Processing and analysis of serum antibody binding signals from Printed Glycan Arrays for diagnostic and prognostic applications

Marko I. Vuskovic*

Department of Computer Science,
College of Science San Diego State University,
San Diego, CA 92182, USA
E-mail: marko@cs.sdsu.edu
*Corresponding author

Hongyu Xu

Transaction Analytics Department,
Fair Issac Corporation,
San Diego, CA 92130, USA
E-mail: hongyuxu@fico.com

Nicolai V. Bovin

Carbohydrate Chemistry Laboratory,
Shemyakin Institute of Bioorganic Chemistry,
Russian Academy of Sciences,
117997 Moscow, Russia
E-mail: bovin@carb.ibch.ru

Harvey I. Pass and Margaret E. Huflejt

Department of Cardiothoracic Surgery,
New York University,
School of Medicine,
New York, 10016 NY, USA
E-mail: Harvey.Pass@nyumc.org
E-mail: Margaret.Huflejt@nyumc.org

Abstract: Procedures for data pre-processing, quality control, data analysis, evaluation and visualisation of the new high-throughput biomarker platform based on Printed Glycan Arrays (PGA) are presented in this paper. PGAs are similar in concept to DNA arrays but contain deposits of various carbohydrate structures (glycans) instead of spotted DNAs. PGA biomarker discovery for the early detection, diagnosis and prognosis of human malignancies and viral diseases is based on the response of the immune system as measured by the level of binding of anti-glycan antibodies from human serum to the glycans on the array. Procedures related to PGA data processing are herein demonstrated

in a pilot study of cases representing 50 sera from patients with malignant mesothelioma and a control sample of 65 sera from high risk subjects exposed to asbestos without symptoms of disease.

Keywords: PGA; printed glycan arrays; serum antibodies; diagnosis and prognosis of cancers; malignant mesothelioma; bioinformatics; immunoruler; molecular biomarker discovery.

Reference to this paper should be made as follows: Vuskovic, M.I., Xu, H., Bovin, N.V., Pass, H.I. and Huflejt, M.E. (2011) 'Processing and analysis of serum antibody binding signals from Printed Glycan Arrays for diagnostic and prognostic applications', *Int. J. Bioinformatics Research and Applications*, Vol. 7, No. 4, pp.402–426.

Biographical notes: Marko I. Vuskovic, DrSc is Professor of Computer Sciences at San Diego State University and Director of SDSU Robotics and Neural Networks Laboratory. He received his doctorate of technical sciences from the University of Zagreb in 1975 in control engineering and computer science. His research interests are in the area of pattern recognition and discovery of glycan-based biomarkers for early detection, diagnosis and prognosis of cancers.

Hongyu Xu, MSc is an analytical scientist at FICO, San Diego. She received her MSc Degree in Computer Science from San Diego State University in 2004. Her research interests are in the areas of computer programming and system analysis, machine learning, statistical modelling, feature selection and classification. She was an active developer of bioinformatics programs for quality control, preprocessing and analysis of data from Printed Glycan Arrays. Her recent work is in analysis and pattern recognition for payment card fraud detection.

Nicolai V. Bovin, DrSc is Professor and Head of Carbohydrate Chemistry Laboratory at Shemyakin Institute of Bioorganic Chemistry (IBCh), Moscow, Russian Federation. He received his MS Degree from the Department of Chemistry, Moscow State University, and PhD and DrSci Degrees from IBCh. In 1976, he joined the Laboratory of Carbohydrates and Glycoproteins led by Prof. A.Ya. Khorlin at IBCh and has started his work on oligosaccharide synthesis. His current research interests include carbohydrate/protein and carbohydrate/carbohydrate interaction; synthesis of oligosaccharides and neoglycoconjugates; supramolecular chemistry; glycoarrays, natural anti-carbohydrate antibodies; medicinal chemistry: influenza therapy and diagnostics, transplantation, oncodiagnostics and oncotargeting.

Harvey I. Pass is Professor of Surgery and Cardiothoracic Surgery for NYU Langone Medical Center and School of Medicine, and Director of the Thoracic Oncology for the NYU Clinical Cancer Institute. His laboratory is the home of the NCI-funded Early Detection Research Network Biomarker Discovery Laboratory for Mesothelioma, and the Clinical CORE for the NCI's recently funded Mesothelioma Pathogenesis Program Project. He has developed one of the first and largest organised, prospective tissue archives with accurate matching demographics, and has used this resource for ongoing collaborative studies of the molecular biology of thoracic malignancy.

Margaret E. Huflejt PhD Assistant Professor, Department of Cardiothoracic Surgery, New York University, School of Medicine. She received the PhD in biochemistry from the University of California at Berkeley, and specialised in protein chemistry and glycobiology at the University of California at San

Francisco. Her main research interests include pathology of malignant transformation, glycobiology-based biomarkers for early detection of cancer and cancer risk and therapeutic applications of carbohydrate-binding antibodies. She leads the Tumor Glycome Group of the Bellevue Thoracic Surgery Laboratory of NYU SoM led by Dr. Harvey Pass, where she continues developments in diagnostic glyco-immunology and glyco-genomics.

1 Introduction

The early detection and diagnosis of cancers in their preclinical state before the disease exhibits symptoms is essential for successful treatment with existing therapeutic approaches, such as surgery, radiation, and chemotherapy. In addition, a quantitative and reliable prognostic measurement of cancer progression is vital for successful disease management and patient stratification (Trademark Publications, 2008). For these reasons the development of effective biomarkers for cancer detection, diagnosis and prognosis has become the ultimate goal of many biomedical researchers in academia, clinical institutions and the diagnostic industry. As a result, the last two decades have concentrated on a variety of molecular biomarkers which can roughly be classified into two major platforms: nucleic acid biomarkers (Sidranski, 1997; Brown and Botstein, 1999) and protein biomarkers (Hutchens and Yip, 1993; Wright, 2002; Issaq et al., 2002).

These platforms are based on identifying the expressed genes and proteins in cancer cells in human tissue or body fluids. Although these platforms have gained considerable attention but their adoption into standard clinical practice is limited by the:

- high cost associated with the technology
- time required for test procedure
- narrow targeting of the test to a particular disease, e.g., cancer type
- substantial variability due to non-homogeneity of tissue samples, rapid degradation of tissue samples between sampling and hybridisation and very small depositions of DNA on the chip.

In the last five years a new biomarker platform has emerged based on glycan arrays (Bovin and Huflejt, 2008), that challenges the limitations of the nucleic acid and protein based platforms. The Printed Glycan Arrays (PGA) are similar to DNA microarrays, but contain deposits of various carbohydrate structures (glycans) instead of spotted DNAs. Most of these glycans can be found on the surfaces of normal human cells, human cancer cells, and on the surfaces of many human infectious agents such as bacteria, viruses, and other pathogenic microorganisms. Transformation of cells from healthy to pre-malignant and malignant is associated with the appearance of abnormal glycosylation on proteins and lipids presented on the surface of these cells. The malignancy-related abnormal glycans are called tumor-associated carbohydrate antigens (TACA), (Hakomori, 2002). There is growing evidence (Aarnoudse et al., 2006) that numerous TACAs are immunogenic, and that the human immune system can generate antibodies against them. Since multiple glycans arrayed on PGAs are either known TACAs or closely related structures, the antibodies present in human sera that bind to glycans on PGA can indicate

the status of response of the immune system to human malignancies (Huflejt et al., 2005a, 2005b). A prototype of PGA with a library of 200 glycan structures was built at Scripps Research Institute, La Jolla, California, under the auspices of the Consortium of Functional Glycomics (CFG) (Blixt et al., 2004). Further development of the PGA with 211 glycans was conducted at Cellexicon, Inc., La Jolla, in collaboration with Shemyakin Institute of Bioorganic Chemistry, of the Russian Academy of Sciences, Moscow, Russia. The "second generation" of PGA was used in several pilot studies sponsored by the National Cancer Institute. Research and improvement of PGA technology and its relevance in diagnostic and prognostic applications is currently continuing in the Glyco-Medical Group of the Thoracic Surgical Laboratory at the New York University, School of Medicine, in collaboration with the Shemyakin Institute. The new PGA chip has over 300 carbohydrate structures, most of which were selected and synthesised based on previous research experiences in various pilot studies (Huflejt et al. 2005a, 2005b, 2005c; Arun et al., 2007), which extend to breast, ovarian, and lung cancers and malignant mesothelioma.

The focus of this paper is to present methods and approaches for the processing and analysis of PGA data used in the above mentioned pilot studies, and in the studies that will follow. The paper briefly describes the principle of PGAs, i.e., the procedure of measuring the level of binding of human antibodies against glycans on the array, the basic procedures for data preparation, pre-processing and quality control, the procedures for diagnostic data analysis and evaluation, and finally the procedures for prognostic data analysis and evaluation. All these approaches are demonstrated on data obtained from malignant mesothelioma sera archived by one of the co-authors (HIP). The serum samples contained 65 high risk subjects exposed to asbestos and 50 subjects diagnosed with malignant mesothelioma. The performance of chosen parameters and methods was evaluated through cross-validation and bootstrapping procedures, discussed later in this paper.

