Chapter 4

Plasma Glycoprotein Profiling for Colorectal Cancer Biomarker Identification by Lectin Glycoarray and Lectin Blot

4.1 Introduction

Colorectal cancer (CRC) is the third most common cancer worldwide with an estimated one million new cases and a half million deaths each year, largely due to the insidious onset of the disease [1]. Although colorectal cancer accounts for 14% of all cancer-related deaths, the overall 5-year survival rate could be around 90% if the cancer was diagnosed and treated early, while the tumors were still localized [2]. The most widely used screening procedures for CRC include fecal occult blood testing, flexible sigmoidoscopy, double-contrast barium enema, and colonoscopy [3-5]. However, as these screening methods are not pleasant, there is patient resistance to undergoing the procedure. Thus, serum- or plasma-based screening methodologies for CRC detection are likely to be far more acceptable than the current screening options and would likely greatly increase the percentage of the population screened.

There is currently great interest in proteomics-based plasma biomarkers with utility for the detection of cancer, as well as to follow the efficacy of therapeutic intervention [6]. The plasma proteome may contain not only normal plasma proteins, but also the different cleavage products and proteins with different post-translational modifications that may result from a given disease state [7]. Within the plasma proteome, the plasma or serum glycoproteome is both functionally important and perhaps the most abundant post-translational modification [8-23]. Many existing cancer biomarkers are glycoproteins, such as Her2/neu in breast cancer, prostate-specific antigen (PSA) in prostate cancer, CA125 in ovarian cancer, and carcinoembryonic antigen (CEA) in colorectal, bladder, breast, pancreatic and lung cancer. Although CEA remains the most widely used serum glycoprotein biomarker in CRC, poor sensitivity and specificity precludes its use for the detection of CRC [24]. Other tumor antigens that have been proposed as serum glycoprotein biomarkers of colorectal cancer include CA 19-9, CA 242 , CA-195, CA 50, CA 74-2, and TIMP-1(tissue inhibitor of metalloproteinase-1) [25-32]. However, as with CEA, these biomarkers have poor sensitivity and specificity, and are not suitable for screening or diagnostic purposes. Alterations in either the level of or type of glycosylation has been shown to influence cellular processes such as growth, differentiation, transformation, adhesion, metastasis and immune surveillance of tumors[33-35]. In particular, aberrant sialylation in cancer cells is thought to be a characteristic feature associated with malignant properties including invasiveness and metastatic potential. An increase in sialylation is commonly observed in various tumors, which may be due to either an increased activity of sialyltransferases or increased numbers of possible sialylation sites on *N*-linked carbohydrates [36]. Recent studies in prostate cancer have also shown changes in fucosylation associated with progression of the disease [37, 38]. Thus, it is of great interest to identify and validate plasma glycoproteins with altered glycosylation whose function may reveal insight into critical events in cancer progression and may have utility as potential markers for cancer detection. In this study, we utilized multi-dimensional liquid chromatography protein separation followed by lectin glycoarrays for screening *N*-glycosylation pattern changes in plasma from patients with colorectal cancer. Bioinformatics analysis of the data facilitated the identification of plasma glycoproteins that possess altered glycosylation. These potential biomarkers were subsequently validated with a second set of independent plasma samples. These glycoprotein biomarkers may have utility for the detection of colorectal cancer.

4.2 Materials and Methods

4.2.1 Plasma Samples

Human plasma samples were collected through a four institutional consortium (Dana Farber Cancer Institute, MD Anderson Cancer Center, St. Michael Hospital, Toronto, Ont, Canada, and University of Michigan Medical Center) of the Early Detection Research Network (EDRN). Human subjects were identified prior to endoscopy and informed consent obtained prior to sample collection procedures specified in a protocol approved by Institutional Review Boards at all collaborating Institutions. The samples were collected, handled, shipped, stored, and managed according to standard operating procedures as specified in the protocol document. Samples were labeled with bar coded subject identification number and tracked from collection through assay via a relational database containing de-identified demographic and clinical data located at the Bioinformatics Unit at Dartmouth Medical College. The samples were stored in a professional repository facility at -80°C until use. The plasma was obtained from 6 patients with colorectal cancer (two stage II, two stage III and two stage IV), 5 samples from patients with colonic adenomas (polyp size with 0.3-1.3 cm), and 9 samples from patients with normal colonoscopes for use in a blinded set to screen *N*-linked glycosylation pattern changes on plasma glycoproteins and 30 plasma samples (10 of each) for use in a testing set. All subjects that donated plasma for this study were between 50-76 years of age with 16 Caucasians and 4 African Americans. The plasma was aliquoted into 0.5 ml aliquots, frozen, and then stored at -80°C until assayed.

4.2.2 Preparation of Glycoprotein Samples for Lectin Glycoarrays or Lectin Blot

4.2.2.1 Delipidation and Immunodepletion of the Plasma Samples

The plasma samples were delipidated by centrifugation for 15 min at 20,000 × g, and the lipid containing upper layer was removed before depletion. 250 μL of the delipidated plasma was depleted using the ProteomeLab IgY-12 LC10 proteome partitioning kit (Beckman Coulter, Fullerton, CA). This procedure enables simultaneous removal of twelve highly abundant proteins from human plasma, including albumin, IgG, α1- antitrypsin, IgA, IgM, transferrin, haptoglobin, α1-acid glycoprotein, α2-macroglobin, apolipoprotein A-I, apolipoprotein A-II, and fibrinogen. Using optimized buffers for sample loading, washing, eluting, and regeneration, the resulting flow-through (unbound) fraction and the eluted (bound) fraction were collected separately during a total of 75 min IgY affinity separation cycle. The final depleted fraction was buffer exchanged into 2 mL Concanavalin A (Con A) binding buffer (20 mM Tris, 0.15 M NaCl, 1 mM Mn^{2+}, and 1mM Ca^{2+}, pH 7.4) with a 10,000 Da molecular weight limit Amicon Ultra-15 centrifugal filter (Millipore, Billerica, MA). The protein concentration of the final concentrated fraction was determined using a Bradford protein assay (Bio-Rad, Hercules, CA) with BSA as a standard. The concentrations of the immunodepleted plasma samples were approximately 1.5-2.0 mg/mL.

4.2.2.2 *N*-Glycoprotein Enrichment with ConA Affinity Capture

ConA columns were prepared by adding 1.5 mL of the agarose-bound lectin (Vector Labs, Burlingame, CA) into 5 mL polypropylene columns (Pierce Biotechnology, Rockford, IL). The columns were first equilibrated with 5 mL of the binding buffer before use. A total of 500 μL of the immunodepleted plasma was loaded onto an equilibrated column. After incubating for 30 min, the unbound proteins were washed out with 6 mL of the binding buffer, and the captured proteins were eluted with 4 mL of the elution buffer (20 mM Tris, 0.5 M NaCl, 0.5 M methyl-R-D-mannopyranoside pH 7.4). The protein recovery of the lectin column was determined based on the Bradford protein assay, using BSA as the standard.

4.2.2.3 HPLC Separation of Glycoproteins

25 μg of the enriched *N*-glycoproteins (corresponding to around 60 μL original plasma) was separated by NPS-RP-HPLC at a flow rate of 0.5 mL/min on a 33 × 4.6 mm ODS III column (Eprogen, Darien, IL, USA) using a ProteomeLab PF2D system (Beckman Coulter, Fullerton, CA, USA). The separation was performed using a water (solvent A) and acetonitrile (solvent B) gradient as follows: (1) 5 to 25% B in 1 min; (2) 25 to 31% B in 2min; (3) 31 to 37% B in 7 min; (4) 37 to 41% B in 8 min; (5) 41 to 48% B in 7 min; (6) 48 to 58% B in 3 min; (7) 58-75% B in 1 min; (8) 75 to 100% B in 1 min. Proteins eluted from the column were collected by an automated fraction collector (FC 204 BE, Beckman-Coulter) controlled by an in-house-designed DOS-based software

program and 32 Karat software (Beckman-Coulter). The 32 Karat software was also used to calculate the peak area of each protein fraction.

4.2.3 Lectin Glycoarrays

After completely drying, the protein fractions were resuspended with 15 µL printing buffer (65 mM Tris-HCl, 1% SDS, 5% DTT, and 1% glycerol) in 96 well plates, and then arrayed on nitrocellulose slides (Whatman, Keene, NH) using a non-contact piezoelectric printing device (Nanoplotter 2.0, GeSiM, Germany). 2.5 nL of each fraction were arrayed on the nitrocellulose slides in spots that were 450 µm in diameter and 600 µm apart. The printed slides were dried for one day after being blocked overnight with 1% BSA in phosphate buffered saline with 0.1% Tween 20 (PBS-T). The blocked slides were first incubated with biotinylated lectin for 2 hrs and then with 1 µg/mL streptavidin conjugated to Alexaflor555 fluorescent dye (Invitrogen, Carlsbad, CA). After being washed and dried, slides were scanned in the green channel using an Axon 4000A scanner. Image analysis was performed using the GenePix 6.0 software (Molecular Devices, Sunnyvale, CA).

4.2.4 Statistical Analysis of Lectin Glycoarray Data

4.2.4.1 Principal Components Analysis (PCA)

Principal components analysis (PCA) was performed for data visualization, which was carried out using log-transformed and normalized array spot intensities. The leading two eigenvectors of the sample covariance matrix were used for visualization. In this study, 20 plasma samples (processed in duplicate when using ConA, AAL, PNA and triplicate when using SNA and MAL) were placed in a two dimensional scatter plot using

PCA. Sample pairs falling close together in the scatter plot are more similar in terms of their overall patterns of normalized glycoform abundances.

4.2.4.2 Hierarchical Clustering

An unsupervised hierarchical clustering (HC) procedure was used without any prior knowledge of grouping to find criteria appropriate for classifying the cases according to the glycosylation pattern from glycoarrays. To do this, the normalized array spot intensities were log transformed, and the pair-wise Pearson correlations were used to carry out HC in which more closely correlated pairs of samples were joined at a lower point on the dendrogram. The scale on the dendrograms was 100 -100 × r, where r is the Pearson correlation coefficient. In the HC analysis, the replicate averages of the 20 distinct biological specimens were used.

4.2.4.3 Z-statistics

For differential abundance analysis, Z-statistics for each protein detected by each lectin were calculated. Comparisons were made of normal versus adenoma, normal versus cancer as well as adenoma versus cancer. Based on the Bonferroni correction for two-sided testing of 36 peaks, Z values of ≥ 3.2 or ≤ -3.2 could be deemed to have significantly different glycosylation levels at a 95% significance level.

4.2.5 SDS-PAGE and Lectin Blot

To identify and validate the glycoproteins of interest as markers of CRC, protein fractions from NPS-RP-HPLC were further separated by 1-D SDS-PAGE using the Mini-Protean cell (Bio-Rad, Hercules, CA) at 80V. The resolved proteins were stained with colloidal Coomassie (Invitrogen) or transferred onto a polyvinylidene fluoride (PVDF) membrane (Bio-Rad). The PVDF membrane was blocked with 5% w/v BSA (Roche, Indianapolis, IN) in PBS-T overnight at 4°C and then incubated with biotinylated AAL and SNA respectively (2 μg/mL in PBS-T containing 3% BSA) for 1 hr at room temperature. The membrane was then washed and incubated with a 100 ng/ml streptavidin-HRP in PBS-T containing 3% BSA. After washing, the signal was visualized using a chemiluminescence detection system (ECL, Pierce) and detected on blue sensitive autoradiography film (Marsh Bio Products, Rochester, NY). Potential glycoprotein biomarkers are detected by lectin blot experiment and the corresponding colloidal blue stained bands were identified by nano-LC MS/MS.

4.2.6 Protein Digestion

4.2.6.1. Tryptic Digestion and N-deglycosylation of NPS-RP-HPLC Fractions

The NPS-RP-HPLC fractions with significantly different glycosylation were dried completely, denatured in 40μL of 100 mM NH_4HCO_3 buffer (pH 7.8), then reduced with 1 mM dithiothreitol (DTT) for 45 min at 56°C and alkalized with 15mM iodoacetamide (IAA) for 1 h at room temperature in the dark. The proteins were then digested with 1-2 μg of TPCK-treated trypsin (Promega, Madison, WI, USA) for 18 h at 37°C. The reaction mixture was then heated for 10 min at 95°C to stop the action of trypsin. 1-2 μL of PNGase F (New England BioLabs, Ipswich, MA) were added to half of the tryptic digest mixture from each fraction to start the N-deglycosylation at 37°C for 12 h. The other half was stored at -80° for later use.

4.2.6.2. Tryptic Digestion and *N*-deglycosylation of SDS Gel Bands

The glycoprotein bands from the colloidal Coomassie blue-stained SDS-PAGE gel were carefully excised. The gel pieces were placed in siliconized Eppendorf tubes (Sigma) and destained 3 times with 200 µl 200 mM ammonium bicarbonate and 40% acetonitrile at 37°C for 30 min each and lyophilized completely in a SpeedVac (Thermo). The dried gel pieces were first deglycosylated by incubating with 10 µL of the PNGase F solution (Sigma) overnight at 37°C followed by trypsin digestion overnight at 37°C. The liquid from the gel piece was transferred to a new tube for nano-LC MS/MS analysis.

4.2.7 LC-MS/MS for Protein Identification and Glycosylation Site Determination

A MS4B MDLC system (Michrom Bioresources, Auburn, CA) interfaced with a linear ion trap mass spectrometer (LTQ, Thermo, San Jose, CA) was used to analyze the tryptic digests from SDS-PAGE. The injected peptide sample was first desalted on a trap column (150 µm × 50 mm, Michrom Bioresources Inc, Auburn, CA) with 3% solvent B (0.3 % formic acid in acetonitrile) at 50 µl/min for 5 min and then released and separated on a nano column (150 µm × 150 mm, Michrom) using a 45 min gradient from 3% B to 95% B at 0.3 µl/min. The resolved peptides were directly introduced into an ESI ion source with the spray voltage set at 2.6 kV.

