Translation of disease associated gene signatures across tissues

Adetayo Kasim*

Wolfson Research Institute for Health and Wellbeing,
Durham University Queen's Campus,
Stockton-on-Tees, TS17 6BH, UK
Email: a.s.kasim@durham.ac.uk
*Corresponding author

Ziv Shkedy and Dan Lin

Interuniversity Institute for Biostatistics and Statistical Bioinformatics,
Hasselt University,
Diepenbeek, Belgium
Email: ziv.shkedy@uhasselt.be
Email: dan.lin@uhasselt.be

Suzy Van Sanden

Janssen Pharmaceutica NV,
Beerse, Belgium
Email: svsande1@its.jnj.com

Josè Cortiñas Abrahantes

European Food Safety Authority,
Parma, Italy
Email: jose.cortinasabrahantes@uhasselt.be

Hinrich W.H. Göhlmann and Luc Bijnens

Janssen Pharmaceutica NV,
Beerse, Belgium
Email: hgoehlma@its.jnj.com
Email: lbijnens@its.jnj.com

Dani Yekutieli

Department of Statistics and Operational Research,
Tel Aviv University,
Tel Aviv, Israel
Email: yekutiel@math.tau.ac.il

Michael Camilleri

Mayo Clinic College of Medicine,
Rochester, MN, USA
Email: camilleri.michael@mayo.edu

Jeroen Aerssens and Willem Talloen

Janssen Pharmaceutica NV,
Beerse, Belgium
Email: jaerssen@its.jnj.com
Email: wtalloen@its.jnj.com

Abstract: It has recently been shown that disease associated gene signatures can be identified by profiling tissue other than the disease related tissue. In this paper, we investigate gene signatures for Irritable Bowel Syndrome (IBS) using gene expression profiling of both disease related tissue (colon) and surrogate tissue (rectum). Gene specific joint ANOVA models were used to investigate differentially expressed genes between the IBS patients and the healthy controls taken into account both intra and inter tissue dependencies among expression levels of the same gene. Classification algorithms in combination with feature selection methods were used to investigate the predictive power of gene expression levels from the surrogate and the target tissues. We conclude based on the analyses that expression profiles of the colon and the rectum tissue could result in better predictive accuracy if the disease associated genes are known.

Keywords: gene expression; joint modelling; surrogate markers; biomarker; target tissue; surrogate tissue; features selection; classification and class prediction and irritable bowel syndrome.

Reference to this paper should be made as follows: Kasim, A., Shkedy, Z., Lin, D., Van Sanden, S., Abrahantes, C.J., Göhlmann, H.W.H., Bijnens, L., Yekutieli, D., Camilleri, M., Aerssens, J. and Talloen, W. (2015) 'Translation of disease associated gene signatures across tissues', *Int. J. Data Mining and Bioinformatics*, Vol. 11, No. 3, pp.301–313.

Biographical notes: Adetayo Kasim is currently a Research Statistician at Wolfson Research Institute for Health and Wellbeing, Durham University, UK. He received a BSc degree in Mathematical Sciences (2002) from University of Agriculture Abeokuta, Nigeria. He holds MSc in Applied Statistics (2005) and MSc in Biostatistics (2006) from Universiteit Hasselt, Belgium. He received a PhD in Statistical Bioinformatics (2010) from Universiteit Hasselt. His current research interests include application of statistical methods in genomics, clinical and social trials, health inequality, global health and anthropology.

Ziv Shkedy is an Associate Professor of Biostatistics and Bioinformatics at Hasselt University, Belgium. He is a co-author of numerous publications applying statistical methods to infectious diseases data, non-clinical experiments in early drug development and the analysis of microarray and gene expression data. Over the last 15 years, he has collaborated with European organisations (ECDC, EMCDDA) on many projects relating to infectious diseases and with pharmaceutical partners on clinical, non-clinical and early drug development projects.

Dan Lin holds a PhD degree in Bioinformatics from Hasselt University, Belgium, where her research focused on analysis of ‘omics’ data from early drug development experiments. She currently works as biometrician at Zoetis (formerly Pfizer Animal Health) research and development for discovery and clinical studies for biological, pharmaceutical and diagnostic veterinary medicine.

Suzy Van Sanden is currently working as a statistician on a European health economics and market access team at Janssen, a subdivision of J&J. Her main focus there is the analysis of patient follow-up studies and observational databases. She obtained a PhD in Statistics at the University of Hasselt in 2008, where she mainly worked on the analysis of genomics data.

