Approaches Towards Clinical Proteomic Studies in Blood

A dissertation presented

By

Christina Orazine

To

The Department of Chemistry and Chemical Biology
In partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in the field of
Chemistry

Northeastern University
Boston, Massachusetts
Approaches Towards Clinical Proteomic Studies in Blood

By

Christina Orazine

ABSTRACT OF DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Chemistry in the Graduate School of Arts and Sciences of Northeastern University
ABSTRACT

This thesis is based in the area of clinical proteomics and relies primarily on LC/MS analytical methods to search for disease associated changes in blood. Computer programming was investigated as a solution to issues of data management and interpretation.

Chapter 1 introduces the analytical methods employed in this thesis including HPLC, mass spectrometry and protein fractionation methods. Background information regarding the disease processes of breast cancer and metabolomic disorders is also provided.

Chapter 2 describes a mouse plasma proteomic analysis aimed to uncover protein biomarkers of the growth of human breast cancer cells implanted in immunocompromised mice. M-LAC fractionation removed the serum albumin and enriched the glycoproteins prior to tryptic digestion and LC/MS analysis. Several proteins were observed to have significant changes in abundance including changes in the abundance of EGFR. Samples analyzed using the LTQ-FT provided high mass accuracy data which was used to discover tumor specific proteins with peptide sequences of human origin.

Chapter 3 describes testing of serum from gastric bypass patients, who developed hyperinsulinemia, through the use of multiple testing methods in hopes of finding the origin of the disease. Two types of abundant protein depletion, 12-protein depletion and 2-protein depletion, were used in separate experiments prior to M-LAC fractionation, tryptic digestion and LC-MS analysis for one part of the study. Of the possible proteins involved in the disease symptoms Vitamin D Binding protein (VDBP) appears as
particularly important. In the next experiment the peptidome of the serum samples was investigated to find endogenous peptides which might play a role in hyperinsulinemia. Apolipoproteins A-1, A-IV and C-III all show reduced endogenous enzymatic cleavage and are known to be associated with lipid transport and type 2 diabetes. A third experiment investigated the possibility of an autoimmune cause for the complications through the use of Western Blotting. Preliminary results from the experiment indicate that there is an autoimmune response or cause of hyperinsulinaemia in gastric bypass patients.

Chapter 4 deals with interpretation of the data produced by clinical proteomics experiments through the use of specialized computer programs. ProteinCenter speeds up identifying the biological roles of the proteins discovered during the analysis. We also explore the possibility of utilizing the LC/MS data from proteomic experiments without identifying proteins to discover disease or health patterns in the chromatographic output.

Chapter 5 discusses what improvements are imminent in the field of clinical proteomics in regards to sample quality and processing time. In the immediate future computer programming will play a significant role in the process of analyzing LCMS data.
ACKNOWLEDGEMENTS

Over the course of any long journey the traveler will meet many people who will help him or her along. I have been particularly fortunate in this regard in that there have been many people who offered guidance and support when it was needed most. Many other people have simply made the path more pleasant and less toilsome.

I would (I suspect unexpectedly) like to thank Dr. Krull for his insistence on proper notebook keeping. While it took awhile to fully implement your recommendations, I now hold my notebook to be one of my greatest resources.

Jian Zhang is someone to whom I owe great deal. Jian took the time to teach me, and was kind regardless of all else. Many times Jian would ask questions and then send me off to find the answer, and because of his efforts I am confident of my future success.

I must thank Tomas Retjar for his patience with my queries and his support through the last few years. Billy Wu was another person whom I thank for withstanding a barrage of questions. I only wish there was time to ask more questions of both of you.

It is very important that I recognize the unwavering support of Dr. Hincapie, whose belief in my abilities made me believe I could succeed. Thank you for allowing me to work alongside you.

My lab mates through the years were some of the most valuable resources. Hiato has been a trusted source of advice and information since the beginning and holds my highest regard. Majlinda first welcomed me into Dr. Hancock’s lab and taught me the M-LAC procedures. There have been many other colleagues through the years on whom I have leaned upon for support. I remember and appreciate everyone.
In truth though, my greatest debt of gratitude, within Northeastern University, goes to Dr Hancock for his enormous wealth of patience and the tremendous amount of guidance he offered to me. You have my true respect in regard to your ability to deal successfully with students, faculty and business.

To my parents and grandparents,

Throughout my life my parents were always my biggest supporters and taught me that education was the key to success. My father also was the one who rescued me the several times when I hopped on the wrong train or forgot my keys back on my desk. My grandfather was anxiously awaiting the completion of my graduate studies, but was placated by the arrival of his great-grand-daughter. Grandpa ran the best restaurant in Cambridge and filled my stomach on many occasions when I could not afford to otherwise do so. It is ultimately to both my grandfather and my grandmother that I owe my belief in myself and commitment to succeed.

To my friends,

Thank you. I am incredibly dependent on my friends and I am incredibly fortunate to have so many of the best people on the planet to support me. Generally I make only a few friends at each point in my life, but they stick around.

To Dan and Ylana,

Thank you, Dan for the breaks when I could laugh and forget the stress for awhile, and for helping me to hang on to that little tiny piece of my soul that’s left. Let’s pray it regenerates, but for now you have enough soul for all of Long Island and me too. Thank you, Ylana, for the late night pep talks. You are a true friend who was always on my side whether I was right or wrong. I would never have made it without you.
To Leatrice and Nicole,

Our get-togethers, while infrequent, have been another source of solace. You guys understood why it was so hard to go from a QC laboratory to a graduate lab.

Pam and Rob Cinelli are a couple who worked behind the scene to support me. They took up the slack when I was late to the barn, and never complained when I ran out of money; Pam even brought dinner out to the barn when I was studying for exams while watching over my horse during a bought of founder.

To my neighbors in Lincoln,

Thank you, Francine, for commiserating about academic life. Thank you also, for knocking on the window at 4 AM so that I could look at Jupiter (if it was Jupiter, I’m a bit bleary at 4 AM). Thank you Joe, Denise and Bob for all of your help. Thank you, to Dave and Bob of Hanson’s Farm and thank you, to everyone else that provided support and advice, prior to and during my years at school.

To my husband Quentin and his children Quentin Jr., and Kayla,

Thank you all for all your support. Quentin and I went to high school together but never dreamed we would meet again and start a romance ending in a marriage that we know will last forever. Quentin has been a staunch supporter and my personal IT staff. My computer has been infected and my files corrupt, but it only took an email or phone call to get everything fixed and good as new.
To my family, my friends, and especially my daughter Lena
<table>
<thead>
<tr>
<th>Table of Contents</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>2</td>
</tr>
<tr>
<td>Acknowledgements</td>
<td>5</td>
</tr>
<tr>
<td>Dedication</td>
<td>9</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>11</td>
</tr>
<tr>
<td>List of Figures</td>
<td>15</td>
</tr>
<tr>
<td>List of Tables</td>
<td>17</td>
</tr>
<tr>
<td>List of Abbreviations</td>
<td>19</td>
</tr>
<tr>
<td>Chapter 1. Introduction to Clinical Proteomics and Proteomic Analysis</td>
<td>23</td>
</tr>
<tr>
<td>Methods</td>
<td></td>
</tr>
<tr>
<td>1.1 Clinical Proteomics</td>
<td>24</td>
</tr>
<tr>
<td>1.2 Biological Mechanisms of Cancer and Diabetes</td>
<td>26</td>
</tr>
<tr>
<td>1.2.1 Genetic Markers</td>
<td>27</td>
</tr>
<tr>
<td>1.2.2 Growth Factors</td>
<td>28</td>
</tr>
<tr>
<td>1.3 The Problem of Dynamic Range in Blood When Searching For Biomarkers</td>
<td>29</td>
</tr>
<tr>
<td>1.4 Analytical Methods for Proteomic Analysis</td>
<td>30</td>
</tr>
<tr>
<td>1.4.1 Two-Dimensional Gels</td>
<td>30</td>
</tr>
<tr>
<td>1.4.2 Use of Microarray in Proteomics</td>
<td>32</td>
</tr>
<tr>
<td>1.4.3 LC-MS/MS Based Proteomic Experiments</td>
<td>33</td>
</tr>
<tr>
<td>A. Ionization Methods</td>
<td>35</td>
</tr>
<tr>
<td>B. Mass Analyzers</td>
<td>37</td>
</tr>
<tr>
<td>i. Quadrupole Mass Spectrometers</td>
<td>38</td>
</tr>
<tr>
<td>ii. TOF and QTOF Mass Spectrometers</td>
<td>39</td>
</tr>
</tbody>
</table>
### Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>iii. Ion Trap Mass Spectrometer</td>
<td>41</td>
</tr>
<tr>
<td>iv. The FT-MS and Orbitrap</td>
<td>43</td>
</tr>
<tr>
<td>C. Variations Of LC-MS Proteomic Experiments</td>
<td>47</td>
</tr>
<tr>
<td>i. Analysis of Intact Proteins (Top down Proteomics)</td>
<td>47</td>
</tr>
<tr>
<td>ii. Analysis of Enzymatically Digested Proteins (Bottom-Up Proteomics)</td>
<td>47</td>
</tr>
<tr>
<td>iii. Two Dimensional LC-MS</td>
<td>48</td>
</tr>
<tr>
<td>iv. Orthogonal Separation Techniques</td>
<td>50</td>
</tr>
<tr>
<td>v. Multidimensional Protein Experiments</td>
<td>51</td>
</tr>
<tr>
<td>1.4.4 Sample Preparation Methods for Proteomic Experiments</td>
<td>52</td>
</tr>
<tr>
<td>i. Depletion</td>
<td>52</td>
</tr>
<tr>
<td>ii. Lectins</td>
<td>53</td>
</tr>
<tr>
<td>1.5 Social/ Medical Importance of Breast Cancer and Diabetes</td>
<td>53</td>
</tr>
<tr>
<td>1.6 Proteomic Analysis of Breast Cancer</td>
<td>55</td>
</tr>
<tr>
<td>1.6.1 Relevance of Estrogen Receptor Positive Breast Cancer</td>
<td>55</td>
</tr>
<tr>
<td>1.6.2 Effect of Estrogen and Tamoxifen on Estrogen Receptor</td>
<td>55</td>
</tr>
<tr>
<td>1.6.3 Importance of EGFR (Epidermal Growth Factor)</td>
<td>57</td>
</tr>
<tr>
<td>1.7 Proteomics of Serum from Gastric Bypass Patients who have</td>
<td>59</td>
</tr>
<tr>
<td>Developed Hyperinsulinemia</td>
<td>59</td>
</tr>
<tr>
<td>1.7.1 Introduction to Hyperinsulinoma Condition Of Gastric Bypass</td>
<td>59</td>
</tr>
<tr>
<td>Patients</td>
<td></td>
</tr>
<tr>
<td>1.7.2 Depletion Methods Used in the Study</td>
<td>60</td>
</tr>
<tr>
<td>1.7.3 Peptidomic Analysis of the Plasma of Gastric Bypass Patients</td>
<td>62</td>
</tr>
</tbody>
</table>
**Table of Contents**

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.7.4 Investigation of an Autoimmune Cause of the Complications</td>
<td>63</td>
</tr>
<tr>
<td>1.8 Bioinformatics</td>
<td>64</td>
</tr>
<tr>
<td>1.8.1 Growth of Informatic and Bioinformatic Software for Proteomics</td>
<td>64</td>
</tr>
<tr>
<td>1.8.2 Quantitation of Data Via Software Programs</td>
<td>65</td>
</tr>
<tr>
<td>1.8.3 Software Designs to Aid in the Interpretation of Protein Identifications</td>
<td>68</td>
</tr>
<tr>
<td>1.8.4 Graphic Comparisons of Mouse and Human Proteomic Data Utilized to Discover Disease Patterns</td>
<td>68</td>
</tr>
<tr>
<td>1.9 References</td>
<td>70</td>
</tr>
</tbody>
</table>

Chapter 2 A Proteomic Analysis of the Plasma Glycoproteome of MCF-7 Mouse Xenograft: A Model System for the Detection of Tumor Markers

2.1 Benefits of Model System

2.1.1 Cancer and Diabetes

2.1.2 The Nude Mouse

2.2 Experimental

2.2.1 Materials

2.2.2 Xenograft Study

2.2.3 Materials and Methods for Depletion of IgG and Fibrinogen

2.2.4 Affinity Capture of Glycoproteins with M-LAC

2.2.5 Digestion of Glycoprotein Fraction

2.2.6 Reversed Phase Desalting Using HPLC

2.2.7 Proteomic Analysis by NanoLC-MS/MS

2.2.8 NanoLC-LTQ-FT Analysis
### Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.2.9 Data Processing and Analysis</td>
<td>90</td>
</tr>
<tr>
<td>2.2.10 Relative Quantitation</td>
<td>90</td>
</tr>
<tr>
<td>2.3 Results</td>
<td>91</td>
</tr>
<tr>
<td>2.3.1 M-LAC Fractionation</td>
<td>91</td>
</tr>
<tr>
<td>2.3.2 IgG and Fibrinogen Depletion Results</td>
<td>93</td>
</tr>
<tr>
<td>2.3.3 Preparation of Tryptic Peptide Digests for LC/MS Analysis</td>
<td>95</td>
</tr>
<tr>
<td>2.3.4 LC/MS Analysis of the Sample Set</td>
<td>96</td>
</tr>
<tr>
<td>2.3.5 Data Normalization through Various Methods</td>
<td>99</td>
</tr>
<tr>
<td>2.3.6 Initial Measure of Either Up or Down Regulation of Protein</td>
<td>102</td>
</tr>
<tr>
<td>Levels in Plasma in the Data Set</td>
<td></td>
</tr>
<tr>
<td>2.3.7 Identification of Tumor Specific Peptides</td>
<td>106</td>
</tr>
<tr>
<td>2.4 Discussion</td>
<td>107</td>
</tr>
<tr>
<td>2.4.1 The Nude Mouse Xenograft as a Model for Breast Cancer</td>
<td>107</td>
</tr>
<tr>
<td>2.4.2 The Plasma Proteome and Secreted Kinases</td>
<td>108</td>
</tr>
<tr>
<td>2.4.3 Ephrin</td>
<td>110</td>
</tr>
<tr>
<td>2.4.4 Differential Quantitation of Selected Murine Kinases</td>
<td>112</td>
</tr>
<tr>
<td>2.4.5 Characterization of Tumor (Human) Specific Proteins</td>
<td>115</td>
</tr>
<tr>
<td>2.5 Conclusion</td>
<td>119</td>
</tr>
<tr>
<td>2.6 References</td>
<td>120</td>
</tr>
<tr>
<td>Chapter 3 Proteomic and Peptidomic Analysis of Serum from Patients</td>
<td></td>
</tr>
<tr>
<td>Having Hyperinsulinemia After Gastric Bypass Surgery</td>
<td>126</td>
</tr>
<tr>
<td>3.1 The Challenge of Diabetes</td>
<td>127</td>
</tr>
<tr>
<td>3.1.2 Proteomic Analysis of Plasma from Gastric Bypass Patients</td>
<td>128</td>
</tr>
</tbody>
</table>
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1.3 Peptidome Analysis</td>
<td>130</td>
</tr>
<tr>
<td>3.1.4 Autoimmune Investigation</td>
<td>130</td>
</tr>
<tr>
<td>3.2 Experimental</td>
<td>131</td>
</tr>
<tr>
<td>3.2.1 Proteomic Experiments</td>
<td>131</td>
</tr>
<tr>
<td>3.2.1.1 Materials</td>
<td>131</td>
</tr>
<tr>
<td>3.2.1.2 Preparation of Human Serum Samples</td>
<td>132</td>
</tr>
<tr>
<td>3.2.1.3 Depletion of Pooled Samples for Proteomic Analysis</td>
<td>132</td>
</tr>
<tr>
<td>3.2.1.4 M-LAC Fractionation of Depleted Samples</td>
<td>133</td>
</tr>
<tr>
<td>3.2.1.5 Tryptic Digestion and Reversed Phase Desalting Using HPLC</td>
<td>133</td>
</tr>
<tr>
<td>3.2.1.6 Proteomic Analysis by Nano LC-MS/MS</td>
<td>133</td>
</tr>
<tr>
<td>3.2.1.7 NanoLC-LTQ-FT Analysis</td>
<td>133</td>
</tr>
<tr>
<td>3.2.1.8 Data Analysis</td>
<td>134</td>
</tr>
<tr>
<td>3.2.2 Peptidome</td>
<td>135</td>
</tr>
<tr>
<td>3.2.3 Auto Antigen</td>
<td>135</td>
</tr>
<tr>
<td>3.2.3.1 Depletion Method</td>
<td>135</td>
</tr>
<tr>
<td>3.2.3.2 SDS-PAGE</td>
<td>136</td>
</tr>
<tr>
<td>3.3 Results And Discussion</td>
<td>137</td>
</tr>
<tr>
<td>3.3.1 Overview of Results</td>
<td>137</td>
</tr>
<tr>
<td>3.3.2 Proteomic Analysis</td>
<td>138</td>
</tr>
<tr>
<td>3.3.2.1 Twelve Protein Depletion</td>
<td>138</td>
</tr>
<tr>
<td>3.3.2.2 Depletion of 2 Abundant Proteins from Serum Samples</td>
<td>141</td>
</tr>
<tr>
<td>3.3.3 Peptidome</td>
<td>143</td>
</tr>
</tbody>
</table>
## Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3.4 Autoimmune Analysis of Patient Serum</td>
<td>151</td>
</tr>
<tr>
<td>3.4 Conclusion</td>
<td>153</td>
</tr>
<tr>
<td>3.5 References</td>
<td>156</td>
</tr>
<tr>
<td>Chapter 4 New Developments in Bioinformatics using Computational Methods</td>
<td>159</td>
</tr>
<tr>
<td>4.1 Elimination of the Post Identification Proteomic Data Bottleneck</td>
<td>160</td>
</tr>
<tr>
<td>using Bioinformatic Tools</td>
<td></td>
</tr>
<tr>
<td>4.1.1 Programming Aimed at Improving the Speed of Interpreting</td>
<td>162</td>
</tr>
<tr>
<td>Biological Meaning in a Clinical Proteomics Study</td>
<td></td>
</tr>
<tr>
<td>4.1.2 Maximizing the Amount of Information Gained from Proteomic Data</td>
<td>164</td>
</tr>
<tr>
<td>4.1.3 Current Methods of Data Handling</td>
<td>166</td>
</tr>
<tr>
<td>4.2 Overview of Mouse Breast Cancer Study as an Example of a Large Biomarker Study</td>
<td>166</td>
</tr>
<tr>
<td>4.2.1 A Summary of the Treatment Groups Described in Chapter 2</td>
<td>167</td>
</tr>
<tr>
<td>4.2.2 ProteinCenter as a Tool to Interpret Complex Proteomic Data</td>
<td>168</td>
</tr>
<tr>
<td>4.3 Results</td>
<td>169</td>
</tr>
<tr>
<td>4.3.1 Summary of Mouse Result</td>
<td>169</td>
</tr>
<tr>
<td>4.3.2 GO Terms Comparison</td>
<td>174</td>
</tr>
<tr>
<td>4.3.3 EGFR</td>
<td>176</td>
</tr>
<tr>
<td>4.3.4 Analysis of the PFAM Domains in the Nude Mouse</td>
<td>177</td>
</tr>
<tr>
<td>4.3.5 Serpin</td>
<td>178</td>
</tr>
<tr>
<td>4.4 The Importance of Protein Pathways Found in the Mouse Data</td>
<td>181</td>
</tr>
</tbody>
</table>
Table of Contents

4.4.1 A Comparison of Biological Samples Based on the Entire LC/MS Data Set 182
4.4.2 Investigation of Disease Patterns Found in the Mouse Plasma Study 183
4.5 Experimental Procedures for the Study of Human Diabetic Plasma 188
4.6 Results of Pattern Investigation in a Diabetes Study 189
4.6.1 The Potential of Bioinformatics to Enable Disease Pattern Searching 191
4.7 Conclusion 192
4.8 References 195

Chapter 5 Future Directions 198
5.1 Analysis of Clinical Plasma and Serum Samples through the Use of Proteomic and Peptidomic Methods 199
5.2 Investigation into the Use of Computer Programming for Interpretation of Proteomic Data 200
5.3 References 202
## List of Figures

### Chapter 1

<table>
<thead>
<tr>
<th>Figure</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Two Important Methods of Ionizing Molecules for Analysis with Mass Spectrometry.</td>
<td>33</td>
</tr>
<tr>
<td>1.2</td>
<td>Quadrupole Mass Analyzer Design.</td>
<td>36</td>
</tr>
<tr>
<td>1.3</td>
<td>Fundamental Design of a TOF Analyzer.</td>
<td>38</td>
</tr>
<tr>
<td>1.4</td>
<td>Schematic of the LTQ.</td>
<td>43</td>
</tr>
<tr>
<td>1.5</td>
<td>Ion Excitation in FT-MS.</td>
<td>41</td>
</tr>
<tr>
<td>1.6</td>
<td>A Cutaway View of the Orbitrap Showing an Ion Orbiting the Inner Electrode.</td>
<td>43</td>
</tr>
<tr>
<td>1.7</td>
<td>Design of Mouse Xeonograph Experiment.</td>
<td>57</td>
</tr>
<tr>
<td>1.8</td>
<td>Protein Depletion Column used to Extend Dynamic Range.</td>
<td>59</td>
</tr>
<tr>
<td>1.9</td>
<td>Process of Peptide Identification.</td>
<td>63</td>
</tr>
</tbody>
</table>

### Chapter 2

<table>
<thead>
<tr>
<th>Figure</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Experimental Design for the Study of the Ncr Nude Mouse Xenograph as a Model of Human Breast Cancer.</td>
<td>82</td>
</tr>
<tr>
<td>2.2</td>
<td>Tumor Growth of MCF-7 Xenografts for the 4 Treatment Groups Over the Course of 6 Weeks.</td>
<td>92</td>
</tr>
<tr>
<td>2.3</td>
<td>Results of Immunoglobulin Depletion Experiments.</td>
<td>94</td>
</tr>
<tr>
<td>2.4</td>
<td>Selection of the Appropriate M/Z Range for Peak Area Quantitation.</td>
<td>105</td>
</tr>
<tr>
<td>2.5</td>
<td>Ephrin A3 Peptides Discovered in the Proteomic Analysis of Mouse Plasma during the LC-MS Analysis</td>
<td>111</td>
</tr>
<tr>
<td>2.6</td>
<td>Effects of Estrogen on Mouse (M) and Mouse with Tumor (Tu).</td>
<td>112</td>
</tr>
</tbody>
</table>

### Chapter 3
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Illustration of Gastric Bypass.</td>
<td>127</td>
</tr>
<tr>
<td>3.2</td>
<td>Comparison of Proteomic Results After 2 Protein Depletion to Those of the Peptidome.</td>
<td>148-149</td>
</tr>
<tr>
<td>3.3</td>
<td>Results of the Autoimmune Western Blot Experiment.</td>
<td>153</td>
</tr>
<tr>
<td></td>
<td>Chapter 4</td>
<td></td>
</tr>
<tr>
<td>4.1</td>
<td>Experimental Design for the Study of the Ncr Nude Mouse Xenograph as a Model of Human Breast Cancer.</td>
<td>167</td>
</tr>
<tr>
<td>4.2</td>
<td>Venn Diagram of the Proteins Identified in Single Analyses of the 6-Week Estrogen With Tumor Group and Estrogen Without Tumor Group.</td>
<td>171</td>
</tr>
<tr>
<td>4.3</td>
<td>Proteins Identified in Estrogen Tumor (6-Week) Only.</td>
<td>172</td>
</tr>
<tr>
<td>4.4</td>
<td>Proteins Identified in Estrogen Tumor (6-Week) and Estrogen No-Tumor (6-Weeks).</td>
<td>173</td>
</tr>
<tr>
<td>4.5</td>
<td>Analysis of the GO Cellular Component Annotations.</td>
<td>174</td>
</tr>
<tr>
<td>4.6</td>
<td>The EGFR Protein Card from Protein Center.</td>
<td>176</td>
</tr>
<tr>
<td>4.7</td>
<td>Comparison of Serpins Present in Mice having a MCF-7 Tumor Implant Versus those Without a Tumor Implant.</td>
<td>178</td>
</tr>
<tr>
<td>4.8</td>
<td>Cellular Location and Protein Interactions of some Proteins Found in the Mouse Breast Cancer Model.</td>
<td>180</td>
</tr>
<tr>
<td>4.9</td>
<td>Replicate Extracted Ion Chromatograms from the Mouse Xenograft Proteomic Study.</td>
<td>185</td>
</tr>
<tr>
<td>4.10</td>
<td>Comparison of Patterns Found in the LC/MS Analysis of Human Plasma</td>
<td>190-191</td>
</tr>
</tbody>
</table>
List of Tables

Chapter 1

Table 1.1. Attributes of Several Popular Mass Spectrometers 34

Chapter 2

Table 2.1. Abundant Proteins Found in Both the Nude Mouse Control Sample and a Human Non Disease Sample 98
Table 2.2. Representative Results of Fetuin Peak Area Analysis 99
Table 2.3. Differential Quantitation of Proteins in 6 Week Old Mice with MCF-7 Implants 101
Table 2.4. Differential Quantitation of Murine Signaling Proteins Over Time in the Different Treatment Groups 103
Table 2.5. Tumor Specific (Human) Proteins Identified in the 6-Week Group of Mice Treated with Estrogen 118

Chapter 3

Table 3.1. Shift in the Glycosylation of Vitamin D Binding Protein 139
Table 3.2. Proteins Up or Down Regulated in Symptomatic Post GP Compared to Obese Control. 144
Table 3.3. Composite Table of Peptidome and 2 Protein Depletion Proteomic Results. 150
List of Abbreviations

**2D LC MS** - Two dimensional liquid chromatography coupled to a mass spectrometer

**2D DIGE** - Two-dimensional difference gel electrophoresis

**2D HPLC** - Two dimensional high performance liquid chromatography

**2D PAGE** - Two dimensional polyacrylamide gel electrophoresis

**AAT** - Alpha-1-antitrypsin

**AC** - Alternating current

**AcCN** - Acetonitrile

**ADP** - Adenosine diphosphate

**APCI** - Atmospheric pressure

**ATCC** - American Type Culture Collection

**ATP** - Adenosine triphosphate

**BCA** - Bichinchoninic acid

**BMI** - Body Mass Index

**BRCA1** - Breast cancer 1

**BRCA2** - Breast cancer 2

**BSA** - Bovine serum albumin

**CA125** - Cancer antigen 125

**CCP** - Complement control protein modules

**CID** - Collision-induced dissociation

**Con-A** - Concanavalin A

**CPAS** - Computational proteomics analysis system

**CSF1R** - Macrophage colony-stimulating factor receptor

**Da** - Dalton

**DC** - Direct current

**DNA** - Deoxyribonucleic acid

**DNA-PK** - DNA-dependent protein kinase 4

**EDTA** - Ethylenediaminetetraacetic acid

**EGFR** - Estrogen growth factor receptor

**ELISA** - Enzyme linked immunosorbent assay
List of Abbreviations

EMBL-European Molecular Biology Laboratory
ER+-Estrogen receptor positive
ESI-Electrospray ionization
FT-MS-Fourier transform mass spectrometry
GEM-Genetically engineered model
GLP-1-Glucagon-like peptide-1
GO-Geneontology
HCl-Hydrochloric acid
HER-2-Human estrogen receptor 2
HLPP-Human Liver Proteome Project
HPLC-High performance chromatography
HPRD-Human Protein Reference Database
HRP-Horse radish peroxidase
HSA-Human serum albumin
HuPO-Human Proteome Organization
ICAT-Isotope coded affinity tag
ICR-Ion cyclotron resonance
IGF-I-Insulin-like growth factor
IGF-Insulin growth factor
IGF-IR-Insulin-like growth factor receptor
IGF-R-Insulin growth factor receptor
IPI-International Protein Index
IRS2-Insulin receptor substrate 2
kDa-Kilodalton
kV-Kilovolt
LC-Liquid chromatography
LIFR-Leukemia inhibitory factor receptor
m/z-mass-to-charge ratio
MALDI-Matrix assisted laser desorption ionization
mg-Milligram
List of Abbreviations

**MIPS** Interaction-Mammalian Protein-Protein Interaction Database
**M-LAC**-Multi-lectin affinity chromatography
**mRNA**- micro ribonucleic acid
**MS/MS**-Tandem mass spectrometry (two stages)
**MWCO**-Molecular weight cutoff
**MW**-Molecular weight
**NCBI**-National Center for Biotechnology Information
**NF**-kappa B-nuclear factor kappa-light-chain-enhancer of activated B cells
**NRDB**-Nonredundant database
**PBS**-Phosphate buffered Saline
**PDB**-Protein Data Base
**PFAM**-Protein families
**PHLD**-phosphatidylinositol-glycan-specific phospholipase D
**PI3 kinase**- Phosphoinositide 3-kinase
**PI3-K**-Phosphoinositide-3 kinase
**PIR**-Protein Information Resource
**PNA**-Polynucleic acid
**ppm**-Part-per-million
**PSA**-Prostate specific antigen
**PTM**-Post translational modification
**QTOF**-Quadrupole time of flight
**RF**-Radio frequency
**RPLC**-Reversed phase liquid chromatography
**RPM**-Rotations per minute
**s/n**-Signal to noise
**SCR**-Short consensus repeats
**SDS-PAGE**-Sodium dodecylsulfate- polyacrylamide gel electrophoresis
**Serpin**-Serine protease inhibitors
**SGD**-Saccharomyces Genome Database
List of Abbreviations

**Shc 1**- (Src homology 2 domain containing) transforming protein 1,
**SILAC**- Stable isotope labeling with amino acids in cell culture
**SOP**- Standard operating procedure
**TAIR**- The Arabidopsis Information Resource
**TBST**- Tris-Buffered Saline Tween
**TFA**- Trifluoroacetic acid
**TIC**- Total ion chromatogram
**TOF**- time of flight
**TRIS**- tris (hydroxymethyl)-aminomethane
**Trisma**- tris (hydroxymethyl)-aminomethane
**t**- Time
**V**- Volts
**WGA**- Wheat germ agglutinin
**z**- Charge
**ω**- Angular frequency
Chapter 1. Introduction to Clinical Proteomics and Proteomic Analysis
Methods
1.1 Clinical proteomics

While proteomics refers to the global study of proteins, clinical proteomics aims to study proteins in the hopes of advancing the study and treatment of diseases. Research in genomics has centered on the sequencing the human genome in hope of discovering the source of all human function and dysfunction. Cancer treatment includes drugs which are aimed at specific biological function such as inhibiting cell division with Tamoxifen or preventing new blood vessel growth with Avastin. When it was realized that genes are only the blueprints to proteins and that the function of the finished product is subject to changes in amino acid sequence and post translational modifications, researchers refocused on studying proteins as the next step towards the goal of understanding and alleviating human disease.

