An Integrated Approach Utilizing Liquid Separations, Protein Microarrays and Tandem Mass Spectrometry Towards Understanding Phosphorylation, Glycosylation and Humoral Response Changes in Cancer.

by

Tasneem H. Patwa

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Chemistry)
in The University of Michigan
2008

Doctoral Committee:

Professor David M. Lubman, Chair
Professor Masato Koreeda
Professor Jairam K. Menon
Assistant Professor Kristina I. Hakansson

To my family

Acknowledgements

I would like to express my sincere gratitude to my research advisor, Professor David M. Lubman for his invaluable and continuous support throughout the course of my Ph.D study. I would also like to thank my committee members, Dr. Masato Koreeda, Dr. Jairam Menon and Dr. Kristina Hakansson for serving on my dissertation committee and for their support and advice.

This work would not have been possible without collaboration with Dr. Diane Simeone, and Dr. Dean Brenner at the University of Michigan. They provided samples used in all humoral response experiments and the glycoprotein microarray studies. A special thanks also goes to Dr. Fred R. Miller at Wayne State University for continuous supply of AT1 and CA1a cell lines for the phosphoprotein experiments. I would especially like to thank Dr. Jia Zhao, Dr. Yanfei Wang and Dr. Yinghua Qiu for their collaboration. My graduate school experience would not have been the same without the friendship and support of Kendra Reid, Amberlyn Wands, Sarah Bethune and all colleagues in my lab.

And last but not least I would like to thank my parents, my brother and the rest of my family as well as my fiancé Ricardo Lira for always being by my side, supporting me and challenging me to do my best. For this I will always be grateful.

Table of Contents

List of Figures

Figure

List of Tables

Chapter 1

Introduction to proteomics and protein microarrays for post translational modification analysis and humoral response studies

1.1. Emergence of proteomics

The deciphering of the human genome has provided valuable information about the numbers of genes and proteins present in human cells. DNA expression differences between normal vs. diseased cells has shown that such studies can provide key information about the pathways that are altered upon disease progression. However with the knowledge about the human genome also came evidence that DNA expression does not necessarily show the true picture when it comes to understanding the state of a cell at any particular time. Studies have shown that DNA and mRNA levels do not necessarily positively correlate.[1] Furthermore mRNA expression levels are not the true indicators of protein expression.[2, 3] Proteins are the product of DNA transcription and RNA translation and are therefore the functional units of cellular processes. After a protein is translated it can also go through considerable post translational modifications. It is believed that there are 50,000 to a million proteins in a cell from a higher organism and these proteins have a dynamic range higher than 10^{10} orders of magnitude.[4-7] Furthermore protein expression levels in all human cells are not the same. Studying protein expression in cells at a global scale has become a major challenge for scientists today due to inherent nature of proteins in a cell as well as available techniques for

studying different types of proteins. Another obstacle to protein analysis is the fact that it can not be cloned into larger quantities in a similar fashion as DNA.

Over the years since global protein expression profiling has become possible, two-dimensional gel electrophoresis (2D-GE) has emerged as a popular technique.[8] In this technique proteins from a cellular lysate (be it from a cell line, tissue or bodily fluids such as serum, plasma or urine) are separated initially by isoelectric focusing according to the protein isoelectric point (pI) using a thin gel strip and subsequently by their molecular weight using a polyacrylamide gel electrophoresis (PAGE). 2D-PAGE methods are robust and have low detection limits in the picomole (and even lower) range but suffer from multiple drawbacks. Proteins with extremely high or low pIs tend to precipitate in a PAGE gel and are therefore out of the realm of this technique.[9] Frequently 2D-GE is coupled offline with electrospray ionization (ESI) and matrix assisted laser desorption ionization (MALDI) forms of mass spectrometry to determine thousands of proteins that are manually or robotically excised from the spots.[10, 11] Because such excision steps are pre-dominantly done by hand or a robot, contamination due to keratins from skin and hair often pose a problem. Comparison of the same gel run in different labs or on different days has also shown significant variation making cataloging of multiple experiments for comparison quite difficult. Proteins have a large dynamic range in terms of their molecular weight. Because the gel separations are restricted to a physical gel surface, separation of both the very high and low molecular weight proteins is difficult. In order to separate very high molecular weight proteins the gel needs to be run for a long time where this could result in lower molecular weight proteins running off the gel and not being analyzed.

1.2. All-liquid proteomics platforms

Due to the number of problems associated with gel electrophoresis discussed earlier, there has been increased interest in alternate techniques for protein expression profiling particularly methods that are all-liquid based. All liquid based techniques are favorable because they can be made almost completely hands-free and integrated with mass spectrometry without too much sample preparation. All liquid techniques have also divided protein expression profiling research and proteomics research into two parts: Bottom-up and Top-down proteomics. In bottom-up proteomics all proteins in a lysate are first digested and then using multiple dimensions of separations they are analyzed and finally identified by mass spectrometry. The most popular bottom-up technique currently is MudPiT originally developed by John Yates 3rd.[12] This Multi-Dimensional Protein Identification Technique first separates all digested peptides by strong cation exchange chromatography and then by reversed-phase HPLC. The resulting fractionated peptides are immediately transferred to a mass spectrometer for identification. Sophisticated software has been developed that is able to analyze many storage servers worth of data that can be obtained by such experiments. Such software is able to determine the identification of the protein from each of the peptide identified and sequenced. Bottom-up techniques are high throughput in nature but have some drawbacks. By digesting proteins in the first stages of the experiment critical information about the proteins is lost i.e. its intact molecular weight which could give information about potential post translational modifications or isoform information. In addition, MuDPiT experiments can

also result in high false positive identifications due to the homology that exists in multiple different proteins.

Top down methods currently being developed eliminate some of the drawbacks presented by bottom-up methods. In top-down methods, all proteins in a lysate are first separated (by chromatographic techniques or by high resolution mass spectrometry such as the FTICR) and then analyzed individually for identification and structural information. In earlier work isoelectric focusing using devices divided into chambers were created to separate and isolate proteins within a certain pI range in specific chamber.[13-15] The proteins in each chamber were then removed and further separated using other liquid base techniques such as RP-HPLC. Problems with these techniques include sample loss due to protein adhering to membrane surface separating chambers as well as poor resolution leading to the same protein appearing in multiple adjacent chambers. Alternately weak anion exchange based techniques have proved to be a good 1^{st} dimension of separation. Chromatofocusing is a weak anion based technique where proteins are separated by their isoelectric points.[16] A unique feature of this technique involves the titration of a start buffer with an elution buffer which results in a gradual change in the pH of the column. The result is the elution of proteins bound to the column from the proteins with the highest pI to the proteins with the lowest pI. Column resolution has been shown to be within 0.2 pH units. Collecting intact proteins according to their pI can provide interesting information about potential protein modifications because it has been shown that modification can result in a change in the protein's pI. Such information would be destroyed by protein digestion in bottom-up methods. The column-based nature of CF

enables the direct coupling of this technique with a second dimension of separation such as RP-HPLC.

In fact CF and RP-HPLC based systems have been commercialized by Beckman Coulter as the PF2D system.[17] In such a system proteins are fractionated and collected by pI in the first dimension. They are then transferred automatically to a non-porous silica reversed phase column for separation in the second dimension. Fractions from the second dimension can be collected by time or by peak depending on user preferences. Non-porous silica is particularly favorable because it increases the number of times the column can be used since clogging of the pores by large proteins is eliminated. It also demonstrates better peak characteristics enhancing the peak capacity and resolution of the separation.

1.3. Post-translational modifications in proteomics

Protein expression profiling work provides relevant information about proteins that change as a function of time, disease or other variables being studied. However other studies have shown that even more important than the protein expression levels in a cell are the modifications present on proteins involved in cellular pathways. Proteins can be modified with a variety of chemical groups ranging from phosphates, glycans, ubiquitin, oxides, methyl, nitrate and sulfates. Protein phosphorylation and glycosylation have been implicated in a variety of signaling pathways and changes in these modifications have been shown to be involved in disease progression. However studying these two modifications presents a great challenge due to reasons described herein.

Reversible phosphorylation is a key and important mechanism that is involved in a range of cellular processes such as cell growth, differentiation and apoptosis via a

variety of signaling pathways. One phosphorylation can trigger a domino effect where a signal travels via multiple phosphorylation events to affect a certain outcome. At any point in time at least one-third of all proteins are thought to be phosphorylated at a serine, threonine or tyrosine residue. It has been estimated that there at roughly 100,000 potential phosphorylation sites in the human proteome.[18] However as of now only a few thousand have been found. One reason for this lack of well-characterized phosphorylation site information is the presence of signaling molecules in very low abundance in the cells. In addition, the stoichiometry in which these molecules are phosphorylated is even lower. Therefore, while phosphorylated proteins, when digested, can be identified, the identification of phosphorylation sites is very difficult. Phosphopeptides do not ionize well because their signal is suppressed by non-phosphorylated peptides. To bypass this problem, research is ongoing in developing methods to isolate and enrich the phosphorylated peptides. Immobilized metal affinity chromatography (IMAC) has been popularly used to enrich phosphopeptides.[19, 20] In this method activated metal chelators bind phosphate groups on phosphopeptides while other peptides can be washed away. The phosphopeptides bound to the chelators can then be eluted out for further analysis. IMAC technologies have been commercialized into easy use ziptip formats. However coenrichment of other acidic peptides containing aspartic and glutamic acid groups hinders phosphorylation site analysis. Current work on amphoteric oxide-based solid phases for phosphopeptide enrichment appear promising.[21, 22] In these methods titanium or zirconium dioxide based solid phases are utilized to enrich phosphate containing peptides. Coenrichment of other acidic peptides is said to be reduced by using competitor acid group containing compounds such as hydroxyl-cinnamic acid. However

these competitor compounds are platform friends in the case of MALDI based instruments but often results in significant precipitation in HPLC based systems and therefore cannot be used.

Glycosylation, which is the attachment of sugar moieties to a protein, is the most complex type of protein post translational modification. Over 50% of all proteins have been estimated to be phosphorylated at any one time.[23] Glycosylation can change a protein's conformation significantly thereby affecting the proteins' activity. Glycoproteins are known to be involved in a variety of cellular and intercellular processes such as molecular recognition, fertilization and embryonic development, inflammation, cell adhesion, immune defense and inter- and intra-cellular signaling. There are two main types of protein glycosylations: N-glycosylation occurs when a glycan is attached by an N-acetylglucosamine to the amide group of an asparagine within a Asn-X-Ser/Thr consensus motif where X can be any amino acid but proline.[24] The other type of glycosylation is the O-linked type where a glycan is attached to the protein via an N-acetylgalactosamine to serine or threonine residues.[25] Glycoproteins are also difficult to study but for very different reasons. Glycoproteins that play a key role in signaling are often very high molecular weight proteins that are not easily isolated for further studies. Glycopeptides are not easily detectable by mass spectrometry because of their lower ionization efficiency compared to non-glycosylated peptides. Furthermore, glycopeptide signal intensities are often suppressed by non-glycosylated peptides particularly when the glycan structure on the peptide ends with negatively charged sialic acid residues.[26] Glycan heterogeneity is also another challenge when studying glycoproteins as one protein could have multiple different glycan structures associated to

it. This has led to development of techniques such as lectin affinity chromatography for selective enrichment of glycopeptides.[27] In addition, glycosylation sites can be easily determined by cleaving glycans from the protein using enzymatic means such as PNGAse F.[28] Cleavage results in addition of a hydrogen on the glycopeptide which can easily be monitored by mass spectrometry. Determination of the glycan structure on the other hand presents a very complicated problem. Glycan structures can be simple to very complex with various forms of branching possible. In order to study these structural complexities cleaved glycans need to be analyzed individually using multiple stages of mass spectrometry.[29, 30] This requires higher quantities of glycans than may be available in samples from natural sources.

At the bioinformatics end mass spectrometry data needs to be extensively analyzed to ensure that any results being obtained are confident protein identifications due to the nature of homology between many proteins. The large amounts of data obtained in proteomic experiments need huge amounts of server space and lots of computer analysis time. Many advances in the field of proteomics informatics have been made over the last 5 years.[31] However, because of the need for extensive sample preparation and individual experiments to study glycans or phosphorylations in different proteins, as well as the large amount of time needed for analysis of data obtained, the currently available strategies are becoming less and less high throughput in their real sense.

1.4. High-throughput protein microarrays

With the advent of microarray technology there is hope for reinstating the high throughput nature of protein profiling. Microarray technology was initially used to profile

DNA expression and interaction. However recent work has focused on applying this microarray technology to complex protein samples. Arrays can be made by fixing a membrane on a glass slide surface. The most popular membranes used include nitrocellulose and polyvinylidene diflouride membranes. Chemical derivatization of glass slide surfaces is also popular such as epoxy, aldehyde, poly-L-lysine and amine surfaces. Proteins can be arrayed on these surfaces using contact or non-contact mechanisms. While contact printing mechanisms are robust and cheaper, non-contact printers are much more reproducible and reliable in scientific studies. A typical microarray is the size of a microscopic slide and it can accommodate up to 10000 spots depending on the array format and protein spots diameter being arrayed. A protein array contains immobilized protein spots. Each spot can contain a set of "bait" molecules.[32, 33] These baits can range from a variety of molecules such as antibodies [34, 35], a cell or phage lysate [36, 37], a recombinant protein or peptide [38-40], a drug [41, 42], or a nucleic acid.[43, 44] The array is hybridized with either a probe (labeled antibody or ligand), or an unknown biological sample such as a cell lysate or serum sample. If the probe or biological samples are tagged with a signal-generating molecule such as a flourophore then positive and negative spots whose intensity corresponds to the extent of binding between the arrayed spot and probe result. An image of the resulting array can be captured by commercially available scanners and can be analyzed and interpreted using supporting software.

Currently protein microarrays have been used in a range of different applications. They can be divided into forward and reverse phase microarrays depending on whether the analyte is in the solution phase or immobilized on the surface. In forward phase arrays

the surface is arrayed with capture molecules, typically antibodies. Each array is processed with one test sample. Multiple antibodies can therefore be arrayed on one slide to see if a test sample has proteins reactive with all the arrayed antibodies. As a result, multiple analytes are measured at once. Such an approach has been used to identify mouse monoclonal antibodies that demonstrate the highest sensitivity for recombinant interleukin-4 detection.[45] Another very sophisticated example of an approach where forward phase arrays are used is a study where multiple antibodies involved in signaling pathways were probed with serum from various patient groups.[35, 46-48] Differential responses when signals from each individual group were compared highlighted key signaling proteins that demonstrated alterations in expression as a function of disease. While such studies are critical in highlighting potential markers of disease, it is important to note that only proteins whose antibodies are arrayed on slides can be probed for differential expression. Novel proteins that may be good markers of disease but that have previously not been implicated in diseases may therefore go unquestioned in these studies.

In a reverse phase array individual test samples are arrayed in each spot such that multiple samples are analyzed at the same time. Each array is then processed with a detection molecule such as an antibody resulting in a specific measurement being taken across hundreds if not thousands of samples. Reverse phase arrays have been used to generate SH2 binding profiles for phosphopeptides, recombinant proteins and entire proteomes.[49] Tissue microarrays are another kind of reverse phase arrays where tissue samples preferably from laser capture micro-dissection experiments are arrayed on slides and probed for proteins of interest. One such study assessed levels of cell survival and apoptotic proteins in breast cancer tissue.[50] Another approach that has been developed

over the last few years is a modified reverse-phase array approach where proteins from a biological medium (cell line, tissue, serum) are initially separated by a chromatographic technique. These proteins are then arrayed on slides after which they are probed for qualities of interest such as post translational modifications of various type [51] as well as immune response from the arrayed proteins particularly if the proteins originate from a diseased sample. [52, 53]

1.5. Dissertation outline

This dissertation attempts to integrate the positive attributes of liquid separations, protein microarrays and mass spectrometry to study disease progression. Chromatofocusing and non-porous reversed-phase HPLC are used to sufficiently isolate proteins into distinct fractions before subjecting them to microarray analysis for assessing protein phosphorylation levels in cellular proteins and humoral response in cellular proteins and tissue proteins. In addition other liquid separation methodologies, particularly protein enrichment by reduction of complexity using affinity chromatography and lectin enrichment chromatography are used to study glycosylated proteins in human serum samples.

The first four chapters detail the development of an all liquid separations techniques, protein microarray and mass spectrometry strategy that can be used to highlight post translational modification changes as a function of cancer. Chapter 2 explains the utility of combining CF and NPS-RP-HPLC to protein microarrays to study phosphorylation changes in breast cancer. The technique is applied to two breast cancer cell lines AT1 and CA1a from the xenograft model of breast cancer resulting in the identification of proteins from key cellular processes expressing differential

phosphorylation. Chapters 3, 4 and 5 are an overview of how such a technique can be modified to study glycosylation changes in pancreatic and colon cancers respectively. Instead of the CF/NPS-RP-HPLC platform, affinity chromatography for removal of the top 12 abundant proteins from serum is coupled to lectin enrichment and NPS-RP-HPLC after which proteins are arrayed on microarrays to assess changes in glycosylation states as a function of cancer. Statistical analysis illustrates that such a methodology is successful in distinguishing between normal and disease groups. Chapter 6 is a story of how this 2D-liquid separations-protein microarray-MS/MS integrated technique can be used to exploit the naturally present humoral response to disease in order to highlight potential panels of biomarkers for pancreatic cancer. Using this approach a panel of 9 proteins is shown to distinguish between normal and cancer serum with good sensitivity and specificity. This chapter illustrates the importance of choice of statistical method to the nature of results obtained. Chapter 7 is an overview of a modification of the humoral response measuring technique that enhances the sensitivity of humoral response measurements.

Viewed together these chapters can be considered as a complete overview of how separation technologies together with protein microarrays can be utilized to study complex biological problems using a range of biological materials (cell lines, tissues, serum) with particular focus on cancers.

1.6. References

[1] Holland, M. J., *J Biol Chem* 2002, *277*, 14363-14366.

[2] Gygi, S. P., Rochon, Y., Franza, B. R., Aebersold, R., *Mol Cell Biol* 1999, *19*, 1720-1730.

[3] Griffin, T. J., Gygi, S. P., Ideker, T., Rist, B., *et al.*, *Mol Cell Proteomics* 2002, *1*, 323-333.

[4] Hochstrasser, D. F., Sanchez, J. C., Appel, R. D., *Proteomics* 2002, *2*, 807-812.

[5] Wilkins, M. R., Sanchez, J. C., Williams, K. L., Hochstrasser, D. F., *Electrophoresis* 1996, *17*, 830-838.

[6] Jacobs, J. M., Adkins, J. N., Qian, W. J., Liu, T., *et al.*, *J Proteome Res* 2005, *4*, 1073-1085.

[7] Issaq, H. J., Chan, K. C., Janini, G. M., Conrads, T. P., Veenstra, T. D., *J Chromatogr B Analyt Technol Biomed Life Sci* 2005, *817*, 35-47.

[8] O'Farrell, P. H., *J Biol Chem* 1975, *250*, 4007-4021.

[9] Aebersold, R., *J Am Soc Mass Spectrom* 2003, *14*, 685-695.

[10] Shevchenko, A., Wilm, M., Vorm, O., Mann, M., *Anal Chem* 1996, *68*, 850-858.

[11] Gygi, S. P., Corthals, G. L., Zhang, Y., Rochon, Y., Aebersold, R., *Proc Natl Acad Sci U S A* 2000, *97*, 9390-9395.

[12] Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., *et al.*, *Nat Biotechnol* 1999, *17*, 676-682.

[13] Wall, D. B., Kachman, M. T., Gong, S., Hinderer, R., *et al.*, *Anal Chem* 2000, *72*, 1099-1111.

[14] Kachman, M. T., Wang, H., Schwartz, D. R., Cho, K. R., Lubman, D. M., *Anal Chem* 2002, *74*, 1779-1791.

[15] Zhu, Y., Lubman, D. M., *Electrophoresis* 2004, *25*, 949-958.

[16] Yan, F., Subramanian, B., Nakeff, A., Barder, T. J., *et al.*, *Anal Chem* 2003, *75*, 2299-2308.

[17] Wang, Y., Wu, R., Cho, K. R., Shedden, K. A., *et al.*, *Mol Cell Proteomics* 2006, *5*, 43-52.

[18] Zhang, H., Zha, X., Tan, Y., Hornbeck, P. V., *et al.*, *J Biol Chem* 2002, *277*, 39379-39387.

[19] Andersson, L., Porath, J., *Anal Biochem* 1986, *154*, 250-254.

[20] Sykora, C., Hoffmann, R., Hoffmann, P., *Protein Pept Lett* 2007, *14*, 489-496.

[21] Larsen, M. R., Thingholm, T. E., Jensen, O. N., Roepstorff, P., Jorgensen, T. J., *Mol Cell Proteomics* 2005, *4*, 873-886.

[22] Kweon, H. K., Hakansson, K., *Anal Chem* 2006, *78*, 1743-1749.

[23] Apweiler, R., Hermjakob, H., Sharon, N., *Biochim Biophys Acta* 1999, *1473*, 4-8.

[24] Vance, B. A., Wu, W., Ribaudo, R. K., Segal, D. M., Kearse, K. P., *J Biol Chem* 1997, *272*, 23117-23122.

[25] Hang, H. C., Bertozzi, C. R., *Bioorg Med Chem* 2005, *13*, 5021-5034.

[26] Annesley, T. M., *Clin Chem* 2003, *49*, 1041-1044.

[27] Geng, M., Zhang, X., Bina, M., Regnier, F., *J Chromatogr B Biomed Sci Appl* 2001, *752*, 293-306.

[28] Carr, S. A., Huddleston, M. J., Bean, M. F., *Protein Sci* 1993, *2*, 183-196.

[29] Kuster, B., Krogh, T. N., Mortz, E., Harvey, D. J., *Proteomics* 2001, *1*, 350-361.

[30] Harvey, D. J., *Proteomics* 2001, *1*, 311-328.

[31] Deutsch, E. W., Lam, H., Aebersold, R., *Physiol Genomics* 2008.

[32] Liotta, L., Petricoin, E., *Nat Rev Genet* 2000, *1*, 48-56.

[33] MacBeath, G., *Nat Genet* 2002, *32 Suppl*, 526-532.

[34] Lal, S. P., Christopherson, R. I., dos Remedios, C. G., *Drug Discov Today* 2002, *7*, S143-149.

[35] Haab, B. B., *Mol Cell Proteomics* 2005, *4*, 377-383.

[36] Wang, X., Yu, J., Sreekumar, A., Varambally, S., *et al.*, *N Engl J Med* 2005, *353*, 1224-1235.

[37] Paweletz, C. P., Charboneau, L., Bichsel, V. E., Simone, N. L., *et al.*, *Oncogene* 2001, *20*, 1981-1989.

[38] Dexlin, L., Ingvarsson, J., Frendeus, B., Borrebaeck, C. A., Wingren, C., *J Proteome Res* 2008, *7*, 319-327.

[39] Pavlickova, P., Schneider, E. M., Hug, H., *Clin Chim Acta* 2004, *343*, 17-35.

[40] MacBeath, G., Schreiber, S. L., *Science* 2000, *289*, 1760-1763.

[41] Humphery-Smith, I., Wischerhoff, E., Hashimoto, R., *Drug Discov. World* 2002, *4*, 17-27.

[42] Hardiman, G., *Pharmacogenomics* 2007, *8*, 1639-1642.

[43] Katilius, E., Flores, C., Woodbury, N. W., *Nucleic Acids Res* 2007, *35*, 7626-7635.

[44] Li, Y., Lee, H. J., Corn, R. M., *Nucleic Acids Res* 2006, *34*, 6416-6424.

[45] Wang, L., Cole, K. D., Peterson, A., He, H. J., *et al.*, *J Proteome Res* 2007, *6*, 4720-4727.

[46] Shafer, M. W., Mangold, L., Partin, A. W., Haab, B. B., *Prostate* 2007, *67*, 255-267.

[47] Sanchez-Carbayo, M., Socci, N. D., Lozano, J. J., Haab, B. B., Cordon-Cardo, C., *Am J Pathol* 2006, *168*, 93-103.

[48] Miller, J. C., Zhou, H., Kwekel, J., Cavallo, R., *et al.*, *Proteomics* 2003, *3*, 56-63.

[49] Machida, K., Thompson, C. M., Dierck, K., Jablonowski, K., *et al.*, *Mol Cell* 2007, *26*, 899-915.

[50] Cowherd, S. M., Espina, V. A., Petricoin, E. F., 3rd, Liotta, L. A., *Clin Breast Cancer* 2004, *5*, 385-392.

[51] Pal, M., Moffa, A., Sreekumar, A., Ethier, S. P., *et al.*, *Anal Chem* 2006, *78*, 702-710.

[52] Taylor, B. S., Pal, M., Yu, J., Laxman, B., *et al.*, *Mol Cell Proteomics* 2007.

[53] Yan, F., Sreekumar, A., Laxman, B., Chinnaiyan, A. M., *et al.*, *Proteomics* 2003, *3*, 1228-1235.

Chapter 2

A novel analysis scheme for assessing phosphorylation changes of high and medium abundant proteins in pre-malignant and malignant breast cell lines using 2D liquid separations, protein microarrays and tandem mass spectrometry

2.1. Introduction

Breast cancer is the most frequently diagnosed cancer in women. More than 200,000 new cases of breast cancer, with over 41,000 deaths, were expected in the United States in 2006.[1] Breast cancer related deaths have declined by approximately 2.3% from 1990 to 2002 primarily due to earlier detection awareness as well as improved treatment. While the five-year survival rate has increased to 98% for local-regional disease, it is only 26% for women with distant metastases.[1] Understanding the molecular mechanisms that underlie breast cancer development and progression to malignancy may uncover better therapeutic targets with potential utility to further decrease breast cancer mortality.

Aberrations in cellular signaling pathways have been associated with cancer development and progression, as cancer cell survival and proliferation rates increase, and as cancer cells become increasingly evasive to the immune system.[2-4] Growth factor signals are propagated from the cell surface intracellular milieu by signaling pathways,

involving a variety of kinases such as membrane receptor kinases (EGFR, VEGF) and cytoplasmic kinases (ERK, MEK, Ras, PI3-K and mTOR).[5] In cancer, these signaling pathways are often dysregulated, resulting in a phenotype characterized by unfettered cell growth and increased invasive potential. Cellular signaling is largely controlled by transient, post-translational modifications of signaling proteins, which alter their ability to bind and interact with downstream effectors.[4-6] Protein phosphorylation is one such modification that primarily acts as a molecular switch to activate or deactivate cellular signaling cascades.[4, 7, 8] A recent review by Krueger *et al*. lists several phosphorylated proteins that are known to contribute to oncogenesis or are used in the context of a cancer biomarker.[9] Proteins from all cellular compartments are represented in this list including histones, HDACs, MAP kinases, Akt, PTEN, EGFRs and ILK.

A variety of techniques have been used to study phosphorylation expression on a large scale.[10] One such technique involves incubation of cells with radioactive ^{32}P followed by 2D gel electrophoresis.[11] Although able to detect a wide dynamic range of phosphoproteins, this method requires handling of radioactive orthophosphate which makes it less favorable. In addition, the dependence on turnover rates at which the orthophosphate is incorporated into proteins may reduce sensitivity of this technique. The use of monoclonal and polyclonal antibodies specific to phosphorylated proteins to detect global phosphoprotein patterns on gels[12] circumvents the use of radiolabels. However, current available phosphoserine-specific and phosphothreonine-specific antibodies are not always reliable and cannot detect phosphoproteins where steric hindrance prevents antibody binding. More recently, a novel small molecule phosphosensor dye has been reported for detecting phosphoproteins on both gel and microarray platforms.[13-16] This

dye is able to detect phosphotyrosine, serine and threonine residues and can discriminate between thiophosphorylation and sulfation.

Gel-based methods have been considered the method of choice in studying global protein expression, but more recently developed techniques have focused on liquid-based methods due to the ease of coupling to mass spectrometers for protein identification. The liquid-based method most frequently used for phosphoprotein analysis in complex samples involves shotgun proteomics where a complex protein mixture is first digested and enriched for phosphopeptides.[17-19] An enrichment step is often necessary since phosphopeptide ionization is typically suppressed in the presence of many non-phosphorylated peptides present in a complex sample. The enriched peptides are then analyzed by LC-MS/MS with comprehensive database searching to confirm identity and elucidate the phosphorylation site. A variety of enrichment methods have been developed ranging from immobilized metal affinity chromatography[20, 21] to amphoteric oxide based enrichment, frequently using titanium or zirconium dioxide,[22, 23] as well as antibody based enrichment. While shotgun proteomics is a high throughput method at the experimental front, it is very time-consuming at the analysis end since data must be closely examined for possible false positives and negatives.

Quantitation of differentially expressed proteins by mass spectrometry is a further challenge because in addition to the inefficient ionization and suppression of phosphopeptide ions, efficient labeling methodologies are needed in order to make quantitation possible. Currently available methodologies such as SILAC[24, 25], ICAT[26] or iTRAQ[27, 28] can be used. SILAC involves stable isotope labeling of proteins as they are produced in cultures therefore introducing problems of turnover rate

differences between proteins. Furthermore, ICAT may not be successful because it requires the presence of a cystein residue for labeling and the frequency of occurrence of a phosphopeptide with a cystein residue can be considered very low. In addition, completeness of this labeling reaction is difficult to monitor especially when multiple samples are being processed at the same time. Labeling at the peptide level also eliminates intact protein information making quantitation ambiguous especially for phosphoproteins that may exist as multiple isoforms or that have homology with other cellular proteins because a peptide could belong to more than one protein. Furthermore, in clinically relevant samples, even labeling may not be sufficient to detect very low levels of phosphoproteins. Label free approaches where mass spectrometric signals are directly compared to obtain quantitative information are also being developed.[29] However such approaches require mass spectrometers with very high mass accuracy and experiments with very precise and high reproducibility in order to ensure that the quantitative information is accurate.

To overcome some of these limitations, we have been developing the coupling of comprehensive 2D-liquid separation methods to protein microarray technology. We have used this strategy previously to assess the phosphorylation status of all proteins in a cell line that was treated with a specific protein kinase inhibitor.[30] While that study was successful in highlighting phosphorylation changes caused by experimentally perturbing a specific biological pathway, there are currently no reports investigating such changes in naturally occurring disease states. A study comparing phosphorylation status in disease states may have utility in elucidation of pathways that play a role in the progression of disease.

A xenograft model of human breast disease progression has been developed from the MCF10A breast epithelial cell lines. Selected cell lines within the series are representative of normal, pre-malignant and malignant phenotypes.[31-33] T24 c-Ha-ras oncogene-transfected MCF10A cells (MCF10AneoT) form small, flat nodules in Nude/Beige mice which persist for the life span of the host and sporadically progress to carcinomas. A variant cell line (MCF10AT1), derived from one xenograft, not only forms simple differentiated ducts which persist in xenografts and sporadically progress to carcinoma, but also forms intermediate proliferative lesions resembling proliferative disease without atypia, atypical hyperplasia, and carcinoma in situ. By establishing cells in culture representing different stages in progression of MCF10AT through atypical hyperplasia to carcinoma, interruption of progression has been made possible. These cell lines continue to progress when reimplanted in vivo in immune deficient mice but are sufficiently stable in vitro to provide the tools essential for the genetic analysis of progression. MCF10AT cells express estrogen receptor (ER) and estradiol (E2) accelerates progression of the premalignant xenograft lesions. Fully malignant variants (MCF10CA lines), some of which are metastatic have also been recently derived. Although many cancers are ER- negative and E2 independent, the early stages of disease may have been E2 responsive. It is hypothesized that E2 independent carcinomas (represented by the MCF10CA lines in the model) may constitutively express proteins that are altered by E2 in earlier premalignant stages (such as MCF10AT1). Attempting to identify genes expressed constitutively in malignant MCF10CA variants that are also induced by E2 in premalignant MCF10AT1 cells are currently in progress.

In this study we compared the phosphoproteome of pre-malignant (MCF10AT1) and malignant (MCF10CA1a c11) cell lines using a 2-dimensional liquid-phase separation method coupled to protein microarray technology. These two particular cell lines were chosen because they are both ER positive and therefore similar to each other despite the different phenotype they present. The naturally occurring, arrayed proteins were probed with the small-molecule phosphosensor dye, ProQ Diamond and anti-phosphotyrosine antibodies. The strategy enabled us to detect and identify differentially expressed phosphoproteins and to determine specific changes associated with the premalignant and malignant phenotypes.

2.2. Experimental Section

Sample Preparation/Cell lines: The premalignant AT1 cell line (MCF10AT1) and malignant CA1a cell line (MCF10CA1a c11) were both derived from the MCF10A human breast cell line and were maintained and prepared as previously described [31, 33].

Cell lysis, buffer exchange and protein quantitation: Cells were mixed with lysis buffer containing 7 M urea, 2 M thiourea, 100 mM dithiothreitol (DTT), 2% n-octyl G-D-glucopyranoside (OG), 10% glycerol, 10 mM sodium orthovanadate, 10 mM sodium fluoride (all from Sigma, St. Louis, MO), 0.5% Biolyte ampholyte (Bio-Rad, Hercules, CA), and protease inhibitor cocktail (Roche Diagnostics, GmBH, Mannheim, Germany) with vortexing at room temperature for 1 hr. Cellular debris and other insoluble materials were removed by centrifuging the mixture at 80000 x g for 1 hr 15 min. The supernatant was subjected to buffer exchange in order to replace the lysis buffer with start buffer

(composition described later) for chromatofocusing using a PD-10 G-25 column (Amersham Biosciences, Piscataway, NJ). The protein concentration was determined using the Bradford Protein Assay kit with bovine serum albumin (BSA, Bio-Rad) standard.

Chromatofocusing (CF): The CF experiment was performed using a Beckman System Gold model 127 pump and 166 UV detector module (Beckman Coulter, Fullerton, CA) with a HPCF-1D prep column (250 mm L x 4.6 mm ID, Eprogen, Darien, IL). A linear pH gradient was generated using a combination of start buffer (SB) composed of 6 M urea, 25 mM BisTris, and 0.2% OG and elution buffer (EB) containing 6 M urea, 0.2% OG, and 10% polybuffer 74 (Amersham Biosciences). Saturated iminodiacetic acid (Sigma) was used to adjust the pH of SB at 7.2 and EB at 3.9. The column was first equilibrated in SB until the pH of the column was the same as start buffer by monitoring with a post detector online assembly of a pH-flow cell (Lazar Research Laboratories, Los Angeles, CA). After equilibration, ~10 mg of sample was loaded onto the column at a low flow rate to allow for interactions of the proteins with the binding sites. Once a baseline was achieved, solvent flow was switched to EB and the flow rate was set to 1 mL/min for CF fraction collection at the intervals of 0.2 pH units along the linear gradient, where the elution profile was recorded at 280 nm. At the end of the gradient, the column was flushed with 1 M sodium chloride (Sigma) to remove any proteins still bound to the column. All collected samples were stored at -80^{o}C until further analysis.

Non-porous silica reversed-phase HPLC: Each CF fraction was loaded onto a non-porous silica reversed-phase (NPS-RP) HPLC column for further separation. An ODSIII-E (8 x 33mm) column (Eprogen, Inc., Darien, IL) packed with 1.5 um non-porous silica

was used to achieve high separation efficiency. The separation was performed at a flow rate of 1 mL/min using a water/acetonitrile solvent system (A was 0.1% TFA in deionized water and B was acetonitrile and 0.1% TFA) and the gradient used was: 5-15% B in 1 min, 15-25% B in 2 min, 25-31% B in 3 min, 31-41% B in 10 min, 41-47% B in 3 min, 47-67% B in 4 min, 67-100% B in 1 min, followed by maintaining the system at 100% B for 3 min. Separation was monitored at 214 nm using a Beckman 166 model UV detector (Beckman-Coulter). Purified protein peaks were collected in deep-well 96 well plates using an automated fraction collector (model SC 100; Beckman-Coulter), controlled by in-house-designed DOS-based software. The column was maintained at 60^{o}C during separation to enhance reproducibility, speed and resolution. Following protein fractionation, the samples were stored at -80^{o}C until further use.

Protein microarrays:

1. **Array spotting:**

All fractions were transferred to shallow-well print plates (Bio-Rad) and were lyophilized to dryness. The samples were resuspended in printing buffer, consisting of 62.5 mM Tris-HCl (pH6.8), 1% w/v sodium dodecyl sulfate (SDS), 5% w/v dithiothreitol (DTT) and 1% glycerol in 1X PBS, and were left agitating on an orbital shaker overnight. Printing was accomplished by depositing 5 droplets of ~500 pL each per fraction using a piezoelectric dispenser (Nanoplotter 2, GeSiM). Distance between spots was maintained at 600 μm and spot sizes were found to be ~450 μm. Prior to processing all slides were kept sealed in a dessicator.

2. **Array processing with ProQ Diamond dye:**

Slides were blocked overnight in 1% BSA (Roche) in 1X PBS-T (0.1% Tween 20). They were then incubated for 1 hr in ProQ Diamond phosphoprotein gel stain (Invitrogen). The slides were then washed in destaining solution (Invitrogen) 3 times for 10 min each, then rinsed with nuclease free water and dried by centrifugation. The slides were scanned in the green channel using an Axon 4000A scanner, and GenePix Pro 6.0 software (Molecular Devices, Sunnyvale, CA) was used for data acquisition and analysis. Spots were considered to be positively fluorescent if background subtracted intensity of the spot was $\geq$X2 the local background intensity around the spot.

3. Array processing with anti-tyrosine antibodies:

Slides processed and scanned with ProQ diamond dye were rehydrated and then incubated in mouse monoclonal antiphosphotyrosine, 4G10 clone antibody (Upstate, Charlottesville, VA) diluted to 2 ug/mL in probe buffer (5 mM magnesium chloride, 0.5 mM DTT, 0.05% TritonX 100 and 5% glycerol in 1X PBS). After primary incubation the slides were washed (5 times, 5 min each) in probe buffer. Secondary incubation was performed for 1hr using donkey anti-mouse antibody conjugated to fluorescent cy5 at a concentration of 1 ug/mL in probe buffer. The slides were finally washed (5 times, 5 min each) in probe buffer and scanned in the red channel. Once again, spots were considered to be positively fluorescent if background subtracted intensity of the spot was $\geq$X2 the local background intensity around the spot.

Removal of SDS from samples: Prior to digestion and protein identification by mass spectrometry samples were cleaned using Detergent-OUT SDS-300 spin columns (G-Biosciences, St Louis, MO) to remove residual sodium dodecyl sulfoxide (SDS) that was

present during reconstitution into print buffer as per the user guide. In short, spin columns were inverted to re-suspend resin and liquid was drained off by spinning at 1000xg for 10 s. Columns were then equilibrated with 1.5 mL deionized water which was collected in a centrifuge tube and discarded. Sample was then applied to the spin columns and was let to stand for 5 minutes. After the columns were loaded they were centrifuged at 1000xg for 30 s and the SDS-free sample was collected in a centrifuge tube.

Trypsin digestion: The samples were dried down to 10 μL, and then 40 uL of 100 mM ammonium bicarbonate and 10 μL of 10 mM DTT were added to sample. The samples were incubated at 60^{o}C for 20 min to allow for reduction of disulfide bonds. 0.5 uL of TPCK modified sequence grade trypsin (Promega) was added and the samples were incubated at 37^{o}C overnight. Digestion was stopped by adding 1 uL of TFA to the digestion mixture.

Peptide sequencing by LC-MS/MS: Digested samples were separated by a capillary RP column (MagicAQ C18, 0.1 × 150 mm) (Michrom Biosciences, Auburn, CA) on a Paradigm MG4 micropump (Michrom Biosciences) with a flow rate of 300 nL/min. The gradient was started at 3% ACN, ramped to 35% ACN in 25 min, 60% ACN in 15 min, 90% in 1 min, maintained at 90% ACN for 1 min and finally ramped back down to 3% in another 1 min. Both solvents A (water) and B (ACN) contained 0.1% formic acid. The resolved peptides were analyzed on an LTQ mass spectrometer (Thermo, San Jose, CA) with an NANO-ESI platform (Michrom Biosciences). The capillary temperature was set at 200^{o}C, the spray voltage was 2.5 kV, and the capillary voltage was 20 V. The normalized collision energy was set at 35% for MS/MS. The top 5 peaks were selected for CID. Precursor selection was based upon a normalized threshold of 30 counts/s.

MS/MS spectra were searched using the SEQUEST algorithm incorporated in Bioworks software (Thermo) against the Swiss-Prot human protein database with Trypsin as the enzyme. Additional search parameters were as follows: (2) allowing two missed cleavages; (3) possible modifications, oxidation of M and phosphorylation of S, T and Y; (4) peptide ion mass tolerance 1.50 Da; (5) fragment ion mass tolerance 0.0 Da; (6) peptide charges +1, +2, and +3. The filter function in Bioworks browser was applied to set a single threshold to consider peptides assigned with Xcorr values as follows: ≥1.5 for singly charged ions, ≥2.5 for doubly charged ions, and ≥3.5 for triply charged ions.

2.3. Results and Discussion

The overall strategy we used for the large scale analysis of cellular protein phosphorylation status is outlined in Figure 2.1. Fractionation of the sample to reduce complexity, was achieved by separation in two dimensions, initially by chromatofocusing (according to the protein pI), and then by RP-HPLC, according to their hydrophobicity. Fractions were manually collected by peaks and each cell line resulted in approximately 1200 fractions after the complete 2-dimensional separation. The fractionated proteins were then printed onto microarrays and analyzed by hybridization with a universal phosphoprotein stain, and with antibodies specific to phosphorylated tyrosine residues. 140 spots were found to exhibit a positive response to the ProQ dye. Sequence analysis of specific phosphoproteins for confident identification was achieved by peptide sequencing using tandem MS/MS. This combinatorial approach overcomes many of the limitations inherent in single-method analyses. Phosphorylation sites have proved difficult to identify by mass spectrometry alone due to poor ionization efficiency and low abundance

of phosphopeptides. Additionally, mass spectrometric methods are not reliable for assessing global phosphorylation in a time-efficient manner. The proposed strategy is high-throughput in nature and a method of choice in initial screening to find differentially expressed proteins over the whole proteome in a sample of interest.

2D liquid separation and microarray reproducibility: A comparison of the 2-dimensional liquid separation (pH 4.0-7.2) is illustrated in Figure 2.2. On the left is a 2D UV map of the pre-malignant AT1 cell line, while on the right is the same for the malignant CA1a cell line. In the center is the comparison of the two maps. It can be seen that while the overall 2D maps are very similar for both cell lines, several differences are revealed. In particular, many proteins are more highly expressed in the malignant cell line, CA1a in the pH range 6.6-7.0 (corresponding to lanes 13 and 14 in Figure 2). Most of these proteins elute during the 1st half of the HPLC run. Sixty nine proteins were detected in the pH range 6.6-7.0 based upon LC-MS/MS experiments in the malignant CA1a cell line.

Comparative screening of the protein microarrays was achieved using the global phosphoprotein stain ProQ Diamond and antibodies specific to phosphorylated tyrosine residues. To investigate the binding properties of ProQ phosphor-stain and antibodies, protein and peptide standards were printed on SuperAmine slides. The slides were then probed initially with the phosphoprotein stain, ProQ Diamond, followed by a monoclonal anti-phosphotyrosine antibody (Figure 2.3a). While ovalbumin and β-Casein solely contain phosphoserine and phosphothreonine residues and therefore fluoresce green as a result of staining, the phosphotyrosine peptide (pY) mixture appears red. This occurs

because the antibody for phosphotyrosine displaces the ProQ and binds to the phosphotyrosine residues present in that spot. Subsequently, a red fluorescently tagged secondary antibody (in this case, an anti-anti-phosphotyrosine antibody conjugated to cy5) binds to the primary anti-phosphotyrosine antibody resulting in a red spot. A section of microarray generated by spotting of pre-malignant AT1 and malignant CA1a is also shown in Figure 2.3b. It can be seen that several fluorescing protein spots indicate the presence of phosphorylation. More importantly, figure 2.3b shows that the protein contents that were being used in the 2-dimensional separation were sufficient for microarray analysis.

Given the dynamic nature of cellular phosphorylation, we undertook a reproducibility study in order to better indicate the biological relevance of our phospho-profile findings. 3 separately grown CA1a cell line batches and 2 separately grown AT1 cell line batches were independently subjected to the entire analytical strategy, including 2D liquid separations, protein microarray and mass spectrometry. Several pH ranges were selected to assess reproducibility for all samples.

Figure 2.4 illustrates the results obtained. When looking at the chromatofocusing result (Figure 2.4a.), where pH fractions as collected could be monitored online for pH via a pH electrode assembly, it can be seen that for all separations a reproducible pH gradient was obtained. Furthermore, it can be seen for the CA1a cell line that all separated samples resulted in very similar and reproducible separation profiles. Similar separation profiles were also observed for the 2 batches of AT1 cell lines run. However, although the peak patterns were very similar they were not identical as in the case of CA1a. This difference was explained by the fact that while all other samples were loaded at a total protein

content of 4.5 mg, one of the AT1 samples had a lower total protein content of only 3 mg which resulted in an overall lower signal during the acquisition of the chromatogram. A comparison of the two batches of chromatograms suggests some subtle differences between CA1a and AT1 particularly in the higher pH range of about 7.0-6.2 and in the lower pH range around 5.6-5.2.

In order to further assess these subtle differences, selected pH ranges were subjected to NPS-RP-HPLC. Example chromatograms illustrating these separations are shown in Figure 2.4b. A high level of reproducibility is seen in both the independently grown batches of CA1a and AT1 samples analyzed. Furthermore, the subtle differences that were seen in the CF profiles are better visualized in the 2nd dimension. It can be seen that the malignant CA1a cells contains more hydrophilic protein peaks relative to the pre-malignant AT1 cells.

Fractionated samples from the 2nd dimension were arrayed on glass slides and probed with ProQ diamond dye to assess the phosphorylation status of the proteins. It is possible that while the chromatograms appear reproducible, the phosphorylation status of the protein may not be the same, making it necessary to assess reproducibility at the microarray level. Five slides were printed and probed with ProQ dye to assess the reproducibility of the printing and hybridization process. Figure 2.4c shows slide images of spots that were arrayed from selected pH ranges. It can be seen that all spots show consistently similar size and shape indicating that the printing process is consistent and reproducible. Slight variation in background intensities between the slides can be attributed to variation in slide surfaces and experimental variation during hybridization. However, these variations do not alter the number of positive spots of the array and

therefore do not affect the results significantly. Figure 2.4d illustrates sample biological reproducibility data obtained using the 3 CA1a and the 2 AT1 batches. It can be seen that for the pH range 6.4-6.2 there is a phosphorylated protein that elutes around retention time 26 min for all samples of CA1a and AT1 that were analyzed. However, for the pH range 5.2-5.0 there is a phosphoprotein (retention time 28 min) that is present only in CA1a samples. The reproducibility experiment revealed that consistent, differential phosphoprotein expressions were achievable across samples and batches.

It was also important to verify that the proteins present in consistently detected spots on the microarray were in fact the same proteins across samples. To this end, SDS was removed from selected sample fractions to be printed on arrays (as outlined in the methods section), proteins were trypsinized and then analyzed by tandem MS. Table 2.1 shows the protein IDs of the two spots that appeared positive for the CA1a samples in the pH range 6.4-6.2. In all cases, the proteins present in specific microarray spots were the same proteins. These analyses show that the strategy and the techniques are highly reproducible and confirm that the differential expression of specific phosphoproteins is maintained in the MCF10A tumor progression model.

Cell-Associated phosphoprotein profiles: All spots representing the same region of the 2D UV map from the two cell lines were compared to identify differential phosphorylation profiles. Pre-malignant and malignant samples were printed on microscopic glass slides with a chemically modified amine surface for studies with ProQ and antiphosphotyrosine antibodies. For each comparison, at least 5 replicate slides were processed. Of the phosphoproteins whose modification sites were identified, 11 proteins

were seen to be phosphorylated in the pre-malignant cell line but not the malignant cell line, and 16 proteins were seen to be phosphorylated in the malignant cell line but not the pre-malignant cell line. Examples of the differences observed, together with the identity of the protein as determined by tandem mass spectrometry, are illustrated in Figure 2.5. In some cases a protein eluted over multiple peaks due to diffusional broadening during sample collection. These proteins appear in multiple spots in the figures. Furthermore, there were instances where more than one phosphoprotein eluted at almost the same retention time. In these cases both protein identities are shown in the figure. Overall, 51 phosphorylation sites from a total of 27 proteins were identified. In addition, 47 previously reported phosphoproteins were also identified, but no phosphorylation site verification was obtained through the MS/MS data. Although dynamic exclusion was used to ensure that peptides eluting over a longer time were not continuously selected for tandem MS/MS analysis, it is possible that more sites were not identified due to the low signal intensities of phosphopeptides which rendered them undetectable using the top 5 ion peak selection used during our tandem MS runs. Furthermore, 3 phosphoproteins were shown to not be differentially expressed in the two cell lines. All phosphoproteins identified with site validation are listed in Table 2.2, along with information about the number of peptides and protein coverage pertinent to protein identification. Table 2.3 also lists all phosphoproteins identified without site validation. Site verification was not possible for these peptides due to low sample amounts. We did investigate phosphopeptide enrichment using titanium dioxide tips to improve yield, but without improvement. The results presented herein correlate well with previous work where approximately 155 spots in a 2D gel stained positive for phosphorylation using the ProQ

Diamond dye.[34] In another study about 100 proteins showed a change in phosphorylation upon stimulation of fibroblast cells where detection on 2D gels was facilitated by using antiphosphotyrosine and antiphosphoserine antibodies.[35] Our proposed strategy bypasses the problems associated with 2 dimensional gel electrophoresis but provides equivalent and complementary information about protein phosphorylation at the intact protein level which can be useful especially when site verification from mass spectrometric data presents a difficulty due to poor phosphopeptide spectra, which is often the case.

The pie chart in Figure 2.6 shows the cellular distribution of proteins whose modification sites were verified regardless of whether the phosphorylation was found in the pre-malignant or malignant cell line. Interestingly, of the 27 differentially phosphorylated proteins whose modified sites could be verified by tandem mass spectrometry, 18 were nuclear proteins (about 67% of all proteins identified). This trend of differential phosphoprotein expression in the nuclear region was also observed for those proteins whose sites were not verified. Closer examination of the proteins showed that the malignant CA1a cell line exhibited increased phosphorylation of nuclear proteins compared to the pre-malignant AT1 cell line.

It should be noted that a majority of the proteins that were detected and identified as being differentially phosphorylated in this work are of high to medium abundance. In this work we observed 85 differentially expressed spots (corresponding to a total of 75 phosphoproteins of which we were able to identify phosphorylation sites from 27 proteins) although we observed a total of 140 protein spots that responded to ProQ Diamond dye. The information that can be found from this work can therefore shed light

on the downstream effects of phosphorylation signaling cascades. However information about the very first changes that occur in a pathway were not detected since these occur on molecules with very low copy numbers in the cell which are generally below the detection limit of the ProQ dye.

An interesting phenomenon that we observed in our experiments was the shifts in pI due to phosphorylation. For example, in table 2.2 it can be seen the protein Lamin A/C appears multiple times. This protein was seen over more than one pH range. In addition it was found that the phosphorylation sites on the protein that were detectable using the unenriched samples were different for each pH range where the protein was observed. This phenomenon illustrates an important aspect about the effect of post translational modifications on protein pI. Previous work from our lab has shown that addition of a post translational modification on a protein changes the protein pI[36] and microarray data from this study further support these findings.

Functional grouping of phenotype-associated phosphoprotein profiles: Many of the differentially expressed phosphoproteins identified in this study fall under distinct categories with respect to the biological processes in which they are involved. Figure 2.7 summarizes these proteins according to their functional role in cellular processes. The majority of differentially phosphorylated proteins were found to be upregulated in the malignant CA1a cell line. A few key proteins that were found to be more phosphorylated in the non-malignant AT1 appear in a box with broken lines in the same figure. Transcriptional and translational proteins were in the majority, while mitotic and apoptosis-related proteins were also represented. In addition, a separate class of enzymes

as well as proteins that maintain cytoskeletal integrity were observed to change in their phosphorylation state as a function of malignant cellular phenotype. A discussion of some of the known roles of the phosphoproteins identified in this study is given below.

- Apoptotic signaling:

Proteins involved in the regulation of apoptosis are important determinants of cell proliferation and survival in malignant phenotypes. Stimulatory growth factor signaling and inhibitory stress factors initiate signal transduction pathways that regulate apoptosis via altering the phosphorylation of key regulating proteins. Three proteins important in the regulation of apoptosis, Bad, Bax and Acinus were differentially phosphorylated in AT1 and CA1a cells. While phosphorylated acinus was only found in CA1a, Bad and Bax phosphorylated forms were uniquely seen in AT1.

Growth factor induced phosphorylation of BAD protects cells from apoptotic stimuli. PI3K/Akt, Ras/MAPK/Rak, and PKA pathways all phosphorylate BAD. When serines at 112, 136, and 155 are phosphorylated, BAD is bound to an inactive complex.[37] In LNCaP human prostate cancer cells, phosphorylated sites necessary for activity varied with the survival signaling pathway.[38] Because malignant cells would be expected to have diminished sensitivity to apoptotic signals, phosphorylation of BAD in AT1 relative to CA1a suggests that additional sites other than the previously reported critical three serines are phosphorylated in AT1 cells.

The consequence of phosphorylated threonines at 135 and 140 in Bax in AT1 cells is unknown. Both apoptotic and anti-apoptotic activities have been associated with phosphorylation at different sites in other cells. Phosphorylation of serine 184 inhibits pro-apoptotic function of Bax in A549 human lung cancer cells[39] whereas

phosphorylation of threonine 167 in Bax activates apoptotic activity in HepG2 human hepatoma cells.[40]

Acinus, apoptotic chromatin condensation inducer in the nucleus protein (Accesion number Q9UKV3), is also a direct target of Akt and phosphorylation on serines 422 and 573 inhibits apoptosis in HEK293 cells, possibly by preventing caspase-mediated cleavage to a form that is necessary for chromatin condensation and apoptosis.[41] Acinus was uniquely seen to be phosphorylated in only the malignant CA1a cells, according to both the microarray and mass spectrometry data. The phosphorylation site that was identified was located on S1004, as shown in Figure 2.8a. Multiple peptides from the protein were sequenced, some of which were in the a.a 800-900 region of the protein. Interestingly it is known that the active form of the protein is a caspase-cleaved isoform, p17 which consists of the sequence a.a 987-1093. It was thus confirmed that the unprocessed, and therefore inactive, isoform was present in the cell line suggesting the absence of apoptotic chromatin condensation. Suppression of apoptosis may be instrumental to the malignant nature of the cell line.

- Transcriptional regulation:

This study showed that several proteins involved in transcriptional regulation were differentially phosphorylated in the two cell lines. Several histones were more phosphorylated in CA1a. Histones are typically positively charged to hold the negatively charged DNA in its condensed form. Phosphorylation of histones imparts negative charge so that DNA is less tightly bound and is thus available for manipulation. Zfp-36 and nucleolar phosphoprotein p130 are transcriptional regulatory proteins that were seen to be more phosphorylated in the malignant CA1a. SAF-B is a scaffold attachment factor that

regulates the formation of the transcriptosomal complex and is also thought to be a corepressor of the estrogen receptor, a pivotal factor in breast cancer phenotypes. SAF-B is known to decrease cell proliferation by reducing transcription of HSP-27. Interestingly this protein was phosphorylated in the pre-malignant AT1.

- Protein synthesis:

In addition to an increase in transcription-related phosphorylation, a parallel increase was seen in translational proteins in the malignant CA1a cell line compared to the pre-malignant AT1. Protein identifications as confirmed by tandem mass spectrometry showed the expression of large numbers of ribosomal proteins in malignant CA1a compared to AT1. These protein IDs are listed in table 2.4. The higher level of expression of ribosome related proteins suggests increased translational activity in the malignant breast cancer cell line. One of these proteins, 60S ribosomal protein L14, was confirmed to be phosphorylated on serine residue 138. No phosphorylation sites on this protein have previously been reported. When comparing the region of the reverse phase chromatogram where this protein eluted (Fig 2.8B.), it can be seen that distinct and unique peak patterns are evident in both the CA1a and AT1 cell lines. L14 ribosomal protein was only identified in the CA1a cell line.