To identify the eluted peptides, data dependent MS/MS analysis (m/z 400-2000) was performed using MS acquisition software (Xcalibur 1.4, Thermo Finnigan), in which a full MS scan was followed by seven MS/MS scans of the seven most intense precursor ions. All MS/MS spectra were compared against the Swiss-Prot FASTA human protein database using the SEQUEST algorithm incorporated into the TurboSequest feature of Bioworks 3.1 SR1.4 (Thermo Finnigan). By allowing up to two missed cleavages, positive protein identification was accepted for a peptide with X_{corr} of greater than or equal to 3.5 for triply-, 2.5 for doubly- and 1.9 for singly charged ions, and all with ΔCn ≥0.1 [39]. The sequence database search was set to accept the following modifications: carboxymethylated cysteines, oxidized methionines, and an enzyme-catalyzed conversion of asparagines to aspartic acids (0.984Da shift) at an *N*-glycosylation site. Accuracy of the SEQUEST assignment of MS/MS spectra to peptide sequences was validated by the TransProteomics Pipeline which includes both PeptideProphet and ProteinProphet software. In this study, peptides were identified with a probability cut-off of $p \geq 0.99$ and protein identifications were confirmed with probability scores of at least 0.9.

4.3 Results and Discussion

4.3.1 Reduction in the Complexity of Plasma Glycoprotein Mixtures by Immunoaffinity Depletion and Lectin Affinity Enrichment

The strategy undertaken during analysis of protein glycosylation will vary depending on the amount of available sample. It is more challenging to determine alterations in glycan structure when the glycoproteins are present in a complex biological medium at a very wide dynamic range, such as in human serum or plasma. In order to analyze glycoproteins expressed in the midlevel abundance range, the most abundant proteins were immunodepleted from the sample, as show in the schematic flowchart of the proposed method (Figure 4.1). 250 µl of each plasma sample was first delipidated and then immunodepleted to remove the lipids and the 12 most abundant plasma proteins based on an avian antibody (IgY)-antigen interaction. Around 7% of the total protein mass in the plasma samples remained after the immunodepletion step (Table 1). The representative chromatograms resulting from the immunodepletion of plasma from normal, adenoma, and colorectal cancer patients indicate the reproducibility of this step (Figure 4.2A).

Following immunodepletion, the remaining proteins in the flow-through fraction were subjected to ConA affinity chromatography to enrich the concentration of *N*-glycosylated proteins. With broad specificity and high affinity, ConA binds with preference to oligomannosidic, hybrid, and bi-antennary *N*-glycans, either unconjugated or attached to proteins or peptides [40]. O-glycopeptides or glycoproteins that contain exclusively O-glycosylation sites are not bound by this lectin. Approximately 70% of the immunodepleted plasma protein content was recovered by ConA affinity chromatography. By selectively isolating the glycoproteins from the immunodepleted plasma proteome, the procedure achieved a significant reduction in analyte complexity at two levels: first, the immunodepletion of the 12 most abundant proteins significantly increases the dynamic range of detection by approximately 90-fold and, additionally, reduces sample heterogeneity due to the removal of the highly variable IgG, IgA and IgM proteins; second, the subsequent *N*-glycoprotein enrichment step affords another effective means of reducing plasma sample complexity.

25 μg of the enriched *N*-glycoprotein mixtures were further separated by NPS-RP-HPLC into 36 fractions for lectin glycoarray or lectin blot analysis. Figure 4.2B shows the reverse phase chromatograms of the plasma samples from different physiological status (9 normals, 5 adenomas, and 6 colorectal cancers). The reproducibility of these chromatograms indicates that the samples from the plasma from normal subjects, from adenoma patients, and from colorectal cancer patients were very similar at the protein expression level. These results suggest that the analysis of

glycoprotein expression alone does not provide valuable information to differentiate the clinical status of individuals.

4.3.2. Lectin Glycoarrays for Identification of *N*-glycosylation Pattern Changes

To analyze the plasma glycosylation patterns, all fractions containing the separated intact glycoproteins from the NPS-RP-HPLC separation were then arrayed on nitrocellulose slides, in duplicate, as unique spots. Subsequently, the slides were screened to analyze the different glycan structures using five different lectins. The following biotinylated lectins were used: *Aleuria aurentia* lectin (AAL), *Sambucus nigra* bark lectin (SNA), *Maackia amurensis* lectin II (MAL), peanut agglutinin (PNA), and Concanavalin A (ConA). AAL binds fucose linked (α-1, 6) to *N*-acetylglucosamine or (α -1, 3) to *N*-acetyllactosamine related structure. Both MAL and SNA recognize sialic acid on the terminal branches. MAL detects glycans containing NeuAc-Gal-GlcNAc with sialic acid at the 3 position of galactose while SNA binds preferentially to sialic acid attached to terminal galactose in an (α-2,6) and an (α -2,3) linkage at a lesser degree. In contrast, PNA binds desialylated exposed galactosyl (β-1, 3) *N*-acetylgalactosamine. ConA recognizes α -linked mannose including high-mannose-type and hybrid-type structures. The utilization of these five lectins have been proved to be highly successful in covering >95% of *N*-glycan types reported and differentiating them according to their specific structures [41]. Figure 4.3 shows the array image with five different lectins for comparing *N*-glycosylation levels and types in plasma from normal, adenoma and colorectal cancer patients. Since only variations in glycan expression were of interest, all array spot intensities were normalized by dividing the corresponding UV peak area to eliminate protein abundance differences. The normalized array data suggest that the overall levels of protein fucosylation and sialylation are higher in colorectal cancer and adenoma plasma samples as compared to the normal plasma controls.

4.3.3 Statistical Analysis of *N*-glycosylation Pattern Changes

Principal components analysis (PCA) and hierarchical clustering (HC) of the normalized glycoarray data were performed to differentiate the plasma samples in terms of their overall *N*-glycosylation patterns and to relate these patterns to clinical status. In PCA, 20 plasma samples assayed in duplicate (ConA, AAL, PNA) or triplicate (MAL, SNA) were analyzed separately for each lectin. The scores of the first two principal components of the normal, adenoma, and colorectal cancer samples are illustrated in a 2-dimensional scatter plot in which each sample was plotted as an individual point (Figure 4.4A). Closer points corresponded to greater similarity in the patterns of glycoprotein expression over all 36 protein spots on the microarrays. When ConA and SNA were hybridized against the arrays, the normal controls (red) were grouped separately from cancer (blue) and adenoma samples (green), while most cancer and adenoma samples were clustered together. In response to AAL and MAL, the normal and cancer samples generally segregated from each other, whereas the adenoma samples overlapped to some extent with both. The PNA-based microarrays did not provide robust fluorescence intensities for most protein spots, however broadly similar results compared to AAL and MAL arrays were observed. Except for PNA, all the other lectins clearly separate the cancer samples from normal controls. The results of the PCA analysis suggest that lectin glycoarrays may have utility as a diagnostic tool to discriminate the diseased states from the non-diseased states in cancer detection. The excellent concordance among the replicates from the same sample (Figure 4.4B) in PCA results indicates that the lectin glycoarray is a robust strategy for screening *N*-glycosylation changes among the plasma samples from different disease states. As expected, similar results were observed in HC by using the Pearson correlation coefficient for distance metrics (Figure 4.5). The clustering results for fucosylated, sialylated and mannosylated glycan expression generally distinguished the normal plasma samples from the cancer and adenoma samples. The results from the different lectins indicate the effectiveness of using multi-lection detection to differentiate plasma samples of the different clinical states based on *N*-glycosylation pattern changes.

As an alternative means of analyzing the lectin glycoarray data we calculated Z-statistics of each array spot to search for signature proteins that might differentiate the plasma samples of the different clinical states (Table 2A). Comparisons were performed of normal versus adenoma (N/A), normal versus cancer (N/C), and adenoma versus cancer (A/C). Z values of ≥ 3.2 or ≤ -3.2 were selected as differential glycosylation at a 95% significance level. A positive Z value indicates elevated glycosylation and a negative Z value means reduced glycosylation.

4.3.4 Identification of Plasma Biomarkers with Altered *N*-glycosylation

Due to the co-elution during the NPS-RP-HPLC separation, there were cases where more than one protein was observed in certain fractions. In order to determine which co-eluting protein was responsible for the differential responses in the lectin-based microarrays, the fraction with altered glycosylation was further separated by 1-D SDS-PAGE and then analyzed by lectin blot. Since the elevated fucosylation and sialylation levels in colorectal cancer plasma were detected on most of the differentially glycosylated proteins, we chose AAL and SNA in the lectin blot analysis to determine which protein corresponded to the differential fucosylation and sialylation pattern.

The corresponding protein bands in the SDS gel with significantly differential glycosylation pattern in colorectal cancer or adenoma were excised, and then digested with PNGase F and trypsin. Protein identification and the possible glycosylation sites were determined by nano LC-MS/MS. Positive identification was validated by the Trans-Proteomics pipeline which includes PeptideProphet and ProteinProphet software. PeptideProphet software was used to effectively identify correct peptide assignments and ProteinProphet was used to validate the protein identifications. In this study, peptides were identified with probability scores of at least 0.99 with a false positive error rate of 0.0007 and proteins were identified with a probability cut-off of $p \geq 0.9$ which corresponds to a 0.7% error rate [42, 43]. Figure 4.6A shows a representative nano-LC-ESI-MS/MS spectrum of a deglycosylated glycopeptide [$(M+2H)^{2+}$ at m/z 553.20] from complement C4. The localization of the *N*-glycosylation site was determined by a mass increase of 1 Da on the N-X-(S/T) sequence after deamidation of asparagine residue into aspartic acid [44]. The b- and y- series of product ions clearly showed a mass shift indicative of conversion of asparagine to aspartic acid at the original site of *N*-glycosylation. In this case, the mass difference of 115 Da for aspartic acid found for both the b_3-b_2 and y_9-y_8 product ion pairs suggests the original *N*-glycosylation at residue 3. Figure 4.6B shows another example of a peptide [$(M+2H)^{2+}$ at m/z 716.82] from kininogen-1. Again a shift of 115 Da for both b_6-b_5 and y_7-y_6 indicates the *N*-glycosylation at residue 6, while b_5-b_4 and y_8-y_7 yield a difference of 114 Da indicating that the Asn at position 5 was not *N*-glycosylated. The significant differentially glycosylated proteins with their Z statistics are summarized in Table 2A and the corresponding detected glycosylation sites are shown in Table 2B, in which 10 proteins displayed significant differential glycosylation among normal, adenoma, and cancer samples. Three of these proteins showed elevated glycosylation in the case of cancer compared to normal and adenoma and seven had higher glycosylation levels in cancer and adenoma compared to normal. The data suggests that Z-statistic analysis of lectin glycoarrays has utility to identify cancer samples relative to adenoma or normal controls. The potential markers to distinguish colorectal cancer from adenoma and normal identified in this study include the elevated sialylation and fucosylation in complement C3, histidine-rich glycoprotein, and kininogen-1.

4.3.5 Lectin Blot of a Control Set for Detection of Potential Biomarkers for Differentiating the Different Clinical States

The correlation between the potential biomarkers and a diagnosis of colorectal cancer or adenoma has been confirmed in a blinded sample set with 30 plasma samples (10 colorectal cancers, 10 adenomas, and 10 normal). The plasma samples were depleted, enriched and separated by multi-dimensional HPLC separation as described previously and then analyzed by 1-D SDS-PAGE, followed by lectin blotting. Figure 4.7 shows representative protein bands after lectin blot which characterized the CRC samples. As shown, complement C3 in all of the normal and adenoma samples barely responded in either the AAL or SNA blot, but was observed to have significantly elevated fucosylation and sialylation in the colorectal cancer samples. These results from lectin blot analysis were consistent with that obtained from the Z-statistic analysis in which complement C3 in cancer was significantly elevated in response to both AAL and SNA as compared to adenoma and normal. The peak areas of complement C3 in each plasma sample indicated that approximately equal amounts of protein were loaded. Additionally, histidine-rich protein displayed significant differential glycosylation but similar protein expression. In this case, fucosylation was found to be significantly elevated in colorectal cancer samples as compared to both adenoma and normal samples, while similar sialylation was observed in cancer and adenoma. Again, the total amounts of histidine-rich protein were quite similar among the plasma samples from normal, adenoma and cancer patients. These results highlight the potential utility of the altered glycosylation patterns instead of absolute protein expression as markers for cancer detection and further increase the specificity of these potential markers.

4.4 Conclusion

We have described a glycoproteomic strategy for the identification of potential plasma biomarkers in the detection of colorectal cancer. The strategy was based on the reduction of plasma complexity by immunodepletion and subsequent glycoprotein enrichment, the screening of glycan pattern changes by the lectin glycoarray format, and the identification of potential markers with altered glycosylation using lectin blot analysis and nano-LC MS/MS. As indicated by the peak intensities from NPS-RP-HPLC separations, the absolute protein amounts of each clinical state were quite similar so that the plasma glycoproteome alone does not differentiate the clinical status of individuals. By using lectin glycoarrays, normal, adenoma, and colorectal cancer plasma showed distinct clustering results of each state following the PCA and HC analysis. In this study, patients diagnosed with either colorectal cancer or adenomas have dramatically higher levels of sialylation and fucosylation compared to the normal controls. The glycoprotein fractions for analysis were identified using SDS-PAGE and lectin blot coupled with nano-LC MS/MS. The potential markers identified in this study to distinguish colorectal cancer from adenoma and normal include elevated sialylation and fucosylation in complement C3, histidine-rich glycoprotein, and kininogen-1. These results demonstrated the usefulness of this strategy for the identification of the *N*-linked glycan patterns to distinguish individuals in different clinical states as well as the identification of potential biomarkers of CRC in human plasma based upon changes in glycan structure rather than in the protein level.

Table 4.1 The amount of protein processed through the IgY antibody column and recovered in the depleted fraction from 250 μL plasma samples.

Samples	**Protein Amount (mg)**	
	Original Plasma	**Depleted Fraction**
Cancer (n=6)	23.67 ± 3.52	1.68 ± 0.22
Adenoma (n=5)	22.98 ± 3.58	1.61 ± 0.23
Normal (n=9)	21.52 ± 3.81	1.52 ± 0.29

Table 4.2A Z-statistics of differentially glycosylated proteins detected by lectins.

