005) Transcriptional response of *Enterococcus faecalis* V583 to erythromycin, *Antimicrob Agents Chemother*, **49**, 2246-2259.

DOI: 10.2202/1544-6115.1418

Paper II

Statistical Applications in Genetics and Molecular Biology

Volume 10, *Issue* 1 2011 *Article* 37

Smoothing Gene Expression Data with Network Information Improves Consistency of Regulated Genes

Guro Dørum, *Norwegian University of Life Sciences*
Lars Snipen, *Norwegian University of Life Sciences*
Margrete Solheim, *Norwegian University of Life Sciences*
Solve Saebo, *Norwegian University of Life Sciences*

Recommended Citation:
Dørum, Guro; Snipen, Lars; Solheim, Margrete; and Saebo, Solve (2011) "Smoothing Gene Expression Data with Network Information Improves Consistency of Regulated Genes," *Statistical Applications in Genetics and Molecular Biology*: Vol. 10: Iss. 1, Article 37.
DOI: 10.2202/1544-6115.1618
Available at: http://www.bepress.com/sagmb/vol10/iss1/art37

Smoothing Gene Expression Data with Network Information Improves Consistency of Regulated Genes

Guro Dørum, Lars Snipen, Margrete Solheim, and Solve Saebo

Abstract

Gene set analysis methods have become a widely used tool for including prior biological knowledge in the statistical analysis of gene expression data. Advantages of these methods include increased sensitivity, easier interpretation and more conformity in the results. However, gene set methods do not employ all the available information about gene relations. Genes are arranged in complex networks where the network distances contain detailed information about inter-gene dependencies. We propose a method that uses gene networks to smooth gene expression data with the aim of reducing the number of false positives and identify important subnetworks. Gene dependencies are extracted from the network topology and are used to smooth genewise test statistics. To find the optimal degree of smoothing, we propose using a criterion that considers the correlation between the network and the data. The network smoothing is shown to improve the ability to identify important genes in simulated data. Applied to a real data set, the smoothing accentuates parts of the network with a high density of differentially expressed genes.

KEYWORDS: differentially expressed genes, gene network, gene set analysis, microarray data analysis, enrichment analysis

Author Notes: We would like to thank two anonymous referees for their help in improving this manuscript.

1 Introduction

Gene Set Enrichment Analysis (Subramanian et al., 2005) and similar methods have in recent years become a popular way of evaluating gene expression data in light of background knowledge of gene sets. The fundamental idea is that sets of genes with some logical connection should show similarities with regard to expression level, and that differential expression should be evaluated at the gene set level instead of at the individual gene level. The benefit of this strategy may be a reduction in the number of false positives, since small, but consistent changes in expressions at the gene set level may be more reliable than large expression changes for individual genes. The sets of genes may be defined in different ways, but is of course motivated by the assumption of correlated expressions between the members of the set. Examples include metabolic pathways, functional categories and gene ontology levels. Various versions of gene set analysis methods have been proposed by Kim and Volsky (2005), Jiang and Gentleman (2007) and Efron and Tibshirani (2007). See Huang et al. (2009) for a recent review.

It is however known that genes are arranged in complex networks, and gene set methods do not take full advantage of the information contained in these networks. Gene set methods require that the genes are divided into groups, while genes may take part in several reactions and do not necessarily fall into just one group. There may therefore be considerable overlap between groups. By shifting the focus from gene sets to gene networks, we can avoid the division into groups and at the same time make full use of the information about gene dependencies contained in the network distances. The fundamental idea is similar to the idea behind gene set methods, that there is a connection between network distance and gene expression similarity.

A growing number of papers are describing methods that make use of detailed network information for the analysis of expression data. Vert and Kanehisa (2003) presented a method for correlating gene networks and gene expression data. Rahnenführer et al. (2004) used distance between pairs of genes to improve statistical scores for finding active pathways. Hanisch et al. (2002) used information about gene networks to improve clustering of gene expression data. In regression modelling network information has also been used to smooth the estimated regression coefficients. Both Li and Li (2008) and Pan et al. (2010) used the network information as a penalization constraint in the parameter estimation, whereas Sæbø et al. (2008) used the network information to adjust the rotations in the Partial Least Squares regression model in their L-PLS. Shojaie and Michailidis (2009) incorporated network information in a latent variable model and used a mixed linear model framework to test the significance of subnetworks, with a generalisation (Shojaie and Michailidis, 2010) of the method to handle more complex experimental de-

signs and test several contrasts simultaneously. Rapaport et al. (2007) used network information to extract the relevant signals in the gene expression data by removing the high-frequency components, and adapted this to classification.

Our approach is similar to the one in Rapaport et al. in that we aim at smoothing away the part of the gene expression data that represents noise. The goal is to eliminate false positives and accentuate important subnetworks with a high density of differentially expressed genes. The method is demonstrated on data simulated from one fictional and three real networks, and on a data set from a real experiment on *Enterococcus faecalis*.

2 Method

In this section we describe the method of network smoothing. The procedure requires that some type of network information is available for the genes in the expression data. A matrix of distances between genes is extracted from the graph topology. We assume that network distances correspond to similarity between the genes' expression patterns, and refer to this matrix as a similarity matrix. The similarity matrix is then used for smoothing of the gene expression data.

2.1 Similarity matrix

A predefined gene network, containing g genes, can be represented as a simple graph G, with genes as nodes and edges between genes representing some biological relationship. A simple graph is undirected, contains no loops and has at maximum one edge between each pair of nodes. Let i and j represent two nodes in G, and let $i \sim j$ indicate that the two nodes are adjacent (directly connected). The $g \times g$ adjacency matrix $\mathbf{A}$ describes the nodes' neighbourhoods, and the entries a_{ij} are

$$a_{ij} = \begin{cases} 1 & \text{if } i \sim j \\ 0 & \text{else} \end{cases} \tag{1}$$

for $i, j = 1, 2, ..., g$. The $g \times g$ degree matrix $\mathbf{D}$ is a diagonal matrix with the degree, i.e. the number of edges to a node, on the diagonal. Let δ_i be the degree of node i. The entries in $\mathbf{D}$ are

$$d_{ij} = \begin{cases} \delta_i & \text{if } i = j \\ 0 & \text{else} \end{cases} \tag{2}$$

DOI: 10.2202/1544-6115.1618

The $g \times g$ Laplacian matrix (Chung, 1997) is defined as $\mathbf{L} = \mathbf{D} - \mathbf{A}$, where the entries are

$$l_{ij} = \begin{cases} -1 & \text{if } i \sim j \\ d_{ij} & \text{if } i = j \\ 0 & \text{else} \end{cases} \tag{3}$$

There are numerous ways of translating the network topology of G into similarity between genes. Here, we will focus on the diffusion kernel (Chung, 1997, Kondor and Lafferty, 2002) as a similarity measure. The concept of diffusion is closely related to random walks and can be imagined as information, just like a fluid, travelling through the network. The diffusion kernel, or diffusion matrix $\mathbf{S}_\beta$ as we will refer to it here, is defined by the matrix exponential of $\mathbf{L}$:

$$\mathbf{S}_\beta = e^{-\beta \mathbf{L}} = \sum_{i=1}^{g} v_i e^{-\beta \lambda_i} v_i^T \tag{4}$$

where v_i and λ_i are the i'th eigenvector and eigenvalue of $\mathbf{L}$, respectively. $\mathbf{S}_\beta$ depends on a parameter β, where $\beta > 0$, that controls the speed of diffusion through the graph. The diffusion is faster for larger values of β, corresponding to shorter distances between nodes. The diffusion matrix contains numbers between 0 and 1, and all rows and columns sum to 1. Figure 1 shows a fictional network and Figure 2 shows its diffusion matrix $\mathbf{S}_\beta$ when the diffusion parameter β is set to 0.1 and 0.7, respectively. In the diffusion matrix black indicates values close to 0 and corresponds to long distances between nodes, and white indicates values close to 1 and corresponds to short distances. With $\beta = 0.1$ there are large distances also between the closely connected nodes, indicated by close to white diagonal elements and black off-diagonal elements. When β is increased to 0.7, the shorter distances between nodes is particularly apparent for the tightly connected nodes 1 to 6, indicated by the lighter area around these nodes. Node 17 is directly connected to only one node, and the colour of the diagonal element has not changed much from $\beta = 0.1$ to $\beta = 0.7$. The information has only one way of travelling to and from node 17, so it will remain mostly unaffected by an increase in the diffusion.

Increasing β too much will eventually result in all connected genes getting an identical diffusion value. Since the purpose in this paper is to use the diffusion matrix for smoothing, we have chosen an upper limit on β by requiring that the diagonal should always contain the largest number in each row. This is equivalent to requiring that a node should always put most weight on itself. We refer to the upper limit as β_{max}. In section 2.3 we propose a measure for finding the optimal value of β between 0 and β_{max}.

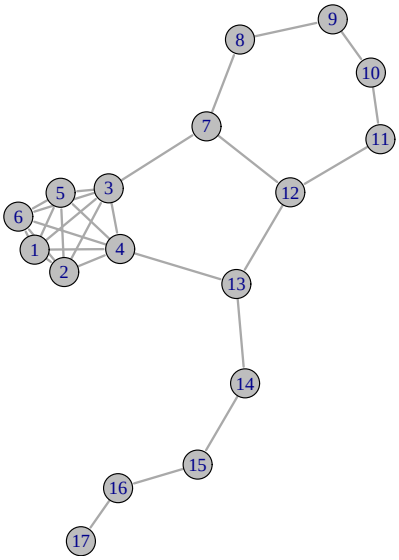
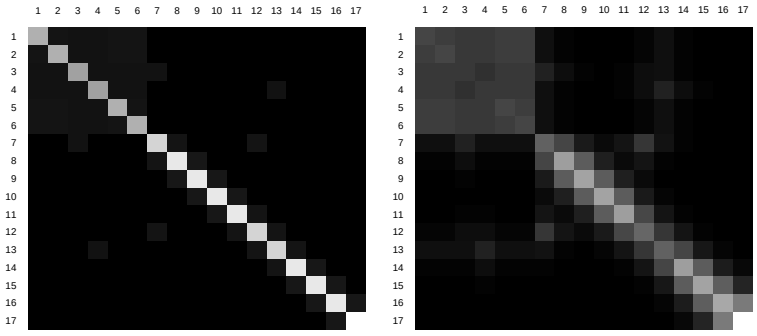


Figure 1: Fictional network.



(a) $\beta = 0.1$ (b) $\beta = 0.7$

Figure 2: Graphical representation of the diffusion matrix for the fictional network. Black indicates large distances and white indicates small distances.

DOI: 10.2202/1544-6115.1618

2.2 Smoothing expression data

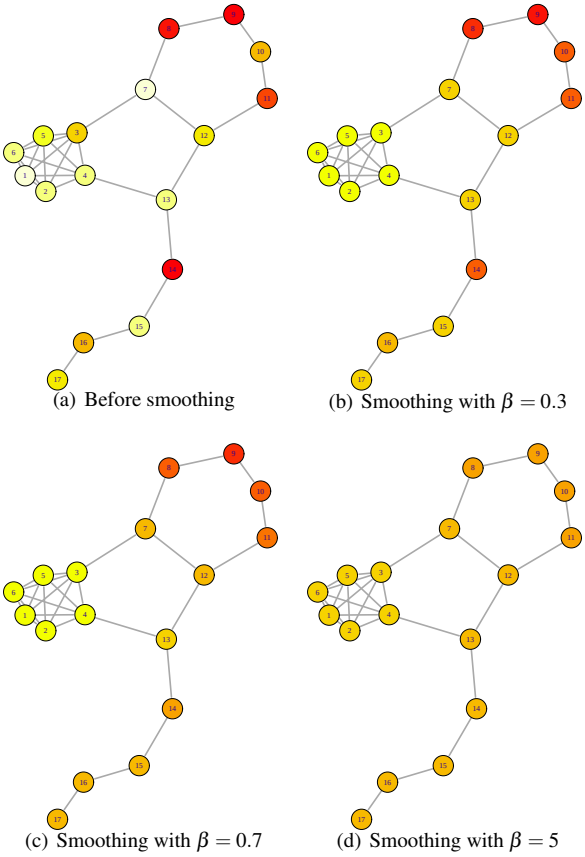
Let $\mathbf{X}$ denote an $n \times g$ matrix of gene expression levels measured for g genes on n samples. A test statistic, e.g. a t-statistic for testing differential expression between two conditions, correlation between the gene and a phenotype vector or a signal-to-noise ratio, is calculated for each gene. Let $\mathbf{t}$ denote the $g \times 1$ vector of test statistics. The absolute values of the test statistics are multiplied with the diffusion matrix $\mathbf{S}_\beta$ to obtain a vector of smoothed test statistics $\mathbf{t}_\beta$:

$$\mathbf{t}_\beta = \mathbf{S}_\beta |\mathbf{t}| \tag{5}$$

The diffusion matrix acts like a weighting matrix since each row sums to 1. The smoothing will give closely connected genes a more similar test statistic and tone down extreme observations. This agrees with the intention of the network smoothing approach; we wish to detect smaller changes within a number of related genes rather than large changes in a few unrelated genes. Nodes without any neighbours are unaffected by the smoothing and keep their original test statistic. Note that since we use absolute values, all smoothed test statistics are positive. We are only detecting whether a gene is differentially expressed, not whether it is up- or down-regulated. However, the direction of regulation may be extracted as the sign of $\mathbf{t}$. Figure 3 shows how the test statistics for the nodes in the fictional network are changing with different levels of smoothing. Each node is coloured by the magnitude of its test statistic, where red is largest and white is smallest. Figure 3(a) shows the test statistics before smoothing. In 3(b) the test statistics are smoothed with a medium value of β, and in 3(c) they are smoothed with β_{max}, which is 0.7 for this network. In Figure 3(d) the degree of smoothing is so extreme that all nodes get an almost identical test statistic. The smoothing implies a sharing of power, where nodes with high test statistics transfer some of their power to their neighbouring nodes. While this may lead to a discovery of subnetworks with more moderate expression, a consequence is that we risk losing individual important nodes. In the next section we propose a way of finding the optimal value of β for a given network.

2.3 Optimal smoothing

The optimal level of smoothing should ideally be determined by the data. Of course, the analyst could alternatively screen a set of values for β between 0 and β_{max} and inspect the results in order to find a level that gives a reasonable outcome in light of prior knowledge, but this may put too much subjectivity into the results and prevent discoveries which do not harmonise with prior knowledge. We therefore seek an optimal level of smoothing solely dependent on the data through some optimality criterion.



(a) Before smoothing (b) Smoothing with $\beta = 0.3$

(c) Smoothing with $\beta = 0.7$ (d) Smoothing with $\beta = 5$

Figure 3: Nodes coloured by the magnitude of their test statistic before and after smoothing with different values of β. Red is largest, white is smallest.

Most statistical smoothing methods depend on one or several regularization parameters controlling the level of smoothing. In some cases, like for prediction models, it is quite straightforward to define some loss function that can serve as a criterion for choosing optimal values for the smoothing parameter(s). An example is non-linear regression, where the width of the moving-average window can be chosen to minimize prediction error. In other problems, like the one issued in this paper, the choice of loss function is not that obvious.

DOI: 10.2202/1544-6115.1618

The aim of our smoothing approach is to identify subnetworks or communities (Newman, 2006) which show a coherent expression pattern that stands out from other parts of the network. The optimality criterion should therefore seek a level of smoothing for which the smoothed gene statistics within communities are more positively correlated than for genes from different communities. A criterion suited for this purpose is the network correlation (Newman, 2002, 2003, 2006). This correlation function is a measure of *assortative mixing*, which measures to what extent adjacent nodes in a network have similar properties. The measure is based on the modularity matrix $\mathbf{B}$ (see e.g. Newman, 2006), which is a matrix representation of the community structure of a network. The modularity matrix is defined as $\mathbf{B} = \mathbf{A} - \mathbf{P}$, where $\mathbf{A}$ is the adjacency matrix as defined in eq. (1), and $\mathbf{P}$ contains the expected number of edges between each pair of nodes. The expected number of edges between nodes i and j if edges are placed at random is

$$p_{ij} = \frac{\delta_i \delta_j}{2m} \tag{6}$$

where δ_i and δ_j are the degrees of the nodes and m is the total number of edges in the network. The network correlation of the smoothed test statistic $\mathbf{t}_\beta$ across the network for a given value of β is defined by

$$r(\beta) = \frac{1}{2m} \mathbf{t}_\beta^T \mathbf{B} \mathbf{t}_\beta \tag{7}$$

Based on this, we define the optimal level of smoothing, β_{opt}, as

$$\beta_{opt} = \text{argmax}_\beta \left(r(\beta) \right) \tag{8}$$

For our fictional network and simulated test statistics, the function $r(\beta)$ is shown in Figure 4. The maximum was found to be $\beta_{opt} = 0.3$ with a network correlation of $r(0.3) = 0.09$. The corresponding smoothed network was shown in Figure 3. As this figure suggests, it is not much to gain when increasing β from 0.3 to 0.7, so it seems reasonable that the optimum is found here.

2.4 Estimation of significance

For each gene we are testing the null hypothesis of the gene being differentially expressed. The test is conditional on the smoothing matrix used. The null distribution of the smoothed test statistics is unknown, and therefore a resampling based method must be adopted for significance testing. Let $\mathbf{X}_r$ denote a $n \times g$ matrix of resampled gene expression data. A $g \times 1$ vector of genewise test statistics $\mathbf{t}_r$ based

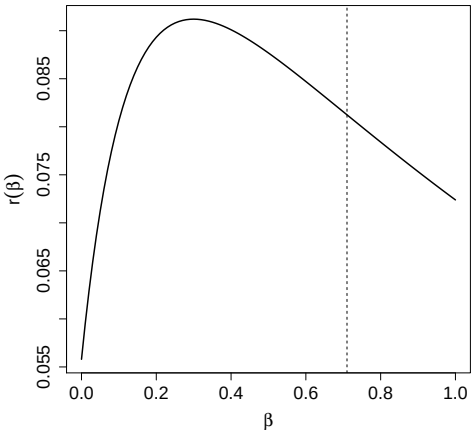


Figure 4: The network correlation $r(\beta)$ for different values of β over the fictional network. The dashed line indicates β_{max}.

on the resampled data, are smoothed with the similarity matrix to obtain smoothed resampled test statistics $\mathbf{t}_{\beta r}$

$$\mathbf{t}_{\beta r} = \mathbf{S}_\beta |\mathbf{t}_r|$$

By repeating this procedure for a large number R of resampled data sets, the vectors of resampled test statistics $\mathbf{t}_{\beta 1}, \mathbf{t}_{\beta 2}, ..., \mathbf{t}_{\beta R}$ give an estimate of the distribution of the smoothed test statistics under the null hypothesis. Let $t_{\beta j}$ denote the observed smoothed test statistic and $t_{\beta rj}$ denote a resampled smoothed test statistic for gene j. A p-value for gene j is calculated as the proportion of resampled test statistics at least as extreme as the observed test statistic

$$p_j = \frac{\#(t_{\beta rj} \geq t_{\beta j}) + 1}{R + 1}$$

The choice of resampling test depends on the type of data. For single-channel data or two-colour data with a common reference (indirect design), a permutation test shuffling a phenotype vector can be used. For two-colour data with direct design, a permutation test exchanging signs can be applied. A problem with the permutation test, for both data types, arises when the number of samples is small, which is often the case for microarray data. Small sample sizes mean that the number of possible permutations is limited, and the accuracy of the estimated

DOI: 10.2202/1544-6115.1618

p-values will be accordingly low. We therefore suggest using a rotation test rather than a permutation test for cases with small sample sizes. The rotation test can rotate samples in all directions while still preserving the correlation structure between the genes (Langsrud, 2005). As shown by Dørum et al. (2009), the rotation test has higher power than the permutation test for small sample sizes.