The potential clinical uses of proteomics are well understood and advancements are anticipated from the large scale effort being currently employed. Prostate specific antigen (PSA) and cancer antigen 125 (CA125) are well-known examples of some of the proteins which are currently in clinical use as biomarkers to aid in the diagnosis of cancer or other diseases. These early examples have issues with poor specificity (PSA) and inability to aid in early diagnosis (CA 125).

Improving the initial diagnosis of disease is only one of several drivers of research into protein biomarkers. Disease specific proteins that can be determined to show changes in amount or structure versus normal response to disease may become a potential target for drug design. Other changes in protein post translational modifications may indicate a specific disease type and thus aid physicians in selecting the correct “personalized” treatment. Proteomic changes may be able to determine whether the
prescribed treatment of a disease is effective, or to project the expected time period of survival, i.e., if one were to find a particular kinase which indicated cancer progression on increasing protein abundance, it would be meaningful to monitor the kinase in order to determine whether treatment was slowing disease progression.

Despite the proven use of protein biomarkers, relatively few have progressed successfully from research to the hospital laboratory. Reasons for this failure are various. In many cases a single easily validated marker may not even exist. If one does exist it may be specific to a certain time point or sub type of the disease. Work in the field continues in the hope that these rare specific biomarkers do indeed exist and that better analytical methods will enable them to be discovered. Other candidate proteins with the potential to be medically useful in disease treatment and detection languish while waiting for the more rigorous and expensive process of validation. Analytical validation includes the requirement that it be possible to reproducibly and accurately detect the protein biomarker in clinical samples in different laboratories. Clinical validation includes finding reproducible results in a large disease sample set in a blinded study.

Because over 60% of the total protein concentration in plasma is albumin, it is nearly impossible to detect lower concentration proteins without employing a form of depletion. Added to the fact of an overabundance of albumin is the additional challenge that the lowest level proteins will have concentrations several orders of magnitude lower than albumin. Depletion and fractionation methods have developed into many diverse forms with the most widely utilized being those based on antibodies.

Other methods simplify analysis of the plasma proteome using dyes, chelated metal cations, lectins, phenylboronates, protein A and G, calmodulin or heparin, to name
a few. Antibodies have a primary function of removing unwanted high abundance proteins.

Plasma consists of an extremely complex mixture of proteins that are released from a variety of tissues including leukocytes and thrombocytes, and thus it is a challenge to understand the site of synthesis of a given protein. For cancer, one anticipates the release of proteins associated with the development of the tumor and associated stromal tissue into blood, and that such proteins can give a signature for early diagnosis. Examples of such processes include changes in the microenvironment of the tumor due to complex interaction networks between the cancer cells and the host. The release of tumor specific proteins can be related to processes such as the increased synthesis and secretion of glycoproteins, the cleavage of matrix or membrane associated proteins and the release of intracellular proteins from cells that have undergone apoptosis.

1.2 Biological Mechanisms of Cancer and Diabetes

Just as important as the search for biomarkers is the investigation of disease mechanisms. There are several well-studied biological mechanisms for cancer and diabetes. For example, upon binding of estrogen to its growth factor receptor triggers tumor growth via several pathways including the phosphorylation of Phosphoinositide 3-kinases (PI3 kinase). Another disease for which one of the underlying mechanisms is well known is diabetes and related metabolic diseases. Elevated insulin levels (hyperinsulinaemia) and depressed insulin levels (hypoinsulinaemia) are both characterized by changes in the amounts of insulin present. Both may be traceable to malfunctioning or absent beta cells. An overabundance of insulin potentially from
expansion of beta cells such as insulinoma will result in hypoglycemia. Type 1 diabetes and Type 2 are linked to low blood insulin and high blood sugar\textsuperscript{7,8}. Hyperinsulinaemia, insulin resistance, post-menopausal breast cancer, colon cancer and kidney cancer have all been connected back to excess body weight defined as overweight (25 to 29.9 kg/m\textsuperscript{2}) or obesity (BMI $\geq$ 30 kg/m\textsuperscript{2}).

Possible pathophysiological and biological mechanism associations between these diseases include insulin-like growth factors, sex steroids, adipokines, obesity-related inflammatory markers, the nuclear factor kappa beta (NF-kappa B) system and oxidative stresses\textsuperscript{9,10,11,12}. Yakar S., et al., 2006, concluded that diet-induced obesity increased tumor prevalence and tumor growth in both male and female mice\textsuperscript{13}. Another link between diabetes and cancer is the effects of certain hormones. Ovariectomy, which reduces estrogen levels and thus the occurrence of estrogen receptor positive breast cancer, was found also to affect both insulin sensitivity and glucose tolerance, which are factors in diabetes\textsuperscript{14}.

1.2.1 Genetic Markers

Cancer can be caused by mutations in oncogenes and tumor-suppressor genes. The mutations are generally isolated, and individual events, although certain gene types which are inherited, may predispose a person to cancer. Recent evidence has indicated the possibility of multiple sites of mutation in oncogenes, tumor-suppressor genes, or microRNA genes in cancer cells\textsuperscript{15,16,17}.

Tumors can contain a spectrum of different genetically-altered cells that can lead to differences in the clinical behavior and the response to treatment of the tumor. The response to chemotherapy, radiotherapy, and other treatments can be different for the
different types of cancer cells leading to difficulties in combating the tumor, i.e., a breakthrough in the treatment of prostate cancer does not necessarily aid in treating brain tumors. For these reasons the study of the biological mechanisms leading to tumor growth are important in the development of rational cancer treatments.

Germ line mutations in the BRCA1 and BRCA2 genes are associated with a 50 to 85 percent lifetime risk of breast cancer, ovarian cancer, or both. It is currently possible to test for these mutations and the development of genetic counseling for those carrying these mutations has begun to be available. In sporadic breast cancer, genetic abnormalities have been identified in several genes (including p53, bcl-2, c-myc, and c-myb), and in some cancers normal genes or gene products (HER-2/neu and cyclin D1) are overexpressed.

1.2.2 Growth Factors

Cells are able to respond to a wide variety of factors that stimulate cell growth, division and adhesion. Gonadal steroid hormones (estrogens, progestins, and androgens), growth factors (epidermal growth factor, transforming growth factors α and β, and insulin-like growth factors I and II), and various cytokines and lymphokines influence the behavior and phenotypic expression of breast cells. These proteins and their binding partners are prime drug targets to stop or control the growth of the tumor. The recognition that these factors influence the growth and dissemination of breast cancer has provided new targets for therapeutic and preventive intervention. Breast cancer also induces neovascularization, which, in turn, facilitates the metastatic process. For these reasons some new cancer treatments have been aimed at the prevention of new blood vessels. Metastatic spreading of cancer is not a random mechanical phenomenon, but requires
systematic interactions between breast cells, stroma, and surrounding normal tissue at both primary and metastatic sites. Adhesion molecules, local mediators, hormones, and growth factors must all act for metastases to develop. On the basis of this new information, diagnosis and treatment have changed. Many new cytotoxic and hormonal agents have emerged from new biologic concepts and are being developed for clinical use to treat malignant tumors.

1.3 The Problem of Dynamic Range in Blood When Searching for Biomarkers

The normal concentration of cancer biomarkers in blood is expected to be in the range of nanograms per milliliter \(^\text{19}\). Neither 2D PAGE, 2D DIGE, LC-MS/MS nor LC/LC-MS/MS has the ability to detect many proteins in this range. Currently, it is predicted that the most disease specific protein changes in blood will be those having the lowest concentrations and will be involved in cellular processes. This fact has re-energized the drive to increase sensitivity and selectivity to aid the search for biomarkers. Serum albumin which makes up over 60% of the protein concentration in plasma has been found to bind low molecular weight disease biomarkers \(^\text{20}\). Despite these considerations the advantages of depleting albumin and other high abundance plasma proteins outweighs any disadvantage in plasma proteomics \(^\text{21}\). Albumin has a molecular mass larger than the filtration cutoff of the kidneys, and thus exists for around 21 days in plasma. The normal healthy concentration range for albumin in humans is between 35-50 mg/ml. This amount decreases with the occurrence of liver disease or malnutrition. Albumin, along with IgG, IgA, transferrin, haptoglobin, and antitrypsin compose the top six most abundant proteins in human plasma \(^\text{19}\). IgG and IgA proteins are an important
part of the immune response. Transferrin and haptoglobin are acute phase proteins involved in inflammation.

On the other end of the spectrum are the low level (pg/mL) growth factors, cytokines and interleukin proteins. Proteins found in the plasma may have various origins prior to being released into the blood stream. Some of the possible sources are tissue leakage proteins, aberrant secretions, and foreign proteins, cleavage of matrix and membrane-bound proteins and release of cytoplasmic proteins from leukocytes and macrophages that have undergone apoptosis.  

1.4 Analytical Methods for Proteomic Analysis

1.4.1 Two-dimensional gels

In many ways, two-dimensional (2D) gels led the quest for biomarkers in proteomics up until the recent maturation of LC-MS technologies. Two-dimensional gel electrophoresis enables the simultaneous separation of thousands of proteins and their isoforms from plasma, tissue or cell lines. This method has been used as a protein separation step followed by Western blotting in a study of hepatocellular cancer aiming to discover autoantibodies. Another study compared the proteins of normal and neoplastic human fibroblasts using 2D gel electrophoresis combined, and they were able to discover novel proteins involved in the conversion of normal fibroblasts to tumorigenicity.

Adoption of the 2D gel methods in the United States has been slower than anticipated due to issues with reproducibility between gels, insoluble proteins that cannot be analyzed, and the difficulty of automating the process of cutting protein spots from the gels.
gel for digestion. Many software packages have been developed to process and analyze 2-D gel images. Processing time for gels varies depending on the exact method and skill of the investigator from approximately 1 day to several days. Spots matched automatically to spots from other gels using software resulted in a lower percentage of variability versus those matched by hand. Furthermore, the number of correct matches decreased significantly as the number of gels increased 27.

The limited dynamic range of 2D gel electrophoresis reduces the method’s ability to resolve proteins having a wide range of concentrations and solubility. Thus depletion and enrichment strategies for serum and plasma are playing an important role in the development of 2D electrophoresis for detection of protein biomarkers. A study by Echan, et al., in 2005 investigated the benefits of 6 abundant protein depletion versus depletion of immunoglobulins or no depletion at all 28. A 6 protein depletion method was able to remove around 85% of the proteins making it possible to load 10-fold to 20-fold more serum or plasma onto the gels. This unfortunately did not result in the detection of low abundance proteins which was attributed in part to proteins which were insoluble within the pH range of the gel. These issues may be solved by running several gels of different pH ranges and molecular weight ranges, or additional depletion such as top 12 protein depletion may also be feasible 29.

An important advantage of 2D gels is the ability to separate isoforms of proteins based on altered isoelectric point. Many diseases are known to be related to specific forms of a protein. Changes in glycosylation are associated with cancer and will cause a shift in the isoelectric points of a protein 30, 31. In spite of these advantages we chose to use an LC-MS-based approach to enable a greater number of protein identifications in a
shorter time period than 2D gel analysis methods would require.

1.4.2 Use of Microarray in Proteomics

An offshoot of DNA microarrays and ELISA technology, protein microarrays analyze protein interactions, typically antibody–antigen recognition. Genetic variations can be detected through the use of DNA microarrays by binding or constructing on a suitable surface such as glass, an array of DNA or PNA (polynucleic acid) of known sequence which will then bind to complementary sequences from a test sample. DNA microarray technology can be used to analyze gene expression levels via DNA or RNA immobilized onto a solid support. A study of 964 breast tumor samples was used to validate DNA microarray use in the long-term assessment of relapse-free survival and sensitivity to chemotherapy.\(^{32}\)

Protein microarrays are composed of proteins which have been immobilized on a solid support, either via a covalent linkage or through the utilization of a streptavidin coated plate and a biotin labeled protein. Either antibodies or antigens may be attached to the support which itself may be composed of glass, polymer or other appropriate material.\(^{33, 34, 35, 36}\)

Similar to the ELISA, the analyte protein may be tagged with a reporter ion, but it is more common to detect bound proteins using a second antibody.\(^{37}\) Useful with a number of different detection methods, isotope-labeled antigens or isotope-labeled antibodies may be added during cell culture. One popular method of protein labeling is SILAC (Stable isotope labeling with amino acids in cell culture) where the cells are cultured with only labeled amino acids.\(^{38}\) Also, some secondary antibodies are available.
as conjugates with horseradish peroxidase (HRP) or dyes\textsuperscript{37}.

\subsection*{1.4.3 LC-MS/MS Based Proteomic Experiments}

There are several forms of LC-MS proteomic experiments, but most approaches attempt to separate a complex mixture of peptides after an enzymatic digest using some form of chromatography in order to facilitate identification in the mass spectrometer. Reversed phase stationary phases continue to be the most popular columns to couple to MS because the elution mobile phases such as an aqueous-formic acid mixture are compatible with electrospray ionization\textsuperscript{39,40}.

Chromatography of whole proteins from complex mixtures like tissue, cell lysates or plasma, can be used to generate fractions for off-line MS analysis. This method allows for the use of those types of chromatography which are unsuitable for use with mass spectrometer either due to high flow rate or mobile phase incompatibility. Prior to injection on to the mass spectrometer samples can be subjected to a desalting step using a reversed phase media.

Many LC-MS/MS proteomic platforms include a prefractionation step prior to tryptic digestion and analysis to reduce sample complexity. Even this extensive amount of separation may not give the necessary peak capacity, which in a plasma proteomic experiment would need to be in the thousands, to resolve all of the components. Multi-dimensional LC experiments consisting of a strong cation exchange column coupled to a reversed phase column offer of increased separation of complex samples. Also, both one-dimensional and 2-dimensional LC experiments can be performed in part or in whole off line from the mass spectrometer. Collection of tryptic peptides into separate fractions
as they elute from a HPLC column can allow mass spectral analysis of only those peaks which are of potential interest in a direct infusion study. This method of analysis saves instrument analysis time on the MS and in processing time as data analysis is only performed on those spectra generated from selected fractions.

A. Ionization Methods

There are several methods to ionize the sample molecules for introduction into the mass spectrometer. Atmospheric pressure chemical ionization (APCI), electrospray ionization (ESI) and Matrix assisted laser desorption ionization (MALDI) are each commonly used methods of ionization which each possess individual advantages and disadvantages. The first, APCI, is useful when working at flow rates in the macro analytical range (around 1 ml/min). This method uses high temperature to vaporize the sample and a flow of inert gas to carry the vapor to the inlet where the sample is ionized by an electrical discharge.

Electrospray ionization (ESI) is typically used with narrow bore columns and low flow rates; in addition, ion trap mass spectrometers are often paired with electrospray ionization sources which produce multiply charged ions which are amenable to fragmentation. Due to the lowered m/z ratio, multiply charged ions effectively extend the mass range of the instrument.
Figure 1.1. Two Important Methods of Ionizing Molecules for Analysis with Mass Spectrometry. (a) Electrospray ionization (ESI), where voltage is applied directly to the mobile phase, is popular in online and some offline proteomic experiments due to its applicability to low flow rate situations encountered when using narrow bore columns. (b) Matrix assisted laser desorption ionization (MALDI) is useful for offline applications where a singly charged molecular ion and noncontinuous introduction of the sample are desired. A pulse from a laser causes the acidified sample to be released from the matrix spot and be drawn into the mass spectrometer by negatively charged plates.

The sample molecules are subjected to an electrical charge before entering the inlet of the mass spectrometer (Figure 1.1a). This charge is sufficient to cause the formation of an aerosol at flow rates in the micro liter range or lower. Multiply charged droplets formed at the outlet of the column break apart due to the repulsion of the negative charges. While the use of 100 µm I.D. or smaller columns effectively increases the analyte concentration and thus aid detection by the mass spectrometer, the
establishment of a stable spray can be problematic.

MALDI (Matrix Assisted Laser Desorption Ionization) is the third common method of ionization and differs significantly from the first two in that it is performed offline from the separation. The sample is mixed with an acidic matrix solution, placed in small droplets onto a reflective metal plate and allowed to dry. After drying, the plate is placed into the mass spectrometer (typically a time of flight instrument) where a laser is fired at the spots causing the sample to be released from the plate for analysis (Figure 1.1b). This process is used to generate a parent ion with a plus 1 charge (M+H). MALDI is useful for confirming the identity of molecules via its molecular weight but can require careful attention because of the difficulty in interpreting the mass of proteins having posttranslational modifications.

<table>
<thead>
<tr>
<th></th>
<th>Quadrupole</th>
<th>Ion Trap</th>
<th>Time-of-Flight</th>
<th>Time-of-Flight Reflection</th>
<th>FTMS</th>
<th>Quadrupole-TOF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.01% (100 ppm)</td>
<td>0.01% (100 ppm)</td>
<td>0.02 to 0.2% (200 ppm)</td>
<td>0.001% (10 ppm)</td>
<td>&lt;0.0005% (&lt;5 ppm)</td>
<td>0.001% (10 ppm)</td>
</tr>
<tr>
<td>Resolution</td>
<td>4,000</td>
<td>4,000</td>
<td>8,000</td>
<td>15,000</td>
<td>100,000</td>
<td>10,000</td>
</tr>
<tr>
<td>m/z Range</td>
<td>4,000</td>
<td>4,000</td>
<td>&gt;360,000</td>
<td>10,000</td>
<td>10,000</td>
<td>10,000</td>
</tr>
<tr>
<td>Scan Speed</td>
<td>–a second</td>
<td>–a second</td>
<td>milliseconds</td>
<td>milliseconds</td>
<td>–a second</td>
<td>–a second</td>
</tr>
<tr>
<td>Tandem MS</td>
<td>MS² (triple quad)</td>
<td>MS⁶</td>
<td>MS²</td>
<td>MS²</td>
<td>MS²</td>
<td>MS²</td>
</tr>
<tr>
<td>Tandem MS Comments</td>
<td>Good accuracy Good resolution Low-energy collisions</td>
<td>Good accuracy Good resolution Low-energy collisions</td>
<td>Not generally applicable</td>
<td>Precursor ion selection is limited to a wide mass range; growing number of applications</td>
<td>Excellent accuracy and resolution of product ions</td>
<td>Excellent accuracy Good resolution Low-energy collisions High sensitivity</td>
</tr>
<tr>
<td>General Comments</td>
<td>Low cost Ease of switching pos/neg ions</td>
<td>Low cost Ease of switching pos/neg ions Well-suited MSⁿ</td>
<td>Low cost</td>
<td>Good accuracy Good resolution</td>
<td>High resolution, MS² high vacuum, super conducting magnet, expensive</td>
<td>Known for high sensitivity and accuracy when used for MS²</td>
</tr>
</tbody>
</table>

Table 1.1. Attributes of Several Popular Mass Spectrometers. Adapted from reference 41.
B. Mass Analyzers

Identification of a protein biomarker during mass spectral analysis relies in part on the mass accuracy, resolution and scan speed of the mass spectrometer used in the experiment. Currently there are several different types of mass spectrometers on the market and each type has its advantages and disadvantages. Different types of mass spectrometers are compared (See Table 1.1) to each other using mass resolving power, mass accuracy, mass range and speed of analysis. The greater an instrument’s ability to separate ions having different m/z values the greater its resolving power will be. To determine resolving power one uses the equation: \( R = \frac{M}{\Delta m} \) where \( R \) is resolving power, \( M \) is the m/z value of one of two known ions differing in mass, and \( \Delta m \) is the mass difference between two ions. The ratio between the measured mass of an ion to its true mass is termed the mass accuracy of the instrument. Instrument mass range is determined by the range of mass-to-charge ratios which can be measured. An instrument’s duty cycle is the fraction of the ions of interest that enter the mass analyzer and are detected. Instrument speed is the number of spectra per unit time that can be generated (spectral generation rate described in Hertz).
A quadrupole mass analyzer, shown in Figure 1.2, consists of four parallel surfaces, usually in the shape of a cylindrical rod, and opposing poles are electronically linked. Radio frequency (RF) and direct current (DC) power sources are utilized to form a two-dimensional electrical field to suspend and guide ions. The ratio of RF to DC is kept constant while RF amplitude and the DC potentials are varied, causing ions which are stable in their trajectories to travel to the detector. To obtain a total ion chromatogram the RF/DC ratio is ramped to systematically detect all the ions within the mass range of the instrument.

The triple quadrupole, which is composed of three quadrupoles arranged in a series, has a mass range of m/z 20 to 3000 and is capable of measuring the exact mass of an analyte. Use of two of the three quadrupole mass analyzers makes it possible to monitor both parent and daughter ions simultaneously. The first and third quadrupoles act as mass filters, and the second quadrupole is the collision cell using Ar, He or N gas (~10^{-3} Torr, ~30 eV) to fragment the parent ions. The second quadrupole works in RF
mode to transmit all of the fragment ions to the last (third) quadrupole to be filtered or scanned fully. Ion guides are normally utilized to help the selected ions maintain their trajectory until they reach the detector.

MS/MS scans provide valuable peptide fragment information in a proteomic experiment and in the quadrupole mass analyzer MS/MS depends on the second quadrupole for collisionally induced dissociation (CID). The resulting fragment ions are scanned in the third quadrupole. MRM experiments can be performed by adjusting the RF/DC frequencies to allow only a selected ion to pass through the first quadrupole for fragmentation.

**ii. TOF and QTOF Mass Spectrometers**

Time of flight (TOF) analyzers form ions in the source and accelerate them through a fixed potential (typically 1-20 kV) into a field-free region, and then are allowed to drift to the detector (as shown in Figure 1.3) this means that ions are separated according to their velocity and will reach the detector at different times. After passing the potential ions with the same charge will have the same kinetic energy and are not affected by any field. The equations which describe the travel of the ions in TOF analyzers are listed below:

\[
K.E = qV \quad \text{Equation 1.1}
\]
\[
qV = \frac{1}{2} mv^2 \quad \text{Equation 1.2}
\]
\[
V = \left(\frac{2qV}{m}\right)^{1/2} \quad \text{Equation 1.3}
\]
\[
T = \frac{L}{\left(\frac{2V}{m/q}\right)^{1/2}} \quad \text{Equation 1.4}
\]
KE is kinetic energy, q is the charge of the ion, V is the applied voltage, m is the mass of the ion, v is the velocity of the ion, t is the time for the ion to travel the length of the flight tube and L is the length of the flight tube. The reflectron TOF employs an electrostatic field at the end of the drift tube into which ions of higher energy will penetrate further than those of lower energy before being reflected into another drift tube. This step focuses ions of the same m/z ratio which previously had slightly different velocities and increases the resolution of the TOF instrument \(^{46}\). Time-of-flight analyzers possess high mass accuracies of 5-50ppm and large dynamic ranges up of \(10^5\) Daltons or more. A time of flight analyzer has a scan speed of about a millisecond. The TOF mass analyzer is easily hyphenated with quadrupole mass analyzers to yield a quadrupole – time of flight MS (QToF) as described in the next section.

The QToF mass spectrometer consists of two quadrupole mass analyzers placed before a time of flight mass analyzer versus the three quadrupoles of the triple quadrupole instrument. Quadrupole Time of flight instruments possess a wide mass range of m/z 20-
20,000, mass accuracies of less than 1ppm and resolution of up to 40,000. Its operation is similar to that of a triple quadrupole instrument in that the first and last analyzer are used as mass filters, while the second is used for fragmentation by CID with nitrogen or argon. Ions produced in the second quadrupole are accelerated orthogonally into a time-of-flight analyzer.

For a scan of all of the ions injected without any fragmentation the first quadrupole operates in radio frequency mode only, and the TOF analyzer separates the ions into a highly resolved and accurate mass spectrum. To obtain MS/MS spectra the first quadrupole is operated in its standard ion filter mode and the second quadrupole is operated as a collision cell as discussed above. The fragment ions produced in the second quadrupole are accelerated using a pulsed electric field which pushes them orthogonally into the field-free drift region of the TOF analyzer.

**iii. Ion Trap Mass Spectrometer**

Although the initial design of the ion trap was introduced in the mid-1980, exploitation of electric fields to trap ions was originally studied by Wolfgang Paul, et al., in the early 1950’s. The quadrupole ion trap which has emerged a popular mass spectrometer for proteomics is a versatile and relatively low-priced mass spectrometer which is useful for the analysis of biomolecules. An ion trap with its two end-cap electrodes and one ring electrode is capable of generating a three-dimensional electric field using alternating current (AC) and DC fields (See Figure 1.4). Ion trap mass spectrometers are often paired with electrospray ionization sources which produce multiply charged ions which are amenable to fragmentation and due to the lowered m/z ratio effectively extend the mass range of the instrument.
Ions having a stable orbit in the trap’s electric field continually oscillate along a three-dimensional trajectory in concentric orbital-like layers. Each orbit contains a specific m/z value and the highest m/z ions are found in the outermost orbits. Changes in the electric field cause the ions to lose stability and to move toward the end-cap electrode and through an opening into the detector component of the mass spectrometer.

Linear ion trap mass spectrometers, such as the LTQ, are capable of high-speed MS and MS/MS cycle times, making them ideal for online proteomic applications. In addition to speed the LTQ offers data-dependent scanning, neutral loss monitoring and parent ion maps. Resolution in both MS and tandem MS is typically R>4000 @ 1000 m/z, allowing the determination of the charge state prior to database searching. A 200 to 4000 m/z mass range in normal operation allows for analysis of large peptides. A typical mass accuracy of 150-500 ppm is a trade-off the fast cycle time.

Ion storage capacity of the ion trap results in high sensitivity because of the increase in signal intensity generated by accumulating ions in the trap. Linear ion trap mass spectrometers can be used to select ions based on their m/z or to trap ions for further analysis. It takes several milliseconds to complete the process of performing a precursor scan followed by several focused scans where a peptide ion is selected for fragmentation. In consequence this limits how many peptides can be selected for fragmentation during a LC-MS experiment \(^47\).
iv. The FT-MS and Orbitrap

In the 1950’s the underlying technology of the Fourier Transform mass spectrometer, the ion cyclon resonance cell, was first described. At that time it displayed its usefulness in the measurement of very small mass differences of ions. Within a superconducting magnet the fixed magnetic field the angular velocity \( \omega_c \) of an ion can be used to determine its mass-to-charge ratio through a conversion to the frequency of the ion’s orbit within the cell as shown in the following equation (Equation 1.7). Ions having the same mass-to-charge ratio and in the same orbit will circumnavigate the cell at the same angular velocity which results in the same frequency (Figure 1.5, part 1). Radio frequency is used to push ions of interest out of their orbits to an orbit having a larger radius for separation and eventually detection of the ions as shown in Figure 1.5, part 2 and part 3.
Detection in the ICR mass spectrometer is non-destructive and depends on the
detection of changes in the electric charges of one of two electrodes located on the outer
walls of the trap caused by the passage of ions. The strength of the signal is dependent
on the induced current and increases with the increase of the post-excitation radius and
ion charge. Ions produced by an electrospray source will be multiply charged with the
result that ions of different charges will produce different changes in the electric charge.
Ions generated in a source (such as an electrospray source) pass through several vacuum pumping stages before introduction into the FT-MS. The ions can be directed into a linear ion trap mass spectral instrument, after each FT MS scan, so as to fragment some ions and provide MS/MS data. In this system the ion trap is located within a superconducting high field magnet. The temperature is close to absolute zero and the vacuum reaches $10^{-10}$ to $10^{-11}$ mBar. Ions are prevented from leaving the ion trap by the trapping plates placed at each end. The resolution of the FTMS can reach to 100,000 with proper tuning and mass accuracies are on the order of 1-2 ppm. The LTQ FT is useful for high accuracy molecular weight determination in the 50-2000 m/z range in both top-down and bottom-up proteomic experiments. Of the mass spectrometers reviewed above, the Fourier transform ion cyclotron resonance mass spectrometer possesses the highest resolution along with ppm mass accuracy.

The LTQ is commonly positioned before the FT mass spectrometer to provide a method of ion fragmentation and to conduct MS/MS experiments.

(Shown in Figure 1.6). A Cutaway View of the Orbitrap Showing an Ion Orbiting the Inner Electrode.
The Orbitrap mass spectrometer is a newer instrument which utilizes centrifugal forces and an electrical field to trap ions as they cycle around a central electrode. This is in contrast to the magnetic and radio frequency fields utilized in other ion trap mass analyzers. The concept is called a Kingdon field, and the concept has been in existence since the 1950’s despite only recently being utilized for mass spectrometry. In response to these forces ions of different mass-to-charge ratios will acquire different orbits which oscillate round the central spindle in orbits that move up and down the spindle. The frequency of the ions is utilized to determine the m/z ratio in a manner similar to that of the FT-MS excepting that the FT-MS utilizes magnetic and radio frequencies. Injection into the trap is achieved through the use of a C shaped trap which uses radio frequency to focus ions and then uses direct current to eject them tangentially into the Orbitrap where they become trapped in a ring-shaped orbit.

The method of ion detection in the Orbitrap may be similar to the non-destructive ion detection method of the FT-MS and depend on the detection of changes in the electric charges of two electrodes or the ions may be allowed to strike an electron multiplier. A Fourier Transformation is used to deconvolute the complex signal resulting from the simultaneous detection of multiple ions. This instrument has a mass accuracy of 1-2 ppm, a resolving power of 200,000 and a dynamic range of 5000. As with the FT-MS a LTQ is commonly positioned before the Orbitrap to perform fragmentation of ions and MS/MS analyses.
C. Variations of LC-MS Proteomic Experiments

i. Analysis of Intact Proteins (Top down Proteomics)

There is an interest in the proteomics field to study post translational modifications of proteins. To accomplish this, “top down” proteomics, where proteins are analyzed intact rather than being enzymatically digested into peptides, i.e., bottom up, was developed\(^{51}\).

Characterization of intact proteins including post translational modifications is achieved through the fragmentation of multiply-charged ions in a mass spectrometer capable of performing MS/MS such as a triple quadropole mass spectrometer. The complex spectra generated by multiply charged proteins limits whole protein analysis to samples of low complexity samples and requires instrumentation capable of very high mass accuracy. An advantage of whole protein analysis over that of peptide analysis is that the information about a protein’s molecular weight including post translational modifications can be lost during enzymatic digestion due to the fact that no actual measurement of molecular weight can be made and that the molecular weight is calculated from a theoretical sequence such as those published on Swiss Prot\(^{52}\).