	ConA			AAL			MAL			SNA			PNA		
Protein ID (Access Number)	**N/A[a]**	**N/C[b]**	**A/C**	**N/A**	**N/C**	**A/C**	**N/A**	**N/C**	**A/C**	**N/A**	**N/C**	**A/C**	**N/A**	**N/C**	**A/C**
Proteins that are significantly different in cancer than those in adenoma and normal															
Complement C3 (P01024)	0.6	-0.09	-1.93	-1.9	**-4.19**	**-3.35**	-2.91	**-5.38**	**-4.05**	-1.72	**-5.77**	**-3.35**	-1.51	**-6.21**	**-4.61**
Kininogen-1 (P01042)	**-4.86**	**-6.48**	0.01	**-5.04**	**-7.22**	**-3.44**	-2.73	**-7.64**	**-4.75**	**-6.68**	**-10.0**	**-3.24**	-1.26	**-4.67**	-2.94
Histidine-rich glycoprotein (P04196)	-1.25	-2.23	-0.53	-0.55	**-4.03**	**-3.64**	0.75	-1.89	-2.86	-1.37	**-3.33**	-2.44	0.84	-0.98	-2.52
Proteins that are significantly different in cancer and adenoma than those in normal															
Alpha-1B-glycoprotein (P04217)	**-3.29**	**-3.94**	-0.75	-3.04	**-6.94**	-1.43	-1.65	-2.65	-0.93	**-3.24**	**-5.13**	-1.65	0.04	**-5.17**	**-4.69**
Hemopexin (P02790)	**-6.41**	**-5.86**	1.32	**-6.68**	**-5.77**	0.18	-2.95	-3.03	-0.12	**-7.62**	**-7.01**	0.58	-1.28	-2.57	-0.8
Complement factor I (P05156)	-2.57	**-3.89**	-0.52	-2.32	**-3.28**	-1.07	-0.98	-2.09	-1.11	**-3.48**	**-5.60**	-1.15	-0.44	**-4.28**	-2.81
Ceruloplasmin (P00450)	**-4.61**	**-4.30**	0.50	**-4.06**	**-4.57**	-0.94	-3.00	-2.52	-0.3	**-5.06**	**-6.65**	0.24	-0.01	**-4.02**	**-3.61**
Afamin (P43652)	**-4.47**	**-4.35**	0.82	**-3.86**	**-4.80**	-2.04	-0.29	-2.11	-2.09	**-4.19**	**-4.34**	-0.74	-2.38	-1.64	-1.30
Alpha-1-antichymotrypsin (P01011)	**-4.21**	**-5.96**	0.89	-3.14	**-5.32**	0.27	-1.13	-1.05	0.47	**-4.07**	**-5.82**	1.08	-1.34	0.48	1.68
Complement C4 precursor (P01028)	**-3.80**	**-5.95**	0.07	**-3.42**	**-6.05**	0.69	-2.56	-2.34	1.59	**-4.22**	**-6.09**	0.95	-1.88	-2.25	0.41

[a], [b]: N: normal; A: adenoma; C: cancer. The highlighted ($Z \geq 3.2$ or $Z \leq -3.2$) correspond to 95% significant level with multiple testing correction.

Table 4.2B Differentially glycosylated proteins identified with the glycosylation sites.

Protein ID (Access #)	MW/pI	Peptide Sequence	Glycosylation site	MH+
Histidine-rich glycoprotein (P04196)	59541.9/7.09	R.VIDF**N***C#TTSSVSSALANTK.D	125	2017.96
		R.HSHN**N***NSSDLHPHK.H	344	1623.74
Kininogen-1 (P01042)	71901.1/6.34	K.LNAEN**N***ATFYFK.I	294	1431.69
Hemopexin (P02790)	51644.3/6.55	R.SWPAVG**N***C#SSALR.W	187	1347.65
		K.ALPQPQ**N***VTSLLGC#TH.-	453	1678.86
Complement factor I (P05156)	65677.6/7.72	K.FLN**N***GTC#TAEGK.F	103	1254.58
Alpha-1B-glycoprotein (P04217)	54239.6/5.58	R.EGDHEFLEVPEAQEDVEATFPVHQPG**N***YSCSYR.T	179	3779.65
Ceruloplasmin (P00450)	122128.6/5.44	K.AGLQAFFQVQEC#**N***K.S	358	1595.69
		K.EHEGAIYPD**N***TTDFQR.A	138	1892.84
Afamin (P43652)	69025.0/5.64	R.YAEDKF**N***ETTEK.S	402	1474.67
		R.DIENF**N***STQK.F	33	1195.56
Alpha-1-antichymotrypsin (P01011)	47621.5/5.33	K.YTG**N***ASALFILPDQDK.M	271	1752.88
Complement C3 (P01024)	187046.9/6.02	K.TVLTPATNHMG**N***VTFTIPANR.E	85	2260.08
		K.HYLMWGLSSDFWGEKP**N***LSYIIGK.D	1617	2841.41
Complement C4 (P01028)	192651.5/6.66	R.FSDGLES**N***SSTQFEVK.K	226	1774.81
		R.GL**N***VTLSSTGR.N	1328	1104.60

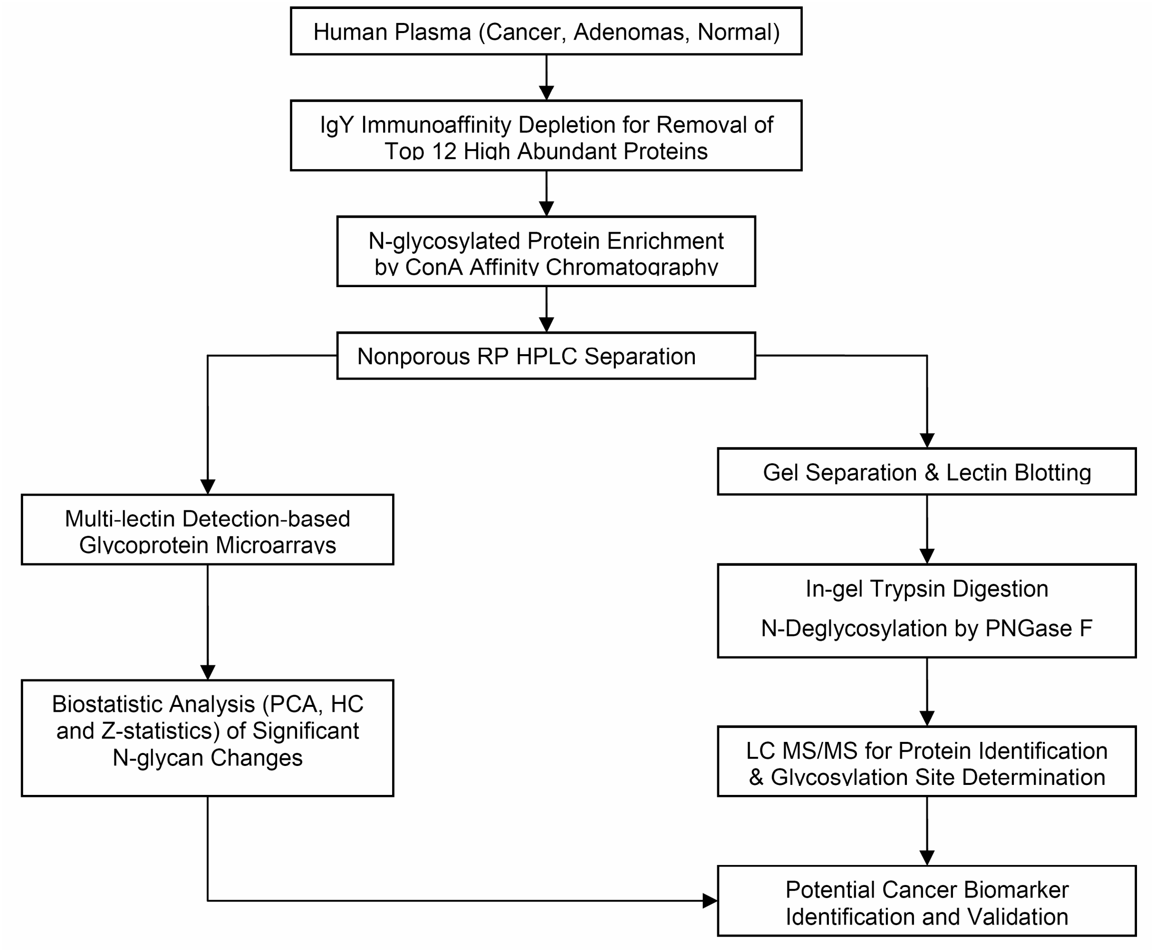
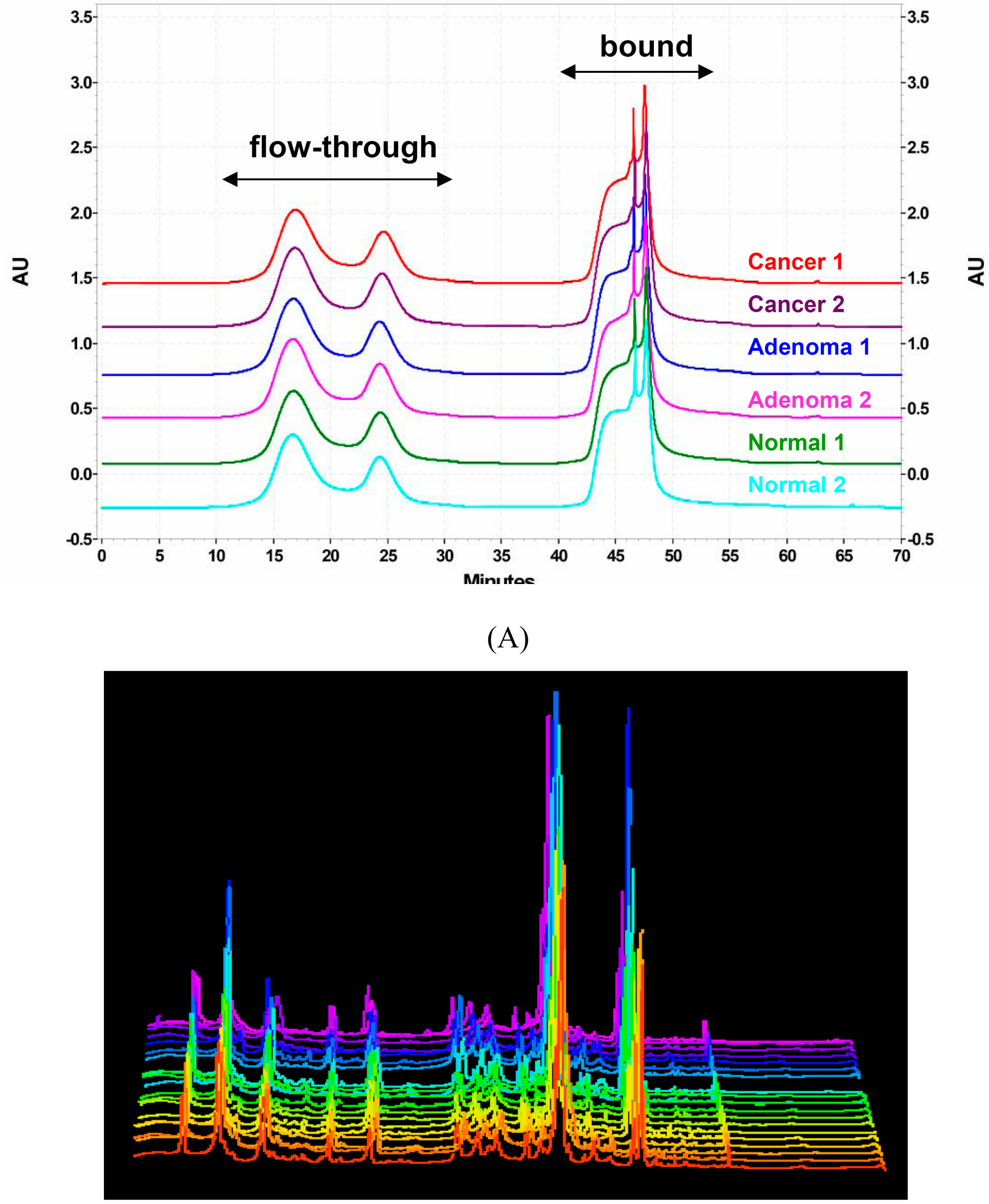

Human Plasma (Cancer, Adenomas, Normal)
IgY Immunoaffinity Depletion for Removal of Top 12 High Abundant Proteins
N-glycosylated Protein Enrichment by ConA Affinity Chromatography
Nonporous RP HPLC Separation
Multi-lectin Detection-based Glycoprotein Microarrays
Biostatistic Analysis (PCA, HC and Z-statistics) of Significant N-glycan Changes
Gel Separation & Lectin Blotting
In-gel Trypsin Digestion
N-Deglycosylation by PNGase F
LC MS/MS for Protein Identification & Glycosylation Site Determination
Potential Cancer Biomarker Identification and Validation

Figure 4.1 Schematic presentation for high throughput analysis of plasma N-glycosylation pattern changes in colorectal



(A)

(B)

Figure 4.2 (A) Representative chromatographic profiles of immunoaffinity depletion of plasma from normal, adenoma, and colorectal cancer patients using ProteomeLab IgY-12 kit. The 12 most abundant proteins are contained in the "bound" fraction while the less abundant proteins in plasma or serum remained in the "flow-through" fraction. (**B)** UV chromatograms of all the plasma samples from colorectal cancer, adenoma, and normal controls. The similarity among these UV chromatograms among different samples indicated that proteins undergo heterogeneity of glycosylation modifications without necessarily changing protein expression.

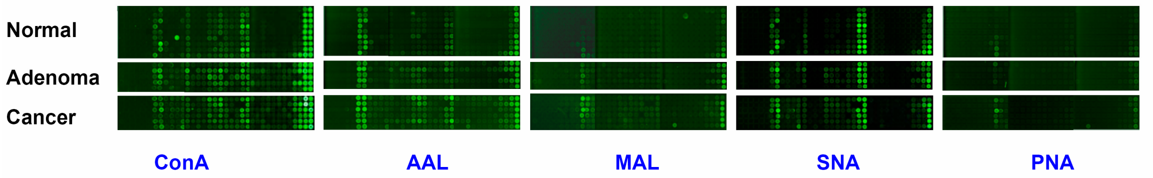


Figure 4.3 Microarray images of N-glycosylated proteins separated from NPS-RP-HPLC with different lectins.

