Proteomics for Cancer Biomarker Discovery

by

Samuel Louis Volchenboum

B.S. Biochemistry, University of Illinois at Urbana-Champaign (1991)
B.S. Honors Biology, University of Illinois at Urbana-Champaign (1991)
M.D., Mayo Medical School, Rochester, Minnesota (1998)
Ph.D. Molecular Biology, Mayo Graduate School, Rochester, Minnesota (1998)

Submitted to the Department of Health Sciences and Technology
in partial fulfillment of the requirements for the degree of

Masters of Science in Biomedical Informatics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2007

Author . . . .
Department of Health Sciences and Technology
May 11, 2007

Certified by . . . .
Isaac Kohane, M.D., Ph.D.
Associate Professor
Thesis Supervisor

Accepted by . . . .
Martha L. Gray, Ph.D.
Director, Harvard-MIT Division of Health Sciences and Technology

Proteomics for Cancer Biomarker Discovery

by

Samuel Louis Volchenboum

Submitted to the Department of Health Sciences and Technology
on May 11, 2007, in partial fulfillment of the
requirements for the degree of
Masters of Science in Biomedical Informatics

Abstract

Background: If we are to successfully treat cancer, we must understand the biologic underpinnings in conjunction with early diagnosis. Genome-wide expression studies have advanced the research of many cancers. Nevertheless, understanding which genes are expressed in a tumor is not equivalent to knowing which proteins are being produced. Proteomics hold great promise for careful examination of the proteins in complex biologic fluids and tissues, and it may be possible to detect disease from a patient's serum, long before it would otherwise be clinically evident. Although there have been steady advances in all the steps of a proteomic analysis, much remains to be standardized. Because of some high-profile problems with the initial analysis of ovarian cancer proteomic data, early exuberance has now been tempered and replaced by a more methodical approach to these studies. **Hypothesis:** My hypothesis in this thesis is that proteomics is a valuable tool in the diagnosis and study of cancer, as will be demonstrated in several steps. **Methods:** First, I describe the current field of proteomics, specifically as it applies to early detection of cancer and biomarker discovery. I lay out the current state-of-the-art technologies for preparing samples and enumerating the proteins in complex fluids and tissues, giving special treatment to the main threats to validity-chance and bias. I also describe the bioinformatic tools necessary for analyzing the large amounts of data produced. Through the example of a mouse model of colorectal carcinoma, I demonstrate the steps involved in a proteomic study, from procuring samples to peptide and protein determination to bioinformatic analysis. Finally, I discuss these findings in light of the proteomic considerations discussed earlier. **Results:** From this work, I discovered that proteomic profiling can describe the proteins in serum from mice both with and without colon cancer. Furthermore, I developed a naive Bayes classifier that could distinguish between the serum of mice with colorectal carcinoma and their normal litter-mates. **Contributions:** Through this work, I have contributed the following. I described the field of proteomics with special emphasis on cancer biomarker discovery and early detection. I enumerated the challenges and pitfalls to developing early detection schemes for cancer based on high-dimensional proteomic analyses. I described a set of experiments on mice harboring a gene mutation that predisposes them to colorectal carcinoma. I detailed the bioinformatic analysis of this data, including the development of a naive Bayes classifier to differentiate the cancerous state from the normal state. Finally, I discussed the caveats of the current work, in reference to the initial discussion on the challenges and pitfalls of early detection schemes and cancer biomarker discovery.

Thesis Supervisor: Isaac Kohane, M.D., Ph.D.
Title: Associate Professor

Acknowledgments

I would like to acknowledge the contributions of Raju Kucherlapati and his Gastroenterology Research Group at the Harvard Medical School - Partners Healthcare Center for Genetics and Genomics. A hearty thank you to Ken Hung, who performed the wet-bench work and prepared and submitted the plasma samples, and David Saraccino, who performed the mass spectrometry analysis. I would like to thank Alvin Kho for his tireless and patient assistance as I negotiated this foray into biomedical informatics. I am grateful for the continual valuable input from Zak Kohane, my mentor and advisor. I would like to thank Patrick Winston for his vision of making MIT students better communicators. Finally, I continue to be indebted to my wife and family for their patience and support while I pursued this degree.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Background

In order to successfully treat cancer, we must understand its biology as well as be able to diagnose it before it has spread. The completion of the sequencing of the human genome and concurrent explosion in expression array technologies have led to unprecedented exuberance regarding the potential to understand and cure disease. Nevertheless, this excitement is tempered by the realization that while quantitating gene expression levels is an important piece of the puzzle, an essential component will be studying the actual expressed protein complement of cells and tissues, a field known as proteomics [7]. The science of proteomics involves not only the acquisition of data about the expression of proteins in cells and tissues, but also the analysis of protein expression and integration of these data with existing knowledge, both in related and disparate fields. The aim of this thesis is to describe the field of proteomics, giving details about both the technology as well as the bioinformatics with a special emphasis on cancer biomarkers. Included will be an example of my own work illustrating one potential application of this technology - identifying biomarkers for human cancer using mouse models. Following this example will be a description of the pitfalls involved.

1.1 Proteomic Technologies and Data Analysis

Until recently, the study of proteins and their interactions was a time-intensive and expensive endeavor often requiring the production of specific antibodies to the proteins in question. Relatively recent technological advances in the sensitive and specific separation and identification of proteins have transformed proteomics into a high-throughput science. Discussed below are the three major components of any proteomics experiment: sample preparation, protein separation, and protein identification.

1.1.1 Sample Preparation

Before proteins can be separated, the sample must first be collected and prepared. For a proteomic study, the sample may be fluid such as serum, plasma, urine or seminal fluid. The study of solid tumors and tissues represents a challenge, as contamination from surrounding normal cells and tissues can make it difficult to interpret results. One way to minimize stromal contamination is through laser capture microdissection (LCM), a method by which a relatively pure population of tumor cells can be isolated for further study [14].

Although not commonly discussed at great length in most methods sections, the sample isolation and preparation is a critical factor in a proteomics experiment. The central principle of most experiments is to keep constant all variables between groups, except for the particular agent being examined, thus allowing any difference between the groups to be attributed to the agent itself. Bias is a significant threat to validity in proteomics experiments, and it can often be explained by errors or inconsistencies during sample preparation [33, 34]. When samples are handled differently, this introduces noise into the signal for one of the compared groups. When this happens, it is very difficult, if not impossible, for post-processing or post-experimental data manipulation to validate the results. Some important steps in preparation include types of collection tubes, the time from collection to spinning and/or freezing, differences in storage temperature, number of freeze/thaw cycles, and/or all the factors involved in the analysis on the mass spectrometry (MS) instrument. Reproducibility in results does not assure that there is no bias, as the same inconsistencies may be present from one experiment to the next. Therefore, in the design of an experiment, there must be careful planning to ensure consistent sample collection and preparation between and among the study and control groups [4].

1.1.2 Protein Separation

Separation of proteins is a key step in any proteomic experiment. The greater the separation achieved prior to protein or peptide identification, the greater the resolving power of the study. Common methods of protein separation include one- and two-dimensional electrophoresis and liquid chromatography [7]. While one-dimensional electrophoresis (1-DE) separates proteins based on their sizes, two-dimensional electrophoresis (2-DE) separates first based on the isoelectric point (pI) followed by separation based on size. Gel-based separation methods are rapidly becoming replaced by methods that involve peptide separation by liquid chromatography (LC) techniques linked to a mass spectrometer [10]. The most basic form of LC involves separation of digested peptides through a C18 resin. More complex separations can be achieved through the use of other in-line methods, such as a cation exchange resin (multi-dimensional protein identification technology, MudPIT). These separations can also be run "off-line," that is, not in series with the MS. Performing these methods off-line is advantageous as it allows sample manipulation and optimization between dimensions [10].

1.1.3 Peptide and Protein Identification

Protein identifications are almost always achieved through mass spectrometry. A mass spectrometer consists of an ion source, a gas analyzer that can measure the mass-to-charge ratio (m/z), and a detector that can measure the number of ions at each m/z value [1]. The two techniques commonly used to volatilize and ionize the proteins and peptide source are electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI). Because ESI ionizes the peptides right out of a solution, it is easily coupled to liquid-based separation techniques, such as LC. MALDI, on the other hand, sublimates and then ionizes the samples out of a solid crystalline matrix and is therefore suited to more relatively simple peptide mixtures.

The mass analyzer measures the mass-to-charge ratio of the ionized analytes. The four main types of mass analyzers are the ion trap, time-of-flight, quadrupole, and Fourier transform ion cyclotron (FT-MS). While ion traps are relatively inexpensive and sensitive, a disadvantage is their low mass accuracy. On the other hand, the FT-MS captures ions under high vacuum in a high magnetic field and has high resolution, mass accuracy, sensitivity and dynamic range. Nevertheless, the expense and operational complexity of the FT-MS has slowed its acceptance. Whereas ESI is usually coupled to ion traps and used to generate fragment ion spectra, MALDI is usually coupled to TOF analyzers that measure the mass of intact peptides.

The peak patterns in the collision-induced (CID) spectra are compared against a comprehensive protein sequence database to compile a ”hit list” of peptide matches. These peptide matches are then analyzed to produce the most probable list of proteins that could have been the source of the particular set of peptides.

One of the difficulties inherent to high-throughput proteomics is filtering the peptide assignments to derive a list of highly-likely correct identifications. In small experiments, an expert in MS can examine the spectra and determine which peaks are likely to lead to robust identifications. Because of time and personnel constraints, this method is infeasible for larger experiments. In these cases, peptide lists can be compared against databases of known protein identifications, though this may lead to many false positives.