- Mitosis:

Malignant cells tend to have increased rates of mitosis due to their proliferative nature. Proteins involved in mitotic spindle formation appeared to be differentially phosphorylated between the two cell lines. One such protein, Stathmin (Op18) was uniquely phosphorylated in only the pre-malignant AT1. This protein regulates the microtubule filament system by destabilizing microtubule assembly.

Nuclear migration protein (NudC) and microtubule associated protein (MAP4) are involved in correct formation of mitotic spindle. NudC is also involved in cytokinesis and cell proliferation. A higher expression of phosphorylated NudC could be indicative of the malignant nature of the CA1a cell line.

Heat Shock protein beta-1 is a stress related protein which is found in the cytoplasm but which colocalizes with mitotic spindles and migrates to the nucleus during stress. Increased phosphorylation of this protein in the malignant cell line could act as a signal for localization to a particular part of the cell.

Nuclear envelope disintegration is an integral component of mitosis. Lamins provide a framework for the nuclear envelope and may also indirectly interact with chromatin. In both cell lines, different forms of Lamin were confirmed to be phosphorylated by tandem mass spectrometry. Lamins are known to be extensively phosphorylated prior to nuclear disintegration during the mitosis process. Six phosphorylation sites were found on Lamin A/C in the CA1a malignant cell line, of which 2 had not been previously reported (S17 and S18). In addition, 7 sites were found on Lamin A/C in the pre-malignant AT1 cell line. Three of these sites were the same as the ones found in the CA1a cell line, while 4 were unique, of which 1 was predicted to be phosphorylated although no experimental evidence has been previously reported. Lamin phosphorylation is involved in regulation of Lamin interactions making the differential phosphorylation of this protein between the two cell lines particularly noteworthy.

- Enzymes:

Few proteins involved in anabolic or catabolic enzymatic processes showed phosphorylation differences between AT1 and CA1a cells. However, 2 examples with

relevance to cancer progression were aromatase and alpha enolase. Alpha enolase (MBP1) is a multifunctional enzyme playing a role in many processes, including glycolysis and growth control. When MPB1 binds to the c-myc promoter, it acts as a transcriptional repressor. Alpha-enolase has been implicated as a potential diagnostic marker for many cancers. In this study, MPB1 was identified in a phosphorylated form in only the AT1 cell line. Aromatase catalyzes the conversion of testosterone to estradiol. It has been reported that a kinase activity may be involved in the regulation of this catalytic process.[42] In this study, a phosphorylated form of aromatase was uniquely found in the pre-malignant AT1 cell line. It is plausible that phosphorylation of this enzyme renders it inactive. Consequently, the absence of estradiol in the pre-malignant AT1 may reduce the proliferative capability of the cell line.

Differential expression of proteins in pI range 7.0-6.6: Both the first and second dimension chromatograms suggested an increased level of protein in the higher pH separation range in the malignant CA1a cell line as compared to the pre-malignant AT1 cell line. Proteins from the range 7.0-6.6 in the malignant, CA1a cell line were analyzed and identified by tandem MS to see if this increase was specific to any particular class of proteins. Figure 2.9 shows a chromatogram of the 2^{nd} dimension separation in the 7.0-6.8 pH range. Interestingly, most identified proteins were ribosomal proteins and other proteins that regulate ribosomal function and genesis as shown in table 2.4. A majority of the proteins identified are known to be phosphorylated and often times the presence or absence of phosphorylation determines their location or activation status in the cell. Furthermore, a large number of positive spots in the microarray (which suggested that the protein in the spot was phosphorylated) corresponded to the fractions analyzed in this

high pH region as mentioned earlier. We were unable to locate the phosphorylation sites on all of these proteins, partly due to the low sequence coverage as most ribosomal proteins have low molecular weights. The theoretical iso-electric points of these ribosomal proteins are beyond the detection and separation capabilities of CF (between pH 8 and 11). It is likely that there appeared to be a higher expression of these proteins in the malignant CA1a because in fact these proteins were phosphorylated in the malignant cell line and therefore acquired a lower pI that made them detectable using the separation scheme used in these experiments.

2.4. Conclusion

We have presented a protein microarray approach coupled to 2D liquid separations for studying phosphorylation differences in a model of breast cancer progression. A comparison of pre-malignant versus malignant breast cells has not been previously reported using the strategy described here. A total of 51 phosphorylation sites in 27 different proteins were confirmed using tandem mass spectrometry and the status of these proteins was found to be specifically associated with the cellular phenotype. 48 additional previously known phosphoproteins were identified without site confirmation. The ontological association of the differentially expressed phosphoproteins included mitosis, apoptosis suppression and translational control. The research presented here illustrates the use of protein microarrays together with mass spectrometry as complementary tools to study phosphoproteins in complex samples. The microarray is often able to detect the presence of phosphorylation not detected by mass spectrometry without using enrichment techniques. When sample amounts are too low to permit enrichment the inability to detect phosphorylation by mass spectrometry becomes a critical issue, making the protein

microarray strategy a valuable alternative means of detecting high to medium abundance phosphoproteins which play a pivotal role in cellular phenotype. Site mapping by mass spectrometry would subsequently be needed for complete characterization; however the strategy outlined above can be used as an effective and rapid initial screen.

Table 2.1: Protein IDs and peptides identified for selected microarray spots that were reproducibly positive from pH range 5.4-5.3 (as shown in figure 2.4c).

Sample	Protein ID	Score	Peptides sequences	Coverage
Ca1a_spot1	Lamin-A/C	400	15	25
Ca1a_spot2	Lamin-A/C	410	19	33
	Protein disulfide-isomerase A3 precursor	370	15	31
Ca1a_spot1	Lamin-A/C	450	20	33
Ca1a_spot2	Protein disulfide-isomerase A3 precursor	380	15	29
	Lamin-A/C	340	17	30
Ca1a_spot1	Lamin-A/C	410	18	33
Ca1a_spot2	Lamin-A/C	480	20	34
	Protein disulfide-isomerase A3 precursor	360	14	30
AT1_spot1	Lamin-A/C	410	19	32
AT1_spot2	Protein disulfide-isomerase A3 precursor	240	11	21
	Lamin-A/C	180	8	14

Table 2.2: Phosphoproteins identified with confirmation of phosphorylation sites. Additional information was obtained from the Swissprot database.

Accesion number, Phosphoprotein	pH range	Pep. identified	% coverage	peptide + site	previously reported site	AT1	CA1A	cellular location
P50914 60S ribosomal protein L14	7.0-6.8	4	18	S138 (AALLKApSPK)	none		X	nucleus
Q14978 Nucleolar phosphoprotein p130	7.0-6.8	2	3	S303 (pSLGTQPPK)	pT607, pT610, pS623, pS643, pS698		X	nucleus
P83731 60S ribosomal protein L24	7.0-6.8	2	13	T24 (pTDGKVFQFLNAK)	pT83, pS86		X	
Q9BQ48 39S ribosomal protein L34	7.0-6.8	3	26	S89 (pSLSH)			X	mitochondria
Q9Y3U8 60S ribosomal protein L36	7.0-6.8	4	27	T17 (VpTKNVSK)			X	
P62318 Small nuclear ribonucleoprotein Sm D3	7.0-6.8	1	7	S93 (NQGpSGAGRGK)			X	nucleus
Q13428 Treacle protein	7.0-6.8	4	4	S959, S964 (IAPKApSMAGApSSSK)	pT173, pS890, pS1034, pS1151, pS1299, pS1301, pS1394		X	nucleus
Q9Y6Q3 Zinc finger protein 37 homolog	6.8-6.6	2	4	T234 (QDKIQpTGEKHEK)	none		X	nucleus
P02545 Lamin A/C	5.4-5.2	39	55	S17, S18 (SGAQApSpSTPLSPTR), S390, T394 (LRLpSPSPpTSQR)	S22, S390, S392, S652. By similarity S407, S496, T505, S507, T510,		X	nucleus
Q09666 Neuroblast differentiation associated protein AHNAK	5.2-4.8	2	4	T2727 (VpTFPKMKIPK)	S264, S312	X		nucleus
P07355 Annexin A2	5.2-5.0	3	7	S84 (ELApSALK)	S18, Y24, S26	X		plasma membrane
P11511 Cytochrome P450 19A1	5.2-5.0	1	2	T391 (KGpTNIILNIGR)	none	X		membrane
Q14562 ATP-dependent helicase DHX8	5.2-5.0	2	5	T914, T915 (DEMLpTpTNVPEIQR)	none		X	nucleus
Q02539 Histone H1.1.	5.2-5.0	2	6	T151 (KSVKpTPK), T203 (pTAKPK)	none	X		nucleus

Q03252 Lamin B2.	5.2-5.0	13	20	S402, S401, S400 (ATSpSpSpSGSLSATGR)	similarity S427	X		nucleus
P02545 Lamin A/C	5.2-5.0	36	56	S390, S392, T394 (LRLpSPpSPpTSQR) S403, S404, S406, S407 (ApSpSHpSpSQTQGGGSVTK)	S22, S390, S392, S652. By similarity S407, S496, T505, S507, T510,	X		nucleus
P02545 Lamin A/C	5.2-5.0	32	48	S390, S390, T394, (LRLpSPpSPpTSQR)	S22, S390, S392, S652. By similarity S407, S496, T505, S507, T510,		X	nucleus
P84103 Splicing factor, arginine/serine-rich 3	5.2-4.8	6	32	S108 (RRpSPPPR), S126, S128, S130 (pSRpSLpSR)	Extensively phosphorylated on serine residues in the RS domain	X		nucleus
Q09666 Neuroblast differentiation associated protein AHNAK	5.0-4.8	11	9	T2727 (VpTFPKMKIPK)	experimental S264, S312		X	nucleus
Q07812 Apoptosis regulator BAX, membrane isoform alpha.	5.0-4.8	2	6	T135, T140 (pTIMGWpTLDFLR)	none	X		membrane
P16403 Histone H1.2	5.0-4.8	3	18	S112 (KAApSGEAK)	similarity S36		X	nucleus
P08779 Cytokeratin 16	5.0-4.6	3	8	Y249 (EELApYLR)	none	X		cytoskeleton
P05787 Cytokeratin 8	5.0-4.8	20	41	S43 (VGSpSNFR)	S24, S74, S432, S451. By Similarity S9, S13, S22, T26, S27, S34, S37, S43, S417, S424, S475, S478	X		cutoskeleton
Q15424 Scaffold attachment factor B	5.0-4.8	2	1	S383, S384, S389 (MpSpSPEDDpSDTK)	by similarity S344	X		nucleus
P28001 Histone H2A.a	4.8-4.6	3	27	T120 (pTESHHK)	S2, T121		X	nucleus
O14929 Histone acetyltransferase type B catalytic subunit	4.8-4.6	2	5	S350, Y351 (pSpYRLDIKR)	none	X		nuclear in S Phase otherwise cytoplasmic
P08621 U1 small nuclear ribonucleoprotein 70 kDa	4.8-4.6	12	26	S226 (YDERPGPpSPLPHR)	S226		X	nucleus

P08670 Vimentin.	4.8-4.6	25	56	S54, S55 (SLYApSpSPGG VYATR), S28, Y29 (pSpYVTTSTR)	S5, S7, S8, S9, S10, S39, S42, S56, S72, S73, Y117, S420, S430, T458, S459. By similarity S25, S26, S34, S47, S51, S66, S83	X	
Q9UKV3 Apoptotic chromatin condensation inducer in the nucleus	4.6-4.4	8	6	S1004 (TAQVPpSPPR)	S240, S243, S365, S386, S388, S657, S661, S676, S1004, T414, T682. By similarity S384	X	nucleus
P20700 Lamin B1.	4.2-4.0	12	23	S395 (LSPSPSpSRVT VSRASSSR)	T20, S23, S391	X	nucleus
P02545 Lamin A/C	4.2-4.0	32	47	S17, S18 (SGAQApSpSTP LSPTR), S94 (KTLDpSVAK), S390 , S392, T394 (LRLpSPpSPpTS QR)	S22, S390, S392, S652. By similarity S407, S496, T505, S507, T510,	X	nucleus

Table 2.3: Previously known phosphoproteins also identified as differentially expressed in this study without confirmation of phosphorylation site(s). All additional information provided was obtained from the Swissprot database.

Accesion #, Phosphoprotein	pH range	# Peptides identified	previously reported site	Protein found in AT1	Protein found in CA1A	cellular location
P16989 DNA-binding protein A (Cold shock domain protein A) (Single-strand DNA binding protein NF-GMB).	7.0-6.8	3	experimental S34		X	nucleus, cytoplasm
P83731 60S ribosomal protein L24 (Ribosomal protein L30).	7.0-6.8	1	by similarity T83, S86		X	
P67809 Nuclease sensitive element binding protein 1 (Y-box binding protein 1) (Y-box transcription factor) (YB-1) (CCAAT-binding transcription factor I subunit A) (CBF-A)	7.0-6.8	3	expermental S102, Y162; by similarity S314		X	cytoplasm, nucleus (during stress)
P68104 Elongation factor 1-alpha 1 (EF-1-alpha-1) (Elongation factor 1 A-1) (eEF1A-1) (Elongation factor Tu) (EF-Tu).	6.8-6.6	4	experimetal Y29		X	cytoplasm, nucleus (during stress)
P04406 Glyceraldehyde-3-phosphate dehydrogenase, liver (EC 1.2.1.12) (GAPDH).	6.8-6.6	4	experimental Y42		X	cytoplasm
P13647 Keratin, type II cytoskeletal 5 (Cytokeratin 5) (K5) (CK 5) (58 kDa cytokeratin).	6.8-6.6	2	by similarity S64		X	cytoskeleton
P39023 60S ribosomal protein L3 (HIV-1 TAR RNA binding protein B) (TARBP-B).	6.8-6.6	3	experimental Y307		X	cytoplasm
P09651 Heterogeneous nuclear ribonucleoprotein A1 (Helix-destabilizing protein) (Single-strand binding protein) (hnRNP core protein A1).	6.8-6.6	2	T138, Y347. By similarity S6, 311.		X	nucleus, cytoplasm
P51991 Heterogeneous nuclear ribonucleoprotein A3 (hnRNP A3).	6.8-6.6	2	by similarity S355, 375		X	nucleus
P62753 40S ribosomal protein S6 (Phosphoprotein NP33).	6.8-6.6	2	by similarity S235, 236, 240, 244, 247		X	
P16403 Histone H1.2 (Histone H1d).	6.4-6.0	10	by similarity S35	X		nucleus
P62807 Histone H2B.a/g/h/k/l (H2B.1 A) (H2B/a) (H2B/g) (H2B/h) (H2B/k) (H2B/l).	6.4-6.2	8	experimental S14 (by STK4)	X		nucleus
P61604 10 kDa heat shock protein, mitochondrial (Hsp10) (10 kDa chaperonin) (CPN10) (Early-pregnancy factor) (EPF).	6.2-6.0	2		X		mitochondria
P23246 Splicing factor, proline-and glutamine-rich (Polypyrimidine tract- binding protein-associated splicing factor) (PTB-associated splicing factor) (PSF)	6.0-5.6	21	phosphorylated on multiple serine and threonine residues during apoptpsis	X		nucleus
P00441 Superoxide dismutase [Cu-Zn] (EC 1.15.1.1).	5.8-5.6	5		X	X	cytoplasm
P30101 Protein disulfide-isomerase A3 precursor (EC 5.3.4.1) (Disulfide isomerase ER-60) (ERp60) (58 kDa microsomal protein) (p58) (ERp57) (58 kDa glucose regulated protein).	5.4-5.2	18			X	endoplasmic reticulum
P04083 Annexin A1 (Annexin I) (Lipocortin I) (Calpactin II) (Chromobindin-9) (p35) (Phospholipase A2 inhibitory protein).	5.4-5.2	8	experimental S5, Y21, T24, S27, Y207	X	X	cytoplasm

P07355 Annexin A2 (Annexin II) (Lipocortin II) (Calpactin I heavy chain) (Chromobindin-8) (p36) (Protein I) (Placental anticoagulant protein IV) (PAP-IV).	5.4-5.2	8	experimental S18, Y24, S26		X	extracellular matrix
Q12906 Interleukin enhancer-binding factor 3 (Nuclear factor of activated T- cells 90 kDa) (NF-AT-90) (Double-stranded RNA-binding protein 76) (DRBP76) (Translational control protein 80)	5.4-5.2	8	experimental S62, T592		X	nucleus
P05783 Keratin, type I cytoskeletal 18 (Cytokeratin 18) (K18) (CK 18).	5.4-5.2	2	by similarity S7, T8, S18, S31, S34, Y36, S42		X	cytoskeleton
P13647 Keratin, type II cytoskeletal 5 (Cytokeratin 5) (K5) (CK 5) (58 kDa cytokeratin).	5.4-5.2	8	by similarity S64		X	cytoskeleton
P02538 Keratin, type II cytoskeletal 6A (Cytokeratin 6A) (CK 6A) (K6a keratin).	5.4-5.2	11	by similarity S60		X	cytoskeleton
P05787 Keratin, type II cytoskeletal 8 (Cytokeratin 8) (K8) (CK 8).	5.4-5.2	9	Similarity S9, S13, S22, T26, S27, S34, S37, S43, S417, S424, S475, S478 ; experimental S24, S74, S432, S451		X	cytoskeleton
P10599 Thioredoxin (ATL-derived factor) (ADF) (Surface associated sulphydryl protein) (SASP).	5.4-5.2	3	experimental T100		X	cytoplasm
Q92934 Bcl2-antagonist of cell death (BAD) (Bcl-2 binding component 6) (Bcl- XL/Bcl-2 associated death promoter) (Bcl-2-like 8 protein).	5.2-5.0	4	by similarity S75, S99, S118, S134	X		mitochondria, cytoplasm after phosphorylation
P50402 Emerin.	5.2-5.0	4	experimental Y59, Y74, Y85, Y95, Y99, Y161, Y163, Y167,	X		nucleus
P06733 Alpha enolase (EC 4.2.1.11) (2-phospho-D-glycerate hydro-lyase) (Non-neural enolase) (NNE) (Enolase 1) (Phosphopyruvate hydratase) (C-myc promoter-binding protein) (MBP-1) (MPB-1)	5.2-5.0	3	experimental Y44, Y287	X		cytoplasm, ncleus
P29966 Myristoylated alanine-rich C-kinase substrate (MARCKS) (Protein kinase C substrate, 80 kDa protein, light chain) (PKCSL) (80K-L protein).	5.2-4.8	5	Experimental S159, S163, S167, S170 ; S46, S118, S135, S262 by similarity	X	X	cytoskeleton
Q02543 60S ribosomal protein L18a.	5.2-5.0	1	experimental Y63	X		cytoplasm
P09651 Heterogeneous nuclear ribonucleoprotein A1 (Helix-destabilizing protein) (Single-strand binding protein) (hnRNP core protein A1).	5.2-5.0	5	similarity S6, S311 ; experimental T138, Y347	X		nucleus, cytoplasm
P08579 U2 small nuclear ribonucleoprotein B".	5.2-5.0	12	experimental Y151	X		nucleus
Q13242 Splicing factor, arginine/serine-rich 9 (Pre-mRNA splicing factor SRp30C).	5.2-5.0	4	experimental S189 ; by similarity S204, 211, 216	X		nucleus
P16949 Stathmin (Phosphoprotein p19) (pp19) (Oncoprotein 18) (Op18) (Leukemia-associated phosphoprotein p18) (pp17) (Prosolin) (Metablastin) (Pr22 protein).	5.2-4.8	6	experimental S16, 25, 38, 63	X		cytoplasm
O60506 Heterogeneous nuclear ribonucleoprotein Q (hnRNP Q) (hnRNP-Q) (Synaptotagmin binding, cytoplasmic RNA interacting protein) (Glycine- and tyrosine-rich RNA binding protein) (GRY-RBP)	5.0-4.8	4	experimental Y373	X		nucleus, cytoplasm
P05783 Keratin, type I cytoskeletal 18 (Cytokeratin 18) (K18) (CK 18).	5.0-4.8	15	by similarity S7, T8, S18, S31, S34, Y36, S42	X		cytoskeleton
P08729 Keratin, type II cytoskeletal 7 (Cytokeratin 7) (K7) (CK 7) (Sarcolectin).	5.0-4.8	18	similarity S14	X		cytoskeleton

Q02543 60S ribosomal protein L18a.	5.0-4.8	1	experimental Y63		X	cytoplasm
Q12906 Interleukin enhancer-binding factor 3 (Nuclear factor of activated T- cells 90 kDa) (NF-AT-90) (Double-stranded RNA-binding protein 76) (DRBP76) (Translational control protein 80)	5.0-4.6	7	experimental S62, T592	X		nucleus
P13647 Keratin, type II cytoskeletal 5 (Cytokeratin 5) (K5) (CK 5) (58 kDa cytokeratin).	5.0-4.6	17	by similarity S64	X		cytoskeleton
P15924 Desmoplakin (DP) (250/210 kDa paraneoplastic pemphigus antigen).	4.8-4.6	4	experimental S22, 176, 2024, 2209, 2815, 2820, 2825; probable S2849	X		cytoskeleton
P68104 Elongation factor 1-alpha 1 (EF-1-alpha-1) (Elongation factor 1 A-1) (eEF1A-1) (Elongation factor Tu) (EF-Tu).	4.8-4.6	6	experimental Y29		X	cytoplasm, nucleus (during stress)
Q9NZM5 Glioma tumor suppressor candidate region gene 2 protein (p60).	4.8-4.6	3	none		X	nucleus
P28001 Histone H2A.a (H2A/a) (H2A.2).	4.8-4.6	3	experimental S2, by similarity T121	X	X	nucleus
P61978 Heterogeneous nuclear ribonucleoprotein K (hnRNP K) (Transformation up-regulated nuclear protein) (TUNP).	4.8-4.6	7	experimental T3, S116, S284	X		nucleus, cytoplasm
P02538 Keratin, type II cytoskeletal 6A (Cytokeratin 6A) (CK 6A) (K6a keratin).	4.8-4.6	9	similarity S60	X		cytoskeleton
P08729 Keratin, type II cytoskeletal 7 (Cytokeratin 7) (K7) (CK 7) (Sarcolectin).	4.8-4.6	17	similarity S14	X		cytoskeleton
P27816 Microtubule-associated protein 4 (MAP 4).	4.8-4.6	3	by similary S253, S643 ; experimental S280, S507, T521, S696, S787, S825		X	nucleus
Q9Y266 Nuclear migration protein nudC (Nuclear distribution protein C homolog).	4.8-4.6	8	experimental S274, S326		X	nucleus, cytoplasm
P22626 Heterogeneous nuclear ribonucleoproteins A2/B1 (hnRNP A2 / hnRNP B1).	4.8-4.6	4	experimental S344		X	nucleus
Q13242 Splicing factor, arginine/serine-rich 9 (Pre-mRNA splicing factor SRp30C).	4.8-4.6	4	experimental S189 ; similarity S204, S211, S216		X	nucleus
Q14134 Tripartite motif protein 29 (Ataxia-telangiectasia group D-associated protein).	4.8-4.6	8	Constitutively phosphorylated by PKC on serine/threonine in A431 cells	X		cytoplasm
P30050 60S ribosomal protein L12.	4.6-4.4	2	experimental Y14		X	
P68104 Elongation factor 1-alpha 1 (EF-1-alpha-1) (Elongation factor 1 A-1) (eEF1A-1) (Elongation factor Tu) (EF-Tu).	4.2-4.0	6	experimental Y29		X	cytoplasm, nucleus (during stress)
P04792 Heat-shock protein beta-1 (HspB1) (Heat shock 27 kDa protein) (HSP 27) (Stress-responsive protein 27) (SRP27) (Estrogen-regulated 24 kDa protein) (28 kDa heat shock protein).	4.2-4.0	1	by similarity S26; experimental S15, 82, 83		X	cytoplasm (interphase), nucleus (heat shock)
P31948 Stress-induced-phosphoprotein 1 (STI1) (Hsc70/Hsp90-organizing protein) (Hop) (Transformation-sensitive protein IEF SSP 3521).	4.2-4.0	4	experimental Y354		X	nucleus, cytoplasm
P06753 Tropomyosin alpha 3 chain (Tropomyosin 3) (Tropomyosin gamma) (hTM5).	4.2-4.0	6	experimental T252		X	cytoskeleton

Table 2.4: Early eluting proteins from pH 7.0-6.6 identified from the malignant CA1a cell line. Any non-experimental information was obtained from the Swissprot database.

Protein	Acc #	Peps seqd	Phospho sites identified	Function	Known phosphos
Nucleolar RNA helicase II (Nucleolar RNA helicase Gu) (RH II/Gu) (DEAD-box protein 21).	Q9NR30	6		cofactor for c-Jun-activated transcription	pS71, pS89, pS121, pS171
Proliferating-cell nucleolar antigen p120	P46087	6		regulation of the cell cycle	pS58, pS181, pT195, pT603, pS732
Ribosome biogenesis regulatory protein homolog	Q15050	4		Involved in ribosome biogenesis	
60S ribosomal protein L38	P63173	1			
60S ribosomal protein L36a-like	Q969Q0	2			
60S ribosomal protein L36a	P83881	4			
Nucleolar phosphoprotein p130	Q14978	2	S303 (pSLGTQPPK)	Related to nucleologenesis	pT607, pT610, pS623, pS643, pS698
60S ribosomal protein L29	P47914	1			
Protein CGI-117	Q9Y3C1	2		May play a role in the regulation of mRNA stability	pS25, pS234, pS391
Plasminogen activator inhibitor 1 RNA-binding protein	Q8NC51	2			
60S ribosomal protein L26-like 1	Q9UNX3	8			
60S ribosomal protein L24	P83731	2	T24 (pTDGKVFQFLN AK)		pT83, pS86
39S ribosomal protein L34	Q9BQ48	3	S89 (pSLSH)		
DNA-binding protein A	P16989	4		Binds to the GM-CSF promoter. Seems to act as a repressor	pS34
Nuclease sensitive element binding protein 1	P67809	5		Contributes to the regulation of translation by modulating the interaction between the mRNA and eukaryotic initiation factors	pS101, pY161, pS313
60S ribosomal protein L28	P46779	3			
40S ribosomal protein S6 (Phosphoprotein NP33)	P62753	1		controlling cell growth and proliferation through the selective translation of particular classes of mRNA	pS235, pS236, pS240, pS244, pS247
60S ribosomal protein L21	P46778	3			
60S ribosomal protein L13 (Breast basic conserved protein 1)	P26373	2			

H/ACA ribonucleoprotein complex subunit 3	Q9NPE3	1		Required for ribosome biogenesis and telomere maintenance	
60S ribosomal protein L27	P61353	3			
Histone H1.0	P07305	4		necessary for the condensation of nucleosome chains into higher order structures	
60S ribosomal protein L34	P49207	2			
60S ribosomal protein L38	P63173	3			
60S ribosomal protein L35a	P18077	6		bind to both initiator and elongator tRNAs	
60S ribosomal protein L36	Q9Y3U8	4	T17 (VpTKNVSK)		
Protein HT031	Q9Y3Y2	2			
60S ribosomal protein L31	P62899	6			pY103, pY108
NADH-ubiquinone oxidoreductase subunit B14.5a	O95182	3		Transfer of electrons from NADH to the respiratory chain	
Probable ribosome biogenesis protein RLP24	Q9UHA3	2		Involved in the biogenesis of the 60S ribosomal subunit	
40S ribosomal protein S11	P62280	6			
40S ribosomal protein S23	P62266	4			
Probable U3 small nucleolar RNA-associated protein 11	Q9Y3A2	7		Involved in nucleolar processing of pre-18S ribosomal RNA	
60S ribosomal protein L8	P62917	6			pY132
Histone H1x	Q92522	3		Histones H1 are necessary for the condensation of nucleosome chains into higher order structures	pS31
Cytochrome c oxidase polypeptide Vb, mitochondrial precursor	P10606	3		mitochondrial electron transport.	
40S ribosomal protein S24	P62847	4			
Mitochondrial 28S ribosomal protein S14	O60783	4			
Cytochrome c	P99999	4		Suppression of the anti-apoptotic members or activation of the pro-apoptotic members of the Bcl-2 family	
40S ribosomal protein S27-like protein	Q71UM5	2			
60S ribosomal protein L17	P18621	5			
60S ribosomal protein L35	P42766	3			
40S ribosomal protein S26	P62854	2			

Histone H1.1	Q02539	1		Histones H1 are necessary for the condensation of nucleosome chains into higher order structures	
60S ribosomal protein L23a	P62750	2		This protein binds to a specific region on the 26S rRNA	
Heterogeneous nuclear ribonucleoprotein A3	P51991	7		Plays a role in cytoplasmic trafficking of RNA	pS355, pS375
Heterogeneous nuclear ribonucleoprotein A1	P09651	10		Involved in the packaging of pre-mRNA into hnRNP particles	pS5, pT137, pY346
Histone H1.2	P16403	2		Histones H1 are necessary for the condensation of nucleosome chains into higher order structures	pS35
40S ribosomal protein S19.	P39019	3			
Signal recognition particle 14 kDa protein	P37108	2		crucial role in targeting secretory proteins to the rough endoplasmic reticulum membrane	
Heterogeneous nuclear ribonucleoprotein G	P38159	7		RNA-binding protein which may be involved in pre-mRNA splicing	pS58, S165, pS208
60S ribosomal protein L14	P50914	4	S138 (AALLKApSPK)		
THO complex subunit 4	Q86V81	4		Acts as chaperone and promotes the dimerization of transcription factors	
60S ribosomal protein L23	P62829	4			
Heterogeneous nuclear ribonucleoproteins A2/B1	P22626	11		Involved with pre-mRNA processing	pS259, pS344
60S ribosomal protein L30	P62888	3			pS9
ATP synthase coupling factor 6, mitochondrial precursor	P18859	3		one of the chains of the nonenzymatic component (CF(0) subunit) of the mitochondrial ATPase complex	
40S ribosomal protein S25	P62851	4			
60S ribosomal protein L10-like	Q96L21	4		May play a role in compensating for the inactivated X-linked gene during spermatogenesis	
Mitochondrial 39S ribosomal protein L27	Q9P0M9	2			
40S ribosomal protein S20	P60866	3			pT9
60S ribosomal protein L32	P62910	7			
40S ribosomal protein S15	P62841	3			
60S ribosomal protein L10	P27635	2			
40S ribosomal protein S8	P62241	5			

Heterogeneous nuclear ribonucleoprotein C-like dJ845O24.4	O60812	4		nucleosome assembly by neutralizing basic proteins such as A and B core hnRNPs	
Heterogeneous nuclear ribonucleoproteins C1/C2	P07910	2		Binds pre-mRNA and nucleates the assembly of 40S hnRNP particles	pS253, pS260, pS299
Small nuclear ribonucleoprotein Sm D3	P62318	1	S93 (NQGpSGAGRG K)		
Treacle protein	Q13428	4	S959, S964 (IAPKApSMAGA pSSSK)	May be involved in nucleolar-cytoplasmic transport	pT173, pS890, pS1034, pS1151, pS1299, pS1301, pS1394

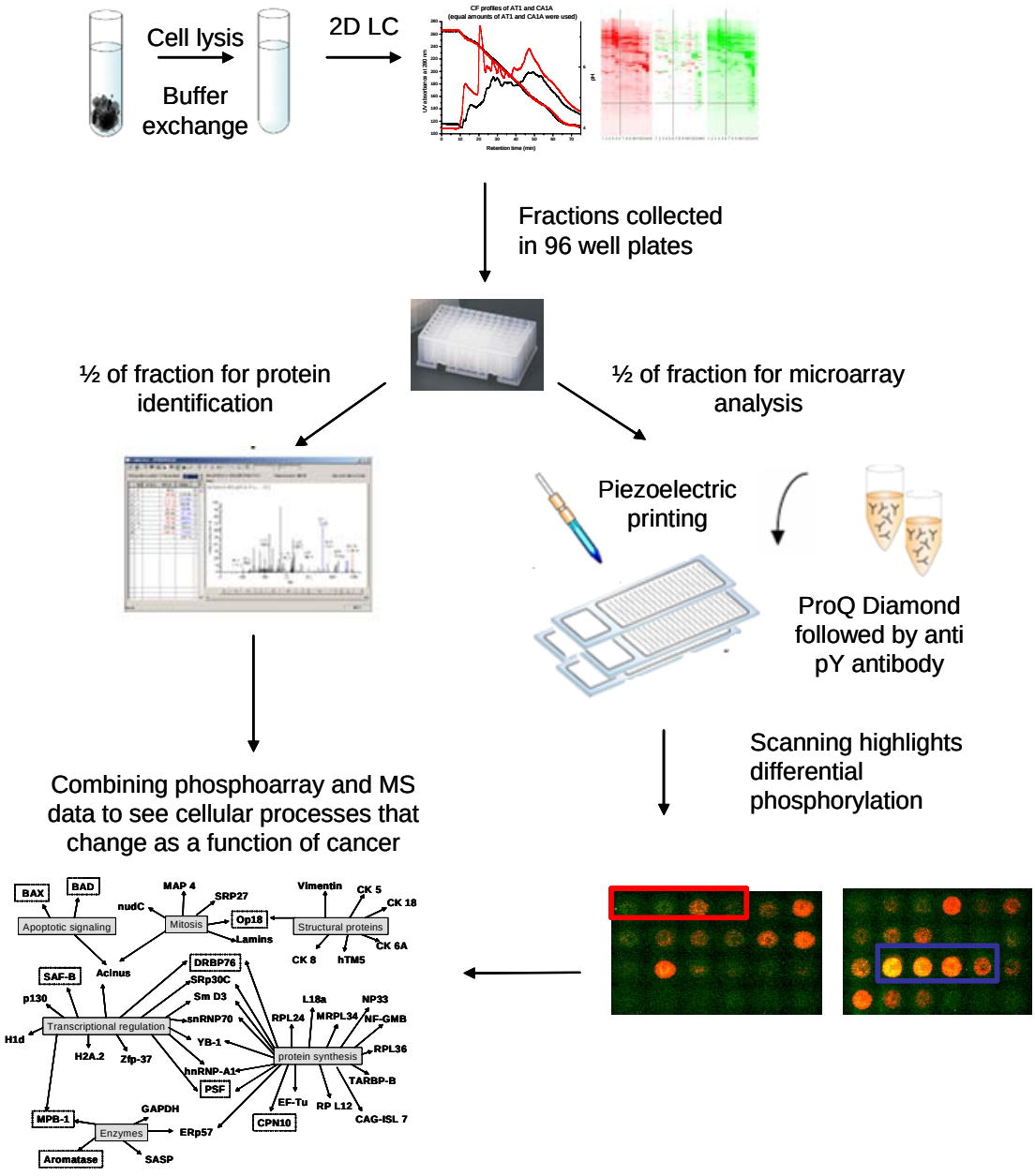


Figure 2.1: Microarray strategy for global evaluation of phosphorylation changes as a function of disease

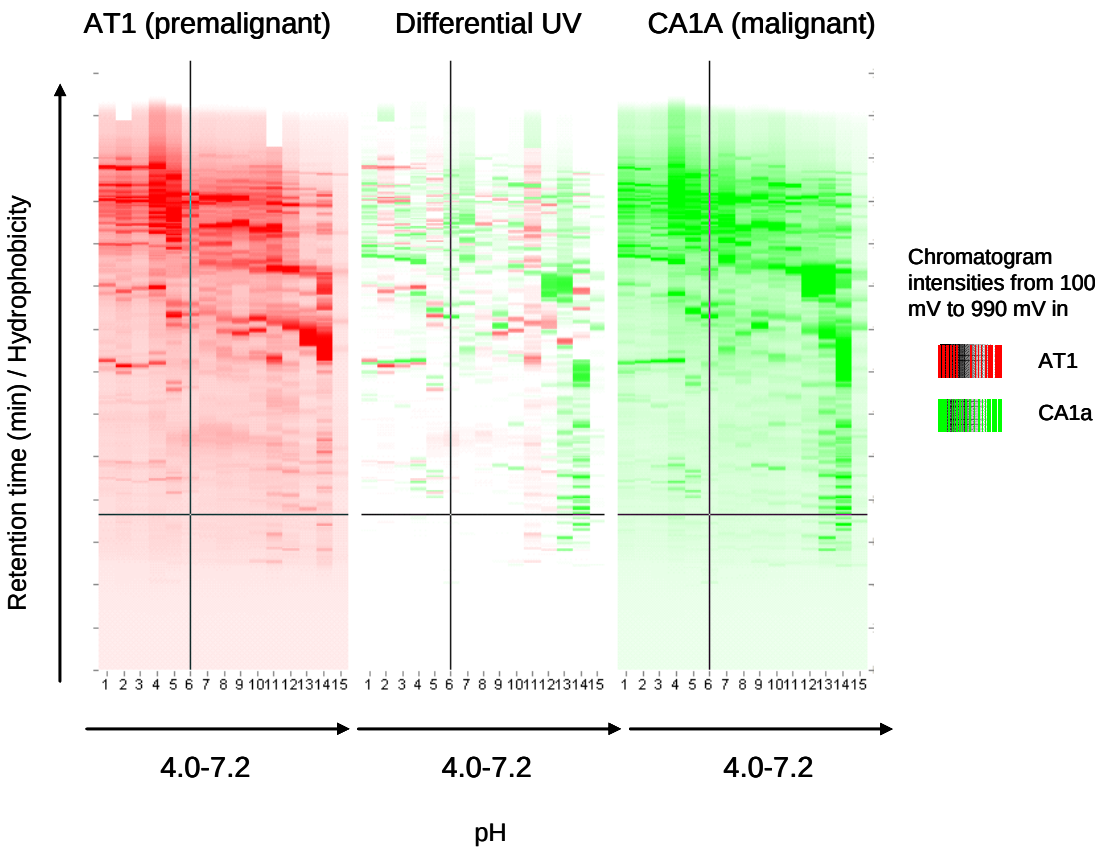


Figure 2.2: 2D liquid separation of pre-malignant AT1 and malignant CA1a cell lines. Each lane represents a pH fraction different by 0.2 units. Vertical axis refers to the retention time during the separation. Intensity of the bands corresponds to peak heights which ranged from a value of 100 mV to 990 mV. Difference between premalignant and malignant sample appears in the middle panel

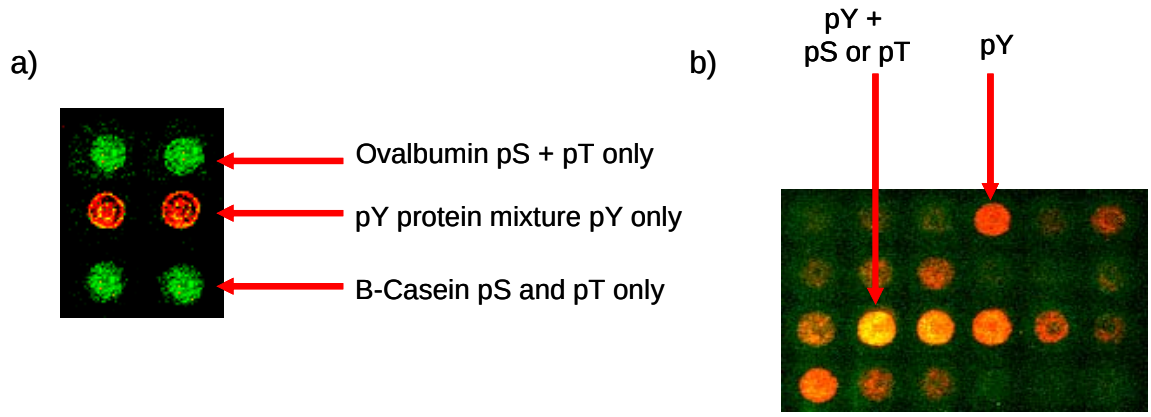
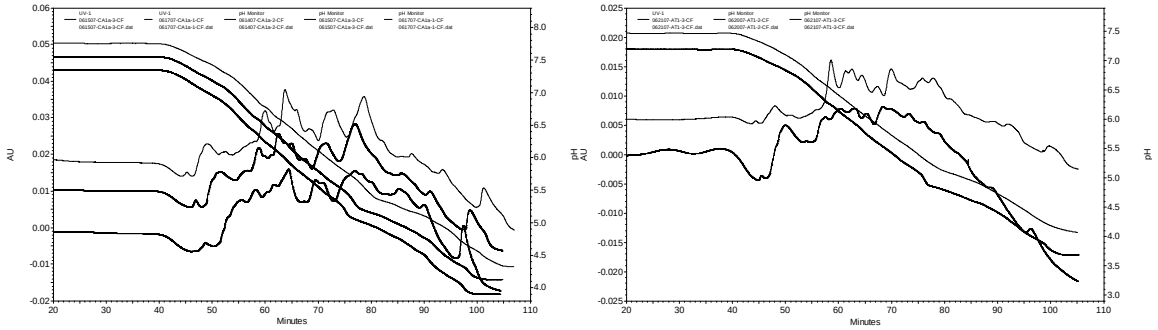


Figure 2.3: Detecting phosphoproteins on microarrays using ProQ Diamond dye and anti-phosphotyrosine antibodies. (a) A study done with standards where ovalbumin, B-casein and a mixture of tyrosine phosphorylated proteins were used. Notice that when probed with both ProQ and anti pY antibody, solely pY proteins appear red, mixture of pY and pS or pT appear yellow and solely pS or pT appear green. (b) An image of a section of a protein microarray containing fractionated proteins from a malignant breast whole cell lysate.

Figure 2.4:

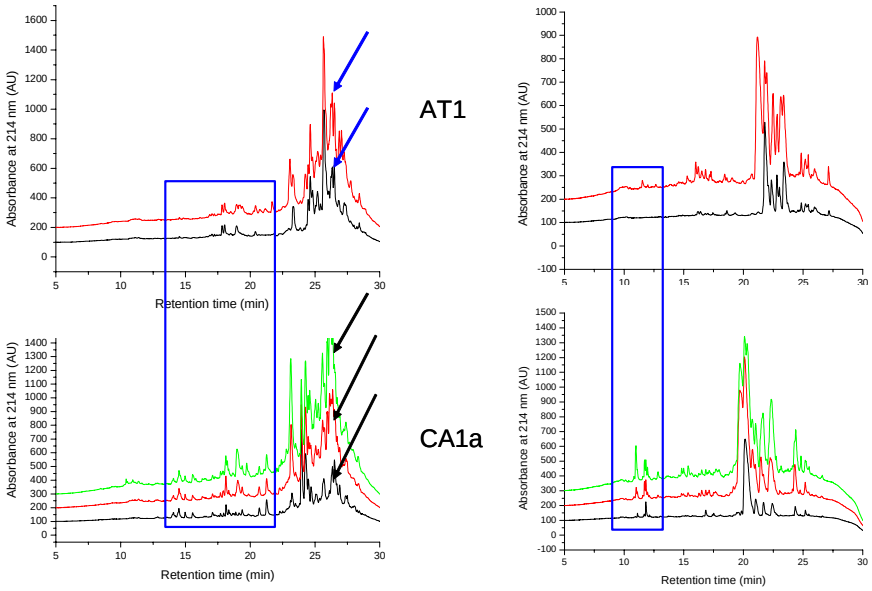
a)



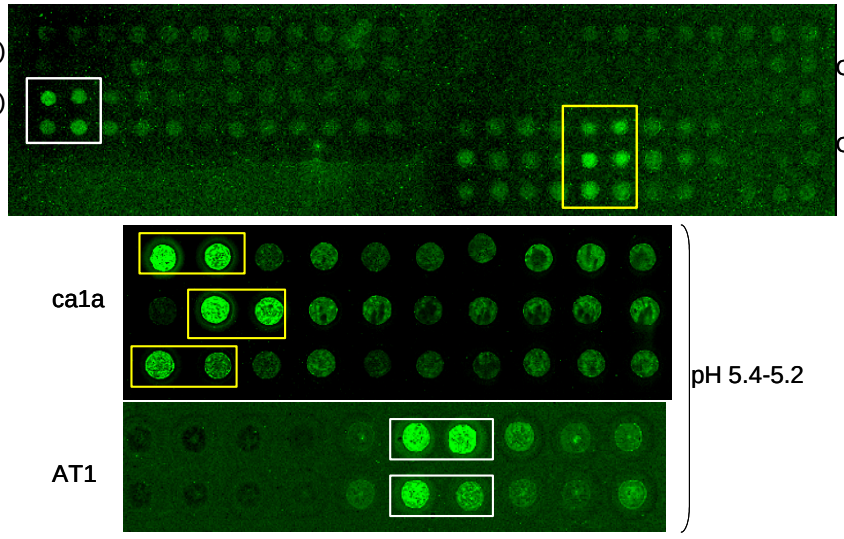
From CF fraction 6.4-6.2

From CF fraction 7.2-7.0

b)



c)



d)

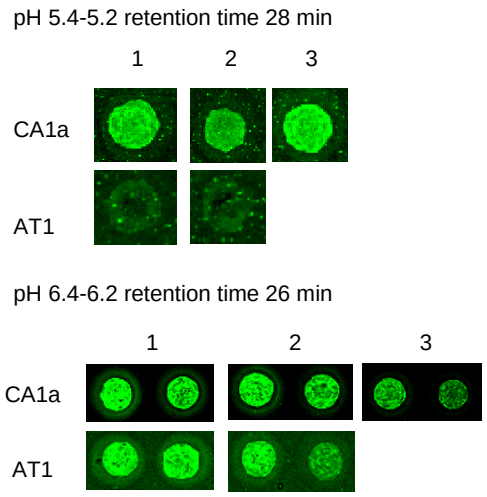


Figure 2.4: Comprehensive study to assess reproducibility of the method. (a) CF chromatograms of 3 ca1a separations are shown on the left and of 2 AT1 separations are shown on the right. In all cases 4.5 mg of sample was loaded (one AT1 separation was performed with only 3 mg of total protein). Co-plotted with the chromatograms are pH profiles to illustrate that the pH gradient was consistent in all separations. (b) 2nd dimension chromatograms of all batches of cell lines for pH ranges 6.4-6.2 and 7.2-7.0. Arrows along the chromatogram illustrate peaks that are shown in subsequent microarray data. (c) Array images of samples from pH fractions 7.2-7.0, 6.4-6.2 and 5.4-5.2 to illustrate reproducibility throughout the separation space. (d) Sections of microarray data showing an example of reproducible positive spots that are unique to ca1a (pH 5.4-5.2, retention time 28 min) and that are found in all cell lines (pH 6.4-6.2, retention time 26 min). Peaks corresponding to the positive spots found in all cell lines are indicated by arrows in fig 2.2b.

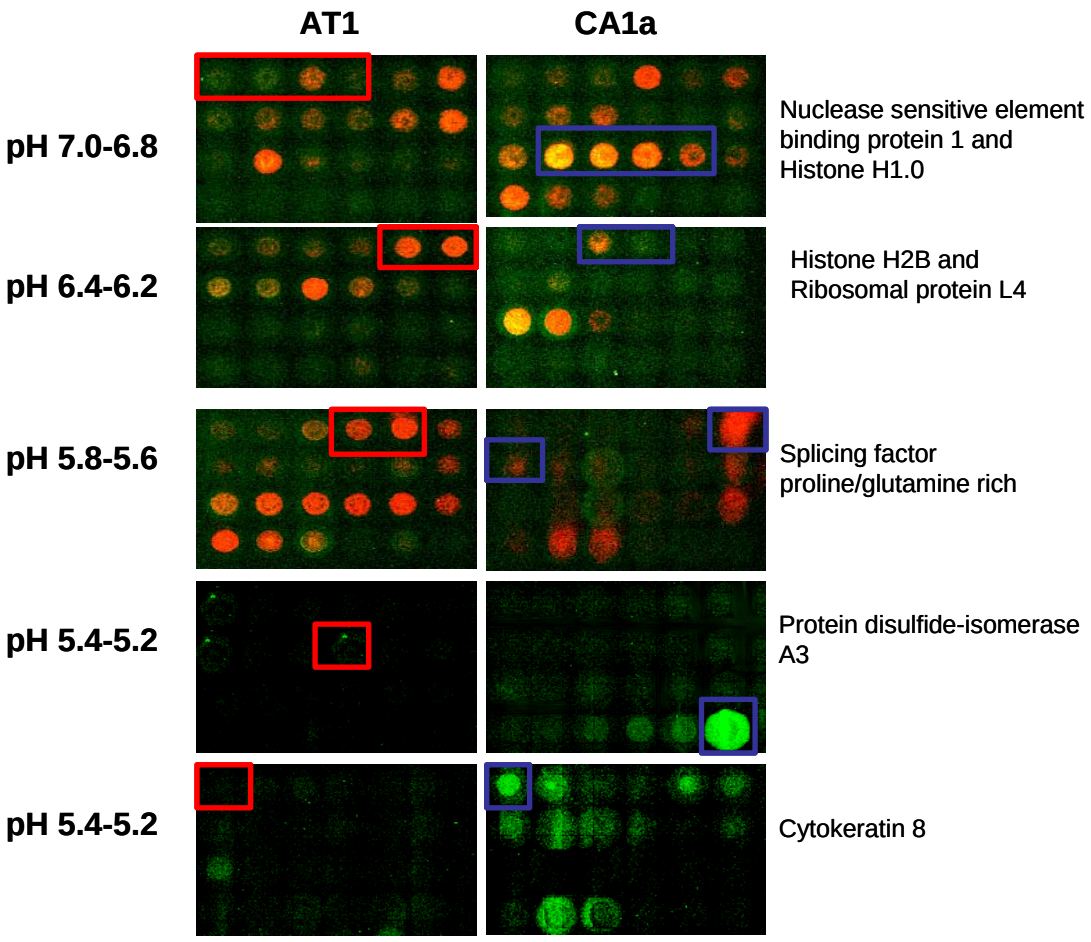


Figure 2.5: Selected microarray images showing comparison of spots where differential phosphorylation was observed between pre-malignant and malignant breast cell lines over different pH regions. Protein IDs as determined by tandem mass spectrometry are shown beside the image. For some proteins, multiple consecutive spots light up due to diffusional broadening during peak collection. In some cases more than one phosphoprotein was identified in the same collected fraction. In such cases, both phosphoproteins are listed.

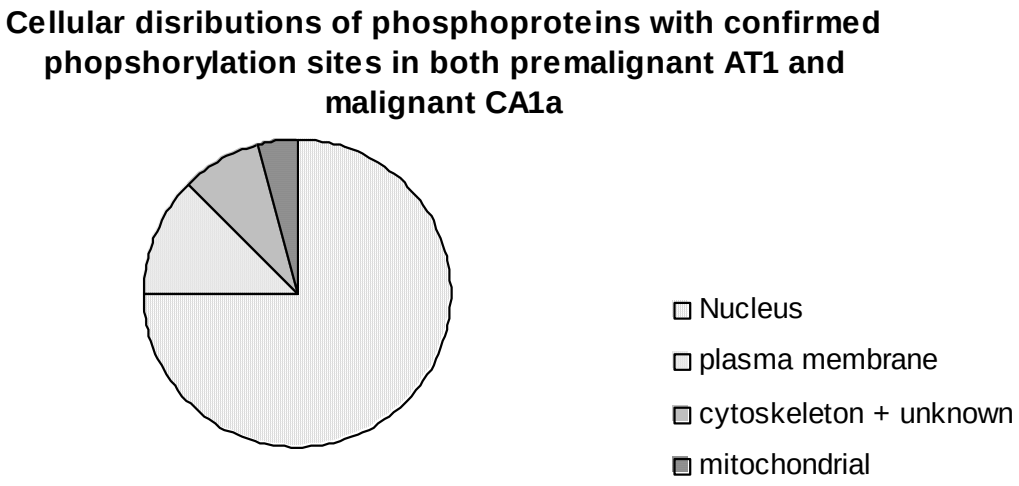


Figure 2.6: Pie chart illustrating subcellular location of phosphoproteins whose phosphorylation sites were confirmed by mass spectrometry in both the AT1 and CA1a cell line combined. Closer examination showed a majority of these phosphoproteins to be present in the CA1a cell line (see table 2.2).

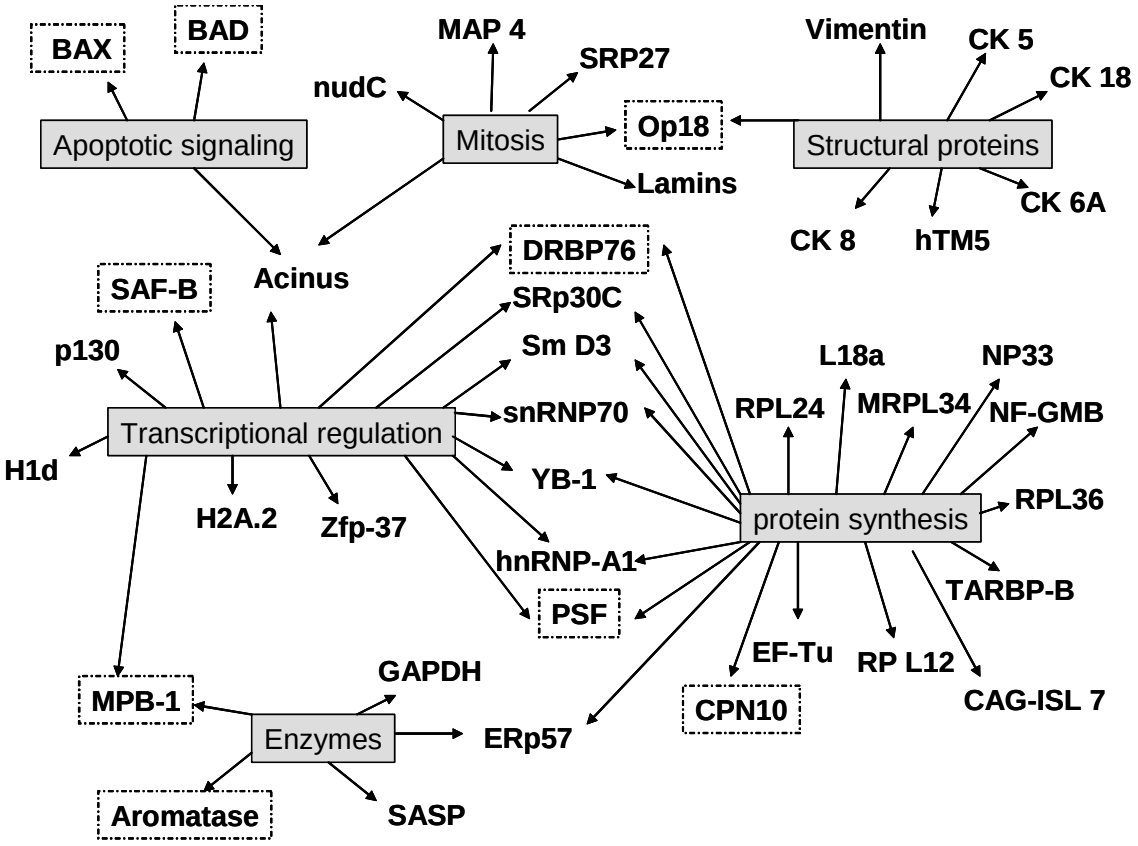


Figure 2.7: Functional classification of proteins differentially phosphorylated in the pre-malignant and malignant breast cell lines. Majority of phosphoproteins were found in the malignant, CA1a. In cases where a phosphoprotein was found in AT1 and not CA1a, it appears in a box with broken lines.