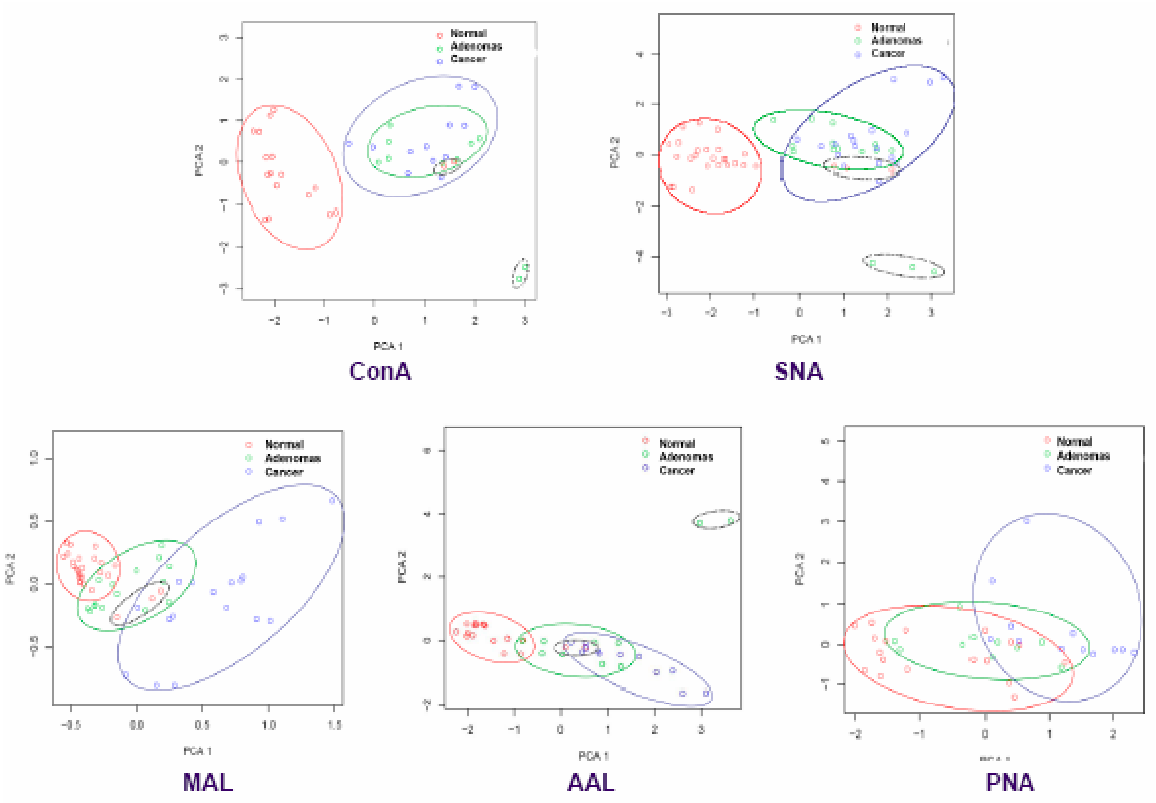


Figure 4.4 (A) PCA plot for normalized glycoprotein microarray data derived from the replicates of healthy individuals, adenoma, and colorectal cancer patients. Ovals indicate the areas where the data points of the three groups are distributed.

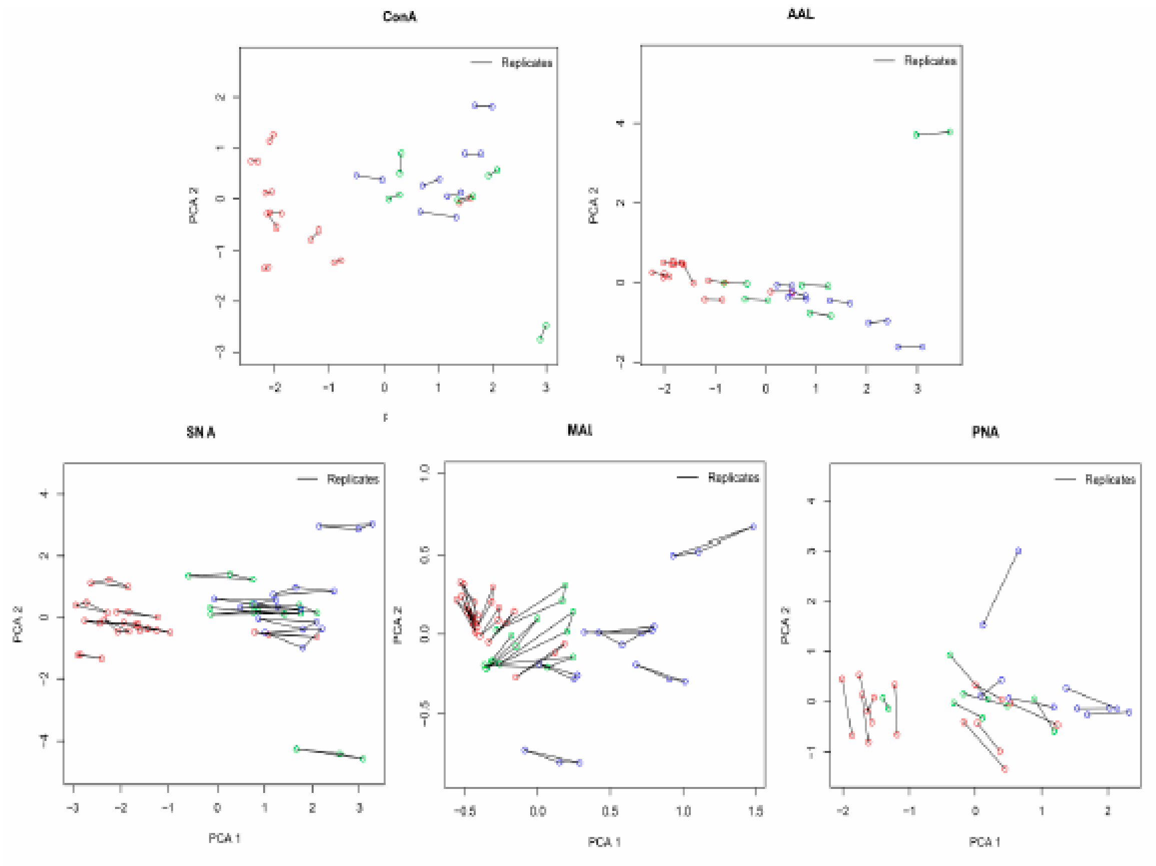


Figure 4.4 (B) Reproducibility demonstration of PCA for normalized glycoprotein microarray data derived from the replicates of healthy individuals, adenoma, and colorectal cancer patients.

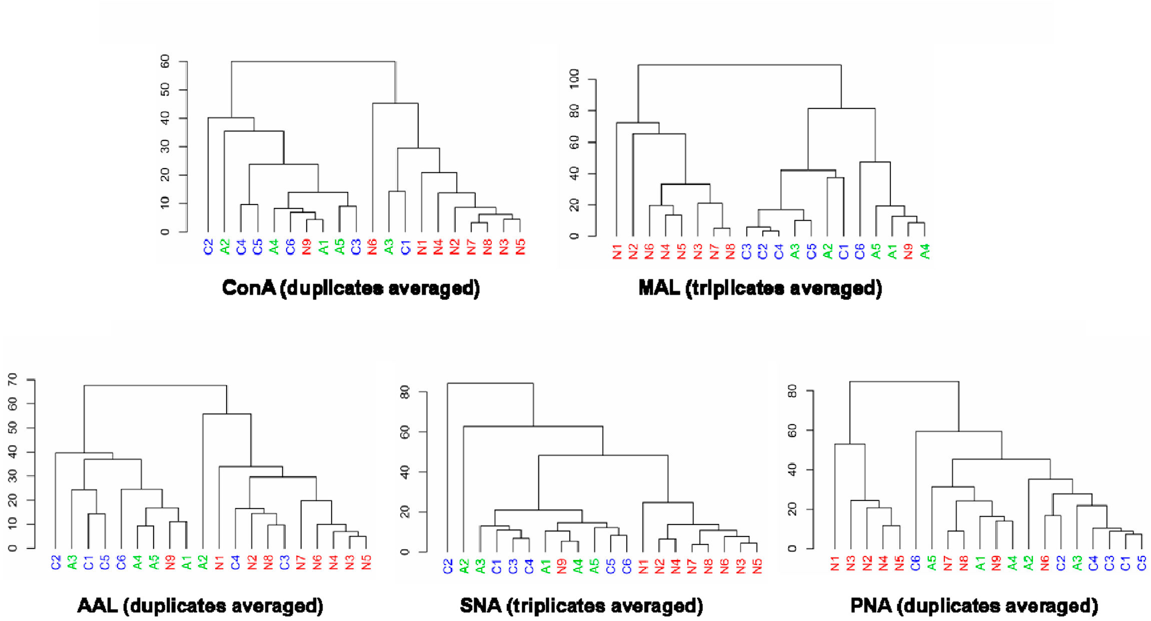
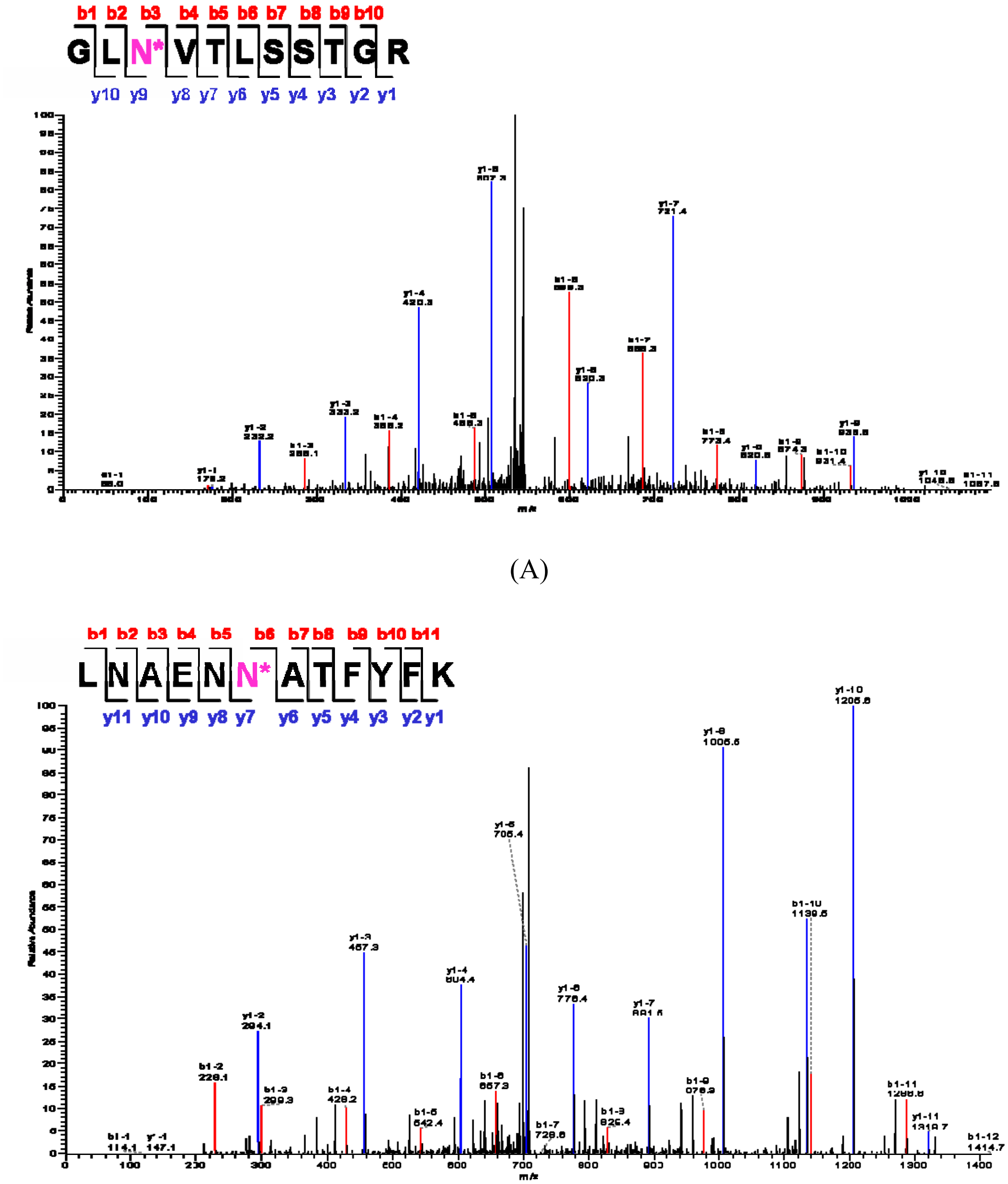


Figure 4.5 Unsupervised hierarchical clustering of glycoprotein microarray data distinguishes colorectal cancer (c1-c6) from adenoma (a1-a5) and normal controls (n1-n9). The clustering method was the average linkage, and the dissimilarity was obtained from the Pearson correlation coefficient.



(A)

(B)

Figure 4.6 Nano LC-MS/MS mass spectra of (A) doubly charged N-glycosylated peptide GLN*VTLSSGH (m/z = 553.28) from complement 4 and (B) doubly charged N-glycosylated peptide LANENN*ATFYFK from kininogen-1. The asterisk (*) denotes the site of N-glycosylation as determined from the tandem mass spectrum.

