BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 13 2012, pages 1669–1676
doi:10.1093/bioinformatics/bts242

Genome analysis Advance Access publication April 28, 2012

Statistical analysis of glycosylation profiles to compare tissue type and inflammatory disease state

Catherine A. Hayes[1,*], Szilard Nemes[2] and Niclas G. Karlsson[1]

[1]Department of Medical Biochemistry, Sahlgrenska Academy and [2]Division of Clinical Cancer Epidemiology, Department of Oncology, University of Gothenburg, Gothenburg, Sweden

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Glycosylation is one of the most important post-translational modifications of proteins and explains some aspects of the diversification of higher organisms not explained by template-driven synthesis. For glycomics to mature as much as genomics and proteomics, the necessary tools need to be developed and tested. Liquid chromatography-mass spectrometry is one of the gold standards for oligosaccharide analysis and leads to large amounts of data, not easily interpreted manually. We present a study on the testing and validation of statistical analysis tools to aid the structural elucidation of these analyses as well as using the results to answer biologically relevant questions.

Results: We show the usefulness of data reduction and statistical analysis in the interpretation of complex glycosylation data. The reduction does not result in the loss of importance of the glycosylation information as shown by comparison of control and disease samples in two tissue types.

Contact: catherine.hayes@medkem.gu.se

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 16, 2011; revised on April 2, 2012; accepted on April 23, 2012

1 INTRODUCTION

Glycosylation is regarded as one of the most important co- and post-translational protein modifications. There are two main types, bonded to the nitrogen of an asparagine in the consensus sequence Asn-X-Ser/Thr (*N*-linked) or to the oxygen of a serine/threonine, with a favourable placed proline in the -1 or +3 position (*O*-linked). Oligosaccharide attachment to protein sites explains some aspects of the functional diversification of higher organisms beyond the template driven gene-encoded polypeptide synthesis. The role and impact of glycosylation is complex and not very well understood and many studies have provided profiles from glycosylated proteins (glycome and/or glyco-proteome) highlighting a number of different causative factors (Marino *et al*., 2010; Reis *et al*., 2010; Vigerust, 2011), e.g. underlying proteins, tissue location, disease state, presence of pathogens, immune response system, etc. It is reasonable to assume that changes in glycosylation are viable targets for biomarker-based research projects (Packer *et al*., 2008).

Genomic and proteomic studies often make use of microarray data to explain changes in phenotype, manifestation of a disease state or to explain a certain biological pathway. These data are quite large, with thousands of microscopic array spots being recorded. Statistical methods have been developed to deal with the interpretation of this type of data. Mass Spectrometry is a powerful technique in the structural elucidation of glycans (An and Lebrilla, 2010), and in particular, liquid chromatography–mass spectrometry (LC-MS) which allows separation of isomeric structures (Novotny *et al*., 2008). Glycome data from LC-MS can be data-rich and difficult to transform into useful 'knowledge', either analytical or biological. Unlike the proteomic domain, oligosaccharide structural databases, and the tools to interrogate them are still in their infancy (Campbell *et al*., 2011; Hayes *et al*., 2011).

O-linked glycosylation is a major part of the structure of mucins in the mucous and usually account for more than 50% of the mass of the protein. As mucins mainly have a protective function on epithelial surfaces as well as in secretions (e.g. stomach lining, saliva; Amerongen *et al*., 1995; Corfield *et al*., 2000) changes in their glycosylation can be indicative of inflammation, lessened lubrication, pathogen attack, etc. (Ohtsubo and Marth, 2006). Blood group antigens are expressed on *O*-linked structures, and the presence or absence of the *Sec* fucosyltransferase (FUT2) determines the secretor status of individuals (Sidebotham *et al*., 1995).

A data-set of published *O*-linked oligosaccharide LC-MS data from different tissue types and/or inflammatory disease states (Table 1) was compiled to compare normal and inflammatory glycosylation response in different tissues. By converting LC-MS data into a mass spectrometry average composition (MSAC) metric that describes each sample, statistical analysis could be simplified and facilitate screening of large sets of sugar-data that would be difficult by manual interpretation. To achieve this for LC-MS data each sample was reduced to a monosaccharide composition that describes the normalized number of sugar residues and modifications per glycan molecule. This method could also be applied to MALDI-TOF data not taking isomers into consideration, or for glycopeptide data as well as other types of glycosylation (e.g. *N*-linked, glycosyaminoglycans etc). One advantage of reducing the data to this type of metric is the elimination of misinterpretation of structure due to re-arrangements (Wuhrer *et al*., 2011). Another advantage is reducing the need for full structural elucidation of samples.

2 METHODS

2.1 Datasets

Datasets were compiled (Table 1) using raw and/or pre-calculated monosaccharide compositions. Access to the data was through Dr Karlsson

*To whom correspondence should be addressed.