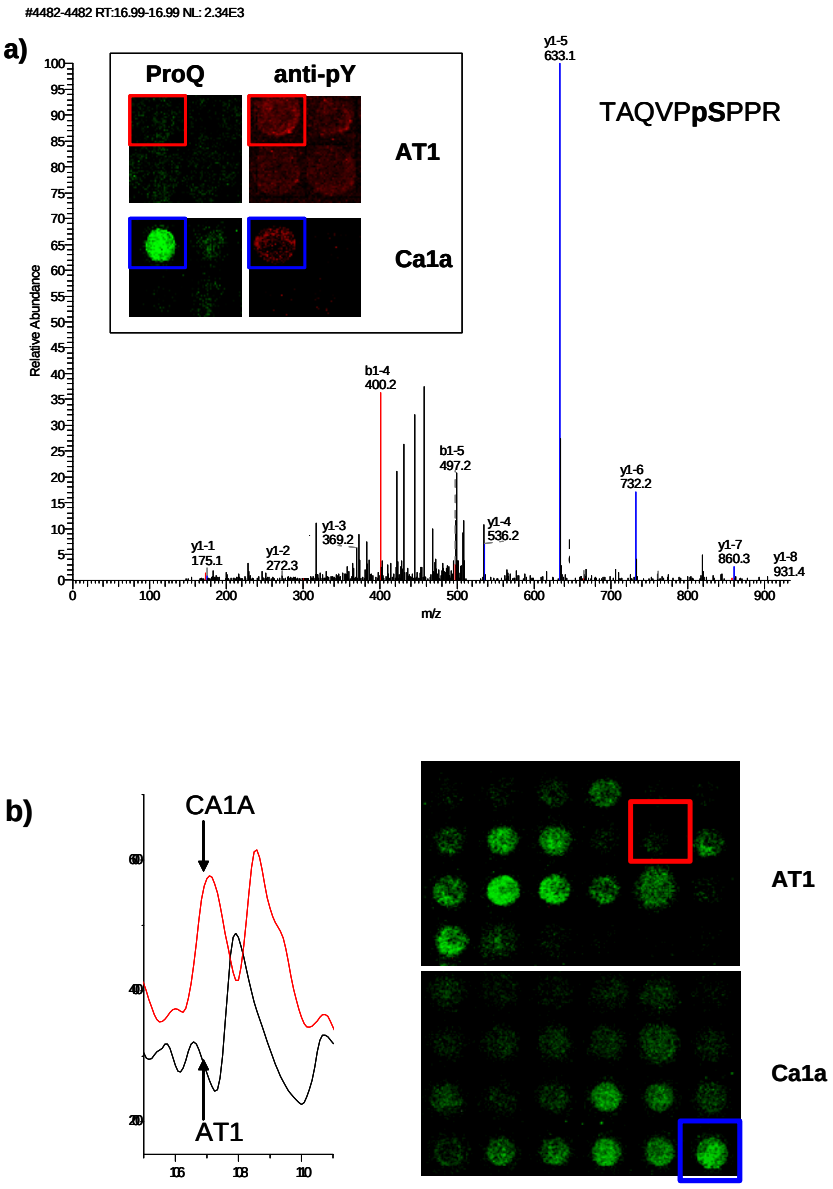


Figure 2.8: (a) Tandem mass spectrum (with +1 ion series highlighted) of selected phosphopeptide from apoptotic condensation inducing factor with inset showing phosphorylation difference between AT1 (boxed in red) and CA1a (boxed in blue) as seen on the microarray. (b) Microarray image together with complementary portion of reverse phase chromatogram where 60S ribosomal protein L14 was found to be phosphorylated in only CA1a.

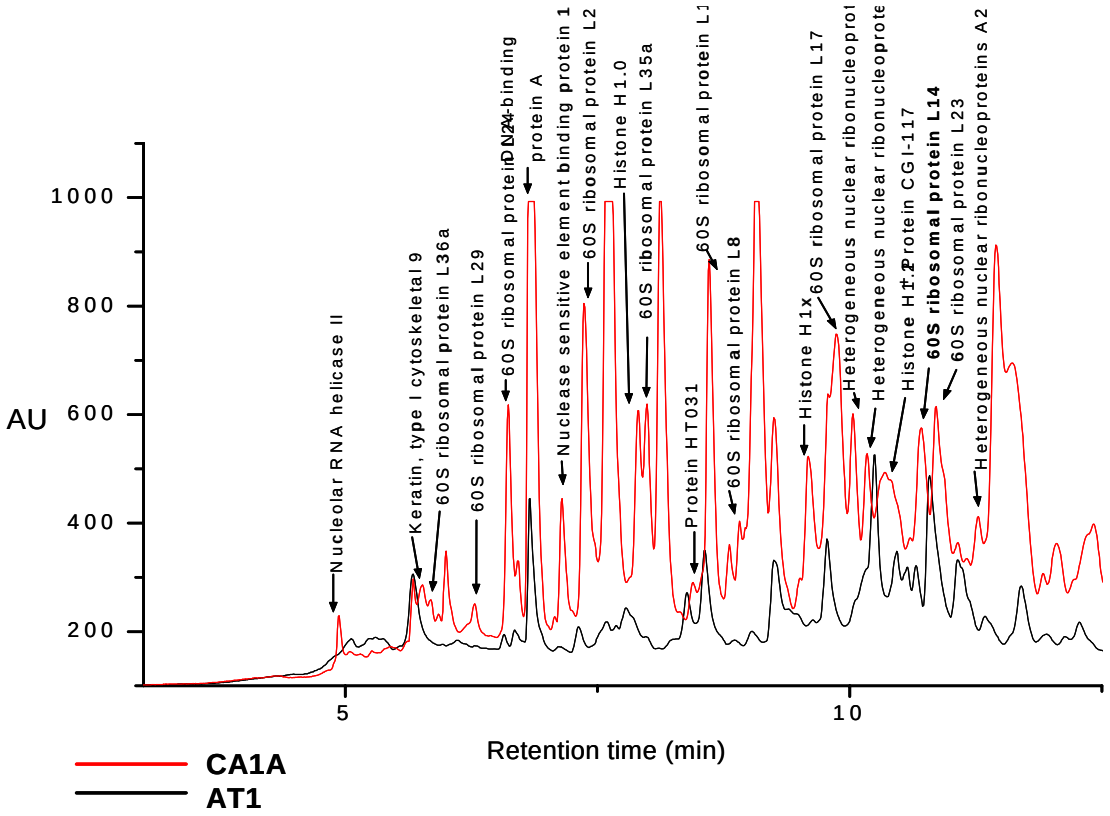


Figure 2.9: Reverse phase chromatogram of pH fraction 7.0-6.8 from malignant cancer cell line, CA1a and pre-malignant cell line AT1. IDs as determined by tandem mass spectrometry are shown for each peak in CA1a

2.5. References

[1] *American Cancer Society* 2006.

[2] Wolf-Yadlin, A., Kumar, N., Zhang, Y., Hautaniemi, S., *et al.*, *Mol Syst Biol* 2006, *2*, 54.

[3] Yu, Y., Hao, Y., Feig, L. A., *Mol Cell Biol* 2006, *26*, 6372-6380.

[4] Hunter, T., *Harvey Lect* 1998, *94*, 81-119.

[5] Chow, S., Minden, M. D., Hedley, D. W., *Exp Hematol* 2006, *34*, 1183-1191.

[6] Spickett, C. M., Pitt, A. R., Morrice, N., Kolch, W., *Biochim Biophys Acta* 2006.

[7] Cohen, P., *Eur J Biochem* 2001, *268*, 5001-5010.

[8] Perkins, N. D., *Oncogene* 2006, *25*, 6717-6730.

[9] Krueger, K. E., Srivastava, S., *Mol Cell Proteomics* 2006, *5*, 1799-1810.

[10] Morandell, S., Stasyk, T., Grosstessner-Hain, K., Roitinger, E., *et al.*, *Proteomics* 2006, *6*, 4047-4056.

[11] Tavares, A., Cimarosti, H., Valentim, L., Salbego, C., *Neuroreport* 2001, *12*, 2705-2709.

[12] Kaufmann, H., Bailey, J. E., Fussenegger, M., *Proteomics* 2001, *1*, 194-199.

[13] Goodman, T., Schulenberg, B., Steinberg, T. H., Patton, W. F., *Electrophoresis* 2004, *25*, 2533-2538.

[14] Martin, K., Steinberg, T. H., Goodman, T., Schulenberg, B., *et al.*, *Comb Chem High Throughput Screen* 2003, *6*, 331-339.

[15] Steinberg, T. H., Agnew, B. J., Gee, K. R., Leung, W. Y., *et al.*, *Proteomics* 2003, *3*, 1128-1144.

[16] Martin, K., Steinberg, T. H., Cooley, L. A., Gee, K. R., *et al.*, *Proteomics* 2003, *3*, 1244-1255.

[17] Cantin, G. T., Yates, J. R., 3rd, *J Chromatogr A* 2004, *1053*, 7-14.

[18] Cantin, G. T., Venable, J. D., Cociorva, D., Yates, J. R., 3rd, *J Proteome Res* 2006, *5*, 127-134.

[19] MacCoss, M. J., McDonald, W. H., Saraf, A., Sadygov, R., *et al.*, *Proc Natl Acad Sci U S A* 2002, *99*, 7900-7905.

[20] Andersson, L., Porath, J., *Anal Biochem* 1986, *154*, 250-254.

[21] Ndassa, Y. M., Orsi, C., Marto, J. A., Chen, S., Ross, M. M., *J Proteome Res* 2006, *5*, 2789-2799.

[22] Kweon, H. K., Hakansson, K., *Anal Chem* 2006, *78*, 1743-1749.

[23] Larsen, M. R., Thingholm, T. E., Jensen, O. N., Roepstorff, P., Jorgensen, T. J., *Mol Cell Proteomics* 2005, *4*, 873-886.

[24] Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., *et al.*, *Mol Cell Proteomics* 2002, *1*, 376-386.

[25] Gruhler, A., Olsen, J. V., Mohammed, S., Mortensen, P., *et al.*, *Mol Cell Proteomics* 2005, *4*, 310-327.

[26] Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., *et al.*, *Nat Biotechnol* 1999, *17*, 994-999.

[27] Thompson, A., Schafer, J., Kuhn, K., Kienle, S., *et al.*, *Anal Chem* 2003, *75*, 1895-1904.

[28] Sachon, E., Mohammed, S., Bache, N., Jensen, O. N., *Rapid Commun Mass Spectrom* 2006, *20*, 1127-1134.

[29] Hoffert, J. D., Pisitkun, T., Wang, G., Shen, R. F., Knepper, M. A., *Proc Natl Acad Sci U S A* 2006, *103*, 7159-7164.

[30] Pal, M., Moffa, A., Sreekumar, A., Ethier, S. P., *et al.*, *Anal Chem* 2006, *78*, 702-710.

[31] Dawson, P. J., Wolman, S. R., Tait, L., Heppner, G. H., Miller, F. R., *Am J Pathol* 1996, *148*, 313-319.

[32] Miller, F. R., *J Mammary Gland Biol Neoplasia* 2000, *5*, 379-391.

[33] Santner, S. J., Dawson, P. J., Tait, L., Soule, H. D., *et al.*, *Breast Cancer Res Treat* 2001, *65*, 101-110.

[34] Stasyk, T., Morandell, S., Bakry, R., Feuerstein, I., *et al.*, *Electrophoresis* 2005, *26*, 2850-2854.

[35] Godovac-Zimmermann, J., Soskic, V., Poznanovic, S., Brianza, F., *Electrophoresis* 1999, *20*, 952-961.

[36] Zhu, K., Zhao, J., Lubman, D. M., Miller, F. R., Barder, T. J., *Anal Chem* 2005, *77*, 2745-2755.

[37] Datta, S. R., Ranger, A. M., Lin, M. Z., Sturgill, J. F., *et al.*, *Dev Cell* 2002, *3*, 631-643.

[38] Sastry, K. S., Smith, A. J., Karpova, Y., Datta, S. R., Kulik, G., *J Biol Chem* 2006, *281*, 20891-20901.

[39] Xin, M., Deng, X., *J Biol Chem* 2005, *280*, 10781-10789.

[40] Kim, B. J., Ryu, S. W., Song, B. J., *J Biol Chem* 2006, *281*, 21256-21265.

[41] Hu, Y., Yao, J., Liu, Z., Liu, X., *et al.*, *Embo J* 2005, *24*, 3543-3554.

[42] Balthazart, J., Baillien, M., Ball, G. F., *J Steroid Biochem Mol Biol* 2001, *79*, 261-277.

Chapter 3

Screening of glycosylation patterns in serum using natural glycoprotein microarrays and multi-lectin fluorescence detection

3.1. Introduction

Glycoproteins are the most heterogeneous group of modifications known in proteins. Glycans show a high structural diversity reflecting inherent functional diversity. N- and O-oligosaccharide variants on glycoproteins (glycoforms) can lead to alterations in protein activity or function that may manifest itself as overt disease.[1, 2] Many clinical biomarkers and therapeutic targets in cancer are glycoproteins,[3-5] such as CA125 in ovarian cancer, Her2/neu in breast cancer and prostate-specific antigen (PSA) in prostate cancer. In addition, the alteration in protein glycosylation which occurs through varying the heterogeneity of glycosylation sites or changing glycan structure of proteins on the cell surface and in body fluids have been shown to correlate with the development and/or progression of cancer and other disease states.[6] Identification of glycoprotein isoforms is becoming increasingly important to the diagnosis and management of human diseases as more diseases are found to result from glycan structural alterations such as I-cell disease, and congenital disorders of glycosylation leukocyte adhesion deficiency type II.[7]

There are approximately 100 human glycan-binding proteins i.e. lectins according to genomic analysis.[8] In addition, the variety of lectin protein folds suggests that there may be additional lectin groups not yet discovered.[8, 9] A high throughput technique that can assess a diverse range of glycosylation states would facilitate research in this area. Furthermore, global screening of glycoprotein profiles in varied biological states can also potentially provide valuable information regarding key pathways that make that state unique.[10, 11]

Protein microarrays have proven to be useful as a high throughput screening method for whole cell lysates, fractionated proteomes, tissues and antigen-antibody reactions.[12-15] Increased interest in glycoproteomes has also sparked related research in the microarray field. A majority of current efforts have focused on carbohydrate microarrays.[8, 11, 16] In this approach various glycan-type structures are arrayed on a range of surface chemistries such as nitrocellulose, glass and dextran-type surfaces after which they are screened in parallel for binding. Such studies are critical in assessing antibody specificity to glycans and determining currently uncharacterized glycosylation structures that elicit responses in cells.[8, 11, 17] However oligosaccharides are difficult to synthesize due to varied stereochemistries, limited availability of enzymes for alternate synthesis strategies, and due to problems with purification when isolating naturally occurring oligosaccharides. Furthermore, the low mass and hydrophilic nature of most oligosaccharides makes non-covalent immobilization difficult for some glycans.[11] This problem has been overcome by successful covalent attachment of glycans to solid surfaces using film-coated photoactivable surfaces[18] and array coupling via flexible linker molecules.[19] Although carbohydrate arrays provide valuable information about

carbohydrate-interacting proteins, they do not allow us to directly study changes in glycosylation in real biological systems.

Current technologies for glycan analysis such as mass spectrometry[20], lectin affinity chromatography[21, 22] and western blotting are time consuming and some, such as mass spectrometry, require expertise and are technically difficult.[23] Lectin arrays can be used for rapid profiling of glycan expression patterns of various glycoproteins. Current studies using lectin arrays have focused on assessing specificity of arrayed lectins[24, 25] as well as changes in lectin binding of whole cell *E. coli* lysates that have undergone a treatment with sialyllactose to see changes in bacterial adhesion to cells.[18] Data from lectin arrays can be useful in determining the most appropriate lectins for glycoprotein enrichment as well as removal of undesirable glycoproteins. However the lectin array platform does not allow one to screen whole glycoproteomes in a way that can enable one to study both changes in overall glycoprotein patterns as well as changes in an individual protein's glycan expression within that glycoproteome.

A novel strategy presented here involves modifying the lectin array approach, making it more useful as a method that can study the total, as well as individual glycoprotein profiles of naturally produced glycoproteins. The strategy employs a liquid fractionated protein microarray approach to screen all glycoproteins in a complex sample on a single array. Glycoproteins are first enriched on a general lectin column and then separated by reverse-phase HPLC. The separated proteins are then arrayed on nitrocellulose slides and probed with lectins with a wide range of binding specificities. The glycoprotein-lectin interaction is assessed using a biotin-streptavidin system. As an example, we demonstrate the potential utility of this approach to identify serum biomarkers in pancreatic diseases.

This method allows us to profile the distribution of glycans in the human glycoproteome and also to study the changes in glycan expression on a global scale and on individual glycoproteins since each glycoprotein sample is a unique spot on the array.

3.2. Experimental Section

Standard preparation: Fetuin from fetal calf serum, asialofetuin from fetal calf serum, porcine thyroglobulin, bovine ribonuclease B, α-acid glycoprotein and human transferrin were purchased from Sigma. A stock solution of 20 mg/mL was made by dissolving standards in de-ionized water. A dilution series was made for each of the standard glycoproteins with the following final concentrations: 2, 1.6, 1.2, 1, 0.8, 0.6, 0.5, 0.4, 0.2, 0.1, 0.05, and 0.025 mg/mL. The dilutions were made directly into printing buffer (composition described in 2.5) to avoid drying and reconstitution in order to minimize sample loss.

Serum Samples: Serum was obtained at the time of diagnosis following informed consent using IRB-approved guidelines. Human normal serum was collected at University of Michigan under the auspices of the Early Detection Network (EDRN). Pancreatitis serum was obtained from patients with chronic pancreatitis who were seen in the Gastroenterology Clinic at University of Michigan Medical Center. Pancreatic cancer serum was obtained from patients with a confirmed diagnosis of pancreatic adenocarcinoma who were seen in the Multidisciplinary Pancreatic Tumor Clinic at the University of Michigan Comprehensive Cancer Center. 40 cc of blood was provided by each patient. The samples were permitted to sit at room temperature for a minimum of 30 minutes (and a maximum of 60 minutes) to allow the clot to form in the red top tubes and

then centrifuged at 1,300 x g at 4 ^{o}C for 20 minutes. The serum was removed, transferred to a polypropylene capped tube and frozen. The frozen samples were stored at -70 ^{o}C until assayed. All serum samples were labeled with a unique identifier to protect confidentiality of the patient. None of the samples were thawed more than twice before analysis. Samples were matched for age and sex to remove this variable from the analysis.

Lectin affinity glycoprotein extraction: An agarose bound lectin, Wheat Germ Agglutinin,(WGA) was purchased from Vector Laboratories (Burlingame, CA, USA). Agarose bound WGA was packed into a disposable screw end-cap spin column with filters at both ends. The column was first washed with 500 µl binding buffer (20 mM Tris, 0.15 M NaCl, pH 7.4) by centrifuging the spin column at 500 rpm for 2 min. Protease inhibitor stock solution was prepared by dissolving one complete EDTA-free Protease inhibitor cocktail tablet (Roche, Indianapolis, IN) in 1 ml H_2O. The stock solution was added to binding buffer and elution buffer at a ratio of (v/v) 1:50. 50 µl serum sample diluted with 500 µl binding buffer was loaded onto the column and incubated for 15 min. The column was centrifuged for 2 min at 500 rpm to remove the non-binding fraction. The column was washed with 600 µl binding buffer twice to wash off the non-specific binding. The captured glycoproteins were released with 150 µl elution buffer (0.5 M N-acetyl-glucosamine in 20 mM Tris and 0.5 M NaCl, pH 7.0) and the eluted fraction was collected by centrifugation at 500 rpm for 2 min. This step was repeated twice and the eluted fractions were pooled.

RP-HPLC separation of lectin-bound glycoproteins: The enriched glycoprotein fraction was loaded onto a nonporous silica reverse phase high-performance liquid

chromatography (NPS-RP-HPLC) column for separation. High separation efficiency was achieved by using an ODSIII-E (4.6x33 mm) column (Eprogen, Inc., Darien, IL) packed with 1.5 μm non-porous silica. To collect purified proteins from NPS-RP-HPLC, the reversed-phase separation was performed at 0.5 mL/min and monitored at 214 nm using a Beckman 166 Model UV detector (Beckman-Coulter). Proteins eluting from the column were collected by an automated fraction collector (Model SC 100; Beckman-coulter), controlled by an in-house designed DOS-based software program. To enhance the speed, resolution and reproducibility of the separation, the reversed phase column was heated to 60 ^{o}C by a column heater (Jones Chromatography, Model 7971). Both mobile phase A (water) and B (ACN) contained 0.1% v/v TFA. The gradient profile used was as follows: 5% to 15% B in 1 min, 15% to 25% B in 2 min, 25% to 30% B in 3 min, 30% to 41% B in 15 min, 41% to 47% B in 4 min, 47% to 67% B in 5 min and 67% to 100% B in 2 min. De-ionized water was purified using a Millipore RG system (Bedford, MA).

Glycoprotein microarray: Purified and separated glycoproteins, or glycoprotein standards (from 2.1), were printed on nitrocellulose slides (Whatman Schleicher & Schuell BioScience, Keene, NH) using a non-contact printer, Nanoplotter 2.0 (GeSiM, Germany). Prior to printing, the proteins were dried down in a 96-well plate and resuspended in 15 μL of printing buffer with stirring overnight at 4 ^{o}C. The printing buffer contained 65 mM Tris-HCl, 1% SDS, 5% dithiothreitol (DTT) and 1% glycerol. Each spotting event resulted in approximately 500 pL of sample being deposited by a piezoelectric mechanism. The event was programmed to occur 5 times per spot to ensure that approximately 2.5 nL were being spotted per sample. Each sample was further spotted as 9 replicates. The resulting spots were ~450 μm in diameter and the spacing

between spots was maintained at 600 μm. After printing the slides were allowed to dry for 24 hours. Blocking was achieved by incubation with 1% Bovine serum albumin (BSA) and 0.1% Tween-20 in 1X phosphate buffered saline (PBS) overnight. Blocked slides were probed with biotinylated lectin in a solution of PBS-T (0.1% Tween 20 in 1X PBS). The lectins used in the study were biotinylated Peanut Agglutinin (PNA), *Sambucus Nigra* bark lectin (SNA), *Aleuria Aurentia* (AAL), Concanavalin A (ConA) and *Maackia Amurensis* lectin II (MAL), all purchased from Vector Laboratories (Burlingame, CA, USA). The working concentration of all lectins used was 5 μg/mL except for SNA, which was used at 10 μg/mL as per vendor recommendation. After primary incubation all slides were washed with PBS-T 5 times for 5 minutes each. Secondary incubation was achieved with a streptavidin-AlexaFluor555 conjugate (Invitrogen, Carlsbad, CA) in a working concentration of 1 μg/mL containing 0.5% BSA, 0.1% Tween-20 in 1X PBS. After secondary incubation the slides are washed 5 times for 5 minutes each in PBS-T and completely dried using a high-speed centrifuge (Thermo Electron Corp., Milford, MA). The dried slides were scanned using an Axon 4000A scanner in the green channel and GenePix Pro 3.0 software (Molecular Devices, Sunnyvale, CA) was used for data acquisition and analysis.

Protein digestion by trypsin: Fractions obtained from NPS-RP-HPLC were concentrated down to ~20 μL using a SpeedVac concentrator (Thermo, Milford, MA) operating at 45 °C. 20 μl of 100 mM ammonium bicarbonate (Sigma) was then mixed with each concentrated sample to obtain a pH value of ~7.8. 0.5 μl of TPCK modified sequencing grade porcine trypsin (Promega, Madison, WI) was added and vortexed prior to a 12-16 hour incubation at 37 ^{o}C on an agitator.

Glycan cleavage by PNGase F and glycan purification: For glycan cleavage and purification, glycoproteins were dried down completely and redissolved in 40 µl 0.1% (w/v) RapiGest solution (Waters, Milford, MA) prepared in 50 mM NH_4HCO_3 buffer, pH 7.9 to denature the protein. Protein samples were reduced with 5 mM DTT for 45 min at 56 ^oC and alkylated with 15 mM iodoacetamide in the dark for 1 h at room temperature. 2 µl enzyme PNGase F (QA-Bio, Palm Desert, CA) was added to the samples and the solutions were incubated for 14 h at 37 ^oC. The glycans released were purified using SPE micro-elution plates (Waters) packed with HILIC sorbent (5 mg). The micro-elution SPE device was operated using a centrifuge with a plate adaptor (Thermo). Protein and detergent were removed during this step. Glycans were further cleaned by a graphitized carbon cartridge (Alltech, DeerWeld, IL) to remove salt. 25 % ACN with 0.05 % TFA was used to elute the carbohydrates.

Mass spectrometry

Protein identification by LC-MS/MS: Digested peptide mixtures from NPS RP HPLC collection were separated by a capillary RP column (C18, 0.3 x 150 mm) (Michrom Biosciences, Auburn, CA) on a Paradigm MG4 micro-pump (Michrom Biosciences) with a flow rate of 5 µl/min. The gradient started at 5% ACN, was ramped to 60% ACN in 25 min and finally ramped to 90% in another 5 min. Both solvent A (water) and B (ACN) contain 0.1% formic acid. The resolved peptides were analyzed on an LTQ mass spectrometer with an ESI ion source (Thermo, San Jose, CA). The capillary temperature was set at 175 ^oC, spray voltage was 4.2 kV and capillary voltage was 30 V. The normalized collision energy was set at 35% for MS/MS. MS/MS spectra were searched using the SEQUEST algorithm incorporated in Bioworks software (Thermo) against the

Swiss-Prot human protein database. One mis-cleavage is allowed during the database search. Protein identification was considered positive for a peptide with X_{corr} of greater than or equal to 3.0 for triply-, 2.5 for doubly- and 1.9 for singly charged ions.

Glycan structure analysis: MS and MS^2 spectra of glycan samples were acquired on a Shimadzu Axima QIT MALDI quadrupole ion trap-ToF (MALDI-QIT)(Manchester, UK). Acquisition and data processing were controlled by Launch-pad software (Karatos, Manchester, UK). A pulsed N_2 laser light (337 nm) with a pulse rate of 5 Hz was used for ionization. Each profile resulted from 2 laser shots. Argon was used as the collision gas for CID and helium was used for cooling the trapped ions. The TOF was externally calibrated using 500 fmol/ul of bradykinin fragment 1-7 (757.40 m/z), angiotensin II (1046.54 m/z), P14R(1533.86 m/z), and ACTH(2465.20 m/z) (sigma). 25 mg/ml 2,5-dihydroxybenzonic acid (DHB) (LaserBio Labs, France) was prepared in 50% ACN with 0.1% TFA. 0.5 µl glycan sample was spotted on the stainless-steel target and 0.5 µl matrix solution was added followed by air drying.

3.3. Results and Discussion

The glycoprotein microarray strategy: The methodology presented here and illustrated in figure 3.1, is a potential approach that can be used to study differences in glycans expressed on unique glycoproteins in complex biological samples. Following the strategy, serum is first purified and enriched for glycoproteins using a general lectin column. The enriched glycoproteins are further separated on a reverse-phase HPLC column. The resolved glycoproteins are then arrayed on nitrocellulose slides as unique protein spots after which they are screened for different glycan structures using five

different lectins. The lectin-binding event is visualized using a scanner by employing a biotin-streptavidin-alexafluor555 scheme. Differential glycosylation patterns can consequently be observed using image analysis software.

Standard glycoprotein microarrays: To determine the feasibility of using a glycoprotein microarray to study separated pre-purified glycoproteins, initial studies were done using standards with known glycan structures in order to assess the specificities of the lectins used, the quality of the processed arrays as well as to determine the range in which a linear response was observed for the concerned standard proteins.

Five standard glycoproteins were used to assess the feasibility of a glycoprotein microarray strategy. A dilution series of each glycoprotein was made using concentrations ranging from a blank with no sample to 2 mg/mL. Each dilution was printed in 9 replicates to assess the variability of spots from the same sample during a print run. Each sample spot on the array was achieved by depositing 5 droplets of approximately 500 pL each resulting in a total volume of 2.5 nL per spot by a piezoelectric mechanism. Consequently the standards spotted ranged from 0-2.5 ng.

Table 3.1 describes the binding specificities of the biotinylated lectins used for glycan detection. Five separate lectins were used for the analysis. ConA recognizes α-linked mannose including high mannose-type and mannose core structures. Both MAL and SNA recognize sialic acid on the terminal branches, while SNA binds preferentially to sialic acid attached to terminal galactose in an (α-2,6) and to a lesser degree, an (α-2,3) linkage.[26] MAL could detect glycans containing NeuAc-Gal-GlcNac with sialic acid at the 3 position of galactose.[27] In contrast, PNA binds de-sialylated exposed galactosyl (β-1,3) N-acetylgalactosamine. In fact, sialic acid in close proximity to the PNA receptor

sequence will inhibit its binding. AAL recognizes fucose linked (α-1,6) to N-acetylglucosamine or to fucose linked (α-1,3) to N-acetyllactosamine. Use of the combination of these five lectins should be highly successful in covering a majority of N-glycan types reported and differentiating them according to their specific structures.

Lectin specificity studies: The specificity of purchased lectins was assessed to ensure that they did not bind non-specifically. Five standard glycoproteins were used for this study, fetuin, asialofetuin, thyroglobulin, ribonuclease B and transferrin. The printed glycoprotein standards were incubated with biotinylated lectins for binding. The bound biotinylated lectins were subsequently detected with streptavidin conjugated to AlexaFluor555. This sandwich-type detection scheme was employed because the very specific biotin-streptavidin interaction should improve signal to noise ratio significantly. Figure 3.2 shows the images obtained when slides were probed with each of the lectins. Background fluorescence was at a minimum with the processing conditions used. Data illustrated in Fig. 3.3A supports previously reported glycan structures corresponding to the glycoproteins used in this study. It is known that the abundant glycan structures of bovine fetuin are sialylated, bi- and tri-antennary complex-type N-glycans (core non-fucosylated). The sialic acid residues are found in both (α-2,3) and (α-2,6) linkages.[28] Abundant glycans in asialofetuin include asialo-bi and asialo-tri antennary N-linked oligosaccharides. Dominant porcine thyroglobulin glycans include disialylated biantennary N-linked oligosaccharides with core fucose[29] and oligomannose N-linked oligosaccharide with 5-9 mannosyl residues.[30] The glycan of ribonuclease B is high mannose type i.e. $Man_{5-9}GlcNac_2$.[31] The dominant glycan in transferrin is sialylated, biantennary complex-type N-glycan.[32]

As shown in Fig. 3.3A, Con A binds strongly to thyroglobulin and ribonuclease B since both of their glycans contain oligomannose N-linked oligosaccharide with 5-9 mannosyl residues. Transferrin, fetuin and asialofetuin bind weakly to Con A as mannose residues are only present in their core structure and not in the exposed branches. SNA binds fetuin, thyroglobulin and transferrin, which have all been reported to possess sialic acid moieties on their glycans, while MAL only bound to Fetuin and porcine thyroglobulin, which have sialic acid attached in an (α-2,6) position. These two lectins can therefore be used to discriminate between sialic acid residues in an (α-2,3) vs (α-2,6) linkage due to the more specific interaction of MAL. 2-3 vs 2-6 sialylation of Lea antigens has been implicated in pancreatic cancer[33], supporting the use of multiple lectin detection schemes in microarray formats for explicit differentiation of glycan structures. This further shows the importance of using multiple lectin detection schemes in microarray formats for explicit differentiation of glycan structures. PNA bound to only asialofetuin since it is the only standard used that has de-sialylated, exposed galactosyl (β-1,3) N-acetylgalactosamine residues in its glycan structure. This lectin was also found to be the most specific lectin used. As shown in Fig 3.2. and Fig. 3.3A, AAL binds strongly to porcine thyroglobulin which is the only standard used whose main structure consists of disialylate, biantennary N-linked oligosaccharide with core fucose. There might be very low abundant fucosylated glycan attached to transferrin as reported in previous data[32] where only 2% of tranferrin glycans are reported to be fucosylated. The abundance might be below the detection limit of this lectin since the highest concentration of transferrin spotted on the slide in this work was only 2 mg/mL, which corresponds to 0.12 fmols absolute amount of the fucosylated transferrin based on previously reported data.

Linearity of response and detection limits of lectins: In all cases where standard proteins elicited response, the limit of detection was found to be between a concentration of 0.05-0.1 mg/mL. This corresponds to an absolute protein content of between 125 pg to 250 pg. On average, glycoproteins fall in the molecular weight range of about 50 kDa. Consequently, 125-250 pg translates into a 2.5 to 5 fmols detection limit. Mass spectrometric glycan structure determination often requires higher amounts of sample due to the need for multiple sample handling steps as well as MS^n fragmentation requirements for complete structural information. In the case of MAL where only fetuin was found to bind, the limit of detection was much higher at almost 1mg/mL protein concentration corresponding to 2.5 ng or 50 fmol total protein content. In this study all protein spots were measured to be approximately 450 μm in diameter. If the printing buffer composition is changed so that spots spread out to a lesser degree across the array surface, the density of sample per spot area could be increased possibly resulting in lower limits of detection.

To determine the linearity of response to individual lectins for each of the standard proteins, curves were generated based on the fluorescence response of all printed spots and their replicates. In addition to the 9 replicates on each slide, data points were collected from two processed slides for each lectin in order to assess the variability between slide images processed in the same manner and on the same day. We found that all proteins showed a linear response to each of the lectins within a 0.025 - 1 mg/mL concentration range. However linearity of response was optimal in a range of 0.025 - 0.5 mg/mL. Figure 3.3 shows some of the standard curves that were obtained. It was noticed that all standard curves were unique to the standard protein that was being used to

generate it. This is not surprising since a lectin does not measure quantity of a protein spotted but reflects the extent to which a particular glycan structure is expressed on that protein. To illustrate this we determined the dominant glycan structures on Ribonuclease B and Transferrin by tandem mass spectrometry. Figure 3.4 shows glycan structures and their corresponding mass spectra. Based on the mass spectra it is evident that ribonuclease B has a mannose-rich glycan structure not present in transferrin. This explains our findings in Fig. 3.3A, where even at the same concentration of standards, ribonuclease B responds to ConA to a much greater degree than transferrin.

Although some of the standards used contained sialic acid residues on their respective glycans, MALDI-based tandem mass spectrometry was often not sufficient to determine their presence (Transferrin in figures 3.3 and 3.4). The inability to detect sialic acid moieties on glycans due to in-source decay as ions transit from the MALDI target to the ion-trap has previously been reported.[34] In order to stabilize the fragile sialic acid moiety, modifications need to be made on the carboxyl group such as esterification[35] and permethylation[36] which require large amounts of sample and are often not feasible for biological, clinically relevant samples due to poor recovery for samples of low abundance. Using a glycoprotein microarray strategy together with mass spectrometry therefore appears to be a more complete means to characterize glycan structures on proteins.

Variation between spots and slides: Fluorescence values for all spots were used to assess the spot variability and reproducibility. We found standard deviations for all proteins and their binding to lectins to be within 10% of the mean value, all replicates considered. Standard deviation values were within 5% of mean when only replicates on

the same slides were taken into consideration. Our data suggests that printing occurs reproducibly and the variation between slides is most likely due to slight heterogeneity between slide surfaces and small differences in sample handling during slide processing. We have found that this small variation in handling is not significant enough to be problematic. PNA was the only lectin that showed much higher standard deviation (almost 20-25% from the mean). However upon closer examination we concluded that there was a great degree of difference between the spots in the two slides and when each slide was analyzed separately the standard deviation values again fell within 5-10% of the mean for all standards that showed a response to PNA (data not shown).

From the study with standards we have been able to show that glycoprotein microarrays can potentially be used to study differences in glycosylation states of individual proteins in more complex biological samples.

Studies with serum samples: Since studies with standards were successful in terms of reproducibility and sensitivity, we attempted to enrich and pre-fractionate glycoproteins from human serum and make a glycoprotein microarray to see if differences were evident in sera from biologically distinct states. In this case, we examined sera samples from patients who were not diagnosed with pancreatic disease or were diagnosed with chronic pancreatitis or pancreatic cancer. Such a strategy could be used with a wide range of biological samples following appropriate sample preparation protocols.

As illustrated in figure 3.1, serum was first purified for glycoproteins using Wheat Germ Agglutinin (WGA). WGA can bind oligosaccharides containing terminal N-acetylglucosamine or chitobiose as well as sialic acid residues, structures that are common to many serum and membrane glycoproteins. The purified and enriched

glycoproteins were then separated in a second dimension by non-porous reverse phase HPLC. This separation resolved the enriched glycoproteins into approximately 30 fractions. When 2.5 mg (~50 μL raw serum) serum proteins were enriched, approximately 100 μg of glycoproteins were typically recovered. Only half of this sample was run in the second dimension. After considering recovery from the reverse phase column and the number of fractions collected in the second dimension, it can be estimated that each fraction contained an average of 1-2 μg of protein (this amount is proportional to the height of relative peaks). All collected fractions were dried down and resuspended in 15 μL of printing buffer so that the working concentrations of the glycoproteins printed were in the range of 0.07-0.13 mg/mL. This range falls between the concentrations that were used for the standard glycoproteins ensuring similarity in parameters used in both studies.

To see if there were any changes in glycosylation patterns between sera from different biological states WGA enriched glycoproteins from normal and pancreatitis serum were fractionated and spotted on nitrocellulose slides. The reverse-phase chromatogram of enriched glycoproteins from the two sera samples showed some differences in peak heights. In addition to confirming the concentration difference shown by the different peak heights, the glycoprotein microarray also indicated a different glycosylation pattern for the observed differences. Fig. 3.5A shows the reverse phase chromatogram highlighting differences between the two samples.

Based on the peak heights alone it seemed that the peak highlighted in red is 2 to 3 times overexpressed in normal serum compared to pancreatitis serum. However microarray data in Fig. 3.5B indicated that response to some of the lectins for the same peak was

often more than 2 to 3 times in the normal serum compared to pancreatitis. To verify that this trend was due to change in glycan expression and not protein concentration, all data was normalized using integrated peak areas. After normalization it was found that the peak concerned expressed almost twice as much mannosylation and fucosylation while all other glycan structures assessed did not change significantly. μLC-MS/MS analysis identified the peak as a complement factor H precursor.

Additionally, the peak highlighted in orange showed another interesting trend. Although the peak height was less than two times higher in the pancreatitis serum compared to the normal serum, normalized response to AAL was more than 6 times higher in the normal sample as shown in Fig. 3.5C. This suggests that the protein concerned is much less fucosylated in chronic pancreatitis. Furthermore, the protein showed almost 4-fold higher expression of mannose on its glycans in normal vs. pancreatitis serum as seen by the normalized fluorescence intensities with Con A. Response to SNA, MAL and PNA was not significantly different for the same protein between the two samples. We found that normalization was necessary for a more accurate picture of differential glycan expression and in order to subtract any differences caused by overall protein abundance.

The glycoprotein shown in Fig. 3.5C was identified by tandem mass spectrometry and found to be α_1-acid glycoprotein precursors 1 and 2. This protein and changes in its glycosylation state have been implicated in various disease pathologies including pancreatitis where higher levels of the protein was seen in severe pancreatitis.[37, 38] Although our study as presented in this report, cannot claim that α_1-acid glycoprotein is a significant marker of pancreatitis, it does show that our novel strategy has the potential as a method that can identify such important markers.

Another separate study was done to see if any difference was apparent in enriched glycoproteins from normal versus pancreatic cancer sera. Pancreatic cancer is currently difficult to diagnose at an early stage due to lack of early diagnostic markers, and in some patients may be difficult to differentiate from chronic pancreatitis.[39, 40] We observed more differences between normal and cancer serum than we did between normal and pancreatitis. Figure 3.6 shows sections of arrays comparing normal and pancreatic cancer serum glycoproteins. In all data shown, reverse-phase chromatograms indicated similar protein amounts since peak heights and widths were comparable. Furthermore all data was normalized by peak area to nullify effects due to concentration difference. It can be seen from the bar graphs that sialic acid was more abundant in selected cancer serum glycoproteins compared to normal serum glycoproteins (Fig. 3.6A and 3.6B). Specifically, antithrombin-III precursor showed a 2.3 fold higher expression of α-2,6 linked sialic acid (as shown by SNA data in Fig. 3.6A) while all other glycans assessed did not change between normal and cancer sera. Also, a leucine-rich alpha-2-glycoprotein precursor showed a 3-fold higher expression of mannose and a 2.5-fold higher expression of fucose in addition to a 6.5-fold higher expression of α-2,6 linked sialic acid (Fig. 3.6B). Conversely some peaks showed higher mannosylation and sialylation in normal serum compared to cancer serum (Fig. 3.6C and 3.6D). An alpha-2-macroglobulin precursor had 37-fold more mannosylation and 28-fold more sialylation (Fig. 3.6C) while complement precursors showed 6-fold higher mannosylation and 5-fold higher sialylation in normal compared to cancer sera (Fig. 3.6D). Table 3.2 summarizes lectin and mass spectrometry data for all analyzed proteins from serum samples that are discussed in Fig. 3.5 and 3.6. Details about MS/MS data and peptides identified are Table 3.3.

While a particular glycosylation was more abundant in cancer versus normal for some proteins, for example α(2,6) sialylation in Antithrombin III, the trend was reversed for other proteins, such as α(2,6) sialylation in complement precursors . If all the proteins were studied together without prefractionation, such differences would not be highlighted because equal but opposite responses would cancel each other out. Our two step strategy involving fractionation prior to array production addresses this potential problem.

Although currently a proof of concept experiment, the strategy presented in this report can be used to identify changes in glycosylation in serum proteins that represent different biological states, and may serve as a novel approach to the identification of clinically useful serum biomarkers. At present, we are investigating global changes in glycosylation profiles in sera from multiple patients with various stages of pancreatic diseases to see if significant differences are evident, particularly to identify glycosylation changes unique to patients with pancreatic cancer.

3.4. Conclusion

We have presented a novel strategy that can be used to profile glycosylation patterns in complex biological samples. Unlike previous methods that can only assay known glycoproteins by using antibody microarrays or unfractionated complex mixtures that make it difficult to distinguish between the glycoproteins that may be causing a different response, our strategy starts with an enrichment step followed by a separation, allowing us to assess glycosylation patterns of individual proteins. This gives us the capability to monitor global glycosylation pattern changes as well as identify potential new proteins whose glycosylation changes are essential in biologically important states since each

glycoprotein is a unique spot on the array. The data presented here provides an example of how our approach can be used to identify different glycosylation patterns in sera from patients with different diseases of the pancreas. The study also showed that glycoprotein microarray data can provide information that reverse-phase UV data cannot. Particularly, we showed that proteins with the same retention time and similar peak heights showed an altered glycan structure distribution after normalization using integrated peak areas. The strategy can be used in large scale on biological samples to determine critical differences for diagnostics as well as large-scale glycoproteome screening.

Table 3.1: Biotinylated lectins used for glycan detection and their specificities

Biotinylated Lectin	Glycan structure detected
Concanavilin A (ConA)	α-linked mannose
Maackia Amurensis II (MAL)	sialic acid in an (α-2,3) linkage
Aleuria Aurantia (AAL)	fucose linked (α -1,6) to N-acetylglucosamine or to fucose linked (α -1,3) to N-acetyllactosamine
Sambucus Nigra (Elderberry) bark (SNA)	sialic acid attached to terminal galactose in (α-2,6), and to a lesser degree, (α-2,3), linkage
Peanut Agglutinin (PNA)	galactosyl (β-1,3) N-acetylgalactosamine

Table 3.2: Protein IDs of data shown in Fig.3.5 and 3.6 as identified by μ-LC-MS/MS with change in glycan expression based on microarray data. All data was background subtracted and normalized based on UV peak areas. N: Normal, P: Pancreatitis, C: Cancer.

Fig.	Protein ID	% Cov	MW	Con A	AAL	SNA	MAL	PNA
5B	P08603 Complement factor H precursor (H factor 1)	21	139034	2x in N	2x in N	No change	No change	No change
5C	P02763 and P19652 Alpha-1-acid glycoprotein 1 and 2 precursors	55 27	23497 23588	4x in P	4x in N	No change	No change	No change
6A	P01008 Antithrombin-III precursor (ATIII).	31	52569	No change	No change	2.3x in C	No change	No binding
6B	P02750 Leucine-rich alpha-2-glycoprotein precursor (LRG).	34	38155	3x in C	2.5x in C	6.5x in C	No binding	No binding
6C	P01023 Alpha-2-macroglobulin precursor (Alpha-2-M).	41	163175	37x in N	Only in N	28x in N	Only in N	Only in N
6D	CO3_HUMAN P01024 Complement C3 precursor	61	187046	6x in N	3x in N	5x in N	No binding	No binding

Table 3.3: Detailed results from tandem mass spectrometry experiments done on proteins discussed. Information about peptides detected, Xcorr and coverage are included.

	Protein ID	% Cov			Theoretical MW
	Sequence	MH+	Charge	XC	Ion series hit
Fig. 6A	**ANT3_HUMAN P01008 Antithrombin-III precursor (ATIII)**	**31.25**			**52569.9**
	K.FDTISEK.T	839.42	1	1.73	10/12
	K.ATEDEGSEQKIPEATNR.R	1874.87	2	2.64	19/32
	R.KELFYK.A	827.47	1	1.63	8/10
	K.SKLPGIVAEGR.D	1126.66	1	1.42	11/20
	K.LPGIVAEGR.D	911.53	1	2.45	11/16
	K.ELFYK.A	699.37	1	1.86	6/8
	R.DDLYVSDAFHK.A	1309.61	1	2.53	13/20
	K.LQPLDFK.E	860.49	1	1.50	8/12
	R.FATTFYQHLADSK.N	1528.74	2	3.23	19/24
	R.FRIEDGFSLK.E	1211.64	1	1.50	11/18
	K.GDDITM*VLILPKPEK.S	1684.92	2	2.40	16/28
	K.EQLQDMGLVDLFSPEK.S	1848.91	2	4.56	23/30
	K.NDNDNIFLSPLSISTAFAM*TK.L	2315.12	2	2.10	14/40
Fig. 6B	**A2GL_HUMAN P02750 Leucine-rich alpha-2-glycoprotein precursor (LRG).**	**33.72**			**38155.10**
	R.WLQAQK.D	773.43	1	1.65	8/10
	K.LQVLGK.D	657.43	1	1.72	8/10
	R.GPLQLER.L	812.46	1	1.79	8/12
	K.ALGHLDLSGNR.L	1152.61	1	2.11	12/20
	R.VAAGAFQGLR.Q	989.55	1	1.85	12/18
	R.YLFLNGNK.L	968.52	1	1.86	10/14
	R.TLDLGENQLETLPPDLLR.G	2037.09	2	3.62	22/34
	K.ENQLEVLEVSWLHGLK.A	1894.01	2	4.58	21/30
	HEP2_HUMAN P05546 Heparin cofactor II precursor (HC-II) (Protease inhibitor leuserpin 2) (HLS2).	**18.04**			**57035.20**
	K.VSMMQTK.G	824.40	1	1.59	9/12
	R.EVLLPK.F	698.45	1	1.89	8/10
	R.SVNDLYIQK.Q	1079.57	1	1.48	9/16
	K.GPLDQLEK.G	899.48	1	1.86	11/14
	R.MLFDK.N	653.33	1	1.62	6/8
	R.LNILNAK.F	785.49	1	1.63	9/12
	R.NFGYTLR.S	870.45	1	1.69	9/12
	K.NYNLVESLK.L	1079.57	1	2.38	12/16
	R.IAIDLFK.H	819.50	1	1.61	8/12
	R.EYYFAEAQIADFSDPAFISK.T	2312.08	2	4.19	23/38

Fig. 6C	**A2MG_HUMAN P01023 Alpha-2-macroglobulin precursor (Alpha-2-M).**	**40.77**		**163175.90**	
	K.AFTNSK.I	667.34	1	1.41	8/10
	K.TFAQAR.A	693.37	1	1.69	8/10
	K.SLNEEAVKK.D	1017.56	1	2.01	12/16
	R.TTVMVK.N	678.39	1	1.47	8/10
	K.SLNEEAVK.K	889.46	1	1.46	10/14
	K.SIYKPGQTVK.F	1120.64	1	2.13	11/18
	K.GVPIPNK.V	724.44	1	1.51	8/12
	R.TGTHGLLVK.Q	925.55	1	1.47	10/16
	R.LVHVEEPHTETVR.K	1545.80	1	2.31	14/24
	R.DLKPAIVK.V	883.56	1	1.44	8/14
	K.DNSVHWER.P	1042.47	1	1.68	9/14
	K.HYDGSYSTFGER.Y	1418.60	1	1.77	11/22
	K.LPPNVVEESAR.A	1210.64	1	2.60	13/20
	R.GEAFTLK.A	765.41	1	1.47	8/12
	K.YNILPEK.E	876.48	1	2.17	10/12
	K.AIGYLNTGYQR.Q	1255.64	1	1.83	11/20
	R.SASNMAIVDVK.M	1134.58	1	2.17	14/20
	R.QTVSWAVTPK.S	1116.61	1	2.06	12/18
	K.VDLSFSPSQSLPASHAHLR.V	2049.05	2	4.54	22/36
	R.HNVYINGITYTPVSSTNEK.D	2137.06	2	4.29	28/36
	K.LHTEAQIQEEGTVVELTGR.Q	2110.08	2	6.52	29/36
	K.NEDSLVFVQTDK.S	1394.68	1	2.57	16/22
	R.VGFYESDVMGR.G	1259.57	1	2.30	15/20
	K.GHFSISIPVK.S	1084.62	1	1.86	13/18
	R.TEVSSNHVLIYLDK.V	1617.85	2	5.29	22/26
	K.MVSGFIPLKPTVK.M	1416.83	2	2.91	17/24
	K.QQNAQGGFSSTQDTVVALHALSK.Y	2387.20	2	3.97	22/44
	K.DTVIKPLLVEPEGLEK.E	1780.01	2	4.77	21/30
	R.IAQWQSFQLEGGLK.Q	1604.84	1	2.59	14/26
	R.TEHPFTVEEFVLPK.F	1672.86	2	4.31	19/26
	R.VSVQLEASPAFLAVPVEK.E	1884.05	2	3.37	21/34
	R.LLIYAVLPTGDVIGDSAK.Y	1845.04	2	4.70	27/34
	K.ALLAYAFALAGNQDK.R	1565.83	2	3.61	22/28
Fig. 6D	**CO3_HUMAN P01024 Complement C3 precursor**	**60.61**		**187046.90**	
	R.AEDLVGK.S	731.39	1	1.62	7/12
	R.FYHPEKEDGK.L	1249.59	2	2.83	16/18
	K.GPLLNK.F	641.40	1	1.42	6/10
	R.WEDPGK.Q	731.34	1	1.50	7/10
	R.SVQLTEK.R	804.45	1	1.77	8/12
	K.SGSDEVQVGQQR.T	1289.61	2	3.77	18/22
	K.YELDK.A	667.33	1	1.73	6/8
	R.ASHLGLAR.S	824.47	1	1.95	10/14
	K.YELDK.A	667.33	1	1.41	6/8
	K.KLVLSSEK.T	903.55	1	2.28	10/14

R.ASHLGLAR.S	824.47	1	1.62	7/14
K.VTIKPAPETEK.R	1212.68	1	1.70	14/20
R.TKKQELSEAEQATR.T	1618.84	2	3.24	21/26
R.FYHPEKEDGKLNK.L	1604.81	2	3.09	17/24
R.IFTVNHK.L	858.48	1	1.45	8/12
R.NEQVEIR.A	887.46	1	1.98	8/12
K.KQELSEAEQATR.T	1389.70	2	3.72	20/22
K.SDDKVTLEER.L	1191.59	1	1.44	11/18
R.HQQTVTIPPK.S	1148.64	1	1.58	12/18
K.LVLSSEK.T	775.46	1	1.62	9/12
R.EALKLEEK.K	959.54	1	2.83	11/14
K.QELSEAEQATR.T	1261.60	1	2.25	15/20
R.EALKLEEK.K	959.54	1	2.10	10/14
R.LKGPLLNK.F	882.58	1	1.76	9/14
R.YISKYELDK.A	1158.60	1	1.95	10/16
R.SEETKENEGFTVTAEGK.G	1855.86	2	5.19	25/32
R.FLYGK.K	627.35	1	1.56	6/8
K.SGQSEDRQPVPGQQMTLK.I	1985.97	2	4.80	25/34
K.AAVYHHFISDGVR.K	1471.74	2	3.95	19/24
K.TGLQEVEVK.A	1002.55	1	1.59	10/16
R.HQQTVTIPPK.S	1148.64	1	1.73	11/18
K.KGYTQQLAFR.Q	1211.65	1	1.98	12/18
K.GQGTLSVVTM*YHAK.A	1507.76	2	3.64	17/26
K.RQGALELIK.K	1027.63	1	1.74	10/16
R.VVLVAVDK.G	842.54	1	1.48	8/14
K.LSINTHPSQKPLSITVR.T	1891.08	2	3.86	18/32
R.EGVQKEDIPPADLSDQVPDTESETR.I	2755.29	2	4.48	24/48
K.RIPIEDGSGEVVLSR.K	1626.88	2	5.09	23/28
K.EDIPPADLSDQVPDTESETR.I	2214.01	2	5.62	26/38
K.GVFVLNK.K	776.47	1	2.18	8/12
K.GQGTLSVVTMYHAK.A	1491.76	2	3.88	18/26
K.GYTQQLAFR.Q	1083.56	1	1.98	12/16
R.TFISPIK.C	805.48	1	1.82	10/12
R.IPIEDGSGEVVLSR.K	1470.78	2	3.77	23/26
R.QGALELIK.K	871.53	1	1.64	11/14
R.EGVQKEDIPPADLSDQVPDTESETR.I	2755.29	2	4.05	23/48
R.YYTYLIM*NK.G	1224.60	1	1.87	10/16
K.FYYIYNEK.G	1139.54	1	2.33	10/14
K.TIYTPGSTVLYR.I	1370.73	2	2.86	18/22
R.TVM*VNIENPEGIPVK.Q	1655.87	2	4.46	21/28
K.EDIPPADLSDQVPDTESETR.I	2214.01	2	4.40	23/38
R.LESEETM*VLEAHDAQGDVPVTVTVHDFPG K.K	3266.55	2	2.59	13/58
R.IHWESASLLR.S	1211.65	1	2.13	14/18
R.TVMVNIENPEGIPVK.Q	1639.87	1	1.94	13/28
R.VPVAVQGEDTVQSLTQGDGVAK.L	2198.13	2	5.45	29/42
R.ILLQGTPVAQM*TEDAVDAER.L	2173.08	2	5.12	26/38
K.QKPDGVFQEDAPVIHQEM*IGGLR.N	2580.29	2	2.69	13/44
R.QPSSAFAAFVK.R	1152.61	1	1.75	14/20
R.EVVADSVWVDVK.D	1345.70	2	4.06	18/22