In the following we have adopted a similar notation to Langsrud. The rotation test assumes that the rows of $\mathbf{X}$ are multinormal and independent, i.e. that each array $\mathbf{x}_i \sim N_g(\mu, \Sigma_x)$ and that the arrays are independent. The rotation test is however robust to deviations from normality, as shown by Dørum et al. and Wu et al. (2010). By a random rotation of $\mathbf{x}_i$ we get $\mathbf{x}_{ir} \sim N_g(\mathbf{0}, \Sigma_x)$. The rotated genes have expectation 0, but the covariance matrix is maintained. To perform a random rotation of the data, we start by noting that the matrix of gene expressions $\mathbf{X}$ can be decomposed by the QR decomposition

$$\mathbf{X} = \mathbf{X}_Q \mathbf{X}_U$$

where $\mathbf{X}_Q$ is an orthonormal matrix of size $n \times n$, and $\mathbf{X}_U$ is an upper triangular matrix of size $n \times g$ with positive diagonal elements. $\mathbf{X}_Q$ represents the orientation, while $\mathbf{X}_U$ represents the configuration and is a sufficient statistic for Σ_x. Let W be a $n \times n$ matrix of random standard normal distributed data. A QR decomposition of $\mathbf{W}$ gives $\mathbf{W} = \mathbf{W}_Q \mathbf{W}_U$, where $\mathbf{W}_Q$ is an $n \times n$ random rotation matrix. A random rotation matrix multiplied with another rotation matrix is still a random rotation matrix, and a rotated data matrix $\mathbf{X}_r$ can therefore be generated as

$$\mathbf{X}_r = \mathbf{W}_Q \mathbf{X}_Q \mathbf{X}_U = \mathbf{Q} \mathbf{X}_U \tag{9}$$

where $\mathbf{Q} = \mathbf{W}_Q \mathbf{X}_Q$. The rotations are conditioned on Σ_x, so this procedure makes it possible to account for covariances between genes without having to estimate Σ_x. See Dørum et al. and Langsrud for more details on the rotation test.

2.5 Simulated data

Gene expression data were simulated based on the structure of four different networks: the fictional network in Figure 1 with two additional disconnected subgraphs, and the three real pathways "Energy metabolism", "Lipid metabolism" and "Glycan biosynthesis metabolism" from the bacterium *E. faecalis*. Each network can be represented as a graph G with g nodes (genes). An additional node, the "pheno node", was connected to three adjacent nodes in G. Nodes that are highly correlated with the pheno node are considered important. The pheno node has of course no biological meaning, and is used here purely as a statistical approach for

associating nodes with the phenotype. Simulating data this way gives us control of the dependencies between nodes. All nodes that to some extent are connected to the pheno node, will be correlated with it. Let 1st order neighbours denote nodes that are directly connected to the pheno node, 2nd order neighbours denote nodes that are connected to the pheno node through the 1st order neighbours, and so on. Detached nodes refer to nodes that are completely disconnected from the pheno node.

For the fictional network, data were simulated from three different scenarios, where in each scenario the pheno node was connected to different nodes. Hence, all scenarios consider different nodes as important. The three scenarios can be seen in Figure 5. In scenario 1, the pheno node is connected to nodes 1, 2 and 6, which are located in a very dense part of the network where all nodes are directly connected to each other. Nodes 3-5 are the 2nd order neighbours in this scenario, while nodes 7 and 13 are the 3rd order neighbours. In scenario 2, the pheno node is connected to nodes 9, 10 an 11 in a less dense part of the network. The 2nd order neighbours are 8 and 12, and the 3rd order neighbours are 7 and 13. In scenario 3, the pheno node is connected to the same part of the network as in scenario 1, but this time only to node 1. The 2nd order neighbours are now nodes 2-6, while the 3rd order neighbours are still nodes 7 and 13. In all scenarios, nodes 18-22 and 23-27 are completely irrelevant for the pheno node since they are disconnected from the part of the network connected to the pheno node. These nodes were used to compute the type I error rate (false positives).

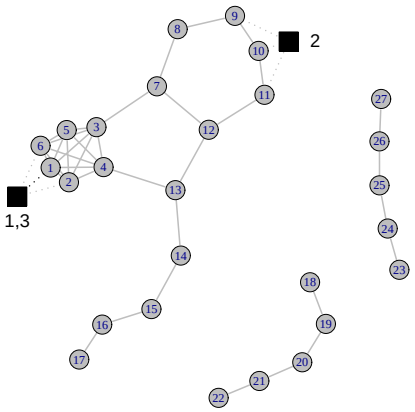
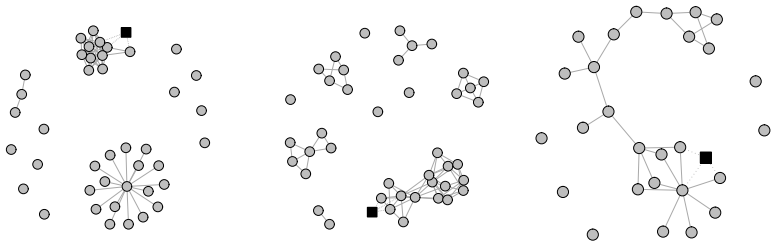


Figure 5: The fictional network used in the simulation study and the three scenarios data were simulated under. The square node represents the pheno node.

DOI: 10.2202/1544-6115.1618



(a) Energy metabolism (b) Lipid metabolism (c) Glycan biosynthesis metabolism

Figure 6: The three *E. faecalis* networks with a connected pheno node (square) used in the simulation study.

The structure of the three *E. faecalis* pathways with a pheno node connected to three adjacent nodes are shown in Figure 6. Energy metabolism has 41 nodes, Lipid metabolism has 42 nodes, and Glycan biosynthesis metabolism has 27 nodes. They are three rather different networks and contain several subgraphs that are not associated with the pheno node.

The correlations between the nodes, including the pheno node, were constructed with an exponential correlation model (see e.g. Diggle et al., 1994)

$$r_{ij} = e^{-\alpha b_{ij}} \qquad \text{for } i,j = 1,...,g+1 \tag{10}$$

for some $\alpha > 0$, where b_{ij} denotes the distance between node i and j. We measured distance as the inverse diffusion distance, where b_{ij} is the inverse of the ij'th element of the diffusion matrix with $\beta = 1$. With this model, increasing the distance between genes in the diffusion matrix will decrease the correlations towards zero. The rate of decrease is faster for larger values of α. We chose $\alpha = 0.3$ and used random signs on the correlations to simulate both activation and inhibition between the nodes. Table 1 gives the average correlations between the pheno node and its 1st, 2nd and 3rd order neighbours, plus the average correlation within the 1st order neighbours, between 1st and 2nd order neighbours and within the 2nd order neighbours, for each scenario/network.

Let $\mathbf{R} = \{r_{ij}\}$ denote the $(g+1) \times (g+1)$ matrix of correlations between the g nodes and the pheno node. For each network, gene expression data for $n = 20$ samples were simulated as follows. Let $\mathbf{U}$ be a $n \times (g+1)$ matrix of standard normal data, where the first g columns represent the genes and the last column represents the pheno node. A dependency structure between the genes was introduced by multiplying $\mathbf{U}$ with $\mathbf{R}^{\frac{1}{2}}$, an upper triangular matrix obtained by Cholesky decomposition

Table 1: Average correlations between the pheno node and its 1st, 2nd and 3rd order neighbours, within the 1st order neighbours, between 1st and 2nd order neighbours and within the 2nd order neighbours. Note that in scenario 3 there is only one 1st order neighbour, and in Energy metabolism there are no 3rd order neighbours.

	pheno-1st	pheno-2nd	pheno-3rd	1st-1st	1st-2nd	2nd-2nd
Scenario 1	0.72	0.41	< 0.01	0.78	0.48	0.41
Scenario 2	0.57	0.08	< 0.01	0.38	0.17	0.04
Scenario 3	0.47	0.06	< 0.01	–	0.57	0.48
Energy	0.44	0.12	–	0.30	0.31	0.43
Lipid	0.48	0.11	< 0.01	0.46	0.21	0.19
Glycan	0.49	0.03	< 0.01	0.24	0.1	0.09

of $\mathbf{R}$. The transformed data $\mathbf{Z} = \mathbf{U} \cdot \mathbf{R}^{\frac{1}{2}}$ have the distribution

$$\mathbf{Z} \sim N_{g+1}\left(\begin{bmatrix} \mu_{\mathbf{x}} \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \sigma_{\mathbf{xy}} \\ \sigma_{\mathbf{xy}} & \sigma_y^2 \end{bmatrix} \right) \tag{11}$$

where Σ_{xx} are the gene covariances, $\sigma_{\mathbf{xy}}$ are the gene-pheno covariances, and σ_y^2 is the pheno node variance. Note that since σ_y^2 and the gene variance σ_x^2 were both set to 1 for simplicity, we have that $\Sigma_{xx} = \mathbf{R_{xx}}$ and $\sigma_{\mathbf{xy}} = \mathbf{r_{xy}}$, where $\mathbf{R_{xx}}$ and $\mathbf{r_{xy}}$ are the gene-gene and gene-pheno correlations, respectively. The first g columns of $\mathbf{Z}$ constitute the gene expression matrix $\mathbf{X}$. The last column, which represents the expression of the pheno node, was rounded to either 0 or 1 to form a phenotype vector $\mathbf{Y}$ representing two phenotypes (e.g. diseased and healthy).

A total of 100 data sets were simulated from each scenario/network. A two-sample t-test comparing the population means of the two phenotypes was performed for each gene. The value of β_{max} was determined to 0.7 for the fictional network, 2.2 for Energy metabolism, 0.4 for Lipid metabolism and 0.5 for Glycan biosynthesis metabolism. Diffusion matrices for ten equally spaced β values between 0.001 and β_{max} were found, and the t-statistics were smoothed with all ten diffusion matrices. Since this is a case of indirect design data with a quite large sample size, statistical significance for the genes were estimated with a permutation test shuffling the phenotype vector.

Power was calculated as the proportion of the 100 data sets in which the most important nodes were found to be significantly expressed (p-value ≤ 0.05). Power was calculated separately for the 1st, 2nd and 3rd order neighbours of the pheno node. Power was also calculated for the detached nodes as an estimate of the type I error rate. In addition, the optimal smoothing parameter β_{opt} was found for each simulated data set.

DOI: 10.2202/1544-6115.1618

2.6 Stress response in *E. faecalis*

This microarray experiment was performed in order to investigate the transcriptional response of the bacterium *Enterococcus faecalis* V583 to bile stress. RNA was extracted from bacteria treated with bile and from untreated cultures as a reference, and reverse transcribed. Labelled cDNA was then hybridised to mutual slides in a direct design experiment. RNA was obtained in two separate growth experiments, and samples were collected after 10, 20, 60 and 120 minutes. At each time point four arrays were used, and the time points were analysed separately. For details about labelling, hybridisations and data pre-processing, see Solheim et al. (2007).

The differential expression between treated and untreated bacteria was measured as log_2(signal treated) - log_2(signal untreated). Loess normalisation implemented in the LIMMA package for R (Smyth and Speed, 2003) was used to correct for intensity dependent trends in the data. To make the samples as independent as possible, an ANOVA model incorporating the unwanted effects of batch and dye were fitted to the data, and the resulting residuals were used in the following analysis. Though this normalisation will not remove all dependencies between the samples, it should be sufficient for the purpose in this paper.

From the KEGG PATHWAY database (Kanehisa and Goto, 2000), 83 metabolic and non-metabolic pathways in *E. faecalis* were downloaded and converted into graphs. These graphs were merged together to one large graph with the R package KEGGgraph (Zhang and Wiemann, 2009), removing redundant nodes and edges due to overlapping pathways. The resulting graph consisted of 800 nodes and 1306 edges where each node represents a gene product (two nodes may represent the same gene product). The 800×800 diffusion matrix $\mathbf{D}$ was calculated for the graph. Rows and columns in $\mathbf{D}$ corresponding to nodes that we did not have gene expression data for were removed, reducing $\mathbf{D}$ to a 633×633 matrix. To make each row in $\mathbf{D}$ sum to 1 again, the values that were removed from a row were added to the row's diagonal. This preserves the network information as much as possible even if some genes are not spotted on the array.

A t-statistic for testing the expected expression log-ratio to be different from zero was calculated for each of the 633 genes. Because of the small number of samples, the estimated variance of each gene was stabilised by adding the 90th percentile of the estimated variances for all genes (Efron et al., 2001). The t-statistic for gene j was computed as

$$t_j = \frac{\bar{x}_j}{\sqrt{\frac{v_j + \tilde{v}}{2n}}} \tag{12}$$

where $\bar{x}_j$ is the average log-ratio for gene j over all four arrays, v_j is the estimated variance of the gene, and $\tilde{v}$ is the 90th percentile variance estimate. The t-statistics were smoothed with the diffusion matrix with $\beta = 0.1$, which was both β_{max} and β_{opt} for this network. Without the upper limit, β_{opt} would have been 0.17. Due to the small sample size, a rotation test was used to compute p-values. The false discovery rate (FDR) was controlled by computing adjusted p-values with the Benjamini-Hochberg method (Benjamini and Hochberg, 1995).

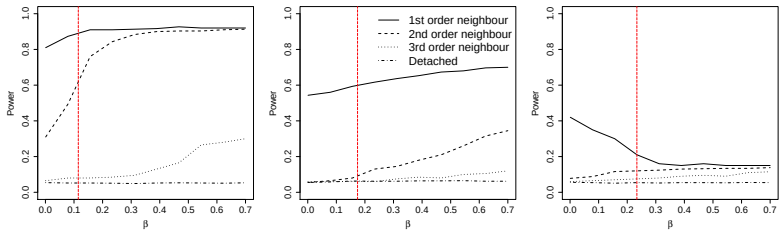
To shift the focus from individual genes to subnetworks, Gene Set Enrichment Analysis (GSEA) was performed on the individual KEGG pathways that made up the network used for smoothing. The pathways were required to have at least 5 members, so only 56 of the 83 pathways were tested. Gene set enrichment scores were computed based on the smoothed t-values (we refer to Subramanian et al. (2005) on how to compute enrichment scores). Each gene set was assigned a p-value computed with a rotation test. See Dørum et al. (2009) for details about GSEA with rotation test. To correct for multiple hypothesis testing, FDR q-values (Storey, 2002) were calculated for each gene set. Since all smoothed test statistics are positive, we are only interested in gene sets with high positive enrichment scores. We therefore used a one-sided version of the approach for computing q-values in Subramanian et al.. GSEA was also performed on non-smoothed absolute t-values for comparison.

3 Results

3.1 Simulated data

For each network/scenario the 100 simulated data sets were analysed as described in section 2.5. Figure 7 shows the average power for the 1st, 2nd and 3rd order neighbours, plus the detached nodes, as a function of the smoothing parameter β in the three scenarios of the fictional network. The average β_{opt} over all simulations are also indicated in the figure. Smoothing with $\beta = 0.001$ approximately corresponds to no smoothing. The nodes have different power in the three scenarios also without smoothing. This is a result of using diffusion distances in the simulation of correlations, meaning that the correlations between nodes depend on the topology of the network. The power without smoothing can be seen in connection with the correlations in Table 1. Networks with high correlation between the pheno node and its neighbours have high power. The benefit of smoothing for the 1st order neighbours seems to be similar for both scenario 1 and 2, while scenario 1 has a higher increase in power for the 2nd order neighbours. Scenario 1 has higher correlations both between the 1st and 2nd order neighbours and within the 2nd order

DOI: 10.2202/1544-6115.1618



(a) Scenario 1 (b) Scenario 2 (c) Scenario 3

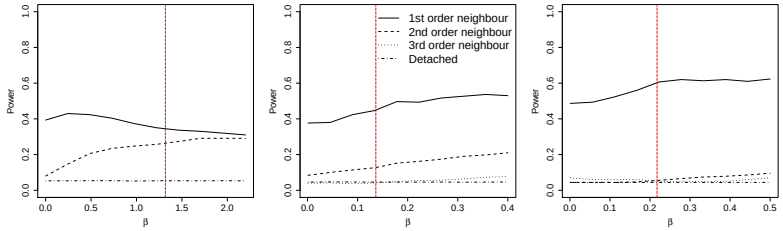
Figure 7: Estimated power for different values of the smoothing parameter β for the three scenarios of the simulated network. The vertical red line indicates β_{opt}, the optimal β according to the network correlation criterium.

neighbours. As the power is growing rapidly with increasing β, the average β_{opt} is quite low for scenario 1. A higher β gives a higher power for the most important nodes, but may deteriorate the correlation between nodes in other communities in the network. The benefit of smoothing for the 3rd order neighbours is also largest in this scenario, but the degree of smoothing required to detect these nodes is above β_{opt}.

In scenario 3, the pheno node was connected to only one node, meaning that there is only one important gene in a subnetwork of less important genes. As can be seen in the plot, the 1st order node quickly looses its power. The 2nd and 3rd order neighbours only have a small benefit of the smoothing; even though the correlations between 1st and 2nd order neighbours is quite large, there is only one 1st order neighbour to borrow power from. However, this scenario has the highest value of β_{opt}. It can be interpreted as there is more to gain from increasing the power for the 2nd and 3rd order neighbours, than keeping the power for the one 1st order neighbour.

Higher values of β were tried for scenario 1 and 2, resulting in a decrease in power for both the 1st and 2nd order neighbours (results not shown). The detached nodes have a power of approximately 0.05 for all networks, indicating that the method controls the type I error satisfactorily.