2 Printed Glycan Arrays

A Printed Glycan Array (PGA) consists of a glass slide coated with a chemically reactive surface on which various glycan structures are covalently attached using standard microarray contact printing technology. A PGA slide contains several sub-arrays of the entire, presently available glycan library in form of microscopic glycan deposits of size 50–100 microns. The version of the PGA used to generate data presented in this paper has two concentrations of glycans (10 and 50 μM) and eight replicates for each concentration, thus resulting in an array of 16 sub-arrays, each containing 211 deposits of different glycan structures, and biotin spots used as a printing control.

The measurement of binding of human Anti-Glycan Antibodies (AGA) to arrayed glycans is achieved as described in Huflejt et al. (2009). Briefly, the PGA slide is first incubated with the subject's serum, allowing the binding of serum antibodies to glycans in PGA deposits. Serum IgG, IgM and IgA immunoglobulins bound to printed glycans are visualised simultaneously with the 'combo' biotinylated secondary goat anti human IgG, IgM and IgA antibodies (Pierce Biotechnology, Inc., Rockford, IL), and streptavidin-Alexa^{555} (Invitrogen/Molecular Probes, Carlsbad, CA). Fluorescence signal intensities that correspond to antibodies bound to printed glycans are scanned

at 90% laser power, and quantified with ImaGene software (BioDiscovery, Inc., El Segundo, CA). The total relative fluorescence signal intensity values (appx. range: 1000–32,000,000 Relative Fluorescence Units) are used for further data processing and analyses.

Figure 1 provides an example of an image obtained by the laser scanner. The figure is an excerpt from the slide presenting only one sub-array. The process of complete PGA processing, beginning with slide printing and ending with data analysis is shown in the block diagram in Figure 2. The rest of this paper will concentrate on the blocks on the right-hand side of the diagram.

Figure 1 Excerpt from a developed PGA shows one of 2×8 replicate sub arrays with fluorescent intensity spots which correspond to the library of 211 different glycan structures

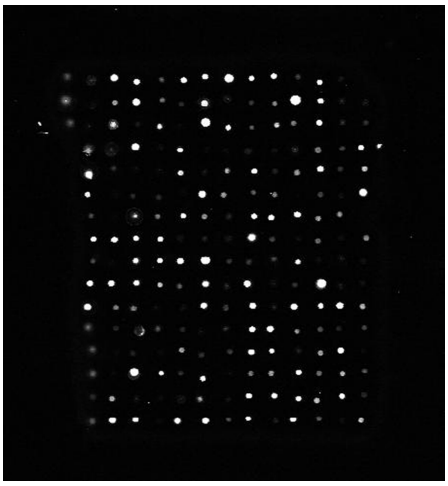
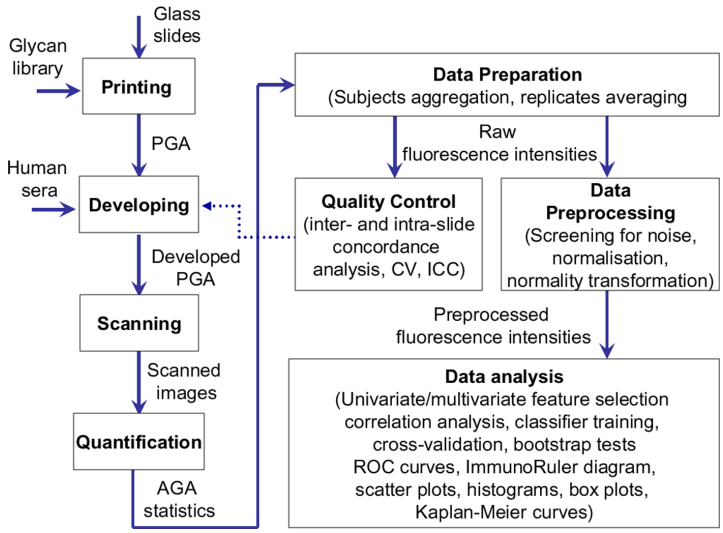


Figure 2 Steps in processing of PGAs and subsequent data analysis (see online version for colours)



3 Data preparation and quality control

Quantification of scanned PGA images is performed separately for each developed slide, i.e., subject. The quantification also includes the first step of quality control: images are visually examined; images with poor quality (high background noise, 'scratches', 'clouds' and irregular shapes of spots, such as a donut-shape and 'bleeding' spots) are rejected; and the development of the sera for the same subject is repeated. Data files obtained by quantification of all successfully developed slides are then aggregated into a single data file which basically contains data matrices of total fluorescence intensities for two concentrations, 10 μM and 50 μM. We have used total intensities instead of mean intensities since the former gives a more adequate measure of the binding level of AGA. The use of total intensities is justified by the fact that the deposition of glycans on PGA chips is very regular, which is tested and verified by examining the 'salt images' of all glycan deposits on each PGA slide that were scanned immediately following the printing.

The fluorescence intensities of bound antibodies for each glycan on each slide are summarised by computing the median across all corresponding replicates. This method of summarisation is more robust in terms of outliers than the mean. The result of the replicate summarisation is n by d raw predictor matrix $\boldsymbol{X}_{\text{raw}}$, where n is number of subjects ($n = 65 + 50$) and d is number of glycans on the PGA library ($d = 211$).

The second step in quality control is the inter-slide quality control, which tests the reproducibility of data obtained for the same subject (same serum) but developed and quantified at different times and on PGA slides from different print batches. In order to quantify this type of reproducibility we used Lin's concordance correlation coefficient (Lin, 1989). Lin's CCC is used instead of the traditional Pearson correlation coefficient as the latter fails to capture differences due to linear bias in scale and location. Since this approach could not have been applied to all subjects due to the cost of slide development, it has been applied to a fraction of randomly selected subjects in control and case samples, whose serum was developed in two different days and on two different batches of PGA slides. The testing of the printed slide batches was however done on a regular basis as a part of the standard operating procedure. For that purpose we have used a serum of two benchmark subjects, which was incubated with each new print batch.

In addition to inter-slide QC we have also performed intra-slide QC that tests the reproducibility of data within R replicated sub-arrays on each slide ($R = 8$). For this purpose we have used the overall concordance correlation coefficient (Barnhart et al., 2002). The OCCC is a generalisation of Lin's CCC through addressing several sources instead of only two. Slides which have OCCC < 0.9 are rejected and the development of the same serum is repeated until a satisfactory level of intra-slide concordance is achieved.

4 Data pre-processing

The data that pass the quality control analyses are generally still noisy and not amenable for direct use in diagnostic and prognostic analysis. Pre-processing steps are thus performed including noise screening, normalisation and normality transformation.

4.1 Noise screening

Noise screening implies cleaning of data from those variables (glycans) that have evidently noisy behaviour and are deemed to be unreliable. There are several sources of noise but we will restrict ourselves to measurement noise that can be assessed through the variation of replicated intensities. The noise screening has involved rejection of those glycans associated with fluorescence intensities that have manifested noisy behaviour consistently for all observations (subjects). We have used three measures as rejection criteria: intensities below the noise threshold, high value of Coefficient of Variation (CV) of replicates, and low Interclass Correlation Coefficient (ICC).

In addition to noise screening, we have also removed 'redundant glycans'. The redundancy of glycans is determined by the Pearson cross-correlation coefficient computed for fluorescence intensities across all observations, for all combinations of glycan pairs.

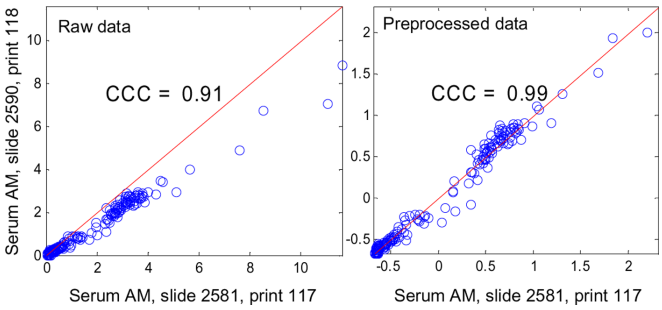
Finally, control signals and signals that correspond to control glycans and non-glycan spots (e.g., biotin spots) are also removed from future diagnostic and prognostic analysis.

4.2 Data normalisation and transformation

The main goal of data normalisation is to reduce the systematic per-slide bias in scale and location. The probe bias was not essential in this study since the quality of printed spots was controlled by examining the salt images mentioned above. We have used intra-array linear normalisation. This normalisation method can remove linear bias with negligible damage to discriminatory information since most of the glycans on the chip (left after noise screening) are non-discriminatory, i.e., class invariant.