The analysis of glycoproteins separated from plasma or serum as in the experiments described in this thesis required the ability to detect multiple proteins and the presence of a high degree of glycosylation.

ii. Analysis of Enzymatically Digested Proteins (Bottom-Up Proteomics)

Enzymatically cleaving proteins prior to analysis with LC-MS is currently the most popular proteomic experiment. The method is applicable to almost any tissue or bodily fluid sample \(^{53-55}\). There are issues with the digestion of formalin fixed tissues
which, although they make up a vast repository of historical pathological samples, tend to
give erroneous results when digested with enzymes and analyzed with LC-MS due to
covalently cross-linked proteins. Recently several laboratories have begun reporting
positive results from their efforts to overcome the problems through the use of optimized
enzymatic digests, ultrasound and/or microwave prior to SDS-PAGE or LC/MS analysis.
Haifeng Xu and fellow researchers were able to generate 4098 Swiss-Prot identifications
with a 1% false discovery rate from liver tissue.

Most separation techniques have the potential to be used to simplify the sample
or to select for an interesting subset of proteins in a bottom-up proteomic analysis of
tissue or plasma. Trypsin remains the most used enzyme for the typical bottom-up
experiment. Reasons for this are multiple and include ease of use, availability of pure
enzyme, specificity of cleavage site (at arginine and lysine) as well as the informatics
tools built up to aid in the identification of tryptic peptides such as Bioworks, which
includes the specificity of trypsin digestion in its search parameters. Over the course of
time, interest in other enzymes has developed in hopes of improving the overall
sensitivity of the MS and to improve sequence coverage. One of these, Lys-C, an
endoproteinase cleaves specifically at amide, ester and peptide bonds at the carboxylic
side of the Lysine residue and not arginine resulting in larger fragments than trypsin.

**iii. Two-Dimensional LC-MS**

The typical reversed phase nanoflow separation of a tryptic digestion of plasma or
tissue results in coelution of more peptides than a mass spectrometer can fragment for
identification at a given time. This situation arises because a digest of plasma or serum is
composed of at least tens of thousands of peptides which can not be resolved using a typical one-dimensional reversed phase LC-MS system. Without fragmentation in the mass spectrometer it is not possible to use sequence information to identify the peptide or its protein of origin. The efforts to overcome this problem can be divided into three different areas which comprise the main parts of a proteomic experiments. Prefractionation of sample proteins, chromatographic separation of proteins or peptides and detection all are important components of the experiments. Prefractionation methods include the enrichment of glycosylated proteins such as performed using the M-LAC column. Detection of coeluting peptides can be enhanced by improving the mass spectrometer’s ability to separate and select peptides for fragmentation and processing speeds. Separation methods are constantly being studied to improved resolution and sensitivity. Although it is possible to employ any number of different HPLC columns, it is the C18 reversed phase column that has found the most widespread use. The octadecyl column (C18) is a well-known stationary phase which is typically rugged and possesses high separation efficiencies which is available in a range of particle shapes, sizes and types. The use of smaller, highly homogenous spherical particles benefits separations by resulting in less diffusion of the analyte due to the reduction of path length and variation of path. Particles are available in various sizes starting from below 2 µm, with 5 µm being commonly employed. Less diffusion equals taller and narrower peaks which allows for increases in peak capacity. These factors can allow for better separation of analytes which in turn allows for a larger number of peptides to be analyzed. Data-dependent scanning is used to select the most intense ions from a precursor ion scan for subsequent fragmentation. Thus a more intense ion which is above the noise level of the
mass spectrometer has a better chance of being selected. The operator may set up the software to perform several scans on the parent ions per precursor scan. A higher number of scans will fragment more ions and result in more peptide identifications per parent scan, but will take a longer amount of time, thus increasing the time for the entire analysis and potentially missing other peptides eluting during the time it takes to do the scans. The number of scans is controlled by the amount of ions contained in the ion trap, the software and the user’s need to balance the number of scans with subsequent analyte elutions and time.

In terms of packing material, changes in stationary phase type are often changes from silica to an organic polymer such as polystyrene. Silica, while having been the most popular particle type for decades, suffers from some drawbacks that persuade column manufactures to continue developing new columns. The two most important drawbacks are the susceptibility to hydrolysis at the silanol group by strong bases and strong acids which can cause loss of the C18 moiety, and the resulting unfunctionalized silanol groups will bind charged molecules such as peptides, thus resulting in sample losses and unpredictable mixed bed separations.

**iv. Orthogonal Separation Techniques**

In an attempt to better separate peptides different types of chromatography may be used in sequence. Separations by ion exchange chromatography will separate analytes based on charge and can be performed either online with the mass spectrometer or offline from it. Online experiments have the advantage of being easily automated, and allow the possibility of setting up multiple analyses on an autosampler. Fractions from the ion
exchange column are eluted onto a reversed phase column and alter binding of the peptide fraction to the column, washed to remove the salts used in the IEX separation before initiation of the organic gradient and placing the column in line with the mass spectrometer. One may also use volatile salts such as ammonium acetate to elute molecules bound to the IEX column which remove the necessity of keeping the column offline during the IEX separation step. Off-line experiments can allow for selection of only peaks which are of greatest interest, thus reducing the analysis time needed on a busy LC-MS system. The use of non-mass spectrometry compatible regents and solvents can be performed off line prior to a solvent exchange or clean up step.

The use of 2-dimensional gels in conjunction with in gel digestion and mass spectrometric identification can have issues with the solubility of membrane proteins. RD Smith performed a study of mouse brain tissue using LC/LC-MS/MS, where cysteinyl peptides were captured on thiopropyl Sepharose resins to enable the identification of 48,000 peptides corresponding to 7792 different proteins. In contrast, a study of transgenic versus wild type rat brain tissue performed using 2D gel technology resulted in visualization of 5093 proteins spots. Comparison of the visual results led to the identification of proteins only of those spots showing differences in abundance. Results such as these spur research utilizing LC-MS methods.

v. Multidimensional Protein Experiments

Another development of 2D HPLC separations termed MudPIT (Multidimensional Protein Identification Technology) was utilized in combination with multiple enzymatic digests to study the yeast proteome. The tissue samples were
digested by Endoproteinase Lys-C prior to separation with a 100 µm i.d. capillary column packed with a layer of cation exchange stationary phase and a layer of reversed phase media which was directly interfaced with a mass spectrometer. First the sample is loaded on to the column with aqueous mobile phase followed by a gradient separation step with organic solvent. Next separation based on charge is achieved with salt steps of a 500mM ammonium acetate buffer. The compiled identifications from three samples analyzed in this experiment totaled 5,540 peptide identifications resulting in the discovery of 1,484 proteins from the \textit{S. cerevisiae} proteome.

1.4.4 Sample Preparation Methods for Proteomic Experiments

\textit{i. Depletion}

Proteomic depletion methods are centered on the use of antibodies developed to remove the abundant proteins in serum or plasma, columns such as Genway’s 12 protein depletion columns and Agilent’s MARS column. The highly specific binding of antibodies can be harnessed to selectively remove abundant proteins such as albumin, immunoglobulins and haptoglobin. As might be expected, this specificity is both an advantage and a draw-back. Antibodies are generally species-specific, meaning that they may not select murine proteins if developed against human proteins. A specific antibody may need to be produced for each target protein and each species. Protein A and Protein G are non antibody-based depletion alternatives. These proteins selectively and strongly bind multi-species immunoglobulins. Protein G binds human, mouse and rat IgGs amongst some other species but will not bind IgA, IgM, or IgD. Protein A also binds IgG from of humans, mice and rats but will bind human IgA and IgM weakly as well.
ii. Lectins

Lectins have been known to be selective for blood types since the 1940’s. These plant derived proteins are able to bind to carbohydrates on the surface of red blood cells and result in the agglutination of the cells. Most lectins are plant derived proteins such as Jacalin, wheat germ agglutin and Sambuca nigra. Although the exact purpose for most lectins is not understood, it is known that some of them participate in cell binding and adhesion. Through the use of non-covalent and steric interactions lectins bind carbohydrate structures in a manner similar to that of antibody antigen interactions. Variable in both origin and form, lectins may be composed of homogenous or heterogeneous subunits. The specificity of lectins toward carbohydrate structures can be utilized to study biological processes. Lectins can be specific for a particular sugar such as sialic acid, mannose or a particular sequence and linkage of sugar molecules in a polysaccharide.

1.5 Social/ Medical Importance of Breast Cancer and Diabetes

The frequency of breast cancer in the United States has been increased over the past three decades. It was estimated that 181,600 new cases of breast cancer were diagnosed in the United States in 1997, and that 44,190 people would die of breast cancer during the same year. In recent years, both incidence and mortality have seen decreases. A woman’s chances of acquiring breast cancer increase with age until menopause. Other risk factors are early exposure to ionizing radiation, long-term postmenopausal estrogen-replacement therapy, and alcohol consumption. About 5 to 10 percent of all breast cancers occur in high-risk families, and there are several familial breast cancer
syndromes, including the breast–ovarian cancer syndrome, the Li–Fraumeni syndrome, and Cowden's disease 64-67.

Links between obesity and cancer risk had been hypothesized in the nutritional literature, dating back to an experimental animal model in the 1940s from Tennenbaum 68, 69. A large number of more recent studies have reinforced this proposal using data from large numbers of people.

In the UK, the Million Women Study reported on the cancer incidence in women and looked for connections between disease and lifestyle 70. High body mass index (BMI) is associated with endometrial cancer, adenocarcinoma of the esophagus, kidney cancer, leukemia, multiple myeloma, pancreatic cancer, non-Hodgkin’s lymphoma, ovarian cancer, breast cancer in postmenopausal women and colorectal cancer in pre-menopausal women.

As the obesity epidemic in America grows at an alarming rate the connections between weight and many diseases has become clear. Cardiovascular disease, hypertension and diabetes are all connected to poor diet and exercise as well as genetic history. The Nurses' Health Study, begun in 1976, collected information regarding medical, lifestyle, and other health-related information from 121,700 female nurses 30 to 55 years of age and continues up to the present time 71.

1.6 Proteomic Analysis of Breast Cancer

1.6.1 Relevance of Estrogen Receptor Positive Breast Cancer

In 2002 and 2003 a decrease in the incidence of estrogen-receptor-positive breast cancer in women aged 50 to 69 was reported in the N.E. Journal of Medicine. The
incidence of breast cancer, estrogen receptor status and age was charted over the course of 28 years. Estrogen-receptor-negative tumors did not show the same downward trend. The authors concluded the most probable reason for the decline of ER+ breast cancer was the effect of changes in the use of hormone replacement therapy 72. Despite treatment advancements breast cancer is associated with high morbidity because the frequency of occurrence has not been reduced. In the U.S., during 2006, 21,920 women were diagnosed with breast cancer (Koman Foundation, 2006). There are high costs in terms of initial treatment, treatment of recurrent cancer, and treatment side effects. Use of a theoretical model of breast cancer can lead to a better understanding of how breast cancer develops and explain the influence of hormones on breast cancer.

Early diagnosis of breast cancer improves outcomes. If early markers of breast cancer could be found in plasma or serum, pre-disease testing could be made less invasive and more frequent. Other driving factors for biomarker discovery in plasma include monitoring disease progress and effect of treatment. Blood is in contact with all organs and thus may carry any protein shed from cells during signaling, growth or cellular death.

1.6.2 Effect of Estrogen and Tamoxifen on Estrogen Receptor

Estrogen is known to promote the growth of the ductal epithelium in the breast, and also causes an increase in the expression of progesterone receptor. Both estrogen and progesterone receptors are ligand activated transcription factors that directly regulate gene expression by shuttling from the cytoplasm to the nucleus with ligand binding 73. There are two separate known genes that code for estrogen receptor proteins. Each gene
produces a distinct protein having a different amino acid sequence. One form of the protein is termed ERα, the other is ERβ. The alpha form is generally found in the luminal epithelium of normal breast tissue but not in the connective tissue surrounding the tumor (the stromal tissue). The beta form can be found in the myoepithelial tissue, luminal epithelial tissue, stroma as well as lymphocytes. Expression of ER alpha changes over the course of the menstrual cycle, and it is highest at the follicular phase. As estrogen levels increase, the level of estrogen receptor alpha in breast tissue decline. Despite this, the level of mRNA stays constant which indicates a post transcriptional reason for the decline, such as increased degradation. The beta form of the estrogen receptor has been observed to increase with increasing estrogen levels. In tissues other than the breast an increase in estrogen causes a decrease in both the ER alpha and progesterone receptor. In human epithelial cells the ratio of ER alpha vs. ER beta is 1:2.74

Breast cancer can be subdivided into two groups, those which are estrogen receptor positive and those which are estrogen receptor negative. Those which are estrogen receptor positive are generally more receptive to treatment especially in the early stages. There are several drugs which inhibit the function of the estrogen receptor. Tamoxifen and fulvestran bind to and inhibit the ER receptor. Aromatase inhibitors such as anastrazole, letrozole, exemestane, which block steroid biosynthesis, are also effective treatments for ER receptor positive cancers.

Tamoxifen irreversibly binds to the estrogen receptor, thus blocking it and inhibiting tumor growth, and may be used to treat all stages of hormone-dependent breast cancer. An important metabolite of Tamoxifen, 4-hydroxytamoxifen (4-OHtam), is actually 100-fold more effective in antagonizing the estrogen receptor.75
The American Society of Clinical Oncology recommends the monitoring of estrogen receptor and progesterone receptor for every case of metastatic breast cancer in both pre and post menopausal women in order to determine suitability of the patient for hormonal therapy. For a breast cancer patient, the discovery of elevated estrogen growth factor receptor levels in biopsy tissue is relatively positive news for the reason that ER+ breast cancer has a history of responding better to treatment 76.

HER-2 is another marker that is measured in all cases of primary invasive breast cancer as deregulation of HER-2 is a marker of poor prognosis. HER -2 is useful in selecting patients who may benefit from anthracycline adjuvant therapy. In breast cancer, where epidermal growth factor and HER2 is regulated by a complex system of hormones, estrogen has been shown to inhibit the expression of HER2 77.

1.6.3 Importance of EGFR (Epidermal Growth Factor)

EGFR is over expressed in some types of breast cancers and interaction with epidermal leads the receptor to undergo dimerization which initiates tyrosine kinase activity, leading to DNA synthesis and cell proliferation. Disruption of the normal activity of EGFR can lead to disregulation of cell division and thus tumor growth. The connection of EGFR to breast cancer is well established and is targeted by several chemotherapies. For example, Gefitinib activity is based on its ability to inhibit the tyrosine kinase activity of epidermal growth factor. Pertuxzumab is a humanized monoclonal antibody created by Genentech to inhibit receptor dimerization 78.

The use of the nude mouse in breast cancer studies eliminates genetic and environmental variations which might result from heterozygous animals and humans. In
addition, removal of the ovaries from the mice produced a murine equivalent to a post menopausal woman but necessitated the supplementation of estrogen to promote maturation of the mammary pad and tumor implantation. In our study half of the ovarectomized mice received an injection of MCF-7 cells and an initial estrogen pellet implant to establish the tumor. The orthotopic placement of the MCF-7 tumor cells in the mammary pad of the nude mice used in this study is considered to be a model of human metastatic disease. A suppressed immune system allows implantation of human tumors which would normally cause an immune response against the implant and alleviates the need to deplete plasma of immunoglobulins.

We performed an analysis of the plasma glycoproteome of the mouse model using multi-lectin affinity chromatography. Fractionation of the pooled mouse plasma was achieved with a mixture of three agarose bound lectins; Concanavalin A (binds α-mannose), Jacalin (binds O linked N-acetylglucosamine), and Wheat germ agglutinin (binds sialic acid). Use of multi lectin affinity chromatography (M-LAC) fractionates the mouse plasma into two different fractions of glycosylated proteins depending on the glycan motifs. Depletion of albumin is achieved for the M-LAC bound fraction due to the fact that albumin is not glycosylated. Both the tumor implanted and the tumor free animals were separated into four groups as illustrated in Figure 1.7, which were treated with estrogen, Tamoxifen, estrogen and Tamoxifen or neither. Tumor growth was measured, and plasma was collected at the beginning, 3-weeks and 6-weeks.
1.7 Proteomics of Serum from Gastric Bypass Patients Who have Developed Hyperinsulinemia

1.7.1 Introduction to Hyperinsulinoma Condition of Gastric Bypass Patients

As the obesity epidemic expands, an increasing number of people seek out surgical means of combating severe weight gain. Roux-en-Y gastric bypass surgery employs two principal means of disrupting the absorption of calories from food. The surgeon reduces the size of the stomach and attaches it to the small intestine via a Y shape. The unnatural shape creates bypass route for food going past part of the small intestine. This causes fewer calories to be absorbed in the small intestine.

A rare side effect of the surgery is that patients develop a post operative hypoglycemia condition 1-2 years after gastric by-pass surgery. Analysis for insulin and
C-peptide revealed elevated amounts of these indicator proteins. Dietary modifications such as frequent small meals failed to alleviate symptoms. Pancreatic tissue removed during pancreatectomy revealed islet hyperplasia and increased β-cell volume although insulinoma was not present. The study designed by the Joslin Diabetes Center included four groups: (1) Asymptomatic- normal recovery after gastric bypass; (2) Obese Control; (3) Symptomatic- develop metabolic complications (hyperinsulinemic hypoglycemia) after gastric bypass, and (4) Normal control Group.

1. 7.2 Depletion Methods Used in the Study

To reduce the dynamic range of the serum samples which would be tested we explored two methods of depletion. The first is the Genway 12 protein depletion column which uses antibodies coupled to a support to deplete the 12 most abundant proteins in plasma/serum. This column has the ability to remove (Figure 1.8) albumin, IgG, IgA, IgM, transferrin, fibrinogen, apolipoprotein A-I, apolipoprotein A-II, haptoglobin, α1-anti-trypsin, orosomucoid (α1 acid glycoprotein) and α2 macroglobulin. Although the 12 protein depletion column offers the promise of enabling the discovery of low level and specific biomarkers, it has numerous issues, like many new technologies. These issues include loss of antibodies over the course of a long study due to cleavage, denaturation of the antibodies, irreversible binding of sample proteins to the column, and the fact that an antibody may not bind all isoforms of the target proteins. Nonspecific binding of other proteins may also contribute to poor reproducibility, although it is ideally minimal with correct use. Loss of binding capacity due to antibody loss may, however, cause an increase in nonspecific binding by exposing the surface of the polymer beads used as a support for the antibodies. Added to the concerns of relatively short column life is the
high cost of procuring a replacement as columns currently sell for more than 15,000 dollars. It is also important to consider the possibility of loss of low-level proteins which are bound to any of the 12 depleted proteins via protein-protein interactions.

Figure 1.8 Protein Depletion Column used to Extend Dynamic Range. An antibody based column purchased from Beckman Coulter Inc. has been developed to remove the top 12 most abundant proteins.
Another option for plasma proteomics involves depletion of only the albumin and immunoglobulins, which can still greatly reduce the dynamic range of the sample for minimal cost. An antibody based anti-albumin column coupled to a Protein G column was used to perform a second analysis on the sample set from the gastric bypass study.

Both experiments included a glycoprotein capture step using the M-LAC column. Trypsin digestion was carried out on both the glycoprotein fraction and the non-glycoprotein fraction in the case of the 12 protein deletion experiment.

1.7.3 Peptidomic Analysis of the Plasma of Gastric Bypass Patients

A third approach was utilized to investigate the possibility of a protein/peptide hormone cause for the syndrome. While the proteomic methods focus on the capture of potential protein biomarkers, they are not capable of identifying endogenous peptide hormones. There is the potential that significant changes identified in the peptidomic content of plasma may aid in understanding the biology of the syndrome and how to treat it. By focusing on only those peptides having a molecular weight less than 10,000 Da, we are examining that portion of the serum containing hormones, cytokines and growth factors. Intestinal hormones such as incretin affect insulin response to glucose. Glucagon-like peptide -1(GLP-1) is a specific example of an important incretin hormone involved in the secretion of insulin\(^79\). Previous experiments performed in our laboratory optimized the analysis of the serum peptidome. By using a method consisting of ultra-filtration and no digestion we were able to see peptide ladders in serum and examine the effects of disease associated proteases\(^80\). Our lab utilizes a Fourier transform-linear ion
trap hybrid mass spectrometer for analysis of the peptidome. This highly accurate instrument can generate accurate peptide identifications with molecular weights with less than a 2 ppm difference versus the theoretical molecular weight.

1.7.4 Investigation of an Autoimmune Cause of the Complications

In diabetes it has been theorized that the body’s own immune system may play an important role in the onset or progression of the disease. Antibodies which recognize and attack one’s own proteins have been noted in several diseases including rheumatoid arthritis, psoriasis and certain classes of diabetes. There are several methods of screening for autoantibodies. From a patient one needs to isolate a fraction that contains the antigen (blood protein) and the screen for binding with an immunoglobulin present in the IgG fraction. Variations of the experimental design generally focus on how the antigen containing portion is immobilized. In most measurements it is both beneficial and necessary to fractionate the depleted plasma fraction. For example, HPLC methods may be used for this step, and the resultant fractions collected in 96 well plates which are suitable for an ELISA after solvent evaporation. There are many reports in the literature of spotting proteins onto derivatized glass plates for use in a microarray format.

In many studies the most popular first line of investigation includes gel-based assays which allow for subsequent analysis by Western blotting. Both 1-dimensional and 2-dimensional formats are applicable. While a 2D separation has greater separation power to facilitate separation into individual proteins and isoforms, it may result in such low concentrations per spot as to be undetectable after Western blotting. Therefore, in
order to perform an initial screen to prove whether there were indeed auto-antibodies in
the serum of patients who developed complications after gastric bypass we chose to use a
1D SDS-PAGE separation.

Prior to the gel separation, the IgG proteins were isolated from the sample using a
column containing immobilized Protein G. The IgG proteins were reserved for later use
and the other protein fraction was separated on with SDS-PAGE and then
electrophoretically transferred to nitro-cellulose. In the next step the nitrocellulose
membrane was incubated with the immunoglobulin enriched fraction, after which a goat
anti-human IgG antibody conjugated to HRP was added to enable detection. If an auto-
antigenic protein is detected using this method, a complementary method must then be
employed for identification of the protein such as an in-gel digestion and LC/MS
analysis. For example, replicates of the samples can be run on the same gel which allows
some lanes to be used for Western blotting and others for staining or in-gel digestion.

1.8 Bioinformatics

1.8.1 Growth of Informatic and Bioinformatic Software for Proteomics

Progress in the field of proteomics will benefit substantially from the
development of informatics programs which enable faster processing of complex data
sets as well as improved visualization. There are many software and web-based tools
which have potential use during the interpretation of proteomic data\textsuperscript{83,84}. These include
PeptideAtlas\textsuperscript{85}, Open Proteomics Database (OPD)\textsuperscript{86}, Global Proteome Machine (GPM)
\textsuperscript{87}, Pedro\textsuperscript{88}, PepSeeker\textsuperscript{89}, and the Proteomics IDEntifications database (PRIDE)\textsuperscript{90} as
well as others. PeptideAtlas, GPM and PRIDE contain extensive databases of millions of peptide identifications for use with mass spectrometer data.

1.8.2 Quantitation of Data via Software Programs

Figure 1.9 Process of Peptide Identification. A peak in the extracted ion chromatogram (a) has been identified as corresponding to the vitronectin peptide LIQDVWGIEGPIDAAFT based on the parent ion molecular detected in the precursor scan weight (b) and fragmentation pattern (c).

Figure 1.9 outlines the steps in determining the protein of origin of a chromatographic peak. To be certain of the identity it is necessary to have MS/MS data as shown in the bottom panel, and MS/MS data can then be matched to the parent ion as is shown in the second panel of the figure. Once a peak has been confidently identified it is
possible to determine the peak area using the Excalibur software and perform quantitative analyses.

The development of software tools to aid in the analysis of changes in protein abundance is another area of importance to proteomics. While it would be ideal to use rigorous peak area analysis for this step it is currently standard practice to use a value based on the number of peptides identified in an LC/MS analysis of a tryptic digest from a protein to estimate abundance. This determination which is known as spectral count does not take into account that a larger protein will produce more peptides than a smaller protein. A recent improvement, mPAI uses a calculation based on the log of the number of peptides identified divided by the number of peptides which could be generated from that particular protein, but was not used in the study reported in this thesis.91

The use of labeling methods such as SILAC (Stable isotope labeling with amino acids in cell culture) and ICAT can improve the quantification of protein amounts by incorporating isotopic labels into biosynthetic proteins.92, 93, 94 Labeling proteins with the SILAC method relies on metabolic incorporation of amino acids with incorporated stable isotopic nuclei. In a typical experiment, two cell populations are grown in identical culture media, except that one of them contains a heavy form of a particular amino acid. While the benefits of labeling are clear in cell culture studies SILAC experiments are not feasible in human studies.

ICAT (Isotope coded affinity tags) reagents are designed for the determination of relative concentrations of cysteine-containing tryptic peptides obtained from digests of a control and an experimental sample. A stable isotope alkyl linker in the ICAT reagent includes a (cleavable) biotin group which allows rapid affinity isolation of labeled
peptides. Newer 8-plex reagents contain tags of eight different molecular weights to enable pooling of multiple samples. This allows for a single analysis to take the place of up to eight thus saving a significant amount of time. Due to the fact that ICAT reagents are restrictively expensive, especially for experiments with a large number of cohorts such as the clinical proteomic experiments described here it was not used. ICAT reagent may also result in side reactions, especially for low level proteins which may inhibit analysis. Non-labeling techniques offer lower costs, higher speed and fewer issues regarding introduced variations such as labeling inconsistencies between proteins.

1.8.3 Software Designs to Aid in the Interpretation of Protein Identifications

Advances in instrumentation and newly developed sample preparation methods are delivering enormous amounts of data per experiment. As the proteomics field moves into larger, more detailed sample sets the problem of utilizing all of the data will only increase. Thus it is of increasing importance to collect and analyze data in a consistent format which also facilitates data sharing amongst researchers who may not be familiar with mass spectrometry.

There are numerous data processing tools 95, 96 available for interpretation of proteomic data. Proxeon’s Protein Center, which we used in this section, brings together peptide identifications, protein identification, gene-ontology, protein information and mPAI quantitation. This information can be used to evaluate what proteins have been identified and what peptides were used to make the identification. Clear, color coding allows simultaneous run analysis showing in which sample the protein was found. Clickable links bring up a protein information page containing the protein sequence along
Differential quantitation (mPai) calculations allow an estimate of abundance for each protein based on a calculation of the number of peptides identified versus the number of peptides generated from a theoretical digest of the protein. This preliminary scan is usually followed up with normalization to an internal standard and peak area analysis for 2 or more peptides. Such calculations can be used as an initial method to find those proteins showing significant changes, and therefore the most likely to be associated with a disease state.

1.8.4 Graphic Comparisons of Mouse and Human Proteomic Data Utilized to Discover Disease Patterns

Not all of the peptides present in the precursor ion scan in an LC-MS analysis are selected for fragmentation and successfully produce fragments suitable for identification. For each precursor scan an LTQ, for example, can select only a limited number of ions (typically seven) for fragmentation and thus identification. Due to post translational modifications (PTMs), such as glycosylation, some of the peptides selected will not produce useable fragmentation data. To take full advantage of LC-MS data for disease detection, researchers have investigated the possibility of utilizing raw LC/MS output and forgoing MS/MS based protein identifications by analyzing peaks in either the total ion chromatogram or in individual m/z chromatograms for disease connections. This process has the potential to detect changes due to disease that may be missed when comparing individual proteins. The proteins which are producing diagnostic peaks in the LC/MS output...
chromatograms are not identified in this analysis method, and more focused experimentation utilized may be to identify the source of the diagnostic peaks.

This investigation utilized the data produced in the mouse xenograft proteomic study and data produced by Madduma Dayarathna proteomic study of human plasma from diabetic, diabetic hypertensive, obese and normal patients. Extracted ion chromatograms (EIC) from LCMS data gathered during the mouse breast cancer study were examined for changes in peak patterns corresponding to tumor growth. The second data set, of the human plasma, was investigated in the same manner for patterns which could be associated with diabetes. While these investigations were successful it is necessary to examine numerous data sets to validate the use of EICs for disease diagnosis, thus, a program was written to automate the comparisons by Dr. Akella as a continuation of this work.
1.9 References


73. Allan GF, Tsai SY, Tsai MJ, O'Malley BW. Ligand-dependent conformational changes in the progesterone receptor are necessary for events that follow DNA binding. Proc Natl Acad Sci U S A. 1992 Dec 15;89(24):11750-4.


81. Burnette WN. "Western blotting": electrophoretic transfer of proteins from sodium dodecyl sulfate-polyacrylamide gels to unmodified nitrocellulose and radiographic detection with antibody and radioiodinated protein A. Anal Biochem. 1981 Apr;112(2):195-203


A Proteomic Analysis of the Plasma Glycoproteome of a MCF-7 Mouse Xenograft: A Model System for the Detection of Tumor Markers

Acknowledgments

I would like to thank Lihua Yu, Dr. Hincapie, Dr. Rejtar and Dr. Wu for technical expertise. All of the mouse experimentation was performed at AstraZeneca under the guidance of Dr. Maureen Hattersley, for which I would like to thank her. I would like to express gratitude to Dr. Jeff Hanke and Dr. Hancock for essential help interpreting the data. I thank AstraZeneca, Inc., for the financial support of this research. I also thank Thermo Electron and GE Healthcare for support with instrumentation and software.

2.1 Benefits of Model System

There are both advantages and disadvantages to studying a disease using an animal model. Animal models are not capable of exactly replicating human disease for numerous reasons, including differences in genetic makeup, lifestyle differences and life span. Before any nonhuman proteomic results can be accepted as being representative of humans it is then necessary to study human subjects to verify the results found in the animal model. In order to replicate a human disease in an animal it is necessary to first initiate the disease of interest in the animal. There are two common ways in which this is accomplished in cancer studies. One is to cause the disease in otherwise healthy animals by such methods as implanting a tumor, injections of tumor causing agents, or controlled exposure to a disease causing substance. The other involves the breeding of animals which are genetically engineered to spontaneously develop the disease being studied.

Implantation of tumors into an animal, such as performed in the mouse xenograph study, allows for the study of human tumor cell lines. This can be advantageous in many respects toward the study of effects of anti-tumor agents which is more realistic in an animal than in a tumor grown in a culture dish. Interactions with the surrounding stroma and associated signaling molecules, which cannot be replicated in vitro, may have important effects on a tumor’s growth and response to environmental factors such as diet or anti-tumor treatment. Animal models, whether genetically-engineered models (GEM) or orthotopically implanted models, are able to be controlled very stringently in terms of genetic background, age, nutrition and environment, resulting in lower levels of variation compared to human subjects.
Using data generated from xenograph models presents the issue of whether human tumor signaling molecules interact with the animal host’s receptor proteins. This has driven ongoing research utilizing animals genetically designed to spontaneously develop tumors or other diseases. Genetically engineered models (GEMs) use genetic promoters that can control the cell type which initiates the disease, and decrease cellular heterogeneity. Despite the potential advantages no GEM has yet to be accepted as a true model of human cancer. In a proteomic experiment there is the additional necessity to have the genomes of the organisms available for database searching met. GEMs do not meet these criteria.