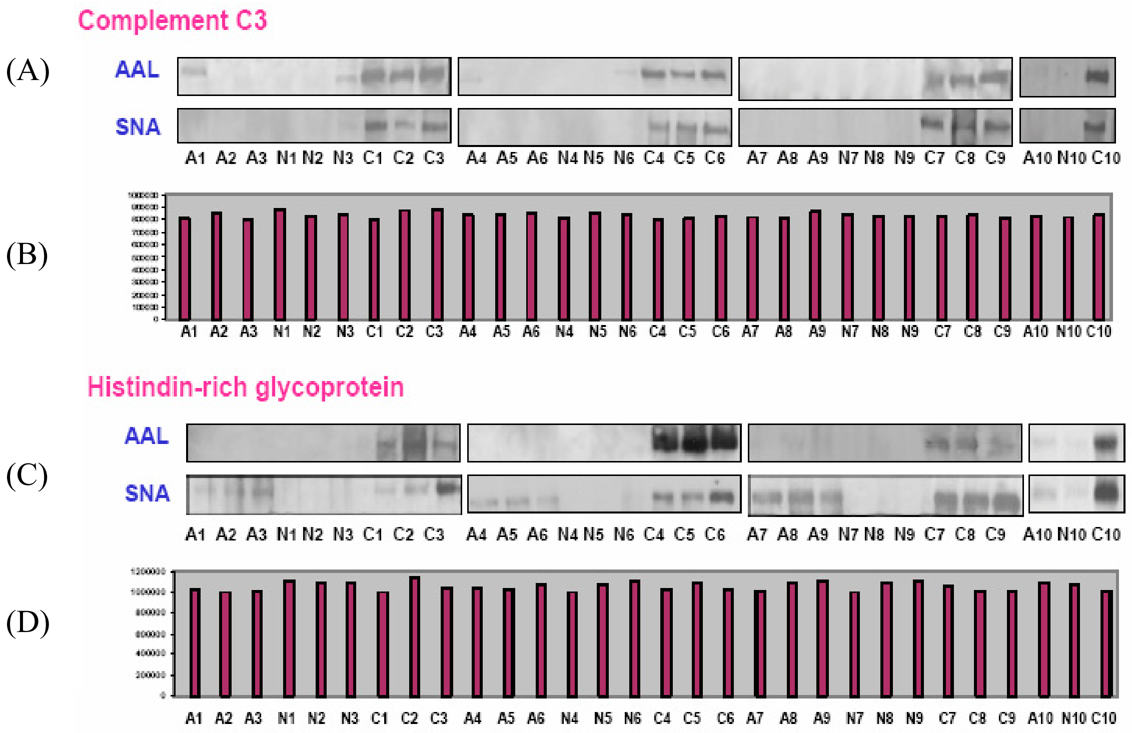


Figure 4.7 Elevated fucosylation and sialylation of complement C3 **(A)** and histidine-rich glycoprotein **(C)** investigated by AAL and SNA blot analysis. The corresponding protein expression levels are shown in **(B)** for complement C3 and **(D)** for histidine-rich glycoprotein respectively.

4.5 References:

[1] Parkin, D. M., Bray, F., Ferlay, J. and Pisani, P., *CA Cancer J. Clin.* 2005, 55, 74-108.

[2] Engwegen, J. Y., Helgason, H. H., Cats, A., Harris, N., *et al.*, *World J. Gastroenterol* 2006, 12, 1536-1544.

[3] Burt, R. W., *Gastroenterology* 2000, 119, 837-853.

[4] Ransohoff, D. F., *Gastroenterology* 2005, 128, 1685-1695.

[5] Kung, J. W., Levine, M. S., Glick , S. N., Lakhani, P., *et al.*, *Radiology* 2006, 240, 725-735.

[6] Wulfkuhle, J. D., Liotta, L. A. and Petricoin, E. F., *Nat. Rev. Cancer* 2003, 3, 267-275.

[7] Feldman, A. L., Espina, V., Petricoin, E. F., Liotta, L., A., and Rosenblatt, K. P., *Surgery* 2004, 135, 243-247.

[8] Novotny, M. V. and Mechref, Y., *J.Sep.Sci.* 2005, 28, 1956-1968.

[9] Plavina, T., Wakshull, E., Hancock, W. S. and Hincapie, M., *J. Proteome Res.* 2007, 6, 662-671.

[10] Yang, Z., Harris, L. E., Palmer-Toy, D. E. and Hancock, W. S., *Clin. Chem.* 2006, 52, 1897-1905.

[11] Qiu, R., Zhang, X. and Regnier, F. E., *J. Chromatogr. B* 2007, 845, 143-150.

[12] Qiu, R. and Regnier, F. E., *Anal. Chem.* 2005, 77, 7225-7231.

[13] Madera, M., Mechref, Y., Klouckova, I. and Novotny, M. V., *J. Proteome Res.* 2006, 5, 2348-2363.

[14] Zhou, Y., Aebersold, R. and Zhang, H., *Anal. Chem.* 2007, 79, 5826-5837.

[15] Zhang, H., Liu, A. Y., Loriaux, P., Wollscheid, B., *et al.*, *Mol. Cell. Proteomics* 2007, 6, 64-71.

[16] Haab, B. B., Geierstanger, B. H., Michailidis, G., Vitzthum, F., *et al.*, *Proteomics* 2005, 5, 3278-3291.

[17] Block, T. M., Comunale, M. A., Lowman, M., Steel, L. F., *et al.*, *Proc. Natl. Acad. Sci. USA* 2005, 102, 779-784.

[18] Steel, L. F., Shumpert, D., Trotter, M., Seeholzer, S. H., *et al.*, *Proteomics* 2003, 3, 601-609.

[19] Drake, R. R., Schwegler, E. E., Malik, G., Diaz, J., *et al.*, *Mol. Cell. Proteomics* 2006, 5, 1957-1967.

[20] Shin, S., Cazares, L., Schneider, H., Mitchell, S., *et al.*, *J. Am. Coll. Surg* 2007, 204, 1065-1071.

[21] Ueda, K., Katagiri, T., Shimada, T., Irie, S., *et al.*, *J. Proteome Res.* 2007, 6, 3475-3483.

[22] Hu, S., Loo, J. A. and Wong, D. T., *Proteomics* 2006, 6, 6326-6353.

[23] Zhang, H. and Chan, D. W., *Cancer Epidemiol. Biomarkers Prev.* 2007, 16, 1915-1917.

[24] Fletcher, R. H., *Ann. Int. Med.* 1986, 104, 66-73.

[25] Duffy, M. J., *Ann. Clin. Biochem.* 1998, 35, 364-370.

[26] Kuusela, P., Haglund, C. and Roberts, P. J., *Br. J. Cancer* 1991, 63, 636-640.

[27] Ward, U., Primrose, J. N., Finan, P. J., Perren, T. J., *et al.*, *Br. J. Cancer* 1993, 67, 1132-1135.

[28] Eskelinen, M., Pasanen, P., Kulju, A., Janatuinen, E., *et al.*, *Anticancer Res.* 1994, 14, 1427-1432.

[29] Lindmark, G., Kressner, U., Bergström, R. and Glimelius, B., *Anticancer Res.* 1996, 16, 895-898.

[30] Carpelan-Holmström, M., Haglund, C., Lundin, J., Alfthan, H., *et al.*, *Br. J. Cancer* 1996, 74, 925-929.

[31] von Kleist, S., Hesse, Y. and Kananeeh, H., *Anticancer Res.* 1996, 16, 2325-2331.

[32] Holten-Andersen, M. N., Christensen, I. J., Nielsen, H. J., Stephens, R. W., *et al.*, *Clin.Cancer Res.* 2002, 8, 156-164.

[33] Hakomori, S., *Adv. Exp. Med. Biol.* 2001, 491, 369-402.

[34] Hakomori, S., *Proc. Natl. Acad. Sci. USA* 2002, 99, 225-232.

[35] Choudhury, A., Moniaux, N., Ulrich, A. B., Schmied, B. M., *et al.*, *Br. J. Cancer* 2004, 90, 657-664.

[36] Orntoft, T. F. and Vestergaard, E. M., *Electrophoresis* 1999, 20, 362-371.

[37] Barrabés, S., Pagès-Pons, L., Radcliffe, C. M., Tabarés, G., *et al.*, *Glycobiology* 2007, 17, 388-400.

[38] Zhao, J., Qiu, W., Simeone, D. M. and Lubman, D. M., *J. Proteome Res.* 2007, 6, 1126-1138.

[39] Qian, W., Liu, T., Monroe, M. E., Strittmatter, E. F., *et al.*, *J. Proteome Res.* 2005, 4, 53-62.

[40] Cummings, R. D. and Kornfeld, S., *J. Biol. Chem.* 1982, 257, 11230-11234.

[41] Patwa, T. H., Zhao, J., Anderson, M. A., Simeone, D. M. and Lubman, D. M., *Anal. Chem.* 2006, 6411-6421.

[42] Keller, A., Nesvizhskii, A. I., Kolker, E. and Aebersold, R., *Anal. Chem.* 2002, 74, 5383-5392.

[43] Yan, W., Lee, H., Deutsch, E. W., Lazaro, C. A., *et al.*, *Mol. Cell. Proteomics* 2004, 3, 1039-1042.

[44] Gonzalez, J., Takao, T., Hori, H., Besada, V., *et al.*, *Anal. Biochem.* 1992, 205, 151-158.

Chapter 5

Serum Glycoproteomics of Esophageal Adenocarcinoma by Multi-Lectin Detection Based Glycoprotein Microarrays and Mass Spectrometry

5.1 Introduction

Esophageal adenocarcinoma, currently the seventh leading cause of cancer-related death in the United States with an overall 5-year survival rates of only 5-15% [1, 2], is increasing faster than any other malignancy in the United States. In 2007, it is estimated that 15,600 new cases of esophageal adenocarcinoma will be diagnosed, with approximately 14,000 deaths. Esophageal adenocarcinoma has been linked with the presence of Barrett's metaplasia. Patients with Barrett's metaplasia have a 30- to 145-fold increased risk of the development of esophageal adenocarcinoma. Although esophageal adenocarcinoma is curable in the early stages, it is usually diagnosed at a late and relatively untreatable stage, frequently after metastasis. A simple screening test for high-risk patients to catch esophageal adenocarcinoma in the early stages may be very effective in improving patient survival. For this reason, there is substantial interest in identification of serum protein markers with utility for the early detection and diagnosis of esophageal adenocarcinoma [4].

Along with the identification of proteins and the determination of their expression level, the analysis of post-translational modifications has become increasingly important in proteomics. Many proteins in human plasma are glycosylated and changes in both the extent of the glycosylation and the glycan structure has been linked to cancer and other disease states. Lectin affinity chromatography is commonly used to enrich and visualize glycoproteins and is used in glycosylation analysis for the purification and concentration of glycoproteins. Lectins have a selective affinity for a carbohydrate or a group of carbohydrates. Because lectin binding is not affected by other features of the glycan, they are highly effective in glycan analysis experiments. In recent studies, lectins have been used in arrays for rapid profiling of glycan expression patterns.

Protein microarrays can be used to detect proteins and monitor their expression levels and for analysis of their post-translational modifications. In protein microarray experiments, proteins are arrayed as spots on a nitrocellulose-coated slide, which can then be probed with monoclonal antibodies, or in the case of glycan analysis, specific lectins. The interaction is then recorded through different detection methods [5]. Protein microarrays have utility for performing high throughput clinical validation of potential biomarkers.

In this study, we employed lectin-based protein microarrays for screening esophageal adenocarcinoma-specific changes in glycan structure on plasma glycoproteins. This strategy, in conjunction with modern separation methods, MS-based technology, and statistical tools for data analysis, has allowed us to identify and characterize plasma glycoproteins with altered glycosylation. These cancer-specific glycan structural changes may have utility for the early diagnosis of esophageal adenocarcinoma.

5.2 Materials and Methods

5.2.1 Samples

Serum was obtained at the time of diagnosis following informed consent. The experimental protocol was approved by The University of Michigan Institutional Review Board. Serum from 20 esophageal patients was obtained from University of Michigan Hospitals. The diagnosis of the esophageal serum was either adenocarcinoma (10 samples) or high-grade dysplasia (10 samples). Sera from 10 healthy subjects were analyzed as normal controls.

5.2.2 Sample Preparation for Protein Microarrays and Lectin Blot

5.2.2.1 Delipidation and Immunodepletion

Prior to immunodepletion, serum samples were delipidated by centrifugation for 15 min at 20,000 × g, after which the lipid-containing upper layer was aspirated off. The delipidated sample was then divided into 250 μL aliquots, which were then immunodepleted using a ProteomeLab IgY-12 LC10 Proteome Partitioning Kit (Beckman-Coulter, Fullerton, CA). This procedure removes the twelve most abundant proteins from human serum, including albumin, IgG, α1- antitrypsin, IgA, IgM, transferrin, haptoglobin, α1-acid glycoprotein, α2-macroglobin, Apolipoprotein A-I, Apolipoprotein A-II, and fibrinogen. Both the flow-through (unbound) fraction and eluted (bound) fraction were collected separately during a 75 min separation cycle. The immunodepleted (flow-through) fraction was concentrated to approximately 2 mL with a 10,000 molecular weight limit Amicon Ultra-15 centrifugal filter (Millipore, Billerica, MA). The Bradford Protein Assay (Bio-Rad, Hercules, CA), using BSA as the standard, was used to determine the protein concentration of the final concentrated fraction, which typically were about 1.5-2.0 mg/mL.

5.2.2.2 N-Glycoprotein Enrichment with ConA Affinity Capture

1.5 mL of agarose-bound lectin (Vector Labs, Burlingame, CA) was added into 5 mL polypropylene columns (Pierce Biotechnology, Rockford, IL) to prepare the Concanavalin A (ConA) columns. Before use, the columns were equilibrated with binding buffer (20 mM Tris, 0.15 M NaCl, 1 mM Mn^{2+}, and 1mM Ca^{2+}, pH 7.4). 500 μL of the immunodepleted serum was diluted to 1.5 mL with binding buffer, and then loaded onto a newly packed ConA column. The column was incubated for 30 min, the unbound proteins were washed out with 6 mL of binding buffer, and the captured proteins were eluted with 4 mL of elution buffer (20 mM Tris, 0.5 M NaCl, 0.5 M methyl-R-D-mannopyranoside pH 7.4). The Bradford protein assay was used to determine protein recovery from the lectin column.