K.VHQYFNVELIQPGAVK.V	1841.99	2	5.01	21/30
R.NTLIIYLDK.V	1092.63	1	2.01	12/16
R.SGIPIVTSPYQIHFTK.T	1787.97	1	1.83	10/30
K.WLILEK.Q	801.49	1	1.43	8/10
R.LVAYYTLIGASGQR.E	1511.82	2	4.35	22/26
R.VPVAVQGEDTVQSLTQGDGVAK.L	2198.13	2	5.40	25/42
K.AGDFLEANYMNLQR.S	1641.77	2	4.18	17/26
R.SNLDEDIIAEENIVSR.S	1816.89	1	3.02	16/30
K.LMNIFLK.D	878.52	1	1.94	9/12
R.AYYENSPQQVFSTEFEVK.E	2166.00	2	5.13	23/34
K.DYAGVFSDAGLTFTSSSGQQTAQR.A	2494.15	2	4.81	25/46
K.KVEGTAFVIFGIQDGEQR.I	1994.03	2	5.91	29/34
R.SNLDEDIIAEENIVSR.S	1816.89	2	4.15	24/30
R.TELRPGETLNVNFLLR.M	1872.03	2	2.36	15/30
R.TMQALPYSTVGNSNNYLHLSVLR.T	2578.31	2	3.83	24/44
K.DFDFVPPVVR.W	1190.62	1	1.66	10/18
R.APSTWLTAYVVK.V	1335.73	2	3.58	18/22
R.AYYENSPQQVFSTEFEVK.E	2166.00	2	4.09	17/34
K.SSLSVPYVIVPLK.T	1401.84	2	2.90	19/24
K.DYAGVFSDAGLTFTSSSGQQTAQR.A	2494.15	2	6.54	28/46
K.SLYVSATVILHSGSDMVQAER.S	2263.14	2	2.83	20/40
R.VPVAVQGEDTVQSLTQGDGVAK.L	2198.13	2	4.81	23/42
K.DYAGVFSDAGLTFTSSSGQQTAQR.A	2494.15	2	5.52	30/46
R.IHWESASLLR.S	1211.65	1	1.40	11/18
R.ILLQGTPVAQMTEDAVDAER.L	2157.09	2	5.28	29/38
R.SNLDEDIIAEENIVSR.S	1816.89	2	2.47	19/30
R.SEFPESWLWNVEDLKEPPK.N	2330.13	2	3.89	22/36
K.EYVLPSFEVIVEPTEK.F	1878.97	2	5.16	25/30
K.VFLDCCNYITELR.R	1588.75	1	1.48	10/24
K.DYAGVFSDAGLTFTSSSGQQTAQR.A	2494.15	2	5.44	25/46
R.SEFPESWLWNVEDLKEPPK.N	2330.13	2	3.61	22/36
K.DSITTWEILAVSMSDKK.G	1923.97	2	4.22	23/32
R.VPVAVQGEDTVQSLTQGDGVAK.L	2198.13	2	4.48	20/42
R.VPVAVQGEDTVQSLTQGDGVAK.L	2198.13	2	4.05	21/42
K.YFKPGM*PFDLMVFVTNPDGSPAYR.V	2765.31	2	3.46	17/46
R.YYGGGYGSTQATFMVFQALAQYQK.D	2679.26	2	5.09	25/46
K.DYAGVFSDAGLTFTSSSGQQTAQR.A	2494.15	2	4.62	22/46
K.VQLSNDFDEYIMAIEQTIK.S	2257.11	2	4.00	23/36
K.SLYVSATVILHSGSDMVQAER.S	2263.14	2	2.61	16/40
K.VQLSNDFDEYIMAIEQTIK.S	2257.11	2	5.22	23/36
K.QDSLSSQNQLGVLPLSWDIPELVNMGQWK.I	3282.65	2	2.87	18/56
K.QLYNVEATSYALLALLQLK.D	2151.21	2	6.07	27/36
R.SNLDEDIIAEENIVSR.S	1816.89	2	4.18	20/30
R.VPVAVQGEDTVQSLTQGDGVAK.L	2198.13	2	4.00	22/42
R.ILLQGTPVAQMTEDAVDAER.L	2157.09	2	2.51	21/38
R.SNLDEDIIAEENIVSR.S	1816.89	2	3.09	16/30
CO4_HUMAN P01028 Complement C4 precursor	**20.30**		**192651.50**	

	K.SHKPLNMGK.V	1011.54	1	2.20	9/16
	R.VEASISK.A	733.41	1	1.59	8/12
	R.NVNFQK.A	749.39	1	1.99	8/10
	R.LFETK.I	637.36	1	1.71	6/8
	K.SHALQLNNR.Q	1052.56	1	1.96	11/16
	K.LGQYASPTAK.R	1035.55	1	1.72	11/18
	R.GLQDEDGYR.M	1052.46	1	1.66	12/16
	K.VLQIEK.E	729.45	1	2.06	8/10
	K.DHAVDLIQK.G	1038.56	1	2.33	12/16
	R.LPMSVR.R	702.40	1	1.47	7/10
	K.ANSFLGEK.A	865.44	1	1.47	8/14
	R.VFALDQK.M	820.46	1	1.84	10/12
	K.LELSVDGAK.Q	931.51	1	1.63	11/16
	R.NFLVR.A	648.38	1	1.46	6/8
	K.ITQVLHFTK.D	1086.63	1	1.72	12/16
	R.VEYGFQVK.V	969.50	1	1.68	10/14
	K.VDFTLSSER.D	1053.52	1	1.40	9/16
	R.TYNVLDMK.N	983.49	1	2.21	10/14
	R.VGDTLNLNLR.A	1114.62	1	1.55	8/18
	R.LTVAAPPSGGPGFLSIERPDSRPPR.V	2574.38	2	2.31	14/48
	R.GSFEFPVGDAVSK.V	1339.65	2	3.84	17/24
	R.GPEVQLVAHSPWLK.D	1560.85	2	4.43	19/26
	K.YVLPNFEVK.I	1108.60	1	2.05	12/16
	R.TLEIPGNSDPNMIPDGDFNSYVR.V	2551.18	2	2.99	20/44
	R.ALEILQEEDLIDEDDIPVR.S	2225.12	2	4.63	24/36
	R.VTASDPLDTLGSEGALSPGGVASLLR.L	2483.30	2	2.75	17/50
	K.EVYMPSSIFQDDFVIPDISEPGTWK.I	2900.37	2	3.28	18/48
Fig. 5B	**CFAH_HUMAN P08603 Complement factor H precursor (H factor 1)**	**21.12**		**139034.80**	
	K.IVSSAMEPDR.E	1104.54	1	1.71	13/18
	R.EYHFGQAVR.F	1106.54	1	2.85	12/16
	R.NGFYPATR.G	925.45	1	1.84	9/14
	K.SPDVINGSPISQK.I	1341.70	1	2.59	15/24
	K.IDVHLVPDR.K	1063.59	1	2.35	12/16
	K.IVSSAMEPDREYHFGQAVR.F	2192.06	2	2.57	17/36
	R.RPYFPVAVGK.Y	1133.65	1	2.00	11/18
	R.KGEWVALNPLRK.C	1410.82	2	3.33	17/22
	R.EIM*ENYNIALR.W	1381.68	2	2.98	16/20
	R.TKNDFTWFK.L	1186.59	1	1.56	10/16
	K.SSNLIILEEHLK.N	1395.78	1	2.64	13/22
	K.CYFPYLENGYNQNHGR.K	1974.86	2	2.48	15/30
	R.NTEILTGSWSDQTYPEGTQAIYK.C	2602.23	2	5.04	25/44
	K.GEWVALNPLR.K	1154.63	1	2.46	13/18
	K.NDFTWFK.L	957.45	1	1.91	9/12
	K.SSIDIENGFISESQYTYALK.E	2265.09	2	5.43	27/38
	K.SPPEISHGVVAHMSDSYQYGEEVTYK.C	2910.33	2	2.84	16/50
	K.SSNLIILEEHLK.N	1395.78	2	2.06	11/22

	K.SIDVACHPGYALPK.A	1470.74	2	1.11	13/26
	K.SSIDIENGFISESQYTYALK.E	2265.09	2	2.81	14/38
	K.SSIDIENGFISESQYTYALK.E	2265.09	2	4.25	22/38
	R.NTEILTGSWSDQTYPEGTQAIYK.C	2602.23	2	3.66	14/44
	R.NTEILTGSWSDQTYPEGTQAIYK.C	2602.23	2	3.10	19/44
	K.SSIDIENGFISESQYTYALK.E	2265.09	2	4.57	21/38
	R.NTEILTGSWSDQTYPEGTQAIYK.C	2602.23	2	2.37	17/44
	K.SSIDIENGFISESQYTYALK.E	2265.09	2	4.97	22/38
Fig. 5C	**A1AG1_HUMAN P02763 Alpha-1-acid glycoprotein 1 precursor (AGP1)**	**54.73**		**23497.80**	
	K.DKCEPLEK.Q	961.47	1	2.14	9/14
	K.SDVVYTDWKK.D	1240.62	2	3.03	15/18
	K.NWGLSVYADKPETTK.E	1708.85	2	5.16	22/28
	K.TEDTIFLR.E	994.52	1	1.58	9/14
	R.YVGGQEHFAHLLILR.D	1752.95	2	5.20	21/28
	K.TYMLAFDVNDEK.N	1445.66	1	3.21	15/22
	K.TYMLAFDVNDEKNWGLSVYADKPETTK.E	3135.50	2	4.38	20/52
	K.EQLGEFYEALDCLR.I	1685.78	2	3.60	19/26
	R.YVGGQEHFAHLLILR.D	1752.95	2	4.71	21/28
	K.SDVVYTDWK.K	1112.53	1	2.53	11/16
	K.WFYIASAFR.N	1160.59	1	2.04	12/16
	A1AG2_HUMAN P19652 Alpha-1-acid glycoprotein 2 precursor (AGP2)	**27.36**		**23588.60**	
	R.SDVMYTDWKK.D	1272.59	1	2.36	11/18
	R.SDVM*YTDWKK.D	1288.59	2	2.90	13/18
	R.SDVMYTDWKK.D	1272.59	2	3.31	16/18
	R.SDVM*YTDWK.K	1160.49	1	1.57	9/16
	R.EHVAHLLFLRDTK.T	1578.88	2	3.10	19/24
	R.YEGGREHVAHLLFLR.D	1796.96	2	3.15	17/28
	R.EHVAHLLFLR.D	1234.71	1	2.60	13/18
	R.SDVMYTDWK.K	1144.50	1	2.68	11/16
	K.TLM*FGSYLDDEK.N	1434.65	2	4.05	19/22
	K.TLMFGSYLDDEK.N	1418.65	1	1.68	9/22
	K.TLMFGSYLDDEK.N	1418.65	2	3.37	19/22

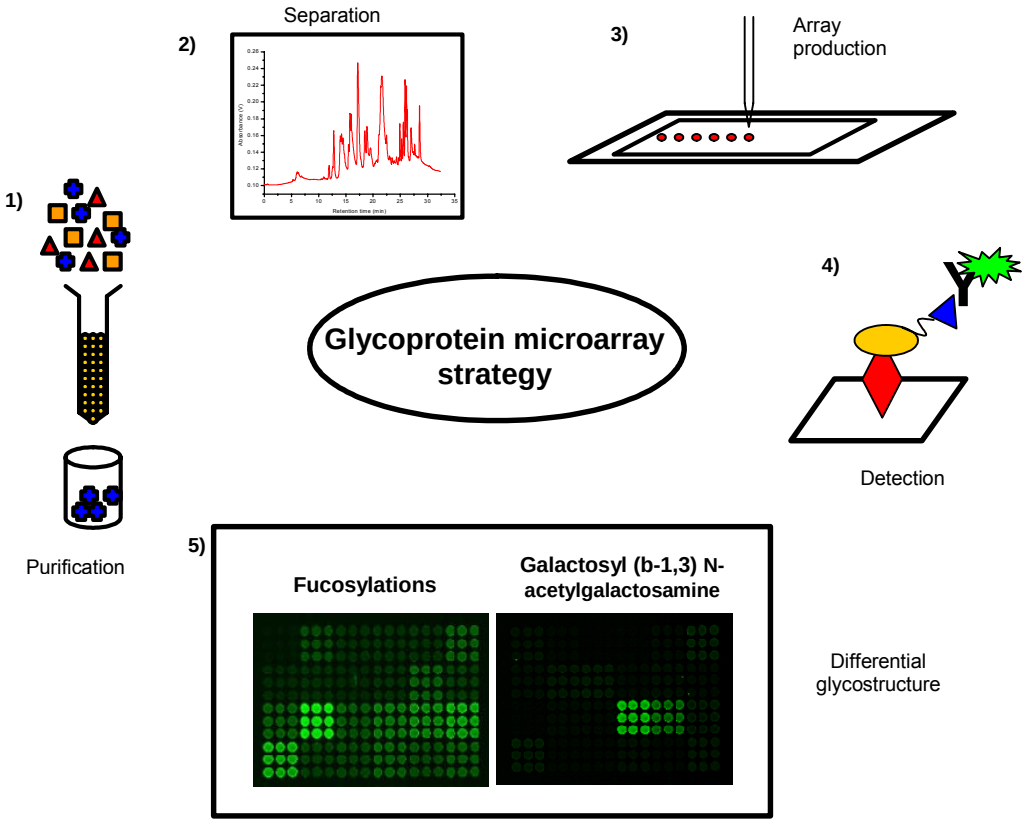


Figure 3.1: Experimental strategy for studying serum glycoproteins. 1) Lectin purification 2) Non-porous reverse phase HPLC separation and fraction collection 3) Microarray production 4) Lectin detection using biotin-streptavidin-Alexafluor555 detection 4) Image acquisition and analysis

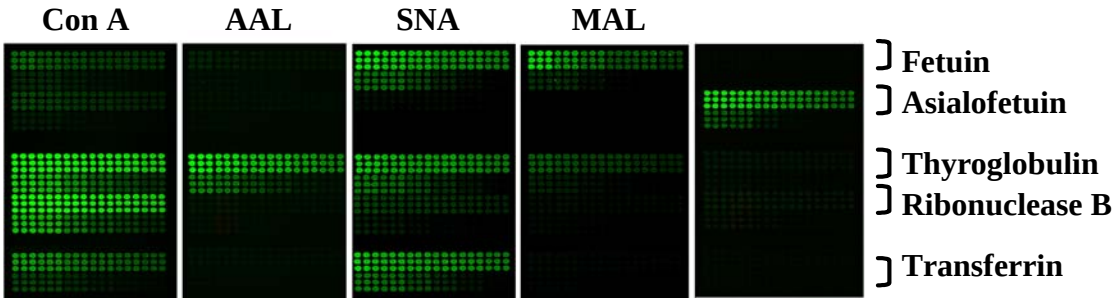


Figure 3.2: Scanned images of printed standard glycoproteins probed with different lectins. Each block bracketed on the right represents a dilution series of indicated standard from 2mg/mL to 0.025mg/mL. Each dilution has been printed as 9 replicates in a 3x3 block.

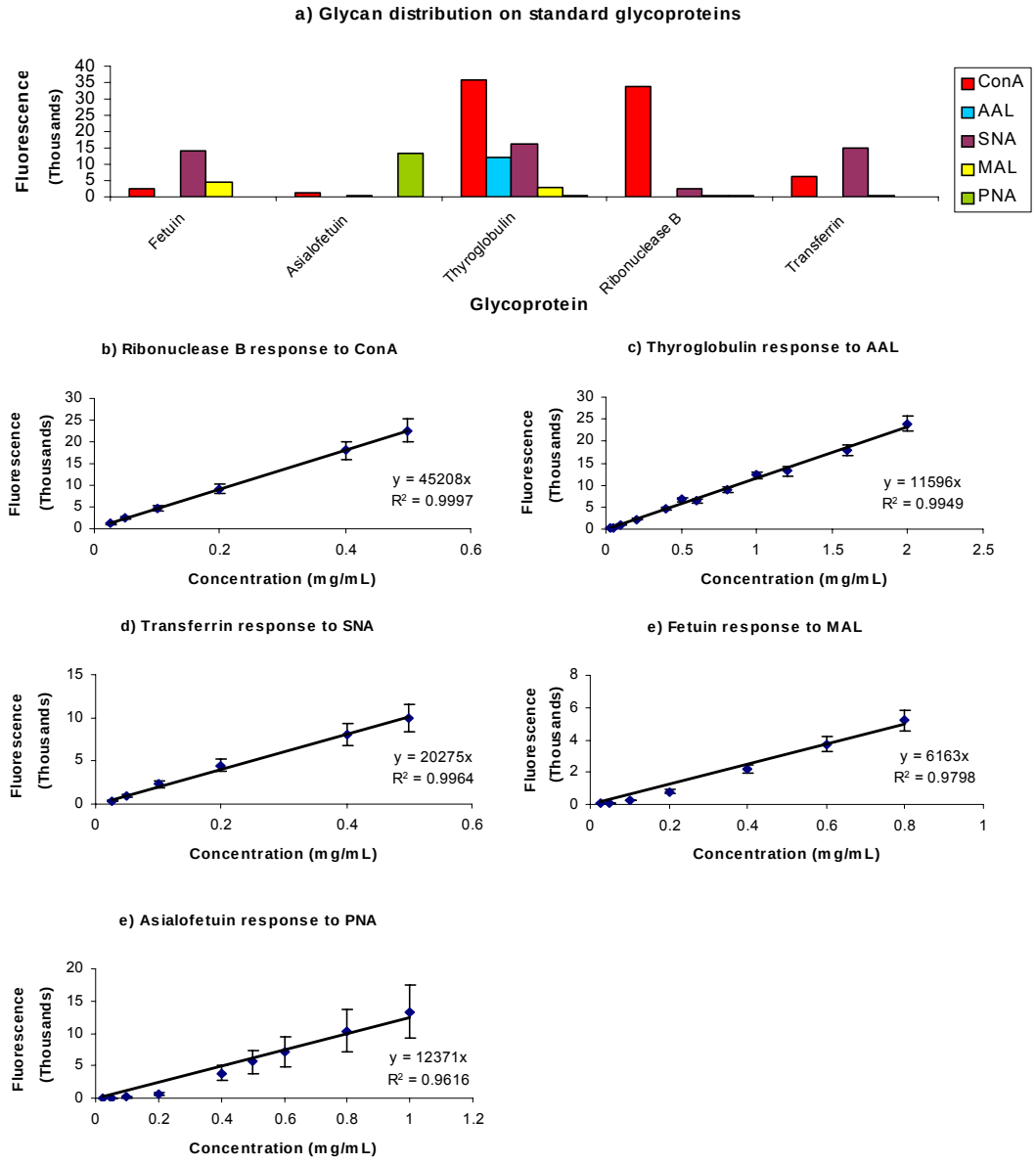


Figure 3.3: Linearity of response in standards a) Glycan distribution on standards printed at 1mg/mL concentration. Standard curve of b) Ribonuclease B in response to ConA c) Thyroglobulin in response to AAL d) Transferrin in response to SNA e) Fetuin in response to MAL f) Asialofetuin in response to PNA using lectin concentration of 5ug/mL

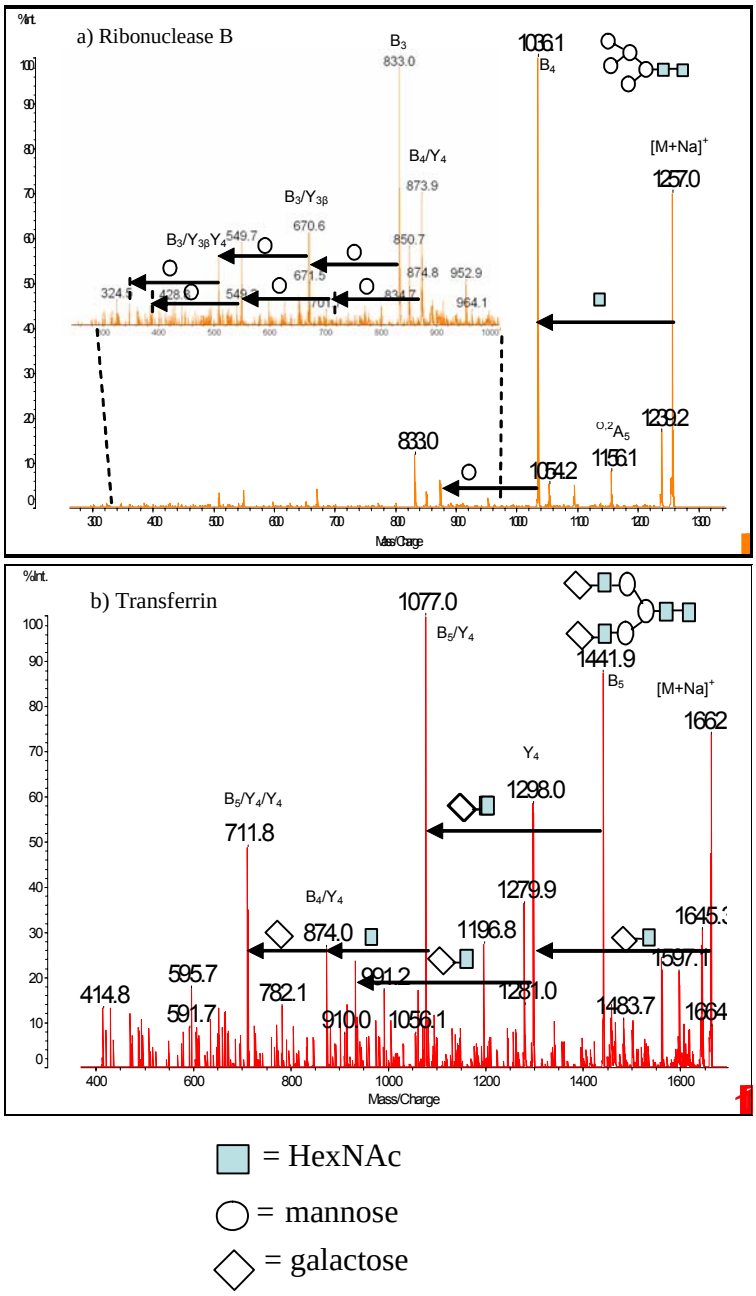


Figure 3.4: Tandem mass spectra of dominant glycan structure in a) Ribonuclease B (precursor ion m/z 1257) b) Transferrin (precursor ion m/z 1663)

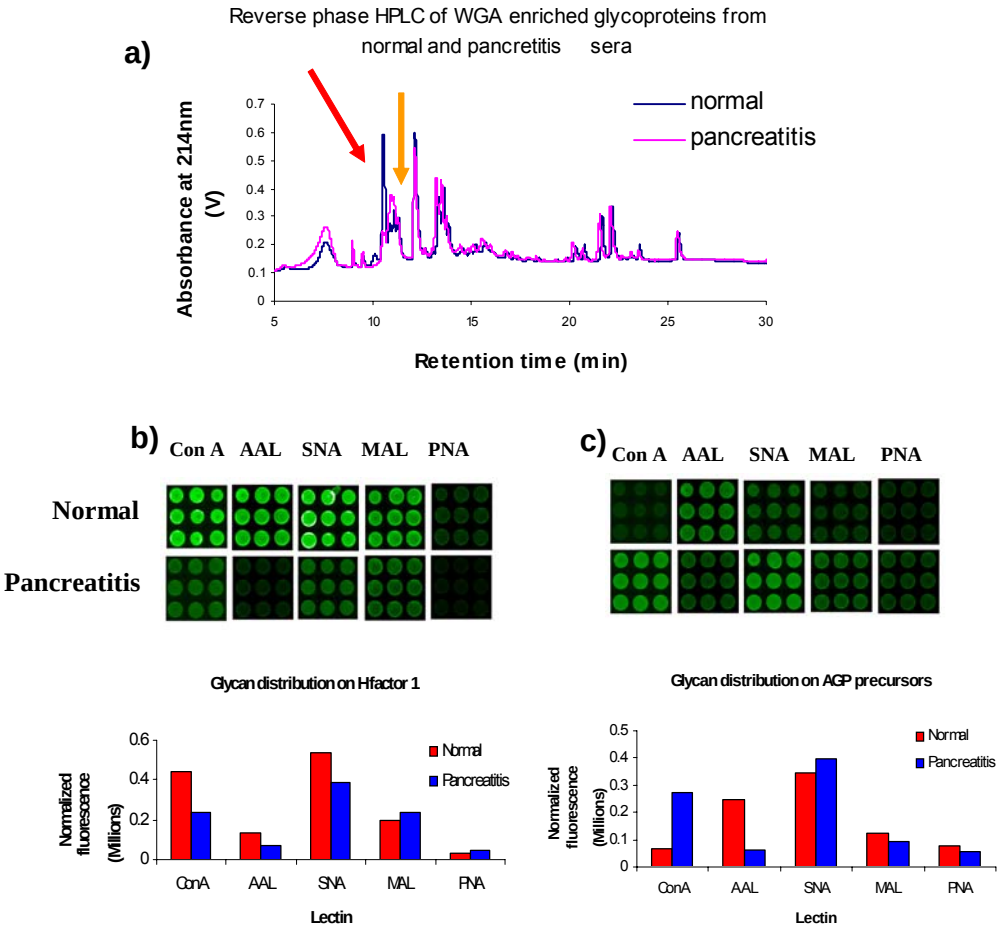


Figure 3.5: Identifying differences in glycosylation from sera of different biological states. **a)** Reverse phase chromatogram of enriched glycopoteins from normal and pancreatitis sera with differences highlighted. Bar graph showing integrated fluorescence values of spots shown in array images after background subtraction and normalization based on UV peak area for peak shown with **b)** red arrow, **c)** orange arrow.

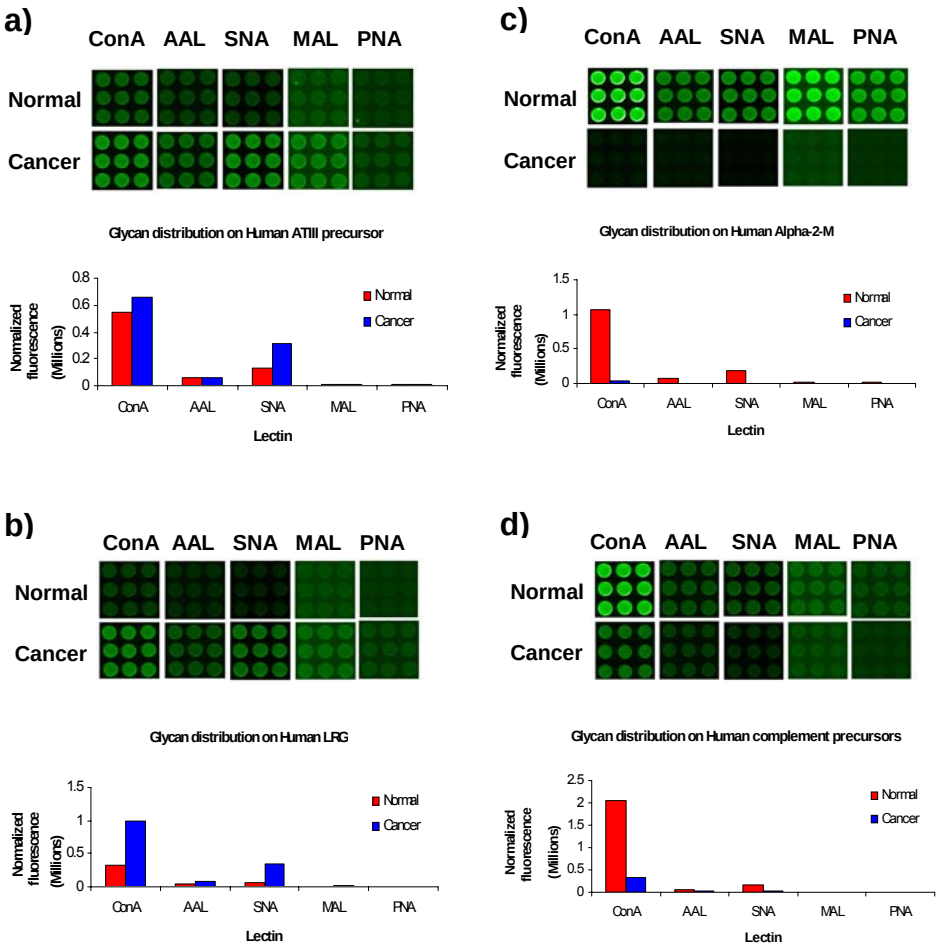


Figure 3.6: Comparison of differential glycosylation patterns in normal vs. cancer serum. All comparisons shown below had approximately the same peak area between cancer and normal sera but glycosylation patterns were different. Each illustration shows sections of microarray images of a protein's binding to the lectins indicated. Bar graphs show integrated fluorescence values of spots shown in the array images after background subtraction and normalization based on UV peak area.

3.5. References

[1] Rudd, P. M., Elliott, T., Cresswell, P., Wilson, I. A., Dwek, R. A., *Science* 2001, *291*, 2370-2376.

[2] Kobata, A., Amano, J., *Immunol Cell Biol* 2005, *83*, 429-439.

[3] Dube, D. H., Bertozzi, C. R., *Nat Rev Drug Discov* 2005, *4*, 477-488.

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

[5] Semmes, O. J., Malik, G., Ward, M., *J Cell Biochem* 2006.

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

[7] Durand, G., Seta, N., *Clin Chem* 2000, *46*, 795-805.

[8] Nimrichter, L., Gargir, A., Gortler, M., Altstock, R. T., *et al.*, *Glycobiology* 2004, *14*, 197-203.

[9] Drickamer, K., Taylor, M. E., *Genome Biol* 2002, *3*, REVIEWS1034.

[10] Adams, E. W., Ratner, D. M., Bokesch, H. R., McMahon, J. B., *et al.*, *Chem Biol* 2004, *11*, 875-881.

[11] Wang, D., Lu, J., *Physiol Genomics* 2004, *18*, 245-248.

[12] Templin, M. F., Stoll, D., Schwenk, J. M., Potz, O., *et al.*, *Proteomics* 2003, *3*, 2155-2166.

[13] Pal, M., Moffa, A., Sreekumar, A., Ethier, S. P., *et al.*, *Anal Chem* 2006, *78*, 702-710.

[14] Yan, F., Sreekumar, A., Laxman, B., Chinnaiyan, A. M., *et al.*, *Proteomics* 2003, *3*, 1228-1235.

[15] Orchekowski, R., Hamelinck, D., Li, L., Gliwa, E., *et al.*, *Cancer Res* 2005, *65*, 11193-11202.

[16] Feizi, T., Fazio, F., Chai, W., Wong, C. H., *Curr Opin Struct Biol* 2003, *13*, 637-645.

[17] Fukui, S., Feizi, T., Galustian, C., Lawson, A. M., Chai, W., *Nat Biotechnol* 2002, *20*, 1011-1017.

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

[19] Schwarz, M., Spector, L., Gargir, A., Shtevi, A., *et al.*, *Glycobiology* 2003, *13*, 749-754.

[20] Wang, Y., Wu, S. L., Hancock, W. S., *Glycobiology* 2006.

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

[22] Qiu, R., Regnier, F. E., *Anal Chem* 2005, *77*, 2802-2809.

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

[24] Kuno, A., Uchiyama, N., Koseki-Kuno, S., Ebe, Y., *et al.*, *Nat Methods* 2005, *2*, 851-856.

[25] Pilobello, K. T., Krishnamoorthy, L., Slawek, D., Mahal, L. K., *Chembiochem* 2005, *6*, 985-989.

[26] Shibuya, N., Goldstein, I. J., Broekaert, W. F., Nsimba-Lubaki, M., *et al.*, *Arch Biochem Biophys* 1987, *254*, 1-8.

[27] Wang, W. C., Cummings, R. D., *J Biol Chem* 1988, *263*, 4576-4585.

[28] Townsend, R. R., Hardy, M. R., Cumming, D. A., Carver, J. P., Bendiak, B., *Anal Biochem* 1989, *182*, 1-8.

[29] Charlwood, J., Birrell, H., Organ, A., Camilleri, P., *Rapid Commun Mass Spectrom* 1999, *13*, 716-723.

[30] Takekawa, H., Ina, C., Sato, R., Toma, K., Ogawa, H., *J Biol Chem* 2006, *281*, 8528-8538.

[31] Joao, H. C., Dwek, R. A., *Eur J Biochem* 1993, *218*, 239-244.

[32] Satomi, Y., Shimonishi, Y., Hase, T., Takao, T., *Rapid Commun Mass Spectrom* 2004, *18*, 2983-2988.

[33] Itai, S., Arii, S., Tobe, R., Kitahara, A., *et al.*, *Cancer* 1988, *61*, 775-787.

[34] Harvey, D. J., Martin, R. L., Jackson, K. A., Sutton, C. W., *Rapid Commun Mass Spectrom* 2004, *18*, 2997-3007.

[35] Powell, A. K., Harvey, D. J., *Rapid Commun Mass Spectrom* 1996, *10*, 1027-1032.

[36] Juhasz, P., Costello, C. E., *J. Am. Soc. Mass Spectrom* 1992, *3*, 785-779.

[37] Ryden, I., Pahlsson, P., Lindgren, S., *Clin Chem* 2002, *48*, 2195-2201.

[38] Uchikov, P. A., Sirakova, I. P., Murdjeva, M. A., Uchikov, A. P., *Folia Med (Plovdiv)* 2000, *42*, 23-30.

[39] Valerio, A., Basso, D., Mazza, S., Baldo, G., *et al.*, *Rapid Commun Mass Spectrom* 2001, *15*, 2420-2425.

[40] Plebani, M., Basso, D., Panozzo, M. P., Fogar, P., *et al.*, *Int J Biol Markers* 1995, *10*, 189-199.

Chapter 4

Using unique lectin binding patterns of glycoprotein microarrays as a tool for classifying normal, chronic pancreatitis and pancreatic cancer sera

4.1. Introduction

Pancreatic cancer, is the fourth most frequent cause of cancer-related death in the USA. It is generally incurable by available treatments and has a 5-year survival rate of <4%.[1] The biologically aggressive nature of this disease, with rapid metastasis, combined with the late clinical presentation of the malignancy has resulted in poor prognosis. Existing markers for pancreatic cancer are not reliable for early diagnosis, distinguishing between pancreatic cancer and chronic pancreatitis, and for the efficient targeting of therapeutics.[2, 3] CA19-9 has been tested for its utility as an early detection marker in pancreatic cancer.[2-4] However, the sensitivity and specificity of this marker is not high, and serum levels are also significantly increased in inflammatory diseases of the pancreas and biliary tract. Therefore, CA 19-9 is not useful for early diagnosis, mass screening, or for distinguishing between pancreatic cancer and chronic pancreatitis. Protein-based serum markers for the early detection of cancer have become an area of increased interest. The glyco-proteome is one of the major sub-proteomes of human

serum. Both N-linked and O-linked glycan variants of glycoproteins on the cell surface and in plasma have been demonstrated to correlate with the progression of cancer and other diseases.[5-9] Changes in glycosylation patterns have been associated with prostate cancer[10, 11], colorectal cancer [12, 13] and breast cancer.[14] The glycosylation of prostate-specific antigen (PSA) secreted by the prostate tumor cell line LNCaP differs significantly from that found in seminal plasma (normal).[10] As glycan differences can distinguish PSA from normal and tumor origins, these differences may have utility for early diagnosis of prostate cancer. Glycosylation changes in a tumor-secreted protein could reflect fundamental changes in enzyme levels (or enzyme activities) involved in the glycosylation pathway. The ability to efficiently profile protein glycosylation variation may ultimately lead to the identification of disease-associated glycan alterations and new diagnostic markers in pancreatic cancer and in other types of cancer.

Protein microarrays are becoming slowly becoming a method of choice in high throughput proteomic analysis due to their ability to screen large numbers of arrayed samples for a property/moiety of interest.[15-17] Current research in this area has focused on a variety of applications ranging from functional analysis to diagnostic-type approaches. Functional approaches typically focus on studying interactions of proteins with a variety of other molecules such as other proteins, lipids, drugs and DNA.[18, 19] Diagnostic applications involve immobilization of antibodies on high density arrays which can be probed with biological fluids or cell lysates to monitor antigen-antibody interactions.[20] In addition, protein microarrays arrayed with naturally produced proteins have been developed to assist in finding novel disease-associated proteins [21, 22] using multi-dimensional liquid-based separation of a proteome, followed by the

arraying of all proteins found in the individual fractions. The resulting microarrays can subsequently be probed with a variety of detection agents, including lectins for glycoprotein detection.

Lectins specifically and reversibly bind glycans with different structural moieties and, thus, have utility in screening glycosylation differences between various samples. Lectin glycoarrays can be used for the rapid profiling of glycan expression patterns of various glycoproteins as illustrated in chapter 3.[23] We have utilized glycoarrays to discern differences in the glycosylation structural patterns of serum glycoproteins specific for pancreatic cancer and chronic pancreatitis. Following immunodepletion to remove high abundance proteins from serum (and to facilitate our ability to detect low abundance glycoproteins), the remaining N-linked glycoproteins were enriched using a general multi-lectin column. These enriched glycoproteins were then separated using non-porous silica reverse phase high performance liquid chromatography (NPS-RP-HPLC). The resolved glycoproteins were then arrayed on nitrocellulose-coated slides and probed with a variety of lectins to screen the glycosylation structures of the serum glycoproteins. The glycoprotein-lectin interaction was assessed using a biotin-streptavidin system that had low femtomole limits of detection.

All data was subjected to bioinformatics analysis to handle and display efficiently the large datasets generated. To compare the overall pattern of glycan expression and not the protein abundance in the samples, each sample was normalized and aligned with the corresponding UV peak area from the RP-HPLC chromatograms. A correlation matrix was obtained by calculating the Pearson correlations among the samples. These correlation matrices were then visualized, using either Principal Components Analysis

(PCA) or Hierarchical Clustering (HC) techniques, allowing multivariate relationships to be explored in order to highlight relationships present in the sample sets. Quantitative measurements were also facilitated, since normalization based on UV peak areas eliminated any concentration dependent variability that existed in the fractionated glycoproteins. Differential glycan expression was calculated by interrogating Z-value information. The individual glycoproteins with altered glycan structures were then identified by mass spectrometry. These glycan structural alterations may have utility for the early detection of pancreatic cancer and for the differential diagnosis of pancreatic cancer and chronic pancreatitis.

4.2. Experimental Section

Serum Samples: Serum was obtained at the time of diagnosis following informed consent using IRB-approved guidelines. Sera were obtained from 6 patients with a confirmed diagnosis of pancreatic adenocarcinoma in the Multidisciplinary Pancreatic Tumor Clinic at The University of Michigan Hospital. These sera were randomly selected from a clinic population that sees, on average, at the time of initial diagnosis, 15% of pancreatic adenocarcinoma patients presenting with early stage (i.e., stage 1/2) disease and 85% presenting with advanced stage (i.e., stage 3/4). Inclusion criteria for the study included patients with a confirmed diagnosis of pancreatic cancer, the ability to provide written, informed consent, and the ability to provide 40 ml of blood. Exclusion criteria included inability to provide informed consent, patient's actively undergoing chemotherapy or radiation therapy for pancreatic cancer, and patients with other malignancies diagnosed or treated within the last 5 years. Sera were also obtained from 8

patients with chronic pancreatitis who were seen in the Gastroenterology Clinic at University of Michigan Medical Center, and from 10 control healthy individuals collected at University of Michigan under the auspices of the Early Detection Research Network (EDRN). The mean age of the tumor group was 65.4 years (range 54-74 years) and from the chronic pancreatitis group was 54 years (range 45-65). The sera from the normal subject group was age and sex-matched to the tumor group. All of the chronic pancreatitis sera were collected in an elective setting in the clinic in the absence of an acute flare. All sera were processed using identical procedures. The samples were permitted to sit at room temperature for a minimum of 30 minutes (and a maximum of 60 minutes) to allow the clot to form in the red top tubes, and then centrifuged at 1,300 x g at 4^{o}C for 20 minutes. The serum was removed, transferred to a polypropylene, capped tube in 1 ml aliquots, and frozen. The frozen samples were stored at -70^{o}C until assayed. All serum samples were labeled with a unique identifier to protect the confidentiality of the patient. The handling of all serum samples was similar in that none of the samples were thawed more than twice before analysis in order to minimize protein degradation and precipitation.

Immunodepletion of high abundance proteins: 125 μL of each serum sample was depleted using the ProteomeLab^{TM} IgY-12 proteome partitioning kit (Beckman Coulter, Fullerton, CA), following centrifugation using a 0.45 μm spin filter for 1 min at 9200 x g, according to manufacturer's protocols. This column facilitates removal of albumin, IgG, α1-antitrypsin, IgA, IgM, transferrin, haptoglobin, α1-acid glycoprotein, α2-macroglobin, apolipoprotein A-I, apolipoprotein A-II and fibrinogen in a single step. The final volume of each serum sample following immunodepletion was concentrated to 500 μl using 15

ml 10kDa Amicon filters (Millipore, Billerica, MA). Protein assays were carried out in a 250 μL transparent 96 well plate (Fisher, Barrington, IL) according to the Bradford assay.

Lectin affinity glycoprotein extraction: Agarose-bound Wheat Germ Agglutinin (WGA) and agarose-bound Concanavalin A (ConA) were purchased from Vector Laboratories (Burlingame, CA, USA). 350 μl agarose-bound WGA and 250 μl agarose-bound ConA were packed into disposable screw end-cap spin column with filters at both ends. The binding and elution process has been described elsewhere.[6] The binding buffer contained 20 mM Tris, 1 mM $MnCl_2$, 1 mM $CaCl_2$, 0.15 M NaCl, pH 7.4. The immunodepleted serum proteins were resuspended in binding buffer, and then passed through the lectin affinity column. The captured serum glycoproteins were released with 250 μL elution buffer (0.3 M *N*-acetyl-glucosamine and 0.3 M Methyl-α-D-mannopyroside in 20 mM Tris and 0.5 M NaCl, pH 7.0). This step was repeated twice and the eluted fractions were pooled.

RP-HPLC separation of lectin-bound glycoproteins: The lectin-enriched glycoprotein fraction was concentrated to ~100 μl with a 10k MW centrifugal filter (Millipore) and re-diluted with de-ionized water. Approximately 30 μg of protein sample was loaded in 800 μl water onto a nonporous silica reverse phase high-performance liquid chromatography (NPS-RP-HPLC) column (ODSII (4.6x33 mm) column (Eprogen, Inc., Darien, IL) packed with 1.5 μm non-porous silica) for separation. The reverse-phase separation was performed at 0.5 mL/min and monitored at 214 nm using a Beckman 166 Model UV detector (Beckman-Coulter). Proteins eluting from the column were collected by an automated fraction collector (Model SC 100; Beckman-Coulter), controlled by an in-house designed DOS-based software program. The reversed phase column was heated

to 60^{o}C by a column heater (Jones Chromatography, Model 7971). Both mobile phase A (water) and B (ACN) contained 0.1% v/v TFA. The gradient profile used was as follows: 5% to 15% B in 1 min, 15% to 25% B in 2 min, 25% to 30% B in 3 min, 30% to 41% B in 15 min, 41% to 47% B in 4 min, 47% to 67% B in 5 min and 67% to 100% B in 2 min.

Glycoprotein microarrays: Purified and separated glycoproteins were printed on nitrocellulose slides (Whatman, Keene, NH) using a non-contact printer, Nanoplotter 2.0 (GeSIM, Germany). Prior to printing, the proteins were dried down in a 96-well plate and resuspended in 15 μL of printing buffer with stirring overnight at 4^{o}C. The printing buffer contained 65 mM Tris-HCl, 1% SDS, 5% dithiothreitol (DTT) and 1% glycerol. Each spotting event resulted in approximately 500 pL of sample being deposited by a piezoelectric mechanism. The event was programmed to occur 5 times per spot to ensure that approximately 2.5 nL were being spotted per sample. The resulting spots were approximately 450 μm in diameter, with the spacing between spots being maintained at 600 μm. After printing, the slides were allowed to dry for 24 hrs. Blocking was achieved by incubation with 1% Bovine serum albumin (BSA) and 0.1% Tween-20 in 1X phosphate buffered saline (PBS) overnight. Blocked slides were probed with biotinylated lectin in a solution of PBS-T (0.1% Tween 20 in 1X PBS). The lectins used in the study were biotinylated Peanut Agglutinin (PNA), *Sambucus Nigra* bark lectin (SNA), *Aleuria Aurentia* (AAL), Concanavalin A (ConA) and *Maackia Amurensis* lectin II (MAL), all purchased from Vector Laboratories (Burlingame, CA, USA). The working concentration of all lectins was 5 μg/mL, with the exception of SNA (10 μg/mL, as per manufacturer's protocols). After primary incubation, all slides were washed with PBS-T 5 times for 5 min each. Detection was achieved using a streptavidin-AlexaFluor555 conjugate

(Invitrogen, Carlsbad, CA) at 1 μg/mL in PBS containing 0.5% BSA and 0.1% Tween-20. The slides were washed 5 times for 5 min each in PBS-T, and then completely dried by centrifugation. The dried slides were scanned using an Axon 4000A scanner in the green channel. GenePix Pro 6.0 software (Molecular Devices, Sunnyvale, CA) was used for data acquisition and analysis.

Data analysis and clustering: All the microarray spot intensities were normalized with corresponding UV peak area. For data visualization, average linkage hierarchical clustering (HC) and principal component analysis (PCA) were used to provide graphical representations of the relationships among the samples. In these unsupervised approaches, 48 serum samples (10 normal, 8 chronic pancreatitis and 6 pancreatic cancer, all processed in duplicate) and the replicate averages of the 24 distinct biological specimens were placed either in a hierarchical relationship (HC) or as points in a 2-dimensional scatterplot (PCA) based on similarities in normalized glycoform abundances. For differential abundance analysis, Z-statistics and Wilcoxon rank sum statistics for each protein detected by each lectin were calculated. Comparisons were made of cancer versus chronic pancreatitis and normal combined, and of chronic pancreatitis and cancer combined versus normal.

Protein digestion by trypsin: Fractions obtained from NPS-RP-HPLC were concentrated down to approximately 20 μL using a SpeedVac concentrator (Thermo, Milford, MA) operating at 45^{o}C. 20 μl of 100 mM ammonium bicarbonate (Sigma) was then mixed with each concentrated sample to obtain pH 7.8. 0.5 μl of TPCK modified sequencing grade porcine trypsin (Promega, Madison, WI) was added and briefly vortexed prior to a 12-16 hour incubation at 37^{o}C on an agitator.

Mass spectrometry

Protein identification by LC-MS/MS: Digested peptide mixtures from NPS-RP-HPLC collection were separated using a reverse phase column attached to a Paradigm HPLC pump (Michrom Bio Resources Inc, Auburn, CA). For nano-LC-ESI-MS/MS experiments, a nanotrap platform (Michrom) was set up prior to the electrospray source. It included a peptide nanotrap (0.2 x 50 mm, Michrom) and a separation column (0.1 mm x 150 mm, C18, Michrom). Peptide sample was injected and first desalted on the trap column with 5 % solvent B (0.3% formic acid in 98% ACN) at 50 μL/min for 5 min. The peptides were then eluted using a 45 min gradient from 5% to 95% B at a flow rate of 0.25 μL /min where solvent A was 0.3 % formic acid in HPLC grade water.

A Finnigan LTQ mass spectrometer (Thermo) was used to acquire spectra. A 75 μm metal spray tip (Michrom) was used and spray voltage was set at 2.5 kV. The instrument was operated in data-dependent mode with dynamic exclusion enabled. The MS/MS spectra on the five most abundant peptide ions in full MS scan were obtained. All MS/MS spectra were searched against the human protein database from SwissProt using SEQUEST algorithm incorporated in Bioworks software (Thermo). Oxidized methionine and N-acetylation were used as variable modifications during the database search. Trypsin was used as a specific protease with two missed cleavages allowed. Positive protein identification was accepted for a peptide with X_{corr} of greater than or equal to 3.0 for triply-, 2.5 for doubly- and 1.9 for singly charged ions. ΔCn cutoff was set as 0.1. Positive protein identification was validated by Trans-Proteomics pipeline. This software includes both the PeptideProphet and ProteinProphet programs that were

developed by Keller et al. (http://peptideprophet.sourceforge.net/).[24] All the reported proteins have an identification probability higher than 95%.

Glycopeptide mapping: Digested peptide mixtures from target glycoproteins were separated by a capillary RP column (C18, 0.2 x 150mm) (Michrom, Auburn, CA) on a capillary pump (Ultra-Plus II MD, Micro-Tech Scientific, Vista, CA). The capillary column was directly mounted to a micro-injector with a 500 nL internal sample loop (Valco Instruments, Houston, TX). The flow from the solvent delivery pump was split pre-column to generate a flow rate of approximately 4 μL/min. The gradient started at 5% ACN, was ramped to 60% ACN in 25 min and finally ramped to 90% in another 5 min. Both solvent A (water) and B (ACN) contain 0.3 % formic acid. The resolved peptides were detected by an ESI-TOF spectrometer (LCT premier, Micromass/Waters, Milford, MA). The capillary voltage for electrospray was set at 3200 V, and for the sample cone at 45 V. Desolvation was accelerated by maintaining the desolvation temperature at 200^{o}C and source temperature at 100^{o}C. The desolvation gas flow was 250 L/h. The data was acquired in "V" mode and the TOF was externally calibrated using a Sodium Iodide and Cesium Iodide mixture. The instrument was controlled by MassLynx 4.0 software. The experimental masses were matched with theoretical glycopeptide masses of target glycoproteins using GlyMod tool (http://www.expasy.ch/tools/glycomod/).

SDS-PAGE and Lectin blotting of separated fractions: The fractions collected from RP-HPLC were further separated by 1-D SDS-PAGE, run in a Mini-Protean cell (Bio-Rad, Hercules, CA) at 80 V. The resolved proteins were transferred onto a PVDF membrane (Bio-Rad). The PVDF membrane was rehydrated in methanol, rinsed, and then blocked in PBS, containing 1% BSA (Roche, Indianapolis, IN) and 0.1% Tween20.

The membrane was then washed in PBS-T 3 times for 1 min, and then incubated with biotinylated *Aleuria aurentia* lectin (5 μg/mL in PBS-T containing 1% BSA) for 1 hr at room temp. Following incubation with the lectin, the membrane was washed 3 times for 2 min each in PBS-T. Detection was with a 200 ng/mL streptavidin-HRP in PBS-T containing 1% BSA. The membrane was washed in PBS-T 5 times for 5 min each followed by one wash with PBS for 5 min. Chemiluminescence was accomplished using an ECL analysis system (Amersham, Piscataway, NJ), and detected on XAR-5 x-ray film (Kodak). The film was digitized using a high resolution digital camera.

4.3. Results and Discussion

Glycoprotein enrichment, depletion and separation: The analytical work flow is illustrated in Fig. 4.1. 10 normal, 8 chronic pancreatitis and 6 pancreatic cancer serum samples were evaluated using glycoprotein extraction followed by liquid separation and microarray spotting of the separated glycoprotein fractions. 125 μl of each serum sample was first reduced in complexity by immunodepletion prior to the lectin extraction step to facilitate detection of lower abundance proteins. The immunodepletion was performed using the IgY-12 column (Beckman-Coulter). This column removes the twelve most abundant serum proteins (albumin, IgG, α1-antitrypsin, IgA, IgM, transferrin, haptoglobin, α1-acid glycoprotein, α2-macroglobin, apolipoprotein A-I and A-II and fibrinogen). Fig 4.2a shows the UV chromatogram of the depletion process where the immunodepleted fraction elutes at around 8 min. Following immunodepletion typically about 8% of total serum proteins are retained in the immunodepleted serum fraction.

Glycoproteins retained in the immunodepleted serum were subsequently enriched using a multi-lectin affinity column composed of WGA and ConA. ConA recognizes α-linked mannose, including high mannose-type and mannose core structures which are common to N-linked glycosylated proteins. WGA can interact with some glycoproteins via sialic acid residues and it also binds oligosaccharides containing terminal *N*-acetylglucosamine.[25] Thus, a majority of the complex type glycans can interact with WGA. Combining these two lectins facilitated the extraction of most of the N-linked glycoproteins in serum. We estimate that approximately 70% protein recovery was achieved from each immunodepleted serum using this lectin affinity column.

Thirty μg of protein from each sample of lectin-enriched glycoproteins were further separated on a non-porous reversed-phase HPLC (NPS-RP-HPLC) C18 column, and the eluting proteins were detected by UV absorption at 214 nm. Fig. 4.2b shows the UV map consisting of the chromatograms of three selected samples. A high level of reproducibility in the UV traces among the different samples in the same group was observed, although slight retention time shifts were observed. These shifts can be associated with the manual nature of peak collection used in the experiments. The time between data collection and beginning of sample run could have varied by 2-3 sec since this was done manually. The UV peak area varied within 10% for serum samples from different individuals. Protein fractions were collected by peak thereby making the collected UV peaks relatively pure compared to the non-immunodepleted sample (especially since sample has already been simplified by immunodepletion). Although occasionally more than one protein per UV peak was observed, it was generally found

that the dominant protein was responsible for the UV absorption or glycan expression change.[6]

Lectin Glycoarrays for Differential Detection of Changes in Glycan Structure: The intact glycoproteins were separated and collected and the peaks were spotted on nitrocellulose slides using a non-contact microarray spotter. The microarrays were then hybridized against various lectins for differential glycan expression analysis. Five lectins (AAL, MAL, SNA, PNA and ConA) were used to detect different glycan moieties. AAL recognizes fucose linked (α-1,6) to N-acetylglucosamine or to fucose linked (α-1,3) to N-acetyllactosamine. MAL can detect glycans containing NeuAc-Gal-GlcNAc with sialic acid at the 3 position of galactose whereas SNA binds preferentially to sialic acid attached to terminal galactose in an (α-2,6) and to a lesser degree, an (α-2,3) linkage.[26, 27] In contrast, PNA binds de-sialylated exposed galactosyl (β-1,3) *N*-acetylgalactosamine. In fact, sialic acid in close proximity to the PNA binding site will inhibit PNA binding. ConA was also used to detect high mannose structures. Greater than 95% of N-glycan types can be covered using these five lectins. The glycoproteins were hybridized with lectins to probe differences in glycan content between normal, chronic pancreatitis and pancreatic cancer sera and the binding was visualized using a biotin-streptavidin-AlexaFluor555 interaction.[23]

Figure 4.3 shows sections of 5 microarrays probed with five different lectins. The left 5 lanes contain normal samples, the middle three lanes contain cancer samples and the right 4 lanes contain the chronic pancreatitis samples. The array data suggests that this particular fraction contains glycan structures consisting primarily of mannose and fucose residues as reaction with ConA and AAL were significant. Further, it appears that

the overall levels of mannosylation and fucosylation are higher in cancer samples compared to normal. However, the raw microarray data in this figure was not normalized and should be analyzed with caution.

As only changes induced by variations in glycan expression are of interest, all array spot intensities were normalized with respect to their corresponding UV peak areas from the chromatograms to mitigate protein abundance differences. The normalized data was used for cluster analysis.

Bioinformatics analysis of the glycoprotein patterns: Bioinformatics analysis of the glycoprotein arrays was performed to determine if there were any lectin response patterns that grouped different disease states together. For data visualization, average linkage hierarchical clustering (HC) and principal component analysis (PCA) were used to provide graphical representations of the relationships among the samples. In these unsupervised approaches, the samples were placed either in a hierarchical relationship or as points in a 2-dimensional scatterplot (PCA) based on similarities in normalized glycoform abundances. All 24 sera (8 chronic pancreatitis, 6 pancreatic cancer and 10 normal), each assayed in duplicate, were analyzed using unsupervised visualization approaches and supervised differential abundance analyses. Separate PCA and HC results were generated for all 48 samples, and for the replicate averages of the 24 distinct biological specimens. The normalized abundances were log transformed, then their pair-wise correlations were used to carry out HC, and their pairwise co-variances were used for PCA. Because results on the set of 48 samples clearly showed good reproducibility between replicates from the same biological source (see Fig. 4.4a-e), for biological inferences the results for the 24 specimen-wise averages were used. The scatter plots

where the duplicate averaged samples tend to cluster separately are illustrated in Figure 4.5 a-e. The pancreatic cancer samples clustered further away from the normal sample than the chronic pancreatitis sample pool especially in response to AAL, ConA and SNA. There were some outliers in the cancer pool that fell into the chronic pancreatitis pool. However, it was seen that this behavior always occurred with sera from the same 3 patients, indicating that the outliers were likely due to individual patient heterogeneity. It was also observed that the glycan expression of chronic pancreatitis serum glycoproteins were more similar to glycoproteins from the normal sera than to glycoproteins from the cancer sera, as shown in the fucosylated (Fig. 4.5a) and high mannose glycan expression (Fig. 4.5b). The hierarchical clustering results of average samples detected by ConA and MAL are shown in Fig.4.5f and 4.5g. The clustering results for fucosylated and sialylated glycan expression patterns generally distinguished the three clinical groups and correlated well with the PCA results. Results with some lectins more clearly distinguished cancer from chronic pancreatitis/normal, while other lectins more clearly distinguished normal from cancer/chronic pancreatitis (shown in Fig.4.5 h-j).

Proteins with altered glycan structures in pancreatic cancer serum: For differential abundance analysis, Z-statistics and Wilcoxon rank-sum statistics were calculated for the normalized array spot intensities from each LC fraction, as detected by each lectin. Comparisons were made of pancreatic cancer versus chronic pancreatitis and normal combined and normal versus chronic pancreatitis and cancer combined. Only Z value higher than 2 or lower than -2 (meaning only <5% would be expected by chance) were considered significant. Positive Z value indicates over-expression and negative Z

value indicates under-expression. The data suggests that all of the lectins have substantial power for identifying cancer samples relative to control or normal samples.