2.1.1 Cancer and Diabetes

The plasma proteome has been a major focus of the Human Proteome Initiative because of the pivotal connection of this proteome to the disease-based initiatives of the organization. In addition, clinical chemistry has made this study a priority in terms of better understanding disease and the search for biomarkers for the early detection of diseases such as cancer. Quantitative proteomics can also be used to characterize the human secretome, and in particular the altered secretion of proteins in disease with the goal of discovering cancer biomarkers and to characterize the tumor secretome. We and others have chosen as a focus the study of the plasma glycoproteome because of the intimate association of this post-translational modification with the secretion process. Altered glycosylation has often been associated with development, progression and metastasis of cancer, and thus the study of the glycoproteome is of undoubted significance to the cancer field. Proteomics
technology has also been applied to plasma analysis in the study of genetically modified animal models\textsuperscript{18} as well as xenografts\textsuperscript{19} but not to the glycoproteome.

For cancer, one anticipates the release of proteins associated with the development of the tumor into the bloodstream, and that such proteins can give a signature for early diagnosis. The release of tumor-specific proteins can be related to processes such as the increased synthesis and secretion of glycoproteins, the cleavage of matrix or membrane associated proteins and the release of intracellular proteins from cells that have undergone apoptosis\textsuperscript{4}. The proteomic analysis of the interstitial fluid from breast tumors has suggested that cell death and the release of cytoplasmic or nuclear associated proteins makes a contribution, albeit small, to the blood proteome of cancer patients\textsuperscript{20}. Recently, we reported on significant changes in the plasma glycoproteome in breast cancer\textsuperscript{15} and autoimmune disease\textsuperscript{21}. However, plasma consists of an extremely complex mixture of proteins that are released from a variety of tissues including leukocytes and thrombocytes, and thus it is a challenge to understand the relevance and source a given protein. In order to begin to tease apart tumor-specific proteins from other plasma proteins, we focused on studying the glycoproteome in an MCF-7 nude-mouse xenograft tumor model to see if we could identify proteins related to tumor biology and progression of the disease. The use of closely controlled animal models greatly reduces the complexity and variation of samples relative to studies in humans\textsuperscript{22-23}. Implanted human tumors also have several advantages over spontaneous tumors as they have distinct origins and genetic backgrounds. For example, the MCF-7 cell line used in this study has an extensive history\textsuperscript{24-33} and was derived from a pleural effusion of a patient with breast cancer\textsuperscript{22}. The cell line is estrogen receptor positive although such
estrogen dependent lines rarely metastasize in nude mice\textsuperscript{33}. It is considered to be
uninvasive in both the Matrigel outgrowth study and in the nude mouse\textsuperscript{24}. Several studies
have attempted to identify proteins likely to be associated with the development of
proliferating MCF-7 cell lines.\textsuperscript{28,32}, such as with mitogenic concentrations of 17-beta-
estradiol (E2)\textsuperscript{29}, or MCF-7 cell lines selected for resistance to anticancer drugs
(mitoxantrone)\textsuperscript{24}, and following treatment with doxorubicin\textsuperscript{31} or adriamycin\textsuperscript{28}. Recently
LC/MS proteomic studies were performed on nuclear fractions isolated from cultured
MCF-7 cells and identified 3715 putative proteins, while a related study of MCF-7
plasma membrane proteins characterized 540 proteins\textsuperscript{30,32} and thus set the stage for
xenograft studies. The study, however, that is closest to our report was reported by, Juan
et al.,\textsuperscript{19} in which the authors studied a xenotransplantation model in the nude mouse of 5
different human cancer cell lines. While this report did not include the MCF-7 cell line,
and only identified a few mouse acute phase proteins and no human proteins, the authors
did speculate on the advantage of using a well controlled inbred animal model to identify
tumor derived serum proteins that were of human origin. We used a more sensitive
proteomics approach based on LC/MS and sample prefractionation focused on the
glycoproteome to see if we could indeed identify tumor-specific proteins based on the
murine host response to tumor growth as well as unique human sequences.
Figure 2.1 Experimental Design for the Study of the Ncr Nude Mouse Xenograph as a Model Of Human Breast Cancer. (a)Initially 56 ovarectimized mice were split into two groups: non tumor and tumor. A MCF-7 tumor was implanted subcutaneously in the mammary fat pad of those mice belonging to the tumor group while the control group was not implanted. Each of the two groups was further split into 4 treatment groups: no treatment, estrogen, tamoxifen or both estrogen and tamoxifen. Estrogen and/or tamoxifen pellets were surgically implanted in mice of the estrogen groups and the estrogen plus tamoxifen groups. For the individual time points plasma from 5 individual mice was pooled for each analysis. Plasma was collected at the beginning of the study, 3 weeks into the study and at 6 weeks.

In our study, ovarectomized mice received an injection of MCF-7 cells, which were implanted with an estrogen pellet. The orthotopic placement of the MCF-7 tumor cells in the mouse mammary pad is considered to be a model of human breast cancer. The animals were separated into groups that were treated with estrogen, Tamoxifen or both. We then performed an analysis of the glycoproteome using multi-lectin affinity chromatography (M-LAC). In contrast to the narrow specificity of antibodies, lectins have a general affinity to glycosylation motifs which are similar across species. Another advantage of our approach is that serum albumin, which makes up approximately 50% of plasma proteins is not glycosylated and is depleted. Thus, the M-LAC platform has been
shown to increase the dynamic range of plasma measurement and provide differential quantitative information of a significant number of plasma proteins. We also used bovine fetuin as an internal standard which allowed us to normalize measurements across the sample set using label free quantitation in the mass spectrometer. Nano LC-LTQ-FT analysis was then used to identify tumor specific proteins based on species specific peptide identification. While the sequence homology between human and mouse is high it is possible to detect species specific peptides using the discriminating power of the high mass accuracy measurement.

We have shown in this study that we could indeed detect tumor-specific proteins derived from the murine host as well as human specific proteins from the tumor. In addition, this study supports the continued development of the M-LAC approach in animal models as well as humans to identify markers of tumor presence and response to therapy. Based on the results of this report we have identified future studies such as a xenograft study with a longer duration that will enable insights into the release of tumor markers into the blood stream and thus facilitate the search for biomarkers for the early detection of cancer.

2.1.2 The Nude Mouse

The nude mouse is ideal for use in xenograph experiments for several reasons. Like most experimental lines of mice they are homozygous and do not contribute variability at the DNA level. Important for this study was the fact that the Balbc nude mice have defective T cells. A defective immune system allows for the implantation of foreign tissue which would otherwise be attacked and destroyed. A secondary benefit is
the lack of IgGs in the blood of the nude mouse. In human plasma IgGs make up a significant portion of the proteins and thus hinder the detection of low level proteins. In addition albumin is non-glycosylated and thus not retained by the lectin column. Thus, in this experiment it was unnecessary to deplete immunoglobulins or albumin. The minimal sample handling used in this study saves time, reagent costs and most importantly avoids issues of sample loss due to irreversible, nonspecific binding to the depletion column.

2.2 Experimental

2.2.1 Materials

Concanavalin A (Con A), Jacalin and Wheat germ agglutinin (WGA) agarose bound lectins were purchased from Vector labs (Burlingame, CA). Disposable polypropylene columns were purchased from Pierce (Rockford, IL). Bovine fetuin and all other chemicals were purchased from Sigma-Aldrich (St. Louis, MO). The BCA protein assay reagent kit was from Pierce (Rockford, IL). Sequencing grade, modified trypsin was purchased from Promega (Madison, WI). 5kDa Amicon molecular weight cut off filters were purchased from Millipore (Billerica, MA). Discovery BIO Wide Pore C18 cartridges (C18, 2 cm x4.00 mm, 3 µm particles) were from Supelco (Bellefonte, PA).

2.2.2 Xenograft Study

The experimental design for the xenograft study is shown in Figure 2.1 (page 83). Estrogen receptor-positive MCF7 cells were obtained from the ATCC and maintained in culture in RPMI media supplemented with 10% fetal bovine serum and 1% L-glutamine at 37°C, 5% carbon dioxide. MCF7 xenografts were established by surgically implanting 8x10⁶ cells into the third mammary fat pad of 6-week old female, NCr Nude mice.
Growth of tumors was supplemented with 0.72mg 60-day-release estrogen pellets (Innovative Research of America, Sarasota, FL) which were implanted subcutaneously on the back of the animals 24 h prior to cell implantation.

When tumors reached an average volume of 120mm³ (day 19 after implantation), the mice were randomized by tumor volume (N=10 per group) into four groups: no treatment (estrogen deprivation), estrogen only, tamoxifen citrate only or estrogen and tamoxifen citrate. In all mice, the original estrogen pellet was removed on the day of randomization. The estrogen only mice were implanted with a new estrogen pellet. The tamoxifen citrate only mice were implanted with a 5mg 60-day-release tamoxifen citrate pellet (Innovative Research, Sarasota, FL). The combined estrogen and tamoxifen treated mice received both estrogen and tamoxifen pellets. Tumor growth was assessed using digital calipers and tumor volume and body weight were measured twice a week. After three weeks of treatment, five mice from each group were sacrificed and tumor tissue and blood was collected from each mouse. After six weeks of treatment, the remaining mice from each group were sacrificed and tumor tissue and blood was harvested. Blood was collected into EDTA microtainer tubes (Beckton Dickinson) then spun at 10 000 RPM for 3 min to separate plasma. The plasma from each group of mice was pooled. All tumor and plasma samples were stored at -80°C. Mice were maintained and sacrificed according to Institutional Animal Care and Use guidelines.

2.2.3 Materials and Methods for Depletion of IgG and Fibrinogen

A pre-packed Protein G column was purchased from Pierce. An anti-fibrinogen column was produced in-house from agarose bound anti-fibrinogen (Sigma-Aldrich, St.
Louis, MO) packed into a 5-ml chromatography column (Applied Biosystems, Foster City, CA) and installed onto a HP 1090 HPLC system (Agilent, Santa Clara, CA). Plasma (50 µL) from mice with no tumor or treatment regimen and a sample of normal human was diluted with 200 µL of loading buffer. Loading buffer consisted of 0.1M Tris, 0.5 M sodium chloride at a pH of 7.4. Elution buffer was 0.2 M glycine at a pH of 2.5 and an equilibration buffer was composed of 0.5 M Tris, 1 M sodium chloride at a pH of 8.0. Glycine buffer (pH 2.5) used to elute retained proteins.

2.2.4 Affinity Capture of Glycoproteins with M-LAC

Pooled samples were randomized prior to analysis to avoid bias. Samples were stored at -75°C until analysis. One hundred microliters of sample was diluted to 500 µL with binding buffer (25 mM Tris, 0.15 M sodium chloride, 1 mM calcium chloride, 1 mM magnesium chloride, 0.055 mM sodium azide at pH 7.4). Bovine fetuin (25 µg/mL) was added to each plasma sample prior to lectin fractionation to serve as an internal standard. Multi lectin columns were gravity packed as previously described with a 1:1:1 mixture of the three lectins. A single M-LAC column was used for each individual sample to avoid possible contamination. The sample was allowed to penetrate into the column for an incubation period of 15 min, and then proteins not bound were washed from the column by the addition of two subsequent 5 mL portions of M-LAC binding buffer, as described above. Washes were collected and subjected to the analysis of protein concentration using a BCA assay. The measurement of protein concentration in the unbound fraction provided an important quality control point. Elution of the glycoprotein fraction was accomplished by the addition of 4 mL of Elution Buffer (25
mM Tris, 0.5 M sodium chloride, 0.2 M methyl-α-mannopyranoside, 0.2 M methyl-α-glucopyranoside, 0.8 M galactose, 0.5 M N-acetyl-glucosamine, 0.05% sodium azide at pH 7.4). A portion of this fraction was also retained for protein concentration analysis.

### 2.2.5 Digestion of Glycoprotein Fraction

The glycoprotein enriched bound fraction was concentrated down to 50 µL using 5kDa Amicon molecular weight cut off filters. Fifty microliters of sample was denatured with 7.2 M guanidine chloride in 0.1 M ammonium bicarbonate (pH 8.0) added to make a concentration of 5.8 M guanidine chloride. The reduction of protein disulfide bonds in the sample was achieved by the addition of 5 mM DTT and incubation at 60°C for 30 min. Protein sulphydryl groups were then alkylated with 15 mM iodoacetamide in darkness for 30 min. The alkylation reaction was quenched by adding a second aliquot of 5 mM of DTT. Samples were diluted to decrease the concentration of guanidine chloride to 1.2 mM with 50 mM ammonium bicarbonate buffer (pH 8.0). Trypsin was added to samples at a 1:40 w/w ratio. Samples were incubated for 18 h at 37 ºC. A second aliquot of trypsin was added at a 1:25 ratio and the samples were incubated at 37°C for an additional 4 h. Digestion was stopped by the addition of formic acid to a final concentration of 1%.

### 2.2.6 Reversed Phase Desalting Using HPLC

Peptides were separated from salts and any undigested material by chromatography on a Discovery BIO C18 column installed on a Shimadzu HPLC (Shimadzu Scientific Instruments, Columbia, MD). Mobile phase A was composed of
0.1% TFA in HPLC grade water. Mobile phase B was composed of 0.1% TFA in HPLC grade acetonitrile. The step gradient method employed for the separation consisted of 3 min long steps of: 0% mobile phase B to wash away salts remaining from the digestion, 30% B to elute peptides to be used for LC/MS analysis and 90% B to wash the column of any undigested proteins or large peptides. The flow rate was set at 1.5 mL/min. Eluting peptides and proteins were monitored 214 and 280 nm. The 30% fraction was collected and concentrated on a speed vacuum to remove AcCN and prepare it for analysis. The organic solvent was removed from each fraction under vacuum and samples were not taken to complete dryness in order to minimize losses. Each fraction was reconstituted to the same volume (20 µL buffer) and then stored at -80°C. For LC/MS analysis a 2.5 µL aliquot was removed.

2.2.7 Proteomic Analysis by Nano LC-MS/MS

All nano LC-MS/MS experiments were performed on an Ettan MDLC system (GE Healthcare, Piscataway, NJ) coupled to a Thermo Finnigan linear ion trap mass spectrometer (Thermo Electron, San Jose, CA). A 15 cm long, 75 µm ID capillary column (purchased from New Objective, Woburn, MA) packed in house with 5 µm, 200 Å pore size Magic C_{18} stationary phase was used for all LC-MS/MS experiments. Mobile phase A and mobile phase B were 0.1% formic acid in HPLC grade water and 0.1% formic acid in HPLC grade acetonitrile respectively. Prior to injection each sample was concentrated to 20 µL. Duplicate injections of 2.5 µL were loaded on to a Peptide Captrap column (Michrom Bioresources, Auburn, CA) using the MDLC autosampler. The trap column was washed with mobile phase A for 10 min at 10 µL/min, and then the
flow rate was reduced to 280 nL/min, the trap column was placed in line with the capillary column and the gradient method was started.

A linear gradient method beginning at 2% Mobile phase B, after 30 min of equilibration, which increased to 40% B over 160 min, to 90% B after an additional 20 min and remained constant at 90% for 20 min comprised the LC separation method. The GE software package Unicorn (GE Healthcare, Piscataway, NJ) allowed for the operation of the Ettan MDLC. The ion transfer tube of the LTQ during the analysis was 245°C and the electrospray ionization voltage was set to 2.0 kV. Normalized collision energy was 35% for MS/MS analysis. MS/MS was triggered automatically by operating in a data dependent mode. The 7 most intense peaks were selected from the full MS scan of 400 to 2000 m/z for MS/MS. Precursor ion exclusion time was 1 min.

2.2.8 NanoLC-LTQ-FT Analysis

An UltiMate NanoHPLC system (LC Packings-Dionex, Marlton, NJ) and LTQ-FT mass spectrometer (Thermo Electron, San Jose, CA) were used for additional nanoLC-MS/MS analyses of samples. The capillary LC column and mobile phases were the same as described above. Electrospray voltage was 1.8 kV. The normalized collision energy was 28% for MS/MS. The ion transfer tube temperature was 245°C. A medium resolution pre-view MS scan was generated after the ions were injected into the ICR cell. The Excalibur software selected the 8 most abundant ions for MS/MS analysis. While the LTQ fragmented these ions, the FT performed a full high resolution MS scan. Precursor ions were excluded from subsequent fragmentation for 1 min.
2.2.9 Data Processing and Analysis

Peptides derived from the internal standard bovine fetuin were identified by searching MS² spectra against a bovine fetuin database downloaded from Swiss-Prot, September, 2005. Murine and human proteins were identified by searching MS² spectra against a murine data downloaded from Swiss-Prot, July, 2005 or a human database which was downloaded from Swiss-Prot, September, 2005. For all database searches trypsin was selected as the enzyme, 2 missed cleavages were allowed. Carbamidomethylation of cysteine was included in the search parameters. Tolerances were set at ± 1.4 Daltons for precursor ion mass and ± 1.0 Daltons for product ion mass. Peptide Prophet Software was used to filter the results to a minimum probability of 95%. Only proteins identified with 2 or more unique peptides were considered.

2.2.10 Relative Quantitation

Preliminary investigation of the complete data set was performed by analyzing the average of the total sequencing events (i.e. spectral count) for each protein. Biologically interesting proteins were selected for manual integration of peak areas. For each protein two high confidence peptides were chosen for peak area quantitation of the extracted ion chromatogram. For the reported peptides the correlation of variation between the replicate runs was 15% or less.

Normalization of the data was achieved by performing the same process of peak area quantitation on 2 peptides having the best signal-to-noise ratio which had been uniquely identified for bovine fetuin. Of the two peptides, the one with the lowest variability was then averaged across all data points (RSD for this peak area was 9.7%).
This average value was utilized for normalization of the data. The overall relative standard deviation was 15% or less between duplicate injections (5% for higher intensity peaks).

2.3 Results

2.3.1 M-LAC Fractionation

We used the multi lectin affinity chromatographic (M-LAC) approach to fractionate mouse plasma into a bound, glycosylated protein rich portion and a flow-through fraction which contained mainly non glycosylated proteins, such as serum albumin. The fractionation was achieved with a mixture of three agarose bound lectins; Concanavalin A (specific for α-mannose type structures), Jacalin (O-linked N-acetylglucosamine), and Wheat germ agglutinin (sialic acid). The combination of multiple lectins has been shown to give more complete capture of the glycoprotein fraction than the use of single lectins. Since the mouse model is immunoglobulin deficient we did not need to deplete the immunoglobulin fraction. In this manner we avoided the use of a depletion step which improves throughput of the study and avoids possible losses of material.
Figure 2.2. Tumor Growth of MCF-7 Xenografts for the 4 Treatment Groups over the Course of 6 Weeks. This indicates that estrogen is required for the continued growth of MCF-7 tumors. The following treatments were used: No treatment, estrogen only, Tamoxifen only or both estrogen and Tamoxifen. The error bars give the range of tumor volumes over the 5 animals in each pool.

In this study, each treatment group contained 5 animals and plasma samples from each group were pooled and the resulting 20 pools, originating from control groups, Tamoxifen and estrogen treatment at different time points, were randomized prior to the M-LAC step (see legend to Fig. 2.1 on page 83 for the experimental design). Bovine fetuin (25 µg) was added as an internal standard to each pool and then loaded onto the M-LAC column. The overall recovery (unbound and bound fraction) of the M-LAC step for all mouse plasma samples was determined by a Bradford protein analysis and averaged
96% with a CV of 10%. The split between the bound and unbound fractions was 9% (CV 28%) vs. 87% (CV of 14%). The variation in amount of the glycosylated fraction is not unexpected in a study which contains both tumor implanted and control animals and different treatment groups, all of which may have effects on the degree of glycosylation of plasma proteins, for example, differences in glycosylation patterns have been noted in certain cancers 8-10.

2.3.2 IgG and Fibrinogen Depletion

In most proteomic studies of complex samples such as plasma depletion of immunoglobulins and fibrinogen is a method of reducing the dynamic range of a sample to enable the detection of lower concentration proteins. In the case of the nude mouse it was discovered that this step is unnecessary and that the lack of an immune response to the implant of the MCF-7 tumor was due to a deficiency of immunoglobulins. This was confirmed by depletion of IgG and fibrinogen from the mouse plasma (Figure 2.3) which revealed the absence of significant amounts immunoglobulin when compared to human plasma. Based on these results, depletion was not performed on the mouse xenograft samples.
Figure 2.3. Results of Immunoglobulin Depletion Experiments. Triplicate injections (single representative runs are shown) of mobile phase A (Tris buffer), human plasma and normal mouse plasma were made separately onto a protein G column installed on a Shimadzu HPLC system. (a) A blank injection of 250 μL mobile phase A on to the protein G column displayed no peak. (b) An injection of mouse plasma (50 μL diluted to 250 μL with mobile phase A) resulted in a diminutive peak. (c) An injection of human plasma (50 μL diluted to 250 μL with mobile phase A) resulted in a large peak.
2.3.3 Preparation of Tryptic Peptide Digests for LC/MS Analysis

At all steps in a proteomic analysis variability can be introduced especially when one considers the complexity and dynamic range of the plasma samples. The trypsin digestion step is no exception, for example, disease related changes in plasma may affect the rate of cleavage of certain proteins. As described in the Experimental section, we used two aliquots of trypsin (ratio of 4% w/w at 0 and 24h) at 37°C to digest the reduced and alkylated glycoprotein fraction (50 µL). In addition, the sample was denatured with guanidine- HCl to minimize incomplete trypsin digestion of associated or aggregated proteins. Despite these precautions our experience is that one cannot expect 100% digestion of all proteins in such a complex sample, and thus we have instituted a reversed phase HPLC (RPLC) cleanup step. Such a step has the advantage of desalting the sample as well as removing partially digested proteins. The latter is particularly important in improving the lifetime of the capillary reversed phase column and the consistency of a series of LC/MS analyses. We also used the UV peak area at 214 nm of the peptide fraction, collected in an isocratic step at 40% organic solvent, to analyze the reproducibility of the digestion and perform any necessary adjustment to the amount of sample analyzed in the LC/MS step. For the entire sample set (20 pools) the coefficient of variation of peak area was 27% and consistent with the expected biological viability of the study. In addition, the RPLC step is relatively fast and the column was reusable, unlike many filtration devices.
2.3.4 LC/MS Analysis of the Sample Set

Since we used the label free approach to give relative quantitation across a sample set, it was important to control bias in the analytical protocol. A key aspect of our approach was to improve column lifetime with our RPLC cleanup step, which also minimizes carryover from sample to sample. It was, however, necessary to perform a blank analysis between each sample analysis. We then monitored variability of LC/MS response over the entire sample set by measuring the integrated base peak from the MS signal between 40 to 160 min (elution window for the majority of peptides) and the amount of variation was again consistent with the degree of biological variation (CV approximately 30%).

In an effort to minimize false positive identifications we used criteria suggested by HUPO for the identification of peptides by MS/MS sequencing, as well as a requirement for a probability greater than 0.9 (Protein Prophet). The final criteria was to only report proteins identified with 2 or more unique peptides, and this gave a total of 1,877 proteins by combining the total data set. In addition, the final time point for the group of 6-week-old mice with implanted tumors receiving either estrogen or Tamoxifen were also analyzed on the hybrid mass spectrometer, linear ion trap - FT MS. The additional analysis by the FT MS system allowed the use of accurate mass as well as retention time information to confirm specific human and mouse protein identifications, such as macrophage colony-stimulating factor 1 receptor, insulin receptor, DNA-dependent protein kinase and vitronectin (data not shown).

Table 2.1 compares the glycoproteome of the control group of the nude mouse with that of the corresponding human glycoproteome. It should be noted that the control
group consists of animals which have been ovarectomized, implanted with the MCF-7 cell line, but then deprived of the estrogen implant. In this situation the growth of the xenograft is not sustained, and thus this group acts as a control. The data shows a high degree of commonality at the level of abundant plasma glycoproteins between the mouse and the human samples. In addition, apolipoprotein CIII and carboxylesterase N could be presumed to be present in the human sample at concentrations below the detection limit for this study. One protein, murinoglobin is species specific, while two proteins, leukemia inhibitory factor receptor (LIFR) and serine protease inhibitor A3K appears to be elevated in this animal model. Many of the proteins listed in the table were also reported in the study of the C57BL mouse plasma glycoproteome\textsuperscript{38} which used an affinity capture of the glycopeptides (common identifications are labeled in Table 2.1). One difference with the athymic mouse model from human and other mouse studies is the very low level of immunoglobulins in the plasma samples of the nude mouse. The strong degree of correlation with previous studies does give confidence about the relevance of this animal model for the study of human disease. These results also confirm that the M-LAC approach can be used as a method for the enrichment of glycoproteins from a variety of mammalian species, which is consistent with the observations of the conservation of general glycosylation motifs across mammals\textsuperscript{33}. 
**Table 2.1. Abundant Proteins Found in Both the Nude Mouse Control Sample and a Human Non Disease Sample.**

<table>
<thead>
<tr>
<th>Spectral count</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>479</td>
<td>Alpha-2-macroglobulin</td>
</tr>
<tr>
<td>89</td>
<td>Hemopexin$d$</td>
</tr>
<tr>
<td>56</td>
<td>Complement C3</td>
</tr>
<tr>
<td>56</td>
<td>Complement factor H$d$</td>
</tr>
<tr>
<td>49</td>
<td>Kininogen$d$</td>
</tr>
<tr>
<td>39</td>
<td>Fibronectin$d$</td>
</tr>
<tr>
<td>51</td>
<td>Fibrinogen</td>
</tr>
<tr>
<td>36</td>
<td>Serum albumin$d$</td>
</tr>
<tr>
<td>26</td>
<td>Apolipoprotein A-I</td>
</tr>
<tr>
<td>25</td>
<td>Beta-2-glycoprotein I</td>
</tr>
<tr>
<td>25</td>
<td>Phosphatidylinositol-glycan-specific phospholipase D1</td>
</tr>
<tr>
<td>24</td>
<td>Alpha-2-HS-glycoprotein$d$</td>
</tr>
<tr>
<td>24</td>
<td>Fetuin-B$d$</td>
</tr>
<tr>
<td>24</td>
<td>Vitronectin$d$</td>
</tr>
<tr>
<td>20</td>
<td>Kallikrein$d$</td>
</tr>
<tr>
<td>17</td>
<td>Inter-alpha-trypsin inhibitor heavy chain H</td>
</tr>
<tr>
<td>16</td>
<td>Prothrombin$d$</td>
</tr>
<tr>
<td>15</td>
<td>Alpha-1-antitrypsin$d$</td>
</tr>
<tr>
<td>14</td>
<td>Alpha-1-microglobulin</td>
</tr>
<tr>
<td>14</td>
<td>Clusterin$d$</td>
</tr>
<tr>
<td>13</td>
<td>Haptoglobin</td>
</tr>
<tr>
<td>12</td>
<td>Afamin$d$</td>
</tr>
<tr>
<td>11</td>
<td>Complement C4$d$</td>
</tr>
<tr>
<td>3</td>
<td>Complement component C9</td>
</tr>
<tr>
<td>10</td>
<td>Complement factor B$d$</td>
</tr>
<tr>
<td>10</td>
<td>Serotransferrin</td>
</tr>
<tr>
<td>9</td>
<td>Ceruloplasmin$d$</td>
</tr>
<tr>
<td>8</td>
<td>Apolipoprotein E</td>
</tr>
<tr>
<td>7</td>
<td>Plasminogen</td>
</tr>
<tr>
<td>7</td>
<td>Serum amyloid P-component$d$</td>
</tr>
<tr>
<td>6</td>
<td>Complement component C8</td>
</tr>
<tr>
<td>6</td>
<td>Myosin</td>
</tr>
</tbody>
</table>

*a. As reported in Yang Z11.*

*b. Measured for control mouse plasma sample as determined by LC/MS analysis of a tryptic digest.*

*c. The following 5 proteins were not found in the human glycoproteome: Apolipoprotein C-III, Leukemia inhibitory factor receptor, Liver carboxylesterase N, Murinoglobulin and Serine protease inhibitor A3K.*

*d. Proteins also found in another mouse glycoproteome study 32.*
Table 2.2. Representative Results of Fetuin Peak Area Analysis

<table>
<thead>
<tr>
<th>Sample</th>
<th>Retention Time</th>
<th>Peak Area</th>
<th>Peak Height</th>
<th>Retention Time</th>
<th>Peak Area</th>
<th>Peak Height</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Tumor, nt</td>
<td>75.05</td>
<td>7.75E+03</td>
<td>9.47E+03</td>
<td>75.17</td>
<td>9.38E+03</td>
<td>8.2E+04</td>
</tr>
<tr>
<td>No Tumor, nt</td>
<td>74.96</td>
<td>9.25E+03</td>
<td>1.15E+04</td>
<td>75.26</td>
<td>7.65E+03</td>
<td>2.55E+04</td>
</tr>
<tr>
<td>Tumor, E+T</td>
<td>68.08</td>
<td>1.76E+04</td>
<td>3.96E+04</td>
<td>75.17</td>
<td>9.38E+03</td>
<td>8.2E+04</td>
</tr>
<tr>
<td>Tumor, E+T</td>
<td>74.31</td>
<td>1.49E+04</td>
<td>3.29E+04</td>
<td>75.26</td>
<td>7.65E+03</td>
<td>2.55E+04</td>
</tr>
<tr>
<td>No Tumor, E</td>
<td>73.98</td>
<td>1.34E+04</td>
<td>3.81E+04</td>
<td>75.82</td>
<td>1.37E+04</td>
<td>3.81E+04</td>
</tr>
<tr>
<td>No Tumor, E</td>
<td>74.28</td>
<td>1.17E+04</td>
<td>2.73E+04</td>
<td>76.37</td>
<td>1.27E+04</td>
<td>2.73E+04</td>
</tr>
<tr>
<td>Tumor, E</td>
<td>75.00</td>
<td>8.87E+03</td>
<td>2.68E+04</td>
<td>75.54</td>
<td>1.13E+04</td>
<td>2.68E+04</td>
</tr>
<tr>
<td>Tumor, E</td>
<td>74.72</td>
<td>8.95E+03</td>
<td>3.08E+04</td>
<td>75.27</td>
<td>1.15E+04</td>
<td>3.08E+04</td>
</tr>
<tr>
<td>Average</td>
<td>74.60</td>
<td>3.50E+03</td>
<td>1.26E+04</td>
<td>74.6</td>
<td>1.17E+06</td>
<td>3.00E+04</td>
</tr>
<tr>
<td>Std Dev</td>
<td>0.45</td>
<td>9.50E+04</td>
<td>3.24E+03</td>
<td>0.45</td>
<td>1.13E+05</td>
<td>7.00E+03</td>
</tr>
<tr>
<td>% CV</td>
<td>0.60</td>
<td>2.72E+01</td>
<td>2.73E+01</td>
<td>0.6</td>
<td>9.68E+02</td>
<td>2.34E+01</td>
</tr>
</tbody>
</table>

1. All samples are from the 6-week time point.
2. Retention time in minutes.
3. Mice with no tumor implant and not receiving either estrogen or Tamoxifen.
4. Mice with tumor and receiving both estrogen and Tamoxifen.
5. Mice with no tumor and receiving estrogen.
6. Mice with tumor and receiving estrogen.
7. Standard Deviation.