Josè Cortiñas Abrahantes is Scientific Officer at European Food Safety Authority. He received the BS degree in Mathematics (1992) from Havana University, a MSc in Biostatistics (1999) and a PhD in Biostatistics (2004) from Universiteit Hasselt. He published on surrogate markers in clinical trials, on multivariate frailty models, incomplete data methods and classification methods for multivariate multi class problems.

Hinrich W.H. Göhlmann studied Biology at the Technische Hochschule Darmstadt, Germany, receiving his Diploma in 1995. He received his doctoral degree for his work on a microarray based whole genome expression analysis of *Mycoplasma pneumoniae* in 1999. Following this, he joined the Department of Functional Genomics at Janssen Pharmaceutical Companies of Johnson & Johnson in Beerse, Belgium as a PostDoc. He currently holds a position as senior principal scientist heading the High Dimensional Biology team in the CREATe department of Core Scientific Technologies. The team consists of experts in the field of next generation sequencing, microarray technology, FACS, High Content Screening and the RNAi platform.

Luc Bijnens holds MSc and PhD degrees in Biology from the University of Antwerp, Belgium and a MSc in Biostatistics from the University of Hasselt, Belgium. He spent the earlier part of his career in academia at the University of Antwerp, Belgium and Kisangani, Democratic Republic of Congo, and later with Bristol Meyers Squibb and the European Organization of Research and Treatment of Cancer. He joined Johnson and Johnson in 1997 where he built a non-clinical biostatistics team that develops statistical methodology and software for R&D.

Dani Yekutieli received his PhD in Applied Statistics from Tel Aviv University in 2002. He is currently a Senior Lecturer at the Department of Statistics and OR in Tel Aviv University.

Michael Camilleri, MD, is a consultant in the Division of Gastroenterology and Hepatology, Department of Internal Medicine at Mayo Clinic in Rochester, Minnesota. He holds the academic rank of Professor of Medicine, Physiology, and Pharmacology, and he is the Atherton and Winifred W. Bean Professor. He is currently Executive Dean of the Department of Development at Mayo Clinic. He attended the University of Malta Medical School. He received a Masters degree in Physiology from University of London, UK.

Jeroen Aerssens is heading the Biomarkers & Translational Pharmacology department in the Community of Research Excellence & Advanced Technology (C.R.E.A.Te) of Janssen Research & Development. He holds a PhD from the University of Leuven (Belgium) and gained over 15 years of R&D experience

within different pharmaceutical companies of Johnson & Johnson. His scientific interest and expertise is in pharmacogenomics and translational biomarker research in various disease areas.

Willem Talloen is a Principal Statistician at Janssen Pharmaceuticals Group. He has a MSc in Biostatistics (Univ. of Hasselt) and a PhD in Biology (Univ. of Antwerp). Before joining J&J in 2005, he worked for the Belgian Public Institute of Health as a statistical consultant. Some of his work resulted in a book, seven book chapters, two patents and more than 35 biological and/or statistical publications. He was nominated for the Janssen Business Excellence awards in 2008. His main research interests lie in the statistical analysis of high-dimensional data generated by -omics approaches.

This paper is a revised and expanded version of a paper entitled 'Translating gene signatures across tissues: a statistical view' presented at the 'International Conference on Bioinformatics & Computational Biology, BIOCOMP 2010', Las Vegas Nevada, USA, 12–15 July 2010.

1 Introduction

The impacts of microarrays in pharmaceutical and biomedical research can be underscored by gains in revealing functions of genes (Aerssens et al., 2008), tumor classification (Golub et al., 1999), drug target identification (Debouck and Goodfellow, 1999) and prediction of the response to therapy (Tusher et al., 2001). Molecular biomarkers for a disease are often investigated by profiling the disease-affected tissue (Alon et al., 1999; Dudoit et al., 2002; Ross et al., 2008). Recent developments in microarray experiments include; multiple outcomes (such as subtype classification and survival prediction, Cai et al., 2010), analysis across multiple experiments/datasets (Xu et al., 2009) and profiling of an alternative tissue, when the disease-related tissue is not readily available or the alternative tissue is more easily accessible.