© The Author 2012. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Table 1. Datasets used in the analysis were compiled from collaborative projects and literature

Tissue	Disease/control	Proteins	Number of samples	References
Lung	Control	MUC5B, 5AC[a]	19/4[b]	Schulz *et al*. (2007)
	CF-exacerbated	MUC5B, 5AC	19/4[b]	
	CF-discharge	MUC5B, 5AC	12	
SMG	Control	MUC5B, 5AC[a]/GP340	3/3[c]	Schulz *et al*. (2005)
	CF	MUC5B, 5AC/GP340	5/5[c]	
	Non-CF LD	MUC5B, 5AC/GP340	5/5[c]	
HBE	CF	MUC5B, 5AC[a]	3	Holmén *et al*. (2004)
	Control	MUC5B, 5AC	3	
Cervix	Pre-ovulation	MUC5B, 5AC, 6, 16[a]	4/2[d]	Andersch-Bjorkman *et al*. (2007)
	Ovulation	MUC5B, 5AC, 6, 16	3/2[d]	
	Post-ovulation	MUC5B, 5AC, 6, 16	3/0[d]	
Saliva	Control/healthy	MUC5B/MUC7	10/26[d]	Karlsson and Thomsson (2009), Thomsson *et al*. (2005)
Colon	Control	MUC2	25	Larsson *et al*. (2009)
	UCi	MUC2	12	
	UCa	MUC2	13	
	IBS	MUC2	5	

Each dataset was divided into categories, based on tissue and/or mucin type, and on health status. Structures were converted to MSAC before being subjected to statistical analysis. CF, Cystic fibrosis; SMG, Sub-mucosal gland; LD, Lung disease; HBE, Human bronchial epithelial Cell Culture; UCi, Inactive ulcerative colitis; UCa, Active ulcerative colitis; IBS, Irritable bowel syndrome.
[a]Mixture of proteins; [b]Adult/children; [c]Mucin fraction/GP340; [d]Secretor/non-secretor.

(co-author) and collaborators in Gothenburg University. Calculation of the MSAC is presented in Supplementary Table S1. LC-MS analysis data of released sugars from the mucin fraction of different tissue types was collected in similar manners and with similar interpretation of the structural data from lung, saliva, cervix and colon. Control and disease samples were available for lung and colon (Table 1). Salivary samples were differentiated into two main mucins, MUC5B and MUC7, and cervix were classified into pre-, post- and during ovulation. The secretor status of the cervix and salivary samples were known. The monosaccharide compositions consisted of average numbers of each of four residues, hexose (Hex), N-acetyl hexosamine (HexNAc), fucose (Fuc), sialic acid (NeuAc) and a common modification, sulfation (Sulf).

2.2 Monosaccharide compositional analysis

Output from data interpretation of a sample comprises a structurally annotated list of mass to charge ratios (m/z) and intensities. Structures were reduced to monosaccharide compositions of HexNAc, Hex, Fuc, NeuAc and Sulf. The monosaccharide compositions were multiplied by percentage intensity for each structure and then each parameter summed over the entire sample to give the MSAC metric. An example of this calculation is given in Supplementary Tables S1b and 3.

2.3 Heat plots

Cluster analysis was used as an exploratory tool to determine the groupings in the dataset and Euclidean distance (L_2 norm) was used as measure of distance between observations. As the aim was to find similar clusters, we used hierarchical clustering with complete linkage. To avoid preferential attention to variables with larger variance prior to clustering, all variables were scaled to mean zero and unit variance. This also facilitated the graphical representation of the dataset as heat maps. The heat maps depict the two-way classification of the observations (i.e. the plots are clustered on both x and y axes).

2.4 Classification methods

To assess the health status of patients as a function of MSAC, we built classification models based on linear discriminant analysis. Discriminant function analysis helped in determining which variables discriminate between the health statuses of the patient samples and offered a general classification matrix. The classification matrix was based on leave-one-out cross-validation. As a more subtle representation of the relationship between MSAC and health status, predictive logistic regression models were built. Due to the high degree of correlation between the monosaccharide compositions, the odds ratios estimated from the full model with all predictors included were unreliable. Thus, we proceeded with model selection based on Akaike Information Criterion and Likelihood Ratio tests (Bozdogan, 1987). The low number of predictors allowed a manual selection, which in all cases yield the same result as automated stepwise regression. Area under the receiver operating characteristic curves (AUC) was used as an external validating tool. The value of AUC ranges between 0.5 (random classification) and 1 (perfect classification). This was followed by the use of non-parametric bootstrapping to estimate the variability associated with AUC values.

2.5 Validation of statistical models

Cross-validation was carried out by iterations of model building using 90% of the training set, and using the remaining 10% to test the outcomes. In this procedure, error bars were close to zero (results not shown). All statistical methods and scripts have been published (Hayes *et al*., 2012).

3 RESULTS

The main objective of this study was the identification of statistical tools that could be used to probe large datasets of oligosaccharide LC-MS data to extract relevant biological evaluations that may not have been easily achievable by manual interpretation. These tools could also be used to identify patterns in the analytical data that may help in structure determination as well as providing insight into important biological events such as inflammation.

3.1 Glycosylation profiles in different tissue types

MS data was compiled for a range of tissue types and mucins. The typical mass spectra from these samples indicate different

