Appendix

A.1 Derivation of Equation (3.3)

Setting $m := |M|$ we write down $S_C(I \setminus M)$, correcting for the effects of the gene exclusion on the mean via δ_μ:

$$\begin{aligned}
S_C(I \setminus M) &= \frac{1}{n_c(k-m)} \sum_{i \in I,\, i \notin M} ||c^i - (\mu + \delta_\mu)||_2^2 \\
&= \frac{1}{n_c(k-m)} \sum_{i \in I,\, i \notin M} \Big\{ \underbrace{||c^i - \mu||_2^2}_{:=A} + \underbrace{2(c^i - \mu)^T \delta_\mu}_{:=B} + \underbrace{||\delta_\mu||_2^2}_{:=C} \Big\}
\end{aligned}$$

The refitting-effect δ_μ is available as:

$$\delta_\mu = \frac{1}{k-m} \sum_{i \in I,\, i \notin M} c^i - \mu \quad = \mu\Big(\frac{k}{k-m} - 1\Big) - \frac{1}{k-m} \sum_{i \in m} c^i = \frac{1}{k-m} \sum_{i \in M} (\mu - c^i)$$

With this we get for B and C, taking into account the factors before the sums:

$$\begin{aligned}
B &= \frac{-2}{n_C(k-m)^2} \sum_{i,j \in M \times M} (\mu - c^i)^T (\mu - c^j) \\
C &= \frac{1}{n_C(k-m)^2} \sum_{i,j \in M \times M} (\mu - c^i)^T (\mu - c^j)
\end{aligned}$$

Rewriting the sum contributing to A we retrieve the first two terms of Equation (3.3):

$$A = \frac{1}{n_c(k-m)} \Big\{ \sum_{i \in I} ||c^i - \mu||_2^2 - \sum_{i \in M} ||c^i - \mu||_2^2 \Big\} = \frac{k}{k-m} S_C(I) - \frac{1}{n_c(k-|M|)} \sum_{i \in M} ||c^i - \mu||_2^2$$

A.2 The `dcoex` software package

Synopsis

The software package `dcoex` is a tool for exploratory microarray data analysis. It identifies groups of genes that are differentially coexpressed between two phenotypically distinct conditions. This provides the user with a means complementing hypothesis generation driven by techniques like clustering or on the basis of differential gene expression. The significance of the results is assessed via a permutation procedure.

Motivation

Looking for groups of differentially coexpressed genes can be motivated as follows: In case a regulatory mechanism is disrupted between two phenotypes, the coregulation of groups of genes to changes by definition. Via the hypothesis that coregulation of genes is reflected in their coexpression, this implies searching for groups of differentially coexpressed genes [85, 108]. In a simplistic view, these groups of genes contain genes causing the phenotypical distinction.

Algorithms attempting to find *differentially* coexpressed genes need to be contrasted to algorithms looking for coexpression such as various clustering and biclustering approaches [24, 137, 145, 156]. More similar is the approach of Dettling et al. [33] who look for groups of genes with differential group-mean expression.

Implementation

The main challenge is screening the large space of candidate groups, for which we have provided an approach in Section 3.2. In [85] we introduced an efficient to compute criteria allowing a fast screening of neighborhoods in a descent algorithm. We implemented the algorithm using the statistical programming language R [70] in form of the package `dcoex`.

Input and output `dcoex` takes two expression matrices as input, one fore each phenotype. The matrices must be compatible, i.e. each row should code for the same gene. Additionally, there are two tuning parameters: One for the target size of the group, another one for the greediness of the search. Further on, a starting point for the descent algorithm can be provided. `dcoex` returns a group of differentially coexpressed genes accompanied by a score quantifying the amount of differential coexpression (Equation 3.4).

Significance assessment On the basis of the differential coexpression score we provide two types of permutation procedures to assess how likely such a score arises by chance. We repeatedly shuffle patient labels (*a*) for all genes at once and (*b*) for each gene separately. The frequency of scores smaller than or equal than the "biological" score then yields an empirical *p*-value. While (*a*) assesses the impact of the group assignment it might be prone to the effect of confounding variables. Version (*b*) is less conservative by removing all gene correlation from the null hypothesis.