Achiron and Gurevich (2006) performed a microarray experiment on Peripheral Blood Mononuclear Cells (PBMC) as an alternative to brain tissue to investigate gene signatures for multiple sclerosis, a chronic inflammatory demyelinating autoimmune disease affecting the central nervous system. Similarly, Le-Niculescu et al. (2008) identified biomarkers for mood disorders using gene expression from the blood. They concluded that blood biomarkers may offer an unexpectedly informative window into brain functioning and disease state. For hepatis C, Asselah et al. (2008) investigated gene signatures to predict response to pegylated interferon plus ribavirin combination therapy in patients with chronic hepatitis C, based on liver gene expression. While their study reported a potentially useful signature, the liver biopsy is difficult to obtain and it is thus hard to transfer such a test into clinical use. Therefore, other investigative searches of gene signatures associated with the response to interferon plus ribavirin combination therapy in patients with chronic hepatitis C used PBMC and also provided reasonable predictability of treatment response (Huang et al., 2008a; Huang et al., 2008b).

In this paper, we investigate gene signatures for Irritable Bowel Syndrome (IBS) using both the disease-affected tissue (Colon) and an alternative tissue (Rectum) based on the previous study from Aerssens et al. (2008), which identified gene signatures for IBS based on gene expression profiling of only the colon tissue. Here, we show that expression profiling of the rectum tissue, a neighbouring tissue, may also be predictive of IBS for the known disease associated gene signatures. Moreover, it is shown that the prediction accuracy can be improved as compared to when featured selection (selection of a subset of genes that are associated with IBS) and classification are based solely on the rectum tissue. Note that our objective is not to replace the colon tissue with the rectum tissue for the diagnosis of IBS, but to show that the rectum tissue could also be important in the clinical diagnosis of IBS. The colon tissue is here considered the 'target tissue', given that abnormalities in bowel function (diarrhoea and/or constipation) are a clinically important characteristic for IBS, whereas the neighbouring rectum tissue is considered as the 'surrogate tissue'.

This paper is organised as follows: Section 2 describes the case study based on IBS. A joint ANOVA model of the two tissues for detecting differentially expressed genes is formulated in Section 3. Section 3.1 presents the application of the joint ANOVA model on the case study. The methodology for classification and class prediction is discussed in Section 4 and Section 4.1 presents the application of the feature selection and classification methods on the case study. The paper is concluded with a discussion in Section 5.

2 The case study: irritable bowel syndrome

The data were derived from 58 subjects, including 34 IBS patients and 24 healthy controls. For a more detailed description of the clinical characteristics of the cohort study, we refer to Aerssens et al. (2008). For each subject, three biopsies were taken for gene expression profiling: two samples from the colon tissue at 10 cm apart and a third sample from the rectum tissue. A detailed description of the procedures of biopsy collection and further sample processing (RNA extraction, biotin labelling, hybridisation on Affymetrix Human Genome U133-Plus2.0 GeneChips) are provided in Aerssens et al. (2008).

Sample processing was performed in four batches, each of which comprised samples from both IBS and healthy subjects. Gene expression summary values for raw GeneChip data were computed using the gcRMA algorithm (Wu et al., 2004), which performs background adjustment, log transformation of the intensities to base two, quantile normalisation and summarisation, taking guanosine-cytidine affinities into account. Presence-Absence call (PANP, Warren et al., 2007) was used for calling the detection of genes absent or present. In total, expression profiles of 21,212 genes were measured for each sample.

3 A joint ANOVA model for the two tissues

The gene expression data consists of two expression matrices, $\boldsymbol{X}^A$ and $\boldsymbol{X}^B$ for the target and surrogate tissues, respectively. Note that there are two replicates per subject for the target tissue and one for the surrogate tissue. Let Z_i ($i = 1,\dots,n$) be an indicator variable representing the disease status of the ith subject given by:

$$Z_i = \begin{cases} 1 & \textit{diseased} \\ 0 & \textit{healthy} \end{cases}$$

We assume that the mean gene expression depends on the disease status for each tissue, i.e.

$$\begin{aligned} E(X_{ij_k}^{A}\,|Z_i) &= \mu_j^A + \alpha_j^A Z_i, \quad k=1,2\,; \quad j=1,\ldots,m\,;\ i=1,\ldots,n \\ E(X_{ij}^{B}\,|Z_i) &= \mu_j^B + \alpha_j^B Z_i \end{aligned} \tag{1}$$

where X_{ijk}^{A} represents the *k*-th replicate of the expression level of gene *j* from the *i*-th subject based on the target tissue and X_{ij}^{B} represents the expression level of the *j*-th gene and subject *i* from the surrogate tissue. We denote the gene-specific disease effects by α_j^A and α_j^B for the target and surrogate tissues, respectively. The parameters μ_j^A and μ_j^B denote gene specific intercepts. Note that the mean structure in (1) implies that the mean expression for each gene from the target tissue is equal for the two samples taken from the same subject. Since gene expression profiling of the subjects were done in four batches, there is a need to account for possible batch effects that may result from change in temperature, incubation time and other unknown factors associated with a specific batch. The batch effects were accounted for in the model as fixed effects. Let F_{il} takes the value of 1 if the sample from subject *i* is processed in batch *l* and 0 otherwise, where $l = 1,2,\ldots w - 1$ and *w* denotes the number of batches ($w = 4$ for our case study). The mean structure in (1) was modified as follows:

$$\begin{aligned} E(X_{ij_k}^{A}\,|Z_i) &= \mu_j^A + \alpha_j^A Z_i + \sum_{l-1}^{w-1} \beta_{jl}^A F_{il} \\ E(X_{ij}^{B}\,|Z_i) &= \mu_j^B + \alpha_j^B Z_i + \sum_{l-1}^{w-1} \beta_{jl}^B F_{il} \end{aligned} \tag{2}$$

where β^A and β^B are vectors of batch effects from the target and surrogate tissues, respectively. To take possible correlation between the three measurements of the same gene into account, we formulate a gene-specific joint model in the following way:

$$\begin{pmatrix} X_{ij_1}^{A} \\ X_{ij_2}^{A} \\ X_{ij}^{B} \end{pmatrix} \sim N \begin{pmatrix} \mu_j^A + \alpha_j^A Z_i + \sum_{l=1}^{w-1} \beta_{jl}^A F_{il} & \\ \mu_j^A + \alpha_j^A Z_i + \sum_{l=1}^{w-1} \beta_{jl}^A F_{il}, & S_j \\ \mu_j^B + \alpha_j^B Z_i + \sum_{l=1}^{w-1} \beta_{jl}^B F_{il} & \end{pmatrix} \tag{3}$$

Here, S_j is a gene-specific covariance matrix. The following covariance structures are assumed.

Model 1: The ‘independent samples’ model. This model assumes that the three measures are independent. Hence, the covariance matrix is given by:

$$S_j = \begin{pmatrix} \sigma_{11}^A & 0 & 0 \\ 0 & \sigma_{22}^A & 0 \\ 0 & 0 & \sigma_{33}^B \end{pmatrix} \tag{4}$$

Model 2: This model allows for correlations between sample from the target tissue. The covariance matrix is given by:

$$S_j = \begin{pmatrix} \sigma_{11}^A & \sigma_{12}^A & 0 \\ \sigma_{21}^A & \sigma_{22}^A & 0 \\ 0 & 0 & \sigma_{33}^B \end{pmatrix} \tag{5}$$

Model 3: The third model takes into account all possible correlations, i.e. between and within tissues correlation of gene expression levels. For this model, the covariance matrix is given by:

$$S_j = \begin{pmatrix} \sigma_{11}^A & \sigma_{12}^A & \sigma_{13}^{AB} \\ \sigma_{21}^A & \sigma_{22}^A & \sigma_{23}^{AB} \\ \sigma_{31}^{AB} & \sigma_{32}^{AB} & \sigma_{33}^B \end{pmatrix} \tag{6}$$

The variance of the expression levels of gene *j* are denoted by σ_{11}^A and σ_{22}^A for the first and second replicates from the target tissue and by σ_{33}^B for the single replicate from the rectum tissue. The covariance between the first and second replicates from the target tissue is denoted by $\sigma_{12}^A = \sigma_{21}^A$. The parameter $\sigma_{13}^{AB} = \sigma_{31}^{AB}$ denotes the covariance between the first replicate from the target tissue and the single replicate from the surrogate tissue and $\sigma_{23}^{AB} = \sigma_{32}^{AB}$ represent the covariance between the second replicate of the target tissue and the single replicate from the surrogate tissue. The parameters of the joint models were estimated using the 'mixed procedure' in SAS.

The main hypotheses for the joint ANOVA models can be stated as:

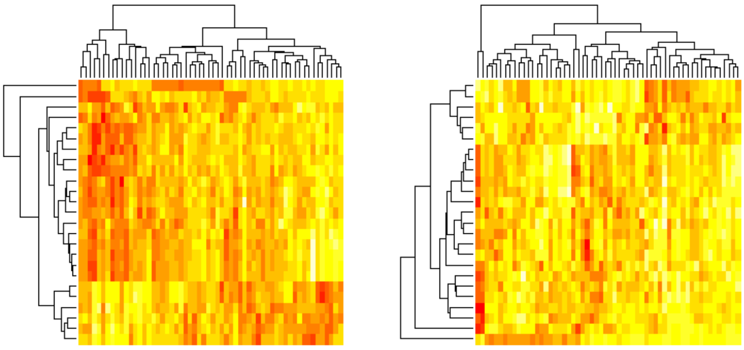
$$\begin{matrix} H_{0_j}^A : \alpha_j^A = 0 & & H_{0_j}^B : \alpha_j^B = 0 \\ & \textit{and} & \\ H_{1_j}^A : \alpha_j^A \neq 0 & & H_{1_j}^B : \beta_j^B \neq 0 \end{matrix} \tag{7}$$