Figure 8 shows the power for the data simulated from real networks. In Energy metabolism we can observe the same behaviour in the 1st order neighbours as for scenario 3; they are losing power when β is increasing. These nodes have to "give up" some of their power to the 2nd order neighbours, which in their turn gain power by the smoothing. Like scenario 3, Energy metabolism has the highest β_{opt} of the three real networks.



(a) Energy metabolism (b) Lipid metabolism (c) Glycan biosynthesis metabolism

Figure 8: Estimated power for different values of the smoothing parameter β for the data simulated from real networks. The vertical red line indicates β_{opt}, the optimal value of β according to the network correlation criterium.

Lipid metabolism and Glycan biosynthesis metabolism have approximately the same benefit of smoothing for the 1st order neighbours. The increase in power for the 2nd order neighbours is minimal for Glycan biosynthesis metabolism. This network has rather low correlations between 1st and 2nd and within 2nd order neighbours. The smoothing seems to have minimal effect on the 3rd order neighbours in all networks (Energy metabolism has no 3rd order neighbours). The level of the detached nodes is satisfactory in all networks. In summary, the effect of smoothing seems to be smallest on Glycan biosynthesis metabolism, which is the most loosely connected of the three real networks. Energy metabolism and Lipid metabolism are both tightly connected networks, and have more benefit from the smoothing.

3.2 Stress response in *E. faecalis*

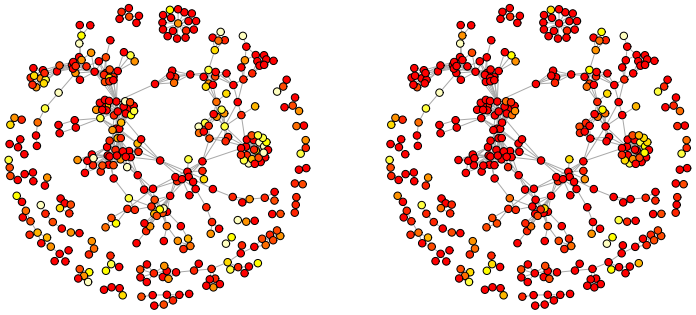
The data from the *E. faecalis* experiment were analysed as described in section 2.6. The number of significant genes (adjusted p-value ≤ 0.05) before and after smoothing are given in Table 2. The network smoothing resulted in more significant genes at all time points.

Table 2: Number of significant genes (adjusted $p \leq 0.05$) in the *E. faecalis* data set before and after smoothing.

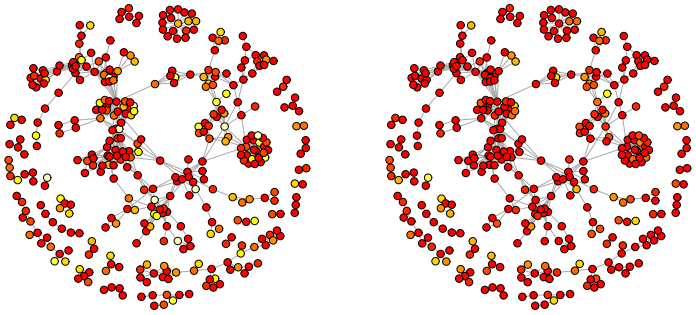
	Time 10	Time 20	Time 60	Time 120
Before smoothing	172	284	51	104
After smoothing	276	394	109	262

DOI: 10.2202/1544-6115.1618

In Figures 9 and 10, the nodes are coloured by the magnitude of their adjusted p-value (divided into ten equally sized intervals between 0 and 1) before and after smoothing with $\beta = 0.1$. For better visualisation, nodes that were not in the expression data and nodes without any edges (which are not affected by the smooth-



Time 10



Time 20

Figure 9: The *E. faecalis* network at time points 10 and 20 before (left) and after (right) smoothing with $\beta = 0.1$. The nodes are coloured by their adjusted p-value. Nodes not in the expression data and nodes without any edges are not shown.

ing), are not included in the plot. Individual nodes with high adjusted p-values surrounded by nodes with small adjusted p-values become more significant after smoothing, while highly significant nodes become less significant after smoothing if they are surrounded by nodes with high adjusted p-values. In other words, the power is more evenly distributed over the subnetwork after smoothing.

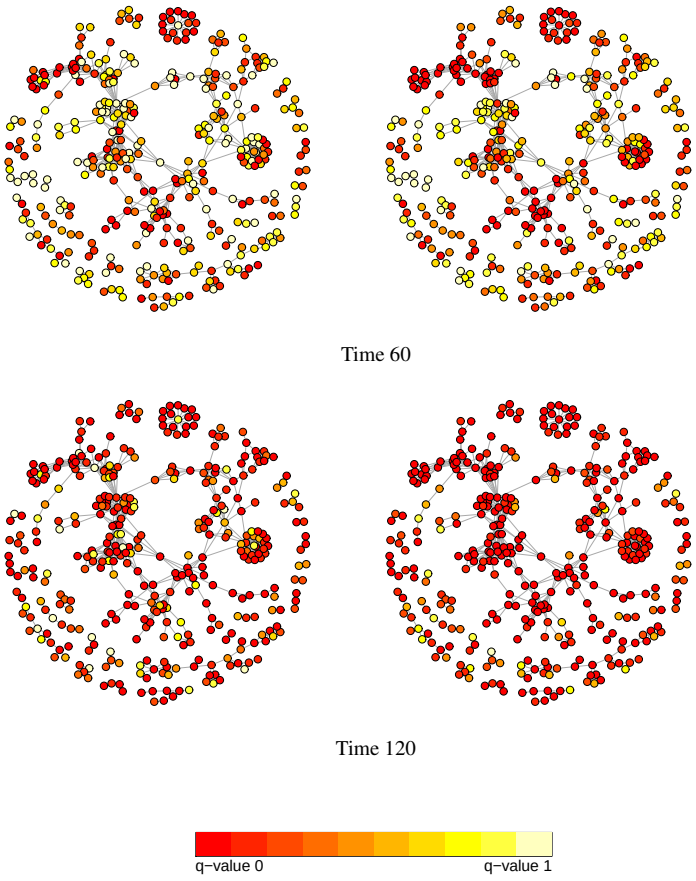
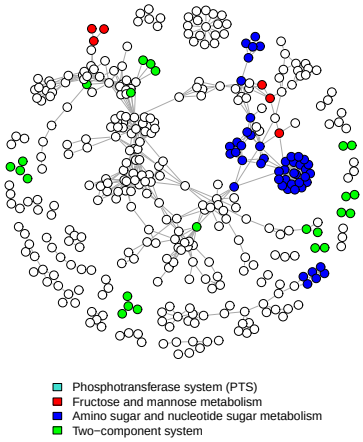
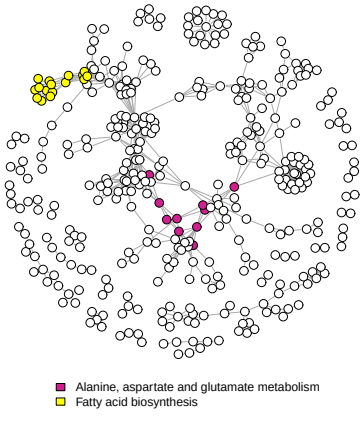


Figure 10: The *E. faecalis* network at time points 60 and 120 before (left) and after (right) smoothing with $\beta = 0.1$. The nodes are coloured by their q-value. Nodes not in the expression data and nodes without any edges are not shown.

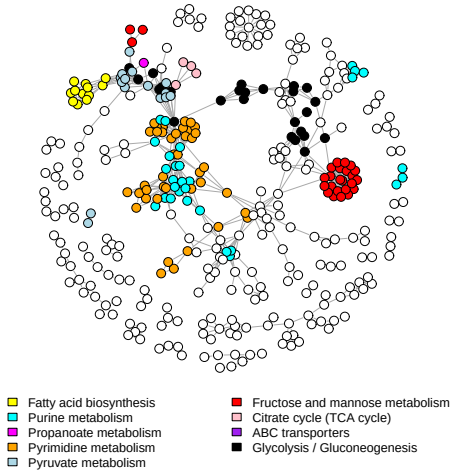
DOI: 10.2202/1544-6115.1618



(a) Time 20

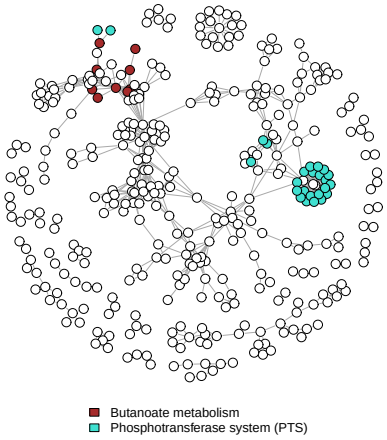


(b) Time 120

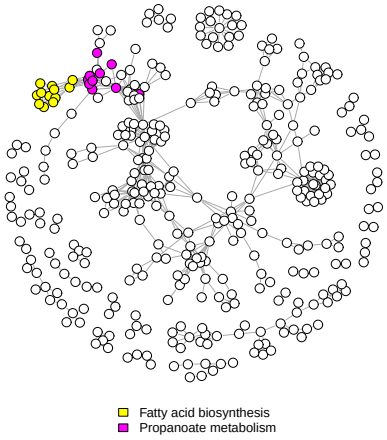


(c) Time 60

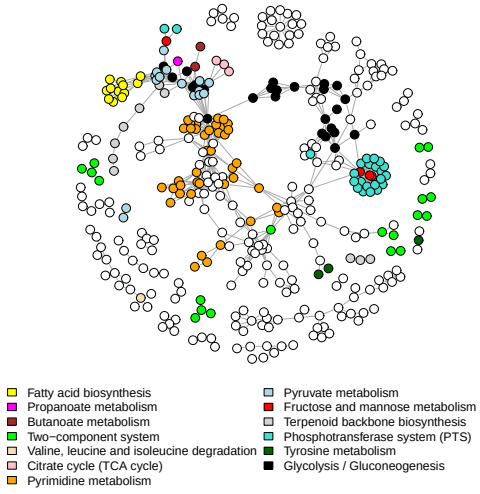
Figure 11: Significant pathways ($q \leq 0.05$) with GSEA on non-smoothed *E. faecalis* data. Note that some pathways are completely overlapping with other pathways and cannot be seen in the plot.



(a) Time 20



(b) Time 120



(c) Time 60

Figure 12: Significant pathways ($q \leq 0.05$) with GSEA on network smoothed *E. faecalis* data. Note that some pathways are completely overlapping with other pathways and cannot be seen in the plot.

DOI: 10.2202/1544-6115.1618

Another interesting observation is that before smoothing the significant pathways differ between the time points, while after smoothing the significant pathways seem to be more consistent over time. All the significant pathways at time 20 and 120 are also significant at time 60, which is not the case before smoothing. The significant pathways (q-value≤ 0.05) from GSEA before and after smoothing are coloured into the graphs in Figures 11 and 12, respectively. There were no significant gene sets at time 10. Interestingly, time 60 is the time point with the least significant genes both before and after smoothing, yet this is the time point with most significant gene sets in GSEA.

4 Discussion

In this paper we have presented a method for smoothing gene expression data with *a priori* network information. The aim of the smoothing was to remove false positives and identify parts of the network where genes are moderately, but coordinately expressed. A similar smoothing approach was used by Rapaport et al. (2007), but with the aim of classifying samples rather than identifying differentially expressed genes. Rapaport et al. considered the smoothing procedure as a spectral decomposition of the graph, where the eigenvectors with large eigenvalues represent the high-frequency noise components and the eigenvectors with small eigenvalues are smoother functions containing the biologically important information. They used two different smoothing methods, one that attenuated eigenvectors with large eigenvalues, and one that completely filtered out the eigenvectors with the largest eigenvalues. The first method is similar to our approach of smoothing with the diffusion matrix, with the degree of attenuation being adjusted by the parameter β. Rapaport et al. showed that filtering out approximately 80 % of the eigenvectors with the largest eigenvalues gave a small improvement in the classification, while attenuating these eigenvectors gave no improvement, at least for large values of β. However, they used values up to $\beta = 50$, while we chose to restrict the value of β to diffusion matrices in which a node puts most weight on itself. This resulted in β's in the range of $0.1 - 2.2$ for the networks used in this paper.

Another important distinction between our method and the approach of Rapaport et al. is that they are smoothing raw data rather than absolute values of the test statistics. Because our network information did not include information about the direction of regulation, we chose to smooth absolute values. If not, we would risk that neighbouring nodes with opposite signs cancel each other out. Absolute test statistics were also used by Saxena et al. (2006) for gene set enrichment analysis in order to identify gene sets with bi-directional changes, and were mentioned

by Efron and Tibshirani (2007) as an option for genewise statistics for gene set testing. A drawback of using absolute values is the risk of accentuating less important subnetworks. Provided that we knew the directions of regulation in the network, we could have kept the signs of the test statistics. An example of such a network is operons, which are sets of genes located adjacently in a bacterial genome and controlled by a common regulatory sequence. Operons appear to be strong indicators of co-regulated genes, and the regulation between all nodes is positive. If a gene in an operon has a test statistic with opposite sign from the rest of the operon, it should be canceled out by the smoothing.

When the type of network is chosen, there are several ways that the network topology could be translated into gene dependencies in the smoothing matrix. In this paper we have only considered diffusion, but other topology descriptors that with some small modifications could act as smoothing matrices are e.g. a matrix of shortest distances between genes (shortest path), or the modularity matrix which was used in the criterion for finding the optimal smoothing.

The simulation study showed that the network smoothing improved the ability to identify nodes associated with the phenotype. The result of the smoothing is that genes with a strong signal share some of their power with their closest neighbours. With an improved ability to identify nodes with a weaker signal comes the risk of losing some important individual nodes. At the same time, this could be thought of as a way of removing false positives; a single important node in a subnetwork of unimportant nodes may be an indication of a false positive gene. The choice of the smoothing parameter β is thus a trade-off between detecting the most important genes, and detecting larger groups of somewhat less important genes with the chance of losing individual genes.

In this paper we used the network correlation as optimality criterion for setting the level of smoothing. Other choices of loss functions may, of course, replace this. There is a vast literature on the problem of finding optimal values for regularization parameters in smoothing problems, for example in spatial statistics, epidemiology and image restoration (Chan and Kay, 1990, Chan and Gray, 1996, e.g.). Some of the suggested criteria are based on generalized cross-validation GCV (Golub et al., 1979) and variants thereof, which probably could be adapted to network smoothing. These will however not take the community structure into account. In this respect, we believe the network correlation criterion used here is a more appropriate measure for the problems discussed in this paper. A thorough comparison of various criteria would be interesting to investigate in future work.

When a gene set method like GSEA is combined with the network smoothing, as was done on the *E. faecalis* data, the smoothing should accentuate important subnetworks that can then be picked up by GSEA. A benefit of using a gene set method after smoothing is that you can interpret results on a gene set level rather

DOI: 10.2202/1544-6115.1618

than on an individual gene level. The analysis of the smoothed *E. faecalis* data with GSEA revealed most significant gene sets at the time point that also had the fewest significant individual genes. The reason may be that GSEA is a type of *comparative* gene set test, meaning that it tries to identify gene sets that stand out from the other sets, compared to a *self-contained* gene set test that looks at each gene set separately (Goeman and Bühlmann, 2007). If the data contains a large number of significant genes, it may be difficult for a gene set to assert itself, as shown in a simulation study by Efron and Tibshirani (2007). See Goeman and Bühlmann, Efron and Tibshirani and Tian et al. (2005) for an in-depth discussion about the different types of gene set tests and the null hypotheses they are testing.

The analysis of the *E. faecalis* data revealed an increased conformity in the significant pathways across the time course after smoothing (Figure 11-12), i.e. all the pathways identified as significant at either time 20 or time 120 were also significant at time 60 in the smoothed data set. In the non-smoothed data set on the other hand, markedly less overlap were observed between the time points. This could imply that the network smoothing produces more robust results. Among the recurring pathways was ”Fatty acid biosynthesis” encoding genes involved in type II fatty acid biosynthesis (FAB). The genes associated with this pathway have previously been linked to bile-induced cell envelope modifications in *E. faecalis* and other Gram-positive bacteria (Taranto et al., 2003, 2006, Le Breton et al., 2002). A considerable overlap is seen between the ”Fatty acid biosynthesis” and the ”Propanoate metabolism” pathways, both of which were identified as significant at time 60 and time 120 after smoothing. In addition, the latter pathway also includes genes coding for two lactate dehydrogenases. Mehmeti and co-workers (Mehmeti et al., 2011) recently identified putative Rex binding sites upstream of both ldh-1, fabI and fabF. The Rex transcriptional regulator has been recognised both as a repressor and an activator, and is responsive to the level of NADH in the cell. Interestingly, putative Rex boxes were also identified upstream of several of the members of the ”Butanoate metabolism” pathway (significant at time 20 and time 60 after smoothing). Overall, these observations may thus be indicative of an effect of bile on genes involved in pyruvate metabolism, which may be related to a shift in the *E. faecalis* NADH/NAD ratio. A potential effect of bile on the metabolic state of the cell in *E. faecalis* was further supported by the significant pathways ”Phosphotransferase system (PTS)” and ”Fructose and mannose metabolism”.

Unfortunately, there are still limited amounts of network information available, and often one can only find information for parts of the genes in the expression data. For the *E. faecalis* data we chose to exclude genes without network information from the analysis, but this is not a requisite for the network smoothing, as these genes could simply be treated as detached nodes and would be unaffected by the smoothing. In comparison, gene set methods such as GSEA exclude genes that are

not member of any gene set from the analysis, and it is also common practice to leave out small gene sets (e.g. less than five genes).

References

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

Chan, K. and A. Gray (1996): "Robustness of automated data choices of smoothing parameter in image regularization," *Statistics and Computing*, 6, 367–377.

Chan, K.-S. and J. Kay (1990): "Smoothing parameter selection in image restoration," in G. G. Roussas, ed., *Nonparametric functional estimation and related topics*, Proceedings of the NATO Advanced Study Institute.

Chung, F. (1997): "Spectral graph theory," *No. 92 in Regional Conference Series in Mathematics. American Mathematical Society*.

Diggle, P., K.-Y. Liang, and S. Zeger (1994): *Analysis of Longitudinal Data*, Oxford University Press.

Dørum, G., L. Snipen, M. Solheim, and S. Sæbø (2009): "Rotation Testing in Gene Set Enrichment Analysis for Small Direct Comparison Experiments," *Statistical Applications in Genetics and Molecular Biology*, 8.

Efron, B. and R. Tibshirani (2007): "On testing the significance of sets of genes," *Annals of Applied Statistics*, 1, 107–129.

Efron, B., R. Tibshirani, J. Storey, and V. Tusher (2001): "Empirical Bayes analysis of a microarray experiment," *Journal of the American Statistical Association*, 96, 1151–1160.