5.2.2.3 NPS-RP HPLC Separation

Using the PF2D system by Beckman-Coulter, 0.5 mL of enriched N-glycoprotein from ConA affinity chromatography (from 62.5 μL original serum) were separated by nonporous silica reverse phase (NPS-RP) HPLC on a 33 × 4.6 mm ODS III column (Eprogen, Darien, IL, USA) at a flow rate of 0.5 mL/min. The separation was performed using a water (solvent A) and acetonitrile (solvent B) gradient, both of which contained 0.1% v/v TFA. The method consisted of a 30 min nonlinear gradient with the following profile: (1) 5 to 25% B in 1 min; (2) 25 to 31% B in 2 min; (3) 31 to 37% B in 7 min; (4) 37 to 41% B in 8 min; (5) 41 to 48% B in 7 min; (6) 48 to 58% B in 3 min; (7) 58 to 75% B in 1 min; (8) 75 to 100% B in 1 min, followed by rinsing and cleaning. The column temperature was maintained at 60°C using a column heater (Beckman-Coulter) to improve speed and resolution. The eluted proteins were collected by an automated fraction collector (FC 204 BE; Beckman-Coulter) controlled by both an in-house-designed DOS-based software program and 32 Karat software (Beckman-Coulter), which was also used to calculate the peak areas.

5.2.3 Lectin Glycoarrays

The resultant fractions from NPS-RP-HPLC were completely dried in 96 well plates and re-suspended in 15 µL of printing buffer (65 mM Tris-HCl, 1% SDS, 5% DTT, and 1% glycerol). These fractions were left to shake overnight at 4°C. The separated glycoprotein fractions were then arrayed onto nitrocellulose slides (Whatman Schleicher & Schuell BioScience, Keene, NH) using a non-contact piezoelectric printing device (Nanoplotter 2.0, GeSiM, Germany) that spotted 2.5 nL of each fraction, in triplicate. The resultant spots were 450 µm in diameter and 600 µm apart. After drying for 1 day, the printed slides were blocked overnight in a solution of 1% BSA in 1X phosphate buffered saline with 0.1% Tween20 (PBS-T), then individually incubated with a biotinylated lectin for 2 hrs. The biotinylated lectins that were used were Peanut Agglutinin (PNA), Sambucus nigra bark lectin (SNA), Aleuria aurentia (AAL), Concanavalin A (ConA), and Maackia amurensis lectin II (MAL). All lectins were purchased from Vector Laboratories (Burlingame, CA), and were used at 5 µg/mL in 1X PBS-T. MAL was used at a concentration of 10 µg/mL, as per vendor recommendations. After incubation with lectin, the slides were washed with PBS-T 5 times for 5 minutes each, and then incubated with 1 µg/mL streptavidin -Alexafluor555 (Invitrogen, Carlsbad, CA) in PBS-T containing 0.5% BSA. The slides were then washed five times with PBS-T for 5 minutes each, and dried in a high speed centrifuge with slide holder (Thermo Electron Corp., Milford, MA). After drying, the slides were scanned in the green channel using an Axon 4000A scanner and analyzed with the GenePix 6.0 software (Molecular Devices, Sunnyvale, CA).

5.2.4 Data Analysis and Clustering

Microarray spot intensities were standardized with their corresponding UV peak areas before statistical analysis. We used hierarchical clustering (HC) and principal component analysis (PCA) to visualize the data with graphical representations of relationships between different samples. The replicate averages of 30 distinct biological specimens were used in HC relationship. In PCA statistic analysis, 30 serum samples processed in duplicate were placed in a two dimensional scatter plot based on similarities in normalized glycoform abundance. Z-statistics and Wilcoxon rank sum statistics for each protein detected by each lectin were calculated for differential abundance analysis. In the analysis, we compared normal samples with high grade dysplasia, normal with adenocarcinoma, and high grade dysplasia with adenocarcinoma. Proteins with Z values of ≥ 3.2 or ≤ -3.2 were selected as significantly different glycosylated proteins with a 95% significance level with multiple testing corrections.

5.2.5 SDS-PAGE and Lectin Blotting

Protein fractions from NPS-RP-HPLC were additionally separated by 1D SDS-PAGE using Mini-Protean cell (Bio-Rad, Hercules, CA) at 80V. The resolved proteins were stained with Coomassie brilliant blue (CBB) or transferred onto a polyvinylidene fluoride (PVDF) membrane (Bio-Rad), which was rehydrated in methanol, rinsed, then blocked with 5% w/v BSA (Roche, Indianapolis, IN) in PBS-T (phosphate buffer saline with 0.1% Tween 10). The membrane was incubated with biotinylated AAL and SNA, respectively, (1 μg/mL in PBS-T containing 3% BSA) for 1 hr at room temperature after washing with PBS-T for 10 min each. The membranes were then washed an additional 6 times for 15 min each in PBS-T and incubated with a 100ng/ml streptavidin-HRP in PBS-T containing 3% BSA. After incubation, the membrane was washed in PBS-T 4 times for 15 min each. We used Blue Sensitive autoradiography film (Marsh Bio Products, Rochester, NY) to detect chemiluminescence, achieved using an ECL analysis system (Pierce). A lectin blot experiment was used to detect potential glycoproteins and the corresponding Coomassie brilliant blue stained bands were identified by nano HPLC MS/MS.

5.2.6 Tryptic Digestion

5.2.6.1 Tryptic Digestion of NPS-RP-HPLC Fractions

The NPS-RP-HPLC fractions with significantly different glycosylation were dried completely, denatured in 40μL of 100 mM NH_4HCO_3 buffer (pH 7.8), then reduced with 1 mM dithiothreitol (DTT) for 45 min at 56°C and alkalized with 15mM iodoacetamide (IAA) for 1 h at room temperature in the dark. The proteins were then digested with 1-2 μg of TPCK-treated trypsin (Promega, Madison, WI, USA) for 18 h at 37°C. The reaction mixture was stopped by addition of 1-2 μL of TFA.

5.2.6.2 In-gel Trypsin Digestion

Gel pieces of the glycoprotein band were carefully cut out of the SDS gel. They were placed in siliconized Eppendorf tubes (Sigma), destained 3 times with 200 μl 200 mM ammonium bicarbonate and 40% acetonitrile at 37°C for 30 min each, and lyophilized completely in a SpeedVac (Thermo). The dried gel pieces were reduced with 50 μl of 10 mM DTT in 0.1 M NH_4HCO_3 for 45 min at 56°C. After the supernatant was removed, the resulting gel pieces were alkylated with 30 μl of 55 mM IAA in 0.1 M NH_4HCO_3 for 1 h at room temperature in the dark. The supernatant was removed, and the pieces were washed with 50% acetonitrile at least 3 times for 10 min each until the gel was shrunken and opaque. They were then lyophilized in a SpeedVac. Trypsin (20 ng/μl) was added to the gel slice until the gel became swollen. It was then incubated at $37^{o}C$ for 18 h. The peptide was extracted by 50 μl 0.1% TFA/60% CAN 3 times. The digestion solution was then lyophilized in a SpeedVac.

5.2.7 Protein Identification

To analyze the tryptic digests from both the NPS-RP-HPLC and SDS-PAGE separation, we used a MS4B MDLC system (Michrom Bioresources, Auburn, CA) interfaced with a linear ion trap mass spectrometer (LTQ, Thermo, San Jose, CA). First, the injected peptide sample was desalted on a trap column (150 μm × 50 mm, Michrom Bioresources Inc, Auburn, CA) with 3% solvent B (0.3 formic acid in 98% CAN) for 5 min at 50 μl/min. It was then sent to a separation column (150 μm × 150 mm, Michrom) where it was separated using a 75 min gradient from 3% B to 95% B. The resolved peptides were then directly sent to an ESI ion source with spray voltage of 2.6 kV.

Data dependent MS/MS analysis (m/z 400-2000) was carried out using MS acquisition software (Xcalibur 1.4, Thermo Finnigan) to identify the eluted peptides. The analysis consisted of a full MS scan followed by seven MS/MS scans of the seven most intensive precursor ions. All MS/MS spectra was searched in the Swiss-Prot FASTA human protein database using SEQUEST algorithm incorporated into TurboSequest feature of Bioworks 3.1 SR1.4 (Thermo Finnigan). A peptide with $X_{corr} \geq 3.5$ for triply-, 2.5 for doubly- and 1.9 for singly charged ions, and all with $\Delta Cn \geq 0.1$ [6], allowing for up to two missed cleavages, was considered positive protein identification. The sequence database search was set to accept the following modifications: carboxymethylated cysteines, oxidized methionines, and an enzyme-catalyzed conversion of asparagines to aspartic acids (0.984Da shift) at an *N*-glycosylation site. Accuracy of the SEQUEST assignment of MS/MS spectra to peptide sequences was validated by the TransProteomics Pipeline which includes both PeptideProphet and ProteinProphet software. In this study, peptides were identified with a probability cut-off of $p \geq 0.99$ and protein identifications were confirmed with probability scores of at least 0.9.

5.3 Results and Discussion

5.3.1 Reduction in the Complexity of Serum Glycoprotein Mixtures by Immunodepletion and Lectin Affinity Enrichment

Glycoproteomic analysis of protein mixtures that have the complexity of serum requires that steps be taken to reduce sample complexity prior to the identification of glycoproteins of interest. In order to analyze serum glycoproteins, 10 sera from each of normal, high grade dysplasia, and esophageal adenocarcinoma patients (250 μl) were delipidated by centrifugation. The delipidated sera were then individually immunodepleted to remove the 12 most abundant serum proteins (including albumin, IgG, α1- anti-trypsin, IgA, IgM, transferrin, haptoglobin, α1-acid glycoprotein, α2-macroglobin, Apolipoprotein A-I, Apolipoprotein A-II, and fibrinogen). Around 7% of the total protein mass in each of the serum samples remained after the immunodepletion step. The representative chromatograms resulting from the immunodepletion of serum from normal, high grade dysplasia and esophageal adenocarcinoma patients in Figure 5.1, where the immunodepleted fraction elutes at around 8 min, indicate the reproducibility of this step. Glycoproteins retained in the immunodepleted serum were subsequently enriched by ConA affinity chromatography. ConA binds with preference to oligomannosidic, hybrid, and bi-antennary *N*-glycans, either unconjugated or attached to proteins or peptides with broad specificity and high affinity [7]. O-glycopeptides or glycoproteins that contain exclusively O-glycosylation sites are not bound by this lectin. Approximately 70% of the immunodepleted serum protein content was recovered by ConA affinity chromatography. By selectively isolating the glycoproteins from the immunodepleted serum proteome, the procedure achieved a significant reduction in analyte complexity at two levels: first, the immunodepletion of the 12 most abundant proteins significantly increases the dynamic range of detection by approximately 90-fold and, additionally, reduces sample heterogeneity due to the removal of the highly variable IgG, IgA and IgM proteins; second, the subsequent *N*-glycoprotein enrichment step affords another effective means of reducing serum sample complexity.

30 μg of the enriched *N*-glycoprotein mixtures were further separated by NPS-RP-HPLC into 36 fractions for lectin glycoarray or lectin blot analysis. High-resolution separation of intact proteins was achieved, with the eluting proteins being detected by UV absorption at 214 nm. Figure 5.2 shows the reverse phase chromatograms of 6 serum samples from the different physiological states. Generally, we observed a high level of reproducibility in the UV traces among the different samples in the same group, although slight retention time shifts were observed. The UV peak area variance was within 15% for serum samples from different individuals. The reproducibility of these chromatograms indicates that the samples from the serum from normal subjects, from patients with high grade dysplasia and from esophageal adenocarcinoma patients were very similar at the protein expression level. These results suggest that the analysis of

glycoprotein expression alone does not provide valuable information to differentiate the clinical status of individuals.

5.3.2 Lectin Glycoarrays for Identification of *N*-glycosylation Pattern Changes

In order to analyze the serum glycosylation patterns, the separated intact glycoprotein fractions from NPS-RP-HPLC were spotted on nitrocellulose slides using a noncontact microarray spotter. Five lectins, *Aleuria aurentia* lectin (AAL), *Sambucus nigra* bark lectin (SNA), *Maackia amurensis* lectin (MAL), Peanut Agglutinin (PNA), and Concanavalin A (ConA), were used to detect different glycan moieties. AAL binds fucose linked (α-1, 6) to *N*-acetylglucosamine or (α -1, 3) to *N*-acetyllactosamine related structure. Both MAL and SNA recognize sialic acid on the terminal branches. MAL detects glycans containing NeuAc-Gal-GlcNAc with sialic acid at the 3 position of galactose while SNA binds preferentially to sialic acid attached to terminal galactose in an (α-2,6) and an (α -2,3) linkage at a lesser degree. In contrast, PNA binds desialylated exposed galactosyl (β-1, 3) *N*-acetylgalactosamine. ConA recognizes α -linked mannose including high-mannose-type and hybrid-type structures. The utilization of these five lectins have been proved to be highly successful in covering >95% of *N*-glycan types reported and differentiating them according to their specific structures [8]. Since only variations in glycan expression were of interest, all array spot intensities were normalized by dividing the corresponding UV peak area to eliminate protein abundance differences. The normalized array data suggest that the overall levels of protein fucosylation and sialylation are higher in esophageal cancer and high grade dysplasia serum samples as compared to the normal serum.

5.3.3 Statistical Analysis of *N*-glycosylation Pattern Changes

Bioinformatic analysis of the glycoprotein arrays was performed to highlight lectin response patterns that grouped different disease states together. To differentiate the serum samples in terms of their overall *N*-glycosylation patterns and to relate these patterns to clinical status, principal components analysis (PCA) and hierarchical clustering (HC) of the normalized glycoarray data were performed. In PCA, the 30 serum samples, assayed in duplicate, were analyzed separately for each lectin. The scores of the first two principal components of the samples from normal subjects, from patients with high grade dysplasia and from esophageal adenocarcinoma patients are illustrated in a 2-dimensional scatter plot in which each sample was plotted as an individual point. Closer points corresponded to greater similarity in the patterns of glycoprotein expression over all 36 protein spots on the microarrays. The excellent concordance among the replicates from the same sample in PCA results indicates that the lectin glycoarray is a robust strategy for screening *N*-glycosylation changes among the serum samples from different disease states. As expected, similar results were observed in HC by using the Pearson correlation coefficient for distance metrics.