1.1.4 Methods of Quantification

Identifying the quantity of proteins in a complex mixture remains an unsolved problem. Most solutions involve some sort of multi-step separation / fractionation followed by a procedure for identifying the components in each separated fraction. For instance, one method is to separate a complex mixture via 2-DE, to quantify the intensity of each spot and then identify the proteins with MS. While both MALDI and ESI-MS/MS are efficient and tested technologies for identifying gel-separated proteins, the results of these studies often show similar lists of proteins. This indicates that this technique has a limited resolving capability [18].

The mainstay for protein identification and quantification remains LC-MS/MS (reviewed here: [28]). Aebersold and Mann identify three hurdles to accurate protein quantification through LC-MS/MS [1]. First, single dimension chromatography does not provide sufficient separation for complex mixtures. Second, the relationship between peak signal intensity and peptide amount is not clear. Finally, there is a large amount of data collected from a mass spectrometer, making analysis cumbersome. Despite the difficulties, each of these technological hurdles has been tackled to such an extent that the platforms emerging constitute a robust system for protein identification and quantification.

Quantification is commonly achieved through site-specific isotope tagging chemistries, lending access to various sub-proteomes. A newer method involves metabolically labeling one class of cells with one stable isotope, while the other is labeled with another. Sensitive MS can then distinguish between the two molecular weights of the differentially labeled proteins, and the differences in protein abundance can be quantitated based on the peak heights. This method, known as stable-isotope labeling with amino acids in cell culture (SILAC), is described in [21].

1.1.5 Pitfalls

Regardless of the technology chosen, great care must be taken to ensure that there is uniformity between experiments. In fact, some researchers advocate several well thought-out small pilot studies to show validity before venturing on to larger studies [33]. As discussed above, the greatest threat to validity is chance and bias, and bias can be minimized by devoting considerable time and effort to planning experiments.

1.2 Bioinformatic Considerations

Any proteomics experiment will generate large amounts of data for analysis. Collecting, organizing, storing, and analyzing this data should be performed in an orderly and methodical manner. The basic approach to each of these steps will be outlined below.

1.2.1 Data storage

Many systems have been developed for the storage (and often concurrent analysis) of MS data, such as the open-source initiative, the Computational Proteomics Analysis System (CPAS) [35]. While CPAS has an established following of devotees, there are some drawbacks to the system, namely the manipulation required to admit certain data formats into the data pipeline. Nevertheless, it is becoming a widely accepted benchmark for MS data storage and analysis.

There are several features desirable in any storage system. First, it must be extensible. In other words, as the technology changes, the system must be easily adaptable to new tools and techniques. Second, it must be fast. The best systems run on distributed hardware, thus improving the performance many times over, when compared to an individual machine. Also, it should promote collaboration and data sharing between and among scientists. Ideally, a system can have both a rigid access system and also a robust method of data sharing.

The primary output of an MS experiment is raw spectral data. In most cases, though, the data is immediately searched and matched to peptides and proteins. Therefore, while the raw data is routinely stored and available for subsequent analysis, the bulk of data retrieval and analysis will be on the peptide and protein matches. In many instances, however, the original spectra will need to be examined for quality. This is especially useful when considering a peptide match with a relatively low score. An expert should be able to examine the MS spectra and make a judgement about the quality of the match. Again, given the relative large amount of data, such a system must be nimble and fast enough to allow real-time manipulation of spectra.

In addition to storing the raw spectral data, the system must be able to house all relevant information about an experiment, including the sample collection and preparation, as well as the details of the separation and MS. This is especially important, as consistency between MS runs is one of the keys to a valid study. Some examples of essential data include the type of specimen, how it was collected, patient demographics, time between collection and freezing, any centrifugation, separation prior to MS (including separation conditions, column details, fraction size), as well as all the details about the MS itself. Failure to include any of these may bias the study and therefore weaken the validity.

1.2.2 Data Analysis

There has been a push towards data ”pipelines.” A data pipeline is a coordinated set of analysis steps run in series, often with little user interaction. For instance, the Trans-Proteomic Pipeline (TPP), an open-source initiative developed at the Institute for Systems Biology (Seattle, WA), is an analysis package that can be integrated into the CPAS system [25]. This set of tools is designed to facilitate the database search, validation, peptide quantification and protein quantification steps and requires that proprietary data be converted into an open-source format (mzXML). One of the drawbacks of this method is that in its initial implementation, it required a commercial license (SEQUEST) to one of the MS platforms in order to do the initial search steps. Subsequent incarnations allow for alternative open-source search platforms (eg. XTandem).

Regardless of whether or not a pipeline is used, analysis of proteomic data usually has two major phases. First, there needs to be a determination as to which peptides and proteins were present in the mixture and (if applicable) in what quantities, as related to some control. Once the peptide and protein identifications are made, the second step is to compare the study and the control population, using appropriate statistical measures.

1.2.3 Pitfalls

As with any biomedical informatic analysis, care should be taken to ensure that the most appropriate metrics are used at each step. In many cases, there is not a clear standard (for example, the distance metric used in a clustering algorithm). In these cases, there may need to be trial-and-error as well as optimization to determine the most appropriate algorithms. Thorough and detailed reporting of methods will help assure that analysis is reproducible.

1.3 Cancer Biomarkers

A biomarker is a chemical or substance in the blood, other body fluid, or tissue that may signal the presence of a disease state. Identification of biomarkers of disease may lead to early detection or prevention. One approach to biomarker discovery involves the separation and analysis of proteins in samples taken from those with and without a particular disease. While analysis of the genome and transcriptome may give hints as to important proteins in the development of disease, only analysis of the proteome will give an accurate representation of protein expression. Many biomarkers for cancer are already known, and these makers are present in high and detectable amounts in the disease state, such as prostate-specific antigen [3]. Tumors are highly vascular, and it is likely that tumor-specific proteins are secreted into the bloodstream. Furthermore, tumors induce a state of inflammation, and the plasma proteome may reflect the body's response to cancer in some specific way. Therefore, it is reasonable to suspect that it will be possible to discover proteins unique to cancer by examining the plasma proteome of organisms with and without tumors. To understand the current state of cancer biomarker discovery, it is useful to understand the history of biomarker discovery in ovarian cancer.

1.3.1 Ovarian Cancer Biomarkers - Lessons learned

In 2002, Petricoin and colleagues reported that they used MS to develop a classifier that could identify serum from patients with ovarian cancer with 100% sensitivity and 95% specficity [31]. Petricoin and colleagues made their data publicly available on their website. In a follow up to this study by Zhu, et al., reported similar results [45] using the Petricoin data set. When applied to another publicly available data set from Petricoin, Zhu reported that their classifier for ovarian cancer had 100% sensitivity and 100% specificity, remarkable for any clinical test. This initial excitement led to predictions that a clinical test would be available as early as 2004. But plans were halted by the FDA when questions were raised about the tests reproducibility and reliability [41, 11, 16]. A review of the Zhu data by Baggerly, et al. showed problems with the analysis, and that the method used performs no better than chance [2]. Liotta, et al., authors of the originally posted public data, explain that this unfortunate situation arose because of poor communication between study groups [27]. They attributed the discrepancy to lack of communication. Furthermore, they claim that they are not surprised by the discrepancy because of the lack of standardization between the two sets of data. Ransohoff briefly summarized the story and comments on serum proteomics as a whole [33], giving special attention to the biggest threats to validity, chance and bias.

Chance and Bias

In Ransohoff's view, the biggest challenge to proteomics is reproducibility, and the biggest threats to reproducibility are chance and bias. To answer whether or not chance (or over-fitting) can explain results, the same experiement can be duplicated under the same conditions (but with different subjects). Bias is more difficult to ascertain. As explained by Ransohoff, even under the most controlled circumstances, bias may be unavoidable. For instance, tumor samples may need to be harvested in the operating room, while normal tissue may not, and this difference may be impossible to overcome. Finally, he suggests that more carefully planned studies may help avoid a similar situation,

> ”The solution is not to post more raw data on the Web, although doing so may be useful for some purposes. Rather, the solution is to provide appropriate attention to the process of design, conduct, and interpretation of research and to thoroughly report that process in rigorously reviewed journal articles. In other words, what needs to be reported is not raw data but rather the process of how patients, specimens, and analyses were handled that lead to those data. Rules of evidence may then be applied to determine whether bias or chance provides an alternate explanation for results.”

1.4 Motivation

The motivation for this thesis is to explore the process of performing a complicated and detailed proteomic analysis. In doing so, each of the above factors will be explored and analyzed. Through a specific example, I will demonstrate some of the problems inherent to these types of analyses, and I will offer some possible solutions. Finally, I will comment on future work I am planning and will outline some of the risk factors and how they are being addressed to help ensure success.

Chapter 2

Mouse Models of Colorectal Carcinoma

During 2005 and 2006, I collaborated with Dr. Raju Kucherlapati and the gastroenterology research group at the Partners Healthcare Center for Genetics and Genomics (HPCGG). I was responsible for working with the mass spectrometry core to collect and analyze data for their project which focussed on mouse models of colon cancer. The following is a summary of that work, with special attention to the topics discussed above.

2.1 Background - Colorectal Carcinoma

Worldwide, there were over 750,000 new cases of colorectal cancer diagnosed in 1990, the most recent year with international estimates [6]. Colorectal cancer is a leading cause of mortality and morbidity, affecting men and women equally, and is the third most frequent cause of cancer-related death. Estimated annual incidence in the US for 2004 was more than 150,000 new cases with over 56,000 deaths [24]. The five-year survival rate for colorectal cancer remains less than 70% [24], with the chance of survival correlating with the stage of disease at the time of detection. Five-year survival is 90% for those with cancer detected at an early stage and under 10% for late-stage disease [5]. Only a minority of colorectal cancer is detected at an early stage [19], probably due to a combination of inadequate screening [13] and the occult nature of the disease.