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

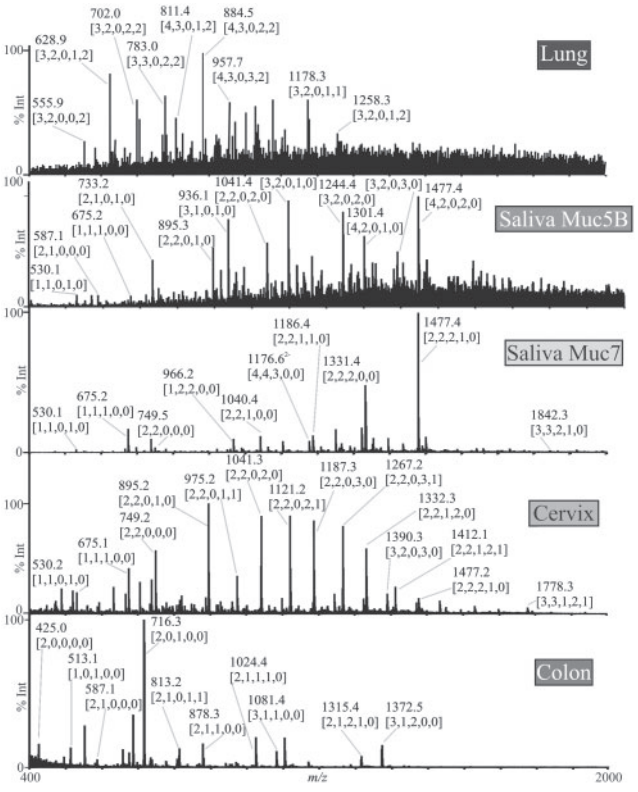

Fig. 1. Typical mass spectrum from each of the five healthy tissue types. Peaks are labelled with m/z (mass to charge ratio) values and structure [HexNAc, Hex, NeuAc, Fuc, Sulf]

Table 2. The most common O-linked cores are presented along with the epitopes for blood group types

Sugar type	Structure	Epitope	Structure
Core 1 (T Antigen)		Blood group O	
Core 2		Blood group A	
Core 3		Blood group B	
Core 4		Tn Antigen	

Secretor status is based on the present or absence of the α-1, 2-linked Fuc (▲) on the Gal (●) of these substructures. Other residues: GalNAc (□) and GlcNAc (■).

glycosylation profiles (Fig. 1). Features such as blood group (Table 2), core type elongation and average monosaccharide composition were identified. Due to the incomplete nature of the structural information in the dataset, it was decided that a metric for average composition should be applied to all data.

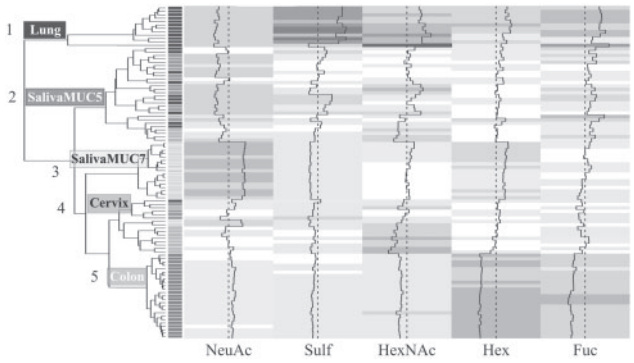

Fig. 2. Heat plot showing the healthy tissue types as categorized using clustering methods. Five main clusters were observed which matched the tissues. The rainbow coloured bar on the left distinguishes the actual categories and points out the outliers in the groups, lung (red), saliva-MUC5B (orange), saliva-MUC7 (yellow), cervix (green) and colon (blue). The heat plot colours are from blue to red (−4.5 to +4.5) normalized to the column mean. The dendogram to the left represents the hierarchical clustering of the samples

3.1.1 Data-set compilation and reduction The data-set was compiled from different sources (Table 1) and reduced to MSAC using the relative intensities and sequences of the solved structures (Supplementary Table S2). Tissue samples included lung, saliva (healthy MUC5B and MUC7), cervix (pre-, post- and during ovulation) and colon. The monosaccharide compositions consisted of average numbers of each of four residues (Hex, HexNAc, Fuc and NeuAc) and one modification (Sulf), repeated for all samples. This reduces the contribution of individual variability due to blood group ABO (Table 2) expression also known as secretor/non-secretor status.

3.1.2 Differential analysis and clustering A heat plot was constructed using the control dataset (Fig. 2). It shows the efficient separation of the samples into five groups by hierarchical clustering (displayed as a dendogram to the left of the figure). Cluster 1 is made up of lung samples characterized by low sialylation and high sulfation. Cluster 2 is made up of 28 samples, 13 salivary MUC5B, 10 lung and 5 cervix (lower sulfation). Cluster 3 is made up of salivary MUC7 samples (high sialylation). Cluster 4 is made up of 16 members, 1 lung sample, 4 salivary MUC5B, 2 salivary MUC7 and 9 cervix samples (intermediate level of sialylation). Cluster 5 is made up of colon samples (short oligosaccharides).

Of the mixed tissue clusters, cluster 2 (mostly MUC5B) belongs to secretor samples from cervix and saliva (one salivary non-secretor). Additional lung samples in this cluster had a secretor-like profile (high in fucosylated structures), but appropriate data was not available to determine the structures. We conclude that these samples are similar to the secretor type profiles of the saliva and cervix, allowing for less than 20% contamination of lung cells with squamous (salivary) cells, (Schulz *et al*., 2007).

Most of the saliva and cervix in cluster 4 were non-secretors and lung samples in this cluster had a non-secretor type profile (lack of fucose-containing structures). It has been shown that pre- and post-ovulation cervix mucins contain less neutral fucosylated structures (Andersch-Bjorkman *et al*., 2007) and more of the NeuAcα2,6GalNAc (Sialyl Tn) and NeuAcα2,3Gal epitopes,