Goeman, J. and P. Bühlmann (2007): "Analyzing gene expression data in terms of gene sets: methodological issues," *Bioinformatics*, 23, 980–987.

Golub, G., M. Heath, and G. Wahba (1979): "Generalized cross-validation as a method for choosing a good ridge parameter," *Technometrics*, 21, 215–223.

Hanisch, D., A. Zien, R. Zimmer, and T. Lengauer (2002): "Co-clustering of biological networks and gene expression data," *Bioinformatics*, 18, 145–154.

Huang, D., B. Sherman, and R. Lempicki (2009): "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic Acids Research*, 37, 1–13.

Jiang, Z. and R. Gentleman (2007): "Extensions to gene set enrichment," *Bioinformatics*, 23, 306–313.

Kanehisa, M. and S. Goto (2000): "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, 28, 27–30.

DOI: 10.2202/1544-6115.1618

Kim, S. and D. Volsky (2005): "PAGE: Parametric analysis of gene set enrichment," *BMC Bioinformatics*, 6.

Kondor, R. and J. Lafferty (2002): "Diffusion kernels on graphs and other discrete input spaces," in *Proceedings of the Nineteenth International Conference on Machine Learning*.

Langsrud, O. (2005): "Rotation tests," *Statistics and Computing*, 15, 53–60.

Le Breton, Y., A. Maze, A. Hartke, S. Lemarinier, Y. Auffray, and A. Rince (2002): "Isolation and characterization of bile salts-sensitive mutants of Enterococcus faecalis," *Current Microbiology*, 45, 434–439.

Li, C. and H. Li (2008): "Network-constrained regularization and variable selection for analysis of genomic data," *Bioinformatics*, 24, 1175–1182.

Mehmeti, I., M. Jönsson, E. Fergestad, G. Mathiesen, I. Nes, and H. Holo (2011): "Transcriptome, Proteome, and Metabolite Analyses of a Lactate Dehydrogenase-Negative Mutant of Enterococcus faecalis V583," *Appl Environ Microbiol.*, 77, 2406–2413.

Newman, M. E. J. (2002): "Assortative mixing in networks," *Phys. Rev. Lett.*, 89, 208701.

Newman, M. E. J. (2003): "Mixing patterns in networks," *Phys. Rev. E*, 67, 026126.

Newman, M. E. J. (2006): "Finding community structure in networks using the eigenvectors of matrices," *Physical Review E*, 74.

Pan, W., B. Xie, and X. Shen (2010): "Incorporating Predictor Network in Penalized Regression with Application to Microarray Data," *Biometrics*, 66, 474–484.

Rahnenführer, J., F. Domingues, J. Maydt, and T. Lengauer (2004): "Calculating the statistical significance of changes in pathway activity from gene expression data," *Statisitcal Applications in Genetics and Molecular Biology*, 3.

Rapaport, F., A. Zinovyev, M. Dutreix, and J. Barillot, E Vert (2007): "Classification of microarray data using gene networks," *BMC Bioinformatics*, 8.

Sæbø, S., T. Almøy, A. Flatberg, A. Aastveit, and H. Martens (2008): "LPLS-regression: a method for prediction and classification under the influence of background information on predictor variables," *Chemometrics and Intelligent Laboratory Systems*, 91, 121–132.

Saxena, V., D. Orgill, and I. Kohane (2006): "Absolute enrichment: gene set enrichment analysis for homeostatic systems," *Nucleic Acids Research*, 34.

Shojaie, A. and G. Michailidis (2009): "Analysis of Gene Sets Based on the Underlying Regulatory Network," *Journal of Computational Biology*, 16, 407–426.

Shojaie, A. and G. Michailidis (2010): "Network Enrichment Analysis in Complex Experiments," *Statistical Applications in Genetics and Molecular Biology*, 9.

Smyth, G. and T. Speed (2003): "Normalization of cDNA microarray data," *Methods*, 31, 265–273.

Solheim, M., A. Aakra, H. Vebo, L. Snipen, and I. Nes (2007): "Transcriptional responses of Enterococcus faecalis V583 to bovine bile and sodium dodecyl sulfate," *Applied and Environmental Microbiology*, 73, 5767–5774.

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

Subramanian, A., P. Tamayo, V. Mootha, S. Mukherjee, B. Ebert, M. Gillette, A. Paulovich, S. Pomeroy, T. Golub, E. Lander, and J. Mesirov (2005): "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *PNAS*, 102, 15545–15550.

Taranto, M., M. Murga, G. Lorca, and G. de Valdez (2003): "Bile salts and cholesterol induce changes in the lipid cell membrane of Lactobacillus reuteri," *Journal of Applied Microbiology*, 95, 86–91.

Taranto, M., G. Perez-Martinez, and G. de Valdez (2006): "Effect of bile acid on the cell membrane functionality of lactic acid bacteria for oral administration," *Research in Microbiology*, 157, 720–725.

Tian, L., S. Greenberg, S. Kong, J. Altschuler, I. Kohane, and P. Park (2005): "Discovering statistically significant pathways in expression profiling studies," *PNAS*, 102, 13544–13549.

Vert, J. and M. Kanehisa (2003): "Extracting active pathways from gene expression data," *Bioinformatics*, 19, II238–II244.

Wu, D., E. Lim, F. Vaillant, M.-L. Asselin-Labat, J. Visvader, and G. Smyth (2010): "ROAST: rotation gene set tests for complex microarray experiments," *Bioinformatics*, 26, 2176–2182.

Zhang, J. and S. Wiemann (2009): "KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor," *Bioinformatics*, 25, 1470–1471.

DOI: 10.2202/1544-6115.1618

Paper III

Rotation gene set testing for longitudinal expression data

Guro Dørum, Lars Snipen, Margrete Solheim, and Solve Sæbø

Department of Chemistry, Biotechnology and Food Science
Norwegian University of Life Sciences, N-1432 Aas, Norway

Abstract

Gene set analysis methods are popular tools for identifying differentially expressed gene sets in microarray data. Most existing methods use a permutation test to assess significance for each gene set. The permutation test's assumption of independent samples is often not satisfied for time series data and complex experimental designs, and in addition it requires a certain number of samples to compute p-values accurately. The method presented here uses a rotation test rather than a permutation test to assess significance. The rotation test can compute accurate p-values also for very small sample sizes. The method can handle complex designs and is particularly suited for longitudinal microarray data where the samples may have complex correlation structures. In addition, the method can test for both gene sets that are differentially expressed and gene sets that show strong time trends. We show on simulated longitudinal data that the ability to identify important gene sets increases when we take the correlation structure between samples into account, and applied to real data the method identifies both gene sets containing differentially expressed genes and gene sets containing genes with strong time trends.

1 Introduction

In the recent years gene set analysis methods have become increasingly popular as a tool for analysing microarray gene expression data. By shifting the focus from individual genes to sets of genes with some biological relation, the results of the analysis may be easier to interpret and more reproducible. Genes in the same gene set are thought to have a similar expression pattern and are defined *a priori*. Examples include metabolic pathways and functional categories. Gene set methods are testing whether a set of genes is enriched, and assigns p-values to each gene set. In addition to easier interpretation of the results, testing gene sets rather than individual genes reduces the number of tests and hence the number of false positives.

Goeman and Bühlmann (2007) introduced the terms *competitive* and *self-contained* gene set tests to differentiate between tests that aim at identifying gene sets that stand out from a collection of gene sets, and tests that look at each gene set by itself. One of the most popular competitive tests is the Gene Set Enrichment Analysis (GSEA), introduced by Mootha et al. in 2003 and modified by Subramanian et al. in 2005. Competitive gene set testing have also been done by e.g. Efron and Tibshirani (2007) and Dørum et al. (2009). Examples of self-contained testing can be found in Goeman et al. (2004), Tian et al. (2005), Jiang and Gentleman (2007) and Wu et al. (2010). See Huang et al. (2009) for a recent review of gene set methods.

Common for gene set methods is that they assign a p-value to each gene set, usually calculated with a permutation test. Correct permutation requires independent samples, which is often not the case in complex microarray designs. In addition, permutation tests require a moderate to large number of samples to calculate accurate p-values. Permutation of genes rather than samples have been shown to seriously underestimate p-values due to the incorrect assumption of independence between genes (Efron and Tibshirani, 2007, Dørum et al., 2009). Dørum et al. introduced GSEA with rotation testing as an alternative to permutation testing. The rotation test can calculate accurate p-values even for small sample sizes. Wu et al. (2010) introduced rotation testing for gene set tests that can also handle complex experimental designs. Their method ROAST can apply different types of gene set statistics. We are developing gene set analysis with rotation testing further by allowing testing of several contrasts simultaneously and by considering complex correlation structures between samples.

In particular, we have in mind gene expression time series data. Microarray experiments performed as time series capture a dynamic picture of gene expres-

sion rather than a snapshot provided by a static experiment. Many authors have dealt with the problem of identifying differentially expressed genes in time series data. Storey et al. (2005) modelled the time course for each gene by splines. Park et al. (2003) fitted ANOVA models and tested interactions between time and treatment. Zhou et al. (2010) estimated the direction in time space with the strongest ANOVA signal of interest. Ma et al. (2009) modelled time course data with functional ANOVA mixed-effect models and tested the interaction between treatment and time. All the above mentioned methods, however, evaluates differential expression on the individual gene level, while we are interested in identifying sets of genes that are differentially expressed. Wei and Li (2008) incorporated gene network information in Markov random field models to identify differentially expressed pathways in time course data.

Time series data can be divided into cross-sectional data, in which each individual is measured only once, and longitudinal data for which individuals are measured repeatedly over time. In this paper we will focus on longitudinal data. The repeated sampling on individuals means that the samples tend to be correlated. In addition, there may be random design factors introducing correlation between samples. The method presented here allows for complex covariance structures for the samples by considering the special correlation structure introduced by time course measurements. Since the method can test several contrasts simultaneously, gene sets can be tested for both differential expression and time trends.

We introduce a gene set analysis method that can handle almost all types of experimental designs, complex covariance structures between samples and a small number of samples, while testing several contrasts simultaneously. The method's power and control of the type I error is demonstrated in a simulation study. A data set investigating stress responses in *E. faecalis* over time is analysed with the aim of identifying both gene sets that are differentially expressed and exhibit a time trend.

2 Model assumptions

2.1 General linear model with correlated errors

Let y_j be a vector of gene expression values for the j'th gene, where $j = 1, 2, ..., g$. For each gene we assume the general model

$$y_j = X\beta_j + e_j \tag{1}$$

where X is a $n \times p$ matrix with n samples and p design factors, β_j is a vector of p parameters, and $e_j \sim N_n(0,V)$. Note that we assume a common covariance structure for all genes.

The rotation test described in section 3 is based on the assumption that the observations are independent. As mentioned by Wu et al. (2010), known dependencies may be removed up front of the rotation testing. In practice these dependencies will be unknown, but by assuming that all genes have the same covariance matrix V, a good estimate of V can be obtained if the number of genes is large. The estimate $\hat{V}$ may be used in a preprocessing step to reduce between-sample dependencies before the rotation testing. Next we describe a method for obtaining a $\hat{V}$ that accounts for complex dependency structures in the data.

2.2 Covariance structure for longitudinal data

In this section we describe a covariance structure, presented in Diggle et al. (1994), for experimental designs with random effects and where measurements are done on the same subjects over time (longitudinal data). We assume that the random variation in the data is due to three different sources of variation. The first source is random design factors that affect all measurements done on a subject, and we denote this variance σ_b^2. For example, if RNA is sampled from different batches, batch may be regarded as a random factor. The second source of variation is serial correlation between measurements done on the same subject at different time points. This correlation is typically weaker for time points that are far apart. We denote this variance σ_t^2. The third source of variation is a random measurement error for measurements done at the same time on the same subject, denoted σ_e^2. The covariance matrix $\mathbf{V}$ is composed of these three sources of variation as

$$\mathbf{V} = \sigma_b^2\mathbf{J} + \sigma_t^2\mathbf{H} + \sigma_e^2\mathbf{I} \tag{2}$$

where $\mathbf{J}$ is a $n \times n$ matrix with ones in positions corresponding to samples with the same level of the random factor, $\mathbf{H}$ is block diagonal with all elements within the same subject specified by a correlation function $\rho(u)$ of the time interval u between the samples, and $\mathbf{I}$ is the identity matrix. An example of a correlation function is the exponential correlation model

$$\rho(u) = e^{-\phi u} \tag{3}$$

for some value of $\phi > 0$.

V is usually unknown and must be estimated. Since we assume an identical covariance structure for all genes, we can use all g genes to compute a restricted maximum likelihood (REML) estimate (see e.g. Diggle et al., 1994). With the reparametrisation $V = \sigma_t^2 W(\alpha)$, the three parameters to estimate are $\alpha_1 = \sigma_b/\sigma_t$, $\alpha_2 = \sigma_e/\sigma_t$ and $\alpha_3 = \phi$. Assuming independence between genes, the restricted log-likelihood function for $\alpha = [\alpha_1, \alpha_2, \alpha_3]$ is

$$\ell^*(\alpha) = -\frac{1}{2}\left[g(n-p)log(\sigma_t^2) + g \cdot log|W(\alpha)| + \frac{1}{\sigma_t^2}\sum_{j=1}^{g} RSS_j(\alpha) + g \cdot log|X^T W^{-1} X|\right]$$

where $RSS_j(\alpha) = (\mathbf{y}_j - \mathbf{X}\hat{\beta}_j)^T \mathbf{W}^{-1}(\mathbf{y}_j - \mathbf{X}\hat{\beta}_j)$. For a given start value of α, $\hat{\beta}_\mathbf{j}$ can be found as

$$\hat{\beta}_j = (\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{-1}\mathbf{y}_j$$

This estimate is inserted into the restricted log-likelihood function, which can then be optimised with respect to α. σ_t^2 is estimated as $\hat{\sigma}_t^2 = \sum_{j=1}^{g} RSS_j(\hat{\alpha})/[g(n-p)]$.

3 Rotation test

3.1 Data preprocessing

The first step in the analysis is to reduce between-sample correlations by multiplying each term of eq (1) with the inverse square root matrix of the estimated covariance matrix $\hat{\mathbf{V}}$

$$\hat{V}^{-1/2} y_j = \hat{V}^{-1/2} X\beta_j + \hat{V}^{-1/2} e_j \tag{4}$$

The model with the transformed data is

$$y_j^* = X^*\beta_j + e_j^* \tag{5}$$

where $e_j^* \sim N_n(0, I)$. For simplicity we will omit the '*' notation, however it is the transformed data that are used in the following, unless otherwise stated.

3.2 Remove nuisance parameters and obtain independent residuals

The next step is to remove the design factors that we are not interested in testing. A common way of handling unwanted design effects is to use an ANOVA type

normalisation by fitting a model with the unwanted effects and use the residuals from this model in the further analysis (Kerr et al., 2000, Dørum et al., 2009). However, these residuals are not independent, and the rotation test (like the permutation test) assumes independence between samples. We therefore employ a method outlined by Langsrud (2005) that removes the nuisance parameters but also obtains independent residuals.

The $n \times p$ design matrix X can be decomposed by a QR decomposition

$$X = X_Q U$$

where X_Q is an orthonormal basis for the column space of X. In order to make X_Q a full $n \times n$ matrix we can add $n - p$ extra orthonormal columns, and at the same time add the same number of rows containing only zeros in the upper triangular matrix U. This corresponds to including some extra non-observed variables without any impact on the response y_j. We now assume that X has been arranged such that all nuisance effects are in the first $p - k$ columns, while the interesting variables are in the last k columns. The interesting variables may be e.g. contrasts for testing linear and quadratic time effects (as for the simulated data in section 4.1). X_Q can be split into

$$X_Q = [X_Q^N, X_Q^B, X_Q^E]$$

where X_Q^N is a $n \times (p-k)$ orthonormal matrix spanning the nuisance space, X_Q^B is a $n \times k$ orthonormal matrix spanning the space of effects we want to test, and X_Q^E are the added extra orthonormal columns. In a similar fashion we can split the $p \times 1$ parameter vector β_j into

$$\beta_j = \begin{bmatrix} \beta_j^N \\ \beta_j^B \end{bmatrix}$$

where β_j^N are the $p - k$ nuisance parameters and β_j^B are the k parameters of interest.

With the QR decomposition of X we can rewrite the model in (5) as

$$y_j = \mathbf{X}_Q \gamma_j + e_j$$

where $\gamma_j = U\beta_j$. The $n \times 1$ vector γ_j contains linear combinations of the param-

eters and has the following structure

$$\gamma_j = \begin{bmatrix} u_{11}\beta_{j1} + u_{12}\beta_{j2} + \cdots + u_{1p}\beta_{jp} \\ u_{22}\beta_{j2} + \cdots + u_{2p}\beta_{jp} \\ \vdots \\ u_{n-p,p}\beta_{jp} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

where the u's denote elements in U. We can split γ_j into

$$\gamma_j = \begin{bmatrix} \gamma_j^N \\ \gamma_j^B \\ \gamma_j^E \end{bmatrix}$$

where γ_j^N contains linear combinations of the $p-k$ nuisance parameters, γ_j^B contains linear combinations of the k interesting parameters, and γ_j^E is the expected effect of the $n-p$ extra non-important variables. Testing $H_0 : \beta_j^B = 0$, i.e. that all parameters of interest are 0, is equivalent to testing $H_0 : \gamma_j^B = 0$. The last element of γ_j^B is $u_{n-p,p}\beta_{jp}$, and for this to be zero, β_{jp} must be zero. The second last element of γ_j^B is $u_{n-p-1,p-1}\beta_{j,p-1} + u_{n-p,p}\beta_{jp}$, and for this to be zero, $\beta_{j,p-1}$ must also be 0, and so on. We estimate γ_j by

$$\hat{\gamma}_j = (X_Q^T X_Q)^{-1} X_Q^T y_j = X_Q^T y_j$$

The observations in y_j are thus projected onto the orthonormal basis spanned by X_Q to separate the various effects. $\hat{\gamma}_j^N$ corresponds to observations along nuisance directions and can be ignored in the subsequent analysis, $\hat{\gamma}_j^B$ corresponds to the estimated effects of interest, and $\hat{\gamma}_j^E$ corresponds to so-called error observations.