Colorectal cancer screening includes digital rectal examination, fecal occult blood testing [38], flexible sigmoidoscopy, and colonoscopy. Because of the costs and discomfort associated with sigmoidoscopy and colonoscopy, computed tomography-based virtual colonography is being explored as an alternative option [15, 22]. Fecal DNA mutational analysis is also being studied [9]. While these newer methods hold promise, the confluence of recent advancements in mass spectrometry and mouse models of colon cancer along with the ever-increasing amounts of genomic and proteomic data have made plasma proteomic analysis an attractive tool for both colorectal cancer biomarker discovery as well as a possible tool for early detection.

2.2 Mouse Model of Colorectal Carcinoma

Much is known about the molecular mechanisms underlying colorectal carcinoma. Inactivation of the adenomatous polyposis coli (APC) gene on chromosome 5q21 has been found to be responsible for familial adenomatous polyposis (FAP) [26, 29], a cancer predisposition syndrome. We have studied a well-characterized intestinal tumor model, $\mathrm{Apc}^{\mathrm{Min}}$ [40], and have identified a novel plasma proteomic profile in mice harboring this mutation [20]. This APC gene mutation makes the mice more susceptible to adenomas of the small intestine. The original APC mutant harbored a stop codon at position 850. We have constructed a new mutation of the APC gene which leads to a larger transcript, due to a stop codon at position 580. We have found that mice with this mutation are more likely to develop carcinomas of the large intestine, and this mouse model should more closely mimic human colon cancer than prior mouse models.

2.3 Study Design

Plasma samples from individual tumors bearing ApcΔ580 mice and their wild-type litter-mates were isolated on two different days. A total of 20 mice were sampled -

10 with the mutation and 10 normal litter-mates. Samples were separated by liquid chromatography, and peptides were identified by FT-MS. Peptide identifications were filtered and used to generate a list of proteins. This list was used to develop a classifier.

2.4 Materials and Methods

2.4.1 Animal Husbandry

Mice were purchased from Jackson Laboratories. Apc Δ580 mice were generated in our laboratory through standard techniques. A similar mouse has been described in the literature [8]. Heterozygous Apc Δ580 mice were mated with wild-type B6 mice. The resulting offspring were screened by PCR of tail DNA using standard methods. Heterozygous Apc Δ580 mice from these matings were used for subsequent studies. Wild-type age- and sex-matched litter-mates were used as controls.

2.4.2 Plasma Harvest and Tumor Quantification

At the time of sacrifice, a lethal coma was induced by intraperitoneal injection of Avertin anesthetic. Blood was removed from the right ventricle by cardiac puncture using a 22-gauge straight needle to prevent hemolysis. Blood was placed in EDTA-coated tubes to prevent coagulation and proteolysis. After centrifugation, the plasma-containing supernatants were removed and aliquoted in individual freezer vials. The vials were initially placed at -20C for one to two hours and then transferred to -80C for storage prior to mass spectrometry analysis. The small and large bowel of the mice were removed and opened longitudinally. The number of tumors was counted using a dissecting microscope.

2.4.3 Plasma Sample Preparation

Twenty-five μl aliquots of either total plasma or glycoprotein-enriched plasma were mixed with 100 μl of 6M urea 1% SDS 100 mM ammonium bicarbonate 10 mM DTT and heated to 37C for one hour. Iodoacetamide was added to 30 mM and the sample placed in the dark for one hour. Residual iodoacetamide was quenched with 2 M DTT. Samples were then diluted to 1.5 ml with 20 μg of Promega sequencing-grade modified trypsin in 5 mM CaCl2, and allowed to digest overnight at 37C. Post-digest, samples were acidified with formic acid to pH ¡3.0. Digests were cleaned up using cation-exchange cartridges, 30 mg MCX cartridges (Waters), and eluted with 250 mM ammonium formate and 6% ammonium hydroxide in 50% acetonitrile. The samples were lyophilized to dryness and re-dissolved in 50 μl 5% acetonitrile 0.1% formic acid.

2.4.4 Mass Spectrometry (MS)

The peptides were separated into 12 fractions using off-line cation-exchange chromatography. The peptide fractions were then normalized and 5-15% of each fraction plated in a 96-well plate, a standard peptide mixture (ovalbumin) was added to each sample at constant concentration, and the peptides from each of these fractions were then separated further over 150 minutes on a 75 μm x 16 cm nanospray chromatography column with direct spray into the FT-MS. The FT was run using a ”top nine” run configuration at 200K resolution. Peptide identifications were made using SEQUEST through the BioWorks Browser 3.2 (Thermo Scientific).

2.4.5 Data Analysis

Data analysis consisted of the following steps: (1) filtering of MS results, (2) data parsing and database storage, (3) peptide searching, protein identification, protein list optimization, and (4) analysis of peptide and protein differences between classes.

Since our program had no existing infrastructure or software for mass spectrometry filtering, data storage, retrieval and analysis, I developed a custom system using Python scripts and MySQL.

Mass Spectrometry, Data Filtering and Database Storage

Sequential database searches were made using the NCBI RefSeq Murine database containing a reverse dummy protein database that assessed the statistical significance of the results. A consensus file for each mouse sample was exported from BioWorks. The data was parsed using custom-designed Python scripts and then stored in a local MySQL database for easy retrieval and analysis.

The database was constructed to allow storage of the data in the most granular fashion. All data elements from the BioWorks output were stored and indexed. The peptide database was keyed on raw peptide sequence. A secondary peptide table was created after replacing all isoleucine residues with leucine residues, as MS cannot distinguish between these two. Also, the peptide sequences stored in this secondary table had all modifications removed. These two simplifications allowed for peptide matching more quickly and easily.

A script was written that allowed the following parameters to be easily varied: (1) peptide cutoff scores based on charge state and XCorr, (2) removing proteins identified by a single peptide, and (3) counting identical peptides with different charge states as unique peptides. By varying these parameters, the false-discovery rate was calculated for a variety of conditions, and the most appropriate parameters were chosen for subsequent data analysis.

Protein Rank List Creation

Given a list of peptides, there are many possible protein lists that could be constructed to explain the list of peptides. I chose the approach of finding the protein list that was most parsimonious. In other words, if three peptides could belong to three proteins or all belong to the same protein, the most parsimonious choice is that they all belong to the same protein. In general, when there are multiple possibilities for a peptide match, the choices are usually from the same family of protein, so this assumption seems to be a safe and valid one.

A script was written that would query each peptide against the most recent Internation Protein Index (IPI) murine database (http://www.ebi.ac.uk/IPI). Every possible protein match was tallied. Once all peptides were queried, the most parsimonious non-redundant protein list was constructed, and the final list was saved.

This list of proteins was utilized to create two matrices for subsequent analysis. Each is a list of non-redundant protein matches with the corresponding twenty-mouse samples. In the first matrix, the protein was scored as either present or absent in that sample (0 or 1). In the second matrix, for each mouse sample, the number of peptides used to identify that particular protein was tallied.

Protein rank lists were constructed by ordering the proteins by two different criteria. In the protein matrix built using the on-off status of the protein, the list was ranked using a frequency difference ratio (FDR), calculated simply as the normalized difference of the frequency of each protein in each class. The number of times a particular protein occurred in the wild-type samples was subtracted from the number of times that protein occurred in the samples from mice with cancer, and the absolute value of this difference was divided by the number in each class. In the second matrix, where the peptide contribution to each protein was tallied, the proteins were ranked according to a Mann-Whitney U statistic. The rank lists of proteins were saved for subsequent analysis.

Hierarchical Clustering and Principal Component Analysis

Principle component analysis (PCA) is a common method used to quantify the dominant global variance structures in high dimensional feature space (eg. proteins or peptides). Using subsets of varying size, PCA was applied to the rank list of proteins. Although all data manipulation was performed using Python scripting, the PCA and hierarchical clustering (HC) was performed in Matlab (http://www.mathworks.com).

Prior to analysis, the data was first normalized to a mean of zero and a standard deviation of one. Hierarchical clustering and PCA were first performed in an unsupervised manner. Subsequently, the same analysis was performed on a subset of the top-ranked proteins.

Naive Bayes Classifier

In order to determine if one class of mice could be distinguished from the other class based only on their MS profiles, a Bayesian classifier was developed and tested. Among the supervised learning methods, Bayesian analysis is commonly employed for very-high dimensional data sets. A classifier is trained on a subset of the data and used to predict the class of unknown samples. Bayesian analysis was accomplished using the Biopython Module for Python (http://www.biopython.org). The protein rank list based on peptide tallies was used for this analysis. A Bayesian classifier was trained on the data. Cross validation was performed by leaving out 10% of the samples on each run and running all possible combinations. In other words, for 20 samples, all 190 combinations of 18 selected from 20 were used to train the classifier and the remaining two were tested. The results from all these trials were averaged. This analysis was repeated starting with the most differentially expressed protein and adding in proteins one at a time from the rank list, until the group contained the top 100 proteins. Finally, a greedy-search algorithm was used to determine a minimal discriminating set of proteins. A script was employed that tested every combination of three proteins from the top fifty (19,6000 sets). This analysis was performed to determine a minimally discriminating set of proteins that could be used to distinguish one class from the other.