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

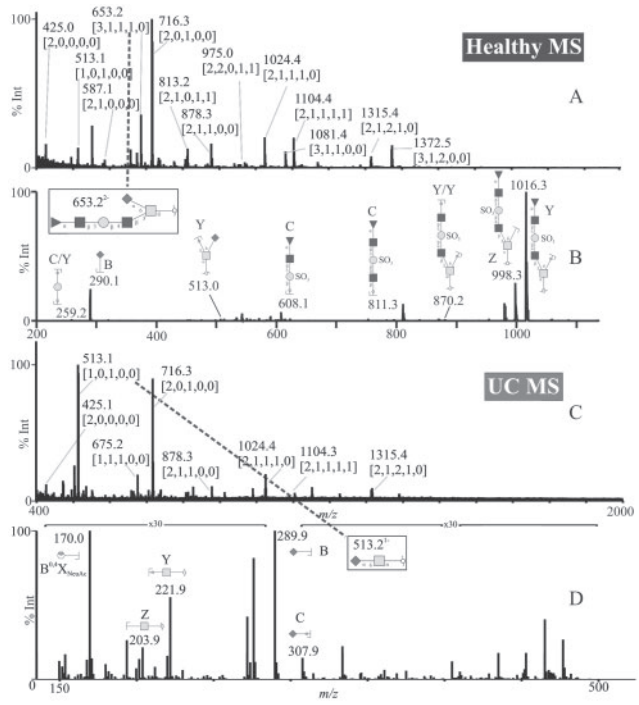

Fig. 3. Mass spectra from typical healthy (**A**) and UC (exacerbated) (**C**) colon samples. Also shown are MS/MS spectra (**B** and **D**) with assigned fragments for typical structures that are representative of the monosaccharide composition as calculated from the univariate probability models for prediction of disease state (Table 3)

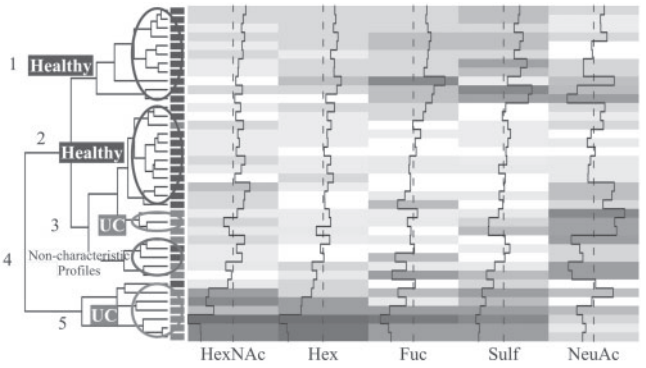

Fig. 4. Heat plot showing the clustering of healthy versus active ulcerative colitis (UC) data. Clustering initially created two groups. One cluster is mainly UC samples. The other cluster can be divided into four daughter clusters, two representing typically healthy samples, one UC profile and one non-characteristic profile. The heat plot colours are from blue to red (−3 to +3) normalized to the column mean. The samples are colour coded to indicate healthy (red) or UC (blue) state

Table 3. Typical monosaccharide compositions for 100% healthy and 100% disease state for both lung and colon (Supplementary Table 4)

	HexNAc	Hex	Fuc	NeuAc	Sulf
Colon: healthy	2.3	0.8	0.6	1.1	0.4
Colon: UC	1.4	0.1	0.1	1.12	0.0
Lung: healthy	4.6	3.0	2.2	0.0	1.8
Lung: CF	2.1	1.6	0.4	1.8	0.0

equivalent to a non-secretor type profile, thereby explaining the presence of the secretor cervix samples in this cluster.

The generalization, that clusters 1 and 2 are made up of blood-group dependent glycosylation (secretors) and that clusters 3, 4 and 5 less so, is re-enforced by these findings. The heat plot (Fig. 2) shows that clusters 1 and 2 are noteworthy by their lower than average amount of NeuAc (blue) and higher than average for other parameters (red). This is reversed in the other three clusters. These data shows that there is both a glycosylation dependency due to individual (secretor status) as well as tissue and protein variation, and that the simplification of the data into MSAC retains this information.

3.2 Glycosylation in disease states versus healthy: COLON

3.2.1 Data-set reduction and clustering A typical mass spectrum from one each of a control and inflammatory disease state (UC) colon sample shows obvious structural differences between the two (Fig. 3A and C). Data from these samples was used to construct a heat plot (Fig. 4).

The analysis results in five clusters, healthy, UC and non-characteristic profiles. Clusters 1 and 2 are composed of mainly control-type colon patients. These profiles have high values for HexNAc, Hex, Fuc and Sulf with varying amounts of sialylation (NeuAc). Clusters 3 and 5 are predominantly UC (1 out of the seven members of group 5 is a control), showing a different profile to that of the control samples, i.e. low overall amounts of HexNAc, Hex, Fuc and Sulf which is in agreement with previous research (Larsson *et al*., 2009). There are varying amounts of NeuAc but on average reduced amounts when compared to the controls. Cluster 4 is a mixture of both control and UC type profiles.

3.2.2 Building of probability models A generalized linear model was used to build a probability curve based on the five parameters. Univariate probability models were built using single parameter data (Supplementary Fig. S1). The probabilities associated with these models were plotted and used to deduce the composition of structures with highest and lowest probability of disease (Table 3) and the predicted structures agreed with typical structures deduced from the mass profile (Fig. 3B and D; Supplementary Table S4).

A likelihood ratio test was used to identify the model that best represented the differences between the control and disease state for the colon, with repeated rounds excluding that with the lowest AUC value. The univariate model based on the Sulf parameter gave the best discrimination. AUC values for the model are given in Table 4.

3.2.3 Validation of probability model Cross-validation was carried out by iterations of model building using 90% of the training set, and using the remaining 10% to test the outcomes. Error bars were close to zero (results not shown). The probabilities for the colon healthy and colon UC samples are shown in Table 4.