3.3 Genewise statistics

The $n-p$ error observations in $\hat{\gamma}_j^E$ can be used to compute the sample variance for gene j. The computations can be done separately for each gene or by some approach introducing smoothing across genes (Wright and Simon, 2003, Smyth,

2004). Let s_j^2 denote the variance estimate for the j'th gene. A t-value for gene j and variable l, where $l = 1, ..., k$, can then be computed as

$$t_{jl} = \frac{\hat{\gamma}_{jl}^B}{s_j} \tag{6}$$

where $\hat{\gamma}_{jl}^B$ is the estimated variable. A genewise F statistic, involving all k variables of interest, can be calculated as

$$F_j = \frac{t_{j1}^2 + t_{j2}^2 + ... + t_{jk}^2}{k} \tag{7}$$

3.4 Gene set statistics

The genewise F statistics are the basis for gene set statistics. In this paper we are using GSEA as the gene set score, but in general any gene set statistic could be used. GSEA starts by ranking the genes based on the F statistics. The rank positions of all members of a gene set are identified before an enrichment score is calculated for each gene set. The enrichment score is normalised to account for gene set size. GSEA makes no inferences on the gene level, so all genes are included in the analysis. This makes GSEA sensitive to smaller changes that are consistent within the gene set. We refer to Subramanian et al. (2005) on how to calculate enrichment scores.

3.5 Assigning significance

In the original GSEA procedure a permutation test is used to generate a null distribution for each gene set, but we have replaced it with a rotation test (Langsrud, 2005). Rotation testing for GSEA was first introduced by Dørum et al. (2009). A rotation test can rotate the genes in all directions and can therefore generate accurate p-values also for small sample sizes. The rotations are conditioned on the covariance matrix so the gene-gene correlations are maintained during the rotation. Permutations can be viewed as rotations restricted to exchange measurement axes. The rotation test assumes multivariate normal distribution for each sample, which in this setting means that the genes on an array have a multivariate normal distribution. However, Dørum et al. and Wu et al. (2010) showed on simulated non-normal data that the rotation test is robust with regard to violations against the assumption of multinormality.

Let

$$\hat{\gamma}_j^{BE} = \begin{bmatrix} \hat{\gamma}_j^B \\ \hat{\gamma}_j^E \end{bmatrix}$$

be the subvector of non-nuisance effects and residuals for the j'th gene. A sample from the null distribution is found by rotation of this $(n-p+k)\times 1$ vector. Let R^* be a valid $(n-p+k)\times(n-p+k)$ rotation matrix, e.g. the Q matrix from a QR decomposition of a $(n-p+k)\times(n-p+k)$ matrix of random standard normal data. Then

$$\hat{\gamma}_j^{BE*} = R^*\hat{\gamma}_j^{BE} \tag{8}$$

is a rotated version, and test-statistics can be computed from the $\hat{\gamma}_j^{B*}$ part of this vector, just as described above. By repeating this procedure for a large number of random rotation matrices, we obtain a sample of test statistics from the null distribution. A p-value is calculated for each gene set as the proportion of this distribution at least as extreme as the observed gene set score.

4 Data

4.1 Simulated longitudinal data

Log-ratios from an experiment comparing two treatments/phenotypes were simulated. Assuming that RNA was extracted from two different batches, batch was included as a random factor with variance σ_b^2. Dye was included as a fixed factor with two levels. Dye and batch are nuisance factors that we want to remove the effect of. Arrays with the same combination of batch and dye followed over time were considered a subject, and each subject was measured at four time points: 1, 2, 6 and 12. Variance due to sampling the same subject over time was denoted σ_t^2. The covariance structure for the samples also included a random error variance σ_e^2 for measurements done on the same subject at the same time. The design matrix included two contrasts for testing linear and quadratic time effects, in addition to dye and an intercept column. Since these data are log-ratios, testing the intercept corresponds to testing for differential expression.

Let n be the number of samples per gene. With four subjects (2 batches $\times$ 2 dyes) measured at four time points there are $n = 16$ samples. The 16×1 expression vector for gene j have the following structure

$$y_j = \left[{y_{j11}}^T, {y_{j12}}^T, {y_{j21}}^T, {y_{j22}}^T\right]^T$$

Each 4×1 vector y_{jbd} contains the measurements for one subject, where $b = 1,2$ denotes batch and $d = 1,2$ denotes dye. Applying the covariance structure described in section 2.2, the 16×16 covariance matrix is the block diagonal

$$V = \begin{bmatrix} R1 & R2 & 0 & 0 \\ R2 & R1 & 0 & 0 \\ 0 & 0 & R1 & R2 \\ 0 & 0 & R2 & R1 \end{bmatrix} \tag{9}$$

The 4×4 matrix $R1$ describes covariance between measurements from the same subject and is composed as

$$R1 = \begin{bmatrix} \sigma_b^2 + \sigma_t^2 + \sigma_e^2 & \sigma_b^2 + \sigma_t^2\rho_{12} & \sigma_b^2 + \sigma_t^2\rho_{13} & \sigma_b^2 + \sigma_t^2\rho_{14} \\ \sigma_b^2 + \sigma_t^2\rho_{12} & \sigma_b^2 + \sigma_t^2 + \sigma_e^2 & \sigma_b^2 + \sigma_t^2\rho_{23} & \sigma_b^2 + \sigma_t^2\rho_{24} \\ \sigma_b^2 + \sigma_t^2\rho_{13} & \sigma_b^2 + \sigma_t^2\rho_{23} & \sigma_b^2 + \sigma_t^2 + \sigma_e^2 & \sigma_b^2 + \sigma_t^2\rho_{34} \\ \sigma_b^2 + \sigma_t^2\rho_{14} & \sigma_b^2 + \sigma_t^2\rho_{24} & \sigma_b^2 + \sigma_t^2\rho_{34} & \sigma_b^2 + \sigma_t^2 + \sigma_e^2 \end{bmatrix}$$

where ρ_{ij} is the correlation between time points i and j. The 4×4 matrix $R2$ describes covariance between subjects from the same batch and has σ_b^2 in all entries. We assume independence between subjects from different batches.

The variance components were set to $\sigma_b^2 = 1$, $\sigma_t^2 = 2$ and $\sigma_e^2 = 3$. For correlation between time points we used the exponential correlation model in eq (3) with $\phi = 1.5$ (corresponding to a correlation of approximately 0.22 between the first two time points). Log-ratios were simulated for $g = 1000$ genes divided into 50 sets of size 20. For gene j a $n \times 1$ vector z_j of random standard normal data was multiplied with the square root matrix of V to get the desired covariance structure

$$y_j = V^{1/2} z_j$$

4.1.1 Estimation of covariance matrix

To examine the accuracy of the REML estimates of the variance components in V for different sample sizes, data were simulated for scenarios with 16, 32 and 64 samples per gene. For the last two scenarios we have 8 and 16 subjects, respectively, where the extra subjects are replicates of the original four subjects. We still assume time dependence only between measurements from the same subject, so covariance between replicates are described by the $R2$ matrix. A dye effect of $\beta_D = 1$ was added to all genes, while a gene effect of $\beta_G = 2$, a linear time effect of $\beta_L = 2$ and a quadratic time effect of $\beta_Q = 1.5$ were added to all genes in the

first two gene sets. One gene set was given only positive effects, while the other was given only negative effects. 500 data sets with each sample size were simulated, and for each data set the variance components were estimated with REML as described in section 2.2. Random uniformly distributed values between 0 and 10 were chosen as start values for α in each simulation.

4.1.2 Effect of including covariance matrix

Next we wanted to explore the effect of incorporating the estimated covariance matrix $\hat{V}$ in the rotation test procedure. Taking into account dependency between samples should increase the power of the method, as long as the assumed covariance structure is reasonable. 500 data sets with a sample size of $n = 16$ were simulated. A dye effect of $\beta_D = 1$ was added to all genes, while linear time effects $\beta_L = \{1, 2, 3\}$ and quadratic time effects $\beta_Q = \{0.5, 1, 1.5\}$ were added to all genes in the first two sets. One gene set was given positive linear and quadratic time effects, while the other was given negative linear and quadratic time effects. All combinations of β_L and β_Q were tried.

The data sets were analysed with the method outlined in section 3, first without the data transformation in section 3.1 and then with the transformation. Two contrasts were tested, one for linear and one for quadratic time effect. Power was computed as the proportion of the 500 data sets in which the two important gene sets were found to be significant.

4.2 Stress response in *E. faecalis*

This microarray experiment was performed in order to investigate the transcriptional response of the bacterium *Enterococcus faecalis* V583 to bile stress. RNA was extracted from bacteria treated with bile and from untreated cultures as a reference, and reverse transcribed. Labelled cDNA was then hybridised to mutual slides in a direct design experiment. RNA was obtained in two separate growth experiments, here referred to as batches. Samples were collected from each batch after 10, 20, 60 and 120 minutes. At each time point four arrays were used ($2 \times$ dye-swap). Arrays with the same combination of batch and dye are considered the same subject, so this is a case of longitudinal data. There were four subjects and a total of 16 arrays. For details about labelling, hybridisations and data preprocessing, see Solheim et al. (2007).

The differential expression between treated and untreated bacteria was measured as log_2(signal treated) - log_2(signal untreated). Loess normalisation imple-

mented in the LIMMA package for R (Smyth and Speed, 2003) was used to correct for intensity dependent trends in the data. A gene was required to be present on at least 3 out of 4 arrays at each time point to be included in the analysis. The number of genes meeting this requirement was 2350. Missing data were imputed by k-nearest neighbours imputation (Troyanskaya et al., 2001) implemented in R.

The design of this experiment is the same as for the simulated data in section 4.1, except that at time 60 three of four samples were dyed with cy3, so the data were analysed unbalanced. In addition to testing for linear and quadratic time effects, we also wanted to test the intercept (i.e. differential expression between treated and untreated bacteria).

Four different types of gene sets were tested in GSEA: functional categories, KEGG pathways, genes classified by first EC number and operons. Gene sets were required to have at least 5 members to be included in the analysis, which resulted in a total of 132 gene sets: 19 functional categories, 59 pathways, 6 EC groups and 48 operons.

Modified t-values (Wright and Simon, 2003, Smyth, 2004) computed for each gene and each contrast were composed to a genewise F-value. In the computation of enrichment scores we chose not to use weighting of the ranked F statistics (i.e. the parameter p in Subramanian et al. (2005) was set to 0). The reason for avoiding weighting is that genes with very high F-values may be too influential on the gene set score if weighting is used. To correct for multiple hypothesis testing, a false discovery rate q-value (Storey, 2002) was calculated for each gene set. Since we are using F statistics we are only interested in the gene sets with a large positive enrichment score, so we used a one-sided version of the approach for computing q-values in Subramanian et al..

In addition to analysing the data as time series, the four time points were analysed as separate experiments. Since in this case we are only interested in testing for differential expression, we used t statistics rather than F statistics as the basis for ranking genes in the computation of enrichment scores. The signs of the t statistics and the enrichment scores in this case give information about the direction of regulation, which we do not get when working with F statistics. In this analysis both large positive and negative enrichment scores are of interest, so the q-values were computed with the approach in Subramanian et al.. In the following, the two analyses will be referred to as the *time series analysis* and the *individual time point analysis*.

5 Results

5.1 Simulated longitudinal data

5.1.1 Estimation of covariance matrix

Table 1 shows the REML estimates for the variance components in V with different sample sizes, averaged over 500 simulations. The estimated batch variance σ_b^2 is very accurate, while the random error variance σ_e^2 is mildly underestimated. The time correlation parameter ϕ is slightly overestimated, meaning that the time correlation between samples is underestimated. However, considering that the estimated time variance σ_t^2 is a little high, the estimated time dependency seems quite decent. The covariance structure for the simulated data may give highly correlated parameter estimates. If the time correlations are weak and the noise variance is small compared to the time variance component, the two parameters σ_t^2 and σ_e^2 are close to being non-identifiable. This was the case for these particular simulations where the estimate of σ_t^2 on average was too small and the estimate of σ_e^2 too big, but the sum of the parameters was correctly estimated with an average of 5.00, and the correlations between the estimates across the 500 simulations was -0.9998. We therefore conclude that the REML-estimation procedure produces good estimates of V to be used in the preprocessing of the data. However, as illustrated in this simulation, the estimates of the individual variance components may be biased and highly correlated if the time dependence is weak. Therefore, if one suspects low time dependence in addition to the fixed linear and quadratic time trends, a simpler covariance structure without time correlations should be chosen.

Table 1: Average REML estimates based on 500 simulations for the components in V, where n is the sample size. The standard deviations (SD) of the 500 estimates are also given.

		$\sigma_b^2 = 1$	$\sigma_t^2 = 2$	$\sigma_e^2 = 3$	$\phi = 1.5$
$n = 16$	Estimate	0.99	2.16	2.84	1.76
	SD	0.08	1.59	1.60	2.10
$n = 32$	Estimate	1.00	2.22	2.78	1.60
	SD	0.06	1.49	1.49	1.55
$n = 64$	Estimate	1.00	2.24	2.76	1.58
	SD	0.05	1.37	1.37	1.33

5.1.2 Effect of including covariance matrix

Two analyses were performed on the simulated data, one where the data were transformed with $\hat{V}$ (as described in section 3.1) and one where this preprocessing step was ignored. This was done in order to see the effect of incorporating covariance between samples on the method's power, i.e. its ability to identify the gene sets with added time effects. Figure 1 shows the estimated power in the two analyses with increasing linear time effect β_L, when the added quadratic time effect was $\beta_Q = 1$. Including the covariance structure in the analysis has increased the method's ability to identify the important gene sets. The same effect from inclusion of the covariance matrix could be seen for the other levels of β_Q, only with lower power for both analyses when $\beta_Q = 0.5$ and higher power when $\beta_Q = 1.5$.

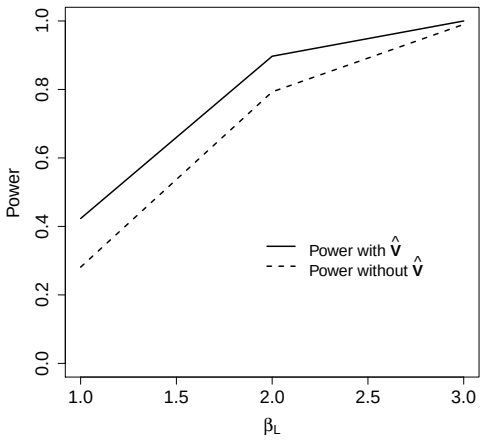


Figure 1: Estimated power with increasing linear time effect (β_L) and quadratic time effect $\beta_Q = 1$, with and without incorporation of the estimated covariance matrix $\hat{\mathbf{V}}$.

5.2 Stress response in *E. faecalis*

In the time series analysis of the *E. faecalis* data we were testing the null hypothesis for each gene set that it was differentially expressed and/or showed a linear

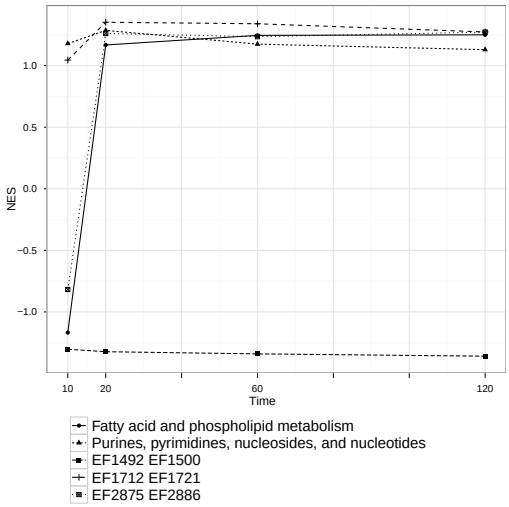
Table 2: Significant gene sets ($q \leq 0.25$) from the time series analysis of the *E. faecalis* data when testing for differential expression and time trend.

Gene set	q	Size*
Functional categories		
Fatty acid and phospholipid metabolism	0.172	33
Purines, pyrimidines, nucleosides, and nucleotides	0.220	56
Pathways		
Pentose phosphate pathway	0.026	18
Fatty acid biosynthesis	0.054	12
Pyrimidine metabolism	0.113	41
Glycine, serine and threonine metabolism	0.132	17
EC		
Transferases	0.132	131
Operons		
EF1712 EF1721	0.033	9
EF2875 EF2886	0.044	11
EF1492 EF1500	0.117	9

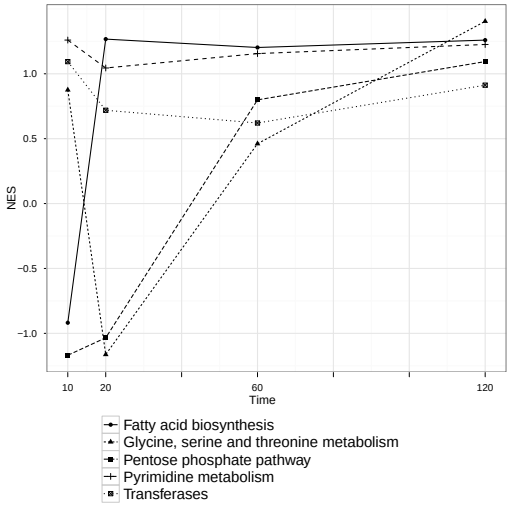
*Genes in the set corresponding to genes in the expression data

and/or quadratic time trend. The significant gene sets (q-value ≤ 0.25) are shown in Table 2. To get a picture of the time trends for the significant gene sets, the normalised enrichment scores from the individual time point analysis were plotted against time. Note that these enrichment scores only reflect differential expression, while the enrichment scores in the time series analysis also reflect time effects. However, the signs of these enrichment scores give information about up- and downregulation. Figure 2 shows the time trends for the significant gene sets in the time series analysis. Since we were testing three contrasts simultaneously, differential expression, linear time effect and quadratic time effect, there may be different causes for a gene set's high score. These time trend plots indicate that both gene sets that are highly differentially expressed over time and gene sets that have a strong time trend are identified by the method.