Data transformations are applied after normalisation, primarily to shorten the distribution tails. We have used a Box-Cox power transform extended for negative arguments in order to handle normalised data values (John and Draper, 1980). The transformation parameter λ is set to 0.2 for all glycans. The value was found to be overall optimal after extensive experimentation with real and simulated data. The effect of normalisation and transformation is illustrated in Figure 3, which shows that the inter-slide concordance correlation coefficient for a benchmark subject was increased from 0.91 to 0.99 after data normalisation and transformation.

Figure 3 Concordance plots for two slides obtained from the serum of the same subject developed on PGA slides from two different print batches (see online version for colours)



5 Diagnostic data analysis

The goal of diagnostic data analysis is to identify a set of features (glycans) and to specify a classification algorithm which effectively enough discriminates between the control and case samples of a given training dataset. The identified set of features and the specified classifier will hopefully be able to accurately classify new observations with unknown class membership. The prediction accuracy of such a classifier can be estimated with various cross-validation and bootstrapping techniques.

5.1 Univariate feature selection

The first step in diagnostic data analysis is to evaluate the discriminatory ability of each glycan separately, which is also known as univariate feature selection, or ranking. Since the data at hand are sampled from an unknown distribution, it is appropriate to use one of the standard non-parametric approaches. We have used Wilcoxon-Mann-Whitney (WMW) ranking in which the ranking of *p*-values is coincident with the ranking by the area under the ROC curve (AUC). The individual ranking of glycans for the mesothelioma assay applied to pre-processed data is shown in Table 1 in which the nine top ranked glycans are represented by their Glycan Identification Number (GID) and by their carbohydrate structure. The corresponding distributions of case and control samples are shown in Figure 4. This figure illustrates the complexity of the discrimination problem: almost all distributions for control and for case samples display multiple modes (at least bimodality) and a relatively high overlap between samples, which is the cause of the relatively low individual AUC values.

Figure 4 Distributions of the case (dotted line) and control sample (solid line) of the top nine glycans listed in Table 1 (see online version for colours)

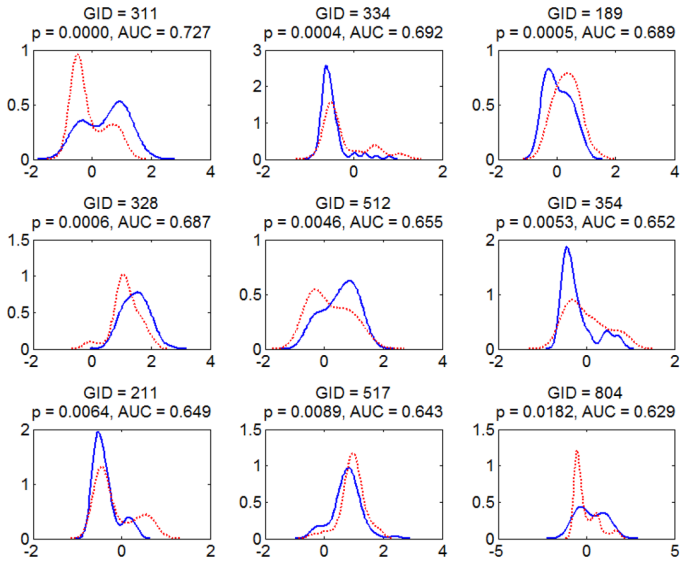


Table 1 Top 9 glycans ranked by Wilcoxon-Mann-Whitney rank sum test in mesothelioma study (65 asbestos exposed, 50 malignant mesothelioma)

Rank	*GID*	*Glycan structure*	*p-value*	*AUC*
1	311	Neu5Acα2-3Galβ1-4Glcβ-sp	0.00003	0.7274
2	334	$(Neu5Ac\alpha2\text{-}8)_3$-sp	0.00043	0.6923
3	189	GlcNAcβ1-6GalNAcα-sp	0.00054	0.6889
4	328	GlcNAcβ1-4(GlcNAcβ1-6) GalNAcα-sp	0.00062	0.6868
5	512	Galβ1-3GlcNAcβ1-3Galβ1-4Glcβ-sp	0.00458	0.6548
6	354	Galβ1-4GlcNAcβ1-6GalNAcα-sp	0.00527	0.6523
7	211	Man1-4GlcNAcβ-sp	0.00638	0.6489
8	517	Galα1-4GlcNAcβ1-3Galβ1-4GlcNAcβ-sp	0.00892	0.6428
9	804	Fucα1-2Galβ1-3(Fucα1-4)GlcNAcβ1-4Galβ1-4Glcβ-sp	0.01822	0.6289

The fact that the individual AUC values are relatively small suggests a necessity for conjunction of the discriminatory information associated with several glycans. A common way to do this is by linear combination (projection) of intensities associated with the top selected glycans:

$$z_i = \boldsymbol{x}_i \boldsymbol{w}, \tag{1}$$

where z_i is the projected intensity for subject i, $\boldsymbol{w} = (w_1, w_2, ..., w_m)^T$ is the projection vector, while $\boldsymbol{x}_i = (x_{ij_1}, x_{ij_2}, ..., x_{ij_m})$ is the row vector of pre-processed fluorescent intensities for m selected glycans. The projection (1) is the basic idea of all linear binary classifiers such as logistic and linear regression, Fisher Linear Discriminant (FLD) or Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), single-layer artificial neural networks, etc., where the class membership of an unlabelled observation z_x is determined by checking the sign of $z_x + w_o$, where w_o is the classification decision point.

After extensive experimentation with real and simulated data we have chosen Multiple Logistic Regression (MLR) as the most appropriate and reasonably efficient projection method for the mesothelioma study.

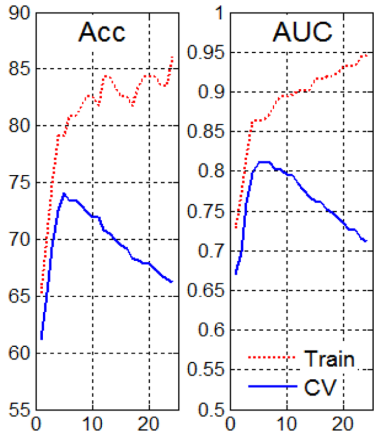
5.2 Cross validation

Once the feature selection and projection method are defined a question remains as to how many features should be used in projection (1). A larger number of variables generally results in better training performance but inevitably leads to over-fitting. The standard approach to address this issue is cross-validation. The results for repeated 10-fold cross-validation are shown in Figure 5. Here we used two performance measures: the accuracy (Acc) and the area under the ROC curve (AUC), which are both functions of projected values $\mathbf{z} = (z_1, z_2, ..., z_n)$ and their class labels $\mathbf{y} = (y_1, y_2, ..., y_n)$, $y_i \in (1, 2)$. The latter performance measure is particularly suitable since it is insensitive to sample imbalance (Fawcett, 2003) and is independent from the choice of decision point w_o, thus covering the performance of a family of classifiers (Hanley and McNeil, 1982; Bradley, 1997). In addition, the AUC is more discriminating than Acc, since it has better

resolution than Acc (Ling et al., 2003). Finally, the AUC has ranking ability, which is an important notion even more fundamental than classification (Flach, 2004; Hand and Till, 2001).

In order to minimise the prediction bias and variability we have used balanced, unbiased repeated 10-fold cross-validation. 'Balanced' refers here to the fact that each fold has the same class distribution as the original training samples. 'Unbiased' refers to the minimisation of the bias due to variable selection by embedding the feature selection into the cross-validation loop (Ambroise and McLachlan, 2002; Simon et al., 2003). The computation of the cross-validated performance measures was carried out by averaging across the folds, instead of pooling, which further reduces the stratification bias (Parker et al., 2007). The number of cross-validation repeats was 100. Figure 5 shows the training and cross-validated performance measures for the number of features that range from $m = 1$ to $m = 24$. The optimal number of features is $m = 5$ for both measures. Clearly over-fitting takes place for $m > 5$. The achieved cross-validated performance measures for 5 features are Acc = 74.1% and AUC = 0.811, while the corresponding training (observed) values for the same number of features are $\text{Acc}_O = 79.1\%$ and $\text{AUC}_O = 0.864$ respectively.