3.2.4 Testing of probability model A subset of colon samples [UC patients with inactive disease, UCi (non-inflammatory state) and IBS

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Table 4. AUC values for best probability models for lung and colon

	AUC (confidence levels)	Average *P* value	
		Control	Disease
Colon	0.935 (0.819–1.00)	0.125	0.760
Lung	0.970 (0.911–1.00)	0.162	0.810

Table 5. Probability data for subset of Colon data

Sample	Average Prob	Max Prob	Min Prob
UCi	0.250	0.953	0.000
IBS	0.108	0.397	0.004

UCi, Inactive ulcerative colitis; IBS, Irritable bowel syndrome.

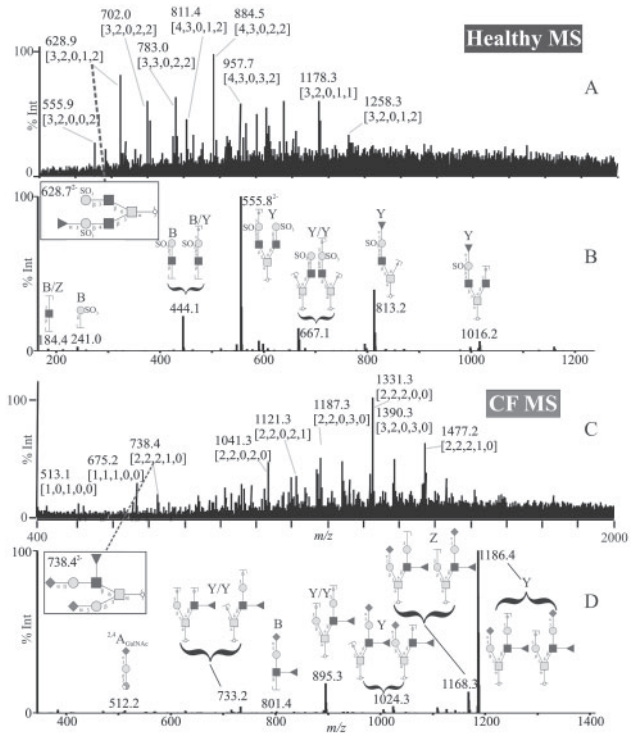

Fig. 5. MS of typical control (**A**) and CF lung (**C**) samples. **B** and **D** show MS/MS spectra with fragments for structures that are representative of the monosaccharide composition as calculated from the univariate probability models for prediction of disease state

(irritable bowel syndrome; non-inflammatory); Table 1] was used to validate the probability models (average probabilities presented in Table 5, full probability results given in Supplementary Table S2a). The results indicate both UCi and IBS samples have control-like glycosylation profiles in keeping with their non-inflammatory states, despite outliers in the UCi data.

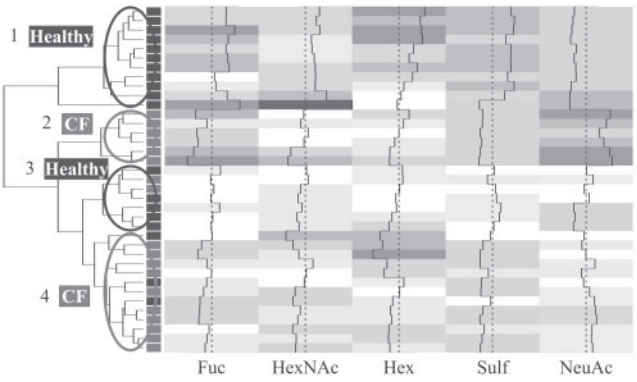

Fig. 6. Heat plot showing the clustering of control versus cystic fibrosis (CF) lung data which produced four clusters, two representing each of the states. Clusters 1 and 2 are made up of healthy and CF patients. Clusters 3 and 4 are mainly healthy and CF, respectively, with outliers in both. The heat plot colours are from blue to red (−4 to +4) normalized to the column mean. The samples are colour coded to indicate healthy (red) or CF (blue) state

3.3 Glycosylation in disease states versus healthy: LUNG

A second study into disease and healthy state was conducted using lung samples, control and cystic fibrosis (CF; inflammatory). Figure 5 shows representative mass spectra from a healthy and a CF patient sample, indicating the differences in structural profile of the two.

Also shown are MS/MS of typical structures from these profiles showing that the control lung samples show elongated structures terminating with fucose and expressing blood-group type antigens (Table 2) whereas the CF samples are shortened and terminated with sialic acid. Heat plot analysis was performed (Fig. 6) and the samples clustered into four clusters, two of each state. Cluster 1 (mainly controls) has lower than average amounts of NeuAc and higher than average values for the rest of the parameters and correspond to cluster 1 of heat plot in Figure 2. Cluster 2 of the lung heat plot contains CF patients. Cluster 3 and 4 are mainly controls and CF, respectively, with outliers in both. Cluster 3 has seven members, two of which are CF and cluster 4 has 13 members with three controls. Cluster 3 contains controls that correspond to the samples found in cluster 2 of the first heat plot in Figure 2. It has been indicated that CF samples have lower amounts of blood-group antigens. Therefore, we can imagine that we are seeing a differentiation dependent on blood grouping in this heat plot.

3.3.1 Building of probability models Univariate probability models (Supplementary Fig. S2) were constructed as before. All parameters did well with NeuAc, Fuc and Sulf scoring the highest. As before probabilities were used to determine the oligosaccharide composition of the structure that represents the extremes of the curve (i.e. 100% probability of healthy or CF state; Tables 3 and 4). These compositions match quite well with typical structures found in a mass spectrometric profile of the samples (Fig. 5).