Example analysis

As an example we present the analysis leading to the results presented in Section 3.3. We assume a normalized expression matrix of the leukemia data stored in the file `expr_mat.rdat`.

```
###=== read in the expression data
> library(dcoex)
> load("expr_mat.rdat")

###=== filter out low variance genes
> inds = v > quantile( apply(expr.mat,1,mad) ,.5)
> expr.mat = expr.mat[inds,]

###=== extract the two phenotypical groups:
> mat1 = expr.mat[,159:176] ##-- normal as control
> mat2 = expr.mat[,1:15] ##-- BCR-ABL translocation

###=== run algorithm with default values and random starting point
> res = dcoex(m1,m2)

###=== significance assessment (takes some time...)
> shuffled.a = dcoex.shuffle(mat1,mat2,times=1000,type="smpl.wise")
> shuffled.b = dcoex.shuffle(mat1,mat2,times=1000,type="gene.wise")
> epval.a = sum(shuffled.a <= res$res) / length(shuffled.a)
> epval.b = sum(shuffled.b <= res$res) / length(shuffled.b)
```

A.3 The docval software package

Synopsis

The software package docval is a tool for the derivation and unambiguous documentation of diagnostic gene expression signatures. It implements versions of the vsn and rma preprocessing schemes for Affymetrix GeneChip microarrays that greatly facilitate external application of a derived signature.

Motivation

Gene expression diagnostic signatures are potentially powerful tools in the area of molecular medicine. Encouraging results have been achieved for diagnosis of disease, prognosis of survival, assessment of risk group and selection of treatment [112, 130, 152]. Before these results can impact clinical practice, it needs to be established that expression signatures hold information complementary to existing prognostic markers. While sharing candidate signatures within the research community can accelerate the process of evaluation, this does not allow for any ambiguity of signatures. As the docval software is designed to minimize ambiguity of diagnosis it can be assumed to greatly facilitate this endeavor.

Implementation and use

We focus our efforts on the documentation of gene expression signatures derived from Affymetrix GeneChip data. The docval package is implemented in the statistical programming language R [70], and depends on the Bioconductor [54] packages affy, vsn and pamr. In the following, we give an example of how docval can be used to derive a diagnostic signature, document it unambiguously and classify external patient data.

Derivation and documentation of a diagnostic signature

As input docval takes an object of class AffyBatch, which is provided by the affy package and can be easily produced from a path to a collection of Affymetrix .cel-files. Then it is able to generate vsn or rma normalized data sets of type exprSet, which is provided by Biobase.

```
###=== read in the data
> setwd("data_dir")
> abo = ReadAffy()
```

```
###=== preprocessing using vsn
> exs.vsn = wrap.val(abo,method="vsn")
###=== preprocessing using rma
> exs.rma = wrap.val(abo,method="rma")
```

The normalized core data exs.[rma,vsn] contains the information needed for the scale adjusting transformation of external patients. At this point one has the option to derive a molecular signature from the core data using the nearest shrunken centroid method [139] as implemented in the package pamr. Prerequisite, however, is a labeling of the arrays of the core data set.

```
###=== lab contains the sample labels
> lab = my.labels

###=== derive the signature
> sig.vsn = pamr.fil(exs.vsn,lab,fil=FALSE)
> sig.rma = pamr.fil(exs.rma,lab,fil=FALSE)

###=== add scale information
> sig.byval.rma = list(sig=sig.rma, params=preproc(description(exs.rma))$val)
> sig.byval.vsn = list(sig=sig.vsn, params=preproc(description(exs.vsn))$val)
> save(sig.byval.rma, file="sig_byval_rma.rdat")
> save(sig.byval.vsn, file="sig_byval_vsn.rdat")
```

Diagnosis of an external patient

At this point, a diagnostic signature together with enough information for unambiguous documentation has been derived. docval can now be used for diagnoses of external patients:

```
###=== load external patient data and the signatures
> abo.extrnl = ReadAffy("external_patient.CEL.gz")
> load("sig_byval_rma.rdat")
> load("sig_byval_vsn.rdat")

###=== transform the external patient to the scale of the core study
> exs.extrnl.rma = wrap.val.add(abo.extrnl,sig.byval.rma$params,method="rma")
> exs.extrnl.vsn = wrap.val.add(abo.extrnl,sig.byval.vsn$params,method="vsn")

###=== diagnose the external patient
> diag.rma = sig.byval.rma$sig(exprs(exs.extrnl.rma))
> diag.vsn = sig.byval.rma$sig(exprs(exs.extrnl.vsn))
```

A.4 Short summary (in german)

Zusammenfassung

Diese Arbeit dreht sich um die Charakterisierung von Krankheiten mit Hilfe von *Genexpressionsdaten*. Solche Daten stellen Zellen auf molekularer Ebene dar und können zur Beschreibung von Krankheiten auf zweierlei Art verwendet werden: Zum einen kann man bekannte Krankheiten genauer und verläßlicher diagnostizieren. Zum anderen kann man versuchen, in stetig wiederkehrenden Expressionsmustern entweder neue Krankheitsentitäten zu entdecken, oder aber aufgrund solcher Muster auf biologishmedizinische Ursachen bekannter Krankheiten zu schließen. Die vorliegende Arbeit enthält methodologisch neue Ansätze für beide Szenarien.