Figure 5 Accuracy of classification and AUC value for combined top WMW-ranked glycans using multiple logistic regression. The diagram shows the training (dotted line) and the repeated 10-fold cross-validated (solid line) accuracy and AUC value for various sizes of glycan sets ranging from one to 24 (see online version for colours)



5.3 Compound feature selection and adjusted ROC curve

Once the optimal number of features has been determined, we would like to establish some kind of ranking of their importance based on an experiment similar to cross-validation instead of the 'observed' ranking given in Table 1. For this we propose the Compound Feature Selection (CFS) algorithm which performs feature selection (in this case the univariate WMW ranking) on many randomly selected subsets of control and case samples, known as a repeated balanced hold-out procedure. The results are compounded and presented in a frequency diagram of presence of features in sets of m selected features in each hold-out iteration. The result for the mesothelioma assay, for $m = 5$ and $B = 1000$ hold-out iterations with balanced hold-out samples of size

$n_1 = n_2 = 35$ is shown in Figure 6. The ranking of features with the univariate CFS algorithm coincides with the WMW ranking: GID = 311, 334, 189, 328, 512.

Figure 6 Compound feature selection for feature sets of size $m = 5$, obtained with 1000 balanced hold-out iterations with equal hold-out samples of 35 subjects. The numbers on the horizontal axis represent column indices of predictor matrix, while the stem numbers correspond to most frequent GIDs. Feature selection is performed by univariate WMW ranking (see online version for colours)

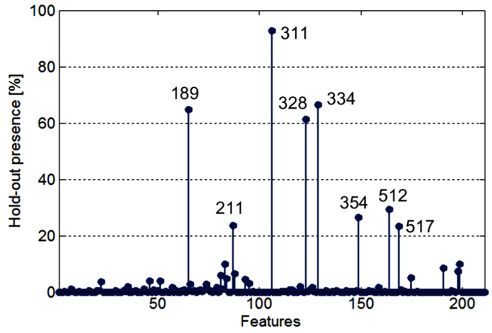
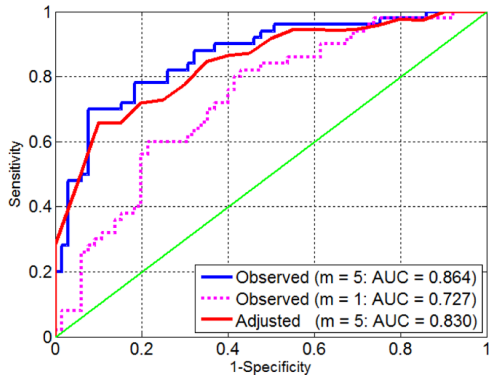


Figure 7 shows the observed ROC curves obtained for the combined top five compound glycans and for a single, top-ranked glycan GID = 311. The figure also shows the adjusted ROC curve[1] obtained by 1000 hold-out iterations with 30–70% split between the test and training samples.

Figure 7 ROC diagram for the mesothelioma assay obtained for top CFS-ranked glycans combined by multiple logistic regression (top solid line). The diagram also shows ROC curve for the single top ranked glycan (dotted line), and the adjusted ROC curve (middle solid line) (see online version for colours)



The procedure for adjusted ROC curve is implemented as follows:

1 Compute the observed ROC curve $s_n = \mathrm{ROC}_O\,(fpr)$ - sensitivity as function of false predictive rate

2 Perform B hold-out iterations. In each iteration do the following:

   a randomly split the data into test and training sets in balanced proportion

   b perform feature selection and projection based on sampled training subset

c derive the training ROC curve $\mathrm{ROC}_T(\mathrm{fpr})$

d derive the validation ROC curve ROC_V based on sampled test subset

e find the curve difference $\Delta(\mathrm{fpr}) = \mathrm{ROC}_\mathrm{T}(\mathrm{fpr}) - \mathrm{ROC}_\mathrm{V}(\mathrm{fpr})$.

3 Find the average of differences across all iterations $\Delta_{\mathrm{avr}}(\mathrm{fpr}) = \mathrm{mean}(\Delta)$

4 Adjust the ROC curve: $\mathrm{ROC}_\mathrm{A}(\mathrm{fpr}) = \mathrm{ROC}_\mathrm{O}(\mathrm{fpr}) - \Delta_{\mathrm{avr}}(\mathrm{fpr})$.

This algorithm is used to derive the estimated predictive ROC curve and AUC value that accounts for feature selection bias. The resulting AUC value is slightly higher than the value obtained by repeated 10-fold cross-validation in Figure 5.

5.4 ImmunoRuler

After the optimal features are selected and the projection vector determined for a given set of training data, it would be useful to visualise the ranking of the results by using some standard measure – the risk score. Therefore we propose a novel visualisation method which we name 'ImmunoRuler' and which presents the risk scores of all subjects in the training set, in an organised graphical way. Such a diagram obtained for the mesothelioma assay is shown in Figure 8. The risk scores are defined as:

$$r_i = \frac{1}{1+\exp(-z_i - w_o)}, \tag{2}$$

where z_i is the projection (1), while w_o is the classification decision point. The risk scores are sorted in ascending order, separately for controls (bars 1 to 65) and cases (bars 67 to 115). The two groups of bars are painted in different colours, and each colour has two shades to indicate quartile ranges. In the case in which the vector $\boldsymbol{w}$ and scalar w_o are estimated by logistic regression, the risk scores r_i can be interpreted as conditional probabilities of belonging to the case sample, and the decision point for classification would be line of equal odds $r_o = 0.5$. The position of the classification decision line can be modified depending on the diagnostic application, which can demand higher specificity or higher sensitivity. A common way is to consider the loss due to misclassification (Adams and Hand, 1999):

$$L = \pi_1 f_1 C_1 + \pi_2 f_2 C_2, \tag{3}$$

where π_k, f_k and C_k are probability of belonging to class k, probability of misclassifying class k, and the cost of misclassifying class k respectively. With the known priors, this translates into:

$$nL = n_1(1-s_p)C_1 + n_2(1-s_n)C_2. \tag{4}$$

If we use the desired ratio between costs of misclassifying controls and cases, $\gamma = C_1 / C_2$, then the corrected value of classification decision point based on minimal loss can be determined by the following maximisation procedure:

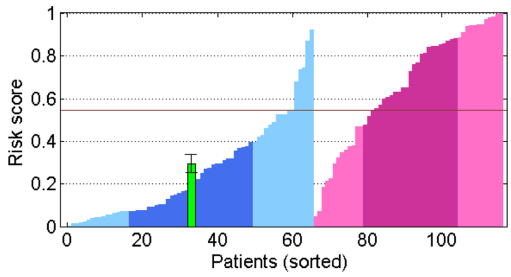
$$w_c = \arg\max_t \left[\gamma n_1 s_p(t) + n_2 s_n(t)\right], \tag{5}$$

where t is the varying decision point which is used to compute sensitivity and specificity. The corrected decision line on the ImmunoRuler diagram becomes:

$$r_c = \frac{1}{1+\exp(w_o - w_c)}. \tag{6}$$

The corrected decision line $r_c = 0.546$ in Figure 8 is determined for equal cost of misclassification of controls and cases, $\gamma = 1$. The corresponding accuracy is 82.6%, specificity 92.3% and sensitivity 70%.

Figure 8 The ImmunoRuler diagram of risk scores obtained for 65 controls (on the left) and 50 cases (on the right). The bar graphs are sorted by the ascending order of risk scores and are painted in darker shade for scores within quartile ranges. The immuno-ruler diagram can be used to classify unlabelled subjects (bar with whiskers) by comparing it with the decision threshold (see online version for colours)



The ImmunoRuler can be used to classify a new, unlabelled subject by computing his/her risk score using the features and the projection vector obtained during the ImmunoRuler training phase, and by projecting the risk score on the background of the trained ImmunoRuler (the wider bar with solid edges). This bar can be plotted with whiskers which indicate standard deviation or MAD of replicates, propagated all the way from pre-processing phase to the projection.

5.5 Significance of observed AUC

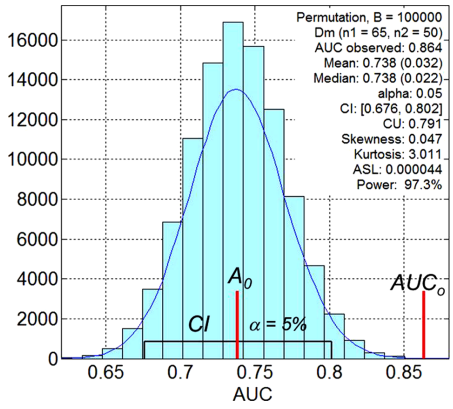
The statistical significance of the observed AUC value can be tested by the nonparametric bootstrap (Efron and Tibshirani, 1993). In order to estimate the distribution of the AUC value under the null hypothesis that the control and case samples were drawn from the same population, we have computed the AUC value for 100,000 randomly permuted training samples. The permutation bootstrap is generally less biased than the bootstrap with replacement. The computation of the AUC replications included feature selection in order to minimise the bias due to variable selection. The resulting empirical distribution is shown in Figure 9. As seen the distribution is very close to a normal distribution, with skewness 0.047 and kurtosis 3.01. The mean value and the standard deviation of the distribution are $A_0 = 0.738$ and $\sigma_0 = 0.032$ respectively. The two-sided confidence interval for 95% confidence level, CI = [0.675, 0.801], agrees with the normality assumption, i.e., CI $\simeq [A_0 - 1.96\sigma_0,\ A_0 + 1.96\sigma_0]$, thus allowing the approximation of the empirical null distribution with the normal

distribution $F_{H_0}(x) \approx \Phi(x; A_0, \sigma_0^2)$. Consequently the achieved significance level can be computed as:

$$\text{ASL} = 1 - F_{H_0}(\text{AUC}_O) \approx 1 - \Phi(\text{AUC}_O; A_0, \sigma_0^2) = 0.000044, \tag{7}$$

where $\text{AUC}_O = 0.864$ is the observed AUC value. Thus, we can conclude that the observed AUC is significantly different from the expected AUC under the null hypothesis obtained by permutation bootstrap, and that there is strong evidence to reject the null hypothesis.