2.5 Results

2.5.1 Characterization of Proteins

Peptide identifications were made according to the following criteria: Delta correlation (ΔCN) ¿ 0.1, and Ranking of Primary Score (Rsp) = 1. The Cross Correlation Score (XCorr) varied based on charge state and several ranges were studied. In all cases, the protein identifications were optimized to create the most parsimonious list of proteins. Table 2.1 shows several samples, and in each case, the XC cut-off varied with the charge state. Also varied was whether or not multiple charge states for the same peptide were considered unique identification.

Cut-Offs	Charge Unique?	1	>1	>2	>3	FPR
1.8/2.0/2.5	Y	1111	368	249	206	2.5%
1.8/2.0/2.5	N	1143	336	227	188	2.5%
2.0/2.5/3.0	Y	762	328	234	201	1.7%
2.0/2.5/3.0	N	789	301	224	182	1.7%
2.5/3.0/3.5	Y	392	273	216	183	0.9%
2.5/3.0/3.5	N	408	257	202	166	0.9%

Table 2.1: **Filtering Criteria and False-Positive Rate**
Results are shown for several sets of filtering criteria. The cut-offs refer to the cross correlation score based on a charge state of +1/+2/+3, respectively. Indicated is whether or not the same peptide with different charge states was considered unique. The number of proteins identified by one, more than one, more than two, and more than three peptides are shown. Finally, the false-positive rate for the given filtering criteria is shown. Further details can be found in the text.

The list of peptides found was used to create a list of proteins. As described above, each peptide was searched against the most recent IPI Murine Database. The protein possibilities were tallied, and the most parsimonious list of proteins to explain the data was created. The number of proteins identified by one unique peptide as well as the number identified by two or more, three or more, or four or more peptides are also shown. Finally, the false positive rate was calculated by dividing the number of reverse hits by the total number of peptide rows. For XC cut-offs of 1.8, 2.0, and 2.5

for charge states +1, +2 and +3, respectively, the false positive rate was 2.5%. This rate decreased to 0.9% for cut-offs of 2.5, 3.0, and 3.5. But even though the false positive rate decreased substantially, so did the number of peptides identified. Based on a false positive rate of 2.5%, as well as published cut-offs for other studies, the parameters chosen for all subsequent analysis were XC cut-offs of 1.8, 2.0, and 2.5 for charge states +1, +2 and +3, respectively.

2.5.2 Protein Rank List

Two protein rank lists were created. In one list, each protein was scored as one or zero based on whether or not a peptide was present in a particular sample. For each protein, a frequency difference ratio statistic was calculated as described above. This statistic is a measure of the differential expression of that protein between the two classes (cancer, normal). The protein list was sorted, and the top 50 proteins are shown in Table 2.2 on page 37. For the second list, each protein was scored by tallying the number of peptides contributing to the identification of that protein in each sample. A Mann-Whitney U statistic was used to calculate the likelihood that the two populations were different, and the protein list was sorted according to the score. The top 50 most differentially expressed proteins (by this metric) are shown in Table 2.3 on page 38. There were nine proteins found unique to the mice with cancer and ten proteins unique to the control mice (Table 2.4 on page 39).

2.5.3 Hierarchical Clustering and PCA

For each protein, the number of peptides assigned to that protein were tallied in each sample. The data was normalized to a mean of zero and a standard deviation of one. These data were used for subsequent clustering and PCA. Using Matlab, hierarchical clustering and PCA were first performed on the entire list of proteins (see Figure 2-1 on page 34). The distance metric used was correlation, and the linkage method was average. Then, a Mann-Whitney U test was used to rank the proteins based on their peptide tally in one class (cancer) vs. the other class (wild-type). The top 50 proteins were subjected to hierarchical clustering and PCA as above (Figure 2-2 on page 35).

2.5.4 Naive Bayes Classifier

One of the goals of this study is to determine if it is possible to distinguish the mice with cancer from the normal controls. Using the peptide tally protein rank lists, a Bayesian classifier was trained on the data, and the accuracy of the classifier was determined by leave-one-out cross-validation, as described in the Methods. The classifier was trained on an increasing number of top-ranked proteins, and the results are shown in Figure 2-3 on page 36. Using only the top-ranked protein yields a predictive accuracy of more than 80%. Adding in sequential top-ranked proteins increases the predictive accuracy to as high as 95% (for 30 proteins). Adding in further proteins does not improve the accuracy. In fact, the accuracy declines as lower-ranked, uninformative proteins are added into the classifier. For comparison, two runs of randomly selected proteins are shown, and as expected, average predictive accuracy is about 50% (chance).

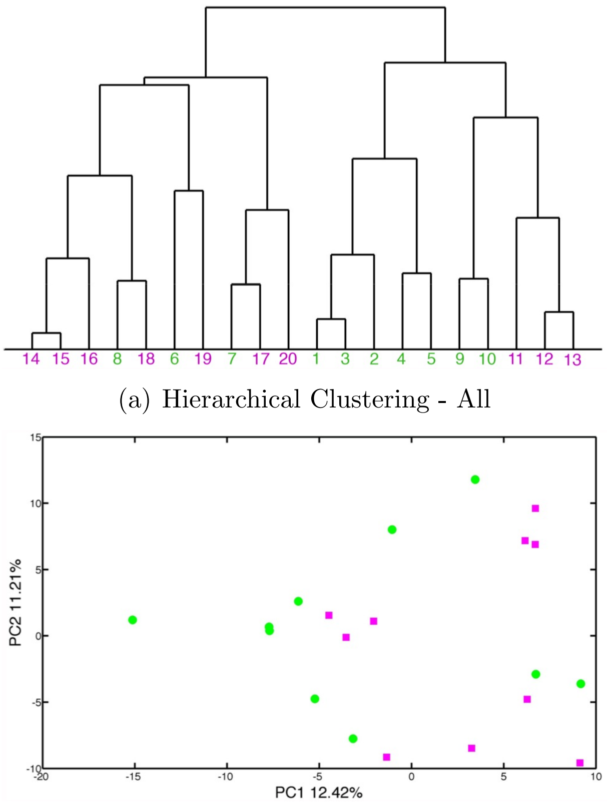
Since the linear combination of features determines the accuracy of the classifier, the top-ranked most differentially expressed proteins may not yield the best classifier. Therefore, a minimal discriminating set of proteins was sought. For the top-ranked 50 proteins, every set of three were chosen (19,600 sets in total) and a Bayesian classifier was trained and tested. The top performing sets of proteins are shown in Figure 2.5 on page 40. Nine sets of three proteins were found that, when used to train a Bayesian classifier, yielded a predictive accuracy of 98.9% or better. The top-ranked protein, intestinal maltase-glucoamylase was found in all sets. But the other two proteins existed throughout the top 50 proteins. It is important to note that the proteins themselves in each of the minimal discriminating sets are not necessarily important in the genesis, maintenance or metastasis of colon cancer. It is the linear combination of these proteins that give the classifier its predictive power. Nevertheless, several of these proteins merit further validation and study.

2.6 Discussion - Mouse Models of Colon Cancer

Treatment advances have helped increase survival in colorectal carcinoma. However, the greatest gains will likely be made through early detection of occult malignancy. A direct correlation exists between survival and early detection, and any technologies that facilitate early identification will lead to increased survival. These experiments demonstrate that mass spectrometry can be used to reliably differentiate plasma samples from mice with and without colorectal carcinoma.

These experiments demonstrate that it is relatively easy to train a classifier to identify the class of an unknown sample. Because of the paucity of samples, cross validation was used to predict the accuracy of the classifier. A greedy search algorithm was used to find the minimal discriminating set of proteins, and results indicate that using as few as three proteins is enough to predict class with 99% accuracy. Further study will be needed to test whether or not these findings generalize to other ApcΔ580 mice, and eventually, human subjects

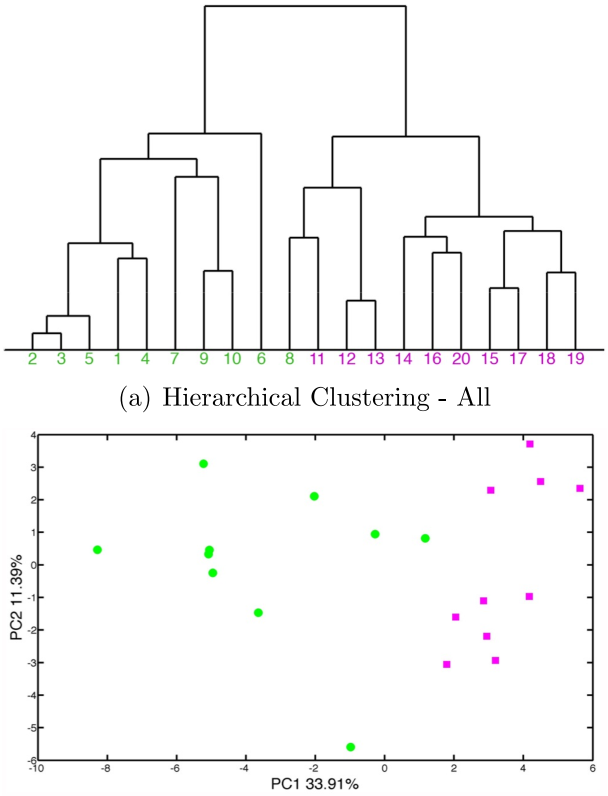
The ultimate goals of these studies are two-fold. First is to identify novel proteins and pathways in tumorigenesis. A second goal is to increase our ability to diagnosis colorectal carcinoma at an earlier stage. Serum testing as a diagnostic marker for occult malignancy will be difficult because of the low abundance of tumor-secreted proteins in the blood. Tests that are not sensitive enough will miss a significant proportion of tumors, while those that are not specific enough will lead to unnecessary invasive confirmatory diagnostic procedures. Using mass spectrometry profiles to identify individuals with early-stage colorectal carcinoma is an attractive prospect. While these results are encouraging, studies will need to be extended to human samples. Ultimately, a diagnostic test could be developed that may quickly and accurately detect malignancy long before it would become clinically evident.