Likelihood ratio test was used to identify the best combination of variables to use in a probability model, resulting in the combination of NeuAc and Sulf being the best predictor of CF disease state in the lung. AUC values for the model are given in Table 4. The average probabilities for the healthy and CF lung samples are

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Table 6. Probability values for subset of lung patient data

Sample	Average Prob	Max Prob	Min Prob
CF-discharge	0.922	0.995	0.624
SMG-control (GP340)	1.000	1.000	1.000
SMG-CF (GP340)	0.999	0.999	0.999
SMG-non-CF LD (GP340)	0.995	1.000	0.980
SMG-control (Mucin)	0.802	0.992	0.564
SMG-CF (Mucin)	0.870	0.999	0.745
SMG-non-CF LD (Mucin)	0.862	0.989	0.615
HBE-control	0.825	0.962	0.576
HBE-CF	0.884	0.937	0.809

CF, Cystic fibrosis; SMG, Sub-mucosal gland; GP340, Glycoprotein 340; LD, Lung disease; HBE, Human bronchial epithelial cells.

tabulated in Table 4 showing that disease states were significantly different from the random score of 0.5. From the ROC curves of the univariate probability models, we can identify the trends with respect to each of the parameters (Supplementary Fig. S2). As has been shown previously (Schulz *et al*., 2005) the levels of NeuAc increase and the other parameters decrease with respect to disease state. The univariate models give monosaccharide composition of an extreme case control-like or CF-like glycosylation profile (Table 3). This shows that the overall length of the structure is halved in CF (HexNAc and Hex), and that the number of fucose residues is reduced 5-fold which could be indicative of loss of blood group type antigens, as well as a change in core type.

3.3.2 Testing of probability model A test set of samples was used to validate the probability model consisting of sub-mucosal gland (SMG) samples and human bronchial epithelial cell culture (HBE) samples (Table 6). Average probabilities as calculated by the model are presented in Table 6. Probability values were high for all samples, indicating a CF-like rather than control-like profile for all samples.

4 DISCUSSION

Glycosylation is recognized as an important post-translational modification involved in cell–cell recognition and communication (Morelle and Michalski, 2005). Changes in glycosylation due to inflammatory response (Mortier *et al*., 2011), can arise in a number of different ways, i.e. biosynthetic alterations of transferases, donors, acceptors or glycosidases; changes in protein expression or cellular location (Lauc *et al*., 2010). De-coding these changes and their contribution to the glyco-profile of a sample will allow greater understanding of the inflammatory response.

The MSAC metric allows the reduction of complicated data, but in contrast to traditional monosaccharide compositional analysis, there is a common baseline among the samples (Core HexNAc). An example of the difference between the two methods of normalization is given in Supplementary Table S3.

4.1 Glycosylation profiles in different tissue types

The control dataset has diverse origins, protein backbones and glycosylation features (Fig. 1). The heat plot displays the efficient separation of the samples into five groups by hierarchical clustering of a glycosylation-based metric (MSAC; Fig. 2), which corresponds to groupings based on tissue type and/or mucin. The simplest of these are the colon (MUC2; Larsson *et al*., 2009) and the two salivary mucins (MUC7 and MUC5B; Thomsson *et al*., 2005). The cervix and lung samples are made up of mixtures of different mucins both containing predominantly MUC5B and to a lesser extent MUC5AC (Karlsson and Thomsson, 2009; Schulz *et al*., 2005; Thomsson *et al*., 2005). In the lung, MUC5B mainly originates from the SMG showing a highly sialylated profile, and MUC5AC from the goblet cells. In healthy lung samples, SMG secretions contribute an estimated 95% of upper airway mucous (Wine and Joo, 2004). Of these secretions, 40% are from mucous cells and 60% from serous cells, producing water, electrolytes and a rich mixture of 'anti' compounds. The data shows increased sulfation and decreased sialylation of the control lung samples indicating a reduced contribution from SMG in the healthy state.

Salivary secretions show glycosylation of the individual mucin entities as well as the contribution from the multiple cell types (Bikker, 2004). Salivary MUC7 is one of the non-gel forming mucins and does not display blood-group antigens as much as MUC5B in saliva (Karlsson and Thomsson, 2009). In studies on cervical mucous from different time points along the menstrual cycle, amounts of MUC5B vary as does the glycosylation profile of the mucous (Andersch-Bjorkman *et al*., 2007) indicating a contribution from proteins that are expressed and underlying transferases involved in glycosylation. During ovulation, the mucous glycosylation is comprised of more neutral fucosylated core 2-type structures (more 'secretor' like) whereas in samples from before and after ovulation the glycosylation is shortened (sialyl Tn type structures) and highly sialylated. The change in glycosylation appears to be limited to terminal structures and does not appear to alter the underlying cores. Differential expression of C3GnT at the Tn branching point of *O*-glycosylation biosynthesis determines appearance of core 3 type structures. Presence of this transferase would result in an increase of core 3 with a decrease in core 1 structures. We can hypothesize from this that a change in glycosylation in this instance arises from a change in the regulation of the transferases as opposed to a change in the actual transferases that are expressed. This has been seen previously with the cellular reorganization of GalNAc transferases that is mediated by the Src tyrosine kinase. This protein causes the GalNAc transferase to relocate from the Golgi to the ER. It has been hypothesized that the GalNAc-T has a lectin domain which competes with the core I gal transferase site blocking the extension of the T antigen (Gill *et al*., 2010).

There are three homogenous clusters, showing mucins from different tissues (Fig. 2) cluster 1 (lung), cluster 3 (salivary MUC7) and cluster 5 (colon). These clusters have different mucin protein backbones (Table 1). Clusters 3 and 5 are made up of single mucins, i.e. salivary MUC7 and colonic MUC2. Cluster 1 is lung samples, but these are made up of a mixture of mucins, MUC5B and 5AC. Therefore, we can make the assumption that the glycosylation profile of the samples in these three clusters is dependent on both tissue and protein.