To investigate whether the gene set scores have successfully captured the important trends for the members of a set, the time trends for the individual genes in



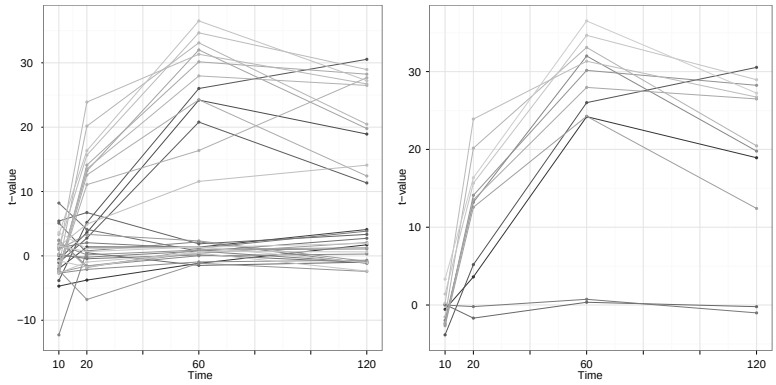
(a) Functional categories and operons



(b) Pathways and EC groups

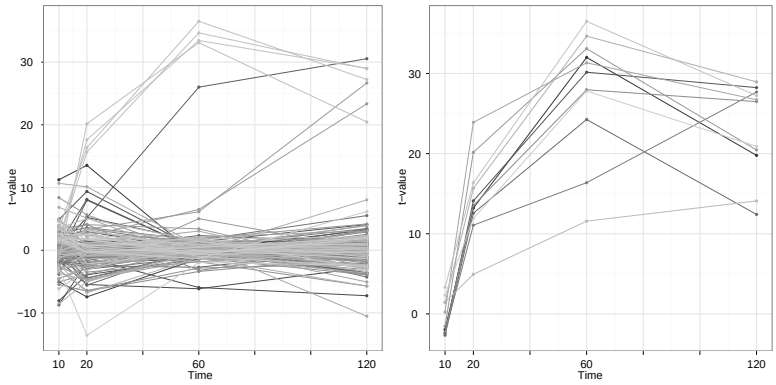
Figure 2: Normalised enrichment score (NES) for each time point separately for significant gene sets in the time series analysis of the *E. faecalis* data.

a few chosen gene sets are plotted in Figure 3. In these plots the modified t-values reflecting differential expression from the individual time point analysis are plotted against time. The y-axis is not directly comparable with the y-axis in Figure



(a) Fatty acid and phospholipid metabolism

(b) Fatty acid biosynthesis



(c) Transferases

(d) EF2875 EF2886

Figure 3: Time trends for all genes in four of the significant gene sets from the time series analysis of the *E. faecalis* data. Each time point shows the genes' modified t-values.

2, nevertheless it is possible to spot a general trend among the genes that seems to correspond with the gene set trend. The genes in the functional category ”Fatty acid and phospholipid metabolism” seem to cluster into two groups, possibly reflecting that this gene set is composed of several groups of genes, where one (or more) of these groups displays a different time trend than the remainder. This notion is supported by the plot of the pathway ”Fatty acid biosynthesis”, where the genes show a similar trend as the group of genes that stand out in the functional category. Moreover, the majority of remaining genes of the functional gene set belongs to the pathway ”Glycerophospholipid metabolism”, for which no significant differential expression/time trend was detected. It thus seems to be mainly type II fatty acid biosynthesis (type II FAS) genes influencing the gene set score for ”Fatty acid and phospholipid metabolism”. Of note, the type II FAS genes also reappear in the operon ”EF2875 EF2886”. The EC group ”Transferases” is a very large gene set, which makes it harder to spot a general trend among the genes in this set.

6 Discussion

Rotation testing for GSEA was first introduced by Dørum et al. (2009). By using a rotation test one can avoid the problem with granularity of p-values that the permutation test suffers from when the sample size is small. GSEA is an example of competetive gene set testing, where the aim is to identify gene sets that stand out from a collection of gene sets. Wu et al. (2010) presented ROAST, which is rotation testing for self-contained gene set testing, where each gene set is tested separately without considering the remaining gene sets. ROAST can handle most experimental designs and correlation between samples. In this paper we have further developed the contributions of ROAST with a method that allows for more complex covariance structures between samples, and that can test more than one contrast simultaneously. These improvements are especially useful for longitudinal data analysis, because these types of data may introduce more complex correlations between samples, and the focus of attention may be gene sets that show both a time trend and differential expression. For the gene set statistic we chose to use the GSEA enrichment score, but in principle any test statistic could be used.

Because we assumed a common covariance structure for all genes, we had a large number of observations to estimate covariance between samples from. An alternative could be to divide genes into non-overlapping groups, either based

on some pre-clustering of the genes or in light of prior biological knowledge, and assume a common covariance structure only for genes in the same group. This would of course result in more parameters to estimate, but the assumption of a common covariance structure could be more correct within these groups. In the REML function we also assumed independence between genes. This is an assumption that conflicts with the gene set methods' hypothesis of correlation between genes in a set. Wu et al. also assumed independence between genes when estimating correlation between samples, but rather than assuming a structure for the correlation, they estimated each element of the covariance matrix directly. Considering gene-gene correlations in the estimation of sample correlations would be an interesting topic for further improvement of the method. Another issue to be further explored is the method's robustness against incorrect assumptions about the correlation structure. In the simulation study we used the same correlation structure for generation and analysis of the data.

The rotation test's assumption of multinormal distribution for each sample is a strong assumption, but as previously argued it has been shown to be robust with regards to deviations from normality by Dørum et al. (2009) and Wu et al. (2010). However, Dørum et al. also mentioned a data set with stronger deviations from normality where the rotation test showed somewhat lower power than the permutation test. To further confirm the robustness of the rotation test, more tests on time series data sets with stronger deviations from normality could be carried out.

A permutation test can be seen as a rotation test that is restricted to exchange measurement axes. Both the permutation test and the rotation test assumes independence between the samples to be permuted/rotated. The procedure for obtaining independent residuals in the linear model in section 3.2 could also be used in relation to a permutation test. The rotation matrix in eq (8) could in principle be replaced by a permutation matrix if the number of samples is large enough.

In the simulation study we observed that the type I error was not controlled properly on a gene level when the estimated covariance matrix $\hat{\mathbf{V}}$ was not included in the analysis (results not shown). A significance level of 0.05 for genewise p-values gave an actual rejection level of over 0.07 for the non-important genes (as an estimate of the type I error). On a gene set level the type I error was controlled for both the analysis with and without inclusion of $\hat{\mathbf{V}}$, however the rejection level for the non-important gene sets was slightly higher when $\hat{\mathbf{V}}$ was not included. This may imply that the actual gain in power from including $\hat{\mathbf{V}}$ is even larger than Figure 1 shows.

The significant gene sets from the analysis of the *E. faecalis* data included

the functional category "Fatty acid and phospholipid metabolism", the pathway "Fatty acid biosynthesis" and the operon "EF2875 EF2886", all encoding genes involved in type II fatty acid biosynthesis. Le Breton et al. (2002) previously identified an *E. faecalis* bile-sensitive mutant corresponding to a gene involved in fatty acid biosynthesis by random gene disruption strategies. Moreover, bile exposure has been reported to trigger changes in the membrane fatty acid composition, and a decrease in membrane fluidity and in the protein:phospholipid ratio in other bacteria (Ruiz et al., 2007, Kimoto-Nira et al., 2009, Taranto et al., 2003). The putative roles of these genes in *E. faecalis* bile response may thus be related to bile-induced modifications in cell membrane properties, and have been thoroughly covered by ourselves and others (Le Breton et al., 2002, Dørum et al., 2009, Solheim et al., 2007, Vebo et al., 2009). Among the significant gene sets was also the pathway "Pentose phosphate pathway". In addition to generating energy through fermentation of sugars, this multifunctional pathway generates reducing equivalents in the form of NADPH, which can be used in reductive biosynthesis reactions within the bacterium, including fatty acid biosynthesis (Huycke, 2002). Moreover, the pentose phosphate pathway also channels pentoses into nucleotide biosynthesis (Huycke, 2002). A potential role of nucleotide biosynthesis in *E. faecalis* bile response was further reflected by the significant functional category "Purine, pyrimidines, nucleosides and nucleotides" and the pathway "Pyrimidine metabolism".

References

P. J. Diggle, K.-Y. Liang, and S. L. Zeger. *Analysis of Longitudinal Data*. Oxford University Press, 1994.

G. Dørum, L. Snipen, M. Solheim, and S. Sæbø. Rotation Testing in Gene Set Enrichment Analysis for Small Direct Comparison Experiments. *Statistical Applications on Genetics and Molecular Biologi*, 8(1), 2009. ISSN 1544-6115.

B. Efron and R. Tibshirani. On testing the significance of sets of genes. *Annals of Applied Statistics*, 1(1):107–129, 2007. ISSN 1932-6157.

J. Goeman, S. van de Geer, F. de Kort, and H. van Houwelingen. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99, 2004. ISSN 1367-4803.

J. J. Goeman and P. Bühlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, 2007. ISSN 1367-4803.

D. W. Huang, B. T. Sherman, and R. A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13, 2009. ISSN 0305-1048.

M. M. Huycke. *Physiology of enterococci. In: The enterococci. Pathogenesis, molecular biology and antibiotic resistance*. ASM Press, Washington D. C., 2002.

Z. Jiang and R. Gentleman. Extensions to gene set enrichment. *Bioinformatics*, 23(3):306–313, 2007. ISSN 1367-4803.

M. Kerr, M. Martin, and G. Churchill. Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, 7(6):819–837, 2000. ISSN 1066-5277.

H. Kimoto-Nira, M. Kobayashi, M. Nomura, K. Sasaki, and C. Suzuki. Bile resistance in Lactococcus lactis strains varies with cellular fatty acid composition: Analysis by using different growth media. *International Journal of Food Microbiology*, 131(2-3):183–188, 2009. ISSN 0168-1605.

O. Langsrud. Rotation tests. *Statistics and Computing*, 15(1):53–60, 2005. ISSN 0960-3174.

Y. Le Breton, A. Maze, A. Hartke, S. Lemarinier, Y. Auffray, and A. Rince. Isolation and characterization of bile salts-sensitive mutants of Enterococcus faecalis. *Current Microbiology*, 45(6):434–439, 2002. ISSN 0343-8651.

P. Ma, W. Zhong, and J. Liu. Identifying differentially expressed genes in time course microarray data. *Statistics in Biosciences*, 1:144–159, 2009. ISSN 1867-1764.

V. Mootha, C. Lindgren, K. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstrale, E. Laurila, N. Houstis, M. Daly, N. Patterson, J. Mesirov, T. Golub, P. Tamayo, B. Spiegelman, E. Lander, J. Hirschhorn, D. Altshuler, and L. Groop. PGC-1 alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34(3):267–273, 2003. ISSN 1061-4036.

T. Park, S. Yi, S. Lee, S. Lee, D. Yoo, J. Ahn, and Y. Lee. Statistical tests for identifying differentially expressed genes in time-course microarray experiments. *Bioinformatics*, 19(6):694–703, 2003. ISSN 1367-4803.

L. Ruiz, B. Sanchez, P. Ruas-Madiedo, C. G. de los Reyes-Gavilan, and A. Margolles. Cell envelope changes in Bifidobacterium animalis ssp lactis as a response to bile. *FEMS Microbiology Letters*, 274(2):316–322, 2007. ISSN 0378-1097.

G. Smyth and T. Speed. Normalization of cDNA microarray data. *Methods*, 31 (4):265–273, 2003. ISSN 1046-2023.

G. K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3:Iss. 1, Article 3, 2004.

M. Solheim, A. Aakra, H. Vebo, L. Snipen, and I. F. Nes. Transcriptional responses of Enterococcus faecalis V583 to bovine bile and sodium dodecyl sulfate. *Applied and Environmental Microbiology*, 73(18):5767–5774, 2007. ISSN 0099-2240.

J. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B - Statistical Methodology*, 64(Part 3):479–498, 2002. ISSN 1369-7412.

J. Storey, W. Xiao, J. Leek, R. Tompkins, and R. Davis. Significance analysis of time course microarray experiments. *PNAS*, 102(36):12837–12842, 2005. ISSN 0027-8424.

A. Subramanian, P. Tamayo, V. Mootha, S. Mukherjee, B. Ebert, M. Gillette, A. Paulovich, S. Pomeroy, T. Golub, E. Lander, and J. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 102(43):15545–15550, 2005. ISSN 0027-8424.

M. Taranto, M. Murga, G. Lorca, and G. de Valdez. Bile salts and cholesterol induce changes in the lipid cell membrane of Lactobacillus reuteri. *Journal of Applied Microbiology*, 95(1):86–91, 2003. ISSN 1364-5072.

L. Tian, S. Greenberg, S. Kong, J. Altschuler, I. Kohane, and P. Park. Discovering statistically significant pathways in expression profiling studies. *PNAS*, 102 (38):13544–13549, 2005. ISSN 0027-8424.

O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–525, 2001. ISSN 1367-4803.

H. C. Vebo, L. Snipen, I. F. Nes, and D. A. Brede. The Transcriptome of the Nosocomial Pathogen Enterococcus faecalis V583 Reveals Adaptive Responses to Growth in Blood. *PLoS ONE*, 4(11), 2009. ISSN 1932-6203.

Z. Wei and H. Li. A hidden spatial-temporal markov random field model for network-based analysis of time course gene expression data. *Annals of Applied Statistics*, 2(1):408–429, 2008. ISSN 1932-6157.

G. Wright and R. Simon. A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics*, 19(18): 2448–2455, 2003. ISSN 1367-4803.

D. Wu, E. Lim, F. Vaillant, M.-L. Asselin-Labat, J. E. Visvader, and G. K. Smyth. ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics*, 26(17):2176–2182, 2010. ISSN 1367-4803.

B. Zhou, W. Xu, D. Herndon, R. Tompkins, R. Davis, W. Xiao, W. H. Wong, and Inflammation Host Response Injury. Analysis of factorial time-course microarrays with application to a clinical study of burn injury. *PNAS*, 107(22): 9923–9928, 2010. ISSN 0027-8424.

Paper IV

Improved preprocessing for rotation gene set testing for longitudinal expression data

Guro Dørum and Solve Sæbø

Department of Chemistry, Biotechnology and Food Science
Norwegian University of Life Sciences, N-1432 Aas, Norway

Abstract

Gene set tests are used in the analysis of gene expression data to identify predefined sets of genes that are differentially expressed. Most existing methods use a permutation test to calculate a p-value for each gene set. Permutation tests are limited to data sets with relatively large sample sizes and independent samples, which is often not found in microarray data. In a previous paper we introduced a gene set test that uses a rotation test rather than a permutation test to compute p-values. The gene set rotation test can handle small sample sizes and complex experimental designs. It is particularly suited for longitudinal data and can identify both gene sets with differential expression and time trends. In order to reduce correlations between samples, the data are preprocessed with an estimated covariance matrix taking random design factors and time correlations into account. Here we propose an improved preprocessing for the gene set rotation test that also considers gene correlations in the estimation of covariances between samples. We demonstrate on simulated data that the improved preprocessing method increases the gene set rotation test's power in identifying gene sets with time trends. The method is also illustrated by application to a real data set investigating stress responses in *E. faecalis*.

1 Introduction

Gene set tests for the analysis of differentially expressed genes have gained increased popularity over the last few years, especially after the introduction of the Gene Set Enrichment Analysis (Mootha et al., 2003, Subramanian et al., 2005). Rather than searching for differentially expressed individual genes, gene set tests aim at identifying sets of genes that are co-expressed. The gene sets are defined prior to the analysis based on biological knowledge. Examples of gene sets include pathways, functional categories and GO categories. Genes that belong to the same set are believed to have correlated expression patterns. Gene set tests can have more statistical power than individual gene tests because they accumulate signals from a number of genes. This also increases the sensitivity towards sets of genes with moderate, but consistent changes in gene expression. Various gene set tests have been introduced by i.e. Goeman et al. (2004), Tian et al. (2005), Efron and Tibshirani (2007), Goeman and Bühlmann (2007), Jiang and Gentleman (2007) and Wu et al. (2010). See Huang et al. (2009) for a recent review.

A p-value for each gene set is usually computed with a permutation test permuting samples. A permutation test is based on the assumption that the samples are independent and identically distributed, and requires a certain number of samples to estimate accurate p-values. This severely limits the types of data sets that can be analysed with gene set methods. Small sample sizes are common in microarray experiments, and depending on the design of the experiment, the samples may not be regarded as exchangeable. Some gene set tests permute genes rather than samples. Since the number of genes is usually very large, the problem with small sample sizes is avoided. However, when we permute genes we implicitly assume that the genes are independent, an assumption that contradicts the fundamental idea behind gene set tests. The genes are grouped because they are believed to have correlated expression profiles. Permutation of genes has been demonstrated by Efron and Tibshirani (2007) and Dørum et al. (2009) to produce severely overoptimistic p-values.

Dørum et al. (2009) introduced rotation testing as an alternative to permutation testing for gene set tests. In contrast to permutation tests, rotation tests (Langsrud, 2005) has no limitations to the maximum number of rotations and can therefore compute accurate p-values also for small sample sizes. Wu et al. (2010) introduced a gene set test with rotation testing that can also handle data from complex experimental designs. The data is represented by a linear model, and the rotations can be done in a subspace orthogonal to the nuisance factors. Correlation between

samples due to random design factors was handled by estimating an empirical covariance matrix as described in Smyth et al. (2005).

In Dørum et al. (Submitted) we presented a gene set test with rotation testing for complex designs that is particularly suited for longitudinal gene expression data. In longitudinal data, individuals have been measured repeatedly over time, and are thus expected to have considerable correlation between samples. Rather than estimating a completely unstructured covariance matrix, as in Wu et al., we assumed a structure for the covariances incorporating dependencies both due to time and random design factors. The components of the covariance matrix were estimated by restricted maximum likelihood, assuming an identical covariance structure for all genes and independence between genes. The estimated covariance matrix was included in a preprocessing step before the gene set rotation test to reduce correlations between samples. The gene set test is also particularly suited for longitudinal data by offering the opportunity to test for several properties in gene sets simultaneously, e.g. differential expression and time effects.

Both Wu et al. and Dørum et al. assumed an identical covariance structure for all genes and independence between genes when estimating the covariance matrix. The assumption of independent genes is very doubtful considering the gene set tests' idea about correlated expression patterns within gene sets. In this paper we present an extension to the method in Dørum et al.. We improve the preprocessing step by considering gene dependencies in the estimation of the covariance matrix. Gene dependencies are modeled with the use of network distances from an *a priori* defined gene network. Rather than estimating dependencies between all genes, the genes are divided into non-overlapping groups and dependencies are assumed only within groups. We refer to Dørum et al. for a detailed description of the gene set rotation test, as this paper focuses only on the preprocessing to remove correlations between samples.

2 Methods

2.1 Gene network

We use gene networks both to divide genes into non-overlapping groups and to model gene dependencies. This section gives a brief introduction to gene networks and some measures for describing the topology of a network.