Figure 9 The empirical distribution of the AUC value obtained with permutation bootstrap under the null hypothesis that the control and case samples are drawn from the same distribution. The replicated AUC values were obtained by performing the feature selection in each of the 100,000 bootstrap iteration in order to minimise the feature selection bias (see online version for colours)



5.6 Multivariate feature selection

Feature selection can be thought of as a function $J_m = f(\boldsymbol{X}, \boldsymbol{y}, m, \varphi)$ which maps an n by d predictor matrix $\boldsymbol{X}$ of explanatory variables, n by 1 vector $\boldsymbol{y}$ of corresponding response variables (labels), number m of features we consider as important and want to select, and a specifier of feature selection method φ, into a set of unique column indices of matrix $\boldsymbol{X}$, $J_m = \{j_1, j_2, ..., j_m\}, 1 \le j_k \le d$. Examples for the specifier φ are WMW, FSFS/BSFS – Forward/backward sequential feature selection (Draper and Smith, 1966) and (Nete and Wasserman, 1974), LARS – Least Angle Regression (Efron et al., 2004), Random Forest Feature Selection (Breiman, 2001), RFE – Recursive Feature Elimination (Guyon et al., 2002), GLL - Generalised Local Learning (Aliferis et al., 2010a, 2010b). WMW is an example of univariate feature selection which consists of independent ranking of all d features based on Wilcoxon-Mann-Whitney rank sum test statistics, and then simply selecting m top ranked features. Other methods are examples of multivariate feature selection approaches, since they combine several features (columns of matrix $\boldsymbol{X}$) into discriminant vector $\boldsymbol{z}$ which is then used in some performance measure such as accuracy or AUC value. The approach can be formally described as:

$$\boldsymbol{z} = \boldsymbol{X}(J_m)\,\boldsymbol{w}_m, \tag{8}$$

where $\boldsymbol{X}(J_m)$ denotes the sub matrix of the predictor matrix $\boldsymbol{X}$, which contains only columns listed in J_m, while $\boldsymbol{w_m}$ is m by 1 projection (or regression) vector obtained by some projection method, such as MLR, SVM, LDA, applied to $\boldsymbol{X}(J_m)$. Besides the classifiers based on linear projection (8) there are other classifiers, such as Naïve Bayes classifier, regression trees (CART, C4.5), Random Forest classifier, k-nearest neighbour classifier etc. which do not fit directly into description (1) and are not considered in this discussion.

The power of multivariate feature selection is based on the fact that some features that can be ranked very poorly in a univariate test, but combined with other highly ranked features can produce a large training effect size, i.e., large training performance measure (Guyon and Elisseeff, 2003). However, multivariate feature selection has to be used with caution since it can easily lead to over-fitting and low cross-validated performance, especially in case of small training samples.

Discrimination based on equation (8) where J_m specifies only linear explanatory variables does not address the polymorphic behaviour of the immuno-response, suggested by the distributions in Figure 4. For example if we assume MLR where the elements of $\boldsymbol{w}$ are regression coefficients estimated from training data $\{\boldsymbol{X}(J_m), \boldsymbol{y}\}$, and $\boldsymbol{x} = (x_1, x_2, ..., x_m)$ is a feature vector of a subject, i.e., a row vector of $\boldsymbol{X}(J_m)$, then it can be shown that the marginal effect of the risk score of the subject with respect to the signal associated with the feature j can be written, based on equation (2), as:

$$\frac{\partial r}{\partial x_j} = r(1-r)w_j. \tag{9}$$

In other words, the marginal effect with respect to any signal does not explicitly depend on a signal associated with any feature, but depends only implicitly through r. If we however suppose for example a feature vector with linear and one interaction term $\boldsymbol{x} = (x_1, x_2, ..., x_m, x_1x_2)$, then:

$$\frac{\partial r}{\partial x_1} = r(1-r)(w_1 + x_2 w_{m+1}), \tag{10}$$

suggesting that the marginal effect of risk score with respect to x_1 explicitly depends on another signal x_2. In light of the polymorphic assumption, we can consider feature $j = 1$ as a discriminatory feature, while feature $j = 2$ can be a hidden variable which is not necessarily discriminatory (not highly ranked in a univariate test) but makes a difference between subjects with different responses measured by binding to glycan $j = 1$. In other words, the feature $j = 2$ provides a polymorphic context for feature $j = 1$ when assessing its marginal effect. Following this reasoning, we will include into feature selection and projection the interaction and quadratic terms, which gives rise to the design matrix $\boldsymbol{X}_D = [\boldsymbol{X}_L \mid \boldsymbol{X}_I \mid \boldsymbol{X}_Q]$ which is generally composed from three matrices: $\boldsymbol{X}_L = [\boldsymbol{u}_1 \mid \boldsymbol{u}_2 \mid ... \mid \boldsymbol{u}_M]$ – the n by M matrix of linear terms, $\boldsymbol{X}_I = [\boldsymbol{u}_1 \circ \boldsymbol{u}_2 \mid \boldsymbol{u}_1 \circ \boldsymbol{u}_3 \mid .... \mid \boldsymbol{u}_1 \circ \boldsymbol{u}_M \mid ... \mid \boldsymbol{u}_2 \circ \boldsymbol{u}_3 \mid .... \mid \boldsymbol{u}_{M-1} \circ \boldsymbol{u}_M]$ – the n by C_2^M matrix of interaction terms, and $\boldsymbol{X}_Q = [\boldsymbol{u}_1 \circ \boldsymbol{u}_1 \mid \boldsymbol{u}_2 \circ \boldsymbol{u}_2 \mid .... \mid \boldsymbol{u}_M \circ \boldsymbol{u}_M]$ – the n by M matrix of quadratic terms. The columns $\boldsymbol{u}_j$ of $\boldsymbol{X}_L$ are pre-selected columns from the original data matrix $\boldsymbol{X}$. The pre-selection is done for practical reasons to reduce the size of the matrix $\boldsymbol{X}_I$ and therefore to reduce the execution time for feature selection algorithms. We used

$M = 60$ top glycans ranked by WMW test. The columns of $\boldsymbol{X}_I$ represent all possible element-wise products (operator 'o') of columns of $\boldsymbol{X}_L$, which are ordered as implied by the expression above.

The equivalent of (8) is now the projection $\boldsymbol{z} = \boldsymbol{X}_D(C_m)\boldsymbol{w}_m$, where C_m is the set of column indices of $\boldsymbol{X}_D$ selected by the specified feature selection algorithm, while $\boldsymbol{w}_m$ is vector obtained by a projection algorithm applied to $\boldsymbol{X}_D(C_m)$. Thus indices from C_m refer either directly to the columns of $\boldsymbol{X}$ (the features) or to products of columns of $\boldsymbol{X}$ (the interaction or quadratic terms).

The performance of multivariate feature selection in the mesothelioma study is shown in Figure 10. The figure presents AUC values obtained by univariate feature selection based on WMW ranking and by multivariate feature selection based on forward sequential feature selection (denoted FWD). Both approaches are used with projection based on MLR, and are applied to design matrix with only linear terms and with linear and interaction terms. The left part of the figure shows training, while the right part shows the cross-validated AUC values. The cross-validation is performed with 10-fold cross-validation repeated 100 times. The AUC values are derived for various values of m ranging from 1 to 7. As expected the training performance is best with multivariate feature selection with linear and interaction terms and the worst with univariate feature selection with linear terms only. The cross-validated performance is, however, best for univariate feature selection with only linear terms (already discussed earlier, see Figure 5), and the next in performance is the multivariate feature selection with linear and interaction terms. The latter has achieved the maximal cross-validated AUC value of 0.77 for $m = 4$ interaction pairs.