(a) Hierarchical Clustering - All

(b) PCA - All

Figure 2-1: **Hierarchical Clustering and PCA of All Proteins**
Peptides were filtered at XC cutoffs of 1.8, 2.0, and 2.5 for charge states +1, +2 and +3, respectively. Peptides with different charge states were considered identical. Single peptide identifications were excluded. For each protein, the number of peptides assigned to that protein were tallied in each sample and the data normalized to a mean of zero and a standard deviation of one. Using Matlab, hierarchical clustering was performed on the entire list of proteins with distance metric correlation and the linkage method average (Panel a). A graph of PC1 vs. PC2 is shown in (Panel b). The percent variance along each principal component is given for PC1 and PC2. Samples 1-10 represent the animals with cancer (green) and samples 11-20 the wild-type animals (magenta).



(a) Hierarchical Clustering - All

(b) PCA - All

Figure 2-2: **Hierarchical Clustering and PCA of Top 50 Proteins**
The protein list was created as in Figure 2-1. Proteins were ranked by a Mann-Whitney U test based on the peptide tally in each class (cancer vs. wild-type). The top 50 proteins were used for this analysis. Using Matlab, hierarchical clustering was performed on the list of 50 proteins with distance metric correlation and the linkage method average (Panel a). A graph of PC1 vs. PC2 is shown in (Panel b). The percent variance along each principal component is given for PC1 and PC2. Samples 1-10 represent the animals with cancer (green) and samples 11-20 the wild-type animals (magenta).

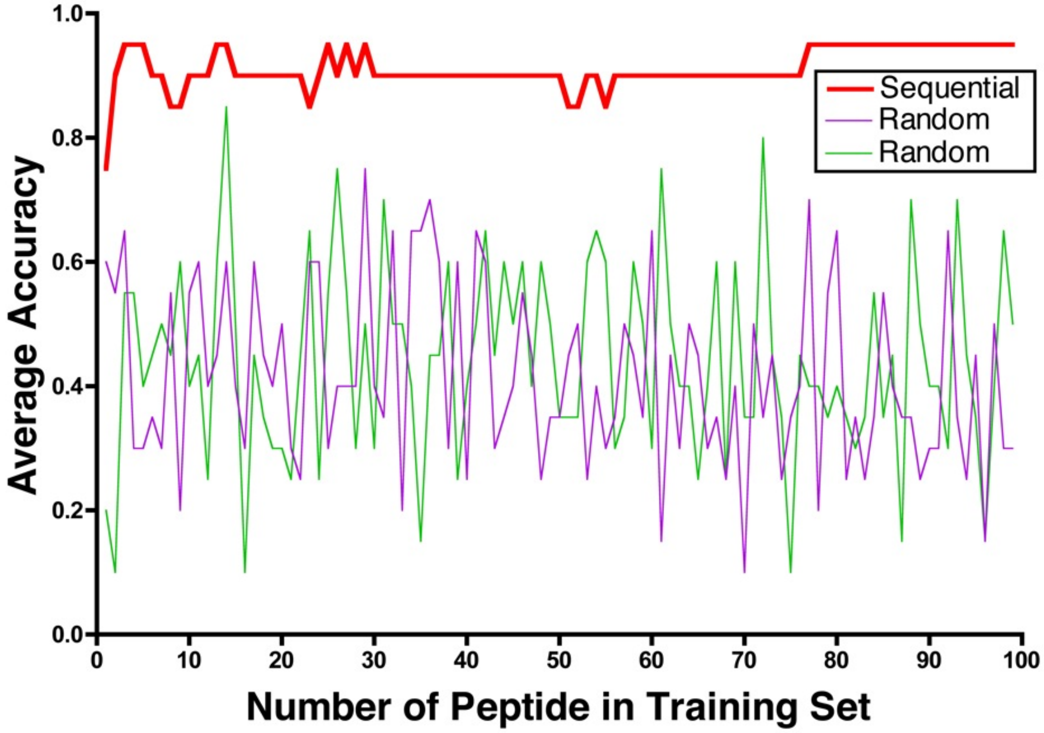


Figure 2-3: **Naive Bayes Classifier**
In order to predict the class of an unknown sample, a naive Bayes classifier was trained on the data. For testing, a leave-two-out analysis was used, in which all permutations of two samples were left out and the classifier was trained on the remaining data. The naive Bayes classifier was trained on the protein list of tallied peptides ranked by Mann-Whitney z-score. An increasing number of proteins were used to train the classifier, and the data are displayed in the graph (top line). For comparison, two runs of randomly selected proteins are shown. While even a small number of proteins predicts with high accuracy, increasing the number of features does not improve the accuracy beyond 95%.

RefSeq	Name	Expression
14861842	G7e protein	Up
94394534	PREDICTED: similar to Ig heavy chain V region 102 precursor	Down
51873060	eukaryotic translation elongation factor 1 alpha 1	Up
84871986	glutathione peroxidase 1	Down
34328108	procollagen type I alpha 1	Down
31981822	cystatin C	Up
71361676	complement factor H-related protein B	Down
47059073	thrombospondin 1	Up
22122483	epidermal growth factor-containing fibulin-like extracellular matrix protein 1	Down
7242197	proteasome (prosome macropain) subunit beta type 1	Down
6755995	WD repeat domain 1	Down
94378746	PREDICTED: similar to Ig kappa chain V-V region L7 precursor	Down
94377281	PREDICTED: maltase-glucoamylase	Down
94378711	PREDICTED: similar to Ig kappa chain V-V region MOPC 41 precursor	Down
6679509	parotid secretory protein	Down
6754698	multiple inositol polyphosphate histidine phosphatase 1	Down
7657429	osteoblast specific factor 2 (fasciclin I-like)	Down
7363449	serum amyloid P-component	Up
94378742	PREDICTED: similar to Ig kappa chain V-IV region S107B precursor	Up
6679749	fibroblast activation protein	Down
21359820	myoglobin	Up
94394520	PREDICTED: similar to Ig heavy chain V-I region V35 precursor	Up
94393196	PREDICTED: similar to Ig heavy chain V region M167 precursor	Down
6679813	FMS-like tyrosine kinase 4	Down
6753966	glycerol-3-phosphate dehydrogenase 1 (soluble)	Down
94378734	PREDICTED: similar to Ig kappa chain V-VI region XRPC 44	Down
111120329	procollagen type I alpha 2	Down
94376463	PREDICTED: similar to Alpha-fetoprotein precursor (Alpha-fetoglobulin) (Alpha-1-fetoprotein)	Down
31982260	insulin-like growth factor binding protein 2	Up
94399993	PREDICTED: similar to zinc finger protein 36-like 3	Up
22094119	myosin XVIIIa	Up
22129037	olfactory receptor 167	Up
30840990	Rho GTPase activating protein 24 isoform 1	Up
15082218	secreted phosphoprotein 24	Up
63543414	PREDICTED: similar *	Up
6754390	inositol 145-triphosphate receptor 1	Up
33859482	eukaryotic translation elongation factor 2	Up
94384435	PREDICTED: similar to Protein C16orf7 homolog (5-day ovary-specific transcript 1 protein) isoform 15	Down
94419495	PREDICTED: similar to Ig heavy chain V region VH558 A1/A4 precursor	Down
7656969	complement factor H-related protein	Up
111607471	amylase 2-1 pancreatic	Down
6754524	lactate dehydrogenase 1 A chain	Up
94378692	PREDICTED: similar to Ig kappa chain V-II region RPMI 6410 precursor	Down
8850219	haptoglobin	Up
6753220	complement component 1 q subcomponent B chain	Up
6755040	profilin 1	Down
63543420	PREDICTED: similar *	Up
27370126	carboxylesterase 5	Down
111074529	procollagen type XII alpha 1	Down
6755821	C-type lectin domain family 3 member b	Down

Table 2.2: **Top 50 proteins as determined by the Frequency Difference Ratio** The on-off status of each protein was tallied for each sample. Any proteins identified by a single peptide were excluded. The Frequency Difference Ratio (FDR) was calculated by subtracting the number of 'on' samples in the normal group from the number of 'on' samples in the cancer group and dividing by the number of samples in each group. The proteins were then ranked by absolute value of the FDR. For expression, 'Up' refers to increased in the mice with cancer, and 'Down' refers to increased in the control mice.