2.3.5 Data Normalization through Various Methods

i. Normalization With UV

Verification of the suitability of normalization between different samples with fetuin as an internal standard was performed by analyzing peak areas of the UV traces of the post trypsin digestion cleanup step at 40% organic mobile phase. After M-LAC fractionation, samples were concentrated to 150 µL and a 50 µL aliquot was digested with trypsin. After trypsin digestion, peptides were separated from unwanted salts and
undigested material by separation with reversed phase chromatography. Analysis of the peak area at 214 nm gives an indication of the amount of protein digested. The samples showed some variability but the %CV was less than 30%. Some variation may have been due to differing protein concentrations from animals being of different ages and/or receiving different hormone treatments.

**ii. Normalization Using Total MS Peak Area Between 40 Min And 160 Min.**

In another effort to look at reproducibility across the study, we measured the MS response (TIC) for each sample was analyzed. The Excalibur software enables a total ion chromatogram measurement as a software function. This analysis resulted in a similar amount of variation as described earlier with a %CV less than 30 (excludes the outlier T, 6, e+t).

**iii. Normalization with Bovine Fetuin Peptides**

Bovine fetuin was added to the mouse plasma as an internal standard before the first step (M-LAC fractionation) of the experiment. In the data analysis step, the fragmentation data was searched against a bovine fetuin database. Peptides identified at high probabilities were used for subsequent normalization. Both the retention time and the m/z ratio were used to locate the selected fetuin peak in the chromatogram. Peak area quantitation was then performed in each run using two different fetuin peptides. A %CV of 10 was obtained for the peak area quantitation of one of the fetuin peptides and a %CV of 30 was obtained for the other peptide. Retention times were reproducible at a %CV of less than 0.6 for both peptides.
Table 2.3. Differential Quantitation of Proteins in 6 Week Old mice with MCF-7 Implants

<table>
<thead>
<tr>
<th>Protein name</th>
<th>Estrogen/control</th>
<th>Estrogen/estrogen and tamoxifen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afamin (O89020)</td>
<td>No Change</td>
<td>2.5</td>
</tr>
<tr>
<td>Alpha-1-antitrypsin 1-1 (P07758)</td>
<td>0.3 /f</td>
<td>No Change</td>
</tr>
<tr>
<td>Alpha-2-HS-glycoprotein (P29699)</td>
<td>0.3 /f</td>
<td>No Change</td>
</tr>
<tr>
<td>Apolipoprotein C-III (P33622)</td>
<td>No Change</td>
<td>3.0 /f</td>
</tr>
<tr>
<td>Beta-2-glycoprotein 1 (Q01339)</td>
<td>0.50</td>
<td>0.4</td>
</tr>
<tr>
<td>C4b-binding protein (P08607)</td>
<td>No Change</td>
<td>No Change</td>
</tr>
<tr>
<td>Carboxypeptidase N (Q9DBB9)</td>
<td>No Change</td>
<td>0.5</td>
</tr>
<tr>
<td>CD5 antigen-like (Q9QWK4)</td>
<td>No Change</td>
<td>0.3 /f</td>
</tr>
<tr>
<td>Epidermal growth factor receptor (Q01279)</td>
<td>0.3 /f</td>
<td>No Change</td>
</tr>
<tr>
<td>Fibrinogen beta chain (Q8K0E8)</td>
<td>No Change</td>
<td>3.0 /f</td>
</tr>
<tr>
<td>Fibrinogen gamma chain (Q8VCM7)</td>
<td>No Change</td>
<td>3.0 /f</td>
</tr>
<tr>
<td>Hepatocyte growth factor activator (Q9R098)</td>
<td>0.4</td>
<td>No Change</td>
</tr>
<tr>
<td>Inositol 1,4,5-trisphosphate receptor type 3 (P70227)</td>
<td>0.5</td>
<td>No Change</td>
</tr>
<tr>
<td>Insulin receptor substrate 2 (P81122)</td>
<td>3.0 /f</td>
<td>3.0 /f</td>
</tr>
<tr>
<td>Inter-alpha-trypsin inhibitor heavy chain H3 (Q61704)</td>
<td>0.4</td>
<td>No Change</td>
</tr>
<tr>
<td>Kininogen-1 (O08677)</td>
<td>No Change</td>
<td>2.4</td>
</tr>
<tr>
<td>Leukemia inhibitory factor receptor (P42703)</td>
<td>0.3 /f</td>
<td>0.5</td>
</tr>
<tr>
<td>Liver carboxylesterase N (P23953)</td>
<td>0.4</td>
<td>No Change</td>
</tr>
<tr>
<td>Macrophage colony-stimulating factor 1 receptor (P09581)</td>
<td>0.3 /f</td>
<td>No Change</td>
</tr>
<tr>
<td>NACAM (P70670)</td>
<td>No Change</td>
<td>0.5</td>
</tr>
<tr>
<td>Phosphatidylinositol-glycan-specific phospholipase D 1 (O70362)</td>
<td>No Change</td>
<td>3.0 /f</td>
</tr>
<tr>
<td>Piccolo protein (Q9QYX7)</td>
<td>0.3 /f</td>
<td>No Change</td>
</tr>
<tr>
<td>Plasma kallikrein (P26262)</td>
<td>No Change</td>
<td>2.4</td>
</tr>
<tr>
<td>Prothrombin (P19221)</td>
<td>0.1</td>
<td>0.4</td>
</tr>
<tr>
<td>Serotransferrin (Q92111)</td>
<td>No Change</td>
<td>5.0</td>
</tr>
<tr>
<td>Serum amyloid P-component (P12246)</td>
<td>No Change</td>
<td>3.0 /f</td>
</tr>
<tr>
<td>Spectrin alpha chain, erythrocyte (P08032)</td>
<td>No Change</td>
<td>0.5</td>
</tr>
<tr>
<td>Vitronectin (P29788)</td>
<td>0.2</td>
<td>No Change</td>
</tr>
</tbody>
</table>

a. Spectral count measured as described in the experimental section.
b. Swiss-Prot accession number.
c. Control is mouse group with tumor and no treatment with either tamoxifen or estrogen.
d. 6-week time point.
e. Expressed as a ratio.
f. If one measurement was below the detection limit the value was arbitrarily set at 3 or 0.3.
g. NACAM is an abbreviation for Nascent polypeptide-associated complex alpha subunit, muscle-specific form.
2.3.6 Initial Measure of Either Up or Down Regulation of Protein Levels in Plasma in the Proteomic Data Set

The MS/MS data was first searched against the murine database downloaded from Swiss-Prot July, 2005. We used spectral counts as an initial measure of either up or down regulation of proteins in the large data set resulting from this study. Proteins which had a substantial differences (ratio of >2 or <0.5) in their spectral counts when estrogen treatment of animals with tumor implant results were compared to other treatments are presented in Table 2.3. The first comparison mimics the effects of tumor growth (control vs. estrogen treated animals) and the effect of Tamoxifen treatment on the tumor (estrogen vs. estrogen and Tamoxifen treated group). While there have not been reports of comparable proteomic studies there is a substantial literature that suggests that our results are consistent with prior studies (see below). We will discuss the biological aspects of these results in the Discussion section.

Through the analysis of spectral count, epidermal growth factor receptor (EGFR) has decreased 3-fold in the tumor bearing group receiving estrogen vs. the group with the same tumor implant but not receiving estrogen and without tumor growth. This result is consistent with other observations on the levels of EGFR in athymic mice tumors (MCF-7) in the presence and absence of estrogen. We also observed that macrophage colony-stimulating factor 1 receptor (CSF1R), a transmembrane kinase receptor, was also down-regulated in the group of mice with growing tumors vs. controls, which is consistent with previous studies. Another example, alpha-1- antitrypsin which was decreased in the estrogen treated group (Table 2.4) and levels of this protein are negatively correlated with growth of MCF-7 cells. Insulin receptor substrate 2 (IRS2)
showed a 2-fold increase on estrogen treatment of the MCF-7 implanted athymic mice which is supported by the report of Lee, et al. In studies with athymic mice with MCF-7 tumors it was found that Tamoxifen is a tumorstatic agent.

**Table 2.4.** Differential Quantitation of Murine Signaling Proteins Over Time in the Different Treatment Groups

<table>
<thead>
<tr>
<th>Description</th>
<th>N</th>
<th>No treatment</th>
<th>Estrogen only</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serine-protein kinase ATM (Q62388)</td>
<td>3</td>
<td>2.7</td>
<td>0.68</td>
</tr>
<tr>
<td>Tyrosine-protein kinase JAK1 (P52332)</td>
<td>2</td>
<td>0.71</td>
<td>0.32</td>
</tr>
<tr>
<td>Tyrosine-protein kinase JAK2 (Q62120)</td>
<td>4</td>
<td>0.50</td>
<td>0.32</td>
</tr>
<tr>
<td>Ephrin type-A receptor 8 (O09127)</td>
<td>2</td>
<td>0.21</td>
<td>0.36</td>
</tr>
<tr>
<td>Interleukin-1 receptor-associated kinase 4 (Q8R4K2)</td>
<td>4</td>
<td>0.16</td>
<td>0.33</td>
</tr>
<tr>
<td>DNA-dependent protein kinase (P97313)</td>
<td>2</td>
<td>0.98</td>
<td>0.82</td>
</tr>
<tr>
<td>Phosphatidylinositol-4-phosphate 5-kinase C2 domain (Q61194)</td>
<td>2</td>
<td>1.2</td>
<td>0.33</td>
</tr>
<tr>
<td>Phosphatidylinositol-glycan-specific phospholipase D1 (O07362)</td>
<td>2</td>
<td>1.1</td>
<td>0.33</td>
</tr>
<tr>
<td>Epidermal growth factor receptor (Q01279)</td>
<td>2</td>
<td>0.70</td>
<td>0.33</td>
</tr>
</tbody>
</table>

- All animals were implanted with the MCF-7 tumor line and received the treatment listed above.
- The sequence of the peptide used for the peak area quantitation are listed here as follows: ATM – TDFCQMLLPYHVLLQDTHESWR; JAK 1 – VLDPSHRLAFFEAAAMMRQVSHK; JAK 2 – QLAWAMHFLIEEKSILHGNVCAK; EPHA8 – LPAAMGCPRALHQLMLDCWHKDR; CSF1R – TVYFFSPWR; IRAK4 – MSDADPSVEMYASAASQCLHEKK; PRKDC – VQWLHSYSQLNHCRTQSPR; P2C3A – VFLWRENYYCLKHPNCLPK, PHLD- FGSSLVSVR and  EGFR -AVNHVCNPLCSSEGCGWPEPR.
- Each reported value is the average of 2 replicates. CV for the replicates is less than 15%.
- Units are reported to 2 significant figures for measured mass spectra ion intensity.
- CSF1R is an abbreviation for macrophage colony-stimulating factor receptor. PHLD is an abbreviation for phosphatidylinositol-glycan-specific phospholipase D. EGFR is an abbreviation for epidermal growth factor receptor.
- N is the number of unique peptides identified.
- SwissProt accession number
It is of interest, therefore, that many of the changes observed after estrogen treatment were not observed in implanted, athymic mice treated with estrogen and Tamoxifen. For example, the levels of EGFR, macrophage CS1R and alpha -1-antitrypsin changed from a strong decrease with estrogen to no change in the case of the estrogen/Tamoxifen treatment (Table 2.4). Conversely IRS-2 went from a 2-fold increase to a 3-fold decrease in the shift to Tamoxifen treatment. Koibuchi, et.al.,\textsuperscript{36} and Molloy, et al.,\textsuperscript{41} have reported similar changes on Tamoxifen treatment in the same mouse model for EGFR and the IRS-1 pathway respectively (measured by immunoassay).

In our approach, the second step in the measurement of the differential quantitation of a selected protein is to measure the peak area of at least two peptides in extracted ion chromatograms. We also examined the corresponding MS/MS spectra (see Figure 2.4) to ensure that there were no co-eluting contaminant peptides within the mass window threshold and required a consistent set of MS/MS spectra across the sample set of different time points and treatments. As described above, bovine fetuin was added as an internal standard prior to the M-LAC fractionation step and data was searched against a bovine fetuin database. While several peptides were identified at a high confidence level, two peptides which gave the best reproducibility across the data set (CV 10% and 30% and retention time CV of less than 0.6%) were examined in more detail. Finally we decided on the use of the peptide, TPIVGQPSIPGGP MW: 1475.72, charge: +2, RT 74.60 min with the lowest CV, for the purposes of normalization of the peak area data. Fetuin peak areas are tabulated in Table 2.2.
Figure 2.4. Selection of the Appropriate M/Z Range for Peak Area Quantitation. (a) The scan number matching the Sequest search results for the peptide is selected from the total ion chromatogram. This scan will be the ms/ms scan which was used to identify the peptide, the precursor ion m/z and the retention time. The precursor m/z should match the calculated m/z from the Sequest results. In this case Vitronectin is identified by a peptide of a MW of 2002.295 with a charge of +2 and probability of 1.0000. The m/z range for peak area quantitation is 1001.61. (b) Using the precursor m/z and the retention time the peak representing the vitronectin peptide can be located in the extracted ion chromatogram.
In Table 2.3 the peak areas of peptides from several proteins of interest are tabulated. These peptides were then measured across a series of time points with the requirement that the retention times met established reproducibility values (RSD < 5%) and the S/N values were >3 (or the raw signal intensity was > 1 x 10⁴). The relative peak areas were either reported as the mean of measured peak areas or as the area recorded for the peptide with the best S/N value (cross-checked with the other peptides for consistency). If the areas disagreed by > 30% then the values were not utilized. The results of these measurements with relation to the biology of this cancer model will be reported in the Discussion section. As an example, EGFR was shown to be decreased approximately 3-fold in the 6-week-old mice which were treated with estrogen (maximum tumor volume) relative to the control group (Table 2.3 pg 102) which was consistent with the spectral count measurement.

2.3.7 Identification of Tumor Specific Peptides

Previous studies on xenografts of human cell lines in athymic mice have been unable to characterize relatively low levels of human proteins in the presence of abundant mouse plasma proteins¹⁹. For the characterization of human specific proteins we therefore selected the sample which represented the mice having the greatest tumor mass and with the highest level of secreted tumor proteins. This group, as can be seen in Figure 2.1 (page 83), is composed of the mice which received estrogen at the 6-week time point with a tumor volume of 736 vs. < 250 mm³ at the 3-week time point. By utilizing a high mass accuracy LC-LTQ-FT instrument we were able to distinguish either human or mouse
specific peptides. The MS/MS fragmentation spectra from these experiments were searched against both a mouse and human sequence database. Peptides, found searching against the human database were then filtered using criteria suggested by HUPO and a Protein Prophet probability of 0.90\textsuperscript{3} and a mass accuracy difference of less than 2 ppm from the theoretical value for the identified sequence. The peptide list generated by the human database search was compared with the list resulting from the mouse database search. Human unique peptide sequences were then checked by BLAST searches against human and mouse sequences for species specificity. A further requirement was the absence of the human protein in a negative control (no tumor control mouse at t\textsubscript{0}). The peptides coding only to mouse proteins and the peptides present in both mouse and human proteins (i.e. homologs) were discarded so that unique peptides identified only in the human database search remained. As an additional precaution any proteins resulting from single peptide identifications were also discarded. This resulted in the identification of 17 proteins by 2 or more high confidence peptides with at least one peptide being unique to human (Table 2.6). The disease association of these proteins will be described in the Discussion Section.

2.4 Discussion

2.4.1 The Nude Mouse Xenograft as a Model for Breast Cancer

The study of the etiology of breast cancer in human subjects is challenged by genetic and environmental factors\textsuperscript{22}. An alternative approach is to study established breast cancer cell lines such as SKBR-3, A431 and MCF-7 with environmental agents\textsuperscript{22-24}. Such studies are restricted by the differences between cell lines relative to the
heterogeneous environment of tumor cells in vivo\textsuperscript{22, 23}. The use of xenograph models in immunosuppressed mice\textsuperscript{22} allows implantation and subsequent study of human tumors.

The nude mouse possesses a compromised immune system which was selected to make it amenable to studies requiring the implantation of foreign tissue. A defective thymus results in malfunctioning T-cells which are unable to produce immunoglobulins. Depletion with protein G resulted in the observation that there were few immunoglobulins in the mouse blood (see Figure 2.3, pg99) and support the decision not to utilize any steps for the depletion of abundant proteins. The lack of IgGs may be a contributor to the discovery of many low abundance proteins in the final data either due to less sample handling or to an intrinsic lower dynamic range in the sample.

In this study ovariectomized, athymic mice received an injection of MCF-7 cells, which were implanted with an estrogen pellet. After establishment of the tumor (see methods section), the mice were randomized into groups and received the treatments shown in Fig.2.1 (page 83). As is shown in Fig. 2.2, the control group showed little growth of the tumor, as did the group receiving Tamoxifen. The latter group serves as a control for the effects of Tamoxifen, and the lack of tumor growth confirms the tumorstatic properties of this drug. The groups of animals of greatest interest were those who received estrogen (maximum tumor growth) and the combination of estrogen and Tamoxifen (intermediate growth, as was also reported by Koibuchi Y, 2000)\textsuperscript{36}.  

108
2.4.2 The Plasma Proteome and Secreted Kinases

Recently there have been a large number of studies of the plasma proteome and advanced methodologies used in a multi-laboratory study have achieved the characterization of 2000 to 3000 proteins. In a similar manner, we were able to identify greater than 1800 proteins with high probability identifications in replicate measurements on individual pools and in the different treatment modalities. With such improved dynamic range of the proteomic measurement it is possible to identify relatively low level proteins including members of signaling transduction pathways. In a similar manner, the Human Proteome Organization (HUPO) plasma proteome study identified 272 kinases.

An important issue for a plasma biomarker study is the mechanism by which proteins associated with signaling pathways are released into the blood compartment and what information can be related to biology of the tumor. In this context the microenvironment of tumor – stromal interactions in carcinoma formation can play a key role in growth factor mediated pathways, but in addition the body can mount significant inflammatory and immune responses. These events can result in processes such as activation of cells of the monocyte-macrophage system, increase in tumor-associated macrophages, inflammation mediated cytokine release and angiogenic factors. It has been shown that any or all of these events can alter the regulation of signal transduction pathways. Therefore it is difficult to attribute the tissue origins of a given blood protein and illustrates the advantage of the mouse xenograft model. In the study reported here we were able to identify several kinases including ephrins which could be identified as being of mouse origin through careful comparison of protein and peptide sequences.
2.4.3 Ephrin

Ephrin A3 is produced by many tissues (brain- critical role in axon guidance and repulsion in development, also in skeletal muscle, spleen, thymus, testis, ovary, small intestine). Mouse and human Ephrin A3 are 96% homologous. This protein could possibly be shed by the tumor due to the fact that it is expressed by MCF-7\textsuperscript{49,50} and in breast cancer\textsuperscript{51}. In 2.6 we list the ephrin receptors characterized in the proteomics study. We found four other ephrin receptors which are closely related in the top 10 matches of the homology search are also observed in the study (A2, A7, A8, B3 and B4). Closely related proteins such as these may have homologous peptides making identification of the protein difficult. Ephrin A3 is the best characterized receptor protein in this proteomics study due to the sequence coverage by the peptides. In Figure 2.5 it can be seen that there is coverage near the N- and C-terminus and the interior of the molecule but not near the transmembrane domain. Eph receptors, which interact with ephrins, are the largest recognized subfamily of receptor tyrosine kinases and have been shown to be involved in pathological processes such as carcinogenesis. It is known that Eph receptors and their ligands are critical for vascular development and are involved in angiogenesis in breast carcinoma. The significance of the function of Eph receptors proteins is rooted in a greatly advanced appreciation of the role of ephrins in the regulation of cell motility, invasion and metastasis.
Table 2.5. Top 10 Matches of the Ephrin A3 Homology Search

<table>
<thead>
<tr>
<th>Accession Number</th>
<th>Description</th>
<th>Length</th>
<th>Blast score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q80VZ2_MOUSE</td>
<td>A4 Eph receptor A4 (CDNA, RIKEN full-length enriched library, clone:M5C109J03 product:Eph receptor A4, full insert sequence) [Ephr4] [Mus musculus (Mouse)]</td>
<td>986 AA</td>
<td>1228</td>
</tr>
<tr>
<td>EPHA7_MOUSE</td>
<td>A7** Ephrin type-A receptor 7 precursor (Tyrosine-protein kinase receptor EHK-3) (EPH homology kinase 3) (MDK-1) [Mus musculus (Mouse)]</td>
<td>998 AA</td>
<td>1226</td>
</tr>
<tr>
<td>EPHA6_MOUSE</td>
<td>A6 Ephrin type-A receptor 6 precursor (Tyrosine-protein receptor kinase receptor EKH-2) (EPH homology kinase 2)</td>
<td>1035 AA</td>
<td>1226</td>
</tr>
<tr>
<td>EPHA8_MOUSE</td>
<td>A8** Ephrin type-A receptor 8 precursor (Tyrosine-protein receptor kinase receptor EEK) [Mus musculus (Mouse)]</td>
<td>1004 AA</td>
<td>1121</td>
</tr>
<tr>
<td>EPHB2_MOUSE</td>
<td>B2 Ephrin type-B receptor 2 precursor (Tyrosine-protein kinase receptor EPH-3) (Nucleotide kinase) (SEK-3) [Mus musculus (Mouse)]</td>
<td>994 AA</td>
<td>1071</td>
</tr>
<tr>
<td>QISP23_MOUSE</td>
<td>B1 Eph receptor B1 [Ephb1] [Mus musculus (Mouse)]</td>
<td>984 AA</td>
<td>1063</td>
</tr>
<tr>
<td>QITY9_MOUSE</td>
<td>B3** Eph receptor B3 [Ephb3] [Mus musculus (Mouse)]</td>
<td>977 AA</td>
<td>1043</td>
</tr>
<tr>
<td>EPHA2_MOUSE</td>
<td>A2** Ephrin type-A receptor 2 precursor (Tyrosine-protein kinase receptor EKH) (EPH homology kinase 2) (MPK-5) (SEK-5) [Mus musculus (Mouse)]</td>
<td>977 AA</td>
<td>927</td>
</tr>
<tr>
<td>EPHB4_MOUSE</td>
<td>B4** Ephrin type-B receptor 4 precursor (Tyrosine-protein kinase receptor MDK-2) (Developmental kinase 2) (Tyrosine kinase MYK-1) [Mus musculus (Mouse)]</td>
<td>987 AA</td>
<td>869</td>
</tr>
<tr>
<td>EPHA5_MOUSE</td>
<td>A5 Ephrin type-A receptor 5 precursor (Tyrosine-protein kinase receptor EKH-1) (EPH homology kinase 1) (Brain-specific kinase) (OEK-7) [Mus musculus (Mouse)]</td>
<td>877 AA</td>
<td>733</td>
</tr>
</tbody>
</table>

** These ephrin proteins were also found in the mouse plasma study.

Figure 2.5. Ephrin A-3 Peptides Discovered in the Proteomic Analysis of Mouse Plasma. There were 7 peptides (shown underlined) in the complete data set. The transmembrane domain (shown in brackets with an asterisk) was not detected.
2.4.4 Differential Quantitation of Selected Murine Kinases

Of the many different types of proteins found in the mouse xenograft proteomic data kinase proteins were selected as the group having the greatest chance of being specific to breast cancer. Kinases are enzymes which add a phosphate group to a substrate protein and thus control signaling in many important biological pathways such as cell division.

Figure 2.6. Effects of Estrogen on Mouse (M) and Mouse With Tumor (Tu). In the figure 0, 3 and 6 stand for the time of sample collection. All of the peak areas measurements are normalized by dividing by 10*5. In this analysis, the corresponding control (mouse, no tumor, no implant or mouse, tumor, no treatment) was subtracted from the reported values for the effects of estrogen on the mouse (M) or mouse with growing tumor (Tu) respectively and the relative errors have been summed. Other proteins which show similar changes such as EPHB3, JAK2 and PRKDC are not shown.
We interrogated the entire data set (>1800 identifications) for proteins associated with signaling pathways and identified 69 kinases (data not shown). This list was curated for disease association and then proteins were selected for label-free differential quantitation (an important criteria was that the candidate was present at a level sufficiently high for robust measurement\(^2\)). As described in the Methods section, the quantitation process included either spectral count values (Table 2.3, page 102) or peak area measurements (Table 2.4, page 104). The changes in protein concentration listed in Table 2.4 are given for the different time points (0, 3 and 6 weeks) for the no treatment, estrogen, Tamoxifen and estrogen plus Tamoxifen treatment group (all groups contain the MCF-7 implant).

Tables 2.3 and 2.4 show that some proteins increased during the time course of the estrogen treatment (insulin receptor substrate 2 (IRS2), leukemia inhibitory factor 1 receptor (LIF), some showed an increase at the intermediate time point and then a decrease (tyrosine-protein kinase JAK2, serine protein kinase ATM and Interleukin-1 receptor-associated kinase 4 (IRAK)) and the rest showed decreases (PI3-K, DNA-dependent protein kinase 4 (DNA-PK), interleukin-1 receptor associated kinase 4 (IL-1RK), JAK1, EGFR, vitronectin). In our quantitative measurements the variability between analytical replicates was less than 10%, but to account for biological variability we generally consider a 2 fold change as significant. The following proteins met this criteria across the time points for the animals with estrogen stimulation (all decreases) - JAK1 (14 fold), PI 3-kinase (below our detection limit after the initial time point) and EGFR (3.7 fold). These proteins showed no significant changes in the control animal.
which suggested that decreased levels of these proteins in the estrogen group could be associated with tumor growth.

The observations of decreases in JAK1 and EGFR are supported by other reports of reduced JAK1 expression in breast cancer tissues vs. matched non-cancer tissues\textsuperscript{53} and serum levels of the extra cellular domain of EGFR was observed to be decreased in breast cancer patients\textsuperscript{54}. Furthermore, studies of MCF-7 cell lines have found the presence of estrogen can suppress expression of EGFR\textsuperscript{55}, which is consistent with our observation that human EGFR was not observable in the estrogen treated group (see next section). An advantage of the xenograft model is that we could identify EGFR with mouse specific peptides (see Table 2.6) and thus demonstrate a non tumor source. While the tissue(s) of origin for the EFGFR observed in the plasma samples is not known, a possible candidate is the adjacent stromal tissue which is known to express growth factors and related receptor kinases\textsuperscript{46, 56, 57}. Future studies with this model system will be required to understand the potential role of stromal proteins in the development of the cancer related plasma glycoproteome, but interestingly such signaling can occur with decreased receptor expression as was observed in this study\textsuperscript{57-60}. To further explore this point Figure 2.6 shows that a comparison of the two treatment groups (animals with no tumor and with tumor) can allow one to estimate the effects of estrogen on the mouse versus the more complex interaction with animal and tumor. Interestingly, the figure shows that there are opposite changes in levels of the kinases for the two groups, which mimics the heterotypic expression reported above and is consistent with the presence of stromal tissue in the tumor bearing animals.
In the estrogen plus Tamoxifen treatment group, in which we observed intermediate tumor growth, a number of proteins, serine-protein kinase ATM, Ephrin receptor A8, interleukin-1 receptor associated kinase (IRAK) and EGFR were decreased, while phosphatidylinositol-4-phosphate 3-kinase C2 domain (PI 3-kinase) showed a significant increase (3-fold). These results were consistent with previous studies which have shown that Tamoxifen treatment of MCF-7 xenografts reduced tumor growth rates, increased apoptosis significantly \(^{61}\), down regulated EGFR, and that this receptor is further decreased by estrogen \(^{36}\). The final treatment group (Tamoxifen) is the least relevant to this study as no tumor growth was observed in this animal group. Table 2.3 (pg 102) shows that substantial increases and decreases were observed with all proteins, except for phosphatidylinositol-glycan-specific phospholipase D1 (PHLD), including a 15-fold increase in PI 3-kinase, which could be attributed to additional, non estrogen-receptor mechanisms \(^{62,63}\). The change over time observed for this key enzyme in the combined treatment group (increase with estrogen plus Tamoxifen treatment vs. decrease with estrogen treatment) could be attributed to such non receptor effects of Tamoxifen.

### 2.4.5 Characterization of Tumor (Human) Specific Proteins

As described in the results section the mouse group with the expected highest level of tumor proteins (6-weeks with estrogen treatment) was selected for the attempted characterization of human specific peptide sequences. While such proteins have not been previously characterized for mouse xenografts, other studies have successfully studied mixed proteomes via the measurement of unique species specific peptides \(^{58}\). Table 2.6 gives a conservative list of human specific peptides that were measured in the FTMS (see
Results section). To meet our conservative criteria each protein had to be identified with at least two peptides with at least one human specific peptide (shown in italics). The table groups the proteins into the following classifications: growth factor and cytoskeletal signaling, immune response, and transcriptional regulation. It was of interest to observe that the proteins discovered had in many cases extensive association with breast cancer (literature associations are listed as footnotes to Table 2.6).

As shown in Table 2.6 many of the lower level proteins were only detected by specific human peptides without identification of corresponding mouse sequences. In the case of abundant proteins (spectral count greater than 20) both mouse and human proteins were observed which suggests both mouse and tumor origins. The remaining proteins, which were present at lower levels (spectral count ≥ 2) are probably glycosylated and in many cases potentially secreted and are thus candidates for tumor specific markers, such as LDL receptor-related protein 1B and ATP binding protein A2. Conversely known tumor associated proteins such as EGFR, JAK, and PI3-K were found in this study only with mouse specific peptides and thus must have a non-tumor origin. While the biosynthesis of such proteins may indeed be tumor related, such as a stromal origin, these proteins can not be directly associated with the tumor unless expression profiling studies are performed to demonstrate a unique site of synthesis. The value of the mouse model in studying human tumor secretion is emphasized by the difficulty in performing such profiling studies in man, and furthermore, even if a protein is expressed and thus measured by expression profiling it may not be secreted.