The hypotheses are to test gene-specific average expression differences between IBS patients and healthy controls for both the colon and rectum tissues. In doing so, we also want to investigate how information shared across the two tissues through the model-based covariance matrix influences the number of differentially expressed genes between IBS patients and healthy control. By incorporating the covariance structure between colon and rectum tissues, the dependence between the two replicates from the target tissue and dependence between the two tissues are taken into account.

3.1 Application to the data

Akaike Information Criterion (AIC) was used to explore the dependencies between samples from the same gene. Model 2 has smaller AIC values than model 1 for 17,583 out of the 21,212 genes in the expression data; model 3 has smaller AIC than model 2 for 18,392 genes out of the 21,212 genes in the expression data. There is also a group of genes that have similar AIC values across all the three models, this group may correspond to the house-keeping genes. At the FDR of 5%, no genes were found to be differentially expressed between the IBS patients and the healthy controls based on gene expression profiling of the rectum tissue, but 30, 176 and 365 genes were found to be differentially expressed between the IBS patients and the healthy controls based on the target tissue when model 1, model 2 and model 3 were used, respectively. Indeed, accounting for both between and within tissue correlations in the model corrects for the obvious dependency between samples that are collected in the same subjects and/or tissue, resulting in fewer differentially expressed genes. Using the most optimum model for each gene, there were 140 differentially expressed genes from the colon tissue, whilst no differential expressed gene was found for the rectum tissue. Figure 1 seems to suggest that gene intensities are more subtle in the rectum tissue as compared to the colon tissue, which may partly explain why there were differentially expressed genes found in the colon tissue, but not in the rectum tissue. It may also be that relevant genes based on gene expression profiling of the rectum tissue were lost in a pool of false positives due to the failure of the gene-by-gene analysis to capture interactions among potential gene signatures.

Figure 1 Heat maps of expression profiling of colon and rectum tissues for relevant genes to irritable bowel syndrome (see online version for colours)



(a) Colon (b) Rectum

Notes: On the rows are the genes and on the columns are the subjects.

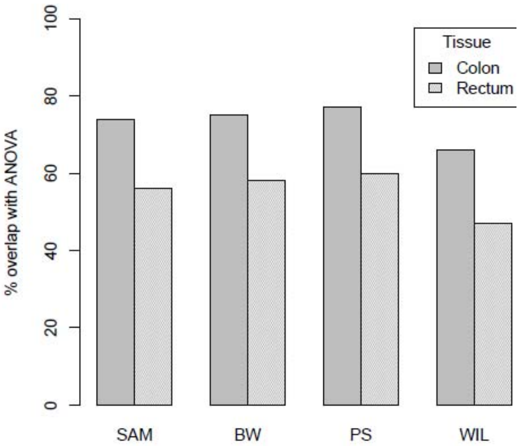
4 Classifications and class prediction

We propose five approaches to predict the disease status (class prediction) using gene expression levels from both the target and surrogate tissues. The first approach is to use the expression levels from the colon tissue for both feature selection and class prediction (T/T); the second approach is to use gene expression level from the rectum tissue for both the feature selection and class prediction (S/S); the third approach is to use the gene expression levels from the colon tissue for feature selection and gene expression level from the rectum tissue for class prediction (T/S), the fourth approach is to use gene expression levels from the rectum tissue for class prediction assuming known gene signature (A/S), whilst the fifth approach is to use gene expression levels from the colon tissue for class prediction assuming known gene signature (A/T). The main difference between the last two and other approaches is that the feature selection step is not required since the relevant features are assumed to be known. To minimise overfitting and over optimistic results, we consider 1000 re-sampled dataset, in which at each re-sampling, the dataset was divided into learning and testing set. The learning set consists of 2/3 of the patients randomly selected from the data, taking into account their disease status. The remaining 1/3 serves as the testing set. Four different classification algorithms with four different feature selection methods were used to estimate the misclassification errors. The feature selection methods used are: Wilcoxon rank sum test, Significant Analysis of Microarray (SAM, Tusher et al., 2001), the Between-Within ratio (BW, Dudoit et al., 2002) and the Prediction Strength (PS, Xiong et al., 2001). The gene selection methods were used in combination with the following classification methods: Diagonal Linear Discriminant Analysis (DLDA), Random Forests (RF; Breiman, 2001) and Support Vector Machines (SVM; Cortes and Vapnik, 1995; Furey et al., 2000). The support vector machines were used with radial kernel (SVMR) and the linear kernel (SVML). Note that for each combination of the classification algorithms and feature selection methods, different subsets of genes were used. Specifically, the top p genes are selected for classification with p = 10, 20, 30, 40, 50, 60, 70, 80, 90, 100. The average misclassification errors for each of the 4 × 4 × 10 combinations of feature selection methods, classification methods and the top p genes were based on 1000 re-sampled datasets.