RefSeq	Name	Expression
94378500	PREDICTED: similar to Maltase-glucoamylase intestinal	Down
11596855	transferrin receptor	Up
94393200	PREDICTED: similar to Ig heavy chain V region 441 precursor	Down
31982300	hemoglobin beta adult major chain	Down
6671650	complement component 1 q subcomponent A chain	Down
9055252	inter alpha-trypsin inhibitor heavy chain 4	Up
67010045	hypothetical protein	Up
14861842	G7e protein	Up
6680175	hemoglobin alpha 1 chain	Down
6680496	inter-alpha trypsin inhibitor heavy chain 3	Up
7363449	serum amyloid P-component	Up
115430101	complement component 4 binding protein	Up
33859636	serine (or cysteine) proteinase inhibitor clade A member 3K	Down
38348520	hypothetical protein	Down
7304875	alpha-2-HS-glycoprotein	Up
94394534	PREDICTED: similar to Ig heavy chain V region 102 precursor	Down
63629958	PREDICTED: apolipop*	Up
33859506	albumin 1	Down
6753822	fibulin 1	Down
94378692	PREDICTED: similar to Ig kappa chain V-II region RPMI 6410 precursor	Down
13624321	coagulation factor XIII beta subunit	Down
6680608	pregnancy zone protein	Down
6679182	orosomucoid 1	Up
30578393	coagulation factor XIII A1 subunit	Down
110347473	apolipoprotein A-IV	Up
51873060	eukaryotic translation elongation factor 1 alpha 1	Up
84871986	glutathione peroxidase 1	Down
7304911	alpha-2-glycoprotein 1 zinc	Down
15011841	glutathione peroxidase 3	Down
110347564	ceruloplasmin isoform b	Down
59709439	serine (or cysteine) proteinase inhibitor clade D member 1	Up
34328108	procollagen type I alpha 1	Down
16418335	leucine-rich alpha-2-glycoprotein	Up
110347406	complement component factor i	Up
6754132	histocompatibility 2 Q region locus 10	Up
8850219	haptoglobin	Up
6671501	apolipoprotein C-IV	Up
31981822	cystatin C	Up
71361676	complement factor H-related protein B	Down
136429	RYP PIG TRYPSIN PRECURSOR	Up
6753220	complement component 1 q subcomponent B chain	Up
94378742	PREDICTED: similar to Ig kappa chain V-IV region S107B precursor	Up
6678093	serine (or cysteine) proteinase inhibitor clade A member 3N	Up
47059073	thrombospondin 1	Up
22122483	epidermal growth factor-containing fibulin-like extracellular matrix protein 1	Down
7242197	proteasome (prosome macropain) subunit beta type 1	Down
6755995	WD repeat domain 1	Down
6679749	fibroblast activation protein	Down
6681257	extracellular matrix protein 1	Down
6754384	inter-alpha trypsin inhibitor heavy chain 2	Up

Table 2.3: **Top 50 Proteins as Determined by the Mann-Whitney U Test** For each protein, the number of peptides found matching that protein were tabulated for each sample. The Mann-Whitney U test was used to determine which proteins are most differentially expressed. For expression, 'Up' refers to increased in the mice with cancer, and 'Down' refers to increased in the control mice.

Found Only in Cancer	
GI	Name
22094119	myosin XVIIIa
22129037	olfactory receptor 167
30840990	Rho GTPase activating protein 24 isoform 1
6753556	cathepsin D
9507247	interleukin 17b
9789931	diaphanous homolog 3
94364089	PREDICTED: similar to ciliary rootlet coiled-coil rootletin
94384421	PREDICTED: similar to Zinc finger protein 469
6679373	serine (or cysteine) proteinase inhibitor clade E member 1
Found Only in Normal	
GI	Name
7710010	cartilage oligomeric matrix protein
6678189	seminal vesicle secretion 5
6754360	insulin receptor
6755198	proteasome (prosome macropain) subunit alpha type 6
21389311	intercellular adhesion molecule
33468869	seminal vesicle protein 2
63485066	PREDICTED: similar to translocated promoter region protein isoform 3
94390964	PREDICTED: hypothetical protein LOC69926
6753966	glycerol-3-phosphate dehydrogenase 1 (soluble)
6755995	WD repeat domain 1
84871986	glutathione peroxidase 1

Table 2.4: **Proteins Found in One Population Only**
The proteins found exclusively in either the mice with cancer or the normal mice are shown. If a protein was identified by a single peptide, it was excluded.

Accuracy	GI (rank)/Name	GI (rank)/Name	GI (rank)/Name
99.7	94378500 (1) Intestinal Maltase-Glucoamylase (predicted similar)	1873060 (25) Eukaryotic Translation Elongation Factor	30578393 (23) Coagulation Factor XIII A1 Subunit
99.4	94378500 (1) Intestinal Maltase-Glucoamylase (predicted similar)	3057893 (23) Coagulation Factor XIII A1 Subunit	110347473 (24) Apolipoprotein A-IV
99.4	94378500 (1) Intestinal Maltase-Glucoamylase (predicted similar)	51873060 (25) Eukaryotic Translation Elongation Factor	6679749 (47) Fibroblast Activation Protein
99.2	94378500 (1) Intestinal Maltase-Glucoamylase (predicted similar)	6671650 (4) Complement Component 1 q	51873060 (25) Eukaryotic Translation Elongation Factor
99.2	94378500 (1) Intestinal Maltase-Glucoamylase (predicted similar)	30578393 (23) Coagulation Factor XIII A1 Subunit	71361676 (38) Complement Factor H-Related Protein B
99.2	94378500 (1) Intestinal Maltase-Glucoamylase (predicted similar)	110347473 (24) Apolipoprotein A-IV	31981822 (37) Cystatin C
99.2	94378500 (1) Intestinal Maltase-Glucoamylase (predicted similar)	51873060 (25) Eukaryotic Translation Elongation Factor	7304911 (27) Alpha-2-Glycoprotein 1 Zinc
99.2	94378500 (1) Intestinal Maltase-Glucoamylase (predicted similar)	7304875 (12) Alpha-2-HS-Glycoprotein	6753822 (17) Fibulin 1
98.9	94378500 (1) Intestinal Maltase-Glucoamylase (predicted similar)	6671650 (4) Complement Component 1 q	110347564 (30) Ceruloplasmin Isoform b

Table 2.5: **Naive Bayes Minimal Discriminating Set**
In order to find a minimal discriminating set of proteins, a greedy algorithm was used. For the top-ranked 50 proteins, every possible combination of three proteins was tested using the leave-two-out analysis. The top ranking sets of proteins are displayed, along with the average accuracy of the classifier, and are all more than 98.9%.

Chapter 3

Proteomics for Cancer Biomarker Discovery

3.1 Cancer is a Significant Cause of Mortality

Between 1974 and 1992, the incidence of cancer in the United States rose from 400 cases per 100,000 to 510 cases per 100,000 Ries; 2007aa. Since then the rate has fallen to 457 per 100,000 (2004), still well above the incidence in 1974. Despite apparent advances in diagnosis, treatment and ancillary care, death rates continued to climb from 199 per 100,000 in 1974 to a high of 215 per 100,000 in 1991. From 1991 until 2004, the mortality rate fell slightly to 186 per 100,000. Cancer is the leading cause of death in those aged 45-64 and the second leading cause of death in both age groups 35-44 and greater than 65 (http://www.CDC.gov). It is the third leading cause of death in young adults. In 2004, cancer was responsible for 1.1 million deaths in the US, making it the second leading cause of death behind heart disease.

3.2 Early Detection Improves Outcome

More than 60% of patients with breast, colon, lung and ovarian cancer have metastatic disease at presentation, severely limiting the success of conventional therapeutics [42]. For example, 2/3 of cases of ovarian cancer are detected only after the disease has spread beyond the peritoneal cavity [36]. Despite this, women usually have little or no specific diagnostic symptoms. With advanced disease, 5-year survival rates are just 35-40%, while survival rates are over 95% when disease is detected when it is still confined to the ovary. Clearly, early detection of ovarian cancer would have a profound impact on the survival of this disease.

3.3 Proteomics for Early Cancer Detection

As described above, work published in 2002 demonstrated that mass spectrometry could be used to distinguish between the sera of two groups of women - those with ovarian cancer and those without. Subsequently, it was found that this early exuberance was premature, and that the validity of the initial data had suffered from bias and chance. Nevertheless, proteomics is emerging as a significant tool in both the early detection of cancer as well as in biomarker discovery.

3.4 Lessons from the Current Work

I describe above a mouse model of colorectal carcinoma. Ten mice with colon cancer were compared to ten of their wild-type litter mates. Their plasma was collected and separated by liquid chromatography and subjected to mass spectrometry to identify proteins. Several thousand peptides were identified in each class (cancer, wild-type), and these peptide lists were used to create non-redundant protein lists. Based on the peptide tally for these proteins, a Bayesian classifier was developed that could reliably distinguish between the cancer and non-cancerous state. Furthermore, several sets of just three proteins were identified that, when used to train a Bayesian classifier, yielded 99% accuracy. This experiment demonstrates some of the pitfalls in designing and implementing a proteomic study, and these are outlined below.

3.4.1 Bias

Bias is broadly defined as "the systematic erroneous association of some characteristic of a group in a way that distorts a comparison with another group." [33] It is difficult to control for bias, and many factors are inherent to the study design [33]. In the current work, there are many opportunities for bias to have been introduced. For instance, the mice sera was collected on two different days and in two batches. These two batches were separated and run on the MS separately. In fact, while unsupervised analysis did not show a natural separation of cancer from non-cancer mice, it did show some separation based on the day on which the study was performed (data not shown). This indicates that there was a significant amount of bias, and that this may have impacted the validity of the results.

In order to control for bias as much as possible, great care must be taken to standardize as many aspects of the experiment as possible. The way in which the mice are handled and sacrificed must be consistent. The way in which the plasma is isolated and frozen must be standardized. More importantly, the sample preparation, liquid chromatography and MS must all be done in a strictly designed fashion, minimizing the variability of each step. Each of the steps from sacrifice to MS can add an element of bias to the final result, and in the current work on colorectal carcinoma, there appears to have been a significant amount of such variability.