In the future we believe that higher levels of stringency could be achieved by performing an integrated protein and glycan analysis, as cancer has been associated with
the generation of abnormal or rare glycan motifs. The corresponding protein analysis has the potential of determining the tissue of origin for the glycan. Our approach to such an investigation of tumor specific glycoproteins will be to use affinity isolations with biospecific ligands such as antibodies or lectins, followed by both protein and glycan characterization using methodology previously described⁶⁴-⁶⁶. We are hopeful that such an approach will identify tumor specific biomarkers in mouse plasma samples that can be translated to human studies.
<table>
<thead>
<tr>
<th>Description</th>
<th>A</th>
<th>B</th>
<th>Peptide</th>
<th>Δ ppm</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Growth factor and cytoskeletal signaling</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fibrinogen gamma chain (S)</td>
<td>5</td>
<td>39</td>
<td>DNCCILDER</td>
<td>-0.74</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>EGFQHLPTGTETFWLGNK</td>
<td>-0.14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>NWIQXYEGFFHLSPTETFWLGNK</td>
<td>-0.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>YLQEIYNSNNQK</td>
<td>-0.85</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>YLQEIYNSNNQKVNLKEK</td>
<td>-0.37</td>
</tr>
<tr>
<td>Vitronectin(S)</td>
<td>5</td>
<td>15</td>
<td>DWHGVPQVDAAMAGR</td>
<td>1.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>INCQKTYLFKGSQYWR (2+)</td>
<td>-0.31</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>INCQKTYLFKGSQYWR (3+)</td>
<td>-1.30</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TSAGTRQPQFSRDWHGVPGQVDAAMAGR</td>
<td>-0.69</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TYLFKGSQYWR</td>
<td>-0.89</td>
</tr>
<tr>
<td><strong>Low-density lipoprotein receptor-related</strong></td>
<td>4</td>
<td>0</td>
<td>ECPGLSYVNEKLYWISSGNTINR</td>
<td>-0.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SGRCISSKWHCDDDDCDSVGCSVHSCFDFNQFR</td>
<td>1.30</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>STLTKAQNQVYSTDILTPNLTDYR</td>
<td>-1.39</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>YSINVQQLSNNPCTRLLMNAGEPEYAIAVNPK</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>DNSNASNMSKQNSLIFRASVTLPINPMAPALEVDQVTNLNK</td>
<td>1.39</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>KELISAVEVFLQDCQYFLDYDNLGMPASGLKQPEASDR</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RQSFIFNFLPSEAVDKIFGIVGVGHCTQR</td>
<td>1.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>IEAVMTSCQGLNENITIMISNTANSQMGR</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PKSAEAKSNQK</td>
<td>-0.43</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TQVLQGHCVKRD</td>
<td>-0.84</td>
</tr>
<tr>
<td><strong>Immune response</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Haptoglobin(S)</td>
<td>3</td>
<td>31</td>
<td>GSFPWQAK</td>
<td>-0.90</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>VGYVYSGWGR</td>
<td>-0.59</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>VTSIQDWWQKTIAEN</td>
<td>-0.57</td>
</tr>
<tr>
<td>Alpha-2-macroglobulin(G)</td>
<td>2</td>
<td>61</td>
<td>QTVSWAVTPK</td>
<td>-0.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>YGAATFTR</td>
<td>-0.53</td>
</tr>
<tr>
<td>Golgi autoantigen, golgin subfamily A member 5(M)</td>
<td>2</td>
<td>0</td>
<td>IMIQDQSEGNLSQNQALQTLQERHEADATLKR</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>NQLTNKTLSNSSQSELENSRLQHTETLQK</td>
<td>-0.32</td>
</tr>
<tr>
<td>Glutamate decarboxylase 2(M)</td>
<td>2</td>
<td>0</td>
<td>MREHPWGGSGQFQPGSAISNMAMNIAR</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>QKGFPVLVSATAGTIVYVAGFDPLLAVADICK</td>
<td>-0.61</td>
</tr>
<tr>
<td>Interferon regulatory factor 7</td>
<td>2</td>
<td>0</td>
<td>VLFGEWLLLGEISSGCGYGLQWLDEARTCFR</td>
<td>-0.72</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>VVGHPSCTFLYGPDDKPVRATDPQQVAFSPAELPDQKQLR</td>
<td>0.71</td>
</tr>
<tr>
<td>Phosphatidylinositol-4,5-bisphosphate 3-kinase</td>
<td>2</td>
<td>0</td>
<td>KPQYYYPPFKIEK</td>
<td>-1.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>WNKLEDVAQLAQLIQWPKLPPR</td>
<td>-1.22</td>
</tr>
<tr>
<td>Protein kinase C and casein kinase substrate</td>
<td>2</td>
<td>0</td>
<td>FREVELKEQVYKHDLNSVAGVK</td>
<td>-0.69</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>HLDSLNSVAGKIAVHDLEQSIARAADAVEDL</td>
<td>-0.23</td>
</tr>
<tr>
<td><strong>Transcriptional regulation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP-binding cassette sub-family A member 2(G,M)</td>
<td>3</td>
<td>0</td>
<td>FLWNLINTDLIKTGRSVVLTHSMEEECALCTR</td>
<td>-0.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>QDNFSLPSGMALLQQLDIDNAACGWIQMSK</td>
<td>-0.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>KLSATIALTGYPAFILDLFEEPTGMDPKAR</td>
<td>-0.37</td>
</tr>
<tr>
<td>Y-box binding protein 2</td>
<td>3</td>
<td>0</td>
<td>GGGGACGGGGAASGPAAGTPSAAGRTQNPATAVSGTPAPPARRQADK</td>
<td>-0.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PVLAIQVVLTVKFWVFVRNGYGFIR</td>
<td>-0.69</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>VLGTWKFVNFVRNGYGFIR</td>
<td>0.23</td>
</tr>
<tr>
<td>Host cell factor(G)</td>
<td>2</td>
<td>0</td>
<td>GTNVMVTHYLFPPPDVPSDDDLGTVVPPQYQK</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>PSLSGAVPLRSLHSAITGPQNNKMYVFGGWVPLMVDDV</td>
<td>-0.83</td>
</tr>
<tr>
<td>Transcription factor 4</td>
<td>2</td>
<td>0</td>
<td>KVPQGPLSSVAPSASTDAYRDSYPSFSS</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SGTHYSTSSCCTPANGTDSIMANRGSGAGCASEGSSQTGDALG</td>
<td>0.19</td>
</tr>
<tr>
<td>Valyl-tRNA synthetase</td>
<td>2</td>
<td>0</td>
<td>GETTLWNPQCDHAGIATQVYVVEKCLWR</td>
<td>-1.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ITPAHDQNYEVQQRHGEAISIMDR</td>
<td>1.02</td>
</tr>
</tbody>
</table>

*a. The following abbreviations are used:*
M-Membrane associated, G-glycoprotein and S-Secreted. As reported in Expasy (expasy.org)

b. The number of peptides found in human database search.
c. The number of peptides found in mouse database search.
d. Human specific peptides are in bold face.
e. Difference between measured peptide MW and theoretical MW

f. Growth factor and cytoskeletal signaling: dynein motors are associated with anchorage independent growth and cytoskeleton signaling. Vitronectin is one of the neoantigens of C5b-9 complement complex which are deposited on the membranes of tumor cells; vitronectin regulates integrin function in breast cancer; LDL receptor related protein 1B acts as a tumor suppressor; and binding of CD44 to ankyrin members with oncogenic signaling.

g. Immune response: phosphatidylinositol 4, 5-bisphosphate (PIP2), protein kinase C and a related IFN regulatory factor (IRF-3) regulates interferon I (IFN1) gene expression in ovarian cancer; mammaries immortalized and increased tumorigenicty; glutamine acid decarboxylase (GAD) is increased in mammary tumor; acute phase proteins, haptoglobin, and alpha-2 macroglobulin in breast cancer.

h. Transcriptional regulation: Initiation transcription factor 2 (Itf2 or E2-2) is elevated in colon cancer and melanoma. Y-box binding protein, and growth factor independence in mammary epithelial cells and valyl-tRNA synthetase was mapped to the MHC class III region (lung tumorigenesis); ATP-binding cassette A2 (ABCG2) or breast cancer resistance protein (BCRP) is associated with multidrug resistance; Host cell factor 2 (HCF-2) is associated with cell proliferation.

2.5 Conclusions

In this study, we report an in-depth plasma proteomic analysis of a mouse xenograft of the human breast cancer cell line MCF-7. We focused on the glycoproteome by using a novel platform named M-LAC (multi-lectin affinity chromatography) which both increased the dynamic range of the study and primarily selected for secreted proteins. A large set of proteins were identified including the following which exhibited more than a 2-fold change under the condition of tumor growth: phosphatidylinositol-4-phosphate 3-kinase, tyrosine-protein kinase JAK1 and epidermal growth factor receptor.

The use of a xenograft mouse model has the potential of distinguishing tumor proteins (human peptide sequences) from host response (murine). In this study we demonstrated for the first time that high accuracy mass measurements could detect tumor-specific proteins derived from the murine host as well as human specific proteins from the tumor.

In addition, this study supports the continued development of the M-LAC approach in animal models as well as humans to identify markers of tumor presence and response to therapy. Based on the results presented in this report we have identified future studies directed at generating insights into the release of tumor markers into blood and thus facilitate the search for biomarkers for the early detection of cancer.
2.6 References


73. Grossoni VC, Falbo KB, Kazanietz MG, de Kier Joffe ED, Utrreger AJ. Protein kinase C delta enhances proliferation and survival of murine mammary cells. Mol Carcinog. 2007 Jan 11; [Epub ahead of print]


Proteomic and Peptidomic Analysis of Serum from Patients Having
Hyperinsulinemia After Gastric Bypass Surgery.

Acknowledgment.
I would like to acknowledge the financial and technical support of The Joslin Diabetes Center. I would like to offer thanks to Dr. Marina Hincapie and Dr Haven Baker for technical guidance. I must also thank Majlinda Kullolli for assistance in the lab.
3.1.1 The Challenge of Diabetes

Obesity and diabetes are closely intertwined conditions which are concurrently increasing in occurrence. Commonly, it is suggested by health professionals that avoidance of obesity will reduce occurrence of diabetes\(^1\). Diabetes is generally separated into Type 1 and Type 2. Type 1 diabetes, where insulin is produced at low levels, is usually diagnosed in children and young adults. Insulin hormone is required for the uptake of carbohydrates into energy\(^2\). Late onset diabetes, termed Type 2, is characterized by normal insulin production and diminished response by the body to insulin\(^3\).
Due to medical and social pressures, obese individuals are often willing to undertake the risks of gastric bypass surgery. Roux-en-Y gastric bypass surgery, as diagrammed in Figure 3.1, reduces the size of the stomach and forms an attachment to the small intestine via a Y shape. While it is generally a beneficial and life improving procedure, some patients develop a metabolic syndrome characterized by abnormally high levels of insulin.

Proteomic analysis of the patients’ plasma or serum may uncover significant changes in types and amounts of proteins related to causes of diabetes. Some researchers believe that a hormone change causes β-cells to become deregulated and begin to overproduce insulin. A larger question remains to be debated as to whether the deregulation is mechanical in nature (i.e. caused by the surgery) or a preexisting disease.

Autoimmune causes of insulinoma, either hyper insulinoma or hypo insulinoma, may be due to genetic predispositioning resulting from an immune system failure to recognize self-proteins and the production of autoantibodies. Over production of self-proteins or alterations to self-proteins such as glycation, in some patients after gastric bypass surgery may initiate an immune response ⁴.

### 3.1.2 Proteomic Analysis of Plasma from Gastric Bypass Patients

Several analytical methods were employed to investigate plasma and serum for protein and peptide disease markers including antibody depletion of very abundant proteins, glycoprotein enrichment and ultrafiltration. In the previous chapter regarding the analysis of the mouse xenograft plasma, it was determined to be unnecessary to deplete the plasma due in part to the lack of immunoglobulins in the Bulbic mouse. In the
case of the human serum samples, removal of immunoglobulins and albumin extends the
dynamic range of the assay to aid in the discovery of low abundance protein biomarkers.
The multi lectin columns are capable of enriching glycoproteins and thus improving the
dynamic range of the measurements. In addition, analysis of the bound fraction, which
contains the majority of the glycoproteins, can benefit from prior depletion of high
abundance glycoproteins such as alpha -1- glycoprotein. Therefore it was determined to
be necessary to evaluate alternative protocols and to adjust our methods to obtain the
most practical and effective results.

In the initial phase, a 12-protein depletion column was used prior to M-LAC
fractionation of pooled plasma samples. Depletion of the 12 most abundant proteins is
theoretically an ideal way to enrich low level biomarkers, but there are currently issues
with column deterioration which result in unpredictable amounts of breakthrough by the
target proteins. Leigh Andersons’ graph of protein abundances points out that plasma
proteins possess an extraordinarily large dynamic range of abundance

After an initial investigation of the nonbound M-LAC fraction from the 12-
protein depletion step, it was determined to use a comprehensive depletion step and focus
on the glycosylated proteins after minimal abundant protein depletion in an effort to
avoid bias in a clinical study.

A more practical, robust approach in which only albumin and IgGs are depleted
was used in a second analysis of the pooled serum. As albumin comprises more than 50%
of the protein in serum its removal alone can greatly aid our ability to detect low
abundance biomarkers. Immunoglobulins which are glycosylated can bind to the M-LAC
column, decreasing its capacity to bind important proteins, thus necessitating their removal.

### 3.1.3 Peptidome Analysis

Although, analysis of plasma proteins is best achieved with proteomic methods such as enzymatic digestion, these methods will not be as useful to identify endogenous peptide hormones due to the fact that digestion of a plasma sample with trypsin may cleave such peptides. The peptidomic method which uses employed in this study utilizes molecular weight cut off filters to remove material having a molecular weight more than 10,000 Da and relies on a Fourier transform-linear ion trap hybrid mass spectrometer to provide the high accuracy molecular weights needed for peptide identification.6

### 3.1.4 Autoimmune Investigation

Diseases typified by the attack on self-proteins by the immune system such as Rheumatoid arthritis and type 1 diabetes are termed autoimmune diseases. The endocrine system which includes hormones responsible for glucose levels such as insulin and glucagon is a common target of autoimmune attack7-9. Recent progress in the understanding of the causes of autoimmune diseases has identified genetic factors responsible for immune tolerance of self proteins7.

In light of these considerations, we performed an experiment designed to screen for an autoimmune cause of the metabolic syndrome observed in those patients having complications after gastric bypass surgery. This experiment includes a fractionation step using protein G to collect immunoglobulins from each of the serum samples due to the
fact that Protein G, protein which binds immunoglobulins via the Fc region and is produced in Streptococcal bacteria, is a cost effective method to separate immunoglobulin G with no species specificities. The immunoglobulin depleted serum was loaded onto a SDS-PAGE gel to separate the proteins according to molecular weight, then electrophoretically transferred to nitrocellulose and finally incubated with the immunoglobulins collected from the disease samples and anti-human IgG horseradish peroxidase (HRP) conjugate enable detection of any bound immunoglobulins.

In light of the severity of the symptoms displayed by these patients and the fact that the surgery induced the disease, a thorough investigation was warranted. If a protein biomarker, peptide marker or autoantigen can be found as the cause in addition of the disease mechanism, it may elucidate one of the body’s paths to metabolic syndromes.

3.2 Experimental

3.2.1 Proteomic Experiments

3.2.1.1 Materials

Concanavalin A (Con A), Jacalin and Wheat germ agglutinin (WGA) agarose bound lectins were purchased from Vector labs (Burlingame, CA). Disposable polypropylene columns were purchased from Pierce (Rockford, IL). Bovine fetuin and all other chemicals were purchased from Sigma-Aldrich (St. Louis, MO). The BCA protein assay reagent kit was from Pierce (Rockford, IL) Sequencing grade, modified trypsin was purchased from Promega (Madison, WI). 5kDa Amicon molecular weight cut off filters were purchased from Millipore (Billerica, MA). Discovery BIO Wide Pore C18 cartridges (C18, 2 cm x4.00 mm, 3 µm particles, 300 Å pores) were from Supelco
Pre-packed Protein G columns were purchased from Pierce (Rockford, IL). Tris-bis 4-12% gradient gels, See Blue protein ladder, 10X SDS-Tris buffer, Loading buffer were purchased from Invitrogen (Carlsbad, CA). Gels were run on an Invitrogen mini electrophoresis unit. Trifluoroacetic acid and Formic acid were purchased from Pierce (Rockford, IL). Micron centrifugal filters with a MWCO of 10,000 Da were purchased from Millipore (Billerica, MA). A Discovery BIO wide pore C18 guard column was used to clean up samples.

3.2.1.2 Preparation of Human Serum Samples

The four groups of serum samples collected by Joslin Diabetes center in this study were age and weight matched and presumably otherwise healthy. Included was a normal control of healthy donors, a normal obese group, and a non-symptomatic post gastric bypass group and symptomatic. Aliquots from each individual of each treatment group were combined for the pooled sample analysis to reduce variability contributions from individuals and frozen at -80°C until testing.

3.2.1.3 Depletion of Pooled Samples for Proteomic Analysis

A HP 1090 HPLC system (Agilent, Santa Clara, CA) was used for depletion of HSA and IgGs. Pooled serum (50 µL) was diluted with 200 µL of loading buffer. Bovine fetuin (150 µL) was added as an internal standard. The depletion column consisted of protein A resin and a separate anti HSA column installed on a HP1090 HPLC system. Loading buffer consisted of 0.1M Tris, 0.5 M sodium chloride at a pH of 7.4. Elution buffer was 0.2 M glycine at a pH of 2.5 and an equilibration buffer was composed of 0.5 M Tris, 1 M sodium chloride at a pH of 8.0. Unretained proteins were
collected for further fractionation with M-LAC. Retained immunoglobulin and HSA were washed from the columns with glycine buffer (pH 2.5) and analyzed for protein concentration by Bradford analysis.

3.2.1.4 M-LAC Fractionation of Depleted Samples

Pooled samples were randomized prior to analysis to avoid bias and fractionated using multi lectin (M-LAC) columns as described in Chapter 2, on Pg 9.

3.2.1.5 Tryptic Digestion and Reversed Phase Desalting Using HPLC

Peptides were separated from salts and any undigested material by chromatography on a Discovery BIO C18 column installed on a Shimadzu HPLC as described in Chapter 2, Pg 9.

3.2.1.6 Proteomic Analysis by Nano LC-MS/MS

All nano LC-MS/MS experiments were performed on an Ettan MDLC system (GE Healthcare, Piscataway, NJ) coupled to a Thermo electron linear ion trap mass spectrometer (Thermo electron, San Jose, CA) using the methods as described in Chapter 2, Pg 9.

3.2.1.7 NanoLC-LTQ-FT Analysis

An UltimMate NanoHPLC system (LC Packings-Dionex, Marlton, NJ) coupled to a LTQ-FT mass spectrometer (Thermo Electron, San Jose, CA) was used for additional
nanoLC-MS/MS analyses of samples. A 15 cm long, 75 µm ID capillary column (purchased from New Objective, Woburn, MA) was packed in house with 5 µm, 200 Å pore size Magic C$_{18}$ stationary phase was used for all LC-MS/MS experiments. Mobile phase A and mobile phase B were 0.1% formic acid in HPLC grade water and 0.1% formic acid in HPLC grade acetonitrile respectively. Prior to injection each sample was concentrated to 20 µL.

A linear gradient method beginning at 2% Mobile phase B, after 30 min of equilibration, which preceded to 40% B over 160 min, to 90% B after an additional 20 min and remained constant at 90% for 20 min comprised the LC separation method. Electrospray voltage was 1.8 kV. The normalized collision energy was 28% for MS/MS. The ion transfer tube temperature was 245°C. A medium resolution preview MS scan was generated after the ions were injected into the ICR cell. The Excalibur software selected the 8 most abundant ions for MS/MS analysis. While the LTQ fragmented these ions the FT performed a full high resolution MS scan. Precursor ions were excluded from subsequent fragmentation for 1 min.

### 3.2.1.8 Data Analysis

Data analysis for all runs was performed using CPAS (Computational Proteomics Analysis System). Human proteins were identified by searching tandem mass spectra against a human genomic database which was downloaded from Swiss-Prot September, 2007. For all database searches trypsin was selected as the enzyme, 2 missed cleavages were allowed. Carbamidomethylation of cysteine was included in the search parameters. Tolerances were set at ± 1.4 Daltons for precursor ion mass and ± 1.0 Daltons for product.
ion mass. Peptide Prophet Software was used to filter the results to a minimum probability of 95%. Only proteins identified with 2 or more unique peptides were considered. Filtering criteria using a protein probability of 0.9, a charge vs. Xcorr value of 1-1.5, 2- 2.5-3-3.8, and a peptide probability 0.7 was used to keep the false positive rate below 10%.

3.2.2 Peptidome

Pooled serum (100 µL) was diluted in 350 µL of water and 100 µL of acetonitrile. Samples were allowed to incubate at room temperature for 30 min and then next spun at 14,000 g for 10 minutes to precipitate undissolved materials. The sample was, next, loaded onto a Micron centrifugal filter device having a molecular weight cut off of 10,000. Samples were spun at 1,500, for 60 min prior to being re-suspended in acetonitrile and water and spun again. A cleanup process utilizing reversed phase chromatography was employed to remove salts. This method was identical to that used for cleanup after tryptic digestion. After this step samples were concentrated down to 20 uL and 10 uL of that was injected onto the FT-MS for peptide identification.

3.2.3 Auto Antigen

3.2.3.1 Depletion Method

A Perceptive BioSystems Biocad Chromatography system (Foster City, CA) was used for depletion of IgGs. Pooled serum (50 µL) was diluted with 200 µL of loading buffer. The depletion column consisted of 1 mL of protein G resin. Loading buffer consisted of 0.1M Tris, 0.5 M sodium chloride at a pH of 7.4. Elution buffer was 0.2 M
glycine at a pH of 2.5 and an equilibration buffer was composed of 0.5 M Tris, 1 M sodium chloride at a pH of 8.0. Unretained proteins were collected for further fractionation with M-LAC. Retained immunoglobulin were washed from the columns with glycine buffer (pH 2.5) and analyzed for protein concentration by Bradford analysis.

### 3.2.3.2 SDS-PAGE

Separation of proteins within the depleted plasma was performed by SDS PAGE. A 4-12% BisTris gel (Invitrogen) was installed in an Invitrogen mini gel system. Protein from each of the four groups (3 µg) was mixed with 8 µL of NUPAGE LDS Sample buffer and 3 µL of 1 M DTT. The samples were heated at 95°C for 5 min. The entire volume of each sample was loaded into individual wells. A standard protein mixture (Novex® Sharp™ Pre-stained, Invitrogen) was added to wells 1 and 5 in the amount of 10 µL. The separation was performed at 150 volts for 1 hr. Proteins were transferred from the PAGE gel to a nitrocellulose membrane (Sigma) in an XCell-II-Blot-Module (Invitrogen) at 30 volts for 45 min. Blocking of nonspecifically binding proteins was achieved by washing the nitrocellulose membrane in a solution composed of 0.5 g of dry milk powder (Fisher scientific), Tris (0.5M), potassium chloride (60mM), sodium chloride (2.8M) and TWEEN (0.1%) at a pH of 7.4 ±0.1. After the blocking step, the nitrocellulose membrane was washed twice in TBST Tris (0.5M), potassium chloride (60mM), sodium chloride (2.8M) and TWEEN (0.1%) at a pH of 7.4 ±0.1. The IgGs collected from the fractionation step were diluted in TBST (15 mL) and incubated with the membrane overnight. Next the membrane was washed twice with TBST prior to addition of Anti-Human Peroxidase antibody produced in rabbit (SIGMA) diluted in 15
ml TBST. After 1 hour for incubation the membrane is again washed twice with TBST before being immersed in PBS [Sodium Chloride (1.37M), Potassium Chloride (0.027M), and Phosphate Buffer (0.119M)]. Pierce SuperSignal West Pico Chemiluminescent Substrate was added to produce a fluorescent signal which was detected with a FlourChem instrument controlled by FlourChem software (Alpha Innotech Corporation, San Leandro, CA).

### 3.3 Results and Discussion

#### 3.3.1 Overview of Results

Since it is commonly believed that new disease specific biomarkers will be masked by highly abundant proteins, antibody based depletion columns such as the ProteomeLab IgY LC 10 Column from Beckman Coulter used in this study have been developed at great costs of labor, time and money. In practice the confounding issues of cost, irreproducibility (see results below) and possible loss of proteins which are bound to the proteins targeted for the depletion limit studies using this method. The cost of a column containing 12 different antibodies is above 10,000 dollars per column. Furthermore as the column ages its binding capacity may decrease resulting in non reproducible results.

Both the 12 protein depletion and the 2 protein depletion resulted in similar protein identifications. Peptides originating from vitamin D binding protein, which is involved in the transport of vitamin D sterols in blood, were discovered in all of the samples which underwent 12 protein depletion and were analyzed on the FT-LTQ with the notable exception of the obese control BD sample.
LC-MS analysis of individual samples on the Thermo-Finnagan LCQ, which had undergone depletion of albumin and IgGs, produced 145 protein identifications or an average of 36 identifications per run. HUPO Xcorr criteria of 1.9, 2.5, and 3.5 for charge states +1, +2, and +3, respectively were used to select highest probability peptide identifications. Proteins identified as being differentially expressed between the obese control and the symptomatic groups are presented in Table 3.2.

Analysis of the peptidome on the FT-LTQ resulted in 112 proteins/peptides identified from 3167 peptides in the aggregate analysis of 4 LC/MS runs. Xcorr criteria of 1.9, 2.2 and 3.85 for charge states +1, +2, and +3, respectively and a Peptide Prophet probability greater than 0.7 were used to select high quality peptide identifications.

3.3.2 Proteomic Analysis

3.3.2.1 Twelve Protein Depletion

In the initial phase of the study 12 protein depletion was used to remove the most abundant proteins prior to M-LAC fractionation. These included: albumin, IgG, IgA, IgM, transferrin, fibrinogen, apolipoprotein A-I, apolipoprotein A-II, haptoglobin, α1-anti-trypsin, orosomucoid (α1-acid glycoprotein) and α-2 macroglobulin. Albumin makes up over half the concentration of protein in human serum. The remaining 11 proteins listed above can overwhelm analysis of lower concentration proteins because they make up about 90% of the remaining protein concentration after albumin is depleted. Analysis in triplicate was performed using both the LTQ and hybrid FT-LTQ instruments.
Serotransferrin is responsible for iron transport, was down regulated in samples from patients having hypoglycemia. A proteomic study of human serum and plasma comparing normal individuals to those with Type 1 diabetes performed by Dr RD Smith’s lab discovered a similar level of down regulation $^{10}$. Another protein discovered to be down regulated in gastric bypass patients who developed complications was the thyroid hormone-binding protein transthyretin. An immunohistochemistry study of the distribution of transthyretin-containing cells in islets of Langerhans in type-2 diabetic and nondiabetic individuals found significantly more transthyretin in the cells from the diabetic patients $^{11}$. This connection to diabetes is significant in that the gastric bypass patients with complications have some similarities to diabetic patients in that the metabolic system is dysfunctional, although in diabetics insulin levels are too low versus the gastric bypass patients who have elevated levels of insulin.

<table>
<thead>
<tr>
<th>Sample group</th>
<th>BD$^a$</th>
<th>UB$^b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>25</td>
<td>6</td>
</tr>
<tr>
<td>Post Gastric bypass Asymptomatic</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>Obese Control</td>
<td>0</td>
<td>106</td>
</tr>
<tr>
<td>Post Gastric bypass Symptomatic</td>
<td>5</td>
<td>11</td>
</tr>
</tbody>
</table>

Table 3.1. Shift in the Glycosylation of Vitamin D Binding Protein

- a. Spectral count of Vitamin D binding protein tryptic peptides found in FT analysis of the M-LAC bound fraction
- b. Spectral count of Vitamin D binding protein tryptic peptides found in FT analysis of the M-LAC unbound fraction

Peptides originating from vitamin D binding protein were discovered in all of the samples analyzed on the FT-LTQ with the notable exception of the obese control BD.
sample. See Table 3.1. There can be a substitution of threonine for lysine at a position of 436 in the G2 form of the protein (Swiss-Prot, 2007) which is of interest due to the fact that threonine is amenable to glycosylation whereas lysine is not. Our data supports this in that the peptide AKLPDATPKELAK, containing a lysine at residue number 436, is found in the non-bound samples. The peptide AKLPDATERELAK was not detected in any of the samples and would not be amenable to CID fragmentation because glycosylation at Thr436 disrupts the formation of b and y ions used by Sequest for identification. The two different peptides originate from trypsin digestion of the different forms of the Vitamin D binding protein that may be present in different individuals.

In the healthy control and the post gastric bypass patients who did not develop high insulin levels after surgery possess 2-fold more M-LAC bound peptides from Vitamin D binding protein than both the symptomatic and obese individuals. It may be theorized that the bound peptides originate from a glycosylated protein containing the threonine substitution.

Although there has been interest in Vitamin D binding protein in relation to the development of metabolic syndromes in humans, this study did not look at whether glycosylation had any importance but future studies planned in the lab may be able to investigate this aspect. Several scientific articles are available which show connections to diabetes susceptibility to alleles type in certain ethnic populations. Furthermore, Vitamin D deficiency is connected to obesity. In 2000 Jacobo Wortsman studied the effects of oral administration of vitamin D on obese individuals and concluded that same subjects had a vitamin D deficiency caused by decreased bioavailability of vitamin D. In addition, long-term deficiency of vitamin D has been associated with diabetes.
involved in the transport of vitamin D sterols in blood, and is important in promoting cell
differentiation \(^{18}\). Genetic polymorphisms which may result in carbohydrate
dissimilarities in Vitamin D-Binding protein are known to occur \(^{18}\). DBP and gelsolin
(also found in this fraction) had increased gene transcription in a transgenic mouse study
of vitamin D metabolism aimed at understanding type 2 diabetes \(^{19}\).

### 3.3.2.2 Depletion of 2 Abundant Proteins from Serum Samples

A total of 145 proteins were identified from all of the runs where the serum had
undergone 2 protein depletion. Proteins identified as being differently expressed between
the obese control and the symptomatic groups are presented in Table 3.2. A ratio higher
than 2.5 is shown as up regulated. One lower than 0.5 is considered down regulated.

- Leucine-rich alpha-2-glycoprotein, Angiotensinogen, N-acetylmuramoyl-L-
  alanine amidase, Alpha-1-acid glycoprotein 1 and Pregnancy zone protein were all
  observed to up regulated 4 fold or more according to spectral count calculations. Several
  of these proteins have known literature connections to diabetes or other metabolic
  syndromes. Angiotensinogen, for example, has been implicated as a biomarker for
  hypertension in a previous study performed in our laboratory \(^{20}\). Inhibition of the renin-
  angiotensin system is a target for the treatment of diabetes \(^{21}\). This system regulates
  blood pressure and is one biological connection of diabetes to the cardiovascular system
  \(^{22}\). Improvement of a patient’s ability to regulate their blood pressure can be
  accomplished through the improvement of blood circulation in skeletal muscles.
  Enhanced blood flow will aid insulin and glucose delivery to insulin-sensitive tissues. In
  fact, efforts have begun to delay the onset of diabetes by blocking the angiotensin II
  receptor using drugs such as rampril and rosiglitazone \(^{21}\).
The protein showing the greatest change, Leucine-rich alpha-2-glycoprotein, is associated with some types of cancer but not diabetes or any related condition \(^{23}\). It is an acute phase protein which is unlikely to be a disease specific protein. As these patients did display hyperplasia (rapid cell growth) of beta cells it may be a general marker of rapid cell growth. N-acetylmuramoyl-L-alanine amidase which showed a significant change in the data breaks down cell walls by hydrolyzing glycopeptides which is another important process of cell division \(^{24}\).