4.1 Application to the data

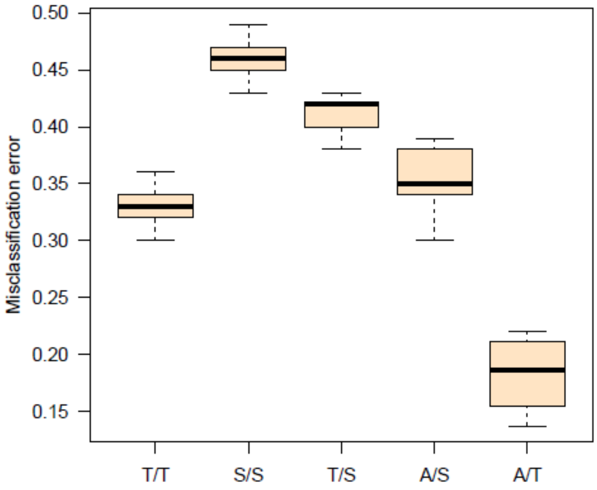
We compared the feature selection methods with the joint ANOVA model by investigating the percentage of genes in the Top 100 that are common between the two approaches. Figure 2 shows that percentages of the overlapping genes were higher for the colon tissue than the rectum tissue. Also, the percentages of the overlapping Top 100 genes were comparable among the feature selection methods, except for the Wilcoxon test. We favoured the feature selection methods because the joint ANOVA model violate the principle of cross-validation since it is based on all the dataset (i.e. both learning and testing sets) and could consequently suffers from over-fitting. It is also more computationally intensive to repeatedly perform it on 1000 re-sampled datasets.

Figure 2 The percentage of overlapping top 100 genes from the feature selection methods: (Significant Analysis of Microarrays: SAM, Between-Within ratio: BW; Prediction Strength: PS; and Wilcoxon test: WIL) and the joint ANOVA model



Ruschhaupt et al. (2004) proposed that classification algorithms should be disentangled from the biology in order to establish what amount of the observed discrimination can be attributed to biological differences. Figure 3 presents an aggregated results for the different approaches considered for class predictions by averaging out both the effects of feature selection, classification methods and the top *p* genes.

Figure 3 Overview of the classification results in terms of misclassification error (see online version for colours)



Notes: T/T denotes feature selection and classification based on the target tissue, S/S denotes feature selection and classification based on the surrogate tissue, T/S denotes feature selection based on the target tissue and classification based on the surrogate tissue, A/S denotes the biologically relevant genes from Aerssens et al. (2008) with classification for expression from the surrogate tissue, A/T denotes the biologically relevant genes from Aerssens et al. (2008) with classification for expression from the target tissue.

The average misclassification error when both feature selection and classification were based on the colon tissue is 33%, the average misclassification errors when both the feature selection and classification were based on the rectum tissue is 46%. Note that the misclassification error based on expression level from the rectum tissue improves to 41% when feature selection was based on the colon tissue and 35% when the features were assumed known based on Aerssens et al. (2008). The misclassification error for the colon tissue was improved to 18% when the features were assumed to be known. This finding suggests that an overall significant improvement in the predictive power of gene expression from the colon and the rectum tissue can be obtained if relevant set of genes are used for classification. Note that the large variation shown for A/S and A/T is because the aggregated misclassification errors were based on four values resulting from the four classification algorithms since the features were assumed to be the known 25 IBS associated gene signatures from Aerssens et al. (2008). The complete tables of misclassification errors for T/T, S/S and T/S based on each combination of feature selection method, classification algorithms and the top *p* genes are provided in the supplementary web appendix.