Nach einer Einleitung, die unter anderem die *Microarray Technik* kurz skizziert, folgt ein ein weiteres einführendes Kapitel. Darin werden Methoden der *statistischen Lerntheorie* beschrieben, die man benutzen kann um aus Beispieldaten Schemata, oder *molekulare Signaturen*, für eine Diagnose abzuleiten. Die Darstellung ist auf die Anwendung statistischer Verfahren auf Microarray Daten zugeschnitten und das Kapitel bildet die theoretische Grundlage der folgenden Arbeit.

Thema des zweiten Kapitels ist die unzweideutige *Dokumentation* einmal hergeleiteter molekularer Signaturen. Die Dokumentation einer Expressionssignatur ist ein notwendiger Schritt, falls diese zwischen Wissenschaftlern und Forschungseinrichtungen ausgetauscht werden soll. Ein solcher Austausch aber muss Tests und Validierungen einer Signatur vorangehen, die ihrerseits für den klinischen Einsatz unerläßlich sind. Wir stellen zwei Methoden vor, die gebräuchliche Strategien der Datenvorverarbeitung ergänzen und demonstrieren eine signifikante Erhöhung der Stimmigkeit von Diagnosen an verschiedenen Datensätzen.

Im dritten Kapitel wird das Konzept der *differentiellen Ko-expression* und der dazugehörige `dcoex` Algorithmus vorgestellt. Eine Gruppe differentiell koexprimierter Gene hat die Eigenschaft in Proben eines bestimmten Phänotyps kohärent exprimiert zu sein, verliert diese Kohärenz allerdings in den Proben eines anderen Phänotyps. Der `dcoex` Algorithmus ist eine Methode solche Gruppen von differentiell koexprimierten Genen in Datensätzen zu finden, wobei in kombinatorisches Optimierungsproblem heuristisch gelöst wird. Gruppen differentiell koexprimierter Gene können nicht nur zur molekularen Charakterisierung unterschiedlicher Phänotypen beitragen. Aus den Gengruppen abgeleitete Informationen kann man zur Formulierung fokussierter biologischer Hypothesen verwenden. Wir demonstrieren dies an einem Leukämiedatensatz.

A.5 List of pulications associated with this thesis

Expression of late cell cycle genes and an increased proliferative capacity characterize very early relapse of childhood acute lymphoblastic leukemia Renate Kirschner-Schwabe, Claudio Lottaz, Jörn Tödling, Peter Rhein, Leonid Karawajew, Cornelia Eckert, Arend von Stackelberg, Ute Ungethm, Dennis Kostka, Andreas E. Kulozik, Wolf-Dieter Ludwig, Gnter Henze, Rainer Spang, Christian Hagemeier and Karl Seege In *Clinical Cancer Research* 2(15), 2006: 4553-61

Traces of molecular disease mechanisms on microarrays D. Kostka, C. Lottaz and R. Spang In *Statistische Methoden in der empirischem Forschung* ISSN 1615-4177, (Hrsg.: J. Kaufmann), Berlin: Schering AG 231-236 (2005)

Finding disease specific alterations in the coexpression of genes D. Kostka and R. Spang In *Bioinformatics* 20(Suppl. 1), 2004: i194-i199 (Proceedings of ISMB 2004)

Patient Classification C. Lottaz, D. Kostka and R. Spang In *Bioinformatics - From Genomes to Therapies* T. Lengauer (Editor), to be published by Wiley-VCH. [in press]

Computational Diagnostics with Gene Expression Profiles C. Lottaz, D. Kostka, F. Markowetz and R. Spang In *Methods in Molecular Biology – Bioinformatics* J. Keith (Editor), to be published by Humana Press. [accepted]

Ehrenwörtliche Erklärung

Hiermit erkläre ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen verwendet habe.

Berlin, Oktober 2006 Dennis Kostka