The reproducibility of the 2-protein depletion experiment was monitored using several methods. Bradford analysis of depleted fractions gave a total average recovery of 86\% (CV of 10\%). This is consistent with expected results and other studies \(^{26,27}\) considering losses from sample transfers. The M-LAC bound portion contained an average of 5\% of the starting protein amount due to the removal of albumin and IgG which corresponds to 20 fold enrichment in glycoproteins. After tryptic digestion, reversed phase HPLC was used to remove salts and undigested material from the samples prior to LC/MS. In this experiment the resultant CV of less than 10\% of the peptide peak area between samples indicates successful and consistent tryptic digestions. Bovine fetuin was added as an internal standard prior to depletion and enables normalization of peak area analysis between different patient samples. An average of 10 fetuin peptides were recovered for both the glycoprotein enriched (M-LAC bound) and glycoprotein depleted (M-LAC unbound) although only 2 fetuin peptides are needed for normalization. Normalization using peptides originating from bovine fetuin was discussed in detail in the previous chapter on page 18.
3.3.3 Peptidome

About 1% of the serum proteome is composed of low molecular weight proteins/peptides. Hormones, cytokines, growth factors and proteolytic fragments of high molecular weight proteins make up portions of the peptidome. Analysis of the peptidome thus provides an exciting opportunity to investigate disease specific peptides and proteases.

From 3167 peptides identified in four aggregate analyses 112 parent proteins were identified with the use of conservative X corr values of 1+, 2+, 3+= 1.9, 2.2, 3.8 and a peptide probability of 0.7 or higher for sequence assignment. Table 3.3 includes proteins identified with both the proteomic methods and the peptidomic methods thus, some potential biomarkers found in the proteomic data sets are not included. Asymptomatic patients are those who underwent gastric bypass surgery and did not develop complications. Those patients who underwent the procedure and developed hyperinsulinemia are described as being symptomatic. Data from the obese control group is presented to provide data regarding the effect of obesity on proteomic and peptidomic concentrations.
<table>
<thead>
<tr>
<th>Protein description</th>
<th>obese control</th>
<th>symptomatic post GB</th>
<th>Ratio change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leucine-rich alpha-2-glycoprotein</td>
<td>2</td>
<td>15</td>
<td>7.5</td>
</tr>
<tr>
<td>Angiotensinogen</td>
<td>1</td>
<td>5</td>
<td>5.0</td>
</tr>
<tr>
<td>N-acetylalumamoyl-L-alanine amidase</td>
<td>2</td>
<td>9</td>
<td>4.5</td>
</tr>
<tr>
<td>Alpha-1-acid glycoprotein 1</td>
<td>4</td>
<td>16</td>
<td>4.0</td>
</tr>
<tr>
<td>Pregnancy zone protein</td>
<td>2</td>
<td>8</td>
<td>4.0</td>
</tr>
<tr>
<td>Apolipoprotein B-100</td>
<td>2</td>
<td>7</td>
<td>3.5</td>
</tr>
<tr>
<td>Afamin</td>
<td>6</td>
<td>20</td>
<td>3.3</td>
</tr>
<tr>
<td>Alpha-2-HS-glycoprotein</td>
<td>10</td>
<td>&gt;3.0</td>
<td></td>
</tr>
<tr>
<td>Apolipoprotein C-II</td>
<td>3</td>
<td>&gt;3.0</td>
<td></td>
</tr>
<tr>
<td>Complement C7</td>
<td>4</td>
<td>&gt;3.0</td>
<td></td>
</tr>
<tr>
<td>Complement component C9</td>
<td>6</td>
<td>&gt;3.0</td>
<td></td>
</tr>
<tr>
<td>Apolipoprotein(a)</td>
<td>4</td>
<td>&gt;3.0</td>
<td></td>
</tr>
<tr>
<td>Alpha-1-acid glycoprotein 2</td>
<td>6</td>
<td>&gt;3.0</td>
<td></td>
</tr>
<tr>
<td>Corticosteroid-binding globulin</td>
<td>3</td>
<td>&gt;3.0</td>
<td></td>
</tr>
<tr>
<td>Antithrombin-III</td>
<td>6</td>
<td>18</td>
<td>3.0</td>
</tr>
<tr>
<td>Heparin cofactor 2</td>
<td>4</td>
<td>&gt;3.0</td>
<td></td>
</tr>
<tr>
<td>Alpha-2-HS-glycoprotein</td>
<td>13</td>
<td>36</td>
<td>2.8</td>
</tr>
<tr>
<td>Ceruloplasmin</td>
<td>23</td>
<td>63</td>
<td>2.7</td>
</tr>
<tr>
<td>Inter-alpha-trypsin inhibitor heavy chain H4</td>
<td>23</td>
<td>60</td>
<td>2.6</td>
</tr>
<tr>
<td>Zinc-alpha-2-glycoprotein</td>
<td>2</td>
<td>5</td>
<td>2.5</td>
</tr>
<tr>
<td>Hemoglobin subunit alpha</td>
<td>4</td>
<td>10</td>
<td>2.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Protein description</th>
<th>obese control</th>
<th>symptomatic post GB</th>
<th>disease/control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complement C1q subcomponent subunit B</td>
<td>2</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>Fetuin-B</td>
<td>4</td>
<td>1</td>
<td>0.3</td>
</tr>
<tr>
<td>Fibrinogen alpha chain</td>
<td>48</td>
<td>&lt;0.0</td>
<td></td>
</tr>
<tr>
<td>Fibrinogen beta chain</td>
<td>23</td>
<td>&lt;0.0</td>
<td></td>
</tr>
<tr>
<td>Fibrinogen gamma chain</td>
<td>33</td>
<td>&lt;0.0</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.2. Proteins Up or Down Regulated in Symptomatic Post Gastric by Pass Compared to Obese Control.

a. Based on spectral counts. Only those with a ratio of change two fold or greater are included.

b. Pools of nine individual samples.

c. The ratio was set to >3 for singletons where the protein was not identified in the control. It was set to <0.0 if it was not identified in the symptomatic group.
Figure 3.2 displays plots of protein spectral counts selected due to their connection to obesity or metabolic disorders which have been normalized to the normal control. The importance of these figures is discussed in the following text.

Endogenous digestion of proteins (proteolysis) is a natural occurrence by which an organism deactivates and breaks down proteins generally through the action of enzymes. A high concentration of enzyme may degrade protein levels to near zero with little reflection on the amount of protein synthesized. Overproduction or under production of an enzyme may be linked to a disease state such as in the case of RAS (rennin angiotensin system) where digestion of angiotensin by angiotensin-converting enzymes (ACE) has been studied in connection with the vascular effects seen in diabetic patients\textsuperscript{27}.

Several of the proteins which display changes, can be correlated with other observations of metabolic conditions. Apolipoproteins A-1, A-IV and C-III are known to be associated with lipid transport and type 2 diabetes, and all show reduced enzymatic cleavage in the peptidome analysis. Apo A-IV is produced by the liver and intestine in response to the ingestion of fat and is found to increase in the blood of murine models of diabetes\textsuperscript{28}. These proteins are associated with one another on a gene cluster that is linked with triglycerides and high density lipoprotein cholesterol synthesis or transport\textsuperscript{29}. This conclusion of genetically linked up regulation in disease is supported by the presence of these proteins in the proteomic data. Apolipoprotein A-1 shows significant up regulation (>2 fold) in the protein analysis, whereas the other proteins showed slight increases.
Alpha-1-antitrypsin, alpha-2-macroglobulin, and inter-alpha-trypsin inhibitor H1 all share a trend of reduced peptide count in the peptidome results versus the proteomic results indicating reduced proteolytic cleavage. While the meaning of these results is not clear it can be noted that these proteins are acute phase proteins that are associated with inflammation. Alpha-1-antitrypsin has some literature connection with type 1 diabetes and may increase 4-fold during acute phase periods when comparing the function of healthy and diseased peritoneal membranes.  

Peptidomic analysis revealed increased endogenous digestion of ceruloplasmin, a protein which has been connected to genetic diseases that causes diabetes mellitus, in the symptomatic patients in comparison to obese and asymptomatic patients. No change in ceruloplasmin abundance was detected in the proteomic results possibly due to the fact that endogenously cleaved peptides might not be detected in database search for tryptic peptides as was performed during the CPAS data analysis.

Spectral count data of the LC/MS analyses are presented in Table 3.3 to permit investigation into changes in protein abundance when the peptidome results are compared to M-LAC bound proteomic results. Only results from gastric bypass patients with complications and results from healthy normal individuals are presented to focus on those samples most likely to exhibit protein abundance changes. Thus this table is unable to answer questions in regards to whether a change is related to obesity of the patients or the surgical procedure (gastric by-pass) they underwent. Changes of two-fold or more between symptomatic results and the corresponding normal results (i.e. peptidomic symptomatic are compared to petidomic control) are considered to be significant.
Of particular interest are those proteins which display opposing trends when comparing peptidome results to those of the M-LAC bound. Several proteins display this trend including several apolipoproteins, complement proteins, inter-alpha-trypsin and alpha 1 anti trypsin. These proteins are highlighted and the abundance changes indicated with arrows. A trend in the peptidomic results of reduced proteolytic cleavage will result in a higher abundance of that protein which may in turn be detected in the proteomic results. These results can indicate a reduction in an enzyme or the production of a protein which is resistant to degradation. Complement C4-A and complement C4-B are proteins formed through the cleavage of a common protein (complement C4) which are linked to susceptibility to autoimmune diseases such as SLE and genetic diseases such as congenital adrenal hyperplasia (CAH)\textsuperscript{32}. 
Figure 3.2. Comparison of Proteomic Results after 2 Protein Depletion to Those of the Peptidome. All results are based on spectral counts that have been normalized with the normal control of either the peptidome results or proteome results.

A. Apolipoproteins and endogenous peptides from apolipoproteins are shown together to exhibit the reduced presence of endogenous peptides relative to peptides produced by tryptic digestion.

B. Protease Inhibitors proteins and endogenous peptides are also shown together and exhibit a similar trend.

C. Complement factor proteins and endogenous peptides did not display the same trend.
Table 3.3. Composite Table of Peptidome<sup>a</sup> and 2 Protein Depletion Proteomic<sup>b</sup> Results

<table>
<thead>
<tr>
<th>Name of Protein</th>
<th>PEPTIDOME&lt;sup&gt;a&lt;/sup&gt;</th>
<th>M-LAC Bound&lt;sup&gt;bc&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Symptomatic</td>
<td>Control</td>
</tr>
<tr>
<td>Alpha-1-antitrypsin</td>
<td>3&lt;sup&gt;a&lt;/sup&gt;</td>
<td>11</td>
</tr>
<tr>
<td>Alpha-1B-glycoprotein</td>
<td>ND&lt;sup&gt;f&lt;/sup&gt;</td>
<td>3</td>
</tr>
<tr>
<td>Alpha-2-HS-glycoprotein</td>
<td>6</td>
<td>16</td>
</tr>
<tr>
<td>Alpha-2-macroglobulin</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>Apolipoprotein A-I</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>Apolipoprotein A-IV</td>
<td>2</td>
<td>23</td>
</tr>
<tr>
<td>Apolipoprotein B-100</td>
<td>ND</td>
<td>5</td>
</tr>
<tr>
<td>Apolipoprotein C-I</td>
<td>ND</td>
<td>1</td>
</tr>
<tr>
<td>Apolipoprotein C-III</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>Apolipoprotein E</td>
<td>ND</td>
<td>4</td>
</tr>
<tr>
<td>Ceruloplasmin</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>Coagulation factor XIII A chain</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>Complement C1s subcomponent</td>
<td>ND</td>
<td>2</td>
</tr>
<tr>
<td>Complement C3</td>
<td>17</td>
<td>22</td>
</tr>
<tr>
<td>Complement C4-A</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>Complement C4-B</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>Complement C5</td>
<td>ND</td>
<td>6</td>
</tr>
<tr>
<td>Complement factor B</td>
<td>ND</td>
<td>1</td>
</tr>
<tr>
<td>Fibrinogen alpha chain</td>
<td>292</td>
<td>315</td>
</tr>
<tr>
<td>Fibrinogen beta chain</td>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td>Fibrinogen gamma chain</td>
<td>ND</td>
<td>9</td>
</tr>
<tr>
<td>Gelsolin</td>
<td>ND</td>
<td>8</td>
</tr>
<tr>
<td>Hemopexin</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>Inter-alpha-trypsin inhibitor heavy chain H1</td>
<td>ND</td>
<td>7</td>
</tr>
<tr>
<td>Inter-alpha-trypsin inhibitor heavy chain H2</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>Inter-alpha-trypsin inhibitor heavy chain H4</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>Leucine-rich alpha-2-glycoprotein</td>
<td>ND</td>
<td>2</td>
</tr>
<tr>
<td>Platelet basic protein</td>
<td>ND</td>
<td>7</td>
</tr>
<tr>
<td>Plexin domain-containing protein 2</td>
<td>ND</td>
<td>1</td>
</tr>
<tr>
<td>Prothrombin</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td>Serum albumin</td>
<td>ND</td>
<td>1</td>
</tr>
<tr>
<td>Tissue factor</td>
<td>ND</td>
<td>4</td>
</tr>
<tr>
<td>Transthyretin</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Vitronectin</td>
<td>ND</td>
<td>2</td>
</tr>
</tbody>
</table>

<sup>a</sup> Single analysis of endogenous serum peptides performed with hybrid LTQ-FT-MS.

<sup>b</sup> Single proteomic analysis performed on LCQ MS.

<sup>c</sup> LC-MS analysis of glycoprotein enriched fraction collected from M-LAC gravity column.

<sup>d</sup> Analysis of serum from normal control

<sup>e</sup> Spectral count

<sup>f</sup> ND- Not detected
3.3.4 Autoimmune Analysis of Patient Serum

Western blot analysis of patient serum for the presence of auto antibodies resulted in detection of antigenic protein bands (See fig. 3.4). These encouraging results indicate the feasibility of screening a patient’s serum proteins with immunoglobulins captured from the patient’s serum and support the possibility of an autoimmune cause of complications after gastric bypass. Western blotting is composed of an initial step of protein separation generally with an SDS-PAGE, electrophoretic transfer to a membrane and a final detection step where the separated proteins are probed with antibodies. In the western blot experiments, as performed for this analysis, the detection of autoimmune reactive proteins was accomplished by incubating the membrane, after electrophoretic transfer of the proteins, in the antibodies separated from the serum of symptomatic patients and then with a horseradish peroxidase anti-human IgG conjugate. A false negative result may occur if there is no detection of proteins in the symptomatic sample due to low concentrations, a low signal-to-noise ratio, loss or blockage of the epitope region of the antibody or if an auto-antibody is not captured by the protein G column. Due to the fact that the detection step includes the use of a mouse anti-human antibody, false positive results can occur if not all of the antibodies are removed from the serum. False positives may also occur in the sense that auto antigens related to other diseases not related to hyperinsulinemia may be detected.

The IgG depleted plasma proteins from both normal and symptomatic patients were separated on an SDS gel and then transferred to cellulose paper prior to being incubated with immunoglobulins captured from the depletion of the plasma from the symptomatic patients. Thus any autoantigens and undepleted IgGs were separated and
Nonspecific binding occurs when the antibodies bind to molecules other than their specific antigen including the membrane and proteins transferred to it from the SDS-PAGE gel. While the design of this experiment has the potential for nonspecific binding to occur it was not present in the result except perhaps as background noise. Nonspecific binding to the membrane of either the human self antibodies or the horseradish peroxidase anti-human IgG conjugate will cause background illumination and interfere with detection of the autoantigens. The strong bands having the shortest migration distance are predicted to be undepleted immunoglobulin due to the fact that they are present in each sample lane and due to that fact that protein G does not deplete all immunoglobulins.

As can be seen in Figure 3.3 there are differences in the protein bands separated from the control sample to the gastric bypass with complications sample and the obese sample to the gastric bypass with complications sample. A protein band visible at in the disease lane is not present in the normal samples. These preliminary results indicate that there is possibly an autoimmune response or cause of hyperinsulinaemia in gastric bypass patients. Analysis of human serum may result in the identification of autoantibodies for use as diagnostic markers in specific autoimmune diseases and in the identification of individual proteins or patterns of protein expression that are deregulated in autoimmune diseases.
3.3 Results of the Autoimmune Western Blot Experiments. IgG depleted serum from symptomatic gastric bypass patients and from control patients who had not undergone gastric bypass were first separated on an SDS-PAGE gel. Proteins were transferred to nitrocellulose and incubated with the IgGs from the disease fraction. Detection was facilitated by incubation with a horseradish peroxidase anti-human IgG conjugate. The bands indicated by the uppermost arrows (A) are predicted to be undepleted IgG. The band indicated by the lower arrow (B) is found only in the disease analysis.

3.4 Conclusion

While many patients who choose to undergo gastric bypass surgery recover successfully and enjoy improvements in their health such as abatement of type 2 diabetes other patients develop life-threatening complications in the form of abnormally high insulin levels. Hypoglycemia, low blood sugar, caused by the increase in insulin can result in unconsciousness, seizures, and brain damage. Thus in an effort to discover the
cause of the abnormal insulin levels we collaborated with Joslin Diabetes Center in order to study serum collected from affected individuals in comparison with serum from unaffected individuals.

Our methods included depletion of abundant serum proteins prior to LC/MS analysis to reduce the dynamic range of the serum samples. Proteomic investigation discovered a possible genetic cause of the symptoms related to vitamin D binding protein exposed through the observation that symptomatic patients tended to have less vitamin D binding protein which was retained by the M-LAC column than healthy controls. The vitamin D binding results would need thorough investigation to be substantiated and might only be present in some individuals (pooled samples were tested) but such investigation is warranted due to the fact that vitamin D is correlated strongly with obesity and metabolic disorders in literature searches.

Analysis of endogenous peptides revealed reduced spectral counts in the peptidome results versus spectral counts for the same protein in the proteomic results indicating reduced proteolytic cleavage. The reduced cleavage may be due to decreased amount of proteolytic enzymes or due to the presence of proteins resistant to cleavage. A detailed investigation into enzymes which digest the proteins discussed in section 3.3.3 and their relation to the endocrine system would prove interesting as would an investigation into amino acid sequence variations which might render a protein resistant to digestion.

The autoimmune screening experiment was able to detect the presence of serum proteins which elicited an autoimmune response. The endocrine system is a common target of autoimmune disorders and an autoimmune form of diabetes (type 1)
destroys the beta-cells of the pancreas which are responsible for producing insulin.

Additional experiments to confirm this finding and to identify the protein eliciting the immune response have the potential to reveal a new autoimmune condition.


3.5 References


Chapter 4 New Developments in Bioinformatics using Computational Methods

Acknowledgements

The work in this chapter is a result of collaboration with several people, and I would like to extend my appreciation to all of them, and especially Dr. Akella and Dr. Rejtar for their efforts to automate the pattern recognition processes. I would also, like to thank Dr. Hincapie and Dr. Hancock for their support. The final portion of this chapter has been possible due to the efforts of the people at Proxeon, especially Dr. Christian Ingrell. Madduma Dayarathna performed the analysis of the diabetic samples that were used in part of the pattern recognition work.

4.1 Elimination of the Post Identification Proteomic Data Bottleneck using Bioinformatic Tools

A proteomic experiment, such as the mouse proteomic study described in the first chapter, performed using a well developed LC-MS platform can potentially result in vast amounts of fragmentation spectra in need of interpretation. Interpretation of spectra by an individual rather than a computer is hampered by many factors including time, skill and bias issues.

Bioworks (Thermo Scientific) and Peptide Prophet (Proteome Software Inc.) both use the SEQUEST algorithm but differ in their means of separating high-quality matches from the overall data set. Quality of the matched peptides can be judged by how close the molecular weight of the theoretical peptide matches that of the actual precursor ions (Mascot) or matches may be compared to multiple theoretical MS/MS spectra, and these matches compared to find which is the best match.

Finding correct protein identifications is one step toward the ultimate goal of discovering a biomarker of disease. The next step includes a comparison of the types and amounts of proteins present in a disease sample versus a normal control. There are various methods in use to convert a list of proteins identified in an LC/MS experiment to a list of potential biomarkers. One may in certain types of experiments label the sample sets by covalently attaching isobaric tags to peptides thereby easing identification in mass spectral data. Labeling, such as ICAT, was not used in the mouse plasma study for several reasons including not only costs and time but the possibility of low abundance proteins being lost during sample transfers due to noncovalent binding to container surfaces or through side reactions during the derivitization. Without mass tags the process
of differential quantitation relies on the use of peak area quantitation spectral counts which is the number of peptides identified from a protein \(^1\). Spectral counting is relatively quick and effective but can suffer from high levels of deviation in the analysis of lower abundance proteins in complex samples such as plasma due the fact that not all peptides will be selected for fragmentation due to time constraints in a LC/MS analysis. Without fragmentation data the peptide, although it was present in the sample analysis and probably ionized, will not be identified or used in spectral count analysis. Another problem with the use of spectral counts may be that larger proteins may produce a larger number of peptides than a smaller protein causing a larger protein to appear more abundant. This can be overcome by using a factor based on the number tryptic peptides divided by the length of the protein.

Interpretation of the proteomic data can be troublesome due to the fact that there may be many proteins which display changes in abundance in relation to non-disease specific processes such as inflammation. The mouse plasma study revealed that even proteins known to be highly associated with tumor growth had not originated from the tumor and thus the origin of a given plasma protein is uncertain.

It may also be feasible to look at data sets generated by proteomic experiments with a wider view such as that utilized by gene ontology which attempts to assess the data set as a whole and determine which cellular processes are up regulated during disease. Going a step further we believe it is possible to compare whole data sets against one another and to build up knowledge, without delving into individual identifications of proteins, to identify features of the data which are associated with disease or health.
The large data set generated by the mouse model presented us with the possibility of comparing chromatographic LC-MS data in addition to the peptide identification step while comparing tumor LC-MS analysis to non-tumor analysis. Due to the fact that any widespread use of this type of analysis must be automated collaboration was established with Proxeon.

Another area which has developed into a bottle-neck is the process of utilizing protein identifications. As research into proteomics continues to identify proteins and their associations protein databases will grow and uncover patterns of gene expression. Relevant information and annotations for protein identifications can be gained from online databases including Swiss-Prot, Human Protein Reference Database (HPRD), and Entrez Gene. One of the next steps of progression for proteomics is to be able to integrate protein and genomic data to enable studies of systems biology.

Improvements to database search accuracy, reproducibility and especially speed are imminent. Not only does Software development continue to bring new progress for data interpretation but processor speed and internet tools advance continually bringing new options to how searching can be performed. Compilation of published and stored proteomic datasets combined with ever increasing amounts of data being streamed over the internet promises to build bridges amongst research groups and fields.

4.1.1 Programming Aimed at Improving the Speed of Interpreting Biological Meaning in a Clinical Proteomics Study

The current development of mass spectrometry based proteomics applications has entered into the phase of enormous data collection like genome sequencing projects in the
last decades of the 20th century. That stage is characterized by the very well developed
technology of protein analysis, robust instrumentation with advanced software tools to
acquire and analyze samples as well as support by bioinformatics applications. One of
the obstacles to fast interpretation of proteomic data is mass spectrometric data
processing based on efficient spectra analysis and database searches followed by data
validation. Secondly, there is a need for improved bioinformatical analysis of retrieved,
validated data. Bioinformatic analysis and datasets comparison is a time consuming and
imprecise process of establishing what is known or novel in these data sets without
proper annotation. Also comparison of list of protein identifications can be a complex
process especially if the comparison is performed between different species or protein
curations obtained in different laboratories at different times or coming from literature
sources. Sophisticated tools are needed that can handle the complexity of these data
including: redundancy (same protein but different accession, or alleles & fragments),
different types of accession codes or outdated accession codes.

There are numerous software tools which have potential use during the
interpretation of proteomic data4, 5. Modern proteomics must face the challenge of
performing bioinformatics analysis and comparison of large datasets.
Distinguishing novel proteins from known in these data sets without proper annotation of
what is the correct protein name and function is a time-consuming and sometimes almost
impossible task. Tools are needed that can handle the complexity of this data, including:
redundancy (same protein but different accession codes or alleles and fragments),
different protein database accession codes or outdated accession codes. To resolve
these issues we have used ProteinCenter a tool that enables efficient data mining and
categorizing of large data sets in a collaboration with Proxeon. The ProteinCenter, database is built using the Sun Java technology and the Microsoft mySQL database technology for optimal performance. ProteinCenter contains a compilation of public sequence databases to form a comprehensive and consistent superset of 10 million protein sequences derived from over 40 million protein records from GenBank, Refseq, EMBL, UniProt, Swiss-Prot, Trembl, PIR, IPI, PDB, Ensembl. Proteins are richly annotated by consolidated annotation from public databases together with high standards annotation from internal computational enrichment of the sequence data. The integrated database is updated regularly depending on its source, enabling tracking of outdated accession keys. All the information in ProteinCenter can be accessed by a web user interface which allows for import and comparison of database dependent search results. Also complex queries and statistical calculation based on the information kept in the database can be performed on the imported search results.

4.1.2 Maximizing the Amount of Information Gained from Proteomic Data

LC/MS clinical projects generate thousands of peptide and protein identifications and thus there are issues with how the data can best be utilized. Generally, most raw proteomic data is only seen by a handful of people. A select portion may be published for all to see, but even then a publication based on such as the AZ study reported here, may only garner interest within the mass spectrometry or separation science communities. Organizations such as Human Proteome Organization (HUPO) are valuable intermediates who aid in the general acceptance and dispersal of proteomic data. HUPO has driven quality standards for proteomic data such as setting the values of cross correlation scores
versus charge state. Without such standards it is impossible to evaluate the quality of data or make comparisons from one data set to another. With its organ based initiatives, including the liver and brain proteome, the organization is able to initiate studies along strict guidelines in multiple laboratories thus generating high quality, organ specific data which can be offered to the medical research community for consideration. For illustration examples will be given.

In order to aid understanding of the roles of proteins in healthy and diseased human liver tissue the Human Liver Proteome Project (HLPP) established SOPs for liver sample collection, pretreatment and analysis. Normal liver tissue was collected from around 100 healthy volunteers. These liver samples were pooled and distributed to 8 laboratories throughout the world for protein expression profiling. Another 78 liver samples were collected in China. These studies had resulted in the identification of 2053 unique proteins with two or more unique peptides by 2006. 

The HUPO brain study proceeded with a set of reference samples which were shipped to nine international laboratories to be analyzed for protein content with different approaches. A portion of the study analyzed murine brain tissue from embryonic, postnatal and 8-week time points using 2-D DIGE using internal standardization and overlapping pH gradients (pH 4-7 and 6-9). 214 protein spots showing developmental stage-dependent intensity differences (> two-fold) were detected, 56 of which were identified. Human postmortem and postnatal murine brain tissue were analyzed with a 2-D LC-MS/MS method and resulted in the identification of 350 human and 481 mouse proteins could be identified by at least two different peptides. Hamacher
acknowledged that the heterogeneity contributed by the different analysis strategies and data formats, presented a real challenge when comparing results.

4.1.3 Current Methods of Data Handling

Currently most of the peptide and protein identifications generated from a LC/MS experiment are reviewed by moderately to highly skilled researchers. Out of the possibly, thousands of proteins those selected for further investigation will be those proteins which show abundance changes or are of interest to the investigators. Recognizing the importance of new or unknown proteins may be nearly impossible when relying on literature searching or on Swiss Prot annotations ⁹.

As corporations and academic research laboratories have embarked on proteomic experiments each has encountered the enormous amount of time required to proceed from sample analysis to complete data interpretation and each has attempted to come up with solutions to speed up the process. From automated matching programs such as SEQUEST to protein interpretation software such as Protein Center, numerous computer programs have been developed to aid the quality and speed of data interpretation at every conceivable point of a proteomic experiment.

4.2 Overview of Mouse Breast Cancer Study as an Example of a Large Biomarker Study

The large data set produced during the mouse xenograph study was analyzed using ProteinCenter. The design of the study is reviewed in Figure 2.1 (page 83) and Figure 4.1. The M-LAC and LC/MS methods are detailed in the second chapter.
Figure 4.1. Experimental Design for the Study of the Ncr Nude Mouse Xenograph as a Model of Human Breast Cancer.

4.2.1 A Summary of the Treatment Groups Described in Chapter 2

Initially 56 ovarectimized mice were split into two groups: non tumor and tumor. A MCF-7 tumor was implanted subcutaneously in the mammary fat pad of those mice belonging to the tumor group while the control group was not implanted (see Figure 4.1). Each of the two groups was further split into 4 treatment groups: no treatment, estrogen, Tamoxifen or both estrogen and tamoxifen. Estrogen and/or tamoxifen pellets were surgically implanted in mice of the estrogen groups and the estrogen plus tamoxifen groups. For the individual time points plasma from 5 individual mice was pooled for each analysis. Plasma was collected at the beginning of the study, 3 weeks into the study and at 6 weeks.
Due to the fact that MCF-7 cell lines are estrogen dependent for growth the control group showed little growth of the tumor, as did the group receiving only Tamoxifen. The Tamoxifen group serves as a control for the effects of Tamoxifen and the lack of tumor growth confirms the tumorstatic properties of this drug. Those animals who received estrogen showed the greatest tumor growth and had the greatest potential of having observable protein changes in their plasma. The MCF-7 tumors of group receiving both estrogen and Tamoxifen had intermediate tumor growth and were best suited for discovery of proteins impacted by the effect of Tamoxifen on a growing tumor.

4.2.2 ProteinCenter as a Tool to Interpret Complex Proteomic Data

ProteinCenter offers several computational tools including Geneontology (GO) analysis which can give an immediate answer regarding quality of the data. For example plasma samples such as the mouse plasma samples resulted in identification of a high number of proteins which were known to be secreted when compared to the whole mouse genome. ProteinCenter presents these comparisons in visual formats which are readily interpreted and understood.