Figure 10 Comparison of training and cross-validated AUC value for various univariate and multivariate feature selection approaches (see online version for colours)

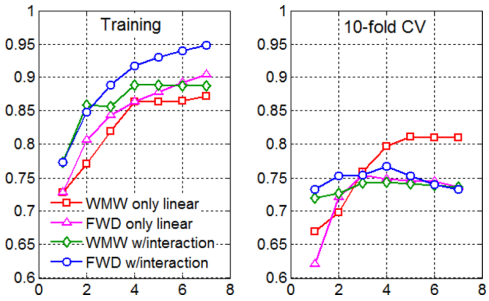
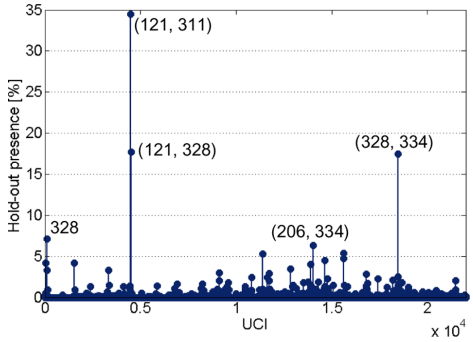


Figure 11 shows the CFS diagram for multivariate forward sequential feature selection for $m = 4$. The top CFS selected columns of the design matrix with linear and interaction terms refer to interaction pairs of GIDs (121, 311), (121, 328), (328, 334), and linear term 328. According to Table 1, the glycans 311, 334 and 328 can be considered as discriminatory glycans with WMW p-value $p \leq 0.00062$, while the glycan 121 has WMW-rank 17 with $p = 0.053$ (not shown in Table 1), and as such cannot be regarded as a discriminatory glycan, but rather as a hidden glycan which provides a polymorphic context to discriminatory glycans 311 and 328. It is interesting to note that CFS with forward sequential feature selection applied to only linear terms and $m = 4$ would select glycans 311, 328, 189 and 334, but not the glycan 121. The carbohydrate structure of glycan 121 is *GalNAca1-O-Ser*.

Figure 11 Compound feature selection based on the forward sequential feature selection used with multiple logistic regression and design matrix with linear and interaction terms. The number of selected features/interaction pairs is $m = 4$. The GIDs indicate the individual glycan and the interaction glycan pairs with the highest frequency of occurrence through 1000 hold-out iterations (see online version for colours)



The permutation bootstrap test with $B = 10{,}000$ iterations applied to FWD feature selection with $m = 4$ has resulted in empirical distribution under the null hypothesis with the mean value $A_0 = 0.84$, standard deviation $\sigma_0 = 0.023$, skewness = 0.026, kurtosis = 2.955, CI = [0.790, 0.880] and achieved significance level ASL = 0.00014. The observed AUC value for the training set is $AUC_O = 0.918$. The observed value for CFS-selected glycan interaction pairs, GID = (121, 311), (121, 328), (328, 334) is $AUC_O = 0.891$, and the corresponding ASL = 0.006. These numbers show that despite the relatively low cross-validated precision and AUC value, there is a solid statistical significance for the AUC values obtained with the multivariate feature selection with interaction terms. Better cross-validated performance will be achieved with larger samples.

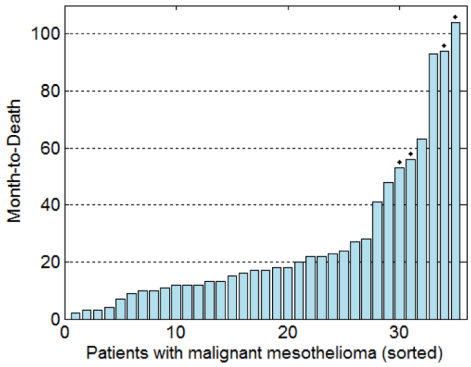
6 Prognostic data analysis

The goal of the prognostic analysis is to identify a set of glycans which best discriminate patients with low or high survival ability among the patients already diagnosed with the disease, here malignant mesothelioma. These glycans can be then used with a trained classifier to predict the survival probability of new patients diagnosed with mesothelioma, or to assess the effectiveness of a treatment in the process of the disease management of the patient. In order to perform the prognostic analysis it is necessary to know the survival times of patients in the training sample. In this study we have used Anti-Glycan Antibody (AGA) immunoprofiles of 35 patients out of 50 cases used in diagnostic analysis. The 35 cases analysed here are presented with malignant mesothelioma that can be histologically defined and staged, while the other 15 patients are presented with very advanced and often un-resectable cancer with clearly very poor prognosis, and therefore were deemed inappropriate for prognostic analysis.

The sorted survival times of the 35 cases are shown in Figure 12. The survival times are expressed in number of months between drawing the blood and the death of the patient. The four marked bars represent patients who were still alive at the time of clinical data collection and they will be treated as censored data in further analysis. The training

sample used in this analysis is relatively small and has limited utility for the plausible identification of prognostic markers. However the analysis that we present here will show that the data have indeed enough power to discriminate between poor and good prognosis groups.

Figure 12 Sorted Month-to-Death of 35 patients with malignant mesothelioma used in prognostic analysis. The patients are sorted by increasing survival time. The four marked bars represent patients still alive at the time of recording their clinical data (see online version for colours)



6.1 Cox proportional hazard regression

A natural approach to relate the predictors associated with various glycans with the survival times is the Cox proportional hazard regression model (Cox, 1972). The Cox PH regression can be built into a classifier, which works in a similar way as proposed by Lopes-Rios et al. (2006) where the gene P16/CDKN2A was used as a binary predictor (homozygous deletion status). This classifier has been cross-validated by the repeated 10-fold cross-validation, with 100 repetitions and the result was rather poor: Acc < 60%, AUC < 0.6, which rendered the classifier and the prognostic glycans ineffective. The same inferior result was obtained for cut-off time of 28 months in labelling the poor and good prognosis groups.

Besides the general failure of the classifier above there is a question of the appropriateness of application of the Cox PH model on PGA data in the first place. Namely, the basic assumption about the predictors to be used in the Cox PH model is the proportionality assumption, or consequently the independence of predictors on survival time, which can not be guaranteed for PGA-based predictors.

6.2 Discriminatory approach

An alternative approach to prognostic analysis is to apply the same discriminatory approach that was used in the diagnostic analysis. For this purpose we have labelled the 35 mesothelioma patients into poor and good prognosis groups by using cut-off value of 28 months. This seems to be a natural value after considering the survival times in Figure 12.

The univariate ranking of prognostic glycans with WMW tests applied to 27 patients labelled with poor prognosis (28 or less months of survival) and 8 patients labelled with good prognosis is shown in Table 2. As seen there are two glycans, GID = 154 and

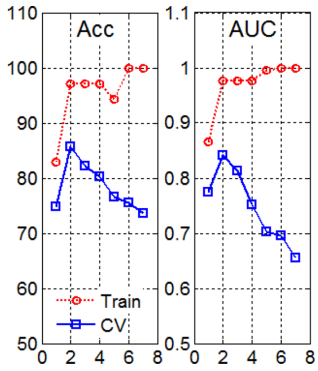
GID = 215 with distinguishing low *p*-values and high AUC values, considering the sample sizes.

The cross-validation results are shown in Figure 13, which is analogous to Figure 5 obtained in the diagnostic analysis. As seen the best cross-validated performance, Acc = 85.7% and AUC = 0.842, is obtained for two glycans.

Table 2 WMW ranking of glycans for samples of 27 poor prognosis and 8 good prognosis patients with malignant mesothelioma

Rank	*GID*	*Glycan structure*	*p-value*	*AUC*
1	154	Glcα1-4Glcβ-sp	0.00204	0.8657
2	215	GlcAβ1-6Galβ-sp	0.00302	0.8519
3	158	Galβ1-4Glcβ-sp	0.04306	0.7407
4	181	Neu5Acα2-6Galβ-sp	0.04306	0.7407
5	207	6-O-Su-Lacβ-sp	0.06204	0.7222

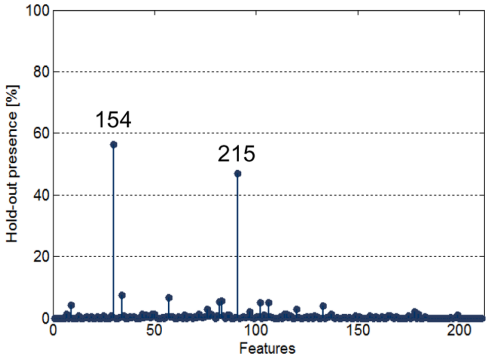
Figure 13 Prognostic accuracy and AUC value for combined top WMW-ranked glycans using multiple logistic regression. The diagram shows the training (dotted line) and cross-validated (solid line) performance values for various sizes of glycan sets ranging from one to 7 (see online version for colours)



The CFS diagram for two-glycan feature sets is shown in Figure 14, which is an analogue to Figure 6. The sharp drop in frequency after the second glycan clearly suggests that only two glycans, 154 and 215 can be considered as prognostic glycans. This finding may further improve after we implement the new generation of PGA arrays which will have an extended library of 300 glycans.