The other two clusters are made up of mainly one type of sample but with outliers from other groups. This may be due in part to the secretor status of the samples. There is also a contribution from the mixture of proteins that are found in these groups. Cluster 2 is mainly salivary MUC5B also found to be a constituent of lung mucous. As there are differing proteins and glands that contribute to the lung mucous, it is not surprising that there is more individual

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

differences between samples and that a certain amount of 'fuzziness' exists between these samples and salivary MUC5B.

Cluster 4 is mainly made up of cervix samples. Again the mucous constituents are a mixture of proteins, with similar to the lung mucous, MUC5B and MUC5AC dominating but additional minor contributions from MUC6 and MUC16. The positioning of cluster 4 is interesting as even if the underlying proteins are similar to those in the lung and salivary MUC5B, the group is found in between salivary MUC7 and colonic MUC2. This implies a similarity to the non-blood group displaying mucins of these two groups and most of these samples are indeed classified as non-secretors.

These results confirm the three main contribution factors to glycosylation profiles in healthy tissue types; (1) tissue location and contributing glands, (2) the underlying apo-protein and (3) expression of the glycosylation biosynthesis components.

4.2 Glycosylation in disease states versus healthy: colon

The next step was the investigation into differences due to disease state within a tissue. We started with an established model, i.e. colon. Glycosylation in the colon is confined to one mucin and is quite homogenous across samples with no blood group display. This is thought to be due to pressure of natural selection of the intestinal flora (Larsson *et al*., 2009). The overall glycosylation profile is one of core 3 type structures with sialic acid on the C-six position of the GalNAc. This is reduced to the sialyl Tn antigen (NeuAc–GalNAc) in inflamed patients. There is hardly any expression of the longer more complex structures as seen in the control samples. The clustering yields three groups. Two clusters are clear in their overall group profile; cluster 1 is healthy and cluster 3 is disease (UCa). The third cluster is a mixed profile. However, on closer examination, we can see that it can again be sub-divided into healthy and disease profiles. Some clinical data was available for these patients, but this did not yield any explanation as to why these particular disease patients did not cluster together with the others in cluster 3. The inactive ulcerative colitis patients (UCi, 14) and IBS (non-inflammatory) were used to test the colon probability model. Both of these subsets fell in the healthy probability range as defined by the sulfate ROC curve. This fits with the previous findings that the glycosylation profile of patients in remission is similar to that of control samples, and therefore the change in glyco-profile is result of the inflammation rather than a direct link to the disease state.

4.3 Glycosylation in disease states versus healthy: lung

The lung data (control and CF) showed that there is a correlated change in glycosylation with disease state, as shown previously (Schulz *et al*., 2007); exacerbated CF patients show an overall decrease in Sulf and Fuc and a concomitant increase in NeuAc. This is clearly demonstrated in the heat plot. There are three clusters, one healthy (cluster 1), one CF (cluster 3) and one mixed profile (cluster 2; Fig. 6). The move from healthy to CF is marked by an overall increase in NeuAc and a decrease in all the other parameters (clearly seen in the difference between clusters 1 and 3). The non-characteristic profile group shows a reverse of this trend for NeuAc; there seems to be an increase in the level of this residue. In general, for the rest of the parameters, the overall trend seems to follow that of 1 and 3. Interestingly, inflammatory response is believe to be mediated by a number of different lectins on the surface of leukocytes including siglecs (sialic acid-binding Ig like lectins) and L-selectin, whose main targets are sialic acid and sulfated/sialylated oligosaccharides, respectively. The probability model that best describes the separation of the two groups (control and CF) is based on the amounts of the both sialic acid and sulfation, which is validated by the importance of these two variables in the inflammation process. The present statistical analysis study also shows a shortening of overall chain length (indicated by the reduction in both Hex and HexNAc from the univariate probability models; Table 3) in the inflamed versus control subjects. This appears to be accompanied by a change from core 3/4 type structures to core 2 type structures. This can be seen in the mass spectra profile but also from Table 3, where the decrease in HexNAc from healthy to CF-type profiles is larger than that of Hex. A simple decrease in chain length, or lactosamine units would mean an equal decrease in both of these residues whereas the observed decrease in HexNAc is 2.5 compared to 1.5 for Hex. This indicates an altered expression profile of core 1 transferase and/or core 3 transferase (Schulz *et al*., 2007) in inflamed lung that is subsequently turned into core 2 and core 4 structures by the GlcNAcβ1-6 transferase.

Subsets of data were used to test the probability model (multivariate based on both NeuAc and sulfate). All of these subsets placed their profiles closer to the sialylated disease state rather than to the healthy state. This observation also includes the SMG samples from CF and control patients. This implies that the difference between disease and non-disease state in CF may be partly explained by an increase in sialylated mucin contribution to the sputum from this gland. Included also in the subset were human bronchial epithelial cells both disease and non-disease state, again both classified as closer to the disease profile by our probability model. We hypothesize that by their very nature cells are in a 'wound' like state and this may explain why the control cells also seem to have inflammatory-type glycosylation.

Overall, we believe that the use of statistics such as these presented here are the next step in understanding the complicated mechanism of inflammatory cause and effect in relation to glycosylation. These tools are particularly useful when large datasets are available and need to be reduced to gain biological insight. Future work will include applying these statistics to other types of sugar data (i.e. *N*-linked) as well as the use of data with more structural information.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Drs Jessica Holmén-Larsson and Kristina Thomsson. Medical Biochemistry, Gothenburg University.