Proteins with significant changes ($P<0.05$) in chronic pancreatitis or pancreatic cancer serum were identified by peptide sequencing using nano LC-MS/MS. Positive protein identification was further validated by the Trans-Proteomics pipeline which includes both PeptideProphet and ProteinProphet software.[24] PeptideProphet automatically validates peptide assignment to MS/MS spectra made by a database search program such as SEQUEST. For each dataset, it calculates the distribution of search scores and peptide properties among correct and incorrect peptides, and uses those distributions to compute for each obtained peptide sequence a probability that it is correct. Only identifications with a TPP protein probability of >95% was considered a true hit. ProteinProphet takes the peptides and search results and statistically validates the identifications at the protein level. The altered protein IDs together with their Z-statistics ($Z>2$ or $Z<-2$) are summarized in Table 4.1. The positive Z score in "cancer" is indicative that this glycosylation is specifically over-expressed in cancer compared to normal and pancreatitis combined. The negative Z score in "normal" is indicative that this glycosylation is under-expressed in the normal sample compared to pancreatitis and cancer combined. Thus, the differences shown are cancer specific.

In certain fractions, more than one protein was identified due to co-elution during LC separation. These fractions were further separated by SDS-PAGE and analyzed by lectin blots to determine which of the co-eluting proteins was responsible for the differential lectin response.

Increased fucosylation and sialylation in pancreatic cancer sera were detected on a majority of the differentially glycosylated proteins, including Hemopexin, Beta-2-glycoprotein 1, serum amyloid P-component, Antithrombin-III and Haptoglobin-related protein. Decreased sialylation was detected on plasma protease C1 inhibitor. This phenomenon has also been previously shown.[6] The immunodepletion of the abundant serum proteins in combination with further separation and lectin detection enabled the observation of glycosylation alteration in less-abundant proteins which had previously been difficult to detect. Some proteins have been suggested to be potential marker proteins in cancer. Beta-2-glycoprotein has been observed to be over-expressed in breast cancer serum [28] and serum amyloid P-component has been found down-regulated in stomach cancer tissue.[29] However, the glycosylation pattern alteration of these proteins has not been widely studied in sera from other cancer types.

Increased sialylation and fucosylation of these proteins in cancer serum lends support to the theory that glycosylation changes may have clinical utility for the identification of markers for early cancer detection. In order to verify the glycosylation changes that we observed, lectin immunoblotting and glycopeptide mapping experiments were also performed on selected LC fractions. AAL lectin was used to examine the fucosylation expression level of target proteins. Fig. 4.6a shows the lectin blot results of fucosylated Antithrombin-III. It was observed to be up-regulated in cancer serum compared to normal and chronic pancreatitis serum. Peptide mapping experiments were performed on the tryptic peptides from the LC fractions using μLC-ESI TOF. As shown in Fig. 4.7a, very similar patterns were observed for the unmodified peptides for cancer versus normal samples from Antithrombin-III. However, an over-expressed fucosylated

mono-sialylated glycopeptide was detected in cancer serum (Fig. 4.7b). This is consistent with the up-regulation of fucosylation and sialylation on Antithrombin-III observed in the microarray experiment (Table 4.1). These results highlight the potential utility of using altered glycosylation patterns, rather than absolute protein levels, as markers for early cancer detection.

Haptoglobin-related protein epitope expression is a clinically important predictor of the recurrence of cancer in patients with early breast cancer, especially in combination with progesterone-receptor status.[30] In our study, the over-expression of fucosylated haptoglobin-related protein in pancreatic cancer serum was also verified by both glycoprotein microarrays and lectin blot experiments. (See Fig. 4.6b and Table 4.1) The up-regulation of fucosylated haptoglobin in pancreatic cancer serum has been reported previously[31], although this protein was removed during the immunodepletion step, thus its glycosylation changes were not analyzed in this study. In a peptide mapping experiment, similar levels of a desialylated glycan structure and a mono-sialylated glycan structure were observed on the peptide NLFLNHSENATAK from haptoglobin-related protein in cancer and normal samples (Fig. 4.8a and 4.8b), where the fully sialylated glycan structure on this peptide was found to be up-regulated in cancer sample as shown in Fig. 4.8c. These results confirmed the increased response of SNA lectin and unaltered response of PNA lectin on this protein (Table 4.1).

The over-expression of fucosylation in hemopexin in hepatocellular carcinoma has been previously reported.[32] In pancreatic cancer serum, significant up-regulation with a Z-score of 6.15 was observed on fucosylated Hemopexin. The lectin blotting experiment verified this alteration where an increased response to AAL was observed as

compared to chronic pancreatitis and normal serum (Fig.4.6c). Over-expressed desialylated and partially sialylated glycopeptides were also observed on Kininogen-1 in pancreatic cancer sera (shown in Fig.4.9). This is consistent with the result from glycoprotein microarrays where an increased response of this protein to SNA, MAL and PNA was detected in cancer samples.

In some cases the glycan moieties that were detected by the five lectins including sialylation, fucosylation, galactosylation and mannosylation were all up-regulated in pancreatic cancer. For instance, the glycosylation of both Serum amyloid P-component and Beta-2-glycoprotein 1 were found to be up-regulated as detected by all five lectins. This may be due to the increased branching of glycans that has been associated with metastasis and has been correlated with tumor progression in human cancers of pancreas, breast, colon and melanomas.[8, 11, 33, 34] Fucosyltransferase 3 has been shown to be over-expressed, and several isoforms of mannosidase have been shown to have decreased expression in pancreatic cancer (as compared to chronic pancreatitis and normal pancreata).[35] Fucosyltransferases increase fucosylation in selected proteins. The over-expression of highly branched glycosylation either implicates increased activity of certain glycosyltransferases (which may lead to increased expression of certain terminal glycans such as sialic acid and fucosyl residues) or to decreased mannosidase activity (leading to decreased trimming of high mannose structures, with corresponding increased branching of the high mannose core) in cancer. However, all that being said, no pancreatic proteins were found to show large glycosylation changes in the present study. In fact, the majority of interesting proteins are secreted by the liver. It is well established that pancreatic cancer is a highly inflammatory neoplasm and, as such, may elicit

inflammatory cytokine production with an associated acute phase response from the liver. This acute phase response may, in fact, contribute to the synthesis of altered glycan moieties on the secreted liver glycoproteins.

4.4. Conclusion

We have demonstrated the utility of glycoprotein microarrays as a tool to differentiate serum samples from patients with pancreatic cancer, chronic pancreatitis or normal subjects. Analysis of multiple normal, chronic pancreatitis and pancreatic cancer sera showed distinct segregation of each state following PCA and HC analysis. Normal and chronic pancreatitis sera were closer in similarity to each other, whereas pancreatic cancer sera were distinct from the other two groups. Sialylation and fucosylation were the dominant glycosylation differences seen to change with progression of pancreatic cancer. Both an increase and decrease in glycosylation levels of different proteins were observed as a function of disease. Many proteins whose glycosylation patterns changed as a function of disease have been previously implicated in cancer. The results from this study confirm previously implicated changes [8, 11] in that not only do protein abundances change as a function of cancer, but more importantly, modifications such as changes in glycosylation patterns of a serum glycoproteome may indicate presence or absence of a disease. This change in disease specific glycosylation was further confirmed for selected proteins by glycopeptide mapping experiments using a μLC-ESI-MS platform that were able to show distinct glycopeptide differences between different sample types. The ability to screen serum glycosylation patterns for sample classification and detect the location of

the altered glycosylations by further mass spectrometric validation may have utility for the early detection of cancer.

Table 4.1: Z values of the altered glycosylations detected by five lectins.(Z>2 or Z<-2 corresponds to P<0.05)

Protein ID / acc #	AAL		MAL		SNA		ConA		PNA	
	Normal	Cancer	Normal	Cancer	Normal	Cancer	Normal	Cancer	Normal	Cancer
Beta-2-glycoprotein 1 (P02749)		2.49		3.3		2.08		2.07	-2.47	2.13
Hemopexin (P02790)		6.15		2.85		3.24		3.01		
Haptoglobin-related protein (P00739)	-3.6	2.82	-2.49		-3.71	2.41	-3.08		-3.63	
Serum amyloid P-component (P02743)	-4.96	4.02	-4.96	2.85	-5.28	4.11	-5.31	3.59	-5.96	3.12
Clusterin (P10909)	-2.22	2.92	-2.52			2.22			-2.08	
Antithrombin -III (P01008)	-3.5	3.18	-2.9	3.28	-2.93	2.58	-3.24	2.63	-3.44	2.56
Kininogen-1 (P01042)	-2.69	4.31		2.39	-2.64	3.98	-3.06	3.95	-2.1	2.14
Plasma protease C1 inhibitor (P05155)						-2.97				

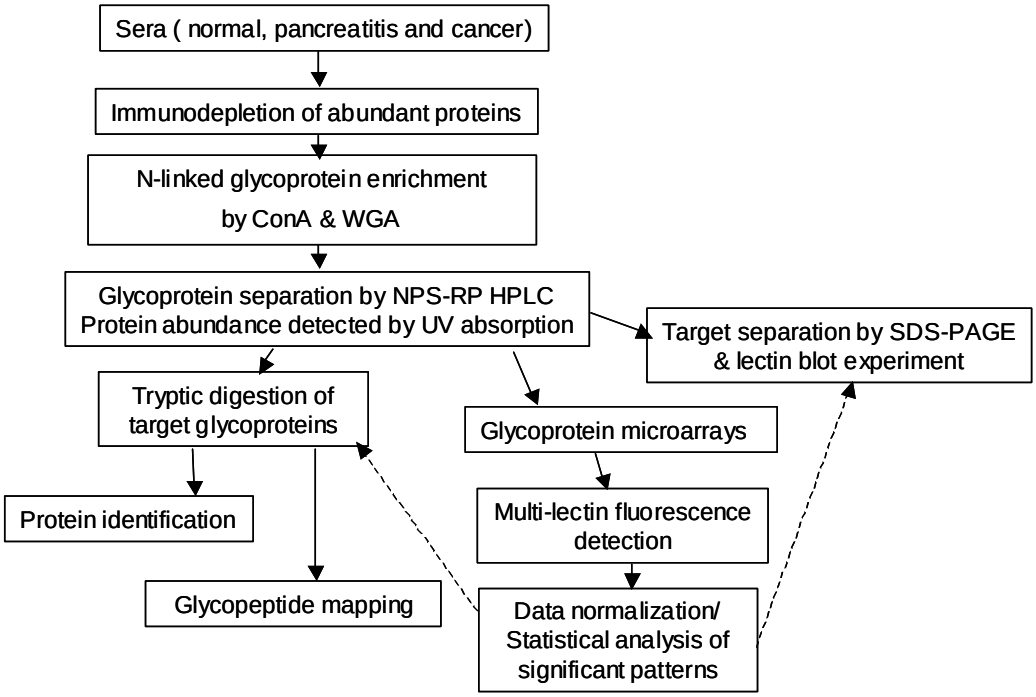
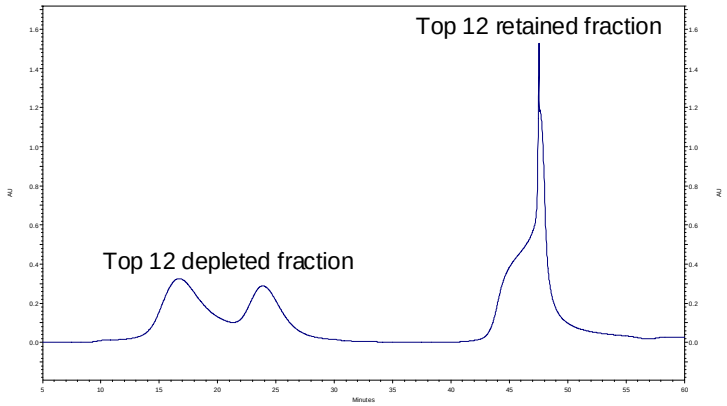


Figure 4.1: Strategy used to screen the glycosylation patterns and characterize the target glycoproteins using samples of normal, chronic pancreatitis, and pancreatic cancer sera.

(a)



(b)

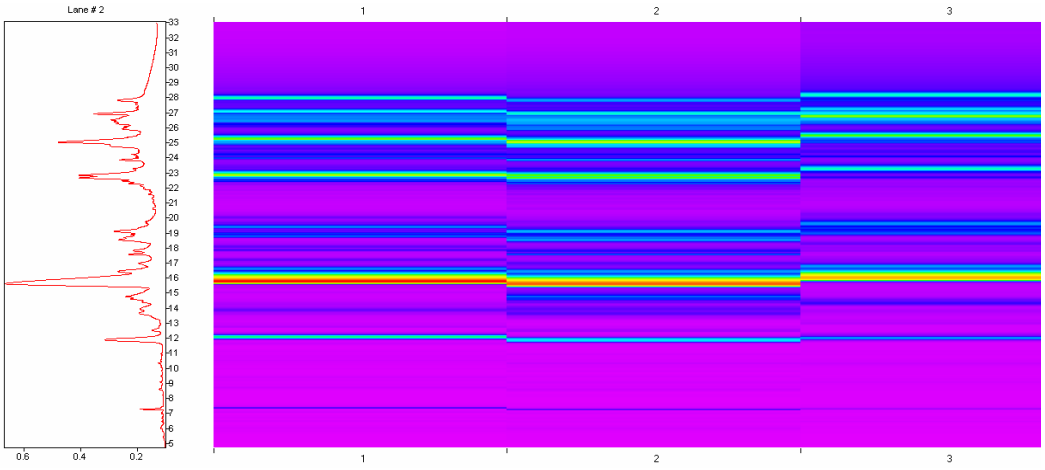


Figure 4.2: (a) UV Chromatogram of 125 µl serum depletion by IgY antibody column to remove the 12 high abundance proteins. During the binding process, the fraction flowing through was collected as the immunodepleted serum fraction, with the abundant protein fraction collected during elution. The absorption was set at 280nm. (b) WGA and ConA selected glycoproteins from three depleted serum samples were separated by NPS-RP C18 column. The UV absorption was at 214nm.

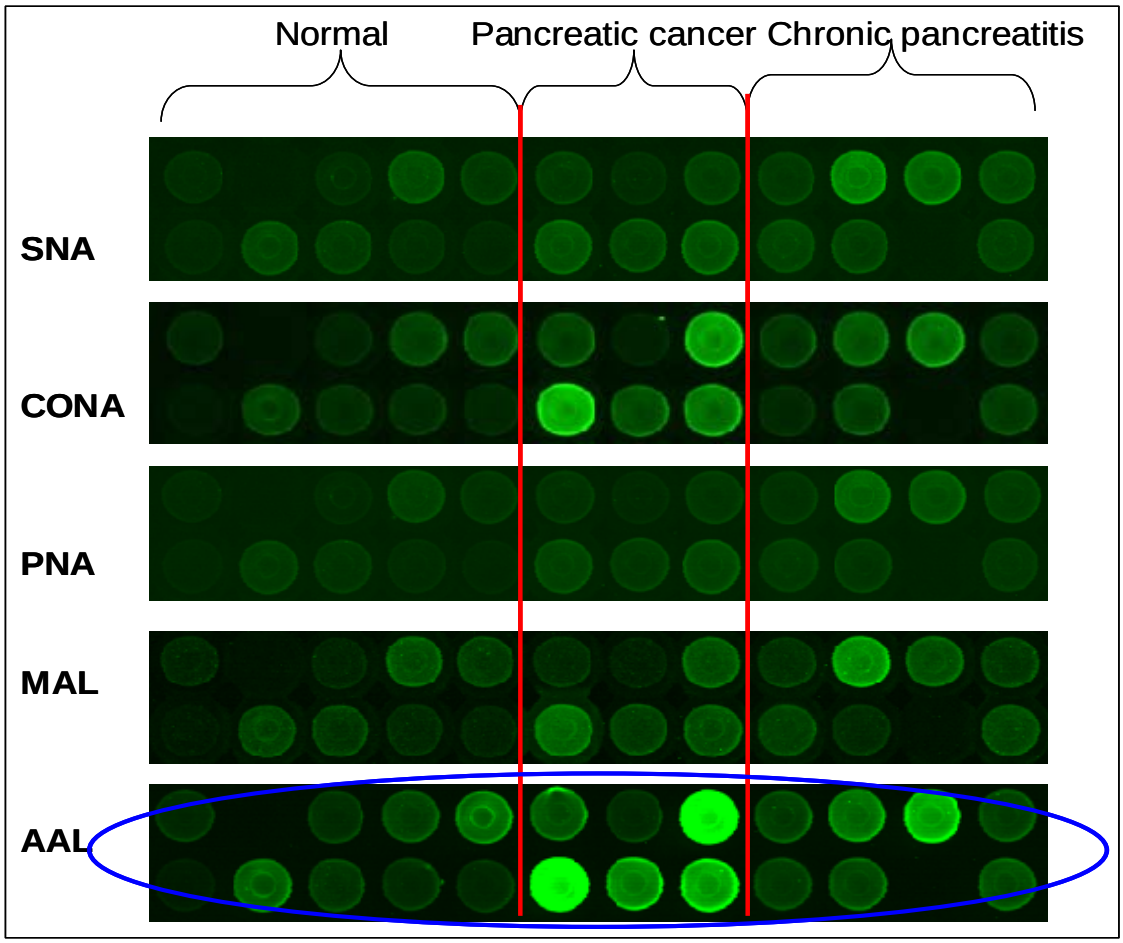
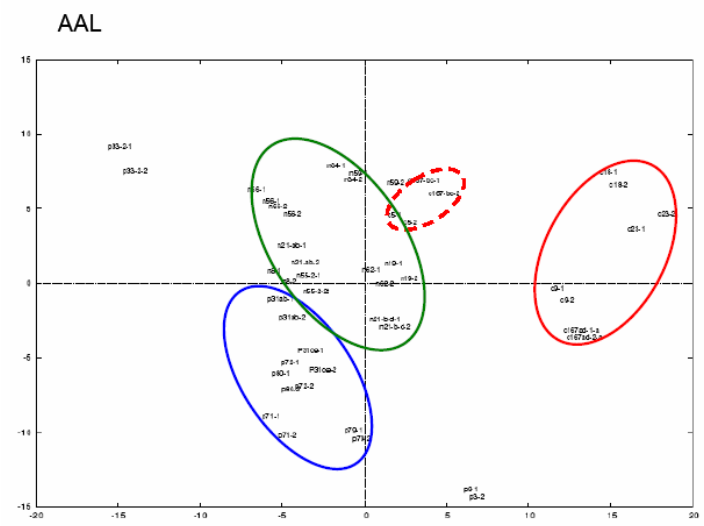


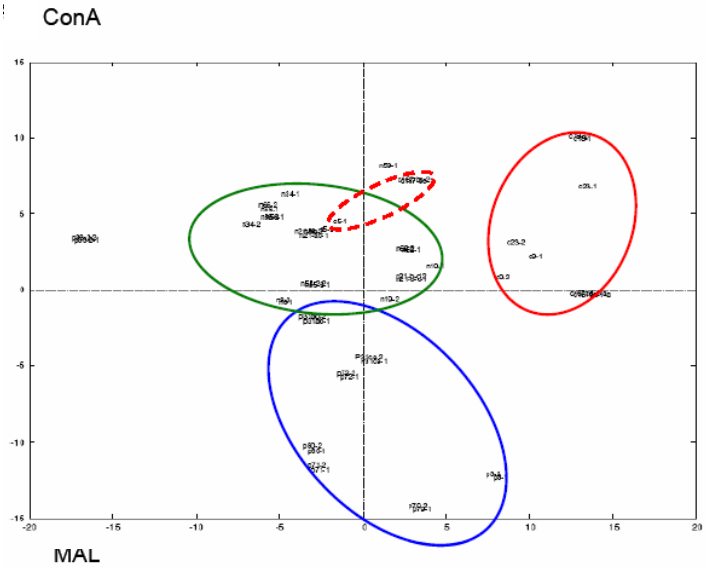
Figure 4.3: Sections of glycoprotein microarray showing comparison of one fraction from NPS-RP-HPLC across all 24 samples. Each panel is a section of identical arrays probed with lectin indicated on the left side of the panel. It was observed that this fraction contained proteins that were predominantly mannosylated and fucosylated. It was also observed that the level of glycosylation (based on raw microarray data) was higher in cancer samples compared to the controls.

Figure 4.4:

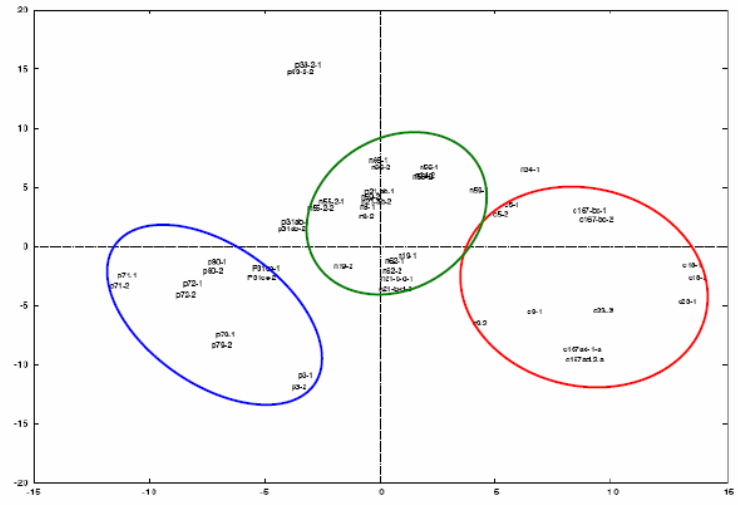
(a)



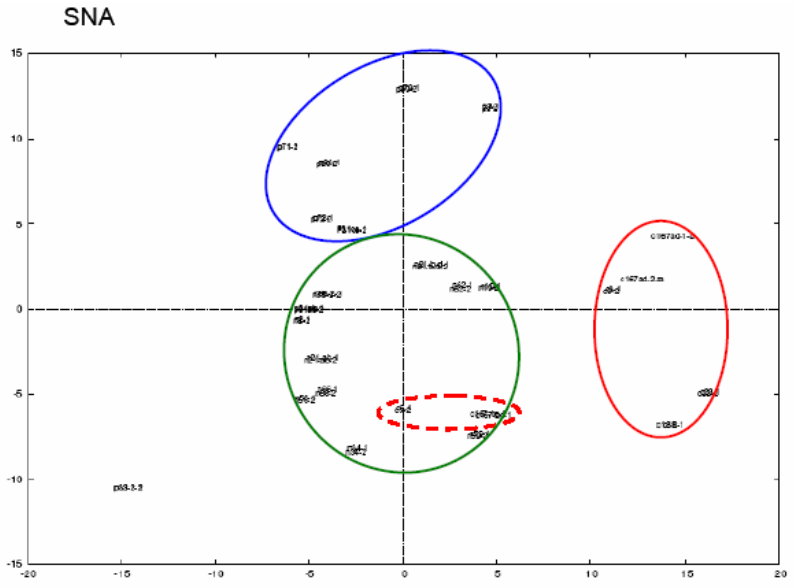
(b)



(c)



(d)



(e)

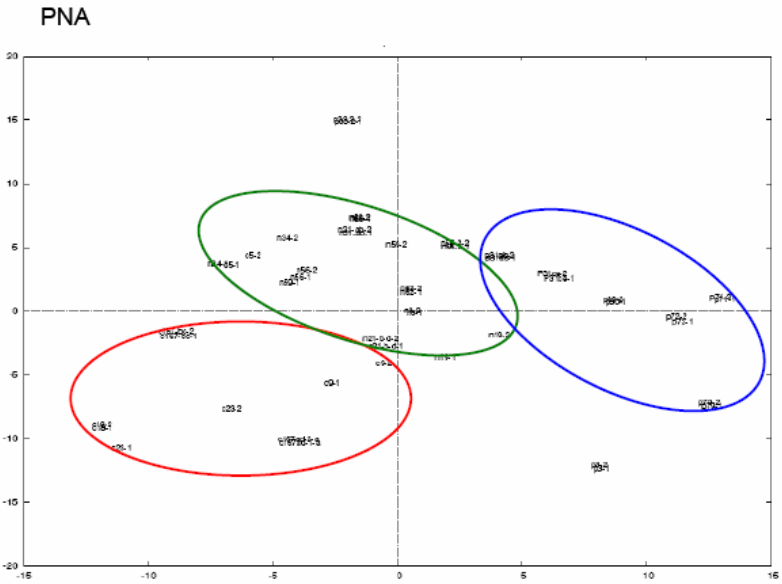
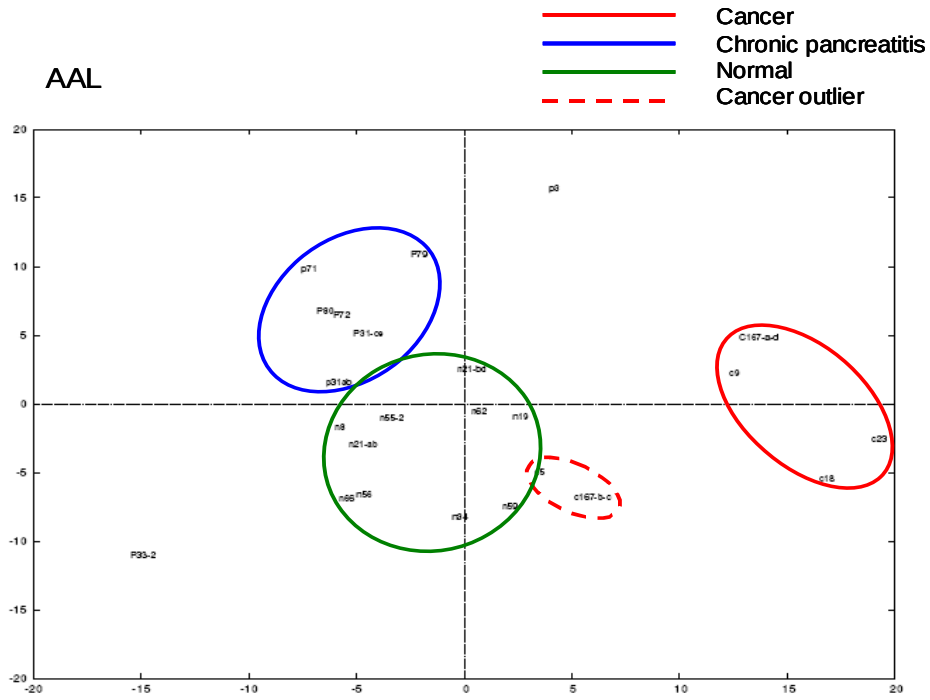


Figure 4.4: The normalized glycoprotein microarray responses to lectins (a) AAL (b) ConCA (c) MAL (d) SNA (e)PNA were visualized by principal component analysis (PCA). 24 serum samples (10normal, 8 pancreatitis and 6 pancreatic cancers), assayed in duplicate, were analyzed without replicate averaging.

Figure 4.5:
(a)


Cancer
Chronic pancreatitis
Normal
Cancer outlier
AAL

(b)

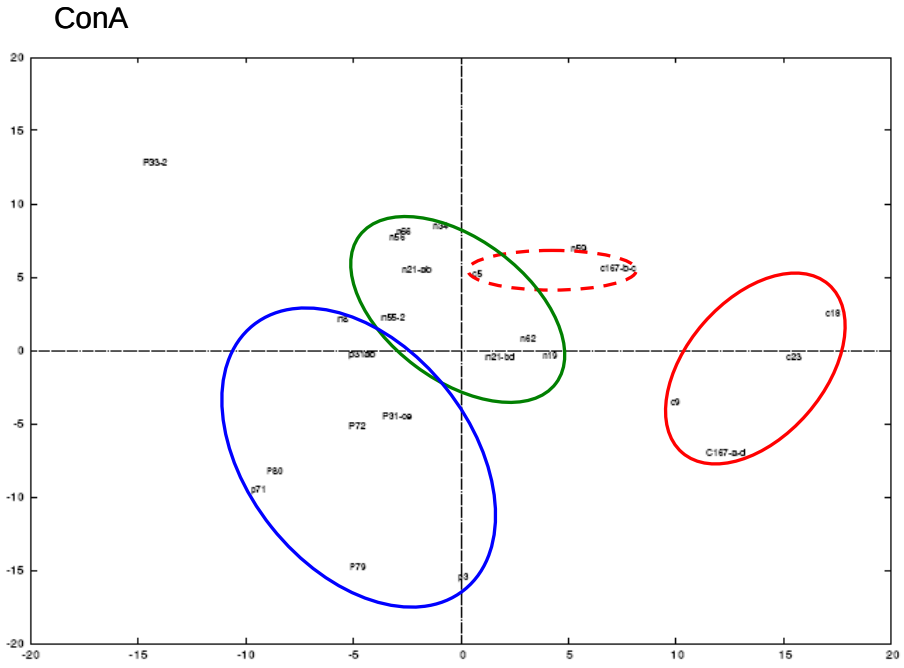
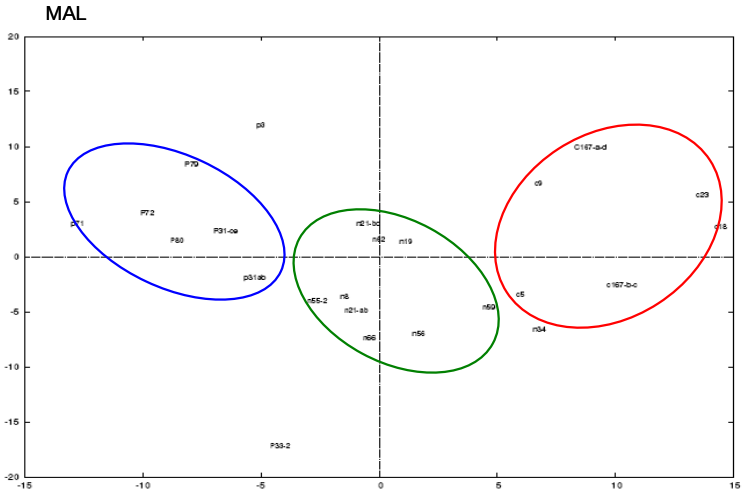
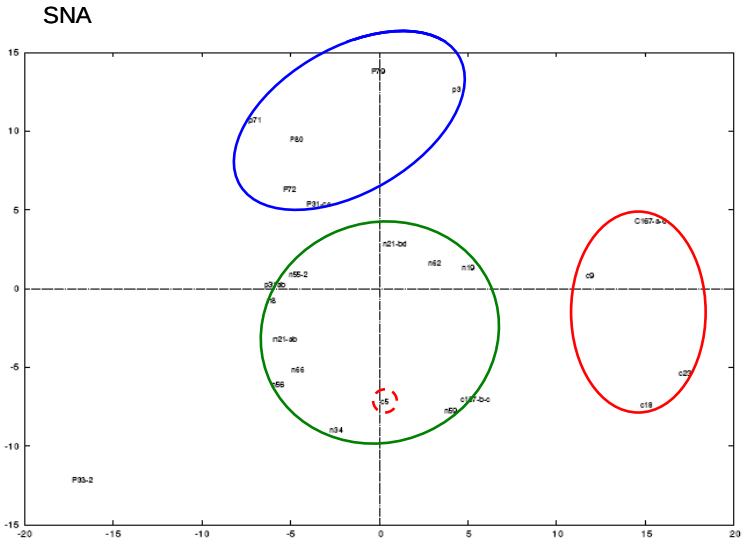

ConA

Figure 4.5:

(c)



(d)



(e)

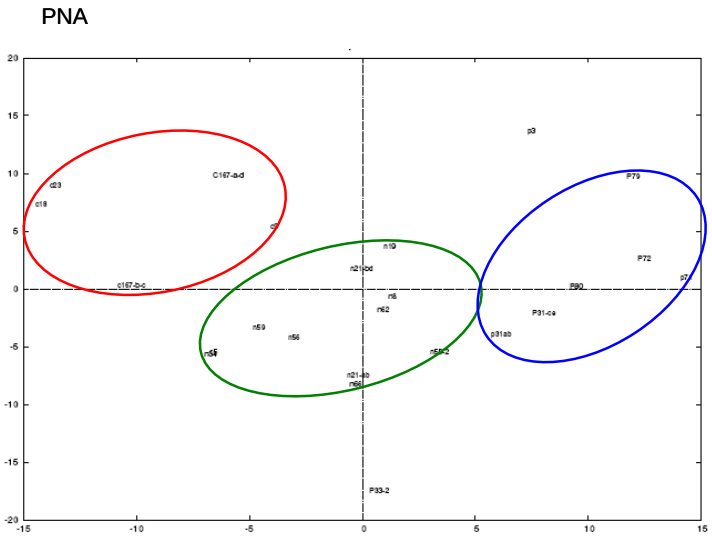
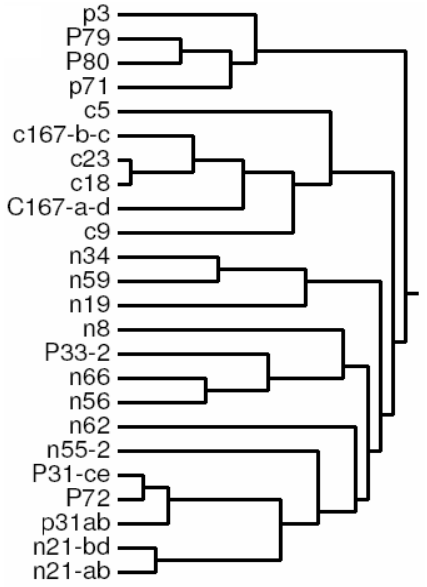
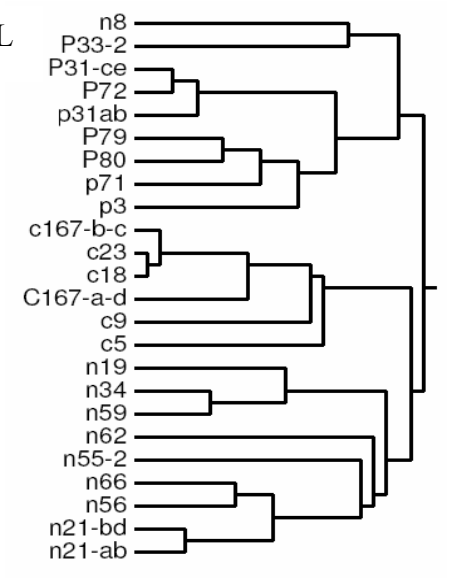


Figure 4.5:

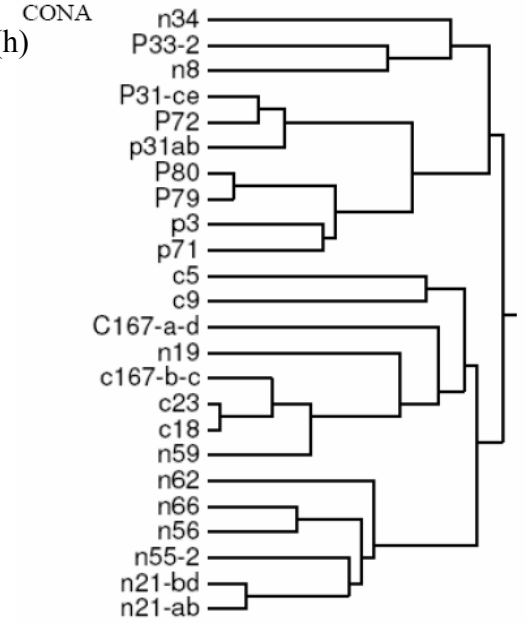
(f) AAL



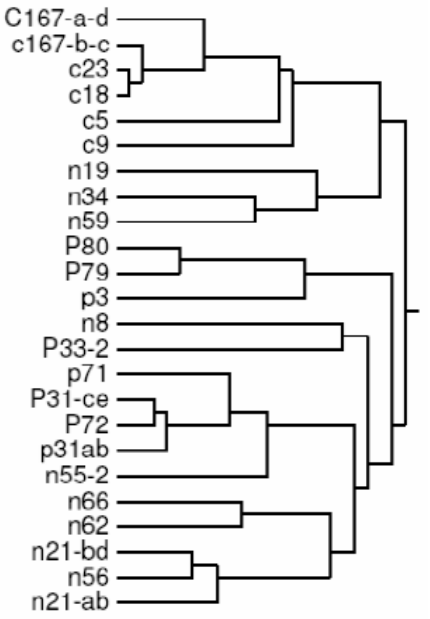
(g) MAL



(h) CONA



(i) PNA



(j)

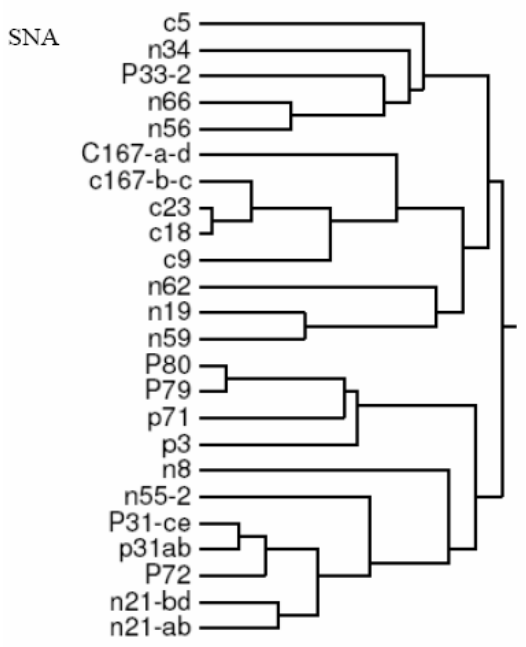


Figure 4.5: The normalized glycoprotein microarray responses to lectins (a) AAL, (b) ConA, (c) MAL, (d) SNA, and (e) PNA were visualized by principal component analysis (PCA). Twenty-four serum samples (10 normal, 8 chronic pancreatitis, and 6 pancreatic cancers) were studied. Average linkage hierarchical clustering (HC) of the array responses to (f) AAL, (g) MAL, (h) ConA, (i) PNA and (j) SNA were shown to provide graphical representations of the relationships among the samples. The figure shows the clustering of serum samples obtained from patients with pancreatic cancer, chronic pancreatitis, or from normal subjects.

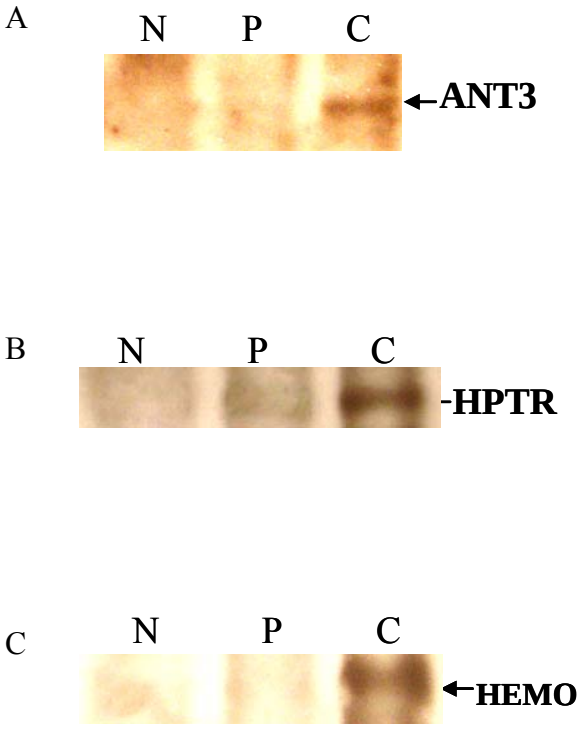
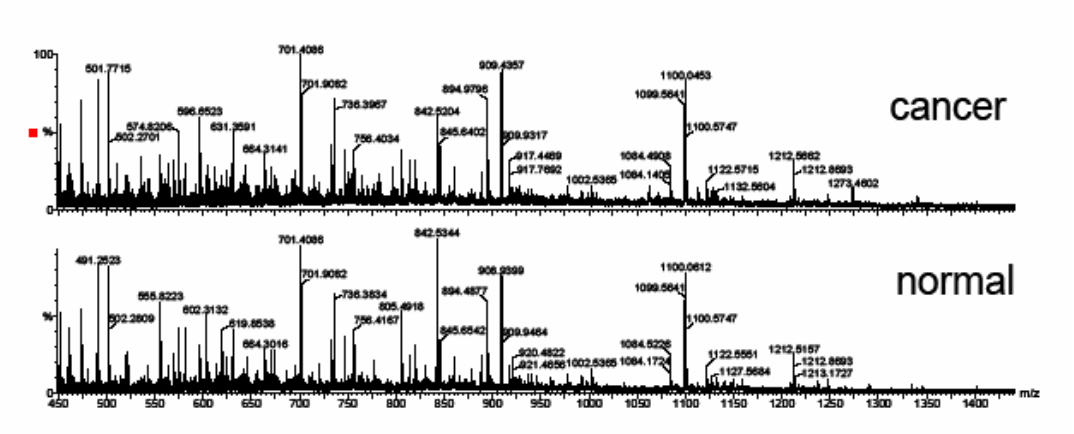


Figure 4.6: AAL lectin blot analysis of (a) Antithrombin-III, (b) Haptoglobin-related protein, (c) Hemopexin in N (normal), P (chronic pancreatitis), and C (pancreatic cancer) serum.

(a)



(b)

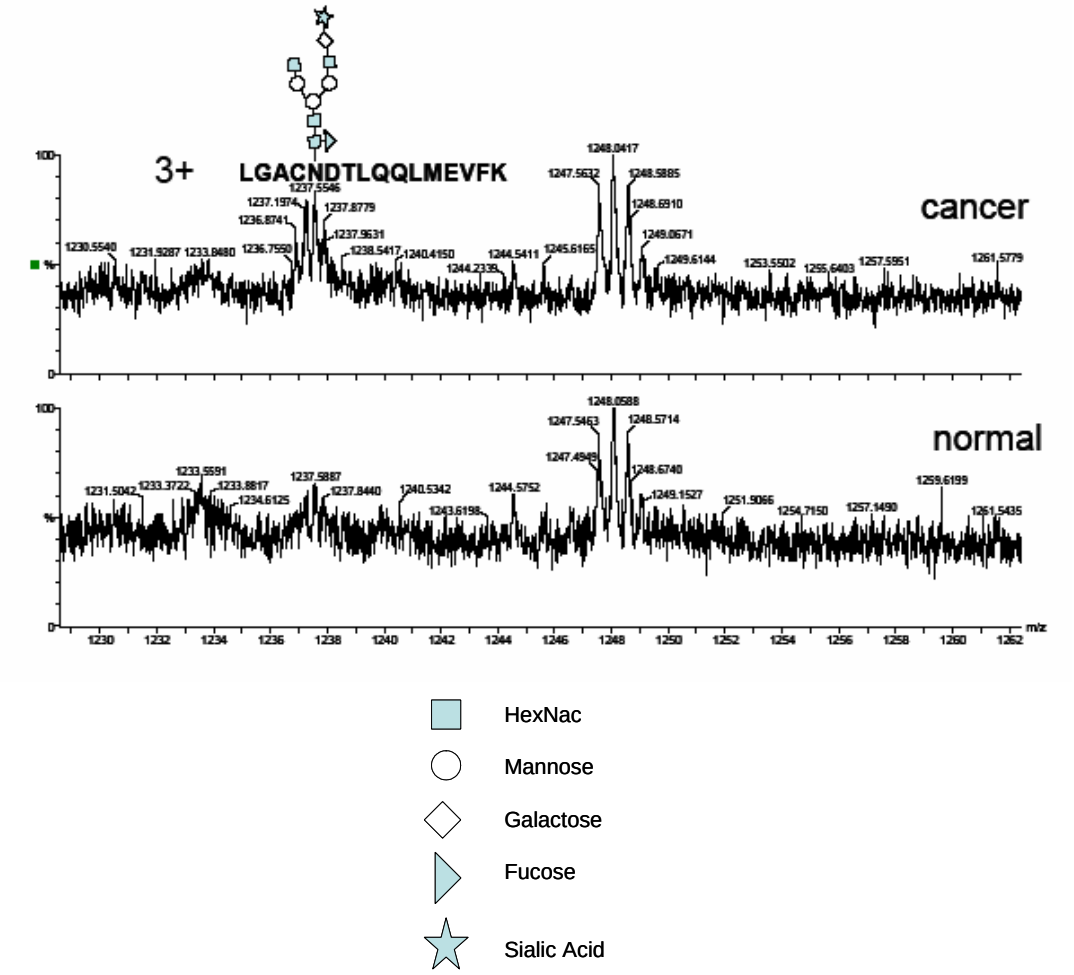


Figure 4.7: Peptide mapping of Antithrombin-III. (a) Very similar patterns of unmodified peptides and (b) altered glycopeptide LGAC*N*DTLQQLMEVFK (124-139) + $(Hex)_1$ $(HexNAc)_2$ $(Deoxyhexose)_1$ $(NeuAc)_1$ + $(Man)_3(GlcNAc)_2$ were detected by µLC-ESITOF in normal and pancreatic cancer serum.

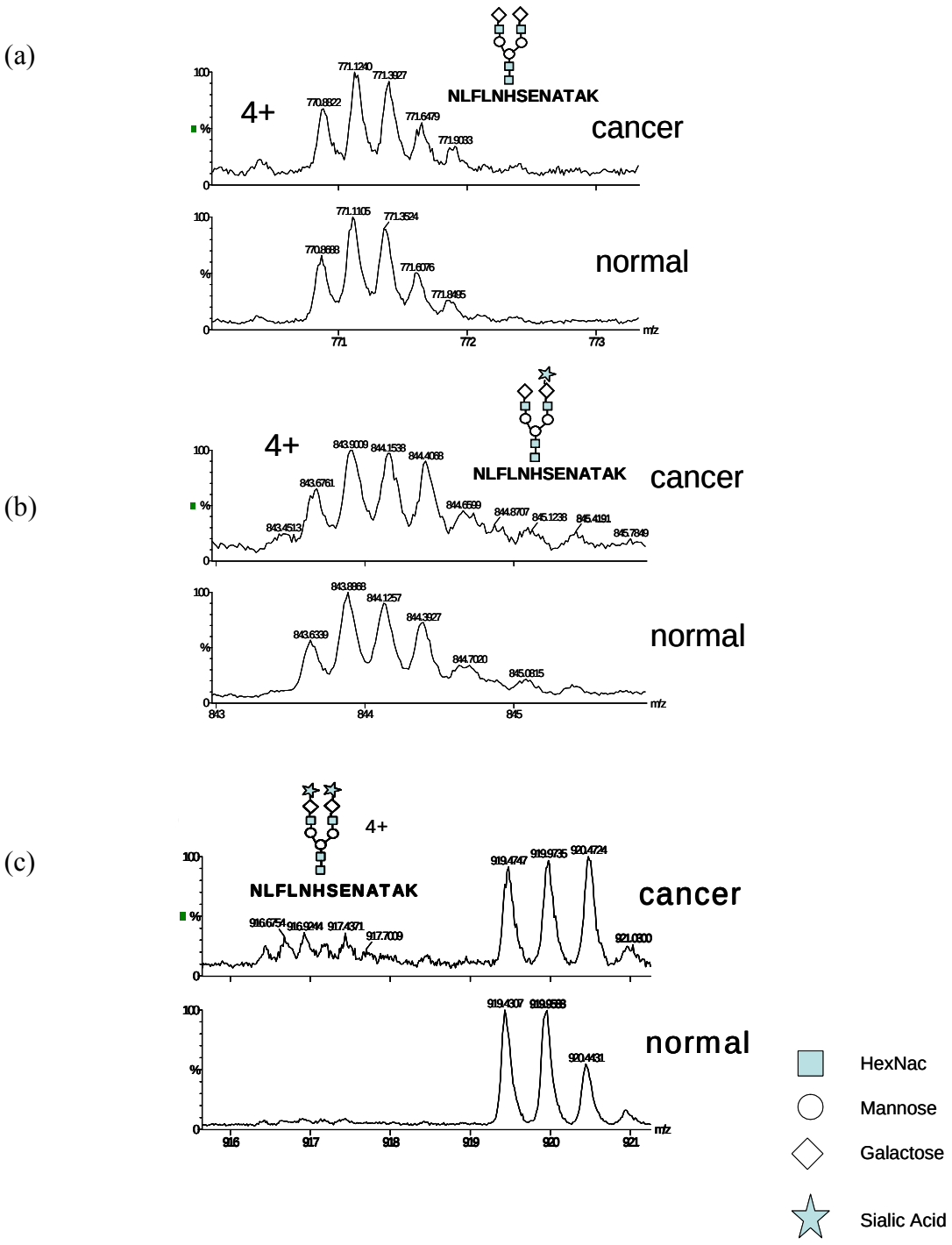


Figure 4.8: Peptide mapping of Haptoglobin-related protein. (a) Glycopeptide NLFL NHSE NATAK(145-157) + (Hex)2(HexNAc)2 + (Man)3(GlcNAc)2, (b) glycopeptides NLFL NHSE NATAK(145-157) + (Hex)2(HexNAc)2(NeuAc)1 + (Man)3(GlcNAc)2, and (c) glycopeptides NLFLNHSENATAK(145-157) + (Hex)2 (HexNAc)2(NeuAc)2 + Man)3(GlcNAc)2 were detected as multiple charged peaks in normal and pancreatic cancer serum.

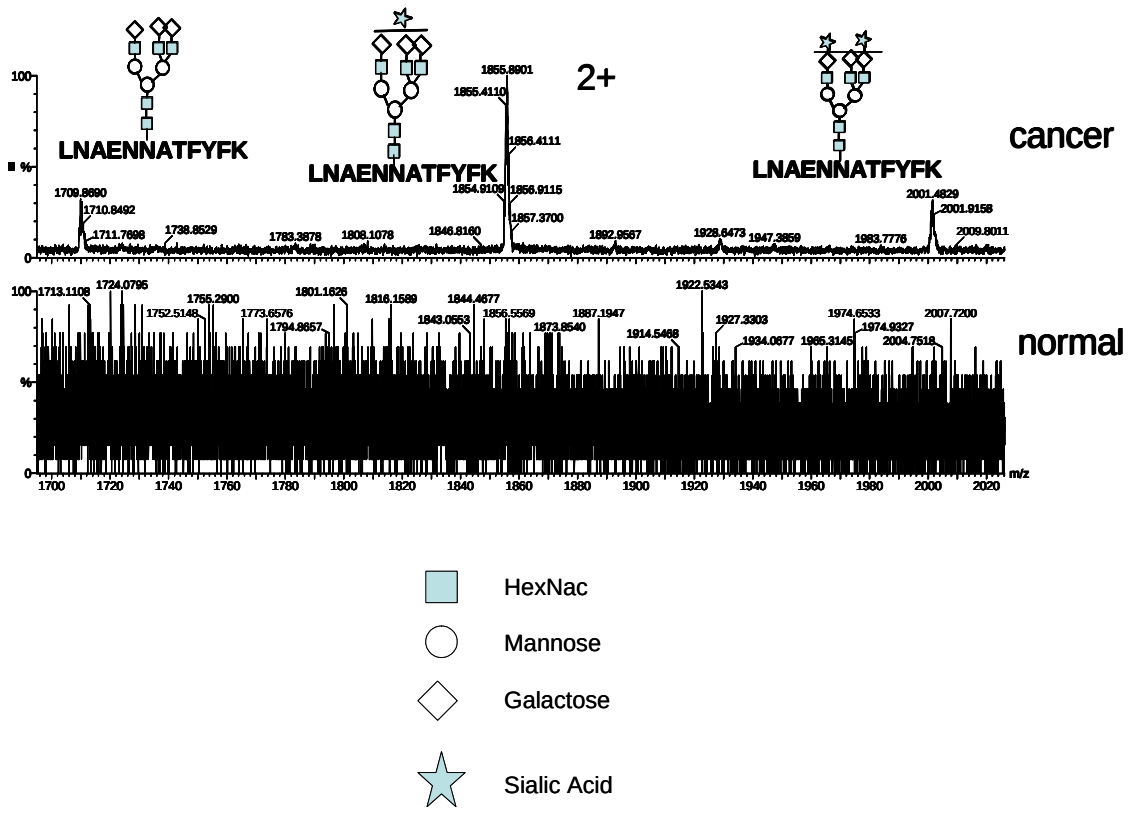


Figure 4.9: Peptide mapping of Kininogen-1 (P01042). Glycopeptide LNAEN NATFYFK(289-300) + Hex)3(HexNAc)3 + (Man)3(GlcNAc)2, LNAENNATFYFK(289-300) + (Hex)3(HexNAc)3(NeuAc)1 + (Man)3(GlcNAc)2, and LNAENNATFYFK(289-300) + (Hex)3(HexNAc)3(NeuAc)2 + (Man)3(GlcNAc)2 were detected as doubly charged peaks.

4.5. References

[1] Jemal, A., Murray, T., Ward, E., Samuels, A., *et al.*, *CA Cancer J Clin* 2005, *55*, 10-30.

[2] Dalgleish, A. G., *Bmj* 2000, *321*, 380.

[3] Duffy, M. J., *Ann Clin Biochem* 1998, *35 (Pt 3)*, 364-370.

[4] Nazli, O., Bozdag, A. D., Tansug, T., Kir, R., Kaymak, E., *Hepatogastroenterology* 2000, *47*, 1750-1752.

[5] Durand, G., Seta, N., *Clin Chem* 2000, *46*, 795-805.

[6] Zhao, J., Simeone, D. M., Heidt, D., Anderson, M. A., Lubman, D. M., *J Proteome Res* 2006, *5*, 1792-1802.

[7] Peracaula, R., Royle, L., Tabares, G., Mallorqui-Fernandez, G., *et al.*, *Glycobiology* 2003, *13*, 227-244.

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

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

[10] Peracaula, R., Tabares, G., Royle, L., Harvey, D. J., *et al.*, *Glycobiology* 2003, *13*, 457-470.

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

[12] Kasbaoui, L., Harb, J., Bernard, S., Meflah, K., *Cancer Res* 1989, *49*, 5317-5322.

[13] Kellokumpu, S., Sormunen, R., Kellokumpu, I., *FEBS Lett* 2002, *516*, 217-224.

[14] Ng, R. C., Roberts, A. N., Wilson, R. G., Latner, A. L., Turner, G. A., *Br J Cancer* 1987, *55*, 249-254.

[15] Uttamchandani, M., Wang, J., Yao, S. Q., *Mol Biosyst* 2006, *2*, 58-68.

[16] Sobek, J., Bartscherer, K., Jacob, A., Hoheisel, J. D., Angenendt, P., *Comb Chem High Throughput Screen* 2006, *9*, 365-380.

[17] Hultschig, C., Kreutzberger, J., Seitz, H., Konthur, Z., *et al.*, *Curr Opin Chem Biol* 2006, *10*, 4-10.

[18] Pal, M., Moffa, A., Sreekumar, A., Ethier, S. P., *et al.*, *Anal Chem* 2006, *78*, 702-710.

[19] Wang, D., Liu, S., Trummer, B. J., Deng, C., Wang, A., *Nat Biotechnol* 2002, *20*, 275-281.

[20] Haab, B. B., Dunham, M. J., Brown, P. O., *Genome Biol* 2001, *2*, RESEARCH0004.

[21] Yan, F., Sreekumar, A., Laxman, B., Chinnaiyan, A. M., *et al.*, *Proteomics* 2003, *3*, 1228-1235.

[22] Davies, D. H., Liang, X., Hernandez, J. E., Randall, A., *et al.*, *Proc Natl Acad Sci U S A* 2005, *102*, 547-552.

[23] Patwa, T. H., Zhao, J., Anderson, M. A., Simeone, D. M., Lubman, D. M., *Anal Chem* 2006, *78*, 6411-6421.

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

[25] Bakry, N., Kamata, Y., Simpson, L. L., *J Pharmacol Exp Ther* 1991, *258*, 830-836.

[26] Shibuya, N., Goldstein, I. J., Broekaert, W. F., Nsimba-Lubaki, M., *et al.*, *Arch Biochem Biophys* 1987, *254*, 1-8.

[27] Wang, W. C., Cummings, R. D., *J Biol Chem* 1988, *263*, 4576-4585.

[28] Alsabti, E. A., Muneir, K., *Jpn J Exp Med* 1979, *49*, 235-240.

[29] Jang, J. S., Cho, H. Y., Lee, Y. J., Ha, W. S., Kim, H. W., *Oncol Res* 2004, *14*, 491-499.