The core of ProteinCenter is a relational database representing protein information in a complex domain model. This database is updated periodically with information from a central server at Proxeon Bioinformatics, which in turn collects and integrates data from a long list of external sources into the domain model. Among the external sources included are NRDB, IPI, UniProt, Ensembl, TAIR, SGD, Flybase, Entrez Gene, Interpro, IntAct, PFAM, MIPS Interaction, Gramene and Gene Ontology. Once the data has been integrated, sequences of Computational Enrichments™ are performed, including protein-
protein similarity, signal peptide and transmembrane domain predictions. The data is accessed from a standard web browser through an application server, which presents a variety of views of the data while at the same time permitting efficient, on-the-fly computation of, e.g., protein and peptide clustering. We will demonstrate the value of this relational database in the analysis of a complex dataset.

4.3 Results

4.3.1 Summary of Mouse Result

By organizing protein identification data under column headings including protein name, gene name, taxonomy and GO annotations ProteinCenter can increase the speed at which data can be analyzed for biological meaning. Two representative analyses, 6-week with tumor and estrogen and 6-week without tumor were selected as having the greatest potential to reveal proteins which had increased or decreased abundance in response to tumor growth and demonstrate the potential of ProteinCenter. In Figure 4.2 in a Venn diagram in displays the number of proteins which are identified in both groups and those which are identified separately. As illustrated in the figure, the majority of the proteins (91) were identified in both analyses. The taxonomy column indicates the genus to which the protein is associated in this case it indicates that the proteins listed are of mouse origin as would be expected due to the fact that this data is the result of searching fragmentation spectra data against a mouse database. In this case the actual taxonomy of the proteins may be human due to the fact that the implanted tumor was human. Searching the fragmentation data against a human database would result in similar protein identifications due to the high degree of homology between humans and mice.
Displayed in the right half of Figures 4.3 and 4.4 are Gene Ontology (GO) results for each protein. GO annotations, such as the output from ProteinCenter, consist of a set of controlled vocabularies encompassing language with a universal meaning for the annotation of molecular characteristics across organisms, and are used to describe the biology of a data set. Of particular interest in Figure 4.3 are proteins which have associations to cell growth such as protease serine 2 (trypsin 2) which was found only in the tumor estrogen data set (Serine proteases are discussed in Section 4.3.5), hepatocyte growth factor and complement factor B. Altered expression of cell organization proteins such as lumican and cyclin may indicate the proliferation of tumor cells. Orosomucoid 2, another protein which can be seen in Figure 4.3, which is involved in protein binding and is also known as alpha1-acid glycoprotein (AGP), is found to be increased in infection, inflammation and cancer. Displayed in Figure 4.4 is a partial list of proteins found in both the tumor bearing animals and the nontumor bearing animals. Of particular interest in this figure is epidermal growth factor receptor (EGFR), which has established connections to breast cancer tumor progression and serine protease inhibitors (Serpins), which are associated with the breakdown of the extracellular matrix. ProteinCenter provides a column for gene name which facilitates the recognition of proteins belonging to the same gene family. While only a small part of the data set is presented here, ProteinCenter can be used to interrogate large data sets and facilitate fast interpretation of biological meaning of individual proteins and entire data sets. Both EGFR and Serpins (serine protease inhibitors) are discussed below in further detail.
Figure 4.2. A Venn Diagram of the Proteins Identified in Single Analyses of the 6-Week Estrogen With Tumor Group and Estrogen Without Tumor Group.
**Figure 4.3. Proteins Identified in Estrogen Tumor (6 Week) Only**
<table>
<thead>
<tr>
<th>Description</th>
<th>S</th>
<th>Data set</th>
<th>Cluster</th>
<th>Gene</th>
<th>AA</th>
<th>AS</th>
<th>Fr</th>
<th>Tax</th>
<th>Molecular Functions</th>
<th>Cellular Components</th>
<th>Biological Processes</th>
<th>TM</th>
<th>SP</th>
<th>Pep</th>
<th>Cell growth</th>
<th>Receptor activity</th>
<th>DNA binding</th>
<th>Enzyme regulator activity</th>
<th>Protein binding</th>
<th>Extracellular</th>
<th>Membrane</th>
<th>Metabolic process</th>
</tr>
</thead>
<tbody>
<tr>
<td>immunoglobulin 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>serum amyloid A 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>leukemia inhibitory factor rec...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>serine (or cysteine) proteinas...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Serum amyloid P-component prec...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>alpha-2-HS-glycoprotein</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coagulation factor X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complement C3 precursor (90E-M...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ig heavy chain V-III region A4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>apoA protein H precursor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ceruloplasmin isoform b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>haptoglobin</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ig heavy chain V region F4 pr.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hemoglobin subunit beta-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mannose-binding lectin (C)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ig kappa chain V-III region T1 p.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ig kappa chain V-IV region XIB...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>serine (or cysteine) proteinas...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>vitaminin</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mannose-binding lectin (A)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>carboxypeptidase N, polypeptid...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S-phosphatidylinositol-1,4,5-tr...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hepatocyte growth factor activ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>serine (or cysteine) proteinas...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inter-alpha-trypsin inhibitor ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>of angiogenin</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>epidermal growth factor receptor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complement Factor I precursor ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complement Factor H precursor ...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>transferrin</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>complement factor B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ig kappa chain V-II region 25...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ig kappa chain V-III region P2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>transcription elongation factor...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ig kappa chain V, C region secreted</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>serine (or cysteine) proteinas...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ig heavy chain V region NQPC 4...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.3.2 GO Terms Comparison

A comparison of the cellular component GO terms generated from the mouse study revealed the prevalence of extra cellular proteins such as would be expected in blood (Figure 4.5). Extra cellular proteins which include those proteins which are involved in intra-cell signaling, cell growth, and extra cellular matrix, are more likely to be found in plasma in high amounts versus cellular proteins. In the case of cancer there is the potential to find extracellular protein biomarkers which indicate the break down of the extracellular matrix in response to tumor growth.\(^{13}\) Cell signaling proteins, in particular, travel in blood to recruit blood vessels from the host to support tumor growth.\(^{14}\)

ProteinCenter analysis of biological process GO terms which are associated with the nude mouse versus the wild mouse revealed increases in the GO terms associated
with inflammatory and immune responses. This is an expected result due to the fact that all of the animals had been ovariectomized and injected with an estrogen pellet. Animals which also received an implant of MCF-7 tumor cells would be expected to show an increase of the number of proteins with inflammatory and immune GO terms due to the growth of the tumor\textsuperscript{15}.

A comparison of the mouse model to normal mice with a focus on molecular function revealed increased representation in several categories including enzyme regulation, protease inhibitor activity and transporter activity. There are numerous reports of increased enzyme activity in cancer including kinases which are required for a tumor’s high level of cell division. Transporter proteins are also necessary in elevated amounts to maintain a rapid level tumor growth. Transporter proteins are responsible for the movement of nutrients across cell membranes and also have the ability to remove from the cell anti-tumor drugs\textsuperscript{16}. 
4.3.3 EGFR

ProteinCenter includes a feature called Protein Card which aims to decrease the amount of time spent on literature searching by providing basic information about each protein and clickable links to external sites such as NCBI. Figure 4.6 shows the Protein Card view from Protein Center for EGFR where the structural features such as the fact...
that EGFR possesses a transmembrane domain, are presented. Also featured in this view are GO annotations, associated key words and functions. Gene ontology annotations are part of an effort to make gene products (i.e. proteins) searchable by computer software by using consistent definitions. Gene products are assigned terms which are from one of ontologies either molecular function, cellular component or biological process.

GO annotations identify the molecular functions of EGFR to include DNA binding, enzyme regulator activity and receptor protein tyrosine kinase. Cellular locations associated with EGFR include nuclear and membrane locations. Cell motility and cell proliferation are important GO annotated biological processes attributed to EGFR. The Protein Card feature of ProteinCenter aided in identifying EGFR’s interactions with other proteins and in producing Figure 4.8(pg 183) where EGFR is illustrated to show it’s interaction with DNA-PK and BRCA-1.

4.3.4 Analysis of the PFAM Domains in the Nude Mouse

Another method of examining the types of proteins identified in the mouse model is to study the PFAM (protein families) domains represented in the nude mouse versus those found in the wild type mouse. PFAM is based on a database of protein sequences which are aligned to reveal conserved sequences across proteins and species. Aligned sequences are then used to determine common ancestry and more importantly for proteomic studies, common function. Those domains which were observed to be more common in nude mice versus wild mice were the Serpin (serine protease inhibitors), trypsin (enzymes that exploit serine in their catalytic activity are ubiquitous) and sushi
Sushi domains are also known as Complement control protein [CCP] modules, or short consensus repeats [SCR], exist in a wide variety of complement and adhesion proteins.

The Proxeon program allowed us to rapidly determine reduced presence of serine protease inhibitor (serpin) proteins, such as alpha-1-trypsin inhibitor, in the tumor bearing mice versus non-tumor bearing mice (Figure 4.7). Serpins are involved in the regulation of plasmin which plays an important role in several biological pathways such as the plasminogen activator system, which is crucial to the process of breaking down the
extracellular matrix to allow tumor growth\textsuperscript{18,19}. An observation of down regulated alpha-1-trypsin inhibitor can be correlated to an up regulation of the plasminogen activator system and the growth of the tumor. This critical pathway in the mouse is being influenced by the MCF-7 human tumor cells as this effect is not observed in the animals with no tumor xenograph. Serpins are a group of glycosylated and secreted proteins which can control the ability of serine proteases to perform their functions and which have been implicated in cancer progression. The plasminogen activator system is regulated in part by alpha-1-trypsin inhibitor and alpha-2-antiplasmin which are capable of inhibiting plasmin by forming stable complexes\textsuperscript{20,21}. Alpha-1-antitrypsin is an acute phase protein which can be produced by tumor cells. In a study of the effect of AAT on cultured MDA-MB 48 breast cancer tumor cells AAT inhibited growth by 61\%\textsuperscript{20}. Alpha-2-antiplasmin, which displayed modest down regulation in the disease analyses, is a proposed anti-tumor agent\textsuperscript{21}. The functions of the Serine protease inhibitors A3C, A3K and A3M have yet to be fully elucidated and also possess highly homologous amino acid sequences according to a BLAST search.

The plasminogen activator system is regulated by EGFR through protein kinase C delta and protein kinase alpha in glioblastoma cells\textsuperscript{22,23}. Biological pathways such as the EGFR and plasminogen activator system are often found to be interconnected and the study of the connections promises to shed new light on tumor growth and on the spread of tumor cells in an organism. Conventional search engines perform the task of protein identification but interpreting bioinformatic data such as the mouse proteomic data presented in this thesis can take months to finalize and prepare for publication. ProteinCenter is a proteomics data interpretation software package which enables
filtering, clustering and statistical bio-informatic analysis of protein identifications. By providing rapid initial classification of protein origins and molecular processes, ProteinCenter allowed us to proceed to the analysis of the biological pathways such as that illustrated in the figure and text below.

Figure 4.8. Cellular location and protein interactions of some proteins found in the mouse breast cancer model. E indicates the protein was found in the estrogen tumor group. T indicates the protein was found in the Tamoxifen tumor group. E+T indicates the proteins found in the estrogen and Tamoxifen tumor group. Downward arrows indicate a reduced amount of the protein from 0 to six weeks. Upward arrows indicate an increased amount of the protein over the time span of the experiment. A horizontal arrow indicates no change. Data used to make this figure and full protein names are given in Chapter 2, Table 2.3.
4.4 The Importance of protein pathways found in the mouse data

Studies such as the one performed on the MCF-7 mouse tumor plasma produce large numbers of protein identifications on which immense amounts of time can be spent in order to learn the purpose of these proteins in relation to normal and disease states. The complete mouse study resulted in the identification of many kinase and kinase related protein which were investigated as to their biological roles and resulted in the construction of the diagram in Figure 4.8. While this diagram is the product of many weeks of literature research, ProteinCenter can speed up the process of understanding protein associations by providing GO classification, family associations and Protein cards.

Kinase proteins, such as those arranged in Figure 4.8, have important roles in cellular activities such as cell growth and cell division. The lipid kinase phosphoinositol-4-phosphate 3-kinase (PI 3-kinase or mouse PK3CA, class II, containing the p110delta and C2 subunits) and related members represent one of the most important regulatory proteins that control key cellular functions and the resulting phosphoinositide products activate a host of signaling proteins \(^{24}\). Several pivotal studies have shown that PI 3K has an integral role in tumorigenesis via association with oncoproteins, by genetic analysis and by mouse transgenic and knockout studies \(^{25}\). Importantly, other studies have shown recruitment of PI 3-kinase with signaling molecules identified in this study: epidermal growth factor receptor \(^{26}\) (EGFR), interleukin-1 receptor associated kinase (IRAK) \(^{27}\), ephrin receptor EphA8, A3 and B2 \(^{28}\) and members of the PI 3-kinase like kinase family (PIKK members DNA-PK and ATM) \(^{29}\). Also, it has been shown that ATM heterozygotes have an increased risk of developing breast cancer \(^{30}\) and are known to be associated with
EGFR. A related protein found in this study is GPI-PLD which cleaves the GPI anchor sequence attached to many cell surface proteins including receptors\textsuperscript{31}. GPI-PLD was down regulated in the estrogen and combined treatment groups, which is consistent with reports that this protein was decreased during inflammation\textsuperscript{32} and which can be related to tumor growth. In conclusion, analysis of this proteomic data set with ProteinCenter software provided insight into additional protein targets for investigation of their importance in cancer progression.

4.4.1 A Comparison of Biological Samples Based on the entire LC/MS Data Set

In the previous section we showed that improved informatics tools such as ProteinCenter will aid in the use of proteomic data for the discovery of potential biomarkers of disease such as in the case of the mouse model where we found proteins associated with the rapid growth of the implanted tumor. In discovery efforts based on LC/MS analysis it is typically observed that as few as 20\% of the peptides observed in a precursor scan are selected for fragmentation and can produce fragments which can be used for identification\textsuperscript{33}. The LTQ ion trap mass spectrometer performs a precursor ion scan first and then seven data dependent MS/MS scans which can allow the identification of up to seven peptide sequences per segment of the chromatographic peak. In addition some peptides selected for fragmentation will not produce useable data due to post translational modifications (PTMs) such as glycosylation which produces few b or y fragments. PTMs play important roles in the function of proteins and are important in the study of many diseases including cancer and diabetes. PTMs can result in missed cleavages during the trypsin digestion process possibly due to steric interference with the
A peptide containing more than one lysine or arginine will contain more than one site for protonation which will inhibit random backbone dissociation of the peptide during collisionally induced fragmentation. In this manner traditional LC/MS analysis will only identify a small portion of the peptides present in a complex proteomic analysis.

In the effort to maximize the use of proteomic data for disease detection several researchers have investigated the possibility of utilizing raw LC/MS data without any MS/MS based protein identifications. In this process peaks in either the total ion chromatogram or in individual m/z chromatograms are analyzed for disease association. When successful, comparisons of raw data can detect changes due to disease that may be missed when comparing lists of individual proteins due to missing fragmentation data as discussed above. This data analysis method will not be able to identify which protein(s) are producing diagnostic peaks in the LC/MS chromatograms and will not give any information regarding disease pathways but has the advantage of accessing much more of the information available in a protein analysis. The solution to this problem is to conduct additional experiments aimed at identifying the source of the diagnostic peaks. As a follow-up study to these approaches, the following figures and discussions show the use of a data visualization tool in Sequest to find disease associated patterns at selected m/z ratios in two data sets.

4.4.2 Investigation of Disease Patterns Found in the Mouse Plasma Study

The LC/MS data from the mouse plasma study was ideal for the observation of patterns due to the high reproducibility of the chromatograms. The fact that the mice were genetically homogenous and lacking IgGs may also have contributed
to the observations. As shown in Figure 4.9, changes in the MZ pattern at 945 units, for example, can be distinguished for a mouse with a growing tumor with estrogen only treatment versus estrogen+Tamoxifen treatment (note: results for two samples are shown for each treatment). Slight changes in MZ values result in significant differences in the patterns indicating both the richness of the data and the possibility of slight variations in MZ values causing non-reproducible data (such as in the case of a MS instrument which has been calibrated poorly). Replicate LC/MS analyses, 2 injections from the sample preparation, are shown to demonstrate the reproducibility of retention times, peak shapes and intensities. Comparison of identified peaks revealed the retention time shift between replicate sample injections to be less than half a minute on average. The majority of peaks are found at retention times corresponding to the increase in organic (acetonitrile) mobile phase from 2% to 40% which occurs over the first 160 minutes of the analysis. These results indicate that under tightly controlled conditions for sampling, sample preparation and sample analysis the LC/MS patterns will be reproducible from sample to sample. Other distinctive tumor patterns can be distinguished at different MZ values (e.g., 681, 781, 821, and 1002 units). Although the m/z range of the LTQ is from 200 to 4000 units the patterns tend to be richest in peaks from around 500 to 1000. This is a range consistent with many tryptic peptides carrying 2 or 3 positive charges.
Figure 4.9. Replicate Extracted Ion Chromatograms from the Mouse Xenograft Proteomic Study. The chromatograms in each figure were generated on an Ettan MDLC system (GE Healthcare, Piscataway, NJ) coupled to a Thermo Finnigan linear ion trap mass spectrometer.

(a) A different pattern of peaks is apparent in the m/z range of 993 ±0.5 when comparing replicate runs of digested glycoprotein enriched plasma from mice with tumor and not receiving estrogen than from mice with tumor and receiving estrogen. The replicate runs show reproducible patterns of peaks.

(b) At an m/z range of 945.0 differences in the number, retention times and intensity of peaks are also evident when comparing chromatograms generated by analysis of plasma from mice with tumor and receiving estrogen to chromatograms from mice with tumor receiving estrogen and Tamoxifen (reduced tumor growth).
The chromatograms in Figure 4.10 were produced during the analysis of plasma from nude mice as described in chapter 2 of this thesis. The mice were divided into groups which received an MCF-7 tumor implant and groups which did not. Both the tumor and non-tumor group were subsequently divided into four sub groups- control, estrogen, tamoxifen and estrogen and tamoxifen (see Figure 4.9). All of the data shown in Figure 4.10 is from the analysis of plasma at the end point of the study (6-weeks).

At some m/z values it was possible to detect patterns of peak intensities and retention times which were specific to a disease or treatment group. Due to the fact that LC/MS runs were performed in nonsequential duplicates this observation could be confirmed as not being due to chance by aligning replicates to demonstrate the reproducibility of the patterns. The pre-selected m/z data shown in these figures was obtained from a series of ion chromatograms where data collection covered all m/z signals over a range of 0 to 2000 units. Mass-to-charge ratios can be selected by hand or with software (such as the CLUE-TIPS software discussed at the end of this chapter) to find a pattern which is distinctive for the disease vs. control sample or to search for expected similarities between related data sets. The m/z patterns represent differences in levels of protein expression due to the presence of the tumor in the animal and the effects of estrogen or tamoxifen on the animals that are detected by mass spectrometry and that change with disease and environmental factors.

Figure 4.10 on pg 192 displays examples of selected m/z values that show signatures potentially related to tumor growth in the mouse which are monitored in the
LC/MS experiment. The samples used in this comparison are from mice receiving estrogen at the six-week time point but differ in that one group has tumor implants and the other group does not. An example of the differences the can be seen at 120 in part a of the figure where prominent peaks can be observed in the disease sample but are absent in the control sample. A large peak at 105 min present in all four runs is presumably derived from a high level plasma protein, which could potentially be used as a landmark to align chromatograms.

The second set of chromatograms shows a comparison of LC/MS chromatograms resulting from analysis of plasma from two groups of mice at the six-week time point which had tumor implants and were receiving estrogen but differed in the fact that one group was receiving Tamoxifen to inhibit tumor growth. Distinct peak patterns are again discernible in each duplicate set of chromatograms when the two treatment groups are compared.

The pattern analysis study utilized 22 LC-MS data files from the mouse data study\textsuperscript{31} at different time points under various treatments (estrogen, tamoxifen and both). During the process of performing peak area analysis of the kinase proteins as discussed in Chapter 2 of this thesis around sixty mass windows were manually examined, with patterns being discernible in most of them, over the course of several days. Automatic analysis of mass windows enables the much faster examination of all of the MZ windows and subsequent comparison of the data. A human examiner of the MZ windows will be able to pick out similar topography in LC/MS data irregardless of retention time shifts of the peaks and shifting chromatographic baselines. For automatic data processing this is a
much more demanding task which includes aligning prominent peaks which are consistent for a MZ window in each analysis.

For the CLUE-TIPS software computer based alignment in chromatographic retention time was performed using the OBI-Warp \(^{36, 37}\) program. Computer based m/z filtering, alignment, denoising and baseline correction are combined in the CLUE-TIPS software suite to generate a set of matrices termed the Knowledgebase which contain Extracted Ion Chromatograms (EIC) for all samples associated with each m/z value. A Tanimoto inter-point map is then constructed by applying Tanimoto distance for all pairs of samples in each m/z-matrix. The Tanimoto distance is a method of calculating the similarity (or dissimilarity) of two data sets by analyzing the numbers of elements (i.e., a chromatographic peaks present at specific retention times) in each data set and the number of elements the data sets have in common. A heat map can then be used to visualize the Tanimoto inter-point map after hierarchical clustering. The CLUE-TIPS publication describes this process of utilizing pattern identification in complex proteomic data sets in further and in greater detail\(^ {38}\).

### 4.5 Experimental Procedures for the Study of Human Diabetic Plasma

Data generated in our laboratory by Dayarathna was used to further investigate disease patterns in humans\(^ {39}\). In this experiment samples of human plasma (purchased from Bioreclamation, Inc.) from individual donors were matched by age, race, and gender. Pooled samples of 5 individuals from a control group (group 1); obese group (group 2); obese plus diabetes and no hypertension (group 3), and obesity plus diabetes with hypertension (group 4) were first depleted of six abundant proteins (albumin, immunoglobulin G, immunoglobulin A, transferrin, alpha-1-antitrypsin, and haptoglobin)
using the MARS column (Agilent, Palo Alto, CA) and then fractionated with M-LAC. The unbound portion was subjected to tryptic digestion and LC-MS analysis in the same manner as the mouse plasma samples.

4.6 Results of Pattern Investigation in a Diabetes Study

Pattern investigation results from a diabetes study performed using human plasma samples for MS analysis are illustrated in Figure 4.10. The samples were analyzed in replicate nonsequential injections on the same analytical system as that used for the mouse xenograft study. The human plasma samples underwent M-LAC fractionation and the nonbound portion was then subjected to trypsin digestion and LC/MS analysis. In the plasma samples from diabetic patients we observed EIC patterns at selected mass values (e.g., 681, and 781 units) that were different from the corresponding profiles for controls or other disease groups. In this study we observed a lower degree of reproducibility between replicates than that observed in the study of the nude mouse xenograft which was related to performance of the nano-flow HPLC system. Alignment of the chromatograms, however, was possible since there is a distinct peak at 83 minutes (see the top panel) and between 111 and 116 minutes there is at least one peak present in the normal sample while in the diabetic analysis there are no prominent peaks in this time period. In the middle panel displays the same four LC/MS analysis at 681amu. A strong peak at 106 minutes dominates all four chromatograms and gives a landmark by which to align the results. It can be seen that a peak at 125 minutes is larger in the normal vs. diabetic analyses. In the final portion of the figure chromatograms from normal, obese, diabetic and diabetic-hypertensive LC/MS analyses are compared at 781 m/z. There is a
low degree of similarity in the four analyses which indicates the possibility of pattern recognition software being able to correctly identify a condition from other similar conditions. Interestingly the highest degree of similarity is seen in the diabetic and diabetic-hypertensive chromatograms although further investigation is needed to confirm these observations. Compared to the mouse LC/MS patterns there is a lower quantity of peaks in the LC/MS patterns from the human study, a fact which may be attributable to a larger effect of the tumor on the proteins secreted into the blood. In this case a direct comparison of the mouse LC/MS patterns to the human LC/MS patterns is not ideal due to the fact that the mouse data is from the glycoprotein enriched M-LAC fraction while the human plasma is glycoprotein depleted M-LAC fraction.
Figure 4.10. Comparison of Patterns Found in The LC/MS Analysis of Human Plasma. At m/z ranges of 781(a) and 681(b) replicate chromatograms of diabetic and normal plasma samples there are distinct differences at both m/zs between disease and normal. The bottom set of chromatograms(c) displays chromatograms of the four sample types at an m/z of 781.

4.6.1 The Potential of Bioinformatics to Enable Disease Pattern Searching

The observation of specific MS patterns associated with glycoproteomic studies results in MS signatures which could be associated with tumor growth as was observed in the mouse plasma study. Due to the fact that it was also possible to observe disease
associated peak patterns in human samples (Figure 4.10). The initial observations in this section formed the basis of the follow-up study where Dr. Akella and Dr. Retjar took the observations described earlier in this chapter and developed a program that allowed automation of the pattern interpretation process. Automated searching by computer software is a complex multi-step process which starts with aligning, noise reduction the LC/MS runs and attempting to match the peaks from run to run. Alignment of LC/MS analyzes of complex proteomic samples is necessary due to natural variability of peptide elution times. CLUE-TIPS utilizes Obi-warp software to align runs by identifying landmark peaks having high confidence tandem MS identifications and noise reduction was accomplished by using a mathematical moving average filter. A Tanimoto inter-point map was generated by arranging mz-matrices into a matrix where m/z values are in rows and similarity values are in columns. Generation of a Tanimoto similarity matrix is beneficial in that it occupies a smaller amount of storage space compared to the output from the LC/MS instrumentation. Applying clustering methods to the Tanimoto inter-point map reveals patterns associated with variability between different LC/MS runs, sample origin and disease state.

4.7 Conclusion

Current LC/MS proteomic platforms are capable of generating extraordinarily large amounts of data and thus, rendering manual review of the data impractical. Additionally, the proteins identified during analysis are a part of complex biological systems which for which new information is continually being gathered as to biological function and disease associations. The time and required database expertise are
formidable barriers to utilizing a significant fraction of the protein identifications. Improved bioinformatic approaches offer an answer to this challenge and numerous programs exist to perform tasks ranging from peptide identification to assigning biological meaning. We have illustrated in this chapter the potential of Proxeon's ProteinCenter program to expedite the process of moving from data acquisition to publication by bringing together a list of identified proteins, peptides and biological information along with clickable links to protein structure and function.

Proteomic LC/MS data, such as that generated during the nude mouse xenograft experiment, contains numerous unidentified peptides due to low level and non-fragmented peptides. This is an example of data generated by costly proteomic experiments, which is not utilized. While increasing the efficiency of the chromatographic separation or increasing the speed of the mass spectrometer both help fragment more peptides, it is unlikely that all proteomics data will successfully be interpreted. As a solution, our group investigated the possibility of applying pattern recognition programming to extracted ion chromatograms to utilize information from all peptides in the sample thus including data from non-fragmented peptides. After the initial phase described here, we collaborated with Dr. Akella and Dr Rejtar which resulted in the pattern recognition program which is termed CLUE-TIPS and is the basis of a paper produced by our group\textsuperscript{38}. In this collaboration Dr. Akella and Dr. Rejtar supplied the bioinformatics and programming expertise. The Clue-TIPS program is beneficial in terms of not only maximizing the information gained from an experiment but also in terms of increased speed and reduced data storage. These factors lead to the possibility of analyzing large numbers of samples for similarities and differences relating to health of a
human subject. An alternative application is to use the program to visualize the quality of LC-MS runs in multi sample experiments.
4.8 References


Chapter 5. Future Directions
5.1 Analysis of Clinical Plasma and Serum Samples Through the Use of Proteomic and Peptidomic Methods

Chapters 2 and 3 describe research which includes the proteomic analysis of clinical blood samples from either human or animal subjects which relied on enrichment of glycoproteins through the use of a multi-lectin affinity chromatography (M-LAC) column to detect low level proteins. The M-LAC column was developed in our laboratory and while the method has been utilized successfully in several studies\textsuperscript{1-5} with the experiment reported in Chapter 2 being the first time M-LAC was used with mouse plasma. Each of the proteomic experiments excepting the auto antibody screening described in Chapter 3, were performed with the use of nano flow LC/MS systems. The resultant data was of very high quality in regards to chromatographic reproducibility and the number of proteins identified and enabled us to utilize the data to discern changes in the concentrations of several proteins and peptides. Chapter 3 deals in part with the detection of endogenous peptides in serum and the development of an approach for the detection of autoimmune proteins. The results from the peptidomic experiments performed on serum from patients having hyperinsulinaemia as described in Chapter 3 provided clues of reduced proteolytic cleavage of acute phase proteins while the autoimmune screening revealed possible reactions of the patient’s immune systems to self-proteins.

The experiments performed in Chapter 2 of this thesis will provide a reference point for future proteomic experiments in regards to the use of M-LAC fractionation with non-human proteins. In addition, data published from the mouse xenograph experiment
will help fill a current lack of data for murine plasma proteomic studies of cancer. It is concluded that M-LAC fractionation of plasma proteins can aid detection of biologically significant proteins in tumor-bearing animals and future studies aimed at increasing the sensitivity of the method to improve characterization of low level tumor markers. Human and murine glycan structures are analogous in respect to carbohydrate composition both of which contain N linked chains of glucose, mannose and N-acetylglucosamine and are expected to be similarly bound by lectins.

The results presented in chapter 3 provide valuable clues as to the cause or progression of hyperinsulinaemia in gastric by-pass patients. Further study is warranted as to both possible involvement of vitamin D binding protein and possible autoimmune causes.

5.2 Investigation into the Use of Bioinformatic Tools for Improved Interpretation of Proteomic Data

Chapter 4 of this thesis deals with the interpretation of data produced by clinical proteomics experiments through the application of novel bioinformatic tools. ProteinCenter, developed by the company Proxeon was utilized to mine protein identifications, resulting from the mouse plasma study described in chapter 2, for proteins or groups of proteins which showed changes corresponding to tumor growth. This program successfully increased the speed of identifying proteins with disease associations and identifying changes in biological processes such as an increase in enzyme activity in the tumor bearing animals receiving estrogen. The second section of this chapter deals with the conception of a computer program which can be used to identify patterns in
LC/MS data with no reliance on fragmentation data and correspondingly, no reliance on protein identifications. Initially data from the mouse plasma experiment was used to establish the feasibility of a pattern detection method which relies on extracted ion chromatograms. An additional investigation using LC/MS data from proteomic experiments on human plasma successfully detected patterns that may be associated with diabetes led to collaboration with Dr. Akella and Dr Rejtar to develop an automated pattern recognition program which is termed CLUE-TIPS\textsuperscript{6}.

Future work in our laboratory will continue to search for the best informatics tools to aid interpretation of proteomic data as quickly and as completely as possible. ProteinCenter is a viable option to aid future researchers in understanding their data due in part to the fact that the program is continuously updated with new protein functions and disease associations thus aiding investigators efforts to stay current in a vast and expanding field. The CLUE-TIPS program offers a unique opportunity for the analysis of numerous samples and the development of a knowledge database capable of detecting disease associated changes in proteomic or metabolomic data from LC/MS experiments or other types of experiments such as NMR.
5.3 References