As an alternative means of analyzing the lectin glycoarray data, we calculated Z-statistics of each array spot to search for signature proteins that might differentiate the serum samples of the different clinical states (Table 1). Comparisons were performed of normal versus high grade dysplasia (N/D), normal versus cancer (N/C), and high grade dysplasia versus cancer (D/C). Z values of ≥ 3.2 or ≤ -3.2 were selected as differential glycosylation at a 95% significance level. A positive Z value indicates elevated glycosylation and a negative Z value means reduced glycosylation.

5.3.4 Identification of Plasma Biomarkers with Altered *N*-glycosylation

Due to co-elution during the NPS-RP-HPLC separation, there were cases where more than one protein was observed in certain fractions. In order to determine which co-eluting protein was responsible for the differential responses in the lectin-based microarrays, the fraction with altered glycosylation was further separated by 1-D SDS-PAGE and then analyzed by lectin blot. Since the reduced fucosylation and sialylation levels in esophageal adenocarcinoma serum were detected on most of the differentially glycosylated proteins, we chose AAL and SNA in the lectin blot analysis to determine which protein corresponded to the differential fucosylation and sialylation pattern.

The corresponding protein bands in the SDS gel with significantly differential glycosylation pattern in esophageal cancer or high grade dysplasia were excised, and then digested with trypsin. Proteins were identified by nano LC-MS/MS. Positive identification was validated by the Trans-Proteomics pipeline which includes PeptideProphet and ProteinProphet software. PeptideProphet software was used to effectively identify correct peptide assignments and ProteinProphet was used to validate the protein identifications. In this study, peptides were identified with probability scores of at least 0.99 with a false positive error rate of 0.0007 and proteins were identified with a probability cut-off of $p \geq 0.9$ which corresponds to a 0.7% error rate [9, 10]. The significant differentially glycosylated proteins with their Z statistics are summarized in Table 1. Most of these proteins showed reduced glycosylation in the case of cancer compared to normal and high grade dysplasia, except for complement C3, which showed elevated fucosylation and sialylation levels in cancer. The data suggests that Z-statistic analysis of lectin glycoarrays has utility to identify cancer samples relative to high grade dysplasia or normal controls.

Table 5.1 Z-statistics of differentially glycosylated proteins detected by lectins.

	ConA			AAL			MAL			SNA		
Protein ID (Access #)	**D/N**	**C/N**	**C/D**	**D/N**	**C/N**	**C/D**	**D/N**	**C/N**	**C/D**	**D/N**	**C/N**	**C/D**
Complement factor H (P08603)	1.742	-2.892	**-4.539**	2.786	**-4.764**	**-5.477**	1.862	0.594	-1.207	0.078	-2.525	-2.343
Hemopexin (P02790)	1.506	**-5.693**	**-8.283**	1.533	**-4.778**	**-6.668**	1.261	**-3.760**	**-5.424**	1.104	**-4.069**	**-4.732**
Alpha-1B-glycoprotein (P04217)	**5.368**	-0.643	**-4.579**	**3.806**	2.652	-2.231	1.954	0	-1.890	**4.767**	-0.187	**-3.526**
Complement factor I (P05156)	**4.158**	-1.772	**-5.231**	0.426	**-3.338**	**-3.191**	-0.238	0.1270	0.360	2.225	-1.166	**-3.260**
Complement component C6 (P13671)	1.276	-2.182	**-3.570**	-1.047	**-4.744**	**-3.521**	0.547	-0.823	-1.414	1.145	-1.820	**-3.225**
Ceruloplasmin (P00450)	1.230	-3.022	**-3.502**	-0.206	-2.657	-2.566	0.222	-0.1510	-0.368	-0.715	**-3.759**	**-3.516**
Afamin (P43652)	0.243	0.734	0.518	-1.113	-0.704	0.168	-0.606	**-5.026**	**-4.044**	0.344	-2.195	-2.842
Complement C3 (P01024)	3.048	1.685	-1.291	1.169	**5.051**	**5.670**	1.662	**5.320**	**6.030**	0.437	**3.449**	**4.459**

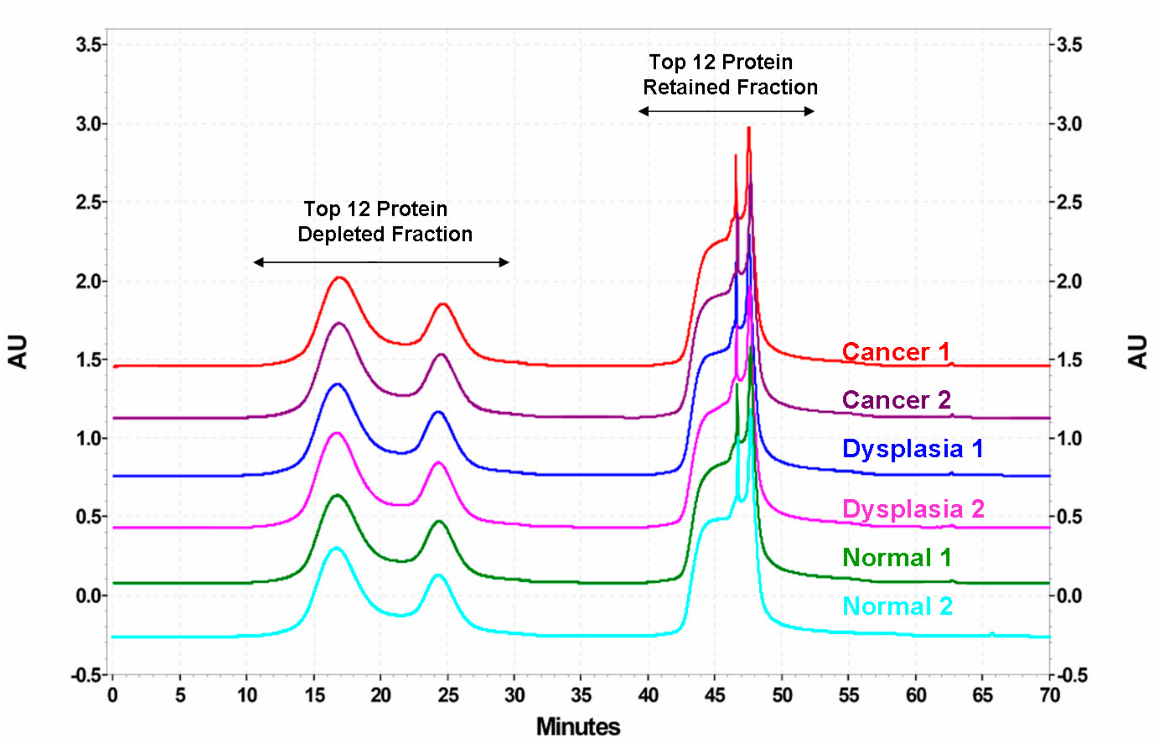
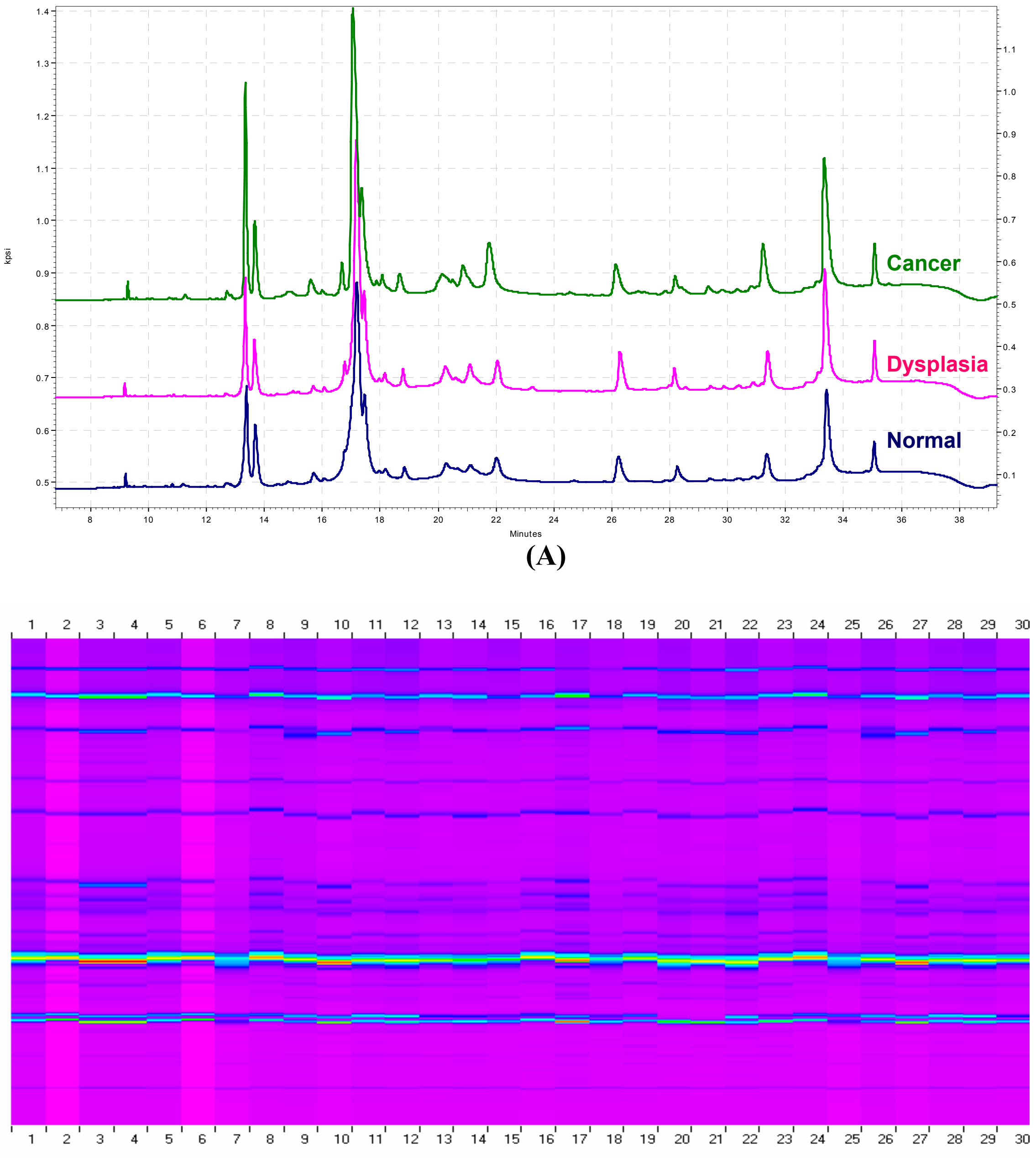


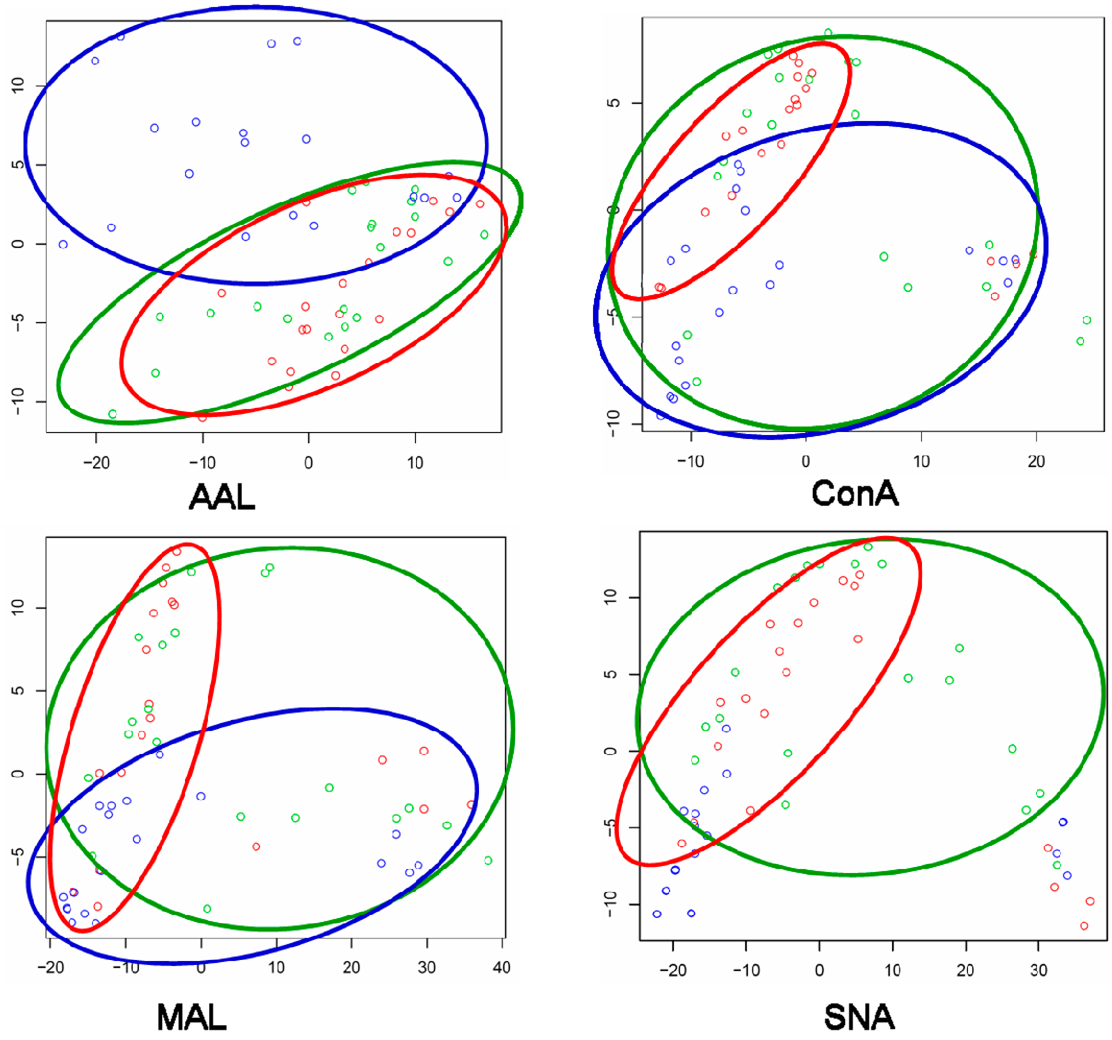
Figure 5.1 Representative chromatographic profiles of immunodepletion of serum from normal, high grade dysplasia and esophageal adenocarcinoma patients using ProteomeLab IgY-12 kit. The 12 most abundant proteins are contained in the “bound” fraction while the less abundant proteins in serum remained in the “flow-through” fraction.