Funding: This study was supported by the Swedish Research Council (621-2010-5322), the Swedish Foundation for International Cooperation in Research and Higher Education (IG2010-2050) and EU Marie Curie Programme grant (PIRG-GA-2007-205302). The mass spectrometer was obtained by a grant from the Swedish Research Council (342-2004-4434).

Conflict of Interest: none declared.

REFERENCES

Amerongen,A.V. *et al*. (1995) Salivary mucins: protective functions in relation to their diversity. *Glycobiology*, **5**, 733–740.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

An,H.J. and Lebrilla,C.B. (2010) Structure elucidation of native N- and O-linked glycans by tandem mass spectrometry (tutorial). *Mass Spectrom. Rev*., **30**, 560–578.

Andersch-Bjorkman,Y. *et al.* (2007) Large scale identification of proteins, mucins, and their O-glycosylation in the endocervical mucus during the menstrual cycle. *Mol. Cell Proteom.*, **6**, 708–716.

Bikker,F.J. (2004) *Salivary Agglutinin: Structure and Function*. Vrije Universiteit Amsterdam, Amsterdam.

Bozdogan,H. (1987) Model selection and akaike information criterion (Aic): the general-theory and its analytical extensions. *Psychometrika*, **52**, 345–370.

Campbell,M.P. *et al.* (2011) UniCarbKB: putting the pieces together for glycomics research. *Proteomics*, **11**, 4117–4121.

Corfield,A.P. *et al.* (2000) Mucins and mucosal protection in the gastrointestinal tract: new prospects for mucins in the pathology of gastrointestinal disease. *Gut*, **47**, 589–594.

Gill, D.J. *et al*. (2010) Location, location, location: new insights into O-GalNAc protein glycosylation. *Trends Cell Biol*., **21**, 149–158.

Hayes,C.A. *et al.* (2011) UniCarb-DB: a database resource for glycomic discovery. *Bioinformatics*, **27**, 1343–1344.

Hayes,C.A. *et al.* (2012) Glycomic work-flow for analysis of mucin O-linked oligosaccharides. *Methods Mol. Biol.*, **842**, 141–163.

Holmen,J.M. *et al.* (2004) Mucins and their O-Glycans from human bronchial epithelial cell cultures. *Am. J. Physiol. Lung Cell Mol. Physiol.*, **287**, L824–L834.

Karlsson,N.G. and Thomsson,K.A. (2009) Salivary MUC7 is a major carrier of blood group I type O-linked oligosaccharides serving as the scaffold for sialyl Lewis x. *Glycobiology*, **19**, 288–300.

Larsson,J.M. *et al.* (2009) A complex, but uniform O-glycosylation of the human MUC2 mucin from colonic biopsies analyzed by nanoLC/MSn. *Glycobiology*, **19**, 756–766.

Lauc,G. *et al.* (2010) Complex genetic regulation of protein glycosylation. *Mol. Biosyst.*, **6**, 329–335.

Marino,K. *et al.* (2010) A systematic approach to protein glycosylation analysis: a path through the maze. *Nat. Chem. Biol.*, **6**, 713–723.

Morelle,W. and Michalski,J.C. (2005) Glycomics and mass spectrometry. *Curr. Pharm. Des.*, **11**, 2615–2645.

Mortier,A. *et al.* (2011) Effect of posttranslational processing on the in vitro and in vivo activity of chemokines. *Exp. Cell Res.*, **317**, 642–654.

Novotny,M.V. *et al*. (2008) Biochemical individuality reflected in chromatographic, electrophoretic and mass-spectrometric profiles. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.*, **866**, 26–47.

Ohtsubo,K. and Marth,J.D. (2006) Glycosylation in cellular mechanisms of health and disease. *Cell*, **126**, 855–867.

Packer,N.H. *et al.* (2008) Frontiers in glycomics: bioinformatics and biomarkers in disease. An NIH white paper prepared from discussions by the focus groups at a workshop on the NIH campus, Bethesda, MD (September 11–13, 2006). *Proteomics*, **8**, 8–20.

Reis,C.A. *et al.* (2010) Alterations in glycosylation as biomarkers for cancer detection, *J. Clin. Pathol.*, **63**, 322–329.

Schulz,B.L. *et al.* (2007) Glycosylation of sputum mucins is altered in cystic fibrosis patients. *Glycobiology*, **17**, 698–712.

Schulz,B.L. *et al.* (2005) Mucin glycosylation changes in cystic fibrosis lung disease are not manifest in submucosal gland secretions. *Biochem. J.*, **387**, 911–919.

Sidebotham,R.L. *et al.* (1995) Influence of blood group and secretor status on carbohydrate structures in human gastric mucins: implications for peptic ulcer. *Clin. Sci. (London)*, **89**, 405–415.

Thomsson,K.A. *et al.* (2005) MUC5B glycosylation in human saliva reflects blood group and secretor status. *Glycobiology*, **15**, 791–804.

Vigerust,D.J. (2011) Protein glycosylation in infectious disease pathobiology and treatment. *Cent. Eur. J. Biol.*, **6**, 802–816.

Wine,J.J. and Joo,N.S. (2004) Submucosal glands and airway defense. *Proc. Am. Thorac. Soc.*, **1**, 47–53.

Wuhrer,M. *et al*. (2011) Mass spectrometric glycan rearrangements. *Mass Spectrom. Rev.*, **30**, 664–680.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016