The observed AUC value for the training set of 35 mesothelioma patients is $AUC_O = 0.977$. In order to determine the statistical significance of this high AUC value we have performed the bootstrap test similar to the test discussed in Section 5.5. The permutation bootstrap test with $B = 10{,}000$ applied to AUC value obtained with WMW feature selection and MLR projection using only linear terms, has resulted in empirical distribution under null hypothesis with mean value $A_0 = 0.86$, standard deviation $\sigma_0 = 0.049$, skewness = –0.057, kurtosis = 2.77, two-sided confidence interval at 95% confidence CI = [0.764, 0.954], and achieved significance level ASL = 0.0092. This test has provided solid evidence that the mean AUC value under null hypothesis is significantly different than the observed AUC value of the training set.

Figure 14 Compound feature selection for feature sets of size $m = 2$, obtained with 1000 balanced hold-out iterations with equal hold-out sample sizes of 7 subjects. The feature selection is performed by WMW ranking and projection by MLR applied to 27 poor prognosis and 8 good prognosis mesothelioma patients (see online version for colours)



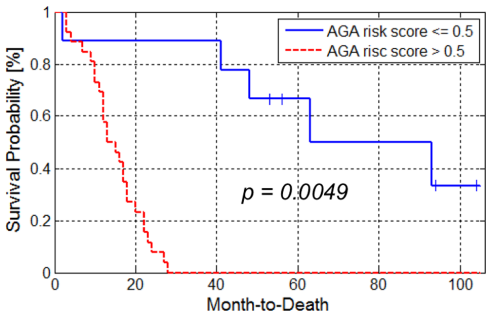
One of the principal goals of the prognostic analysis is to estimate the survival probability of a patient newly diagnosed with disease, or a diseased patient who has undergone a treatment. This can be done with the Kaplan-Meier (KM) estimator (Kaplan and Meier, 1958), which involves two steps:

1 plotting the survival functions for two groups of subjects with given survival times and censored information

2 testing of a new patient using same predictors and assigning the patient to one of the two KM curves.

The predictors used here are based on PGA data from which we are deriving the risk scores based on feature selection and projection discussed in the previous section. If the risk scores are defined by equation (2) and if the projection is performed with MLR, then we can say that patient with risk score equal to or below 0.5 belongs to the good prognosis group, otherwise the patient belongs to the poor prognosis group. The KM curves for the two prognosis groups are shown in Figure 15. The estimator is named "Observed KM" since the partitioning of subjects into two groups was based on the risk scores computed for the training set. The logrank p-value for the estimator is $p = 0.0049$, thus rendering the two KM curves significantly different. The median survival times are 14 and 63 months respectively.

Figure 15 Observed Kaplan-Meier plots obtained for mesothelioma patients with low (≤0.5) and high (>0) AGA risk score (see online version for colours)



Perhaps a more appropriate approach is to derive the KM estimator by a cross-validation technique. We have used leave-one-out approach due to a small training set. The algorithm goes as follows:

1 Label all patients as good prognosis if their survival time is greater than cut-off value, otherwise label them as poor prognosis

2 Remove one patient from the training set

3 Perform feature selection on the rest of the training set

4 Extract predictors associated with selected features

5 Find projection vector and intercept by using MLR

6 Compute the risk score for the removed patient by using training data obtained in Steps 3–5

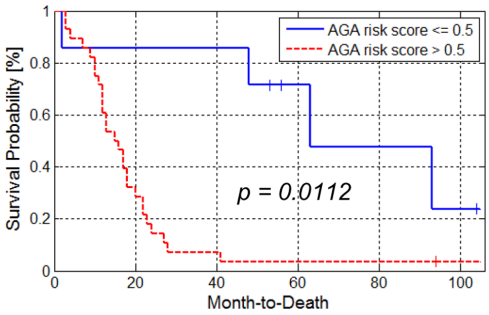
7 Mark the removed patient as good or poor prognosis depending on his or her risk score

8 Repeat Steps 2–7 until all patients from the training set are removed once and labelled

9 Use marked patients and their clinical data (survival times) to construct KM curves.

The LOOCV KM diagram is shown in Figure 16. The good/poor prognosis curves have logrank p-value $p = 0.0112$, still making the two curves significantly different. This diagram can be used to assess the survival probabilities for a new patient by using his or her serum and the risk score computed with projection vector based on prognostic glycans GID = 154 and 215 (determined by the CFS algorithm, Figure 14) and applied to the entire training set.

Figure 16 Cross-validated Kaplan-Meier plots (see online version for colours)



7 Conclusion

In this paper we have systematically presented an array of rigorous approaches for quality control, pre-processing, analysis and evaluation of diagnostic and prognostic data based on printed glycan arrays, which we have used since the inception of the new PGA technology in 2006. The methods are demonstrated on a mesothelioma study performed in the School of Medicine of the NYU, which contained sera of 65 high risk subjects

exposed to asbestos and 50 subjects with malignant mesothelioma. Although this was a case-control study with unspecified intended clinical use and with relatively small samples, we were still able to manifest the existence of diagnostic and prognostic power of PGA data and our analytical approaches. For example the conservative univariate feature selection based on non-parametric Wilcoxon-Mann-Whitney ranking and projection based on multiple logistic regression have resulted in observed values of accuracy and AUC value 79.1% and 0.864 respectively, while the repeated 10-fold cross-validation has yielded 74.1% and 0.811 respectively. The parametric permutation bootstrap with AUC statistics has resulted in achieved significance of the observed AUC value equal to 0.000044. The multivariate feature selection based on sequential feature selection algorithm and multiple logistic regression applied to the predictor matrix with linear and interaction terms has revealed a new glycan, GID = 121 (*GalNAca1-O-Ser*) which is ranked very low in the Wilcoxon test ($p = 0.053$) and which would not have been chosen by univariate or multivariate feature selection approaches applied to only linear terms of the predictor matrix. This glycan is apparently providing a polymorphic context to other highly discriminative glycans, and it has therefore elevated the observed precision and AUC value to 82.6% and 0.918 respectively. The achieved significance of the observed AUC value in parametric bootstrap was 0.00014. The cross-validated precision and AUC value were rather low, 70.1% and 0.77 respectively. These values were low due to the fact that the samples were too small for a multivariate feature selection and for inclusion of interaction terms.

We have also demonstrated the possibility of using the PGA data for prognostic purposes. The analysis is conducted as case-control discrimination where the patients with mesothelioma were labelled as poor prognosis ('case') and good prognosis ('control') depending on their survival times. The cut-off value was chosen to be 28 months. The univariate feature selection with multiple logistic regression applied to only linear terms of the predictor matrix has revealed two distinguished prognostic glycans, which were not seen in diagnostic analysis. The cross-validated accuracy and AUC value were 85.7% and 0.842 respectively. The observed AUC value was 0.977 and the achieved significance level in permutation bootstrap was 0.0092. This is a very encouraging finding despite the very small training sample, which suggests that the PGA based predictors can be used for prognosis as well as for diagnosis.

Although the results presented in this paper are very encouraging, we would still like to acknowledge the fact that the observations are from a small study, and that for potential clinical applications larger and randomised sample sets are needed. The new study is planned for the near future. The new study will be performed on a new version of PGA chips with a library of over 300 glycans.

Acknowledgements

The work presented in this paper was partially supported by NCI grant "Discovery and Clinical Validation of Cancer Biomarkers Using Printed Glycan Array" (Grant No.: U01CA128526 to M.E.H) and NCI/EDRN grant "American/Australian Mesothelioma Consortium" (Grant No.: U01CA111295 to H.I.P.) The PGAs were printed and developed at Cellexicon, Inc., La Jolla, CA. Most of the glycans used in PGA were synthesised and purified in the Carbohydrate Chemistry Laboratory at Shemyakin Institute of Bioorganic Chemistry, Moscow, Russia, under the support of the grant from

RAS Presidium Program "Molecular and Cell Biology". Authors wish to acknowledge very valuable discussions with Professor Richard Levine from Department of Statistics, SDSU, Vladimir Rotar from Department of Mathematics, SDSU, Dr. Ziding Feng from Fred Hutchinson Cancer Research Center, Professor Judith Goldberg from Department of Environmental Medicine, Division of Biostatistics, Langone Medical Center, NYU, and Dr. Gaelle Rondeau from Vaccine Research Institute, San Diego, CA.

References

Aarnoudse, C.A., Garcia Vallejo, J.J., Saeland, E. and van Kooyk, Y. (2006) 'Recognition of tumor glycans by antigen-presenting cells', *Curr. Opin. Immunol*., Vol. 18, pp.105–111.

Adams, N.M. and Hand, D.J. (1999) 'Comparing classifiers when the misclassification costs are uncertain', *Pattern Recognion*, Vol. 32, pp.1139–1147.

