Chapter 21

PROTEOMICS FOR DIAGNOSTIC APPLICATIONS

The Convergence of Technology and the Resulting Challenges

Gordon R. Whiteley

National Cancer Institute - Frederick, Gaithersburg, MD, USA

Abstract: The emergence of proteomics in the post-genomic era has led to a resurgence in the study of and use of proteins for disease diagnosis. While the number of new protein markers has declined over the past 5 years, the use of new and exciting tools such as mass spectrometry, separation techniques and bioinformatics has fueled a search for markers and diagnostic patterns of markers that show great promise. The complexity of validation of these markers and panels of markers is a considerable challenge but the potential exists for the laboratory diagnosis of diseases for which there is no currently available lab test. This new era of proteomic diagnostics will continue to revolutionize our diagnostic arsenal and improve patient outcomes.

Keywords: Proteomics, biomarkers, mass spectrometry, bioinformatics

1. INTRODUCTION

The use of proteins for diagnosis of disease has been in practice for many decades. During the last half of the last century, numerous new markers for disease were discovered in blood. Originally, these were detected by immunological techniques such as immunodiffusion, radioimmunoassay (RIA), hemagglutination, precipitation techniques and enzyme linked immunosorbent assay (ELISA). There were several significant discoveries that advanced this field. The development of the monoclonal antibody by Kohler and Milstein¹ gave the specificity to diagnostic immunoassays that was only dreamed of in earlier years. Innovative techniques such as the use of fluorescent substrates and electrochemiluminescence have further enhanced sensitivity of immunoassays. Discovery of new proteins expanded

³³⁷

G. Spekowius and T. Wendler (Eds.), Advances in Healthcare Technology, 337-348. © 2006 Springer. Printed in the Netherlands.

the list of analytes being approved for use in the diagnostic lab. Some of these such as CEA were initially greeted with very high hopes for their diagnostic capabilities but proved to be less specific than originally hoped and their approved use has been limited. Others such as troponin have proven to be extremely powerful diagnostic tools.

Over the past few years, the discovery and approval of new biomarkers has diminished dramatically. Even with detection technologies that are sensitive the approval of new analytes for diagnosis by the FDA has declined² as shown in Figure 21-1.



Figure 21-1. The declining rate of introduction of new protein tests. The data are plotted to indicate the rate of introduction of new protein analytes in FDA-approved clinical tests (From: N.L. Anderson, N.G Anderson, *MCP* **1**(11), 845-67, 2002).

This decline could be due to several factors: the discovery of the high abundance disease markers is probably nearly complete, the diseases now being researched are more complex and will require a series of markers for diagnosis or we are looking in the wrong place in serum for the presence of these markers or a combination of the above factors.

In the past few years, there have been significant advances in many scientific areas that have led to the new field of proteomics. These advances in the areas of computer science and computing power, computerized databases and database management, separation technologies, biology, and mass spectrometry to name a few have come together to provide tools that are leading to a new wave in disease diagnosis and management. It is the convergence of these diverse technical areas that has demonstrated the

possibilities this technology can yield for disease diagnosis through an identification and understanding of the source and function of new disease markers previously overlooked because of their lack of sensitivity and or specificity or fraction in the serum or tissue.

The progress of the 13 year long human genome project³ led naturally to the study of the products of the genes – the proteins. Although some of the techniques used in proteomics had been developed years before and the term 'proteomics' had been coined originally in 1994, it was not until 1999 that the term became commonly used in literature⁴. Proteomics was defined as referring "to the study of the proteome using technologies of large-scale protein separation and identification". One popular separation technique had been described in the mid 1970's and this was 2 dimensional or 2D gels⁵. The use of 2D gels for protein separation followed by digestion of the proteins into fragments and then analysis of the fragments by mass spectrometry gave an identification of proteins through peptide mass fingerprinting. At the time, there was also talk of making an entire catalog of genes. However, the challenge remained to identify proteins that were expressed in disease and not in unaffected groups.

In the mid 1990's, the technique of laser capture microdissection was described⁶. This allowed for the separation of diseased cells from neighboring normal cells in tissue and a comparison of these two groups of cells could be done in order to determine if the protein content had changed and certain pathways had been triggered disease. The separation was done using a laser activated adhesive coated film that was placed over a tissue section. The cells were examined by a pathologist and diseased cells identified. A laser was then fired that activated the adhesive properties of the coated film and the cells were physically lifted from the slide. The process was then repeated on a second piece of adhesive film but normal cells were picked this time providing a control that matched the diseased cells in every way except for the protein content that had been activated.