3.4.2 Chance

If a study has insufficient power, there can be Type I (false-positive) or Type II (false-negative) errors. But the application of multi-variable models to proteomics (and genomics) introduces another type of chance error-overfitting. When a model is designed to discriminate between two classes, and this model fits the data perfectly, this is likely a case of overfitting. This is a common problem with high-dimensional data, and it occurs when the classifier is ”over trained” for the data. When the classifier is applied to a separate new set of data, it will likely have no discriminatory ability. To test for overfitting, a validation set is preferred. In this experimental design, a separate set of data is ”held back” from the primary analysis. When the analysis is complete and the classifier is derived, it is applied to the separate validation set of data. If the accuracy is the same, then the result is more likely to be valid.

In the current study, there was no validation set, and this significantly weakened the validity of the result.

3.4.3 Peptide and Protein Identification

Proteomics is still an emerging field, and many aspects of these experiments remain to be standardized. The current work illustrates some of the pitfalls in designing and implementing these types of studies.

Animal Husbandry

Ideally, mice from both classes (cancer, wild-type) will be handled and raised in an identical manner. But since the Δ580 mice harbor a germ-line mutation, it is possible that their proteome is different at baseline, irrespective of their tumor burden. This type of bias is very difficult to control. One way to study this would be to repeat the experiment on mice before they have developed tumors.

Sample Isolation

As described above, the mice were sacrificed and bled into tubes containing EDTA. Then the blood was centrifuged to isolate the plasma. The samples were stored at -20C for one to two hours before transferring to -80C. Although this appears to be a consistent and reasonable way to isolate and store plasma, several variables could be responsible for bias and should be optimized and controlled. Any variability in handling time could account for an increased amount of proteolysis. If the samples from mice with cancer are handled any differently (for instance, spending more time on ice while the intestine is examined), then those samples may have more proteolysis and thus will exhibit a different proteome. A small pilot study could be performed to show that samples from both classes are handled the same way and display internal validity.

Protein Digestion and Separation

Again, any variability in the digestion of separation of the samples could lead to bias. Care should be taken to ensure that samples from both classes are handled in exactly the same manner (eg. same amounts of trypsin, same temperature, same digestion time).

Mass Spectrometry

The dynamic range of proteins in serum is 10^{10} [23]. It is likely that any candidate biomarker has been severely diluted in the plasma from its original localized concentration. To overcome this problem, there are several possible solutions as outlined below.

Depletion First, prior to MS, the sample may be depleted. For instance, albumin and immunoglobulin represent a significant amount of the protein in a plasma proteome. If these are removed, then the resolution of the lower-abundance proteins may improve. A criticism of this technique is that any depletion strategy may also remove low-abundance proteins as well, especially if they are using albumin as a carrier protein.

Detection in Primary Tumor Another strategy is to use first use the primary tumor source for biomarker detection and then to attempt to find this biomarker in the plasma sample. This was the method employed by Ding, et al. in their study of cytokeratin 19 in hepatocellular carcinoma [12]. They first showed differential expression in tumor cell lines and subsequently demonstrated detectability in serum samples.

Increased Fractionation One can increase dynamic range prior to MS by performing more extensive fractionation. In the current study, each sample was split into 12 fractions using cation-exchange chromatography. An analysis by Hanash, et al. of the same samples was performed, first pooling and then fractionating the samples into 240 fractions. As expected, their analysis identified more low-abundance proteins (personal communication, data not shown). The main drawbacks of this approach are: (1) it takes much longer to analyze the samples, and (2) it requires much more sample.

Enrichment Strategies An alternative is to target a subset of proteins through enrichment. For instance, targeting an amino acid residue found only in a subset of peptides (eg. cysteine), thus decreasing the overall complexity of the sample [37]. This is often accompanied by isotopic labeling to facilitate quantification. Another method of enrichment is to target phosphorylated peptides [44] or specific N-terminal labeling [17]. One more alternative method is to specifically target the glycoprotein sub-proteome (glycoproteome), since the lower abundance proteins in plasma often arise from outer membrane shedding events [43]. All of these enrichment strategies are imperfect. In some cases, molecules of interest may not be in the enriched fraction. In the current study, there was no enrichment. It is likely that most lower-abundance proteins were not resolved because of the preponderance of high-abundance proteins.

Quantitative Strategies Although not employed in the current study, new methods are being adopted that allow for quantification of proteins from MS experiments. Commonly employed techniques include metabolic labeling using heavy amino acids, enzymatic transfer of ^{18}O from water to peptides or proteins via chemical reactions using isotope-coded affinity tags (reviewed in [1]).

Peptide Identification

One of the challenges for high-throughput experiments is to use databases with large numbers of MS spectra to derive a list of peptides and corresponding proteins [1]. For small datasets, researchers can manually validate each peak and peptide assignment, but this is not feasible for high-throughput work. Instead, incorrect peptide assignments can be removed using various filtering algorithms and by examining the number of hits using a reverse database [32, 30], the technique used in the current work. As the filtering criteria were made more stringent, the number of false hits went down, but this was at the expense of removing some potentially relevant peptides.

Bioinformatic Analysis

As is evident, once the data reaches the point of analysis, there is ample room for bias and chance to affect validity. Furthermore, care must be taken to not introduce further uncertainty into the analysis. A robust data storage and retrieval system is essential for performing reproducible experiments. Ideally, one has access to the raw MS spectra, as examining the spectra for a controversial peptide identification may be useful. In the current work, I developed a database for storage and easy retrieval of MS data, though the raw MS data was not included. A subsequent incarnation of this system should include storage of the raw spectra as well as built-in conversion utilities to open-source formats, such as mzXML.

Constructing the most parsimonious set of proteins from the peptide identifications can be a challenge. In the current work, the SEQUEST algorithm built a list of proteins for each sample. But after the twenty samples were run, it made sense to take advantage of the accumulated data in order to make a protein list that could most simply explain the set of peptides. A possible criticism of this method is that it might make better sense to develop the protein lists individually (for cancer and non-cancer). While this would be easy to implement, I chose to use the most parsimonious explanation for all the data.

Deciding how to tally the proteins is another important decision. Possibilities include analysis at the peptide level or tallying the on-off state of each protein. In addition to using the on-off state of each protein for each sample, I chose to use a tally of the peptide counts for each protein as a semi-quantitative measure. This decision is based on recent work by David States, et al., that showed that the number of peptide hits was proportional to the amount of protein in the mixture [39].

Choosing the appropriate metric for comparing classes is essential. For instance, using a T-statistic would not be appropriate because the data is not from a normal distribution. In the current study, a Mann-Whitney U test was used to compare the two populations, which is a non-parametric test for assessing whether two samples of observations derive from the same distribution.

It is essential to note whether or not an analysis is supervised or unsupervised. In the current work, unsupervised hierarchical clustering and PCA were performed, and unfortunately, did not show a good separation of the data into disease classes. Once the data was ranked, these analyses (now supervised) did show a separation.

Finally, and most importantly, a naive Bayes classifier was trained and used to differentiate with high accuracy the profiles of mice with cancer from those that were healthy. In any such high-dimensional analysis, one must immediately be suspicious of chance as a threat to validity. In this case, it is likely that such a high predictive accuracy is a result of over-training the classifier. A way to test this is to hold out a separate validation set of data and to test the classifier on this set. Unfortunately, as with many experiments, time and cost constraints did not permit this. Another way to help control for over-fitting is to use another artificial intelligence learning method, such as neural networks or support vector machines.

3.5 Future Directions

I plan on applying these techniques to the study of pediatric solid tumors. Initially, I will be studying neuroblastoma, the most common pediatric extra-cranial tumor. I will first perform proteomic analyses of neuroblastoma tissue culture cells followed by analyses of blood and tumors from patients with this disease. From this work, I will learn about the biology of neuroblastoma and develop criteria for early diagnosis and better risk stratification. Ultimately, this work will hopefully improve outcome by providing a method of evaluating disease response to novel therapies.

3.6 Summary

The field of proteomics holds great promise for biomarker discovery and early cancer detection. Currently, there is a need for standardization of everything from sample collection to protein digestion and separation to peptide and protein identification to bioinformatic analysis. There is a growing body of work to support such standardization, and in the coming years, we will likely see many of these areas develop standards of practice, much like the field of genomics saw during the last decade. Carefully planned proteomic analyses can be used for cancer biomarker discovery as well as for developing methods of early detection, a key to extending survival after cancer. Such methods will be applied to an increasing array of tumors, and the data will be a powerful orthogonal adjunct to the current genomic data sets.

3.7 Contributions

Through the current work, I have contributed the following.

- I have described the current field of proteomics as it relates to cancer biomarker discovery and early detection.
- I enumerated the challenges and pitfalls to developing early detection schemes for cancer based on high-dimensional proteomic analyses.
- I described a set of experiments on mice harboring a gene mutation predisposing them to colorectal carcinoma.
- I detailed the bioinformatic analysis of this data, including the development of a naive Bayes classifier to differentiate the cancerous state from the normal state.
- I discussed the caveats of the current work, in reference to the initial discussion on the challenges and pitfalls of early detection schemes and cancer biomarker discovery.

Bibliography

[1] R Aebersold and M Mann. Mass spectrometry-based proteomics. *Nature*, 422:198–207, Mar 2003.

[2] KA Baggerly, JS Morris, SR Edmonson, and KR Coombes. Signal in noise: evaluating reported reproducibility of serum proteomic tests for ovarian cancer. *Journal of the National Cancer Institute*, 97:307–9, Feb 2005.

[3] Y Ban, MC Wang, and TM Chu. Immunologic markers and the diagnosis of prostatic cancer. *The Urologic clinics of North America*, 11:269–76, May 1984.

[4] MS Boguski and MW McIntosh. Biomedical informatics for proteomics. *Nature*, 422:233–7, Mar 2003.