[30] Kuhajda, F. P., Piantadosi, S., Pasternack, G. R., *N Engl J Med* 1989, *321*, 636-641.

[31] Okuyama, N., Ide, Y., Nakano, M., Nakagawa, T., *et al.*, *Int J Cancer* 2006, *118*, 2803-2808.

[32] Comunale, M. A., Lowman, M., Long, R. E., Krakover, J., *et al.*, *J Proteome Res* 2006, *5*, 308-315.

[33] Kim, Y. J., Varki, A., *Glycoconj J* 1997, *14*, 569-576.

[34] Dube, D. H., Bertozzi, C. R., *Nat Rev Drug Discov* 2005, *4*, 477-488.

[35] Logsdon, C. D., Simeone, D. M., Binkley, C., Arumugam, T., *et al.*, *Cancer Res* 2003, *63*, 2649-2657.

Chapter 5

Glycoprotein profiling in plasma samples to elucidate glycoprotein biomarkers of colorectal cancer: An application of natural glycoprotein microarrays and lectin blots

5.1. Introduction

Colorectal cancer is the third most common cancer in the world. It is estimated that one million new cases and half a million new deaths occur due to this disease every year.[1] Colorectal cancer related deaths amount to only 14% of all deaths due to cancer. The five year survival rate in colon cancer can be increased to 90% if the tumor is detected when it is still localized (not malignant and metastatic).[2] Current screening methods for colorectal cancer include fecal testing, sigmoidoscopies, barium enemas and colonoscopies.[3-5] Because some of these methods are invasive or not pleasant, patients are often resistant to these tests. Blood based (serum or plasma) tests are considerably less invasive and therefore a method of choice for colorectal cancer screening possibly increasing the number of patients screened for the disease and therefore diagnosed at a stage early enough for successful intervention.

A large amount of serum based proteomics studies are currently being pursued for elucidation of cancer biomarkers for early detection as well as monitoring the effectiveness of therapy.[6] Plasma and serum typically contain proteins native to this

biological fluid as well as proteins released from diseased cells.[7] Glycoproteins within plasma comprise the most abundant post translationally modified sub-proteome.[8-12] Currently known cancer markers that are glycoproteins include Her2/Neu in breast cancer, prostate-specific antigen (PSA) in prostate cancer, CA125 in ovarian cancer, carcinoembryonic antigen (CEA) in colon, breast, bladder, lung and pancreatic cancers. CEA presents poor sensitivity specificity and therefore cannot be effectively used for early detection.[13] CA19-9, CA242, CA195, CA50, CA74-2 and TIMP-1 have also been proposed as potential markers of colorectal cancer but their sensitivities and specificities are also too low for diagnostic purpose.[14-21]

Changes in abundance and glycan structure have been previously implicated in multiple events during cancer progression ranging from cell growth and differentiation, adhesion, metastasis and immune surveillance.[22-24] Invasiveness and metastatic ability have been linked to changes in sialylation of cancer cells. Increase in sialylation could be due to multiple factors some of which include increased activity of sialyl transferases and increase in the numbers of potential sialylation sites on N-linked glycans.[25] Aberrant fucosylation has also been shown in pancreatic cancer progression.[26, 27] Plasma glycoproteins are therefore important targets because their identity and glycan structure changes could provide insightful information about critical transformations due to cancer progression and because they can be used as potential diagnostic markers of cancers. In this study, multidimensional liquid separation was utilized on immunodepleted plasma followed by natural glycoprotein microarray production with the goal to screen plasma samples for glycan pattern changes as a function of colorectal cancer. Statistical analysis of the data provided a reliable means to identify plasma glycoproteins possessing altered

glycosylation. Statistically significant altered glycoproteins were further validated on a second independent group of plasma samples. These proposed glycoproteins may have utility in the detection of colorectal cancer.

5.2. Experimental Section

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 colonoscopies 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.

Preparation of glycoprotein samples for lectin glycoarrays or lectin blot

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.

***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.

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.

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).

Statistical analysis of lectin glycoarray data

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. The PCA was based on all microarray measurements without selection or weighting. All samples were included in the analysis without selection.

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.

Z-statistics

For differential abundance analysis, Z-statistics for each protein detected by each lectin were calculated. The Z-statistic is the difference in mean levels between two groups being compared (based on log_2 data) divided by an estimate of its standard error. For single comparisons, Z-statistics greater than approximately 2 in magnitude correspond to p-values smaller than 0.05. The Z-statistics of differentially glycosylated proteins detected by lectins together with fold changes both in log 2 and non-log 2 forms are shown in supplementary Table 1. 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.

SDS-PAGE and lectin blot

To identify and validate the glycoproteins of interest, protein fractions from NPS-RP-HPLC were divided into two aliquots and 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) membranes (Bio-Rad). The PVDF membranes were blocked with 5%

w/v BSA (Roche, Indianapolis, IN) in PBS-T overnight at 4°C and then incubated with either biotinylated AAL or SNA (2 µg/mL in PBS-T containing 3% BSA) for 1 hr at room temperature. The membranes were 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). Corresponding colloidal blue stained bands of proteins of interest were identified by nano-LC MS/MS.

Protein digestion

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 trypsin activity. 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.

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.

Mass spectrometry for protein identification and glycosylation site determination

A Paradigm pump 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 gel bands. 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 a nano-ESI ion source with the spray voltage set at 2.6 kV.

To sequence 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). Two missed cleavages were allowed. Protein identification was accepted as positive 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 $\Delta Cn \geq 0.1$. The sequence database search was set to accept the following modifications: carboxymethylated cysteines due to treatment with iodoacetamide, 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 determined 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

Immunoaffinity depletion and lectin affinity enrichment

Determination of alterations in glycan structure presents a challenge when the sample containing the glycoproteins is a complex biological medium. Human serum or plasma contains proteins over a wide dynamic range of approximately 22 orders of magnitude.[28] In addition about 95% of serum comprises of less than 10 high abundance proteins. Signals from glycoproteins of clinical relevance are therefore often beyond the detection limit due to suppression by the high abundance proteins. To facilitate the analysis of glycoproteins expressed in the mid to low level abundance range, the most abundant proteins were depleted from the sample using immunoaffinity chromatography, as shown in the flowchart of the proposed method (Fig. 5.1). 250 µl of each plasma sample was first delipidated and then immunodepleted to remove lipids and the 12 most abundant plasma proteins based on an avian antibody (IgY)-antigen interaction. Following the immunodepletion step, approximately 7% of the total protein mass in the plasma samples remained as shown in table 5.1. Figure 5.2A shows representative chromatograms that demonstrate the reproducibility of the immunodepletion step using 2 normal, 2 adenoma and 2 colorectal cancer samples.

Proteins that did not bind to the solid phase on the immunodepletion column (and were therefore free of the top 12 abundant proteins) were subjected to ConA affinity chromatography. The resulting fraction contained an enriched concentration of *N*-glycosylated proteins. ConA has a broad specificity due to its binding preference to oligomannosidic, hybrid, and bi-antennary *N*-glycans, either unconjugated or attached to proteins or peptides [29]. O-glycopeptides or glycoproteins that contain exclusively O-glycosylation sites were not bound by this lectin. Approximately 70% of the immunodepleted plasma protein content was recovered by ConA affinity chromatography suggesting that the majority of the immunodepleted fraction comprised of glycoproteins. The two-step enrichment procedure achieved a significant reduction in analyte complexity because the immunodepletion of the 12 most abundant proteins significantly increased the dynamic range of detection and reduced sample heterogeneity due to the removal of the highly variable IgG, IgA and IgM proteins. In addition, the subsequent *N*-glycoprotein enrichment step afforded 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 5.2B shows the reverse phase chromatograms of the plasma samples from different plasma samples (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. Slight peak height differences are evident but the overall peak profiles are almost identical in all cases. These results suggest that the analysis of glycoprotein

expression alone may not provide valuable information to differentiate the clinical status of plasma samples. A more detailed analysis of possible glycan structure differences may prove to be more successful for classification.

Lectin glycoarrays for visualizing of *N*-glycosylation pattern across plasma samples

To analyze the plasma glycosylation patterns, all fractions from the NPS-RP-HPLC separation of the plasma samples were arrayed on nitrocellulose slides as unique spots. The slides were then screened in duplicate using five different lectins to analyze the different glycan structures: *Aleuria aurentia* lectin (AAL), *Sambucus nigra* bark lectin (SNA), *Maackia amurensis* lectin II (MAL), peanut agglutinin (PNA), and Concanavalin A (ConA). The binding affinities of these lectins are detailed in the previous two chapters. The utilization of these five lectins has been highly successful in covering >95% of the reported *N*-glycan types and in differentiating them according to their specific structures [30]. Images of sections of slides showing response of all fractions from all samples are illustrated in figure 5.3. It can be seen that just by array spot intensities themselves, control and disease samples can not be easily distinguished necessitating normalization procedures. Because only variations in glycan expression were desired, all array spot intensities were normalized by dividing the corresponding UV peak area to eliminate protein abundance differences since this method proved to work in the studies with pancreatic cancer sera. The normalized array data showed that the levels of protein fucosylation and sialylation were higher in colorectal cancer and adenoma plasma samples as compared to the plasma controls i.e. normal plasma samples.

Statistical analysis of *N*-glycosylation pattern changes

Principal components analysis (PCA) and hierarchical clustering (HC) of the normalized glycoprotein spot intensities were performed to see if the plasma samples could be differentiated and grouped based on their overall *N*-glycosylation pattern differences into their clinical state. For PCA, all 20 plasma samples assayed 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 (Fig. 5.4A). The closer the spots in the PCA space, the greater the similarity in their glycan profiles over all 36 fractions. In the case of ConA and SNA, the normal controls (red) were grouped separately from cancer (blue) and adenoma samples (green), while most cancer and adenoma samples were clustered together. This suggests that the mannose and certain sialic acid structures show a differential expression in normal samples combined compared to the adenoma and cancer samples together. In the case of AAL and MAL, the normal and cancer samples generally segregated from each other, whereas the adenoma samples overlapped with both the normal and cancer sample groups. The PNA microarray data did not provide high fluorescence intensities for most protein spots but showed similar results to AAL and MAL arrays. 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. However the overlap of adenoma samples with cancer samples may hinder the use of such arrays as the sole technique for plasma state classification. The good reproducibility between the replicates from the same sample (Fig. 5.4B) in PCA plots indicates that the lectin glycoarray is a robust strategy for screening *N*-glycosylation changes among the plasma samples from different

disease states since two different slides when processed with the same lectin do not show severe variation due to manual handling and therefore if differences are seen they could confidently be assigned to biological differences in the sample. Similar results were observed in HC by using the Pearson correlation coefficient for distance metrics (Fig. 5.5). 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.

Z-statistics of each array spot were also calculated. While the PCA and HC analysis focused on overall glycan pattern difference between samples, Z-statistics analysis enabled the comparison of individual fractions across samples to see if specific peaks in the reverse phase chromatograms showed differential glycan expression across all samples. Such an analysis can allow for identification of signature peaks and therefore proteins that might differentiate the plasma samples of the different clinical states (Table 5.1A). Comparisons were made for 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 suggests reduced glycosylation.

Mass spectrometric analysis of potential biomarkers that showed altered N-linked glycosylation patterns

Initial attempts at identifying the protein with altered glycosylation in a peak presented a problem. It was found at each fraction from the NPS-RP-HPLC fractionation

contained more than one protein. In some cases this number was as high as even 10 to 15 proteins. Since the only separation used for resolving proteins after glycoprotein enrichment was HPLC, this problem was expected. In order to determine which protein from the co-eluting proteins was responsible for the differential responses in the glycoarrays that were processed with lectins, all fractions that demonstrated altered glycosylation were further separated by 1-D SDS-PAGE and then further analyzed by lectin blotting experiments. To that affect, 1D separated proteins on the gel were transferred onto PVDF membranes which were then probed with lectins of interest. The glycoprotein-lectin interaction was visualized by the biotin-streptavidin system where the streptavidin was conjugated to horse radish peroxidase facilitating imaging of the lectin blots. Because elevated fucosylation and sialylation levels in colorectal cancer plasma were detected in the majority 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.

Bands that corresponded to the protein showing differential glycosylation in the lectin blots were excised from their corresponding SDS gels, and digested with PNGase F and trypsin. Protein identity and the possible glycosylation sites were determined by nano-ESI-LC-MS/MS coupled to the SEQUEST database search. Positive identifications were validated by the Trans-Proteomics pipeline (PeptideProphet and ProteinProphet) software. PeptideProphet software was used to confidently identify correct peptide assignments and ProteinProphet was used to validate the protein identifications obtained through SEQUEST database searches. Peptides were identified as truly positive if they had a probability score of at least 0.99 and 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.[31, 32] Figure 5.6A shows a representative nano-LC-ESI-MS/MS spectrum of the 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 that occurs upon deamidation of asparagine residue into aspartic acid.[33] 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 *N*-glycosylation at residue 3. Figure 5.6B is a tandem MS spectrum of another peptide [$(M+2H)^{2+}$ at m/z 716.82] from kininogen-1. Without tandem MS/MS data the location of the exact glycosylation site would be ambiguous since this particular peptide possesses two asparagine residues. However, because the 115 Da shift is only seen for the b6-b7 and y7-y6 ions but not for the b_5-b_4 and y_8-y_7 ions (which show a shift of 114 Da) it could be concluded that the Asn at position 5 was not *N*-glycosylated and that in fact it was the Asn residue at position 6 that was glycosylated. Glycosylated proteins that showed a significant difference in their glycan expression across sample groups together with their Z statistical scores are summarized in Table 5.2A. The corresponding detected glycosylation sites for these proteins are shown in Table 5.2B. Out of the 10 proteins where differential glycosylation between sample groups was observed, 3 of these proteins showed elevated glycosylation in the case of cancer compared to normal and adenoma combined and seven had higher glycosylation levels in cancer and adenoma combined compared to normal. A majority of the proteins identified did not appear to be specific to colon cancer but are likely due to systemic

changes and may be acute phase proteins or proteins from the liver or pancreas. This observation was particularly interesting since the same trend was seen with the potential glycoprotein markers from pancreatic cancer too (see previous chapter). It is possible that changes in liver protein glycosylation are a characteristic of cancers of the GI tract. Nevertheless, the data suggests that Z-statistical analysis of lectin glycoarrays has potential to identify specific proteins that can distinguish cancer samples from adenoma or normal controls. Specifically, the potential markers that distinguish colorectal cancer from adenoma and normal controls that were identified in this study include elevated sialylation and fucosylation in complement C3, histidine-rich glycoprotein, and kininogen-1.

Lectin blot analysis on a separate set of normal, adenoma and colorectal cancer plasma samples for validation of results obtained from arrays

The diagnostic potential of 2 of the glycoprotein biomarkers identified above was validated using an independent set of 30 plasma samples (10 colorectal cancers, 10 adenomas, and 10 normals). The plasma samples were treated in a similar manner to the training set (depleted, enriched and separated by multi-dimensional HPLC separation as described previously) and then analyzed by 1-D SDS-PAGE and lectin blot analysis. Figure 5.7 shows the protein bands of markers of interest after lectin blot experiments. It can be observed that complement C3 showed minimal reactivity to AAL and SNA in all of the normal and adenoma samples, but significantly elevated reactivity in the colorectal cancer samples suggesting higher levels of AAL and SNA in cancers compared to controls. These results from the lectin blot analysis were consistent with that obtained from the Z-statistic analysis in which complement C3 was significantly elevated in its

response to both AAL and SNA in cancer samples compared to adenoma and normal samples. In order to ensure that the lectin blot intensities were a truly due to more abundant glycan expression on the glycoprotein and not the protein abundance itself the peak areas of complement C3 in each plasma sample were compared. As shown in figure 6 approximately equal amounts of protein were loaded on the SDS gel suggesting that increased glycan abundance was in fact the reason for the differential lectin blot responses seen. Similarly, histidine-rich protein displayed significant differential glycan abundances but similar protein expression. In this case, fucosylation was significantly elevated in colorectal cancer samples compared to both adenoma and normal samples. However, similar sialylation was observed in cancer and adenoma combined compared to normal samples. These differences were consistent with the Z-statistical scores observed in the training set. The results from this study highlight the potential utility of monitoring the altered glycosylation patterns instead of absolute protein expression for cancer detection.

5.4. Conclusion

A glycoproteomic strategy for the identification of potential plasma biomarkers with a potential utility in the detection of colorectal cancer has been described in this chapter. Potential serological markers of colorectal cancer were identified by first reducing plasma complexity using immunodepletion and glycoprotein enrichment. Subsequent RP-HPLC separation followed by array generation and lectin hybridization provided a means of statistical monitoring of glycan pattern changes as a function of disease. Because peak intensities from NPS-RP-HPLC separations reflected similar

protein abundances across plasma samples the plasma glycoproteome alone could not be used to differentiate the clinical status of individuals i.e. control samples form diseased samples. However, by utilizing the glycoprotein microarray strategy outlined in chapter 3, normal, adenoma, and colorectal cancer plasma showed distinct clustering according to clinical status. Glycoprotein fractions that were statistically proven to be different between sample groups were identified using SDS-PAGE and lectin blotting experiments coupled to nano-ESI-LC MS/MS. A validation experiment using an independent set of plasma samples confirmed the results obtained for two of the proteins that were identified. In was concluded that patients diagnosed with colorectal cancer and adenomas demonstrate higher levels of sialylation and fucosylation compared to the normal controls in general. In this study colorectal cancer could be distinguished from adenoma and normal plasma based on elevated sialylation and fucosylation in complement C3, histidine-rich glycoprotein, and kininogen-1. These results demonstrated that *N*-linked glycan patterns can be successfully used to distinguish plasma samples originating from different clinical states. In future work further analysis of sialyl- and fucosyl- glycans needs to be done with greater emphasis on structure elucidation of the glycans. Because higher order mass spectrometry analysis required significantly higher amounts of samples, this attempt will be particularly challenging due to the limited availability of plasma samples.

Table 5.1. The amount of protein processed through the IgY antibody column and recovered in the flow-through fraction from 250 μL plasma samples.

Protein Amount (mg)	**Original Plasma**	**Flow-through 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 5.2A. 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 5.2B. 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.VIDFN*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

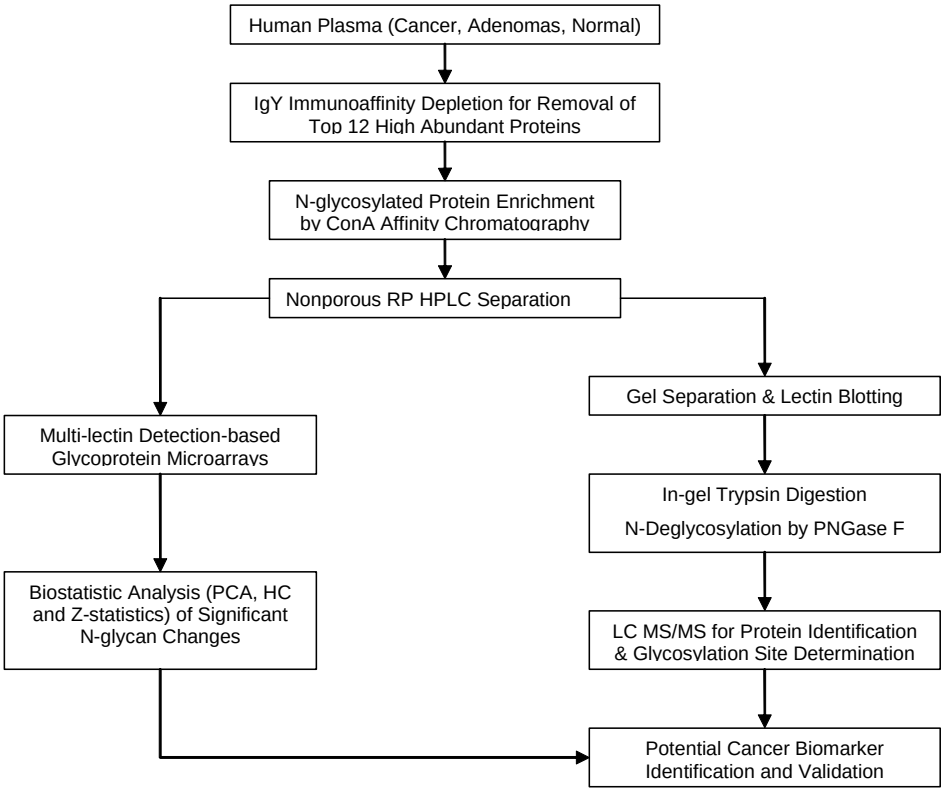
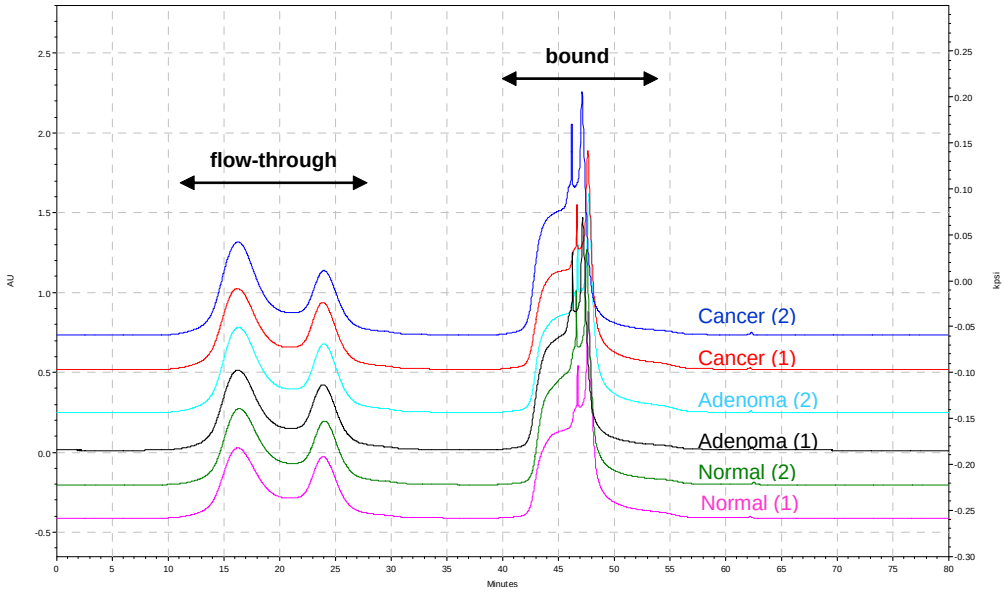


Figure 5.1: Flowchart of overall strategy using high throughput analysis of plasma N-glycosylation pattern changes in colorectal cancer.

a)



b)

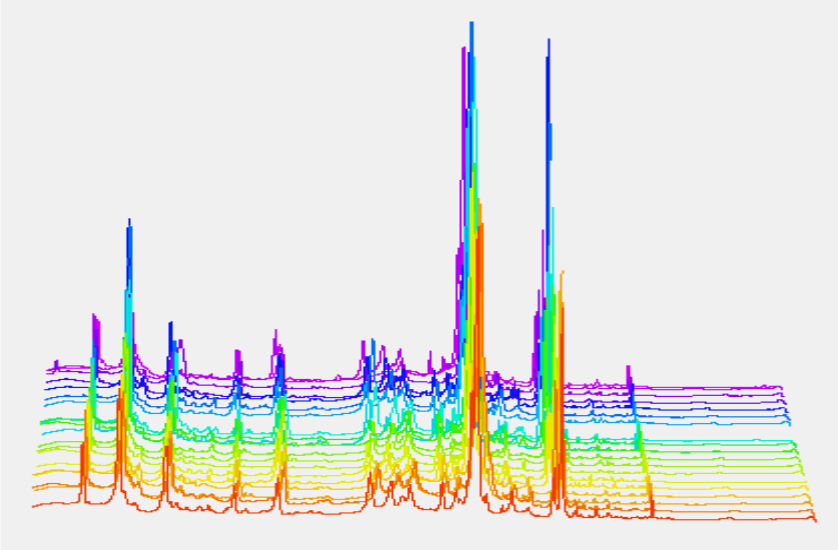


Figure 5.2: (A) Chromatographic profiles of immunoaffinity depletion of plasma from 6 normal, adenoma, and colorectal cancer patients using ProteomeLab IgY-12 kit. The 12 most abundant proteins are contained in the "bound" fraction and the less abundant proteins in plasma or serum remained in the "flow-through" fraction. (B). UV chromatograms of all plasma samples from colorectal cancer, adenoma, and normal controls.

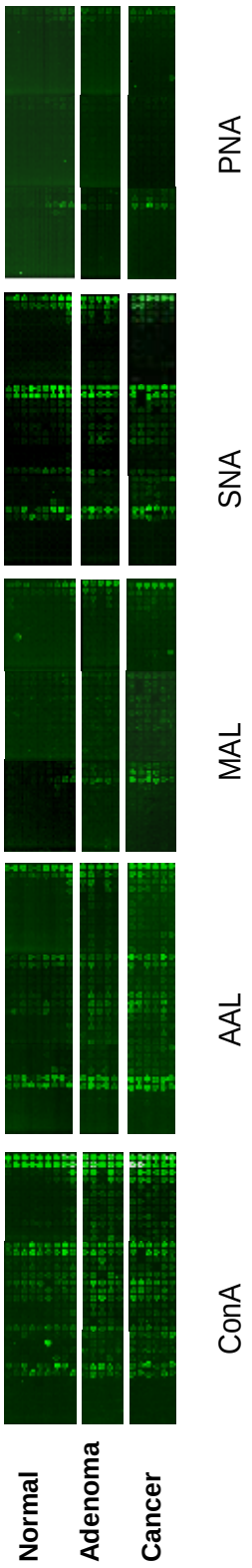
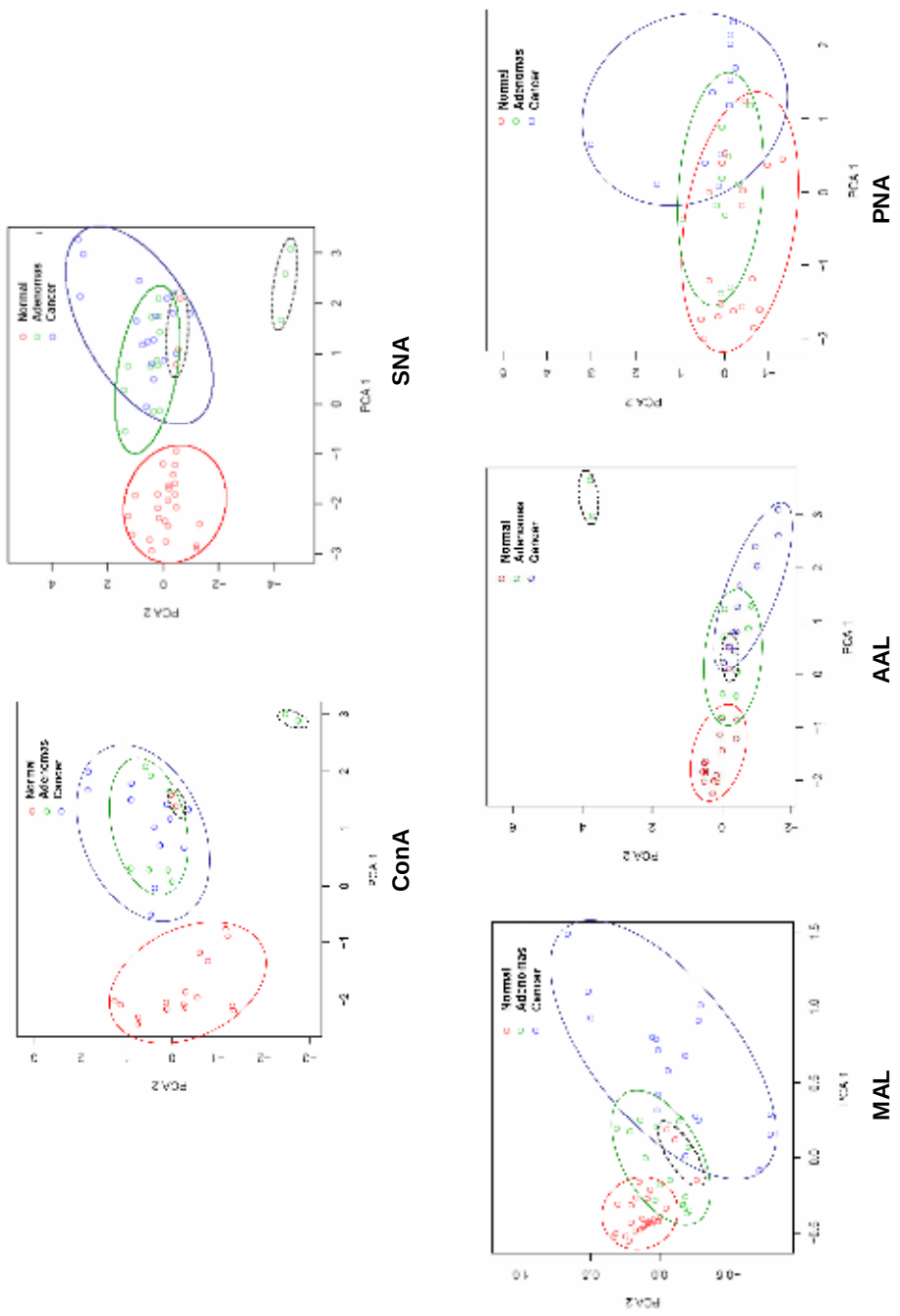


Figure 5.3: Microarray images of lectin response across all collected fractions from all sample groups

Figure 5.4:

a)



b)

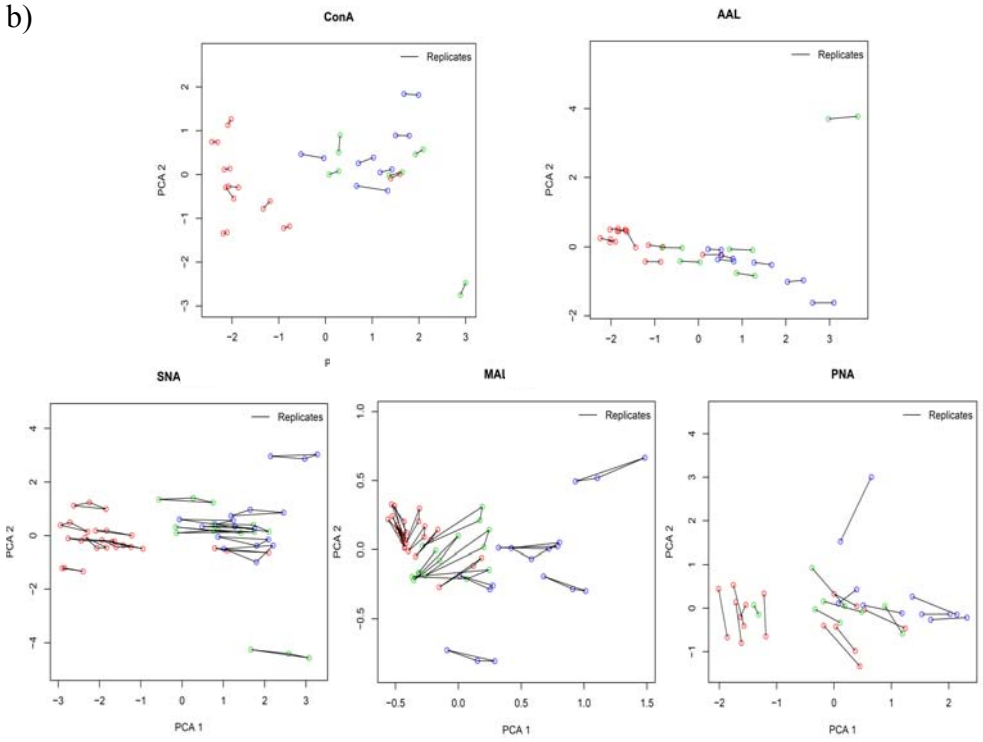


Figure 5.4: (A) Principal components analysis (PCA) plot for normalized glycoprotein microarray data derived from the replicate analysis of healthy individuals, adenoma, and colorectal cancer patient plasma. Circles indicate the areas where the data points of the three groups are clustered. (B)- Reproducibility demonstration of Principal components analysis (PCA) for normalized glycoprotein microarray data derived from the replicates of healthy individuals, adenoma, and colorectal cancer patients.

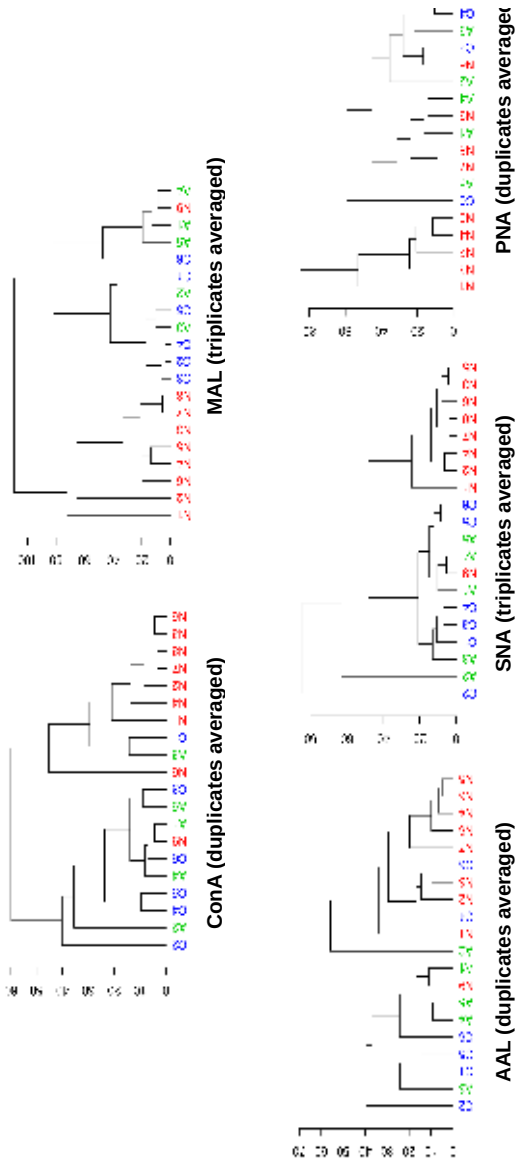
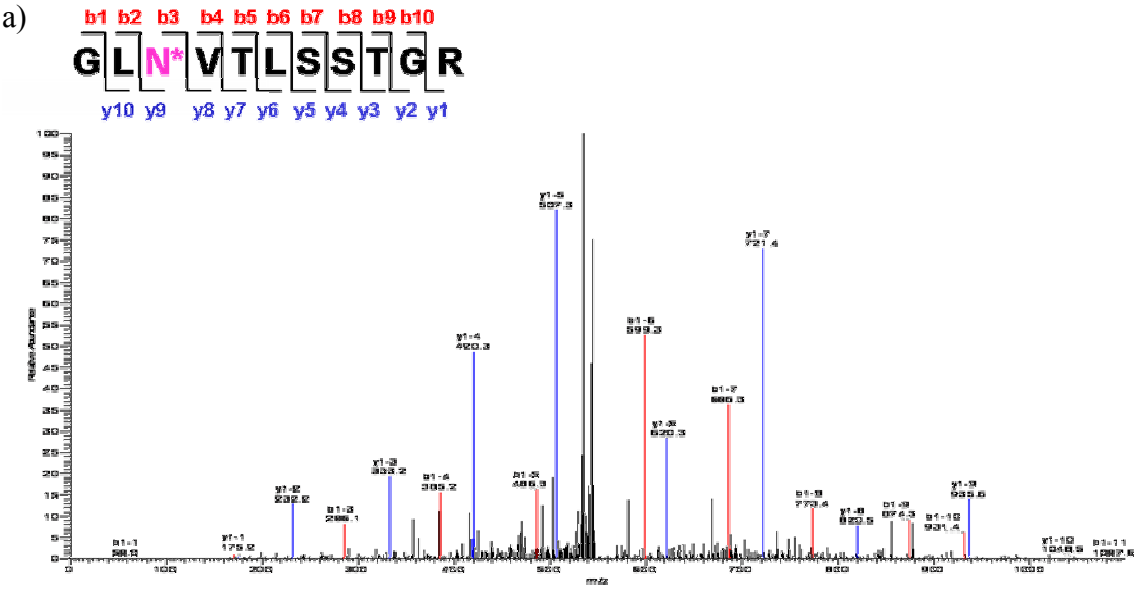


Figure 5.5: Unsupervised hierarchical clustering of glycoprotein microarray data for colorectal cancer (c1-c6) from adenoma (a1-a5) and normal controls (n1-n9). Average linkage was used, and the dissimilarity was obtained from the Pearson correlation coefficient.

a)



b)

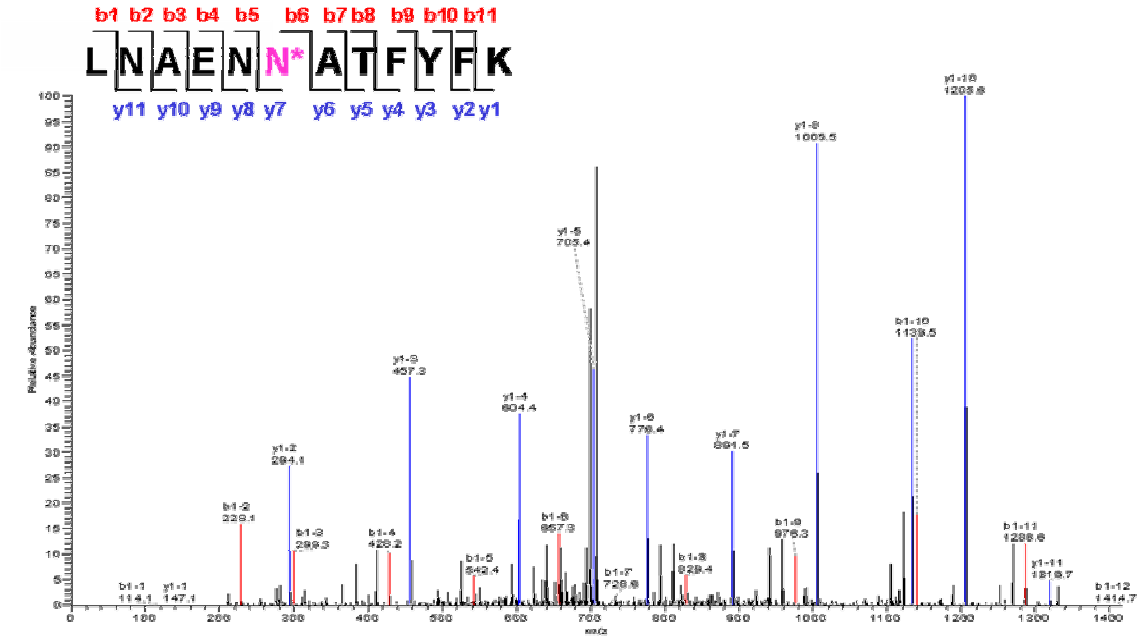
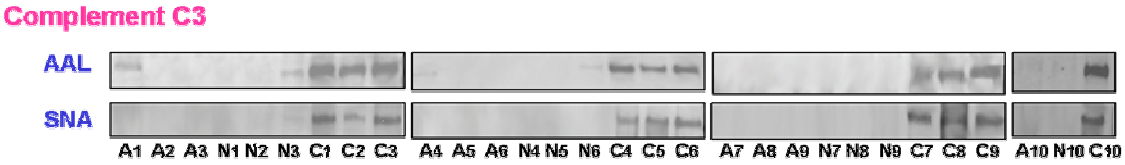
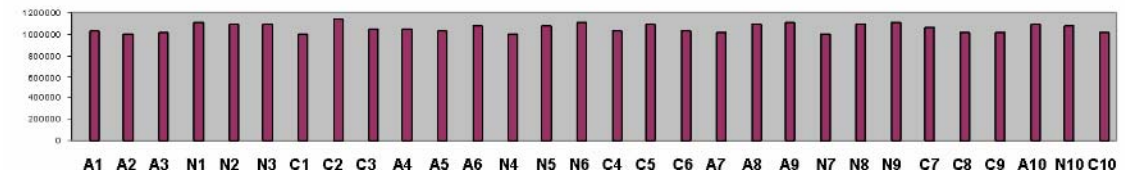


Figure 5.6: Nano LC-MS/MS spectra of (A) doubly charged N-glycosylated peptide GLN*VTLSSGH (m/z = 553.28) from complement 4 and (B) doubly charged N-glycosylated peptide LANENN*ATFYFK from kininogen-1. The asterisk (*) denotes the site of N-glycosylation determined from the tandem mass spectrum. Theoretical location of b ions is indicated by red lines. In most cases these peak intensities were not high enough for detection.

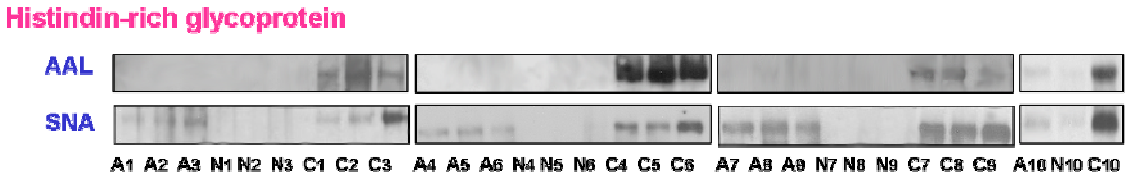
a)



b)



c)



d)

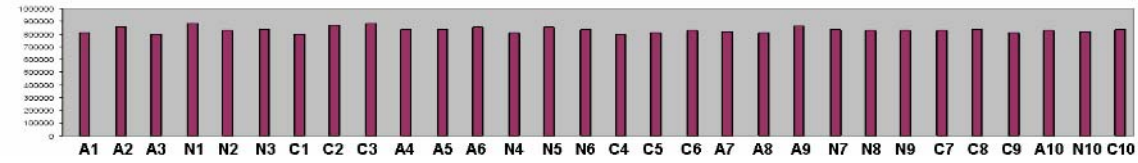


Figure 5.7: Validation study using 30 independent plasma sample to assess fucosylation and sialylation levels using AAL and SNA lectin blot analysis in complement C3 (A) and histidine-rich glycoprotein (C). The corresponding protein expression levels based on chromatogram peak areas are shown in (B) for complement C3 and (D) for histidine-rich glycoprotein.

5.5. References

[1] Parkin, D. M., Bray, F., Ferlay, J., 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] Kung, J. W., Levine, M. S., Glick, S. N., Lakhani, P., *et al.*, *Radiology* 2006, *240*, 725-735.

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

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

[7] Feldman, A. L., Espina, V., Petricoin, E. F., 3rd, Liotta, L. A., Rosenblatt, K. P., *Surgery* 2004, *135*, 243-247.

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

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

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

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

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

[13] Fletcher, R. H., *Ann Intern Med* 1986, *104*, 66-73.

[14] Duffy, M. J., *Ann Clin Biochem* 1998, *35 (Pt 3)*, 364-370.

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

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

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

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

[19] Lindmark, G., Kressner, U., Bergstrom, R., Glimelius, B., *Anticancer Res* 1996, *16*, 895-898.

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

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

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

[23] Hakomori, S., *Proc Natl Acad Sci U S A* 2002, *99*, 10231-10233.

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

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

[26] Barrabes, S., Pages-Pons, L., Radcliffe, C. M., Tabares, G., *et al.*, *Glycobiology* 2007, *17*, 388-400.

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

[28] Anderson, N. L., Anderson, N. G., *Mol Cell Proteomics* 2002, *1*, 845-867.

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

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

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

[32] Yan, W., Lee, H., Deutsch, E. W., Lazaro, C. A., *et al.*, *Mol Cell Proteomics* 2004, *3*, 1039-1041.

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

Chapter 6

A protein microarray approach exploiting the naturally occurring humoral response to identify a potential panel of biomarkers for pancreatic cancer

6.1. Introduction

Major advances in cancer control will be greatly aided by early detection so as to diagnose and treat cancer while it is in an early, curable state. Unfortunately, for pancreatic adenocarcinoma (PDAC), the fourth leading cause of cancer death in the United States[1], effective early detection and screening are not currently available and tumors are typically diagnosed at a late stage, frequently after metastasis. PDAC is generally considered to be largely incurable by available treatment modalities, with a 5-year survival rate of less than 4 percent. Existing biomarkers for this disease are inadequate.[2] CA19-9 has been tested for its utility as an early detection marker in PDAC,[2-5] however, the sensitivity and specificity of this biomarker are not high, and serum levels are significantly increased in inflammatory diseases of the pancreas and biliary tract. Therefore, CA19-9 is not useful for early diagnosis, mass screening, distinguishing between PDAC and chronic pancreatitis, or the targeting of therapeutics. Thus, there is a great need for new biomarkers for PDAC. In the absence of good biomarkers, 80% to 90% of PDAC cases are diagnosed too late in the disease process for

surgical resection to be an effective option. Even among the 10% to 20% of PDAC cases where surgical resection is an option, most patients ultimately die of recurrent or metastatic disease.[6] Identification of novel biomarkers for pancreatic adenocarcinoma may have utility for the detection of this malignancy.

A humoral response to cancer in humans has been well demonstrated by identification of autoantibodies to a number of different intracellular and surface antigens in patients with various tumor types[7-13]. Tumor-specific humoral responses directed against oncoproteins[14, 15], mutated proteins such as p53[16, 17] or other aberrantly expressed proteins have all been described. While it is currently unknown whether the occurrence of such antibodies is beneficial to the patient, knowledge of potential tumor antigens that can evoke tumor-specific immune responses may have utility in cancer diagnosis, in establishing prognosis and in targeted immunotherapy against the disease. In PDAC, autoimmunity has been shown against a number of cellular proteins (or protein isoforms), including MUC1,[18, 19] p53,[17] Rad51,[20] DEAD-box protein 48,[21] two distinct isoforms of calreticulin[22] and one isoform of vimentin.[23] However, in most cases, autoantibodies to specific proteins occur in less than 50% of patient's sera. Therefore, they may not be effective individually for the early detection of PDAC, but rather may have utility as part of a panel.[24]

The strategy of using liquid-based multi-dimensional procedures to separate proteins allows distinct protein containing fractions to be arrayed and interrogated using various types of probes. We have utilized methodology that first employs separation of cell and/or tissue lysates by chromatofocusing, followed by liquid phase separation by nonporous silica reversed phase HPLC (according to hydrophobicity). Thus, a large

number of proteins can be resolved using a liquid-based system. Importantly, liquid-based protein separations are well suited for fractionation of lysates into individual protein fractions, or for purification of individual proteins. Additionally, the separated proteins are maintained in solution, thus facilitating intact protein identification by mass spectrometry and the spotting of individual fractions on protein microarrays with a robotic arrayer. Protein microarrays have been utilized to assess the binding characteristics of multiple samples (probes) simultaneously.[25, 26]

In particular, protein microarrays consisting of arrayed proteins derived from cell line or tumor lysates can be utilized to identify those that have elicited a humoral response.[26] In this study proteins from a pancreatic adenocarcinoma cell line (MIAPACA) were resolved by two-dimensional liquid-based separations, and were then arrayed on nitrocellulose slides. The slides were probed individually with sera from 15 patients diagnosed with pancreatic cancer and 15 normal subjects. The resulting data were analyzed using a rank-based non-parametric test and a z-score to determine humoral response signatures of pancreatic cancer. The PAM (Prediction Analysis for Microarrays) classification algorithm [27] was used to explore the classificatory power of the proteins found to be differential between control and cancer sera. The generalization error of our classification analysis was estimated using leave-one-out cross-validation. From this it was found that if generalized to a new population the classification analysis should predict the serum diagnosis with 86.7% accuracy (4 misclassified samples). Among the 4 misclassified samples, 3 were false positives and only 1 was a false negative resulting in an expected sensitivity of 93.3% and an expected specificity of 80%. Furthermore, recombinant proteins were used to conduct a validation study on some of the proteins

identified and with a separate set of serum samples. The proteins highlighted in this study may have utility as candidate markers of pancreatic cancer.

Protein microarrays for humoral response present analysis issues not present in other microarray platforms. In particular, typically only a subset of patients with a particular tumor type develops a humoral response to a particular antigen, thus resulting in a great amount of variability in the multiple cancer samples analyzed within an experiment. A simple two-sample t-test assesses differences between the average response of the normal and diseased states. However, this will only detect differences resulting from an immune response in a majority of diseased state samples and almost no response in control samples or vice versa. Thus, we used a rank-based test to compare the response between normal and diseased sera. Rank-based tests look for differences in the shape of the distributions which will help identify proteins that are changed in only a few samples. Additionally, in microarray studies the utility of background subtraction has been questioned.[28] In the following, differences in results when using foreground measures alone compared to using local background subtracted measures are also discussed.

6.2. Experimental Section

Sample preparation

Sera-

Fifteen serum samples were obtained from patients with a confirmed diagnosis of pancreatic adenocarcinoma who were seen in the Multidisciplinary Pancreatic Tumor Clinic at the University of Michigan Comprehensive Cancer Center. Sera from the

pancreatic cancer patients were randomly selected from a clinic population that sees, on average, at the time of initial diagnosis, 15% of pancreatic adenocarcinoma patients presenting with early stage (i.e., stage I/II) disease and 85% presenting with advanced stage (i.e., stage III/IV). All sera samples selected for this study were stages III/IV. Inclusion criteria for the study included patients with a confirmed diagnosis of pancreatic cancer, the ability to provide written, informed consent, and the ability to provide 40 ml of blood. Exclusion criteria included inability to provide informed consent, patients' actively undergoing chemotherapy or radiation therapy for pancreatic cancer, and patients with other malignancies diagnosed or treated within the last 5 years. The mean age of the tumor group was 65.4 years (range 54-74 years). The sera from the normal subject group was age and sex-matched to the tumor group. All sera were processed using identical procedures. The samples were permitted to sit at room temperature for a minimum of 30 minutes (and a maximum of 60 minutes) to allow the clot to form in the red top tubes, and then centrifuged at 1,300 x g at 4^{o}C for 20 minutes. The serum was removed, transferred to a polypropylene, capped tube in 1 ml aliquots, and frozen. The frozen samples were stored at -70^{o}C until assayed. All serum samples were labeled with a unique identifier to protect the confidentiality of the patient. None of the samples were thawed more than twice before analysis.

Cell culture and lysis + tissue lysis-

The cells used in this work were from the pancreatic cancer cell line, MIAPACA. The cells were cultured at 37^{o}C in a 5% CO_2-humidified incubator in DMEM growth medium supplemented with 10% fetal bovine serum (FBS) and 1% penicillin/streptomycin (Invitrogen, Carlsbad, CA). When the cells reached ~90% confluence, they were

harvested with a cell scraper and lysed in lysis buffer containing 7 M urea, 2 M thiourea, 100 mM DTT, 2% n-octyl-D-glucopyranoside (OG), 10% glycerol, 10 mM sodium orthovanadate, 10 mM sodium fluoride (all from Sigma, St. Louis, MO), 0.5% Biolyte ampholyte (Bio-Rad, Hercules, CA), and protease inhibitor cocktail (Roche Diagnostics, GmBH, Mannheim, Germany). The lysed cells were centrifuged at 35,000 rpm for 1 hr and were then buffer exchanged into start buffer (6 M urea, 25 mM Bis-Tris, and 0.2% OG) using a PD-10 G-25 column (Amersham Biosciences, Piscataway, NJ) and stored at -80^{o}C until further use. Tissue lysis followed the same protocol as the cell lysis. However the tissue was first cut into many small pieces using a razor blade and it was vortexed vigorously using a bead beater. The beads and resulting tissue mixture were centrifuged and supernatant containing tissue proteins and cellular debris was removed and transferred into another centrifuge tube which was further centrifuged at 35,000 rpm for 1hr at 4^{o}C. The resulting supernatant was buffer exchanged into start buffer and stored for further analysis.

Chromatofocusing (CF)-

Prior to chromatofocusing the extracted protein content from the cell line or tissue sample were assayed using a Bradford protein assay kit, using bovine serum albumin (Bio-Rad) as the standard protein. Chromatofocusing was performed using a Beckman System Gold model 127 pump and 166 UV detector module (Beckman Coulter, Fullerton, CA). A start and elution buffer combination was used to separate lysate proteins according to their pI by generating a linear pH gradient. The start buffer (SB) was composed of 6 M urea, 25 mM Bis-Tris and 0.2% OG (pH 9.0). The elution buffer (EB) contained 6 M urea, 0.2% OG, and 10% combination of Polybuffer 74 and 96 (pH 3.9; Amersham Biosciences).

Saturated iminodiacetic acid (Sigma) was used to adjust the pH of both buffers. A weak anion exchange HPCF-1D prep column (250 mm L x 4.6 mm ID, Eprogen, Darien, IL) was initially equilibrated with start buffer (at 1 mL/min) until a stable baseline was observed. The sample was injected with multiple injections at a low flow rate of 0.5 mL/min to ensure maximum interaction of protein with ion exchange resin. Once a stable baseline was achieved, the solvent flow was increased to 1 mL/min and the mobile phase was switched to EB. Fractions were collected at 0.2 or 0.3 pH unit intervals. The CF profile was monitored at 280 nm wavelength. The pH was monitored with a post detector online pH-flow cell (Lazar Research Laboratories, Los Angeles, CA). After the CF gradient run was completed, the column was flushed with a 1M NaCl solution, followed by deionized water. Finally, the column was flushed with isopropanol and stored in the same until further use. The collected fractions were stored at -80^{o}C until further use. In the case of the tissue proteins the separation pH range was 7.2-4.0 instead of 9.0-4.0.

Non-porous reversed-phase high performance liquid chromatography (NPS-RP-HPLC)-

Each fraction from the first dimension chromatofocusing was further separated in the second dimension by NPS-RP-HPLC, according to protein hydrophobicity. An ODSIII-E (8 x 33mm) column (Eprogen, Inc., Darien, IL) packed with 1.5 μm non-porous silica was used to achieve high separation efficiency. A 0.1% TFA with water (A) and 0.08% TFA with acetonitrile (B) gradient was used in the separation. The following gradient was applied at a flow rate of 1 mL/min and fractions were collected by peak using an automated fraction collector (model SC 100; Beckman-Coulter) in 96-well plates: 5-15% B in 1 min, 15-25% B in 2 min, 25-31% B in 3 min, 31-41% B in 10 min, 41-47% B in 3

min, 47-67% B in 4 min, 67-100% B in 1 min, followed by maintaining the system at 100% B for 3 min. All separations were performed at 60^{o}C and were monitored at 214 nm. All 2D fractions were stored at -80^{o}C until further use.

Microarray printing-

Approximately 30% of the fractionated proteins were transferred to 96-well printing plates (Bio-Rad) and were completely dried using a speedvac concentrator at 60^{o}C. The fractions were then resuspended in printing buffer (62.5 mM Tris-HCl (pH6.8), 1% w/v sodium dodecyl sulfate (SDS), 5% w/v dithiothreitol (DTT) and 1% glycerol in 1X PBS) and were left to shake overnight at 4^{o}C. Slides were printed by transferring each fraction from the plate onto nitrocellulose slides using a non-contact piezoelectric printer (Nanoplotter 2, GeSiM). Each spot resulted from deposition of 5 spotting events of 500 pL each, such that a total volume of 2.5 nL of each fraction was spotted. Each spot was found to be ~450 μm in diameter, with the distance between spots maintained at 600 μm. Printed slides were left on the printer deck overnight to dry and were then stored in a desiccator at 4^{o}C until further use.

Hybridization of slides-

The printed arrays were rehydrated in 1X PBS with 0.1% Tween-20 (PBS-T), and were then blocked overnight in a solution of 1% BSA in PBS-T. Each serum sample was diluted 1:400 in probe buffer (5 mM magnesium chloride, 0.5 mM DTT, 0.05% Triton X-100, 5% glycerol and 1% BSA in 1X PBS) to make a total solution of 4 mL and kept on ice. The slides were hybridized in diluted serum for 2 hrs (1 serum sample per slide). Hybridization was done at 4^{o}C in heat-sealable pouches with agitation, using a mini-rotator. The slides were then washed five times with probe buffer (5 min each), and were

then hybridized with 4 mL goat anti-human IgG conjugated with Alexafluor647 (Invitrogen, Carlsbad, CA) (at 1 μg/mL in probe buffer), for 1 hr at 4^{o}C. After secondary incubation all slides were washed in probe buffer five times, for 5 min each, and were then dried by centrifugation for 10 min. All processed slides were immediately scanned using an Axon 4000B microarray scanner (Axon Instruments Inc., Foster City, CA).