Another key development at about the same time was the commercialization of the surface enhanced laser desorption ionization (SELDI) technology by Ciphergen (www.ciphergen.com). This technology is based on the fractionation of serum on the same surface that is used as the target for matrix enhanced surface desorption ionization (MALDI). The protein arrays consist of a metal surface that has been coated with active binding groups. The variety of groups (anionic exchange, cationic exchange, metal binding surfaces) allows for binding of different fractions from the serum and unbound fractions are then washed away. An energy absorbing matrix is then added directly to this surface and it is then used as the laser target in a mass spectrometer that has been specifically developed to handle these arrays. The surfaces are configured on arrays – 8 spots to an array and a bioprocessor holds12 arrays to give the configuration of a 96 well microtiter plate. This configuration is one familiar with biologists and one that has several robotic and other tools available for sample processing. The arrays are then fed into a low resolution but high sensitivity mass spectrometer that is easy to operate and has software that makes interpretation familiar with not only spectral views but also a 'gel' view that displays data in a manner familiar to biologists. Thus the tools available to biologists now included mass spectrometry.

2. DISCOVERING NEW APPLICATIONS

With these advances came an explosion of potential diagnostic tests analytes, test principles and procedures. One of the earlier publications described the combination of laser capture microdissection and SELDI to demonstrate differences between cancer cells and neighboring normal cells in tissue sections⁷. In this publication, tissue sections had been stained and were microdissected into a lysis buffer. An aliphatic reverse phase SELDI biochip was used to capture proteins from solubilized cancer cells or solubilized control normal cells. The proteins were then crystallized with matrix and mass spectrometry spectra were gathered. A comparison of the spectra from the tumor cells as compared to the normal cells clearly showed differences in up regulation and down regulation of several proteins. This was observed in several conditions including colon cancer, liver cancer, and in prostate cancer where differences were seen when comparing normal cells, prostatic intraepithelial neoplasia cells, tumor cells and stromal cells. The suggestion was made here that protein fingerprinting of early disease lesions was possible based on the differences seen in both spectral and gel views of these cells.

The next step in the process was to extend this finding into looking for these differences in serum. In 2002, the first publication to suggest that serum might indeed carry these differences and these could be observed by mass spectrometry was published⁸. In this publication, a series of sera from ovarian cancer patients and cancer free women were examined by SELDI and the resulting spectra were used to train a computer algorithm to recognize the differences between the two groups. The pattern was then used to classify 116 additional patient samples – 50 with cancer and 66 from cancer free women including several with benign disease such as ovarian cysts. The results of this study were remarkable giving a 100% sensitivity including 18 patients with stage I disease and 95% specificity. This study was followed by similar findings in a wide range of diseases including

21. Proteomics for Diagnostic Applications

prostate cancer⁹, breast cancer¹⁰, renal cancer¹¹, pancreatic cancer¹², and ovarian cancer¹³. In addition, there was also activity in detection of other diseases such as cardiac disease using this technique¹⁴.

However, there was much criticism of this method both from the bioinformatics side¹⁵ as well as the biological perspective¹⁶. The use of bioinformatic methods that were random such as genetic algorithms and thus did not always give absolute reproducibility of patterns was the basis of some criticism. Further refinement of raw spectrum (data processing) was thought to be primitive and needed improvement. Biologically, it was also felt that the source of these peptides that made up patterns was necessary in order to confirm that they were actually associated with the disease being studied. In addition, the reproducibility of the mass spectrometers themselves was at question and whether they could ever be robust enough to be used as a diagnostic device had yet to be answered.

As part of the investigation process and as biologists gained more experience and confidence in using mass spectrometry, a higher resolution instrument was evaluated. While this still used the SELDI-TOF principle and a source manufactured by Ciphergen, the instrument measuring the time of flight was the ABI Q-star which is a quadrapole time of flight instrument. This instrument is able to resolve individual peaks observed in the Ciphergen SELDI instrument into several peaks. Furthermore, the instrument's higher resolution was not subject to drift observed in the Ciphergen mass spectrometer. A repeat evaluation of the ovarian cancer study samples reported earlier was done using this system and the results gave a 100% sensitivity and specificity in the sample group tested¹⁷. While a genetic algorithm with a self-organizing map was used and several patterns were generated, there were at least 4 that gave this excellent result. Furthermore, the overlap of several ions in these patterns indicated their importance in the diagnostic test. The increase in the specificity would be important in further use of the technique because of the low incidence of ovarian cancer in the general population. While the 100% correlation would be ideal, it was noted that this was in a limited number of samples and would need to be further developed and validated in a much larger study in order to be considered as the basis for a diagnostic system.