[5] FT Bosman. Prognostic value of pathological characteristics of colorectal cancer. *European journal of cancer (Oxford, England : 1990)*, 31A:1216–21, Jul 1995.

[6] P Boyle and JS Langman. Abc of colorectal cancer: Epidemiology. *BMJ (Clinical research ed)*, 321:805–8, Sep 2000.

[7] S Chuthapisith, R Layfield, ID Kerr, and O Eremin. Principles of proteomics and its applications in cancer. *The surgeon : journal of the Royal Colleges of Surgeons of Edinburgh and Ireland*, 5:14–22, Feb 2007.

[8] S Colnot, M Niwa-Kawakita, G Hamard, C Godard, S Le Plenier, C Houbron, B Romagnolo, D Berrebi, M Giovannini, and C Perret. Colorectal cancers in a new mouse model of familial adenomatous polyposis: influence of genetic and environmental modifiers. *Laboratory investigation; a journal of technical methods and pathology*, 84:1619–30, Oct 2004.

[9] RJ Davies, R Miller, and N Coleman. Colorectal cancer screening: prospects for molecular stool analysis. *Nature reviews. Cancer*, 5:199–209, Mar 2005.

[10] C Delahunty and JR Yates. Protein identification using 2d-lc-ms/ms. *Methods (San Diego, Calif)*, 35:248–55, Feb 2005.

[11] EP Diamandis. Analysis of serum proteomic patterns for early cancer diagnosis: drawing attention to potential problems. *Journal of the National Cancer Institute*, 96:353–6, Mar 2004.

[12] SJ Ding, Y Li, YX Tan, MR Jiang, B Tian, YK Liu, XX Shao, SL Ye, JR Wu, R Zeng, HY Wang, ZY Tang, and QC Xia. From proteomic analysis to clinical significance: overexpression of cytokeratin 19 correlates with hepatocellular carcinoma metastasis. *Molecular cellular proteomics : MCP*, 3:73–81, Nov 2003.

[13] L Fazio, M Cotterchio, M Manno, J McLaughlin, and S Gallinger. Association between colonic screening, subject characteristics, and stage of colorectal cancer. *The American journal of gastroenterology*, 100:2531–9, Nov 2005.

[14] F Fend and M Raffeld. Laser capture microdissection in pathology. *Journal of clinical pathology*, 53:666–72, Oct 2000.

[15] J Ferrucci and Unknown. Ct colonography for detection of colon polyps and cancer. *Lancet*, 365:1464–5; author reply 1465–6, Apr 2005.

[16] K Garber. Debate rages over proteomic patterns. *Journal of the National Cancer Institute*, 96:816–8, Jun 2004.

[17] K Gevaert, M Goethals, L Martens, J Van Damme, A Staes, GR Thomas, and J Vandekerckhove. Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted n-terminal peptides. *Nature biotechnology*, 21:566–9, Apr 2003.

[18] SP Gygi, GL Corthals, Y Zhang, Y Rochon, and R Aebersold. Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proceedings of the National Academy of Sciences of the United States of America*, 97:9390–5, Aug 2000.

[19] ET Hawk and B Levin. Colorectal cancer prevention. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 23:378–91, Jan 2005.

[20] KE Hung, AT Kho, D Sarracino, LG Richard, B Krastins, S Forrester, BB Haab, IS Kohane, and R Kucherlapati. Mass spectrometry-based study of the plasma proteome in a mouse intestinal tumor model. *Journal of proteome research*, 5:1866–78, Aug 2006.

[21] SI Hwang, DH Lundgren, V Mayya, K Rezaul, AE Cowan, JK Eng, and DK Han. Systematic characterization of nuclear proteome during apoptosis: a quantitative proteomic study by differential extraction and stable isotope labeling. *Molecular cellular proteomics : MCP*, 5:1131–45, Mar 2006.

[22] TF Imperiale. Can computed tomographic colonography become a ”good” screening test? *Annals of internal medicine*, 142:669–70, Apr 2005.

[23] JM Jacobs, JN Adkins, WJ Qian, T Liu, Y Shen, DG Camp, and RD Smith. Utilizing human blood plasma for proteomic biomarker discovery. *Journal of proteome research*, 4:1073–85, Aug 2005.

[24] A Jemal, RC Tiwari, T Murray, A Ghafoor, A Samuels, E Ward, EJ Feuer, MJ Thun, and Unknown. Cancer statistics, 2004. *CA: a cancer journal for clinicians*, 54:8–29, Feb 2004.

[25] A Keller, J Eng, N Zhang, XJ Li, and R Aebersold. A uniform proteomics ms/ms analysis platform utilizing open xml file formats. *Molecular systems biology*, 1:2005.0017, May 2006.

[26] KW Kinzler, MC Nilbert, B Vogelstein, TM Bryan, DB Levy, KJ Smith, AC Preisinger, SR Hamilton, P Hedge, and A Markham. Identification of a gene located at chromosome 5q21 that is mutated in colorectal cancers. *Science (New York, NY)*, 251:1366–70, Mar 1991.

[27] LA Liotta, M Lowenthal, A Mehta, TP Conrads, TD Veenstra, DA Fishman, and EF Petricoin. Importance of communication between producers and consumers of publicly available experimental data. *Journal of the National Cancer Institute*, 97:310–4, Feb 2005.

[28] J Listgarten and A Emili. Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Molecular cellular proteomics : MCP*, 4:419–34, Mar 2005.

[29] I Nishisho, Y Nakamura, Y Miyoshi, Y Miki, H Ando, A Horii, K Koyama, J Utsunomiya, S Baba, and P Hedge. Mutations of chromosome 5q21 genes in fap and colorectal cancer patients. *Science (New York, NY)*, 253:665–9, Aug 1991.

[30] GW Park, KH Kwon, JY Kim, JH Lee, SH Yun, SI Kim, YM Park, SY Cho, YK Paik, and JS Yoo. Human plasma proteome analysis by reversed sequence database search and molecular weight correlation based on a bacterial proteome analysis. *Proteomics*, 6:1121–32, Jan 2006.

[31] EF Petricoin, AM Ardekani, BA Hitt, PJ Levine, VA Fusaro, SM Steinberg, GB Mills, C Simone, DA Fishman, EC Kohn, and LA Liotta. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, 359:572–7, Feb 2002.

[32] WJ Qian, T Liu, ME Monroe, EF Strittmatter, JM Jacobs, LJ Kangas, K Petritis, DG Camp, and RD Smith. Probability-based evaluation of peptide and protein identifications from tandem mass spectrometry and sequest analysis: the human proteome. *Journal of proteome research*, 4:53–62, Feb 2005.

[33] DF Ransohoff. Bias as a threat to the validity of cancer molecular-marker research. *Nature reviews. Cancer*, 5:142–9, Feb 2005.

[34] DF Ransohoff. Lessons from controversy: ovarian cancer screening and serum proteomics. *Journal of the National Cancer Institute*, 97:315–9, Feb 2005.

[35] A Rauch, M Bellew, J Eng, M Fitzgibbon, T Holzman, P Hussey, M Igra, B Maclean, CW Lin, A Detter, R Fang, V Faca, P Gafken, H Zhang, J Whiteaker, J Whitaker, D States, S Hanash, A Paulovich, and MW McIntosh. Computational proteomics analysis system (cpas): an extensible, open-source analytic system for evaluating and publishing proteomic data and high throughput biological experiments. *Journal of proteome research*, 5:112–21, Jan 2006.

[36] L Ries, C Kosary, B Hankey, B Miller, A Harras, and B Edwards. Seer cancer statistics review, 1973-1998. *http://seer.cancer.gov/csr/1975_2004/*, 2001.

[37] M Shen, L Guo, A Wallace, J Fitzner, J Eisenman, E Jacobson, and RS Johnson. Isolation and isotope labeling of cysteine- and methionine-containing tryptic peptides: application to the study of cell surface proteolysis. *Molecular cellular proteomics : MCP*, 2:315–24, May 2003.

[38] JB Simon. Occult blood screening for colorectal carcinoma: a critical review. *Gastroenterology*, 88:820–37, Mar 1985.

[39] DJ States, GS Omenn, TW Blackwell, D Fermin, J Eng, DW Speicher, and SM Hanash. Challenges in deriving high-confidence protein identifications from data gathered by a hupo plasma proteome collaborative study. *Nature biotechnology*, 24:333–8, Mar 2006.

[40] LK Su, KW Kinzler, B Vogelstein, AC Preisinger, AR Moser, C Luongo, KA Gould, and WF Dove. Multiple intestinal neoplasia caused by a mutation in the murine homolog of the apc gene. *Science (New York, NY)*, 256:668–70, May 1992.

[41] L Wagner. A test before its time? fda stalls distribution process of proteomic test. *Journal of the National Cancer Institute*, 96:500–1, Apr 2004.

[42] JD Wulfkuhle, LA Liotta, and EF Petricoin. Proteomic applications for the early detection of cancer. *Nature reviews. Cancer*, 3:267–75, Apr 2003.

[43] H Zhang, XJ Li, DB Martin, and R Aebersold. Identification and quantification of n-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. *Nature biotechnology*, 21:660–6, May 2003.

[44] H Zhou, JD Watts, and R Aebersold. A systematic approach to the analysis of protein phosphorylation. *Nature biotechnology*, 19:375–8, Apr 2001.

[45] W Zhu, X Wang, Y Ma, M Rao, J Glimm, and JS Kovach. Detection of cancer-specific markers amid massive mass spectral data. *Proceedings of the National Academy of Sciences of the United States of America*, 100:14666–71, Dec 2003.