Data acquisition and analysis-

GenePix 6.0 software was used to grid all spots, to determine the median Cy5 single-channel intensities and median local background intensities for each spot. A spot was considered positive if the foreground measure was at least 2X the background intensity measure. Both the foreground data alone as well as the background-subtracted data were considered for analysis. To account for variation between arrays, each array was median-centered and scaled by its interquartile range. After standardization the replicate arrays were averaged. To assess differences between humoral response in cancer and normal sera, the non-parametric Wilcoxon rank-sum test was employed and results patterns were visually assessed to determine if background subtraction was beneficial for this data analysis. Additionally z-score statistics were used on the foreground data to look for subtle differences between the two sera groups. Finally, a classifier was built from the differential proteins found by these methods.

Non-parametric method- A two-sample Wilcoxon rank sum test between cancer and benign sera was run for each spot on the array. Each pH/fraction combination was tested and the p-values were visualized in a grid plot to highlight regions of spots that exhibited differential response between normal and cancer sera. A p-value threshold of 0.05 was used to determine differential proteins for further study.

Z-score method- The standardized data was log transformed after adding a small value to each point to eliminate negative values. Z-score measures were constructed for each spot by subtracting the mean and dividing by the standard deviation of only the control serum samples for that spot. Resulting z-scores were then on the scale of standard deviations from the mean of the control samples. Proteins that had Z-scores of >2 (or < -2) in 20% of the cancer serum samples were determined to be differential and considered for further study.

Prediction Analysis for Microarrays (PAM) classification algorithm and Leave One Out Cross Validation (LOOCV)- The PAM classification algorithm [27], as implemented in R, was used to explore the classificatory power of the proteins found to be differential between control and cancer sera using either the Wilcoxon rank-sum test or the z-score method. From PAM the smallest subset of proteins that gave the lowest error rate were chosen to be used as a classifier. An ROC curve was drawn to illustrate the selection of this 'best' subset of proteins and the area under the curve (AUC) was estimated.

The generalizability of the PAM analysis was estimated using leave-one-out cross-validation (LOOCV) in which each sample was left out of the data set in turn and classified using the remaining samples. Specifically, using only the 29 remaining samples, the same analysis scheme as done above for the full set of 30 samples was repeated, including reselection of differential proteins using Wilcoxon tests and z-scores from the 29 samples[1] and classifier selection using PAM. The resulting classifier was then used to predict the diagnosis of the excluded sample. Each of the 30 samples were predicted in this way and error rates were estimated. ROC curves were drawn to illustrate

[1] The median number of differential proteins across the 30 leave-one-out datasets was 96 (range=[65,109]).

the selection of the 'best' protein subset for classification in each of the 30 leave-one-out cycles and the AUC was estimated.

Heatmaps- Heatmaps were drawn using Cluster and TreeView software.[29] Spots were median-centered across samples and average linkage clustering was used.

Protein identification – nanoLC-ESI-MS/MS-

Individual protein fractions were first dried down to ~10 μL, and then mixed with 40 μL 100 mM ammonium bicarbonate, 10 μL 20mM DTT and 0.5 μL of sequence grade modified trypsin (Promega). The mixture was allowed to incubate at 37^{o}C overnight with agitation after which the digest was stopped by addition of 1 μL TFA (Baker and Baker). The digested sample was loaded on a paradigm desalting column (C18, 5mm x 300 um, Michrom), and was washed at a flow of 50 μL/min using HPLC grade water containing 0.3% formic acid. The desalted peptides were directly eluted onto an analytical column (100 μm x 15cm, C18, Michrom) using a flow rate of 300 nL/min and the following gradient profile, where solvent A was water/0.3% formic acid and solvent B was acetonitrile/0.3% formic acid. The gradient was started at 3% B, ramped to 35% B in 25 min, 60% B in 15 min, 90% in 1 min, maintained at 90% B for 1 min and finally ramped back down to 3% in another 1 min. The eluting peptides were analyzed on a linear ion-trap based mass spectrometer (LTQ, Thermo, San Jose, CA) with an NANO-ESI platform (Michrom Biosciences). The capillary temperature was set at 200^{o}C, the spray voltage was 2.6 kV, and the capillary voltage was 20 V. The normalized collision energy was set at 35% for MS/MS. The top 5 peaks were selected for collision induced dissociation (CID). MS/MS spectra were interrogated using the SEQUEST algorithm in Bioworks software (Thermo) against the SwissProt human protein database. Two missed

cleavages were allowed during the database search. The search threshold was set to 1000 and tolerances were set at 1.4 and 0.00 for peptide and fragment ion tolerances respectively. Protein identification was considered positive when a peptide showed an X_{corr} of greater than or equal to 3.0, 2.5 and 1.9 for triply, doubly and singly charged ions respectively. Only proteins with greater than 10% coverage were considered in the analysis and a minimum of 3 good-scoring peptides were required for positive identification. In the event that more than one protein was found in a fraction, the data was filtered. If the spot of interest was unique and did not lie between adjacent reactive spots then only the highest scoring protein that was not found in adjacent fractions in the separation profile were considered true hits, since the adjacent fractions did not elicit a humoral response. On the other hand if the spot of interest was part of a group of spots eliciting a positive response within a separation profile, the common protein identified in all these spots was considered a true hit.

6.3. Results and Discussion

Experimental Scheme

We spotted native proteins derived from the MIAPACA pancreatic cancer cell line or pancreatic cancer tissue on protein microarrays to characterize a pancreatic cancer-specific humoral response. Such a study has potential utility in identification of novel potential candidate markers of pancreatic cancer. Figure 6.1 illustrates schematically the methodology employed within this study. Proteins from the MIAPACA pancreatic adenocarcinoma cell line or pancreatic cancer tissue were first solubilized, and then separated using two-dimensional liquid-based separation that employs

chromatofocusing (separation according to protein pI) in the first dimension and non-porous reversed-phase HPLC (separation according to protein hydrophobicity) in the second dimension. The MIAPACA cell line was used because it was readily available. Pancreatic cancer tissue was run in order to see if results with the MIAPACA cell line would correlate with the clinically relevant tissue sample. The separated proteins were then arrayed on nitrocellulose slides using non-contact piezoelectric printing. Following printing, slides were hybridized with serum from patients diagnosed with pancreatic cancer or normal subjects. Spots on the slides were statistically evaluated using non-parametric statistical methods to identify proteins that elicit a pancreatic cancer-specific humoral response. It was found that the omission of background subtraction during data acquisition is critical in the identification of a differential humoral response. Furthermore it was seen that tissue samples did not provide results as clear as the MIAPACA cell lines. Reasons for this will be discussed shortly. Proteins that elicited a statistically significant humoral response difference were subjected to classification analysis to obtain a panel of classifiers which were subsequently identified by nano-LC-linear ion trap mass spectrometry. For select identified proteins, a validation study using a separate set of serum samples was attempted where the recombinant protein was arrayed on nitrocellulose slides and probed with serum from a separate cohort of normal and pancreatic cancer patients.

2-dimensional liquid separation

MIAPACA proteins were separated by CF from pH 9.2-3.9, and each CF fraction was subsequently further separated by NPS-RP-HPLC. In the case of the tissue samples separation in the first dimension was restricted to 7.2-4.0 only. Figure 6.2 represents the

2D UV chromatogram from these separations. Along the horizontal axis are all fractions from the first dimension from lower pI to higher pI. The vertical axis represents retention times from the 2nd dimension separations and going up the axis corresponds to increased hydrophobicity. A typical 2D separation across the pH range above results in about 1300 total fractions in the MIAPACA cell line. A majority of these fractions are relatively pure since manual collection by peak is performed. However there are instances when more than one protein is present in the peak particularly for more highly abundant proteins that elute over a longer time. Furthermore it can be seen that the signal intensities of fractions in the MIAPACA separation is considerably higher compared to the tissue samples. This is because the cell line could be easily obtained but tissue samples were more limited in nature and therefore a smaller amount of total protein was available for the 2D separation.

Microarray printing and processing

The separated proteins were printed on nitrocellulose slides and probed with serum from normal individuals and patients diagnosed with pancreatic cancer. The immune response in the sera was visualized using an antihuman-IgG –Alexaflor647 conjugate. Figure 6.3 illustrates portions of the arrays from the MIAPACA cell line printed on nitrocellulose slides to indicate the typical appearance of slides and spot quality, with specific examples of differential humoral response. Tandem mass spectra are also shown to indicate the protein identity present in the spot of interest. It can be seen that spot intensities appear homogenous throughout the spot. However, it was found that some fractions from the separated MIAPACA lysates were not printed on all the nitrocellulose pads due to incorrect calibration of the printing surface and printing errors

that occurred during the print run for the lower pH fractions. Thus, all subsequent data representations indicate these missed spots which were not considered further in the statistical analysis.

Statistical analysis of immune response in MIAPACA cell lines

While looking at the humoral response from one normal and one cancer serum sample may indicate a difference as shown in Figure 6.3, it is critical to assess this response in an adequately sized set of normal and cancer sera to see if the difference is indeed statistically significant. Two analysis approaches were used to analyze the humoral response differences in 15 control and 15 pancreatic cancer sera. The first approach utilized a non-parametric Wilcoxon rank sum test which was repeated using both the locally-derived background-subtracted median spot intensities as well as foreground median intensities without background-subtraction. Results showed that while background subtraction reduced batch effects between slides, a large amount of signal is washed away by background correction. Figures 6.4a and 6.4b show the grid of p-values from the per spot Wilcoxon rank-sum tests between cancer and normal sera. The grid is arranged according to the two dimensional fractionation of the whole cell lysate and colored according to the level of significance and the direction of the difference between cancer and normal sera. Figure 6.4a is the p-value grid from the foreground only analysis and the background subtracted p-value grid is shown in Figure 6.4b. Given the method of fractionation it was expected that in some cases neighboring fractions containing higher abundant reactive proteins would be correlated, thus producing hot (cold) regions. These hot-spots (cold-spots), seen in the foreground-only p-value grid, are missing in the

background-subtracted p-value grid. Thus, it was found that the uncorrected measures were preferable.

In the pancreatic cancer data set, uniform increases or decreases across all cancer samples were not expected. We sought to identify those proteins that elicited a pancreatic cancer-specific humoral response in as few as 20% of the samples or greater. For an alternate view of the changes in the immune response between healthy and cancer diagnosed patients, Z-score plots of each studied pH range were also generated in which z-scores were calculated, per spot, using the mean and standard deviation of only the normal samples. Resulting z-scores were thus on a scale of standard deviations from the mean of the normal samples. Thus, if a fraction had a high z-score it had well above the average normal reactivity at that spot. Likewise, a low z-score indicated that the fraction had well below the normal reactivity. When plotted in grids of spot by sample, patterns could be easily discerned in cancer samples. An example of such a z-score grid is illustrated in Fig 6.4c, where the multiple orange/red fractions across the cancer samples but not control samples is indicative of a protein of interest. Increases or decreases that persist across at least 20% of the cancer samples were pursued for further study.

Comparing statistical analysis results from cell line to tissue samples

Figure 6.4d illustrates the wilcoxon rank sum data for humoral response differences between the normal and pancreatic cancer sera across all printed fractions from the pancreatic cancer tissue samples using the foreground only measures for spot intensities. When comparing this grid diagram to that obtained for the humoral response differences in the MIAPACA cell line fractions (Fig 6.4a) it can be seen that almost no difference of significance is observed with tissue sample data.

This lack of successful results can be attributed to multiple factors. The tissue samples that were analyzed may not have been homogeneous. i.e. all the cells in the tissue sample may not necessarily been cancerous. It is therefore very likely that from the total proteins extracted and separated, only a small percentage were actual cancerous tissue proteins. These proteins could have therefore been below the detection limits of microarray experiments. Such a problem can be alleviated by utilizing a more sophisticated approach toward cancer tissue collection. Laser Capture Micro-dissection (LCM) is capable of extracting very thin layers of cells from tissue samples. If such a technique is utilized to carefully select only cancer cells from available tissue samples a more concentrated batch of cancer-only cells could be obtained to facilitated a more focused humoral response analysis. However such a study would require pooling of multiple samples of cancer cells from different tissue samples to reach the amount of sample that is needed to perform a complete humoral response experiment as outlined in this chapter. Because results with pancreatic cancer tissue were not successful, all subsequent cross validation and classification algorithms were performed on data from the MIAPACA cell line.

Classification and Cross Validation for MIAPACA fractions

The PAM (Prediction Analysis for Microarrays) classification algorithm [27] was used to explore the classificatory power of the proteins from the MIAPACA cell line found to be differential between control and cancer sera. Differential proteins were selected as having (1) a Wilcoxon p-value of 0.05 or less or (2) having over 20% of the cancer samples with a z-score >2 (or <-2). The PAM algorithm selects the most predictive subset of proteins for classification. The best classifier, resulting in the smallest error using the fewest proteins, used 9 proteins, chosen from 98 differential

proteins, and only misclassified 4 samples. The ROC curve shown in figure 6.5a shows the true positive and false positive classification rates associated with this fit. The red points indicate the 30 threshold values considered by PAM, corresponding to 30 subsets of the proteins. The blue circle highlights the chosen threshold which uses only 9 proteins for classification. The area under the curve (AUC) for this ROC curve was estimated to be 0.85.

In an effort to estimate the generalizability of the classification analysis, leave-one-out cross-validation (LOOCV) was used. For the 30 leave-one-out cycles, the median size of the of protein subset chosen for the classifier was 12 proteins (range=[4,83]) which resulted in a median error rate of 4 (range=[2,6]) and an average AUC of 0.82 (range=[0.63,0.96]) for classifier selection. This is comparable to what we found when using all 30 samples.

From predictions of the left out sample, it was found that if generalized to a new population our classification analysis should predict the serum diagnosis with 86.7% accuracy (4 misclassified samples). Among these 4 misclassified samples, 3 were false positives and only 1 was a false negative. This gives an expected sensitivity of 93.3% and an expected specificity of 80%.

We examined how frequently each protein was selected as an important predictor across the 30 LOOCV classifiers built. Two proteins (PH 6.6-6.4, fraction 44 and PH 8.1-7.8, fraction 56) were selected in all 30 LOOCV classifiers. Four other proteins were selected 22 times (PH 6.6-6.4, fraction 38; PH 6.6-6.4, fraction 43; PH 6.6-6.4, fraction 46; PH 7.8-7.5, fraction 42). It is interesting to note that the 9 protein spots selected

initially are among the most common proteins used in the LOOCV classifiers; see table 6.1. Figure 6.5b illustrates the response of all serum groups to these nine proteins.

Figure 6.6 shows the scaled humoral response distribution across all serum samples considered to be differential on a scale of green to red (lowest response to highest response), based upon data from the Wilcoxon tests and z-score plots combined. The 9 proteins spots that comprised the best classifier are indicated by arrows.

Identification and implications of statistically significant proteins identified from MIAPACA cell line

Studying the humoral response to pancreatic cancer has utility to identify potential tumor antigens. These tumor antigens appear as reactive spots on the protein microarray since they bind autoantibodies present in the serum against which the array was hybridized. Since these spots have been identified as specific proteins, important pathway changes as well as key players involved in these changes might be highlighted.

Microarrays printed with fractionated lysates from the MIAPACA cell lines were probed with 15 normal sera and 15 sera from patients diagnosed with pancreatic cancer. Statistical treatment using the Wilcoxon rank-sum statistics, z-score plots and classification analysis of the humoral response to pancreatic cancer cells yielded a panel of 9 spots that showed best specificity and sensitivity in their ability to identify normal and cancer sera correctly. The protein IDs of these are detailed in Table 6.1. In addition, the percentage of cancer samples in which the panel was able distinguish from normal sera is also indicated.

Higher Reactivity in Normal Sera

We found that the humoral response to proteins that are known to be involved in stress was lower in cancer patient sera compared to normal sera. Heat Shock protein beta 1 is a 27 kDa heat shock protein that is expressed in response to environmental stresses or estrogen stimulation. The protein colocalizes with mitotic spindles in dividing cells and is also known to migrate to the nucleus during heat shock. It is also known to be a chaperone protein that inhibits apoptosis and prevents aggregation of actin intermediate filaments.[30] In pancreatic cancer, often characterized by a large amount of inflammation, we found that serum humoral response to HSPB1 was significantly lower than was found in normal sera. While the decreased HSPB1-specific humoral response in pancreatic cancer has not been reported previously, we hypothesize that HSPB1 autoantibodies are detectable in normal serum due to the presence of some naturally occurring process that is disrupted in pancreatic cancer.

Two variants of histone H2, H2A type 1-B and H2A.a, both showed a higher humoral response in normal sera than was apparent in pancreatic cancer sera. A variant of histone has previously been implicated in DNA break repair mechanisms.[31] Such breaks are often caused by external stimuli such as radiation or internal events such as oxidative damage that cause breaks in double-strand DNA. Lack of a humoral response to such proteins in pancreatic cancer suggests that pancreatic cancer patients may be unable to form autoantibodies to proteins that are potentially involved in stress relief mechanisms.

Finally, a key protein, pyruvate kinase, involved in glycolysis was found to elicit higher humoral responses in normal sera as compared to cancer sera. Expression of glycolytic enzymes has been shown to be increased in a variety of cancers, including

pancreatic cancer tissues.[32] However, a humoral response to this enzyme has not been previously reported. Regulator of chromatin condensation and ubiquitin were also more reactive with normal sera compared to cancer sera. Both proteins have been implicated in stress response although the exact mechanism of action is still not understood.

Higher Reactivity in Cancer Sera

A panel of 3 proteins was found to discriminate pancreatic cancer sera with high sensitivity and specificity from normal sera by generating a higher response in cancer samples. Our efforts identified two of these 3 proteins. Phosphoglycerate kinase 1 is a glycolytic enzyme but is also known to be active as a primer recognition protein. PGK1 is known to show antigen activity in other types of cancers[33]. Histone H4 is a nuclear protein that maintains DNA in its proper configuration. As mentioned earlier, certain variants of histones have been implicated in the DNA repair process. Presence of antibodies against histone H4 in cancer sera but not in normal sera may serve as an important indicator of improper DNA regulatory mechanisms in cancer patients.

While none of the proteins discussed were individually able to discriminate clearly between the two clinical groups, used together, as a 9 protein panel, they showed high specificity, sensitivity and selectivity, and may have potential diagnostic utility in the identification of patients with pancreatic cancer. Further validation studies using adequately sized test and trial sets of patient sera with pancreatic cancer would be required for this determination.

Validation using recombinant proteins

For some of the proteins identified as eliciting differential humoral response in pancreatic cancer, we were able to obtain recombinant proteins for further validation

studies using a separate set of 18 normal and 18 cancer sera. Four proteins were used, Phosphoglycerate Kinase (PGK-1), Histone H4, Heat Shock Protein (HSP27) and Pterin-4-alpha-carbinolamine dehydratase. Recombinant proteins were arrayed on nitrocellulose slides and the slides were then probed with samples of 18 normal and 18 cancer sera. In the case of PGK-1 and Histone H4 a differential response similar to that observed in the test set was seen where cancer sera showed an overall higher humoral response compared to the normal sera (Fig. 6.7). On the other hand the Pterin carbinolamine dehydratase and HSP27 did not show a differential humoral response similar to the test set. One possible reason for this lack of differential response could be the nature of the recombinant proteins that were arrayed. It is quite possible that the recombinant protein synthesized in bacteria did not posses key modifications responsible for the antigenicity of the endogenous proteins.

The validation studies showed that PGK-1 and Histone H4 do in fact differentiate normal and cancer sera. However, because the response is not "absent in normal and present in cancer" and there is some overlap with each individual marker, these proteins are not suitable as single biomarkers for diagnostic purposes. However their ability to distinguish normal vs. cancer sera provides important information about possible mechanisms of pancreatic cancer progression and can potentially be used to monitor therapeutic response to the disease.

6.4. Conclusion

A humoral response to tumor proteins may have utility for the detection of the pancreatic cancer. We have used 2-D liquid separation and protein microarrays to study

the humoral response in pancreatic cancer. Several different statistical treatments of results were used to highlight proteins that elicited a differential humoral response pattern between the different clinical groups. It was found that subtraction of background signal from microarray data often eliminated key signals that were able to distinguish between clinical groups, thus foreground measures without background subtraction were used instead for all statistical analyses. Rank-based statistics (Wilcoxon rank-sum tests) highlighted differences between the two clinical groups. Significant variability existed between the measurements obtained with the cancer sera, and z-score statistics were utilized as a complementary statistical tool to further analyze the differences between the cancer and control samples.

The PAM classification algorithm and leave-one-out cross-validation (LOOCV) highlighted a panel of 9 spots that was able to classify groups with high sensitivity and specificity. Furthermore, a separate validation study using available human recombinant proteins was able to substantiate results obtained with LOOCV for phosphoglycerate kinase-1 and histone H4. It is possible that all recombinant proteins used did not provide optimal results because they were not in their active form i.e. the correct isoform or post translational modification was absent. A study comparing the printed protein in the initial study vs. the recombinant protein to verify this hypothesis could not be performed because of insufficient sample from the initial study. Microarray results showed a significantly higher humoral response to a range of proteins in healthy subject sera compared to cancer sera. These proteins are primarily known to be involved in stress response and glycolysis. It was hypothesized that these differences were not due to sample variability (since data was globally normalized to eliminate systemic variations in

slide processing) but rather due to distinct mechanisms that renders these proteins less detectable in pancreatic cancer sera. In addition, proteins that have been previously implicated in cancer progression as well as other novel proteins such as a variety of ribosomal proteins showed higher humoral response in sera from cancer patients compared to healthy subjects.

However, further work using a larger panel of antibody and recombinant protein arrays containing active forms of proteins highlighted in this study together with a much larger sample set of normal and pancreatic cancer sera are necessary in order to validate these proteins as candidate markers of pancreatic cancer. Such work would also require one to assess reactivity to these proteins of sera from other types of cancers in order to ensure that the panel is pancreatic cancer specific.

Table 6.1: Protein identifications of spots that elicited a differential response from normal and cancer sera and were significant according to LOOCV results. All identifications were performed using a nanospray linear ion trap instrument (Thermo, LTQ) and SEQUEST browser. Only proteins that showed at least two high scoring peptides were considered true hits. If the protein was less than 15 kDA one high scoring peptide was considered acceptable.

% times selected	pH fraction	HPLC fraction	Protein Acc #	Protein ID	upregulated/ down regulated	Sequest total Score	% coverage
				Peptide Identified	Charge	Xcorr	
100	**6.6-6.4**	**44**	**P14618**	**Pyruvate kinase isozymes M1/M2**	**Down**	**1978**	**44.82**
				IENHEGVR	2	2.656	
				GSGTAEVELKK	2	3.285	
				PGSGFTNTMR	2	3.539	
				MQHLIAR	2	2.561	
				LNFSHGTHEYHAETIK	2	5.02	
				VFLAQK	1	2.053	
				VNFAMNVGK	2	2.846	
				APIIAVTR	2	2.852	
				ITLDNAYMEK	2	3.36	
				LDIDSPPITAR	2	4.081	
				GDLGIEIPAEK	2	3.867	
				KGVNLPGAAVDLPAVSEK	2	4.454	
				RFDEILEASDGIMVAR	3	5.095	
				GADFLVTEVENGGSLGSK	2	4.037	
				IYVDDGLISLQVK	2	5.089	
				EAEAAIYHLQLFEELR	2	5.399	
				FGVEQDVDMVFASFIR	2	5.737	
100	**8.1-7.8**	**56**	**P00558**	**Phosphoglycerate kinase 1**	**Up**	**359**	**32.69**
				NNQITNNQR	2	3.235	
				VDFNVPMK	2	2.855	
				IQLINNMLDK	2	3.348	
				VSHVSTGGGASLELLEGK	3	3.681	
				YSLEPVAVELK	2	3.182	
				VLNNMEIGTSLFDEEGAK	2	5.062	
				DVLFLK	1	1.937	
				ITLPVDFVTADK	2	2.7	
				VNEMIIGGGMAFTFLK	2	4.029	
				VLPGVDALSNI	2	2.718	
				ALESPERPFLAILGGAK	3	4.558	
73.3	**6.6-6.4**	**38**	**Q8NBJ7**	**Sulfatase-modifying factor 2**	**Down**	**320**	**22**
				FLMGTNSPDSR	2	4.252	
				EATVKPFAIDIFPVTNK	3	3.606	
				SVLWWLPVEK	2	3.456	
				LPTEEEWEFAAR	2	3.245	
				MGNTPDSASDNLGFR	2	4.717	
73.3	**6.6-6.4**	**43**	**P14618**	**Same as 6.6-6.4 fr 44**	**Down**		
73.3	**6.6-6.4**	**46**	**P18124**	**60S ribosomal protein L7**	**Down**	**804**	**37.5**

				TTHFVEGGDAGNR	3	4.142	
				ASINMLR	2	2.589	
				NFAELK	1	2.213	
				SVNELIYK	2	3.209	
				KVLQLLR	2	2.687	
				AGNFYVPAEPK	2	3.519	
				IALTDNALIAR	2	4.012	
				LAFVIR	1	1.674	
				IVEPYIAWGYPNLK	2	4.17	
				EANNFLWPFK	2	2.89	
73.3	**7.8-7.5**	**42**	**Q96A08**	**Histone H2B type 1-A**	**Up**	**180**	**25.98**
				HAVSEGTKAVTKYTSSK	3	4.708	
				EIQTAVRLLLPGELAK	2	2.873	
66.7	**7.8-7.5**	**38**	**P62937**	**Peptidyl-prolyl cis-trans isomerase A**	**Down**	**168**	**17.68**
				FEDENFILK	2	3.333	
				EGMNIVEAMER	2	3.116	
				VSFELFADK	2	2.898	
63.3	**6.4-6.1**	**6**		**Insufficient sample for ID**	**Up**		
53.3	**6.4-6.1**	**4**		**Insufficient sample for ID**	**Down**		
46.7	**9.2-9.1**	**24**	**P46783**	**40S ribosomal protein S10**	**Up**	**70**	**12**
				HPELADK	2	2.621	
				AEAGAGSATEFQFR	2	4.335	
40	**5.1-4.9**	**10**		**Insufficient sample for ID**	**Up**		
40	**8.1-7.8**	**4**	**P28001**	**Histone H2A.a**	**Down**	**200**	**10.85**
				SGRGK	2	2.751	
				AGLQFPVGR	2	3.126	
36.7	**6.4-6.1**	**26**	**P04792**	**Heat-shock protein beta-1**	**Down**	**1500**	**41.46**
				TKDGVVEITGK	2	3.809	
				AQLGGPEAAK	2	3.352	
				QLSSGVSEIR	2	2.665	
				DGVVEITGK	2	3.031	
				PLPPAAIESPAVAAPAYSR	3	5.349	
				RVPFSLLR	2	2.803	
				LATQSNEITIPVTFESR	2	2.754	
				LFDQAFGLPR	2	4.311	
36.7	**6.9-6.6**	**9**	**Q9UNX3**	**60S ribosomal protein L26-like 1**	**Up**	**130**	**12.41**
				KDDEVQVVR	2	3.139	
				FNPFVTSDR	2	3.049	
36.7	**8.7-8.4**	**49**	**P62805**	**Histone H4**	**Up**	**130**	**28.16**
				DAVTYTEHAK	2	2.841	
				DNIQGITKPAIR	2	3.627	
				TLYGFGG	1	2.055	
33.3	**6.4-6.1**	**28**	**P18754**	**Regulator of chromosome condensation**	**Down**	**420**	**20.67**
				DTSVEGSEMVPGK	2	3.98	
				SPPADAIPK	2	2.828	
				VVQVSAGDSHTAALTDDGR	3	5.181	
				LGLGEGAEEK	2	3.063	
				VPELFANR	2	2.94	
				SMVPVQVQLDVPVVK	2	4.8	
				DNNGVIGLLEPMK	2	3.769	

33.3	**6.4-6.1**	**32**	**P61457**	**Pterin-4-alpha-carbinolamine dehydratase**	**Down**	**210**	**13.59**
				LSAEERDQLLPNLR	2	3.566	
				LSAEERDQLLPNLR	3	4.371	
33.3	**6.4-6.1**	**34**	**P02545**	**Lamin-A/C**	**Down**	**760**	**22.74**
				LLEGEEER	2	3.018	
				ITESEEVVSR	2	3.681	
				SGAQASSTPLSPTR	2	4.6	
				LEAALGEAK	2	3.097	
				SLETENAGLR	2	3.271	
				EGDLIAAQAR	2	3.601	
				LQEKEDLQELNDR	3	4.182	
				AAYEAELGDAR	2	3.925	
				LQTMKEELDFQK	3	3.511	
				EAALSTALSEK	2	3.284	
				TLEGELHDLR	2	3.225	
				LADALQELR	2	4.007	
				IDSLSAQLSQLQK	2	4.206	
				LKDLEALLNSK	2	4.318	
				DLEALLNSK	2	2.892	
30	**5.3-5.1**	**15**		**Insufficient sample for ID**	**Up**		
26.7	**5.1-4.9**	**9**	**Q9UK76**	**Androgen-regulated protein 2**	**Up**	**130**	**21.43**
				SSGGREDLESSGLQRR	3	3.735	
				SAGAKSSGGREDLESSGLQR	3	4.369	
				EDLESSGLQR	2	3.076	
				NPPGGKSSLVLG	2	2.971	
26.7	**5.3-5.1**	**20**	**P46109**	**Crk-like protein**	**Up**	**470**	**27.72**
				HGMFLVR	2	2.54	
				IFDPQNPDENE	2	2.766	
				VSHYIINSLPNR	3	3.872	
				VGMIPVPYVEK	2	2.832	
				IHYLDTTTLIEPAPR	2	4.652	
				TALALEVGDIVK	2	4.454	
				TLYDFPGNDAEDLPFK	2	4.177	
26.7	**9.1-8.7**	**15**		**Insufficient sample for ID**	**Down**		

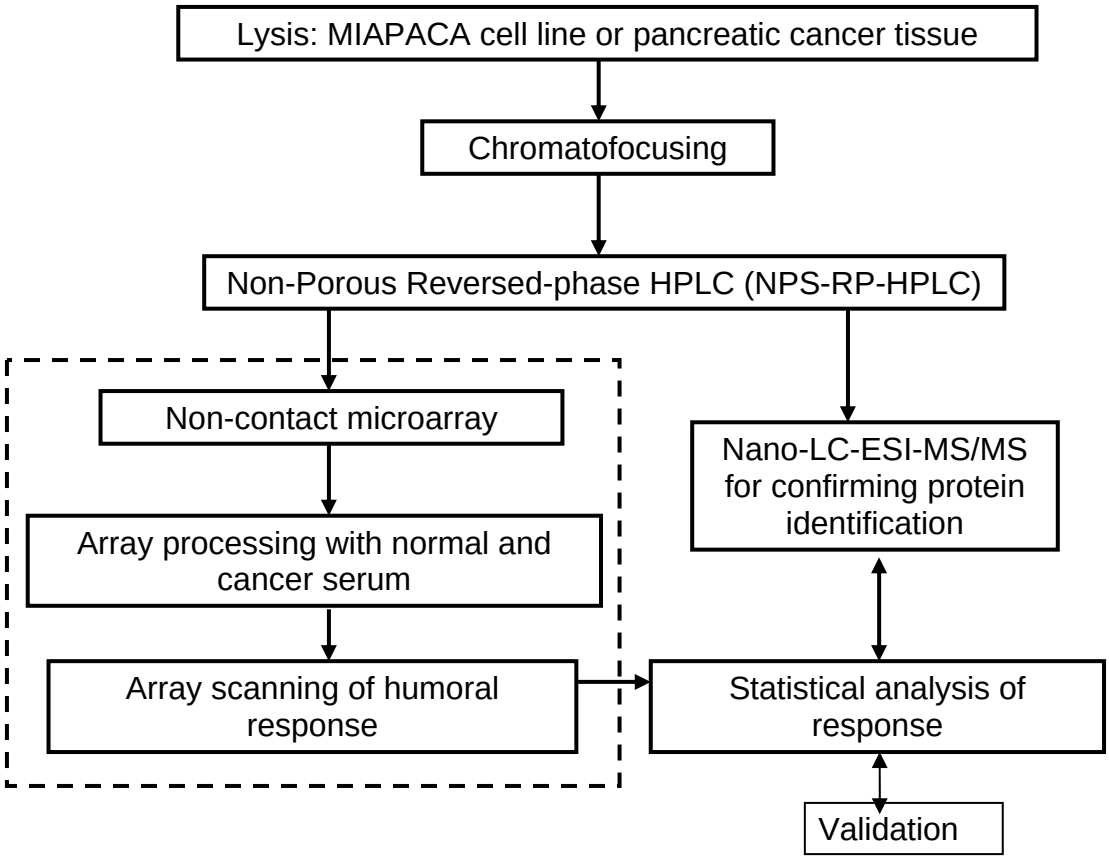
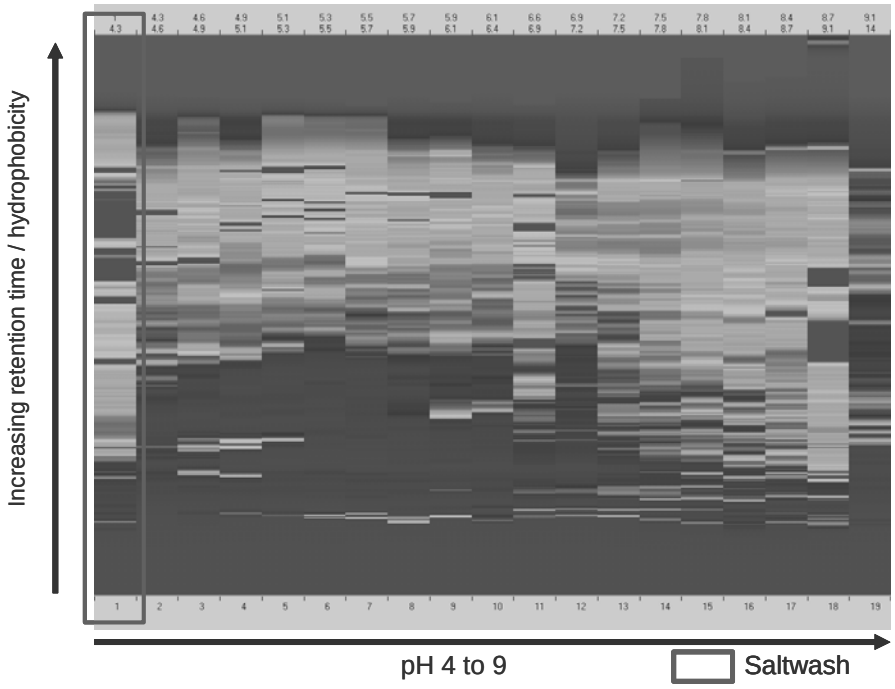


Figure 6.1: Humoral response experimental overview. Proteins are first extracted from cell line and separated in two orthogonal dimensions. Separated fractions are spotted by non-contact means on nitrocellulose slides which are then probed with serum from normal or cancer sera. Antibody-antigen response in detected using anti-human IgG conjugated to a florophore. Following non-parametric analysis proteins of interest are identified by tandem mass spectrometry.

(a)



(b)

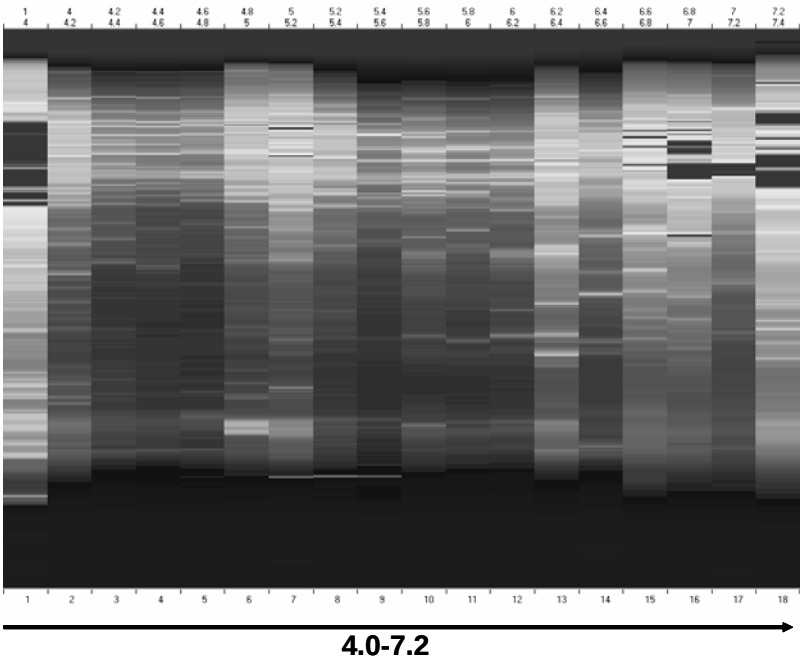
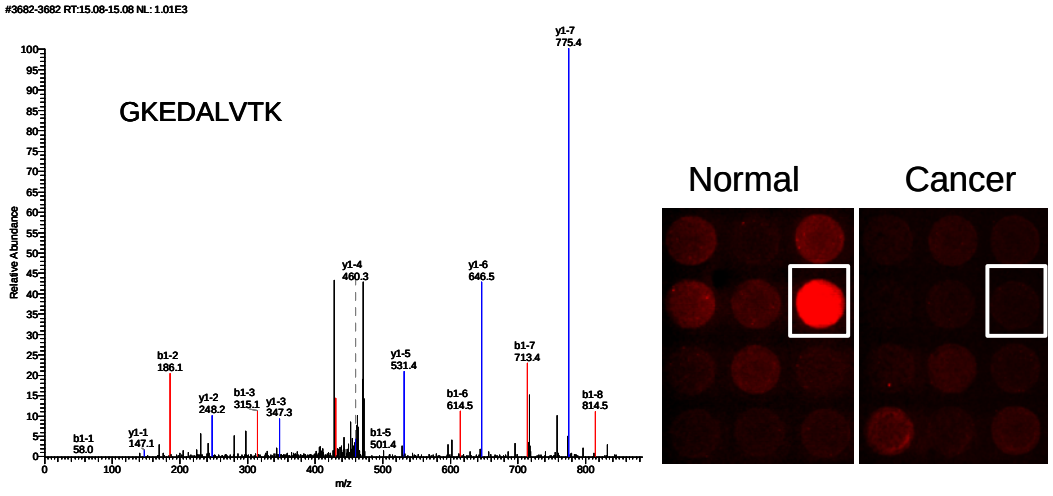


Figure 6.2: 2D UV chromatogram of separated (a) MIAPACA cell lysate and (b) pancreatic cancer tissue. On the horizontal axis are fractions from chromatofocusing starting from the lowest pH going to the highest pH. On the vertical axis is increasing retention time or hydrophobicity of the separated protein.

a)



b)

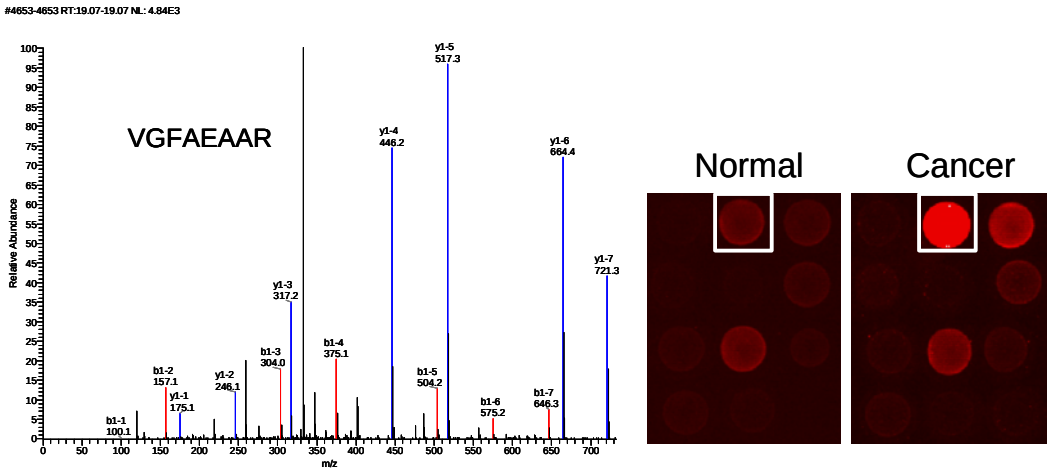
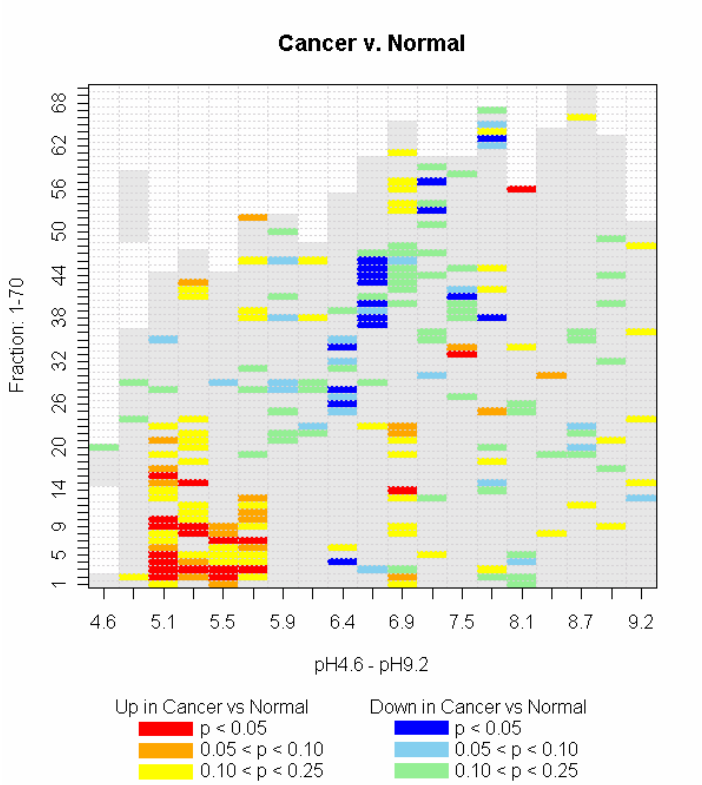
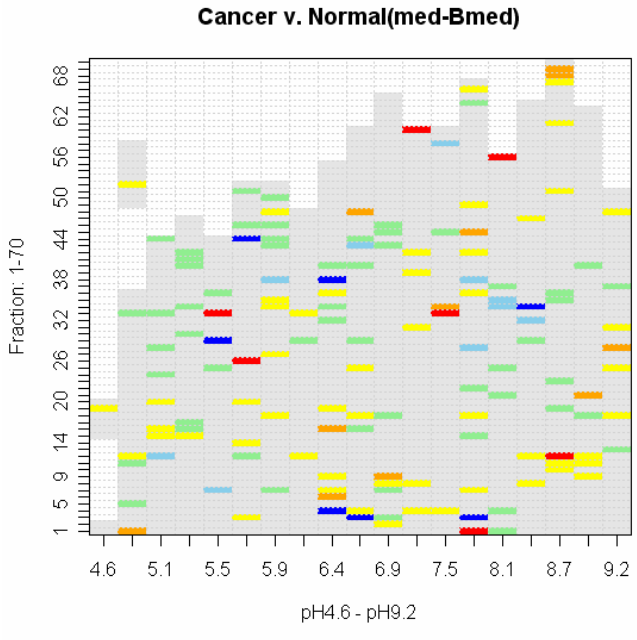


Figure 6.3: Selected microarray shots of differential humoral response as well as selected tandem mass spectrum for sequence confirmation of (a) Fibrillarin and (b) Cathepsin D. Theoretical location of b ions is indicated by red lines. In most cases these peak intensities were not high enough for detection.

a)


Cancer v. Normal
Fraction: 1-70
1
5
9
14
20
26
32
38
44
50
56
62
68
4.6
5.1
5.5
5.9
6.4
6.9
7.5
8.1
8.7
9.2
pH4.6 - pH9.2
Up in Cancer vs Normal
p < 0.05
0.05 < p < 0.10
0.10 < p < 0.25
Down in Cancer vs Normal
p < 0.05
0.05 < p < 0.10
0.10 < p < 0.25

b)


Cancer v. Normal(med-Bmed)
Fraction: 1-70
1
5
9
14
20
26
32
38
44
50
56
62
68
4.6
5.1
5.5
5.9
6.4
6.9
7.5
8.1
8.7
9.2
pH4.6 - pH9.2

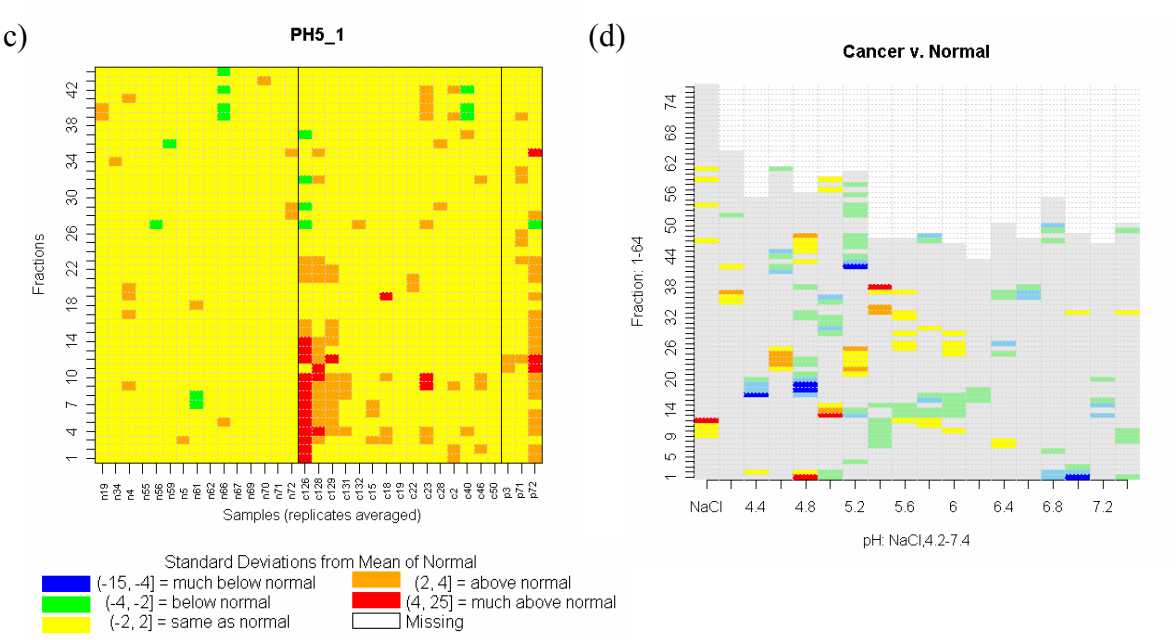
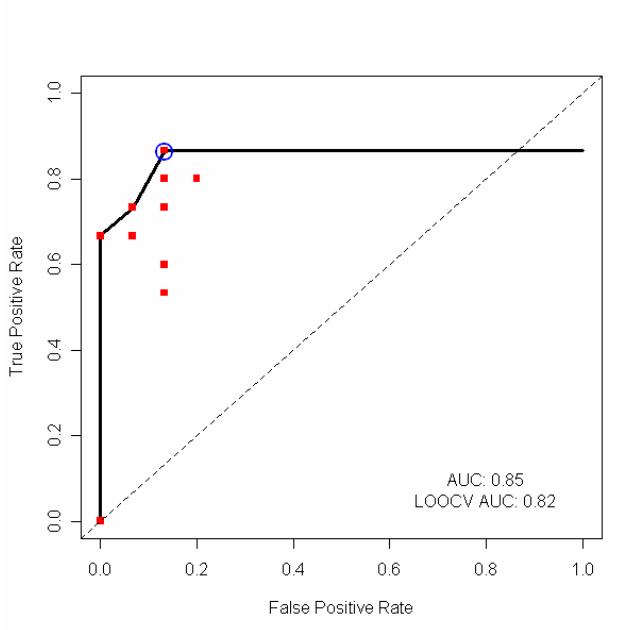


Figure 6.4: All separated fractions showing result with non-parametric Wilcoxon tests (a) without background subtraction and (b) with background subtraction for the MIAPACA cell line. Red and Orange blocks mean significantly higher humoral response in cancer samples compared to normal ($p<0.05$ and $p<0.1$ respectively) and darker and lighter shades of Blue represent higher humoral response in normal compared to cancer ($p<0.05$ and $p<0.1$ respectively). Yellow and green blocks mean $0.1<p<0.25$. (c) z-score plot for proteins separated from pH fraction 5.1-4.9. On the vertical axis are all fraction by increasing retention time and on the horizontal axis are each of the serum samples with which samples were probed. Red and Yellow blocks represents responses significantly higher than the mean of the normal sample ($4<Z<25$ and $2<Z<4$ respectively) while Blue and Green blocks represent responses significantly lower than the mean of the normal sample ($-25<Z<-4$ and $-4<Z<-2$ respectively). (d) All separated fractions from pancreatic cancer tissue showing results with non-parametric Wilcoxon tests. Color codes are the same as for figure 6.4 a and b.

a)



b)

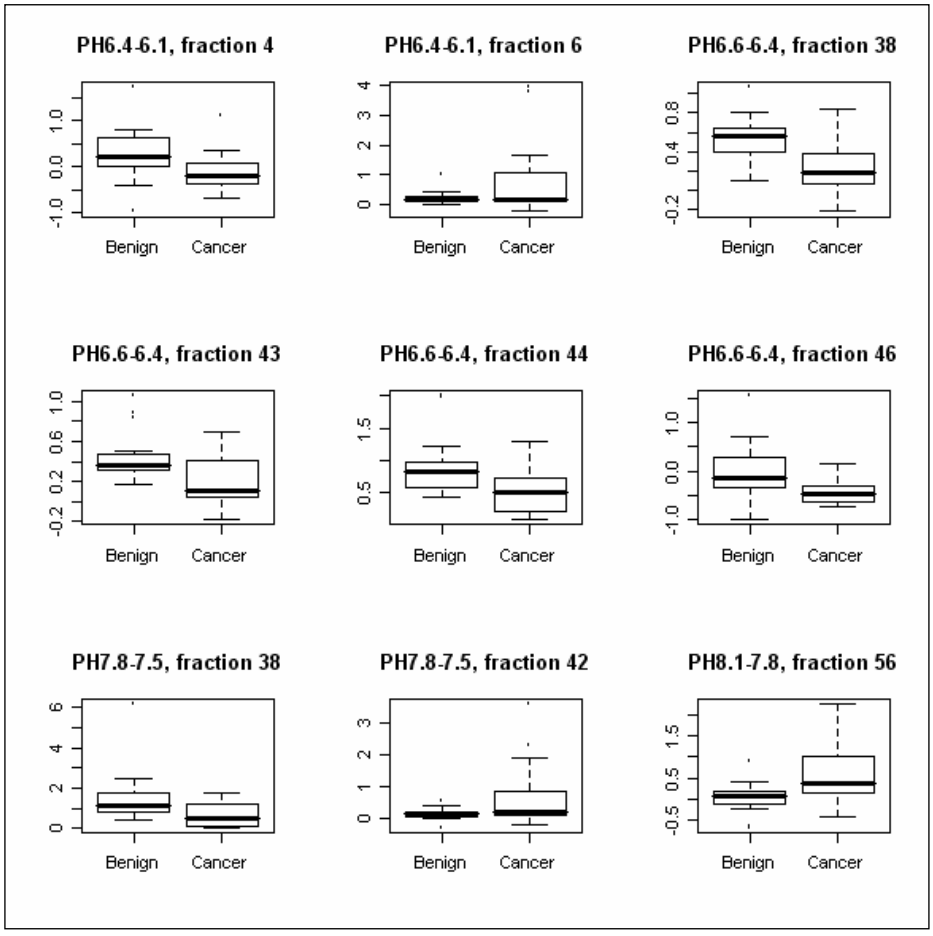


Figure 6.5: (a) ROC curve of 9 protein panel from PAM analysis showing an area under the curve of 0.85. (b) Boxplots of the 9 protein panel classifier built using all 30 samples.

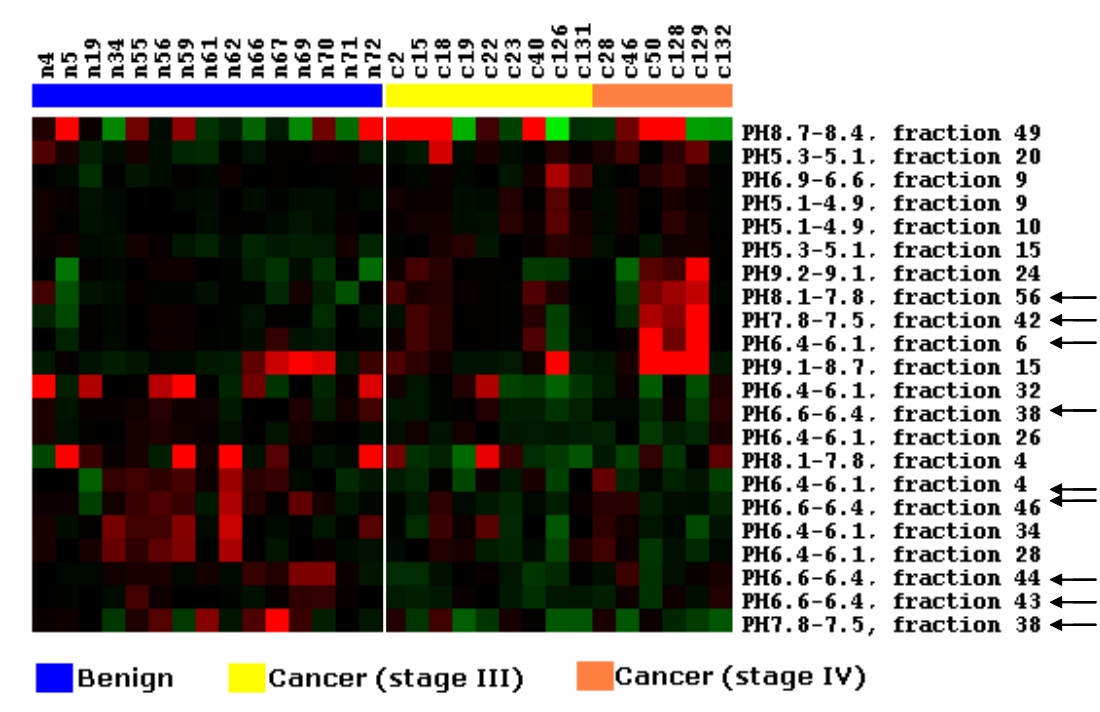
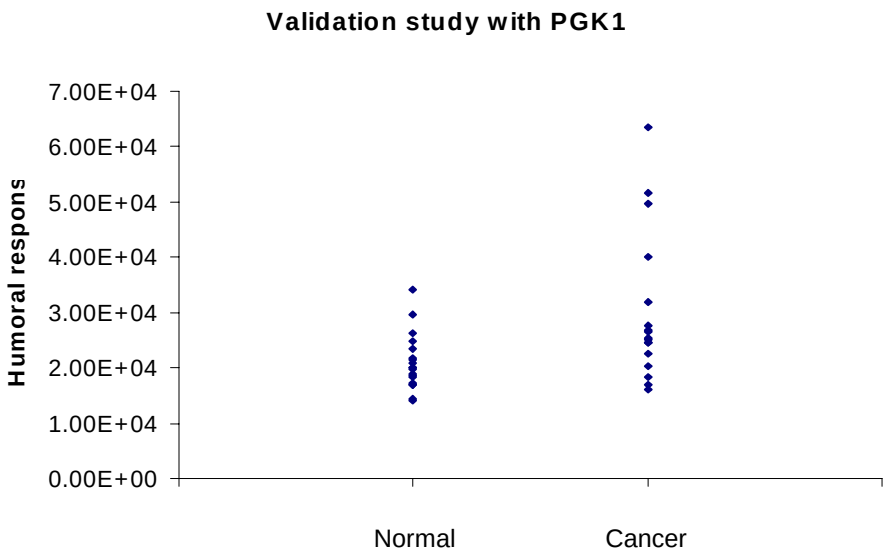


Figure 6.6: Heatmap showing median centered responses of all serum samples to selected proteins of interest. The scale from green to red represents lower response to higher response on a scale of -2 to 2. The arrows in the figure indicate the protein spots that formed the panel of 9 potential markers with highest sensitivity and specificity.

a)



b)

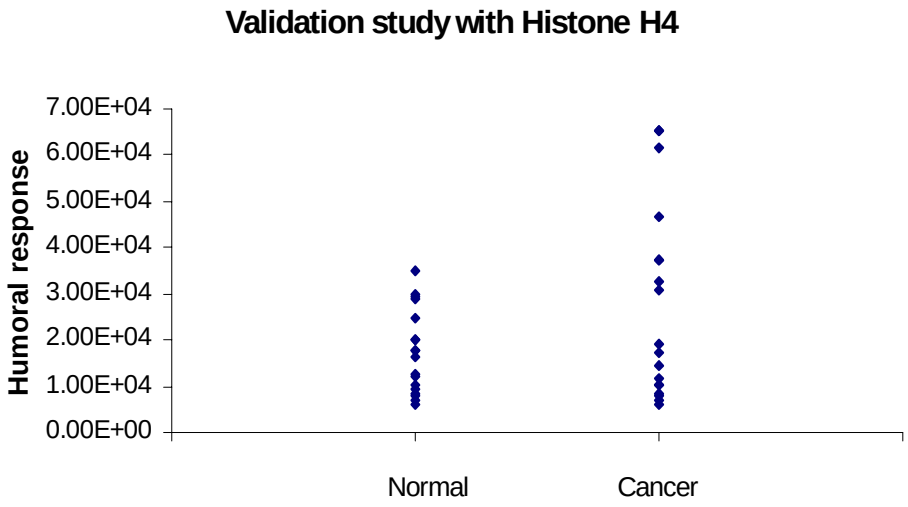


Figure 6.7: Scatterplot illustrating the differential humoral response in recombinant human PGK-1 used for validating initial experimental results.