The bioinformatics analysis of data has also received much attention. It is known that more than 300 groups have downloaded the data from the original Petricoin ovarian study done in 2002 and analyzed it in detail from several points of view. Many of these groups were able to reproduce the original results while some groups believed that the original analysis was flawed. Several suggestions have been made in terms of preprocessing data and analysis including the posting of data in its raw form¹⁸ to allow different

preprocessing methods to be properly evaluated. However, it was pointed out that communication between those analyzing the data and those who produced the data is essential because of the possibility of a mis-interpretation of results based on a mis-understanding of the purpose of the experiment¹⁹. Currently there are multiple data analysis and classification tools in use and being developed both commercially and in academic settings.

In addition to bioinformatics solutions that involve computer algorithms, data visualization tools have been and are continuing to be developed to allow for the handling of large data sets in a manner that can be analyzed by the human eye. The use of these tools can help reduce datasets to simplify the computer process or can be used in an iterative process with bioinformatics tools to confirm patterns as is shown in Figure $21-2^{27}$. In this example, the initial data is examined for quality and any spectra that are found to be of low total ion current for example are eliminated. The entire dataset consisting of all spectra from all patients are then imported into a visualization tool and are colorized by disease group. Using the visualization tool, certain areas are selected as showing discriminating characteristics through visual examination of the data in three dimensions: intensity, mass/charge value and disease group. These areas are then selected and fed to computer algorithms for the selection of individual discriminating values. These values are then confirmed by examining individual spectra from both the disease and normal groups. The process ensures the presence of a pattern rather than artifacts and has been valuable in studies where patterns are difficult to find using bioinformatics tools or where over-fitting can occur.

The next development in proteomics as a diagnostic tool was to determine the principle of the test. As part of this work, the source of diagnostic peptides needed to be reconciled. Basic questions had been raised: why such a small source as an early stage tumor could provide sufficient quantity in the blood by mass spectrometry and why were the small peptides not excreted in the urine as quickly as they accumulated? The answer lay in the discovery that diagnostic protein fragments and peptides appeared to be associated with large and abundant carrier proteins such as albumin²⁰. In this publication, the authors capture albumin out of serum, washed and then caused peptides to dissociate from the albumin through the use of 50% acetonitrile. After a 30mw size exclusion separation, the peptides were examined by mass spectrometry. What was found was a large number of low molecular weight peptides that had been associated with albumin. Identification of some of these peptides has shown their identity to be known linked to cancer such as P53 and BRCA1 and 2. It was proposed that the albumin's long half-life (19 days) acted as a concentrating and

protecting molecule carrying the peptides and preventing clearance by the kidneys²¹.



Figure 21-2. Data Visualization as an Aid to Data Analysis (From: D.Johann et al., *Ann N Y Acad Sci* **1022**, 299, 2004).

This discovery led to the enrichment for albumin in proteomics – the antithesis of that had been done previously. Commercialization of a biomarker enrichment method using the principle of albumin capture has recently been done by PerkinElmer in collaboration with Viva Sciences (www.perkinelmer.com). This method has been optimized as a discovery tool in a 96 well format including albumin capture, elution and dissociation of peptides from albumin, capture and concentration in a C18 Zip plate and direct deposition on a PerkinElmer disposable target plates to be read in the ProTOF orthogonal MALDI instrument. An evaluation of the method for the discovery of a pattern for Alzheimer's disease has been reported²⁵ and a pattern for ovarian cancer diagnosis²⁶. The instrument has the high resolution properties in combination with the ability to scale up discovery. This combination should allow for the large validation studies that will be required to both verify and validate a pattern for disease diagnosis. It is

anticipated that the scale of such studies will need to be massive not only to convince skeptics of the validity of patterns as well as to convince the regulatory agencies that these patterns are effective and should be approved as diagnostic tests.

While proteomic patterns have been widely reported, another tactic in the use of proteomics for diagnosis has been biomarker discovery. A connection between a diagnostic marker and its function in the disease has been advocated by some as a necessary step in order to validate the use of a biomarker in disease diagnosis. In this case, the proteomics approach would not result necessarily in a test itself but would be used as a discovery platform for biomarkers or panels of biomarkers that could be used to detect and monitor disease. One example of this approach is the work of Zhang et al²² where biomarkers for ovarian cancer were discovered using the Ciphergen proteomic platform in an investigation of patients from multiple centers. The three biomarkers identified were apolipoproteinA1, a truncated form of transthyretin and a cleavage fragment of inter-α-trypsin inhibitor heavy chain H4. The addition of these three biomarkers as a panel to CA125 improved the sensitivity and specificity of CA125 alone for detection of disease. The panel approach has been used in diseases such as thyroid disease for many years but the application of this approach to heterogeneous diseases such as cancer is now being investigated and in theory has a sound scientific basis.