A formal testing for the difference in misclassification errors between the different approaches considered for class predictions are presented in Table 1. The analysis was done by using fixed effects model with correlated errors, which accounts for dependency of misclassification errors on feature selection and classification methods. There is a significant reduction in misclassification error rate when expression levels from the rectum tissues for the known gene signatures were used for class prediction (A/S) in comparison to when feature selection and classification were based solely on the rectum tissue (S/S). It is also interesting to see that there is a significant reduction in average misclassification error rate when the expression levels from the colon tissue were used for feature selection, but expression levels from the rectum tissue were used for class prediction (T/S). As expected, using known features also improved the predictive power of the colon tissue (A/T).

Table 1 Estimated differences in misclassification error rate are shown compared to the misclassification error rate when developing and applying the signature on the surrogate rectum tissue (S/S)

Tissue used for features selection	*Tissue used for prediction*	*Estimated difference to the reference (0.46)*	*Std.Err*	*P-values*
Colon	Colon	–0.13	0.002	<.0001
Colon	Rectum	–0.05	0.002	<.0001
Aerssens et al.	Rectum	–0.11	0.009	<.0001
Aerssens et al.	Colon	–0.28	0.008	<.0001

Notes: S/S is used here as the reference with an average misclassification error of 0.46 with a standard error of 0.002.

5 Discussions

The discovery of molecular biomarkers has greatly advanced through the emergence of microarray technology that simultaneously measures expression levels of thousands of genes. Such molecular biomarkers are generally investigated in one particular tissue, thereby missing potential insights into the disease effects in other tissues. As a case study, this paper used data from gene expression profiling of the colon and rectum tissues from IBS patients and healthy controls

First, a gene-specific joint model was formulated to identify differentially expressed genes between IBS patients and healthy controls, whilst accounting for both the within- and between-tissue correlation of each gene. Analysis of the colon tissue resulted in identification of genes that were differentially expressed between the IBS patients and the healthy controls, whereas a similar analysis on the rectum tissue failed to produce differentially expressed genes at the FDR of 5%. A possible conclusion would therefore be that the IBS does not affect gene expression in the surrogate tissue. Lack of differentially expressed genes from the colon tissue after FDR correction may be due to the fact that the gene expression level from the rectum tissue were more subtle than the expression levels from the colon.

Although, no gene was found to be differentially expressed from the rectum tissue after FDR correction, a good discrimination between the IBS patients and healthy controls was obtained when known gene signature was used. As observed by Van Sanden et al. (2007, 2008) the misclassification errors depend on the gene selection methods, classification methods and the number of genes used. Prediction using the rectum tissue was inaccurate when the signatures were derived from the rectum tissue. The gene signatures derived from the colon tissue however worked remarkably better in the rectum tissue. Even more improved results were obtained when known disease-related genes were used for classification based on the expression profiles of the colon and rectum tissues. It has therefore become increasingly clear that the challenge of analysing microarray data is to discover subtle signals in a highly dimensional and noisy dataset.

References

Achiron, A. and Gurevich, M. (2006) 'Peripheral blood gene expression signature mirrors central nervous system disease: the model of multiple sclerosis', *Autoimmunity Reviews*, Vol. 5, No. 8, pp.517–522.

Aerssens, J., Camilleri, M., Talloen W., Thielemans, L., Goehlmann, W.H.H., Van De Den Wyngaert, I., Thielemans, T., De Hooght, R., Andrews, N.C., Bharucha, E.A., Carlson, J.P., Busciglio, I., Burton, D.D., Smyrk, T., Urrutia, R. and Coulie, B. (2008) 'Alteration in mucosal immunity identified in the colon of patients with irritable bowel syndrome', *Gastroenterology and Hepatology*, Vol. 6, No. 2, pp.194–205.

Alon, U., Barki, N., Notterman, D., Gish, K., Mach, S. and Levine, J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal target tissues probed by oligonucleotide arrays, PNAS, 86, 6745–6750.

Asselah, T., Bieche, I., Narguet, S., Sabbagh, A., Laurendeau, L., Ripault, M.P., Boyer, N., Martinot-Peignoux, M., Valla, D., Vidaud, M. and Marcellin, P. (2008) 'Liver gene expression signature to predict response to pegylated interferon plus ribavirin combination therapy in patients with chronic hepatitis C', *Gut*, Vol. 57, No. 4, pp.516–524.

Breiman, L. (1996) 'Bagging predictors', *Random forests. Machine Learning*, Vol. 24, No. 2, pp.123–140.

Cai, Y.D, Huang, T, Feng, K-Y., Hu, L. and Xie, L. (2010) 'A unified 35-gene signature for both subtype classification and survival prediction in diffuse large B-Cell lymphomas', *PLOS ONE*, Vol. 5, No. 9, p.e12726

Cortes, S.S. and Vapnik, V. (1995) 'Support-vector networks', *Machine Learning*, Vol. 20, No. 3, pp.273–279.