6.5. References

[1] Jemal, A., Siegel, R., Ward, E., Murray, T., *et al.*, *CA Cancer J Clin* 2006, *56*, 106-130.

[2] Mann, D. V., Edwards, R., Ho, S., Lau, W. Y., Glazer, G., *Eur J Surg Oncol* 2000, *26*, 474-479.

[3] Ferrone, C. R., Finkelstein, D. M., Thayer, S. P., Muzikansky, A., *et al.*, *J Clin Oncol* 2006, *24*, 2897-2902.

[4] Duffy, M. J., *Ann Clin Biochem* 1998, *35 (Pt 3)*, 364-370.

[5] Boeck, S., Stieber, P., Holdenrieder, S., Wilkowski, R., Heinemann, V., *Oncology* 2006, *70*, 255-264.

[6] Ujiki, M. B., Talamonti, M. S., *Semin Radiat Oncol* 2005, *15*, 218-225.

[7] Brichory, F. M., Misek, D. E., Yim, A. M., Krause, M. C., *et al.*, *Proc Natl Acad Sci U S A* 2001, *98*, 9824-9829.

[8] Le Naour, F., Misek, D. E., Krause, M. C., Deneux, L., *et al.*, *Clin Cancer Res* 2001, *7*, 3328-3335.

[9] Gure, A. O., Altorki, N. K., Stockert, E., Scanlan, M. J., *et al.*, *Cancer Res* 1998, *58*, 1034-1041.

[10] Stockert, E., Jager, E., Chen, Y. T., Scanlan, M. J., *et al.*, *J Exp Med* 1998, *187*, 1349-1354.

[11] Lin, H. S., Talwar, H. S., Tarca, A. L., Ionan, A., *et al.*, *Cancer Epidemiol Biomarkers Prev* 2007, *16*, 2396-2405.

[12] Chatterjee, M., Ionan, A., Draghici, S., Tainsky, M. A., *Omics* 2006, *10*, 499-506.

[13] Draghici, S., Chatterjee, M., Tainsky, M. A., *Expert Rev Mol Diagn* 2005, *5*, 735-743.

[14] Ben-Mahrez, K., Sorokine, I., Thierry, D., Kawasumi, T., *et al.*, *Int J Cancer* 1990, *46*, 35-38.

[15] Pupa, S. M., Menard, S., Andreola, S., Colnaghi, M. I., *Cancer Res* 1993, *53*, 5864-5866.

[16] Winter, S. F., Minna, J. D., Johnson, B. E., Takahashi, T., *et al.*, *Cancer Res* 1992, *52*, 4168-4174.

[17] Raedle, J., Oremek, G., Welker, M., Roth, W. K., *et al.*, *Pancreas* 1996, *13*, 241-246.

[18] Hamanaka, Y., Suehiro, Y., Fukui, M., Shikichi, K., *et al.*, *Int J Cancer* 2003, *103*, 97-100.

[19] Kotera, Y., Fontenot, J. D., Pecher, G., Metzgar, R. S., Finn, O. J., *Cancer Res* 1994, *54*, 2856-2860.

[20] Maacke, H., Jost, K., Opitz, S., Miska, S., *et al.*, *Oncogene* 2000, *19*, 2791-2795.

[21] Xia, Q., Kong, X. T., Zhang, G. A., Hou, X. J., *et al.*, *Biochem Biophys Res Commun* 2005, *330*, 526-532.

[22] Hong, S. H., Misek, D. E., Wang, H., Puravs, E., *et al.*, *Cancer Res* 2004, *64*, 5504-5510.

[23] Hong, S. H., Misek, D. E., Wang, H., Puravs, E., *et al.*, *Biomarker Insights* 2006, *2*, 175-183.

[24] Abu-Shakra, M., Buskila, D., Ehrenfeld, M., Conrad, K., Shoenfeld, Y., *Ann Rheum Dis* 2001, *60*, 433-441.

[25] Cekaite, L., Hovig, E., Sioud, M., *Methods Mol Biol* 2007, *360*, 335-348.

[26] Yan, F., Sreekumar, A., Laxman, B., Chinnaiyan, A. M., *et al.*, *Proteomics* 2003, *3*, 1228-1235.

[27] Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G., *Proc Natl Acad Sci U S A* 2002, *99*, 6567-6572.

[28] Scharpf, R. B., Iacobuzio-Donahue, C. A., Sneddon, J. B., Parmigiani, G., *Biostatistics* 2007.

[29] Eisen, M. B., Spellman, P. T., Brown, P. O., Botstein, D., *Proc Natl Acad Sci U S A* 1998, *95*, 14863-14868.

[30] Sarto, C., Valsecchi, C., Magni, F., Tremolada, L., *et al.*, *Proteomics* 2004, *4*, 2252-2260.

[31] Franco, S., Alt, F. W., Manis, J. P., *DNA Repair (Amst)* 2006, *5*, 1030-1041.

[32] Mikuriya, K., Kuramitsu, Y., Ryozawa, S., Fujimoto, M., *et al.*, *Int J Oncol* 2007, *30*, 849-855.

[33] Shichijo, S., Azuma, K., Komatsu, N., Ito, M., *et al.*, *Clin Cancer Res* 2004, *10*, 5828-5836.

Chapter 7

Enhanced Detection of Autoantibodies on Protein Micorarrays Using a Modified Protein Digestion Technique

7.1. Introduction

Proteome profiling has become a field of increasing interest since protein expression profiles may be a more relevant biological readout of cell systems than transcriptional profiles. The use of protein microarrays facilitates high throughput screening of such protein expression profiles.[1-3] Multiple antibody or antigen probes are located at fixed and unique positions on a microarray chip facilitating interrogation of several thousand sample components simultaneously. Protein chips have emerged in a variety of different formats the most common of which is the antibody microarray where antibodies are immobilized on an array surface to capture proteins of interest.[3-10] Such studies are important when assessing binding properties of already known target proteins, but novel proteins involved in disease progression may be overlooked. Furthermore well-characterized antibodies with high specificities to proteins of interest are difficult to develop and relatively expensive for routine use.

Autoantibody (humoral) response studies can provide critical information about a body's response to disease antigens. In such studies, potential antigens are arrayed on a slide and probed with serum from various classes of patients i.e. normal vs. disease. It is

assumed that there will be antibodies in the serum sample of disease patients produced as a reaction to some of the antigens printed on the arrays. Previous reports have shown the presence of such antigen-antibody reactions as a result of disease .[11-18] In recent work, a novel approach has been demonstrated where natural proteins from a cancer cell or tissue are first resolved by a two-dimensional liquid separation and are then arrayed on thin nitrocellulose microarray slides.[19, 20] All proteins from the cell line are then simultaneously probed with a large number of serum samples to highlight immune responses that can differentiate between normal and disease sera.

One area of difficulty in these types of humoral response experiments is the low signal intensity that is often present in the arrays. While differential responses are observed for certain potential cancer protein markers the response overall is not remarkably high. It is our hypothesis that this weak response could be a result of protein immobilization on the slide which renders the protein unable to move about such that binding sites are blocked from reagent molecules. We propose that reducing the protein size by chemical means may facilitate exposure of these binding sites thereby enhancing the overall sensitivity of the humoral response experiments. It should be noted that pre-treatment with chemical digestion would allow measurement of epitope-sensitive interactions that only requires a certain amino acid sequence and no specific protein confirmation since the protein confirmation will be compromised.

Reduction of the protein size can easily be accomplished by protein digestion. A variety of techniques for protein digestion are currently available, however digestion into very small fragments may completely destroy epitopes where antibody/antigen binding occur, making it important to select enzymes or reagents with care. An ideal digestion

protocol would reduce the protein into a few long peptide fragments rather than many small fragments.

In this report a comparison of two digestion methods vs. undigested proteins is performed after protein fractionation but before printing to assess the improvement in humoral response data (Figure 7.1). A pancreatic cancer cell line (Panc1) was lysed and the extracted proteins were separated in two dimensions. The separated proteins were then either directly arrayed on nitrocellulose slides or were first digested either enzymatically using GluC or chemically using cyanogen bromide (CNBr) and then arrayed on the same slides. The arrays were then processed with serum samples from 10 normal individuals, 10 chronic pancreatitis and 10 pancreatic cancer patients. Humoral response to digested vs. undigested proteins was then compared to evaluate if digestion improved response. Furthermore differences seen between the different serum classes was further interrogated by identification of the protein eliciting the humoral response using LC-MS/MS methodologies.

7.2. Experimental Section

Cell Culture and Sample Preparation and serum collection

Sample Preparation. (a) Cell Culture

Studies were performed using the Panc-1 pancreatic adenocarcinoma cell line (obtained by ATCC). The cells were cultured in Dulbecco's modified Eagle medium supplemented with 10% fetal bovine serum, 100 units/ml penicillin and 100 units/ml streptomycin (Invitrogen, Carlsbad, CA). When the cells reached ~90% confluence, the cells were harvested with a cell scraper.

(b) Cell Lysis

Cell pellets were reconstituted in lysis buffer consisting of 7.5 M urea, 2.5 M thiourea, 4% *n*-octyl-β-D-glucopyranoside (*n*-OG), 10 mM tris(2-carboxyethyl) phosphine (TCEP), 12.5% v/v glycerol, and 1% v/v protease inhibitor cocktail (Sigma, St. Louis, MO). The cell pellets were lysed at room temperature for 1 h, followed by centrifugation at 35 000 rpm at 4 °C for 1 h. The supernatant was buffer exchanged into start buffer (6 M urea, 25 mM Bis-Tris, and 0.2% OG) using a PD-10 G-25 column (Amersham Biosciences, Piscataway, NJ) and stored at -80^{o}C until further use.

(c) Serum collection

Serum was obtained at the time of diagnosis following informed consent using IRB-approved guidelines. Sera were obtained from 10 patients with a confirmed diagnosis of pancreatic adenocarcinoma in the Multidisciplinary Pancreatic Tumor Clinic at The University of Michigan Hospital. These sera were randomly selected from a clinic population that sees, on average, at the time of initial diagnosis, 15% of pancreatic adenocarcinoma patients presenting with early stage (i.e., stage 1/2) disease and 85% presenting with advanced stage (i.e., stage 3/4). Inclusion criteria for the study included patients with a confirmed diagnosis of pancreatic cancer, the ability to provide written, informed consent, and the ability to provide 40 ml of blood. Exclusion criteria included inability to provide informed consent, patient's actively undergoing chemotherapy or radiation therapy for pancreatic cancer, and patients with other malignancies diagnosed or treated within the last 5 years. Sera were also obtained from 10 patients with chronic pancreatitis who were seen in the Gastroenterology Clinic at University of Michigan Medical Center, and from 10 control healthy individuals collected at University of Michigan under the auspices of the Early Detection Research Network (EDRN). The

mean age of the tumor group was 65.4 years (range 54-74 years) and from the chronic pancreatitis group was 54 years (range 45-65). The sera from the normal subject group was age and sex-matched to the tumor group. All of the chronic pancreatitis sera were collected in an elective setting in the clinic in the absence of an acute flare. All sera were processed using identical procedures. The samples were permitted to sit at room temperature for a minimum of 30 minutes (and a maximum of 60 minutes) to allow the clot to form in the red top tubes, and then centrifuged at 1,300 x g at 4^{o}C for 20 minutes. The serum was removed, transferred to a polypropylene, capped tube in 1 ml aliquots, and frozen. The frozen samples were stored at -70^{o}C until assayed. All serum samples were labeled with a unique identifier to protect the confidentiality of the patient. The handling of all serum samples was similar in that none of the samples were thawed more than twice before analysis in order to minimize protein degradation and precipitation.

Separation

Chromatofocusing (CF)

CF separation was performed on an HPCF-1D column (250 × 2.1 mm) (Beckman-Coulter, Fullerton, CA) using the ProteomeLab™ PF2D protein fractionation system (Beckman-Coulter), as described previously.[21, 22] Two buffers were used to generate the pH gradient on the column. The start buffer (SB) solution was composed of 6M urea and 25mM Bis-Tris (pH 7.4). The elution buffer (EB) solution was composed of 6M urea and 10% polybuffer74 (pH 4.0). Both buffer solutions were brought to pH by addition of a saturated solution of iminodiacetic acid. The CF column was pre-equilibrated with SB. After equilibration, 4.5 mg of proteins were loaded onto the CF column and the column was washed with 100% SB to remove material that did not bind

to the column at pH 7.4. Elution was achieved by applying a pH 4.0 elution buffer at a flow rate of 0.2 mL/min. The pH gradient was monitored on-line by a flow-through pH probe (Beckman-Coulter). The UV absorbance of the eluent was monitored on-line at 280nm. The flow rate was 0.2ml/min, with 16 fractions in total being collected in 0.2 pH units in the range of pH 7.0 - 4.0. Each fraction was stored at -80°C until further use.

Non-Porous Silica Reversed-Phase (NPS-RP)-HPLC with sample collection

When the first-dimension separation was completed, the pI fractions collected from the first dimension were separated by NPS-RP-HPLC using an ODSIII (4.6 × 33 mm) NPS column (Eprogen) at a flow rate of 0.5 mL/min and detected by absorbance at 214 nm using a Beckman model 166 UV absorption detector. Proteins eluting from the column were collected by an automated fraction collector (Model SC 100, Beckman), controlled by an in-house designed DOS-based software program. To enhance the speed, resolution, and reproducibility of the separation, the RP column was heated to 65ºC by a column heater (Jones Chromatography, Model 7971, Resolution Systems, Holland, MI). Both mobile phase A: MilliQ® water (Millipore, Billerica, MA), and solvent B: acetonitrile (ACN) (Sigma) contains 0.1% v/v and 0.08% v/v respectively, trifluoroacetic acid (TFA). The gradient was run from 5% to 15% in 1 min, 15% B to 25% in 2 min, 25% to 31% in 2 min, 31% to 41% in 10 min, 41% to 47% in 6 min, 47% to 67% in 4 min, then up to 100% B in 3 min where it was held for 1 min, and then reduced to 5% in 1 min. After the gradient, the column was washed by two fast gradients from 5% B to 100% B in 5 min, 100% B back to 5% B in 1 min. Fractions from the HPLC eluent were collected using a semi-automated in-house program using a Model SC-100 fraction collector. Collected peak fractions were stored at -80ºC for further use.

Protein Digestion by CNBr or GluC

For digestion with CNBr, collected fractions were then dried down and resuspended in 5µL deionized water, 15µL TFA and 5µL 5M CNBr in ACN. The tubes were wrapped in aluminum foil and left overnight at 4 ºC.

For digestion with GluC, collected fractions were dried down to ~10 µL. 40 µL of 100 mM ammonium bicarbonate and 0.1 µg of GluC were added to the same and the mixture was left at room temperature overnight.

Microarray Printing

Fractionated proteins were transferred to 96-well printing plates (Bio-Rad) and were lyophilized to dryness. The fractions were then resuspended in printing buffer (62.5 mM Tris-HCl (pH6.8), 1% w/v sodium dodecyl sulfate (SDS), 5% w/v dithiothreitol (DTT) and 1% glycerol in 1X PBS) and were left to shake overnight at 4^{o}C. Slides were printed by transferring each fraction from the plate onto nitrocellulose slides using a non-contact piezoelectric printer (Nanoplotter 2, GeSiM). Each spot resulted from deposition of 5 spotting events of 500 pL each, such that a total volume of 2.5 nL of each fraction was spotted. Each spot was found to be ~450 µm in diameter, with the distance between spots maintained at 600 µm. Printed slides were left on the printer deck overnight to dry and were then stored, desiccated at 4^{o}C until further use.

Hybridization of slides

The printed arrays were rehydrated in 1X PBS with 0.1% Tween-20 (PBS-T), and were then blocked overnight in a solution of 1% BSA in PBS-T. Each serum sample was diluted 1:400 in probe buffer (5 mM magnesium chloride, 0.5 mM DTT, 0.05% Triton X-100, 5% glycerol and 1% BSA in 1X PBS) to make a total solution of 4 mL and kept on

ice. Each diluted serum sample was used to hybridize a slide for 2 hrs. Hybridization was done at 4^{o}C in heat-sealable pouches with agitation, using a mini-rotator. The slides were then washed five times with probe buffer (5 min each), and were then hybridized with 4 mL anti-human IgG conjugated with Alexaflour647 (Invitrogen, Carlsbad, CA) (at 1 µg/mL), for 1 hr at 4^{o}C. After secondary incubation all slides were washed in probe buffer five times, for 5 min each, and were then dried by centrifugation for 10 min. All processed slides were immediately scanned using an Axon 4000B microarray scanner (Axon Instruments Inc., Foster City, CA) and GenePix Pro 6.0 software (Molecular Devices, Sunnyvale, CA) was used for data acquisition and analysis.

Protein Identification

Proteins were trypsinized in a solution of 100 mM ammonium bicarbonate and 1 mM DTT. The samples digested by trypsin were separated by a capillary RP column (C18, 0.3 × 150 mm) (Michrom Biosciences, Auburn, CA) on a Paradigm MG4 micropump (Michrom Biosciences) with a flow rate of 300 nL/min. The gradient, started at 5% ACN, was ramped to 60% ACN in 25 min and finally ramped to 95% in another 5 min. Both solvents A (water) and B (ACN) contained 0.3% formic acid. The resolved peptides were analyzed on a Finnigan LTQ mass spectrometer (Thermo Electron Corp., San Jose, CA) with a nanoESI ion source (Thermo). The capillary temperature was set at 190°C, spray voltage was 2.6 kV, and capillary voltage was 30 V. The normalized collision energy was set at 35% for MS/MS. The top 5 peaks were selected for CID. Precursor selection was based upon a normalized threshold of 30 counts/s. MS/MS spectra were searched using the SEQUEST algorithm incorporated in Bioworks software (Thermo) against the Swiss-Prot human protein database. The search was performed

using the following parameters: two miscleavages were allowed during the database search; peptide ion mass tolerance 1.50 Da; fragment ion mass tolerance 0.0 Da; Protein identification was considered positive 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.

7.3. Results and Discussion

Protein separation methods

In this study, proteins extracted from the Panc1 pancreatic cancer cell line were separated in two relatively orthogonal dimensions for maximum peak capacity. In the first dimension the proteins were separated according to their isoelectric points by a liquid phase separation technique, chromatofocusing. It has been shown that such a technique is able to separate proteins whose isoelectric points are as low as 0.2 pH units apart or less. In the second dimension each fraction from chromatofocusing was further separated by non-porous reversed phase HPLC. Non-porous particles eliminated column clogging problems associated with separation of large proteins using porous columns. Separation times and quality were also optimized due to the short column length of 33mm and separation temperature of 65^{o}C which reduced diffusional broadening.

Figure 7.2 illustrates the separation quality and reproducibility of each dimension of separation. Figure 7.2A shows three independent CF runs together with the pH profile of each run. It is evident that the pH profiles of all runs are very similar as are the peak profiles. It can also be seen that the peaks and valleys in the profiles correspond to the pH changes and not the retention times so that even if there is a slight change in the pH profile, the proteins eluting over each pH interval will always be the same. Figure 7.2B

illustrates similar reproducibility data generated for the second dimension separations. Separations done on two distinct pH lanes (5.2-5.0 and 6.6-6.4) from CF are shown. Again it was observed that peak profiles were almost identical in all four cases. For the second dimension profile of pH fraction 5.2-5.0 the bottom-most chromatogram has one peak missing which was due to a pressure drop that occurred in the instrument for a short period of time. The reproducibility data suggests that the two-dimensional liquid separation used in this study was robust, reliable and reproducible for further studies.

Digestion protocols:

In humoral response experiments using undigested proteins we separated cellular cancer proteins and printed them on nitrocellulose slides in their intact form. However the overall response of these arrays to serum remained low in most experiments. While some differences in response were statistically different between normal and cancer sera the overall fold differences were not very high. We hypothesized that the low response correlated with the lack of access to binding sites as illustrated in figure 7.3. When the protein is intact the binding site which would potentially react with serum proteins, could be in a sterically unfavorable position, resulting in low binding. If this protein was cleaved into a few pieces it is possible that the autoantibody binding site would be more favorably located during the array hybridization process.

The digestion method chosen to study this hypothesis is critical. A protocol that would cleave the proteins into too many fragments may not be appropriate because this would potentially destroy the binding site. Trypsin is a popular enzyme of choice in attempts at protein identification. However the very property of trypsin that is favorable for PMF experiments (two cleavage sites that lead to many peptide fragments) is not

favorable for the purpose of our study. We have therefore utilized two alternative methods to assess the effect of digestion on the humoral response. Enzymatic digestion using GluC was performed since this enzyme cleaves only after glutamic acid. Chemical digestion using cyanogen bromide (CNBr) was also used. CNBr cleaves after methionine residues and because the occurance of methionine in proteins is very low, cleavage by CNBr should result in very few, long peptides, possibly conserving the binding site for humoral response experiments. Peptides resulting from CNBr and GluC digestion could consequently be up to 4 or more times larger than peptides resulting from trypsin digestion.

Array hybridization

The separated proteins were first divided into 4 fractions and each fraction was used as follows: fraction 1=intact for microarray, fraction 2=digested with GluC for microarray, fraction 3=digested with CNBr for microarray and fraction 4=digested with trypsin for identification. All fractions for microarray analysis were arrayed using a non-contact arrayer, Nanoplotter 2.0E (GeSiM, Germany). The arrays were then probed with serum samples from 10 normal subjects, 10 patients with chronic pancreatitis and 10 patients with pancreatic cancer.

Figure 7.4 illustrates sections of microarray slides probed with serum to compare the three different array-hybridization protocols. The top panel reflects the humoral response to intact proteins arrayed onto nitrocellulose slides. It can be seen that the response is very weak for the fractions from the pH range shown. The middle panel shows the response of GluC digested fractions to serum. In this case all spots showed a positive response. The digestion protocol for GluC involved the addition of 0.1 ug of

gluC into the fraction that was being digested. In printing the arrays of GluC digested fraction, the GluC was also being printed. We concluded that all spots were positive most probably due to non-specific binding to the GluC that was present in the digested sample. The problem could be alleviated by having GluC immobilized on beads prior to digestion so that it can be removed from the sample prior to arraying. Interestingly, the bottom panel shows the humoral response to CNBr digested protein fractions where at least 1 spot shows significantly higher response to serum when digested by CNBr compared to when arrayed as an intact protein. In addition 3 other spots show a slightly higher response when digested compared to when arrayed as intact proteins.

When comparing the 62 fractions that were printed from the proteins from 2 separate pH lanes of the first dimensional separation, it was observed that 10 spots showed a very distinct response to serum when arrayed after CNBr digestion (a >10% improvement). These same spots did not show any response to serum when arrayed as intact proteins. By performing a simultaneous humoral response experiment with both a control (intact proteins) and test (digested with CNBr) set it was confirmed that digestion by CNBr prior to arraying facilitates an improved response making it possible to identify changes in humoral response that may be present between two different classes of sera. We believe that digestion with CNBr shows an enhanced humoral response that is clinically relevant and not just due to non-specific binding because only specific spots showed an enhanced response with this digestion strategy. If the response were non-specific we should have seen an enhancement with all fractions that were printed as was the case with GluC.

Initial study toward differential humoral response in pancreatic cancer using CNBr digested proteins

After confirming that CNBr digestion increases the sensitivity of the humoral response results, 10 normal, 10 chronic pancreatitis and 10 pancreatic cancer sera were hybridized with arrays printed with intact proteins and CNBr digested proteins to see if there was a group specific trend in the humoral response that resulted in improved detection with CNBr digested arrays. To that effect proteins from pH lanes 5.2-5.0 and 6.6-6.4 were fractionated by NPS-RP-HPLC. They were then divided into two parts, one of which was directly arrayed on nitrocellulose slides as intact proteins. The other divided fraction was subjected to CNBr digestion after which it was arrayed on the same nitrocellulose slides as the intact proteins. 30 slides were printed and each was hybridized with serum from the 10 normal, 10 chronic pancreatitis and 10 pancreatic cancer sera.

The arrays were scanned in the red channel to detect a humoral response. Background subtracted data from both sets of experiments i.e. intact protein results for all three sera classes and digested protein results for all three sera classes were then compared to see if there was any differential response between the classes that was more visible in the CNBr digested sample data. A spot intensity of $\geq$ 2 fold than the background intensity was required for the spot to be considered positive. At least 50% of the cancer and pancreatitis spots needed to show a higher response than the 2^{nd} highest normal spot response in order for the response to be considered significantly different in the groups.

5 spots showed a differential humoral response when the spots were digested with CNBr but not when they were arrayed in an intact state. A comparison of some of these

humoral responses is shown in figure 7.5. On the left are all sera responses to intact spots while on the right are sera responses to spots that were digested with CNBr. On the plots 1 refers to normal sera group, 2 refers to the chronic pancreatitis sera group and 3 refers to the pancreatic cancer sera group. For the digested spot responses a broken line is shown to indicate the number of cancer samples that showed a higher response than the second highest normal serum response. It can be seen that in all cases at least 50 % (5/10) of the chronic pancreatitis and pancreatic cancer samples showed higher reactivity in the studied fractions when digested with CNBr compared to when arrayed as intact proteins. Interestingly in figure 7.5 b, c and d the response to the intact spots was generally very low (almost at background fluorescence level) while the same response in the case of the spots digested with CNBr was significantly higher. On the contrary, as shown in figure 7.5 a and e, it was observed that while the overall signal intensities were not very low when intact protein spots were probed with sera from different groups, there was no indication of differential humoral response between the normal and disease sample group. However when digested with CNBr a differential humoral response became very evident despite the slightly lower overall signal intensity.

The 5 spots that were identified as eliciting a differential humoral response were further interrogated by mass spectrometry to determine the proteins present in the spots. Table 7.1 shows the identity of the proteins and relevant information with respect to the protein identification. Out of the 5 spots of interest identification was only possible for three spots. In the other two cases the sample amount was insufficient for conclusive identification.

Two of the three proteins identified were mitochondrial proteins not previously implicated in pancreatic cancer. The third identified protein was ubiquitin-conjugating enzyme E2 variant 1. While this particular protein has not been previously associated with pancreatic cancer, another variant of the ubiquitin-conjugating enzyme has been previously implicated in pancreatic cancer and has been linked to a T-cell mediated recognition of pancreatic cancer cell.[23] An immune response to this class of protein has therefore been shown previously suggesting that the body is responding to the cancer via the ubiquitin pathway. Such a protein could be a potential target of interest both for diagnostics and therapeutic purposes. However due to the small sample numbers used as a proof of concept set for the digested array strategy further experimentation with much larger sample pools is necessary in order to draw more biologically relevant conclusions about pancreatic cancer progression.

7.4. Conclusion

Studies designed to enhance our understanding of the host response to cancer i.e. humoral response can provide key information regarding potential markers of the cancer as well as important pathways critical to the development of the disease. However current efforts at understanding this response using microarray approaches are hampered by low response of cancer proteins on the array surface to serum samples that are used to probe the arrays. Here we present a possible explanation for this weak response and a potential strategy to overcome the problem.

It is possible that upon immobilization to a solid surface the protein on the array is unable to bind to potential binding partners in serum because the binding site on the protein is

sterically hindered. To overcome this problem, we have developed a strategy where protein is first digested using CNBr resulting in a few large fragments that are subsequently arrayed on the microarray surface. Depending on the size of the protein such a digestion could result in 2 to up to 6 fragments. Digestion into large fragments is likely critical to allow fragments to still possess the binding sites in their intact form whereas if digestion into many smaller fragments was performed, the binding site may be destroyed. Preliminary results show that digestion prior to arrays facilitates increased detection sensitivity of the overall humoral response. In addition, when the experiment was repeated with 10 normal, 10 chronic pancreatitis and 10 pancreatic cancer sera, it was observed that out of 62 spots that were arrayed, 10 spots showed stronger signals when arrayed after digestion. Out of these 10 spots, 5 showed a response that was different between normal sera and diseased sera (chronic pancreatitis and pancreatic cancer). In this study a response unique to chronic pancreatitis and pancreatic cancer was observed. This response could be indicative of inflammation rather than cancer since it was seen in both the chronic pancreatitis and pancreatic cancer sera. Regardless, it is very likely that a similar trend will be seen if all pH lanes from the first dimension are interrogated resulting in a larger number of potential proteins that can be validated for diagnostic capabilities. In addition the results from this study are used as a proof of concept for the digested array strategy, but further work with larger sample pools will be needed before any biologically relevant conclusions can be drawn.

Table 7.1: Protein identifications of spots that demonstrated a differential humoral response between the three sera sample groups used with additional information about peptides identified and coverages observed.

pH_fraction number from HPLC run	Spot loci (block, column, row on array)	Acc #	Protein ID Peptide identified	Protein Score Charge	Theoretical MW Xcorr
5.2-5.0_23	861	O60220	Mitochondrial import inner membrane translocase subunit Tim8 A	100	10992
			SKPVFSESLSD	2	3.7
5.2-5.0_36	762	Q13404	Ubiquitin-conjugating enzyme E2 variant 1 (UEV-1)	120	25781
			LLEELEEGQK	2	3.4
			YPEAPPFVR	2	2.7
5.2-5.0_47	812		Insufficient sample for IDs		
5.2-5.0_52	862		Insufficient sample for IDs		
6.6-6.4_73	637	Q13011	Delta(3,5)-Delta(2,4)-dienoyl-CoA isomerase	336	35794
			YQETFNVIER	2	3.2
			EVDVGLAADVGTLQR	2	5.1
			VIGNQSLVNELAFTAR	2	5.1
			MMADEALGSGLVSR	2	5.4

Figures:

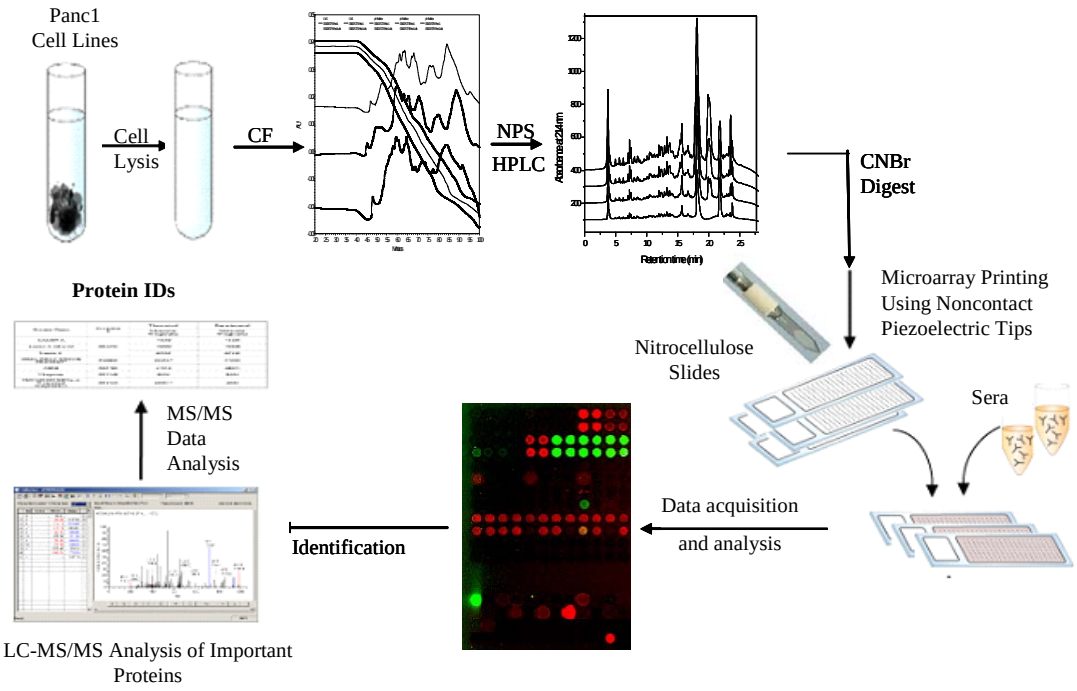
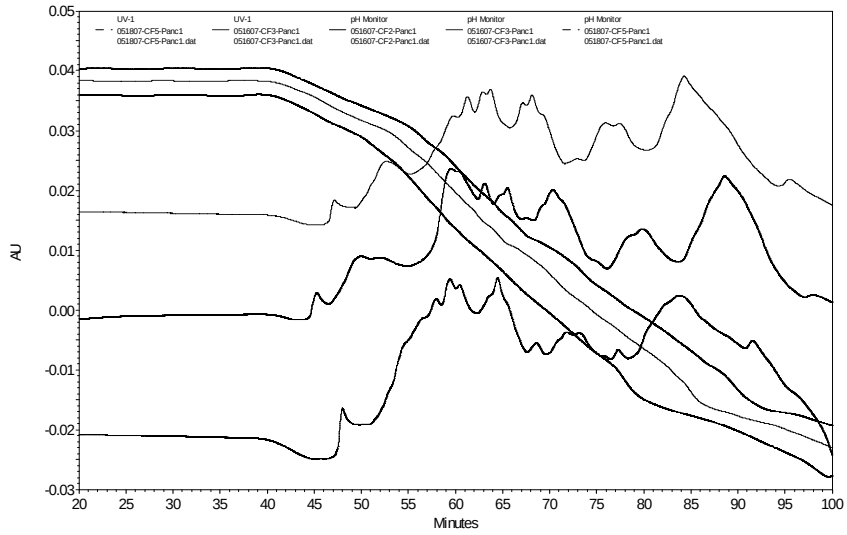


Figure 7.1: Overall workflow of the modified protein microarray strategy. Proteins from a cell line/ tissue are first extracted and separated in two dimensions (chromatofocusing separated the proteins according to their pI and NPS-RP-HPLC separated them according to their hydrophobicity). Separated fractions are split into three parts. One part is digested with trypsin, 1 with CNBr and 1 is left intact. Intact proteins and CNBr digested proteins are arrayed on nitrocellulose slides and probed with serum from different stages of disease (in this case normal, chronic pancreatitis and pancreatic cancer) to visualize humoral response. Tryptic digests of the spots that showed a differential humoral response were then subjected to protein identification using LC-MS/MS.

a)



b)

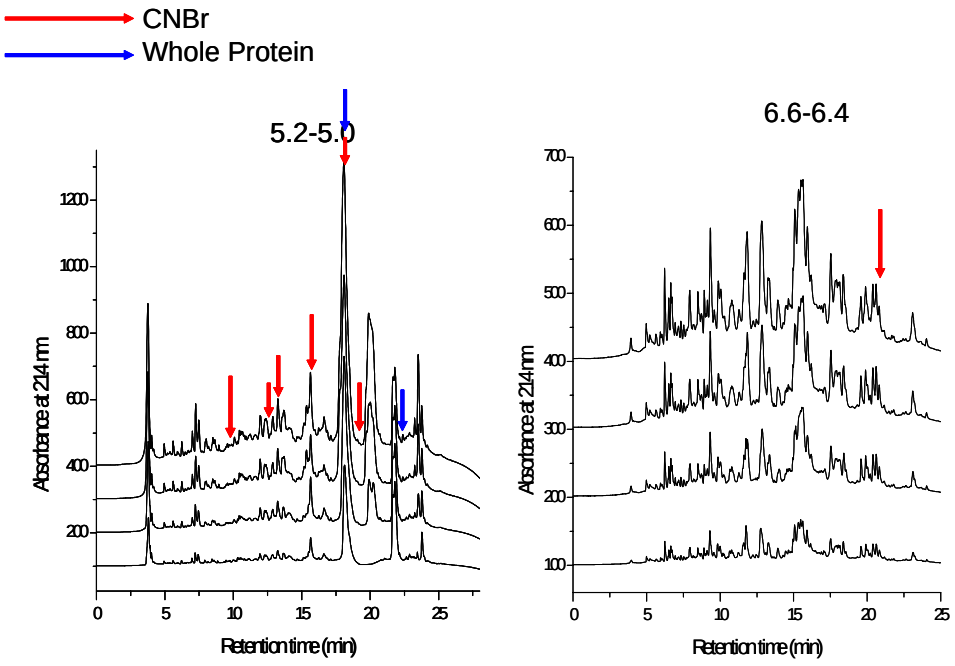


Figure 7.2: Reproducibility of separation methods used. (a) 3 chromatofocusing runs using 4.5 mg of protein lysate from Panc1 cell lines. (b) 4 reversed phase HPLC runs from two distinct pH fractions from the first dimension. Red arrows indicated fractions/peaks that responded to serum when digested by CNBr and Blue arrows indicated fractions/peaks that responded to serum when arrayed in its intact state.

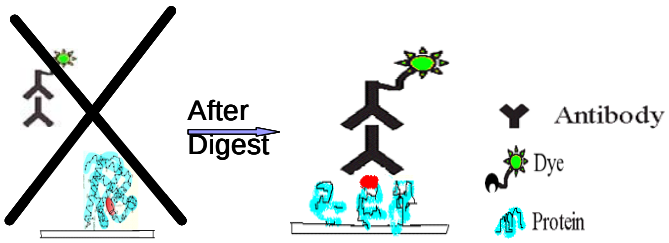


Figure 7.3: Hypothesis about why intact protein microarrays may not show high response signal. Binding site on protein is sterically hindered from serum proteins when the arrayed protein is intact. After digestion with CNBr, fragments with conserved binding sites are more exposed to serum proteins enhancing the signal due to humoral response.

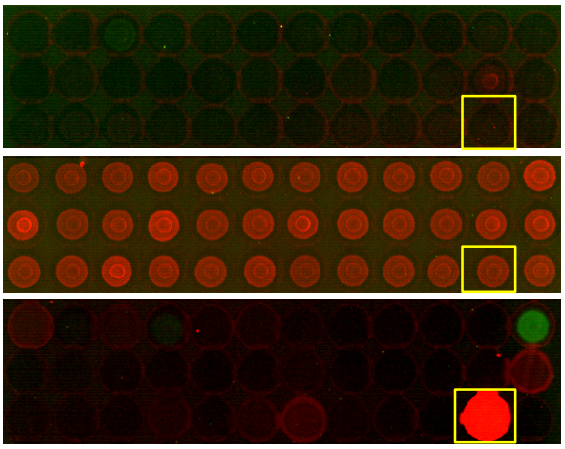
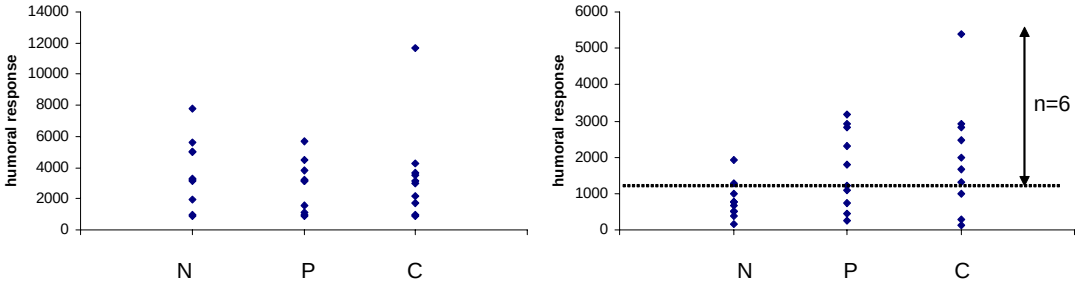
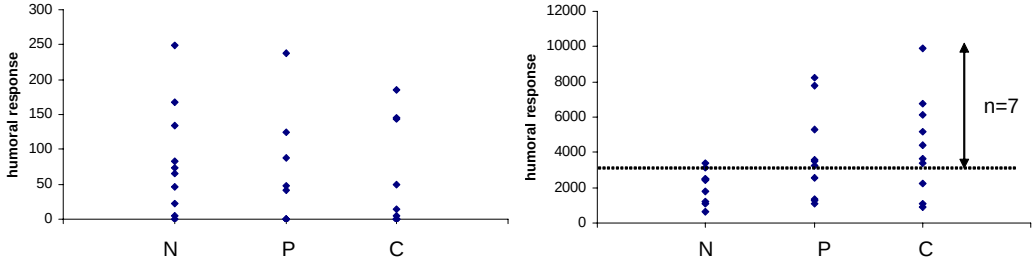


Figure 7.4: Microarray slide section illustrating differences in humoral response using 3 separate arraying methods. The top panel is intact proteins from Panc1 cell lines probed with serum resulting in very low overall response. The middle panel is GluC digested proteins from the same Panc1 cell line probed with serum resulting in a positive response to all arrayed fractions. This binding was non-specific to the GluC present in digested sample. The lower panel shows humoral response to tryptically digested proteins from the same Panc1 cell line. While the overall background is maintained at a low level, spots inside the yellow square illustrate a humoral response that was not present when the same protein in its intact state was probed with serum.

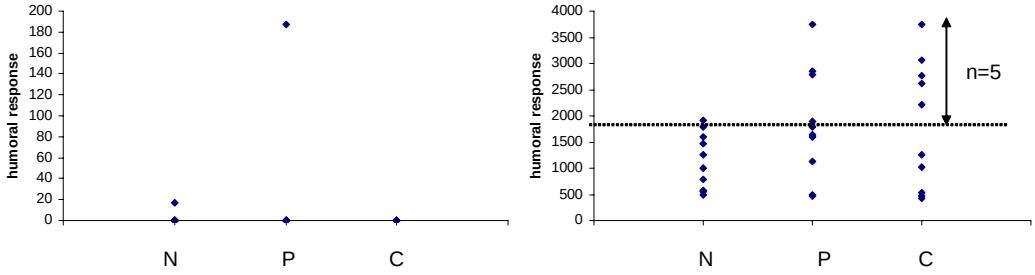
a) (812 and 316)



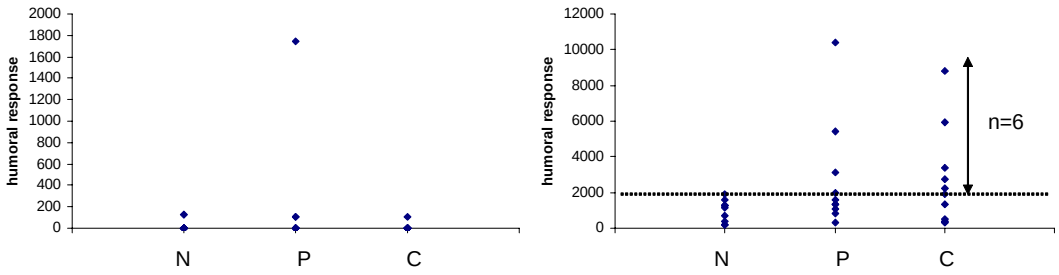
b) (762 and 166)



c) (637 and 533)



d) (861 and 365)



e) (862 and 356)

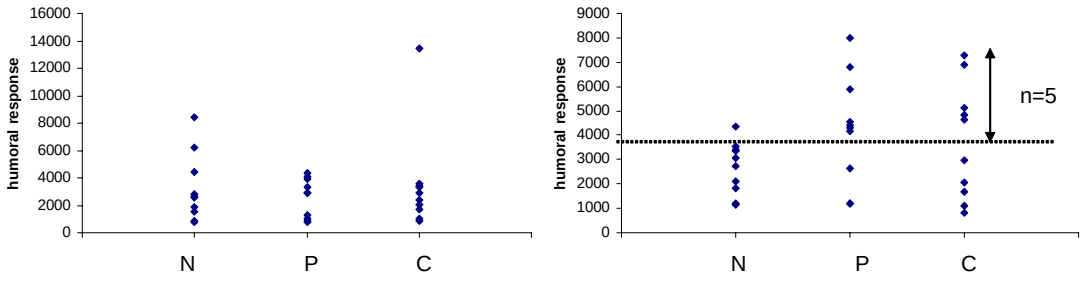


Figure 7.5: Scatter plots illustrating change in humoral response upon protein digestion with CNBr. (a) – (e) show the five spots that demonstrated differential humoral response between normal sera and pancreatitis and pancreatic cancer sera. On the left are scatter plots of all serum sample reactions to the intact spot while on the right are scatter plots of all serum sample responses to the CNBr digested spots. In all plots 1 = normal sera responses, 2 = chronic pancreatitis responses and 3 = pancreatic cancer responses.

7.5. References

[1] Spisak, S., Tulassay, Z., Molnar, B., Guttman, A., *Electrophoresis* 2007, *28*, 4261-4273.

[2] Cekaite, L., Hovig, E., Sioud, M., *Methods Mol Biol* 2007, *360*, 335-348.

[3] Hudson, M. E., Pozdnyakova, I., Haines, K., Mor, G., Snyder, M., *Proc Natl Acad Sci U S A* 2007, *104*, 17494-17499.

[4] Miller, J. C., Zhou, H., Kwekel, J., Cavallo, R., *et al.*, *Proteomics* 2003, *3*, 56-63.

[5] Orchekowski, R., Hamelinck, D., Li, L., Gliwa, E., *et al.*, *Cancer Res* 2005, *65*, 11193-11202.

[6] Shafer, M. W., Mangold, L., Partin, A. W., Haab, B. B., *Prostate* 2007, *67*, 255-267.

[7] Haab, B. B., Zhou, H., *Methods Mol Biol* 2004, *264*, 33-45.

[8] Haab, B. B., *Mol Cell Proteomics* 2005, *4*, 377-383.

[9] Haab, B. B., *Curr Opin Biotechnol* 2006, *17*, 415-421.

[10] Chatterjee, M., Draghici, S., Tainsky, M. A., *Curr Opin Drug Discov Devel* 2006, *9*, 380-385.

[11] Abu-Shakra, M., Buskila, D., Ehrenfeld, M., Conrad, K., Shoenfeld, Y., *Ann Rheum Dis* 2001, *60*, 433-441.

[12] Stockert, E., Jager, E., Chen, Y. T., Scanlan, M. J., *et al.*, *J Exp Med* 1998, *187*, 1349-1354.

[13] Caron, M., Choquet-Kastylevsky, G., Joubert-Caron, R., *Mol Cell Proteomics* 2007.

[14] Bizzaro, N., *Autoimmun Rev* 2007, *6*, 325-333.

[15] Lin, H. S., Talwar, H. S., Tarca, A. L., Ionan, A., *et al.*, *Cancer Epidemiol Biomarkers Prev* 2007, *16*, 2396-2405.

[16] Gao, W. M., Kuick, R., Orchekowski, R. P., Misek, D. E., *et al.*, *BMC Cancer* 2005, *5*, 110.

[17] Chen, G., Wang, X., Yu, J., Varambally, S., *et al.*, *Cancer Res* 2007, *67*, 3461-3467.

[18] Cekaite, L., Haug, O., Myklebost, O., Aldrin, M., *et al.*, *Proteomics* 2004, *4*, 2572-2582.

[19] Yan, F., Sreekumar, A., Laxman, B., Chinnaiyan, A. M., *et al.*, *Proteomics* 2003, *3*, 1228-1235.

[20] Taylor, B. S., Pal, M., Yu, J., Laxman, B., *et al.*, *Mol Cell Proteomics* 2007.

[21] Chong, B. E., Yan, F., Lubman, D. M., Miller, F. R., *Rapid Commun Mass Spectrom* 2001, *15*, 291-296.

[22] Wang, Y., Wu, R., Cho, K. R., Shedden, K. A., *et al.*, *Mol Cell Proteomics* 2006, *5*, 43-52.

[23] Ito, M., Shichijo, S., Tsuda, N., Ochi, M., *et al.*, *Cancer Res* 2001, *61*, 2038-2046.

Chapter 8

Conclusion

This dissertation has described the development and application of an integrated liquid separation, protein microarray and tandem mass spectrometry strategy for global screening of post translational modifications and humoral response changes that occur due to disease progression. Liquid separation techniques are used as an alternative to gel electrophoresis where dynamic range of separation is limited and where manual handling makes protein identification difficult due to introduction of contaminants such as keratins. Fractions collected by liquid separation are then arrayed on a solid surface resulting in a protein microarray. By arraying all fractions on a small surface they can all be probed simultaneously with a reagent that can highlight a property of interest. This makes our proposed strategies robust and high throughput.

The protein microarray can be used to provide information very similar to a 2D gel. However while gels are delicate and can easily break, the microarray is more rugged and not susceptible to the same problems as a gel. Furthermore in order to detect specific proteins on gels using antibodies the proteins on the gels first need to be electrotransferred onto a membrane surface since antibodies cannot permeate the pores in the gel. This step is completely bypassed in microarray experiments making the technique less time consuming and more efficient. In-gel digestion is also avoided in our

proposed strategy because part of the fraction from the liquid separation platforms can directly be coupled to mass spectrometry with minimal sample preparation steps for protein identification.

The strategy presented was successfully applied to study phosphorylation changes in pre-malignant and malignant breast cancer cell lines. Proteins from a premalignant breast cancer cell line AT1 and a malignant breast cancer cell line CA1a were first separated by a 2 dimensional liquid separation technique involving chromatofocusing and non-porous silica reversed phase HPLC. In the 1^{st} dimension proteins were separated according to their iso-electric points and in the 2^{nd} dimension they were separated by their hydrophobicities. The collected fractions (~1200 per cell line) were arrayed on amine and nitrocellulose slides and probed with the universal phosphoprotein binding dye, Pro-Q Diamond followed by antiphosphotyrosine antibodies to highlight phosphoproteins in the separated lysate. Out of the 140 spots that were positive for phosphorylation, 85 showed differential phosphorylation and these spots corresponded to 75 distinct proteins, a majority of which were high to medium abundant proteins. These differentially phosphorylated proteins were identified by tandem mass spectrometry. A total of 51 phosphorylation sites in 27 unique proteins were identified during this process. A majority of the proteins exhibiting differential phosphorylation are known to be involved in transcriptional and translational regulation as well as cytoskeletal integrity and apoptosis. Interestingly these processes are known to change considerably as a function of cancer progression. A study of a much larger scale is warranted to further probe the results obtained in this study. In particular such a study would need to focus on whether

such changes are specific to the cell lines studies or if they are a general characteristic of all types of cancers.

A slightly different liquid separation strategy was applied to study glycosylation patterns across multiple proteins in serum samples. Such patterns in normal vs. cancer sera and plasma were compared to see if classification of a sample based on its glycosylation pattern is possible. The developmental phase of this strategy involved assessing the specificity and sensitivity of lectins and their binding to glycoproteins. After successful attempts at utilizing biotinylated lectin followed by streptavidin conjugated to a fluorescent tag to detect sugar groups on proteins, the strategy was applied to clinically relevant serum and plasma samples from pancreatic and colorectal cancer patients respectively. Samples were depleted by an immunoaffinity column and then enriched for N-linked glycoproteins by a general lectin affinity chromatography step. Enriched glycoproteins were then separated and arrayed on nitrocellulose slides which were subsequently probed with 5 different lectins to assess levels of mannosylation, sialylation, fucosylation and galactosylation in the arrayed spots. Array data was statistically analyzed by principal component analysis (PCA) and hierarchical clustering (HC) and showed distinct and unique grouping of normal, chronic pancreatitis and pancreatic cancer samples with all lectins. In the case of colorectal cancer, normal samples clustered away from diseased samples but adenoma's and cancer plasma samples could not be distinguished from each other based on data from all arrayed samples. Wilcoxon rank sum tests were also performed to determine if individual fractions could be used to uniquely identify sample groups. These tests highlighted potential glycoprotein biomarkers for both pancreatic and colorectal cancer. Potential markers were identified

by mass spectrometry. Interestingly a majority of the proteins identified were liver proteins. A correlation between disease progression and changes in liver protein activity is expected in cancers. Some markers from the colorectal cancer study were validated in an independent set of plasma samples. A similar independent validation is still needed for more pancreatic cancer samples. In addition further studies that utilize antibodies to pull out potential glycoprotein markers from large amounts and numbers of serum samples would be essential to determine if the changes seen are present across a much larger sample pool. More specific glycan changes that are occurring due to disease progression also need to be studied by multiple stages of mass spectrometry in order to assess the linkages in the glycan structures that change due to disease.

Immune response to disease related proteins can also be detected using the integrated approach utilized in this dissertation. When proteins from a diseased cell are arrayed on a solid surface and subsequently probed with serum from normal or diseased patients there is a probability that antibodies against some of the disease proteins that are present in the patient serum will bind the protein against which they were formed if that protein is on the array. Such interactions can be visualized using a secondary anti-human IGg antibody. In order to analyze this immune response in pancreatic cancer, a representative cell line, MIAPACA, was fractionated in 2 dimensions (CF in the 1^{st} and NPS-RP-HPLC in the 2^{nd} dimension). The resulting fractions were arrayed and probed with multiple serum samples from controls and patients diagnosed with cancer. All spots that exhibited a humoral response were statistically analyzed. It was found that non-parametric analysis on foreground only spot intensities resulted in the best result. A panel of 9 potential biomarkers was identified. This panel had an accuracy of ~87%. Sensitivity

and specificity of the panel were found to be 93% and 80% respectively. Validation experiments were also performed on phosphoglycerate kinase and histone H4 and showed results that correlated well with initial experiments. Results with tissue samples were not as promising as those obtained with the cell line. This is likely due to the heterogeneous nature of the tissue samples used where the net amount of protein from cancer cells was probably below the sensitivity of the microarray hybridization method. Future work in this area could be more productive if laser capture micro-dissection is utilized to obtain a concentrated amount of purely cancer cells from multiple tissue samples. Initial work was also done to explore the utility of protein digestion prior to array generation. We hypothesized that arrayed whole proteins may be subject to some steric hinderance. Digesting them prior to arraying has the potential of exposing critical binding sites for humoral response. It was found that by digesting the whole protein by CNBr prior to arraying, humoral response detection sensitivity was significantly improved and overall background was reduced. However further studies across the entire pH range need to be conducted to see if the overall number of proteins eliciting a differential humoral response is increased significantly by this approach.

The strategies described herein can be successfully utilized to study a variety of sample types (cell lines, serum, plasma and tissues) in order to assess modifications and immune responses to disease associated proteins. Consequently it is a robust technique applicable to a diverse set of biological questions that can provide new and complementary information to that obtained using other platforms. However further work needs to be done to stretch the limits of detection and sensitivity of the microarray platform in order to be able to detect significantly lower levels of proteins. In addition,

large scale studies at the clinical level (utilizing a much larger sample pool) are warranted in order to assess the quality of the markers highlighted in the studies described in this dissertation.