Ambroise, C. and McLachlan, G.J. (2002) 'Selection bias in gene extraction on the basis of microarray gene-expression data', *PNAS*, Vol. 99, No. 10. pp.6562–6566.

Aliferis, F.C., Statnikov, A., Tsamardinos, I., Mani, S. and Koutsoukos, D.X. (2010a) 'Local causal and markov. blanket induction for causal discovery and feature selection for classification, Part II: analysis and extensions', *Journal of Machine Learning Research*, Vol. 11, pp.235–284.

Aliferis, F.C., Statnikov, A., Tsamardinos, I, Mani, S. and Koutsoukos, D.X. (2010b) 'Local causal and markov blanket induction for causal discovery and feature selection for classification, Part I: algorithms and empirical evaluation', *Journal of Machine Learning Research*, Vol. 11, pp.171–234.

Arun, B., Vuskovic, M.I., Vasiliu, D., Xu, H., Atchley, D. Chambers, J.R., Bovin, N.V., Sneige, N., Hortobagyi, G.N. and Huflejt, M.E. (2007) 'Immunomodulatory effects of celecoxib in patients at increased risk for breast cancer', 30th San Antonio Breast Cancer Symposium, San Antonio, TX. *Br. Cancer Res. Treatment*, Vol. 106, Suppl. 1, p.S180.

Barnhart, H.X., Haber, M. and Song, J. (2002) 'Overall concordance correlation coefficient for evaluating agreement among multiple observers', *Biometrics*, Vol. 58, pp.1020–1027.

Blixt, O., Head Mondala, S.T., Scanlan, C., Huflejt, M.E. Alvarez, R., Bryan, M.C., Fazio, F., Caralese, D., Stevens, J., Razi, N., Stevens, D.J., Skehel, J.J., van Die, I., Burton, D.R., Wilson, I.A., Cummings, R., Bovin, N., Wong, C-H. and Paulson, J.C. (2004) 'Printed covalent glycan array for ligand profiling of diverse glycan binding proteins', *PNAS*, Vol. 101, pp.17033–17038.

Bovin, N.V. and Huflejt, M.E. (2008) 'Unlimited Glycochip', *Trends in Glycoscience and Glycotechnology*, Vol. 20, pp.245–258.

Bradley, A.P. (1997) 'The Use of the Area under the ROC curve in the evaluation of machine learning algorithm', *Pattern Recognition*, Vol. 30, No. 7, pp.1145–1159.

Breiman, L. (2001) 'Random forests', *Machine Learning*, Vol. 45, No. 1, pp.5–32.

Brown, P.O. and Botstein, D. (1999) 'Exploring the new world of genome with DNA microarrays', *Nature Genetics*, Vol. 21, pp.33–37.

Cox, D.R. (1972) 'Regression models and life tables', *Journal of the Royal Statistical Society Series B*, Vol. 34, No. 2, pp.187–220.

Draper, N.R. and Smith, H. (1966) *Applied Regression Analysis*, 2nd ed., J. Willey & Sons, pp.307–312.

Efron, B. and Tibshirani, R.J. (1993) *An Introduction to the Bootstrap. Monographs on Statistics and Applied Probability 57*, Chapman & Hall/CRC, New York.

Efron, B., Hastie, T., Johnstone, T. and Tibshirani, R. (2004) 'Least angle regression', *Annals of Statistics*, Vol. 32, pp.407–499.

Fawcett, T. (2003) *ROC Graphs: Notes and Practical Considerations for Researchers*, Technical Report, HPL-2003-4, Intelligent Enterprise Technologies Laboratory, HP Laboratories Palo Alto.

Flach, P. (2004) 'Tutorial on the many faces of ROC analysis in machine learning', *21st International Conference on Machine Learning (ICML)*, July, Banff, Alberta, Canada.

Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002) 'Gene selection for Cancer classification using support vector machines', *Machine Learning*, Vol. 46, pp.389–422.

Guyon, I. and Elisseeff, A. (2003) 'An introduction to variable and feature selection', *J. Machine Learning Research*, Vol. 3, pp.1157–1182.

Hakomori, S. (2002) 'Glycosylation defining cancer malignancy: new wine in an old bottle', *PNAS*, Vol. 99, p.10231.

Hand, D. and Till, R. (2001) 'A smple generalization of the area under the ROC curve for multiple class classification problem', *Machine Learning*, Vol. 45, pp.171–186.

Hanley, J.M. and McNeil, B.J. (1982) 'The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve', *Radiology*, Vol. 143, pp.29–36.

Huflejt, M.E., Cristofanilli, M., Shaw, L.E., Reuben, J.M., Fritsche, H.A., Hortobagyi, G.N. and Blixt, O. (2005a) 'Detection of neoplasia-specific clusters of anti-glycan antibodies in sera of breast cancer patients using a novel glycan array', *Proc. Amer. Assoc. Cancer Res*., Vol. 46, p.1313.

Huflejt, M.E., Vuskovic, M. Blixt, O., Xu, H., Shaw, L.E., Reuben, J.M., Kuerer, H.M. and Cristofanilli, M. (2005b) 'Glycan array identifies specific signatures of anti-glycan autoantibodies in sera of breast cancer Patients: diagnostic, prognostic and therapeutic opportunities', *28th Annual San Antonio Breast Cancer Symposium*, TX, *Breast Cancer Res. Treat.*, San Antonio, Vol. 94, p.S85.

Huflejt, M.E., Vuskovic, M., Vasiliu, D., Xu, H., Obukhova, P., Shilova, N., Tuzikov, A., Galanina, O., Arun, B., Lu, K. and Bovin, N. (2009) 'Anti-carbohydrate antibodies of normal sera: findings, surprises, and challenges', *Molecular Immunology*, Vol. 46, pp.3037–3049.

Hutchens, T.W. and Yip, Y-T. (1993) 'New desorption strategies for mass spectrometric analysis of macromolecules', *Rapid Comm. In Mass Spectrometry*, Vol. 7, pp.576–580.

Issaq, H.J., Veenstra, T.D., Conrads, T.P. and Felschow, D. (2002) 'The SELDI-TOF MS approach to proteomics: protein profiling and biomarker identification', *Biochemical and Biophysical Research Communications*, Vol. 292, pp.587–592.

John, N.R. and Draper, J.A. (1980) 'An alternative family of transformations', *Applied Statistics*, Vol. 29, No. 2, pp.190–197.

Kaplan, E.L. and Meier, P. (1958) 'Nonparametric estimation from incomplete observations', *J. Amer. Statist. Assn.*, Vol. 53, pp.457–481.

Lin, L.I. (1989) 'A concordance correlation coefficient to evaluate reproducibility', *Biometrics*, Vol. 45, pp.255–268.

Ling, C.X., Huang, J. and Zhang, H. (2003) 'AUC: a statistically consistent and more discriminating measure than accuracy', *Proceedings of the Eighteenth International Joint Conference of Artificial Intelligence (IJCAI)*, Acapulco, Mexico, pp.519–526.

Lopez-Rios, F., Chuai, S., Flores, R., Shimizu, S., Ohno, T., Wakahara, K., Illei, P.B., Hussain, S., Krug, L., Zakowski, M.F., Rusch, V., Olshen, A.B. and Ladanyi, M. (2006) 'Global gene expression profiling of pleural mesotheliomas: overexpression of aurora kinases and P16/CDKN2A deletion as prognostic factors and critical evaluation of microarray-based prognostic prediction', *Cancer Research*, Vol. 66, No. 6, pp.2970–2979.

Nete, J. and Wasserman, W. (1974) *Applied Lenar Statistical Models, Regression, Analysis of Variance, and Experimental Design*, Richard D. Irwin, Inc., Homewood Ill, pp.382–387.

Parker, B.J., Günter, S. and Bedo, J. (2007) 'Stratification bias in low signal microarray studies', *MBC Bioinformatics*, Vol. 8, p.326.

Sidranski, D. (1997) 'Nucleic acid-based methods for detection of cancer', *Science*, Vol. 278, pp.1054–1058.

Simon, R., Radmacher, M.D., Dobbin, K. and McShane, L.M. (2003) 'Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification', *Journal of the Nationl Cancer Institute*, Vol. 95, No. 1, pp.14–17.

Trademark Publications (2008) *Biomarker Technology Platforms for Cancer Diagnoses and Therapies*, Trimark Publications, Global Information, Inc. Prod. Code TK63076, http://www.the-infoshop.com/report/th63076-diagnoses.html

Wright Jr., G.L. (2002) 'SELDI protein chip MS: a platform for biomarker discovery and cancer diagnosis', *Expert Review of Molecular Diagnostics*, Vol. 2, pp.549–563.

Note

[1]Based on personal communication and advice from Dr. Ziding Feng from Fred Hutchinson Cancer Research Center, February 2008.