A gene network can be constructed based on prior biological information such as pathways or bacterial operons. The network can be represented as a graph

where each node corresponds to a gene and an edge between two nodes represents some relationship between the genes. In networks based on pathways two genes are connected if they take part in successive reactions in the cell, while in networks based on operons two genes are connected if they are located next to each other on the bacterial chromosome and are regulated by a common transcription mechanism. Let G be a graph with g nodes (genes). Further let i and j represent two nodes in G, and let $i \sim j$ indicate that the two nodes are adjacent (directly connected). The $g \times g$ adjacency matrix $\mathbf{A}$ describes the nodes' neighbourhoods, and the entries a_{ij} are

$$a_{ij} = \begin{cases} 1 & \text{if } i \sim j \\ 0 & \text{else} \end{cases} \tag{1}$$

for $i, j = 1, 2, ..., g$. The $g \times g$ degree matrix $\mathbf{D}$ is a diagonal matrix with the degree, i.e. the number of edges to each node, on the main diagonal. Let δ_i be the degree of node i. The entries in $\mathbf{D}$ are

$$d_{ij} = \begin{cases} \delta_i & \text{if } i = j \\ 0 & \text{else} \end{cases} \tag{2}$$

The $g \times g$ Laplacian matrix (Chung, 1997) is defined as $\mathbf{L} = \mathbf{D} - \mathbf{A}$, where the entries are

$$l_{ij} = \begin{cases} -1 & \text{if } i \sim j \\ d_{ij} & \text{if } i = j \\ 0 & \text{else} \end{cases} \tag{3}$$

2.1.1 Dividing genes into non-overlapping sets

In the estimation of the covariance matrix, we simplify the task by assuming dependencies between genes in the same group and independence between groups. To make this assumption as valid as possible, the genes should ideally be divided into groups with strong correlations within groups, and weak or no correlations between groups. To achieve this, we represent the genes with a gene network. The network can be split into smaller, non-overlapping subnetworks with a high density of edges within the network, and few or no edges between subnetworks. Newman and Girvan (2004) introduced the concept of modularity as a way to divide the network into subnetworks based on natural communities. Algorithms that split networks based on modularity, identify parts of the network with fewer edges than expected, and divide the network there. The number of communities to split the network into is not given in advance, but determined from the network itself. It may be the case that no good division of the network exists. The modularity

matrix is defined as $\mathbf{M} = \mathbf{A} - \mathbf{P}$, where $\mathbf{A}$ is the adjacency matrix and $\mathbf{P}$ contains the expected number of edges between each pair of nodes. The expected number of edges between nodes i and j if edges are placed at random is

$$p_{ij} = \frac{\delta_i \delta_j}{2m} \tag{4}$$

where δ_i and δ_j are the degrees of the nodes and m is the total number of edges in the network. For a given division of the network into two groups, let $\mathbf{b}$ be a $g \times 1$ vector where $b_i = 1$ if node i belongs to group 1 and $b_i = -1$ if node i belongs to group 2. The modularity can be expressed as

$$Q = \frac{1}{4m} \mathbf{b}^T \mathbf{M} \mathbf{b} \tag{5}$$

The task is to find the value of $\mathbf{b}$ which maximizes Q. Each of the two groups can again be divided into two groups, and so forth until the divisions no longer give a positive contribution to the modularity. Newman and Girvan (2004) and Newman (2006a,b) present a number of algorithms for splitting networks into communities based on the eigenvectors of the modularity matrix. We chose however to use an algorithm described in Clauset et al. (2004) that does not use the eigenvectors. The algorithm starts by treating each node as a separate community, before joining the two communities that produce the largest increase in modularity. This is repeated until all communities are combined. The optimal division can be traced back to

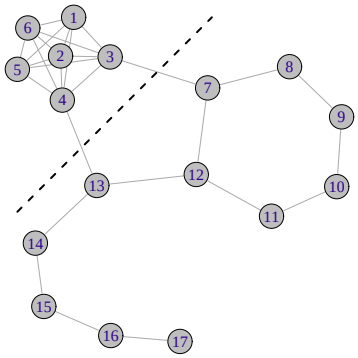


Figure 1: Dividing a network into communities based on modularity. The dashed line indicates the split found with the fastgreedy.community algorithm.

where the largest modularity occurred. The algorithm is implemented in the R function fastgreedy.community in the igraph package (Csardi and Nepusz, 2006), and is very efficient for large networks. In Figure 1 the dashed line indicates the split found by the algorithm in a fictional network. In this case the optimal division was a split into two communities.

2.1.2 Modelling gene dependencies

Gene networks can also be used to estimate dependencies between genes. We assume that the more closely connected two genes are in the network, the more correlated their gene expression should be. One way of measuring distances between genes in a network is to look at the smallest number of edges separating each pair of nodes, i.e. the shortest path between two nodes. In this paper, however, we will focus on the concept of diffusion (Kondor and Lafferty, 2002). Diffusion is closely related to random walks and can be imagined as a fluid travelling through the network. The time that the fluid takes to get from one node to another indicates the distance between the nodes. The diffusion matrix $\mathbf{S}$ is defined by the matrix exponential of the Laplacian matrix $\mathbf{L}$

$$\mathbf{S} = e^{-\beta \mathbf{L}} = \sum_{i=1}^{g} v_i e^{-\beta \lambda_i} v_i^T \tag{6}$$

where v_i and λ_i are the i'th eigenvector and eigenvalue of $\mathbf{L}$, respectively. The parameter β, where $\beta > 0$, controls the speed of diffusion through the graph. The diffusion is faster for larger values of β, corresponding to shorter distances between nodes. We have used $\beta = 1$ for the purpose in this paper. The less fluid that diffuse from one node to another, the further they are apart. We therefore use the inverse diffusion to describe distances. If s is the diffusion between two nodes, then

$$f = 1/s \tag{7}$$

is the distance between the nodes. The correlation between two genes with distance f can be modelled with e.g. the exponential correlation model

$$\gamma(f) = e^{-\gamma f} \tag{8}$$

for $\gamma > 0$. A larger distance in the network will thus give a weaker correlation, and γ controls the magnitude of the correlation.

2.2 Model assumptions

2.2.1 General linear model with correlated errors

Assume that we have $i = 1, ..., m$ subjects measured at n time points for each of $j = 1, ..., g$ genes. A subject in a microarray experiment may refer to arrays with the same levels of experimental factors followed over time. The $n \times 1$ vector y_{ji} contains the measurements over time for subject i and gene j and represents one experimental unit. All experimental units for gene j are arranged in the vector y_j of length $m \cdot n$ with the following structure

$$y_j = \left[y_{j1}, y_{j2}, \ldots, y_{jm}\right]^T \tag{9}$$

We further assume that the g genes are divided into $k = 1, ..., K$ non-overlapping groups of size M_k, e.g. with the network splitting procedure described in section 2.1.1. The observations for all genes in group k can be arranged in a vector of length $N_k = M_k \cdot m \cdot n$ as

$$Y_k = \left[y_1, y_2, \ldots, y_{M_k}\right]^T \tag{10}$$

We assume the following model for group k

$$Y_k = X_k \beta_k + e_k \tag{11}$$

where X_k is an $N_K \times p$ design matrix including design factors plus contrasts of interest, β_k is a vector of p parameters, and $e_k \sim N_{N_k}(0, V_k)$. Note that we assume an identical design matrix X for all genes, so X_k is just X repeated k times. The groups are assumed to be independent.

2.2.2 Covariance structure for longitudinal data

The covariance structure presented here is as given in Diggle et al. (1994). We assume that there are four different sources of variation that account for the total variation between samples. The first source is random design factors, a variance denoted σ_b^2. The second source is serial correlation between samples from different time points measured on the same subject. We expect this correlation to be weaker for time points that are far apart. The time variance component is denoted by σ_t^2. The third source of variation comes from gene dependencies. We also expect this correlation to be weaker for genes that are less related within the gene set, and this variance component is denoted by σ_g^2. The last source of variation is

a random error variance, denoted by σ_e^2. A $N_k \times N_k$ covariance matrix for gene set k including these four sources of variation can be composed as

$$V_k = \sigma_b^2 J_k + \sigma_t^2 H_k + \sigma_g^2 M_k + \sigma_e^2 I_k \tag{12}$$

where J_k is a matrix with ones in positions corresponding to samples with the same level of the random factor, H_k is block diagonal with all elements within the same subject specified by a correlation function $\rho(u)$ of the time interval u between the samples, L_k has all elements specified by the correlation function in eq. (8) and I_k is the identity matrix. We chose to use an exponential correlation model similar to the gene correlation model to describe time correlations. The correlation between two samples measured with time interval u is

$$\rho(u) = e^{-\phi u} \tag{13}$$

for some value of $\phi > 0$. A larger difference in time will lead to weaker correlations between samples.

2.3 Estimating the covariance matrix with REML

V_k is usually unknown and must be estimated, and we use restricted maximum likelihood (REML) for this (see e.g. Diggle). With the reparametrisation $V_k = \sigma_t^2 W_k(\alpha)$, the five parameters to estimate are $\alpha_1 = \sigma_b/\sigma_t$, $\alpha_2 = \sigma_g/\sigma_t$, $\alpha_3 = \sigma_e/\sigma_t$, $\alpha_4 = \phi$ and $\alpha_5 = \gamma$. Assuming independence between gene sets, the restricted log-likelihood function for $\alpha = [\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5]$ is

$$\begin{aligned} \ell^*(\alpha) = -\frac{1}{2}\Bigg[& (N - K \cdot p) log \sigma_t^2 + \sum_{k=1}^{K} log|W_k(\alpha)| + \frac{1}{\sigma_t^2} \sum_{k=1}^{K} RSS_k(\alpha) \\ & + \sum_{k=1}^{K} log|X_k^T W_k(\alpha)^{-1} X_k| \Bigg] \end{aligned} \tag{14}$$

where $N = \sum_{k=1}^{K} N_k$ and $RSS_k(\alpha) = (\mathbf{Y}_k - \mathbf{X}_k \hat{\beta}_k)^T \mathbf{W}_k(\alpha)^{-1} (\mathbf{Y}_k - \mathbf{X}_k \hat{\beta}_k)$. For a given start value of α, $\hat{\beta}_k$ can be found as

$$\hat{\beta}_k = (\mathbf{X}_k^T \mathbf{W}_k(\alpha)^{-1} \mathbf{X}_k)^{-1} \mathbf{X}_k^T \mathbf{W}_k(\alpha)^{-1} \mathbf{Y}_k$$

This estimate is inserted into the restricted log-likelihood function, which can then be optimised with respect to α. Finally, σ_t^2 is estimated as

$$\hat{\sigma}_t^2 = \sum_{k=1}^{K} RSS_k(\hat{\alpha}) / (N - K \cdot p)$$

2.4 Data preprocessing

To reduce between-sample correlations before the rotation test, we multiply each term of eq. (11) with the inverse square root matrix of the estimated covariance matrix $\hat{V}_k$

$$\hat{V}_k^{-1/2} Y_k = \hat{V}_k^{-1/2} X_k \beta_k + \hat{V}_k^{-1/2} e_k \tag{15}$$

The model with the transformed data is

$$Y_k^* = X_k^* \beta_k + e_k^* \tag{16}$$

where $e_k^* \sim N_k(0, I)$. The transformed samples should be approximately independent given that the correct dependence structure is used, and a large number of genes are used in the estimation. The transformed data can be analysed with the gene set rotation test as described in Dørum et al.

3 Data

3.1 Simulated longitudinal data

We simulated gene expression log-ratios from a microarray experiment comparing two phenotypes/treatments. The $g = 1000$ genes were divided into 50 groups of size 20. The genes in a group were assumed to be part of the same operon, meaning that in the network gene 1 is connected to gene 2 which is connected to gene 3, and so on. Note that since we had non-overlapping groups we did not need to use the approach in section 2.1.1 for splitting a network into subnetworks, but the network splitting is used in the real data analysis presented later in this article. Dye was included as a fixed factor with two levels and batch was included as a random factor with two levels. Arrays with the same combination of dye and batch followed over time were considered part of the same subject. There were four subjects measured at the four time points $\{1, 2, 6, 12\}$, resulting in 16 samples per gene. The design matrix included an intercept column (testing the intercept corresponds to testing for differential expression in log-ratios), a dye column and two contrasts for testing linear and quadratic time effects. Covariance between samples was attributed to four sources: batch, time dependency, gene dependency and random error. Gene dependencies were assumed only within groups, time dependencies only within an experimental unit, and batch correlations only between experimental units from the same gene. The covariance matrix for a gene is the

16×16 matrix R with the following structure

$$R = \begin{bmatrix} R_1 & R_2 & R_3 & R_3 \\ R_2 & R_1 & R_3 & R_3 \\ R_3 & R_3 & R_1 & R_2 \\ R_3 & R_3 & R_2 & R_1 \end{bmatrix} \tag{17}$$

The 4×4 matrix R_1 describes the covariance for one experimental unit and is composed as

$$\begin{bmatrix} \sigma_b^2+\sigma_g^2+\sigma_t^2+\sigma_e^2 & \sigma_b^2+\sigma_g^2+\sigma_t^2\rho_{12} & \sigma_b^2+\sigma_g^2+\sigma_t^2\rho_{13} & \sigma_b^2+\sigma_g^2+\sigma_t^2\rho_{14} \\ \sigma_b^2+\sigma_g^2+\sigma_t^2\rho_{12} & \sigma_b^2+\sigma_g^2+\sigma_t^2+\sigma_e^2 & \sigma_b^2+\sigma_g^2+\sigma_t^2\rho_{23} & \sigma_b^2+\sigma_g^2+\sigma_t^2\rho_{24} \\ \sigma_b^2+\sigma_g^2+\sigma_t^2\rho_{13} & \sigma_b^2+\sigma_g^2+\sigma_t^2\rho_{23} & \sigma_b^2+\sigma_g^2+\sigma_t^2+\sigma_e^2 & \sigma_b^2+\sigma_g^2+\sigma_t^2\rho_{34} \\ \sigma_b^2+\sigma_g^2+\sigma_t^2\rho_{14} & \sigma_b^2+\sigma_g^2+\sigma_t^2\rho_{24} & \sigma_b^2+\sigma_g^2+\sigma_t^2\rho_{34} & \sigma_b^2+\sigma_g^2+\sigma_t^2+\sigma_e^2 \end{bmatrix}$$

where ρ_{ij} is the correlation between time points i and j found with eq. (13). The 4×4 matrix R_2 describes covariance between experimental units with the same batch level and has $\sigma_b^2 + \sigma_g^2$ in all entries. The 4×4 matrix R_3 describes covariance between experimental units with different batch levels and has σ_g^2 in all entries. For three genes from the same group, the 48×48 covariance matrix V_k would look like

$$V_k = \begin{bmatrix} R & U_{12} & U_{13} \\ U_{12} & R & U_{23} \\ U_{13} & U_{23} & R \end{bmatrix} \tag{18}$$

The 16×16 matrix U_{12} has all its entries equal to $\sigma_g^2\gamma_{12}$, where γ_{12} is the correlation between gene 1 and 2 found by eq. (8). Likewise, U_{13} has all its entries equal to $\sigma_g^2\gamma_{13}$, where γ_{13} is the correlation between gene 1 and 3, and so on.

In this illustration the variance components were rather arbitrarily set to $\sigma_b^2 = 1$, $\sigma_g^2 = 2$, $\sigma_t^2 = 2$, $\sigma_e^2 = 3$, $\gamma = 0.3$ and $\phi = 0.9$. However, the chosen values for γ and ϕ gave maximum gene and time correlations close to 0.4, which is a reasonable level of time and gene dependence. Log-ratios for the genes of group k were simulated by drawing a $N_k \times 1$ vector z_k of random standard normal data, which was then multiplied with the square root matrix of V_k to get the desired covariance structure

$$Y_k = V_k^{1/2} z_k$$

A dye effect was added to all genes, while a gene effect β_G (corresponding to differential expression between phenotypes), linear time effect β_L and a quadratic time effect β_Q were added to all genes in the first two gene sets. One gene set

was given only positive gene and time effects, while the other was given negative gene and time effects. In the first scenario we added both gene and time effects, in the second scenario we added only time effects and in the third scenario we added only gene effects. The added effects in each scenario are given in table 1.

Table 1: Fixed effects were added in three different scenarios.

	β_D	β_G	β_L	β_Q
Scenario 1	1	2	{0,1,2,3}	0.5
Scenario 2	1	0	{0,1,2,3}	0.5
Scenario 3	1	{1,2,3}	0	0

The components of V were estimated with the REML in eq. (14). We refer to this estimated covariance matrix as $\hat{V}_A$. For comparison, we also estimated a covariance matrix with the estimation method presented in the previous paper, which assumes independence between genes. This covariance matrix is referred to as $\hat{V}_B$. Data preprocessed with each of the two covariance matrices were analysed with the gene set rotation test.

3.2 Stress response in *E. faecalis*

This data set is from an experiment that investigated stress responses in the bacterium *E. faecalis* to bile (Solheim et al., 2007). The expression data for the 2350 genes are log-ratios comparing treated and untreated bacteria. The same data set was analysed in Dørum et al. (2009), so more details about the experimental design can be found there. Samples were collected after 10, 20, 60 and 120 minutes after treatment with bile. The design of the experiment is the same as for the simulated data in the previous section, except for an unbalance at time 60 where three of four samples have the same dye. We also chose to leave out batch from the model in order to avoid over-parametrisation, as previous studies have shown that the batch effect in this data set is very small. The assumed components in the covariance matrix to be estimated were thus σ_t^2, σ_g^2, σ_e^2, ϕ and γ.

We constructed a gene network based on 83 metabolic and non-metabolic pathways from the KEGG PATHWAY database (Kanehisa and Goto, 2000). The pathways were merged to one large network, removing redundant nodes and edges, with the KEGGgraph package in R (Zhang and Wiemann, 2009). The final network consisted of 800 nodes connected by 1306 edges. Of these 800 nodes, 633 corresponded to genes in the data set. These 633 nodes were divided into com-

munities with the R function fastgreedy.community, resulting in 281 communities with sizes ranging from 1 to 57 nodes. The remaining 1717 genes in the data set lacked network information and were regarded as communities of size 1. Gene dependencies were estimated with eq. (8) where network distances were extracted from the diffusion matrix based on all 800 nodes.

Three contrasts were tested, differential expression, linear time effect and quadratic time effect. The 132 gene sets tested included 19 functional categories, 59 pathways, 6 EC groups and 48 operons.

The four time points were also analysed separately. In this case we only tested for differential expression, so the gene set scores were based on t-values rather than F-values. The sign of the gene set score thus give information about up- and downregulation. We refer to the two analyses as the *time series analysis* and the *individual time point analysis*.

4 Results

4.1 Simulated longitudinal data

Table 2 shows the REML estimates for the variance components in $\hat{V}_A$ averaged over 100 simulations. The time variance σ_t^2 is overestimated, while the random error variance σ_e^2 is likewise underestimated. However, the sum of the two parameters was estimated to 5.01 and the correlation between the estimates across the 100 simulations was -0.9989. From the structure of V we observe that if the time correlation function $\rho(u)$ is rapidly decaying as the time intervals increase, the time variance component and the error variance will be nearly non-identifiable. The off diagonal terms that include σ_t^2 will be close to zero, and in this case only the sum of σ_t^2 and σ_e^2 can be estimated. This is likely the case here, and points to the fact that if the time dependencies, beyond the modelled linear/quadratic time effects, are small, it is probably better to leave the time components out of the

Table 2: Average REML estimates based on 100 simulations for $\hat{V}_A$ (assuming dependence within gene sets) and their standard deviations (SD).