Debouck, C. and Goodfellow, P.N. (1999) 'DNA microarrays in drug discovery and development', *Nature Genetic*, Vol. 21, No. 1, pp.48–50.

Dudoit, S., Fridlyand, J. and Speed, T.P. (2002) 'Comparison of discrimination methods for the classification of tumors using gene expression data', *Journal of the American Statistical Association*, Vol. 97, No. 458, pp.77–87.

Furey, T.S., Cristianini, N., Dufy, n., Bednarski, D.W., Schummer, M. and Haussler, D. (2000) 'Support vector machine classification and validation of cancer tissue samples using microarray expression data', *Bioinformatics*, Vol. 16, No. 10, pp.906–914.

Golub, T.R., Slonin, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S. (1999) 'Molecular classification of cancer: class discovery and class prediction by gene expression monitoring', *Science*, Vol. 286, No. 5439, pp.531–537.

Huang, C., Chen, H., Cassidy, W. and Howell, C.D. (2008a) 'Peripheral blood gene expression profile associated with sustained virologic response after peginterferon plus ribavirin therapy for chronic hepatitis-C genotype 1', *Journal of the National Medical Association*, Vol. 100, No. 12, pp.1425–1433.

Huang, T., Tu, K., Shyr, Y., Wei, C.C., Xie, L. and Li, Y.X. (2008b) 'The prediction of interferon treatment effects based on time series microarray gene expression profiles', *Journal of Translational Medicine*, Vol. 6, p.44.

Le-Nicilescu, H., Kurian, S.M., Yehyawi, N., Dike, C., Patel, S.D., Edenberg, H.J., Tsuang, M.T., Salomon, D.R, Nurnberger Jr., J.I. and Niculescu, A.B. (2008) 'Identifying blood biomarkers for mood disorders using convergent functional genomics', *Molecular Psychiatry*, Vol. 14, No. 2, pp.156–174.

Ross, D.T., Scherf, U., Eissen, M.B., Perou, C.M., Rees, C., Spellman, P., Iyer, V., Jefrey, S.S., Van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J.C., Lashkari, D., Shalon, D., Myers, T.G., Weinstein, J.N., Botstein, D. and Brown, P.O. (2008) 'Systematic variation in gene expression patterns in human cancer cell lines', *Nature Genetics*, Vol. 24, No. 3, pp.227–235.

Ruschhaupt, M., Huber, W., Poustka, A. and Mansmann, U. (2004) 'A compendium to ensure computational reproducibility in high-dimensional classification tasks', *Statistical Applications in Genetics and Molecular Biology*, Vol. 3, No. 1.

Tusher, V.G., Tibshirani, R. and Chu, G. (2001) 'Significance analysis of microarrays applied to the ionizing radiation response', *Proceedings of the National Academy of Sciences*, Vol. 98, No. 9, pp.5116–5121.

Van Sanden, S., Lin, D. and Burzykowski, T. (2007) 'Performance of classification methods in a microarray setting: a simulation study', *Biocybernetics and Biomedical Engineering*, Vol. 27, No. 3, pp.15–28.

Van Sanden, S., Lin, D. and Burzykowski, T. (2008) 'Performance of gene selection and classification methods in a microarray setting: a simulation study', *Communications in Statistics - Simulation and Computation*, Vol. 37, No. 2, pp.409–424.

Warren, P., Bienkowska, J., Martini, P.G.V., Jackson, J. and Taylor, D.M. (2007) 'PANP - a new method of gene detection on oligonucleotide expression arrays', *Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering, BIBE'07*, 14–17 October, Boston, MA, USA, pp.108–115.

Wu, Z., Irizarry, R.A., Gentleman, R., Murillo, F.M. and Spencer, F. (2004) 'A model based background adjustment for oligonucleotide expression arrays', *Journal of the American Statistical Association*, Vol. 99, No. 468, pp.909–917.

Xiong, M., Li, W., Zhao, J., Jin, L. and Boerwinkel, E. (2001) 'Feature (gene) selection in gene expression-based tumor classification', *Molecular Genetics and Metabolism*, Vol. 73, No. 3, pp.239–247.

Xu, M., Li, W., James, G.M., Mehan, M.R. and Zhou, X.J. (2009) 'Automated multidimensional phenotypic profiling using large public microarray repositories', *Proceedings of the National Academy of Sciences of the USA*, Vol. 106, No. 30, pp.12323–12328.