(A)

(B)

Figure 5.2 (A) Representative UV chromatograms esophageal cancer, high grade dysplasia and normal serum sample. **(B)** UV chromatograms of all the serum samples from cancer (1-10), dysplasia (11-20) and normal (21-30) serum samples.



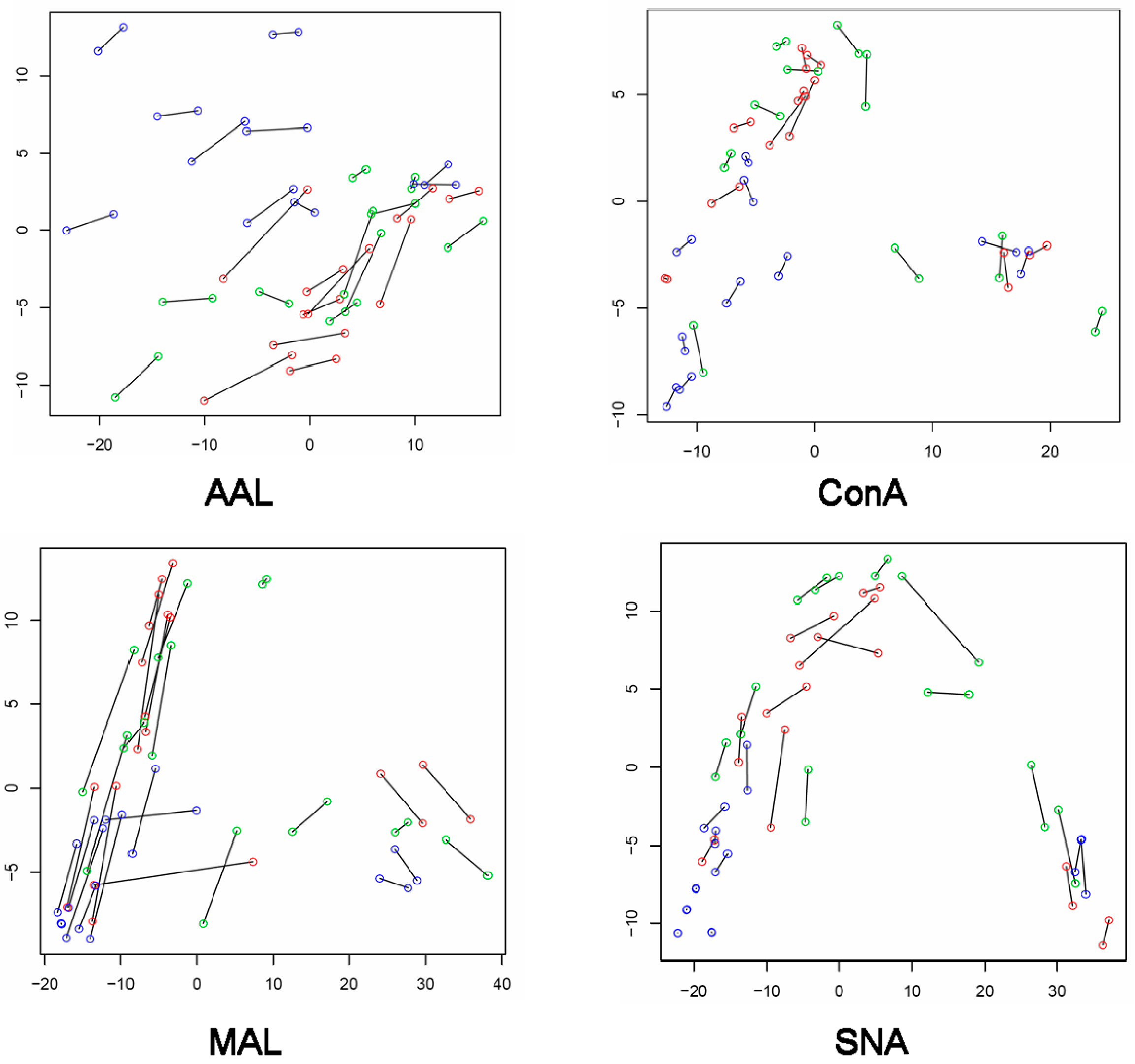
AAL

ConA

MAL

SNA

Figure 5.3 PCA plot for normalized glycoprotein microarray data derived from the replicates of healthy individuals (red), high grade dysplasia (green), and esophageal cancer patients (blue). Ovals indicate the areas where the data points of the three groups are distributed.



AAL

ConA

MAL

SNA

Figure 5.4 Reproducibility demonstration of PCA for normalized glycoprotein microarray data derived from the replicates of healthy individuals (red), high grade dysplasia (green), and esophageal cancer patients (blue).

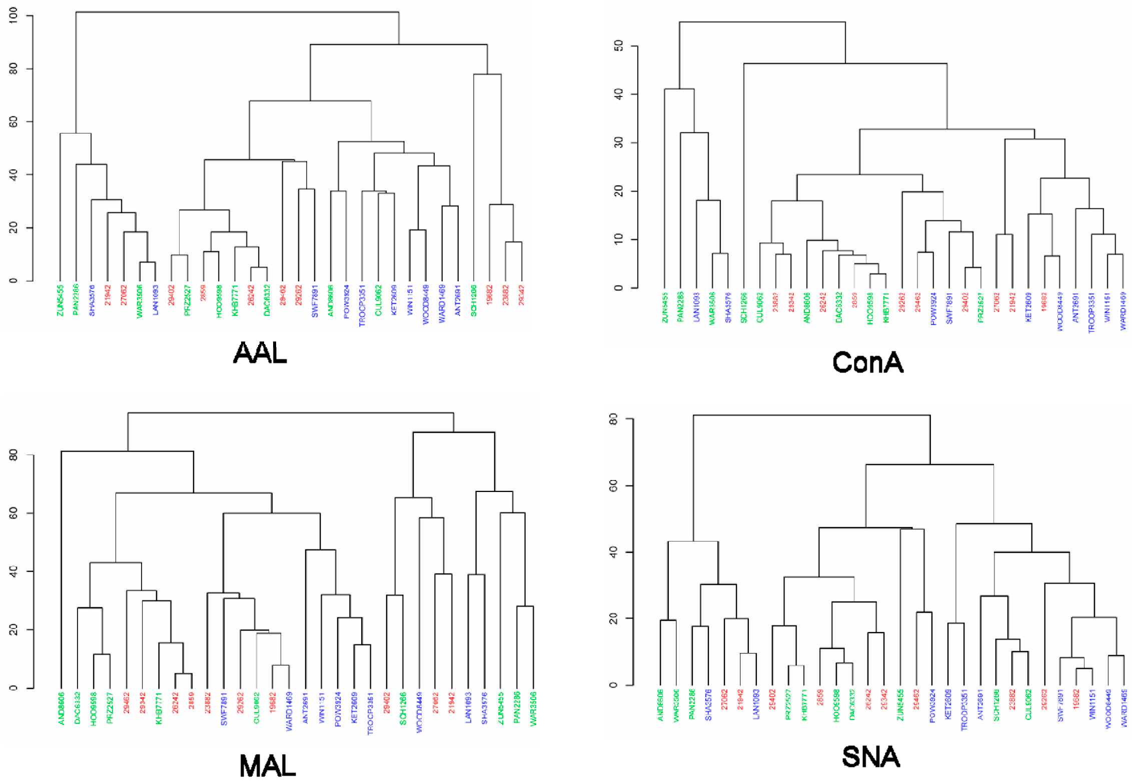


Figure 5.5 Unsupervised hierarchical clustering of glycoprotein microarray data distinguishes healthy individuals (red), high grade dysplasia (green), and esophageal cancer patients (blue). The clustering method was the average linkage, and the dissimilarity was obtained from the Pearson correlation coefficient.

5.4 References:

[1] Jemal, A., Thomas, A., Murray, T. and Thun, M., *CA Cancer J. Clin.* 2002, 52, 23-47.

[2] Parkin, D. M., Bray, F. I. and Devesa, S. S., *Eur. J. Cancer* 2001, 37, S4-66.

[3] Reid, B. J., *Gastroenterol Clin, North Am.* 1991, 20, 817-834.

[4] Wulfkuhle, J. D., Liotta, L. A. and Petricoin, E. F., *Nat. Rev. Cancer* 2003, 3, 267-275.

[5] Cretich, M., Damin, F. and Pirri, G., *Biomol. Eng.* 2006, 23, 77-88.

[6] Qian, W., Liu, T., Monroe, M. E., Strittmatter, E. F., *et al.*, *J. Proteome Res.* 2004, 4, 53-62.

[7] Cummings, R. D. and Kornfeld, S., *J. Biol. Chem.* 1982, 257, 11230-11234.

[8] Patwa, T. H., Zhao, J., Anderson, M. A., Simeone, D. M. and Lubman, D. M., *Anal.Chem.* 2006, 6411-6421.

[9] Keller, A., Nesvizhskii, A. I., Kolker, E. and Aebersold, R., *Anal.Chem.* 2002, 74, 5383-5392.

[10] Yan, W., Lee, H., Deutsch, E. W., Lazaero, C. A., *et al.*, *Mol. Cell. Proteomics* 2004, 3.

[11] Angeloni, S., Ridet, J. L., Kusy, N., Gao, H., *et al.*, *Glycobiology* 2005, 15, 31-41.

Chapter 6

Conclusions

Bacterial cold adaptation in *Exiguobacterium sibiricum* (*E. sibiricum*) 255-15 was studied at a proteomic scale using a 2-D liquid phase separation coupled with mass spectrometry technology. Whole cell lysates of *E. sibiricum* 255-15 grown at 4°C and 25°C were first fractionated according to p*I* by Chromatofocusing (CF), and further separated based on hydrophobicity by nonporous silica reversed phase HPLC (NPS-RPHPLC) which was on-line coupled with an ESI-TOF MS for intact protein *M*r measurement and quantitative interlysate comparison. Mass maps were created to visualize the differences in protein expression between different growth temperatures. The differentially expressed proteins were then identified by PMF using a MALDI-TOF MS and peptide sequencing by MS/MS with a matrix-assisted laser desorption/ionization quadrupole ion trap time-of-flight mass spectrometer (MALDI-QIT-TOF MS). A total of over 500 proteins were detected in this study, of which 256 were identified. Among these proteins 39 were cold acclimation proteins (Caps) that were preferentially or uniquely expressed at 4°C and three were homologous cold shock proteins (Csps). The homologous Csps were found similarly expressed at 4°C and 25°C where these three homologous Csps represent about 10% of the total soluble proteins at both 4°C and 25°C.

Exiguobacterium sibiricum 255-15 has shown significantly improved cryotolerance after liquid broth growth at 4°C and agar surface growth at both 4°C and 25°C compared with liquid broth growth at 25°C (Vishnivetskaya et.al. Cryobiology 2007, 54:234-240). The ability to survive freeze-thaw stress is expected to depend on the physiological state and protein composition of cells prior to freezing. Using 2-D liquid separation and an ESI-TOF MS-based mass mapping technique, we examined the differences in the proteomic profiles of the permafrost bacterium *E. sibiricum* 255-15 grown at two temperatures (4°C and 25°C) and two media (liquid broth and agar surface) before freeze-thawing treatments. In this study, a total of 330 proteins were identified. The cells cultured under the growth conditions associated with the improved cryotolerance have revealed a general down-regulation of enzymes involved in major metabolic processes (glycolysis, anaerobic respiration, ATP synthesis, fermentation, electron transport, and sugar metabolism) as well as in the metabolism of lipids, amino acids, nucleotides and nucleic acids. In addition, eight proteins (2'-5' RNA ligase, hypoxanthine phosphoribosyl transferase, FeS assembly ATPase SufC, thioredoxin reductase and four hypothetical proteins) were observed to be up-regulated. This suggests these eight proteins might have a potential role to induce the improved cryotolerance.

Colorectal cancer (CRC) remains a major worldwide cause of cancer-related morbidity and mortality largely due to the insidious onset of the disease. The current clinical procedures utilized for disease diagnosis are invasive, unpleasant and inconvenient; hence the need for simple blood tests that could be used for the detection of CRC. In this work, we have developed methods for glycoproteomics analysis to identify plasma markers with utility to assist in the detection of colorectal cancer (CRC). Following immunodepletion of the most abundant plasma proteins, the plasma *N*-linked glycoproteins were enriched using lectin affinity chromatography and subsequently further separated by non-porous silica reverse phase (NPS-RP)-HPLC. Individual RP-HPLC fractions were printed on nitrocellulose coated slides which were then probed with lectins to determine glycan patterns in plasma samples from 9 normal, 5 adenoma, and 6 colorectal cancer patients. Statistical tools, including principal components analysis, hierarchical clustering, and Z-statistic analysis, were employed to identify distinctive glycosylation patterns. Patients diagnosed with colorectal cancer or adenomas were shown to have dramatically higher levels of sialylation and fucosylation as compared to normal controls. Plasma glycoproteins with aberrant glycosylation were identified by nano-LC MS/MS, while a lectin blotting methodology was used to validate proteins with significantly altered glycosylation as a function of cancer progression. The potential markers identified in this study for diagnosis to distinguish colorectal cancer from adenoma and normal include elevated sialylation and fucosylation in complement C3, histidine-rich glycoprotein, and kininogen-1. These potential markers of colorectal cancer were subsequently validated by lectin blotting in an independent set of plasma samples obtained from 10 CRC patients, 10 patients with adenomas and 10 normal subjects. These results demonstrate the utility of this strategy for the identification of *N*-linked glycan patterns as potential markers of CRC in human plasma, and may have the utility to distinguish different disease states.

The same strategy has also been used in esophageal cancer biomarker discovery. In this study, 30 samples, 10 from normal, 10 from Barrett's esophagus, and 10 from esophageal cancer, were analyzed. Unlike in colon cancer, glycosylation in esophageal cancer was reduced compared to normal and Barrett's esophagus, but complement C3 still shows the significant elevated fucosylation and sialylation in cancer compared to normal and Barrett's esophagus. These results again confirm the importance of using glycosylation as biomarker in cancer detection.