	$\sigma_b^2 = 1$	$\sigma_t^2 = 2$	$\sigma_g^2 = 2$	$\sigma_e^2 = 3$	$\phi = 0.9$	$\gamma = 0.3$
Estimate	1.00	2.81	2.02	2.20	1.11	0.30
SD	0.07	1.27	0.14	1.29	0.46	0.04

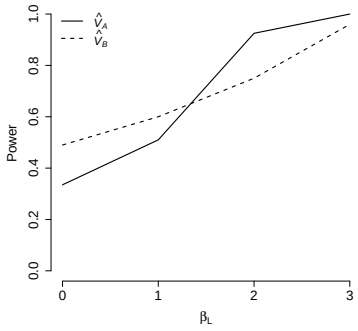
Table 3: Average REML estimates based on 100 simulations for $\hat{V}_B$ (assuming independence between genes) when ignoring gene dependencies.

	$\sigma_b^2 = 1$	$\sigma_t^2 = 2$	$\sigma_e^2 = 3$	$\phi = 0.9$
Estimate	1.00	2.90	2.12	1.11
SD	0.07	1.45	1.48	0.5

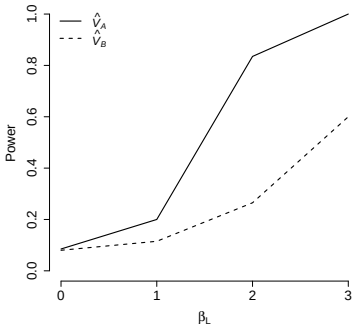
model for V. The time correlation parameter ϕ is slightly overestimated, meaning that the time correlation is slightly underestimated, but this is to a certain degree compensated for by the too large σ_t^2 estimate. Table 3 shows the estimated variances from REML for $\hat{V}_B$ where genes were assumed to be independent. The estimated variance components are more or less identical to those in $\hat{V}_A$ where the gene dependencies were taken into account, so the difference lies in the structure of the two matrices.

Figure 2 shows the power for the gene set rotation test after preprocessing the data with $\hat{V}_A$ and $\hat{V}_B$ in the three different scenarios. In Figure 2(a) we tested for gene (differential expression) and time effects (linear and quadratic) in scenario 1. The power of identifying the two important gene sets depends on the size of the linear time effect. With small linear time effects it is most beneficial not to correct for gene correlations, while for larger time effects it is better to include gene dependencies in the covariance structure. Figure 2(b) shows the results from scenario 2 where only time effects were added, but we tested for both gene and time effects. In this case it is beneficial to correct for gene dependencies irrespective of the magnitude of the linear time effect. Figure 2(c) shows the results of testing only for time effects in the same scenario, and now the power is identical for preprocessing with $\hat{V}_A$ and $\hat{V}_B$. In Figure 2(d) we tested for both gene and time effects in scenario 3 where only gene effects were added. In this scenario it is never beneficial to correct for gene dependencies.

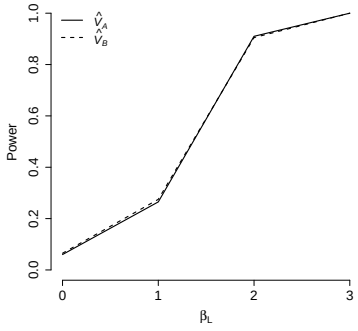
We used the GSEA enrichment score as the gene set test statistic (Subramanian et al., 2005). The enrichment score is calculated by ranking all genes by their test statistic before the position of the gene set members are identified in the ranked list. Gene sets that are clustered at the top or bottom of the list tend to get high scores. Hence, both correlation within gene sets and added gene and time effects are important for the gene set score. Gene sets that are highly correlated can get high gene effects by random chance, while a high linear or quadratic time effect is more unrealistic to get by chance since these effects have more complicated structures. If we had used another type of gene set score that does not consider



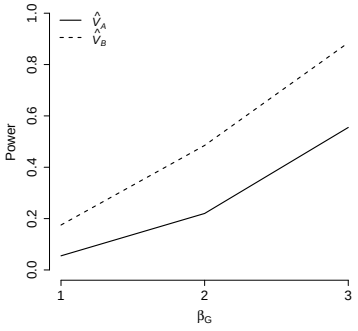
(a) Scenario 1: Testing for gene, linear and quadratic time effects



(b) Scenario 2: Testing for gene, linear and quadratic time effects



(c) Scenario 2: Testing for linear and quadratic time effects



(d) Scenario 3: Testing for gene, linear and quadratic time effects

Figure 2: Power in finding the important gene sets with the gene set rotation test. Each subfigure shows the results for one scenario where the data were preprocessed both with ($\hat{V}_A$) and without ($\hat{V}_B$) gene dependencies. Note that scenario 3 has β_G on the x-axis, while the other scenarios have β_L.

correlations between genes, then we may not have got the same results we see in Figure 2.

When the time effects are small it is mostly gene effects that are reflected in a high gene set score, and in this case it is better to keep correlations to get as high score as possible for the important gene sets. As the time effects increase they contribute more are more to the gene set score. For high time effects it is favourable to remove correlation between genes to remove false positive gene sets that get a high score by accident. In this way the important gene sets with both time and gene effects are easier to identify. The observations for the largest time effects in Figure 2(a) is further verified by the observations from Figure 2(b) where there were no gene effects, but we still tested for gene effects. By removing correlation within gene sets we remove false positives that accidentally get high scores because of their correlation, and the important gene sets with time effects are easier to identify. In comparison, if we do not test for gene effects in this scenario, there is no benefit of correcting for gene correlations either. The observations for the smallest time effects in Figure 2(a) are verified by Figure 2(d). When there are no time effects, only gene effects, the removal of correlations makes it more difficult to identify the sets with gene effects. A general conclusion is that by bringing gene dependencies into the preprocessing, we improve the ability to identify gene sets with strong time effects, but at the same time we make it more difficult to identify gene sets with mainly constant gene effects.

We note however that the type I error is not properly controlled in many of these scenarios when we use the incorrect covariance structure in the preprocessing. This is most notable on a gene level and not so serious on a gene set level. As an estimate of the type I error we used the proportion of times the non-important genes or gene sets came out as significant when we used a p-value cutoff of 0.05. The only scenario where the type I error was controlled properly also with $\hat{V}_B$ is in scenario 2 when we test for only time effects (when the power is identical for $\hat{V}_A$ and $\hat{V}_B$. The power curve for $\hat{V}_B$ should thus be adjusted slightly down for a more fair comparison of $\hat{V}_A$ and $\hat{V}_B$. The type I error was well under 0.05 for $\hat{V}_A$ on a gene set level, indicating that the gene set test may be slightly conservative (this has been noted about GSEA before by Tian et al. (2005) and Goeman and Bühlmann (2007)).

4.2 Stress response in *E. faecalis*

Table 4 shows the estimated covariance parameters in the *E. faecalis* data set. The estimate of γ corresponds to correlations of approximately 0.12 for the closest

Table 4: Covariance parameters in *E. faecalis* data estimated with REML.

σ_t^2	σ_g^2	σ_e^2	ϕ	γ
0.12	0.24	0.11	0.48	0.64

genes in the network, while the estimate of ϕ corresponds to correlations of approximately 0.62 for the closest time points. The gene set rotation test revealed four significant gene sets (FDR q-value$\leq$ 0.25) given in table 5. This is fewer gene sets than what was found with the original preprocessing in Dørum et al. (submitted) that did not consider gene dependencies. Judging by the results we saw in the simulation study, this may be an indication that many of the gene sets found in Dørum et al. contain genes with high gene effects and less distinct time effects, and that the focus now have been directed towards sets with strong time trends. This observation is in a sense verified by Figure 3 showing the time trends for the significant gene sets. The time trends are based on enrichment scores (normalised to account for size, see Subramanian et al. (2005)) from the individual time point analysis. These enrichment scores reflect only differential expression, and their signs indicate whether the genes in the set are mostly upregulated or downregulated. The gene sets seem to have strong time trends, except for maybe the operon EF1712 EF1721. A similar time trend plot in Dørum et al. revealed many significant gene sets without particular time trends. The pathway "Fatty acid biosynthesis" and the operon "EF2875 EF2886" show an almost identical trend, which is not surprising considering they are severely overlapping.

Table 5: Significant gene sets ($q \leq 0.25$) from the time series analysis of the *E. faecalis* data when testing for differential expression and time trends.

Gene set	q	Size*
Pathways		
One carbon pool by folate	0.213	7
Fatty acid biosynthesis	0.226	12
Operons		
EF1712 EF1721	0.026	9
EF2875 EF2886	0.149	11

*Genes in the set corresponding to genes in the expression data

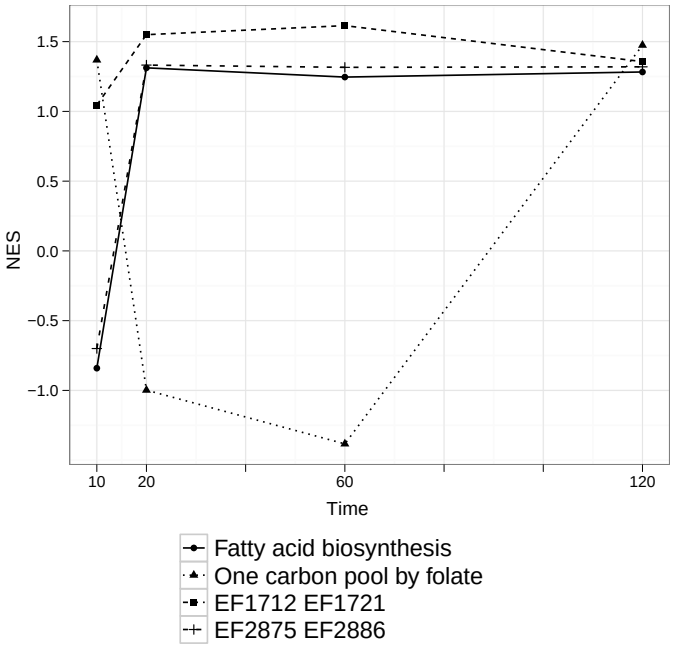
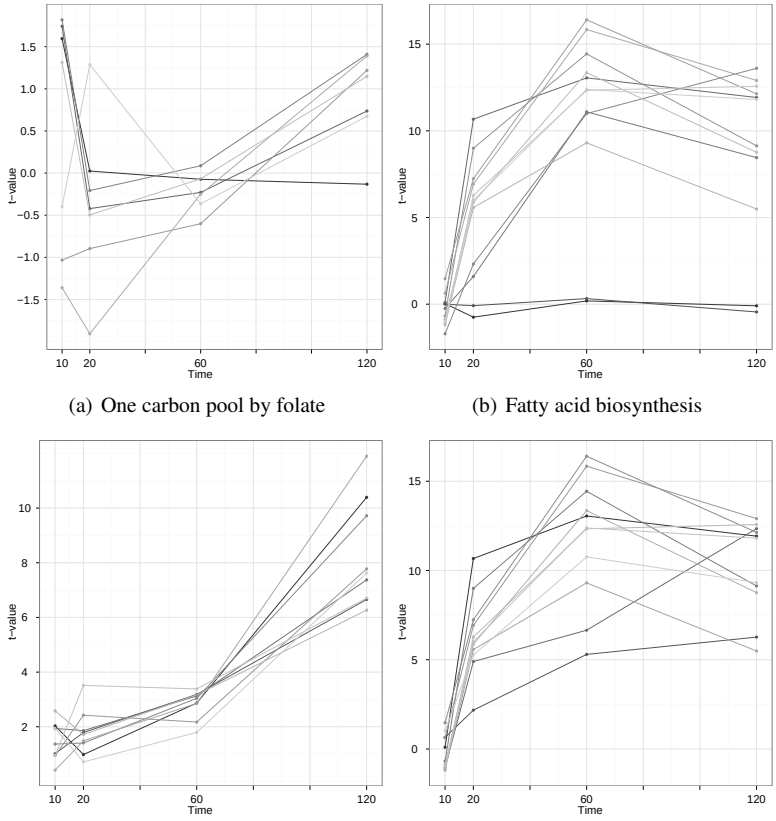


Figure 3: Time trends for the signficant gene sets in the time series analysis of the *E. faecalis* data. The time trend shows the normalised enrichment scores (NES) from the individual time point analysis.

Figure 4 shows the time trends for the individual genes in each of the significant gene sets. In this plot the modified t-values reflecting differential expression in the individual time point analysis have been plotted against time. The time trends for the individual genes seem to reflect the time trends we see on a gene set level in Figure 3 quite nicely. The exception is maybe once again the operon EF1712 EF1721, for which the gene set score does not reflect the constant increase over time that the genes show. The gene set score only increases between the first two time points and then retains the same level.



(a) One carbon pool by folate

(b) Fatty acid biosynthesis

(c) EF1712 EF1721

(d) EF2875 EF2886

Figure 4: Time trends for all genes in significant gene sets from the time series analysis of the *E. faecalis* data. Each time point shows the genes' modified t-values.

5 Discussion

Rotation testing for gene set tests was first introduced in Dørum et al. (2009) as an alternative to permutation testing for small sample sizes. Rotation tests can compute accurate p-values also for small sample sizes. Wu et al. (2010) adapted rotation test for gene set testing to data with complex designs, and in Dørum et al. we presented a gene set rotation test for longitudinal data. In this paper we have improved the preprocessing step in the latter paper to also accommodate gene dependencies. The simulation study indicates that including gene dependencies in the preprocessing changes the focus of the gene set test from sets with differential expression to sets with strong time trends. Although the method in Dørum et al. also could identify gene sets with time trends, the preprocessing presented in this paper has improved the ability to identify these sets.

By introducing gene dependencies into the covariance matrix we run the risk of overfitting the model if the gene dependencies are in fact very small. The division of the network into meaningful groups is therefore important in order to have as much correlation within groups as possible. In the *E. faecalis* data we found that the average correlation between members in each group or community ranged from -0.40 to 0.83, though the most extreme correlations were found in groups with only two members. The smallest correlation within a group was -0.003 (in a group with 3 members). The average correlation between the one-member groups, calculated to 0.00097, can be seen as an estimate of correlation between different groups. Unfortunately the network did not contain information about the direction of regulation, so we only modelled positive correlations. Judging from the negative correlations within some of the groups, it would have been more correct to model negative correlations. However, we emphasise that we use the information that is available, even though it is not complete.

An alternative to grouping genes based on communities in the network, could be to use predefined gene sets that are non-overlapping. In this case the same gene sets could have been used for estimating the covariance matrix and for testing in the subsequent gene set test. However, an advantage of splitting a network into subnetworks is that these can be as large as desired. Algorithms that split networks based on modularity looks for natural communities in the network, and hence will not look for groups of any particular size. Spectral graph partitioning is an alternative approach that splits networks into subnetworks of predefined size by the use of the eigenvectors of the graph's Laplacian matrix (Fiedler, 1973, Pothen et al., 1990).

References

F. R. K. Chung. Spectral graph theory. *No. 92 in Regional Conference Series in Mathematics. American Mathematical Society*, 1997.

A. Clauset, M. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70(6, Part 2), 2004. ISSN 1063-651X.

G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

P. J. Diggle, K.-Y. Liang, and S. L. Zeger. *Analysis of Longitudinal Data*. Oxford University Press, 1994.

G. Dørum, L. Snipen, M. Solheim, and S. Sæbø. Rotation gene set testing for longitudinal expression data. Submitted.

G. Dørum, L. Snipen, M. Solheim, and S. Sæbø. Rotation Testing in Gene Set Enrichment Analysis for Small Direct Comparison Experiments. *Statistical Applications in Genetics and Molecular Biology*, 8(1), 2009. ISSN 1544-6115.

B. Efron and R. Tibshirani. On testing the significance of sets of genes. *Annals of Applied Statisitcs*, 1(1):107–129, 2007. ISSN 1932-6157.

M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(2):298–305, 1973. ISSN 0011-4642.

J. Goeman, S. van de Geer, F. de Kort, and H. van Houwelingen. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99, 2004. ISSN 1367-4803.

J. J. Goeman and P. Bühlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, 2007. ISSN 1367-4803.

D. W. Huang, B. T. Sherman, and R. A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13, 2009. ISSN 0305-1048.

Z. Jiang and R. Gentleman. Extensions to gene set enrichment. *Bioinformatics*, 23(3):306–313, 2007. ISSN 1367-4803.

M. Kanehisa and S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, 2000. ISSN 0305-1048.

R. I. Kondor and J. D. Lafferty. Diffusion kernels on graphs and other discrete input spaces. In *Proceedings of the Nineteenth International Conference on Machine Learning*, 2002. ISBN 1-55860-873-7.

O. Langsrud. Rotation tests. *Statistics and Computing*, 15(1):53–60, 2005. ISSN 0960-3174.

V. Mootha, C. Lindgren, K. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstrale, E. Laurila, N. Houstis, M. Daly, N. Patterson, J. Mesirov, T. Golub, P. Tamayo, B. Spiegelman, E. Lander, J. Hirschhorn, D. Altshuler, and L. Groop. PGC-1 alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34(3):267–273, 2003. ISSN 1061-4036.

M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3, Part 2), 2006a. ISSN 1539-3755.

M. E. J. Newman. Modularity and community structure in networks. *PNAS*, 103 (23):8577–8582, 2006b. ISSN 0027-8424.

M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2, Part 2), 2004. ISSN 1063-651X.

A. Pothen, H. Simon, and K. Liou. Partitioning sparse matrices with eigenvectors of graphs. *Siam Journal on Matrix Analysis and Applications*, 11(3):430–452, 1990. ISSN 0895-4798.

G. Smyth, J. Michaud, and H. Scott. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, 21 (9):2067–2075, 2005. ISSN 1367-4803.

M. Solheim, A. Aakra, H. Vebo, L. Snipen, and I. F. Nes. Transcriptional responses of Enterococcus faecalis V583 to bovine bile and sodium dodecyl sulfate. *Applied and Environmental Microbiology*, 73(18):5767–5774, 2007. ISSN 0099-2240.

A. Subramanian, P. Tamayo, V. Mootha, S. Mukherjee, B. Ebert, M. Gillette, A. Paulovich, S. Pomeroy, T. Golub, E. Lander, and J. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 102(43):15545–15550, 2005. ISSN 0027-8424.

L. Tian, S. Greenberg, S. Kong, J. Altschuler, I. Kohane, and P. Park. Discovering statistically significant pathways in expression profiling studies. *PNAS*, 102 (38):13544–13549, 2005. ISSN 0027-8424.

D. Wu, E. Lim, F. Vaillant, M.-L. Asselin-Labat, J. E. Visvader, and G. K. Smyth. ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics*, 26(17):2176–2182, 2010. ISSN 1367-4803.

J. D. Zhang and S. Wiemann. KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor. *Bioinformatics*, 25(11):1470–1471, 2009. ISSN 1367-4803.