In addition to diagnosis, the possibility of disease progression and treatment monitoring has been an attractive use of proteomics technology. The field of individualized medicine is an area that could predict a patient's response to a therapy thereby guiding the treating physician rather than waiting to see if a patient responds to a therapy 21 . Disease progression could also be monitored by an analysis of changes in an individual's proteomic pattern over time and looking for changes in targeted points in the spectra or by the isolation and identification and quantitation of individual proteins in serial samples. This is currently done in a primitive way in the quantitation of antibodies in infectious disease over the course of time or with tumor markers such as CEA in selected cancers where the initial diagnosis yields an elevated level that is monitored during treatment in serial samples. The mass spectrometry tool is particularly powerful at discrimination of posttranslational modifications that take place in the course of disease and this is generally difficult in immunoassays. A combination of the two techniques has been investigated²³ and could prove to give the advantages of purification by a binding reaction followed by the resolution of the subtle differences in the binding entities by mass spectrometry.

Table 21-1. Recommended Practices for Clinical Applications of Protein Profiling by MALDI TOF Spectrometry.

1. PREANALYTICAL

- Evaluate optimum patient preparation
- Identify optimum procedures for specimen collection and processing
- Analyze specimen stability
- Develop criteria for specimen acceptability

2. ANALYTICAL

- · Prepare calibrators for mass, resolution, and detector sensitivity
- Use internal standards
- Automate specimen preparation
- Optimize methods to yield highest possible signals for peaks of interest
- Identify sequences of peaks of interest
- Develop calibration materials for components of interest
- QC: prepare/identify at least two concentrations of control material
- Evaluate reproducibility (precision)
- Evaluate limits of detection and linearity
- Evaluate reference intervals
- Evaluate interferences such as hemolysis, lipemia, renal failure, acute-phase responses
- Develop materials or programs for external comparison/proficiency testing of analyzers

3. POSTANALYTICAL

- Analyze each spectrum to identify peaks before applying diagnostic algorithms
- Develop criteria for the acceptability of each spectrum based on peak characteristics
- Use peaks rather than raw data as the basis for diagnostic analysis
- Use caution in interpretation of peaks with m/z <1200
- Select peaks with high intensities and sample stability for diagnosis
- Select approximately equal numbers of peaks that increase and decrease in intensity as diagnostic discriminators
- In developing a training set for diagnosis, careful clinical classification of patients is essential
- Clinical validity depends on having a typical rather than highly selected population of patients
- The number of training specimens should be at least 10 times the number of measured values
- Any clinical application should use a fixed training set and algorithm for analysis
- Any analysis should provide a numerical value
- Diagnostic performance should be evaluated with ROC curves to select cutoffs
- A sensitivity analysis should be performed of the necessary precision for accurate diagnostic performance
- There should be QC procedures for daily verification of software performance

*Adapted from G.L. Hortin²⁴.

3. DELIVERING THE PROMISE

All of these new technologies have opened up the possibility for improved diagnostic tests for a wide variety of diseases. However the validation and regulatory challenges are quite great. The validation of a mass spectrometry pattern diagnostic would require the validation and integration of systems and technologies as diverse as proteomics itself: Reagents and the chemistry behind them, robotics processors, mass spectrometers, operational software and diagnostic software along with the clinical end including the sample handling and transport, sample stability and a host of other factors²⁴.

However, the power and potential of the technology will drive the continued innovation, validation and eventual commercialization of the area of proteomics and the eventual arrival of these techniques in the clinical lab will give tools for improved patient outcomes.

ACKNOWLEDGEMENTS

The author would like to thank the NCI/FDA Clinical Proteomics Program staff, both past and current, and the staff of the Clinical Proteomics Lab for their support and insight.

The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organization imply endorsement by the U. S. Government. This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. NO1-CO-12400.

REFERENCES

- G. Kohler, C. Milstein, Continuous cultures of fused cells secreting antibody of predefined specificity, *Nature* 256, 495 (1975).
- N.L. Anderson and N.G. Anderson, The Human Plasma Proteome, MCP 1(11), 845-67, (2002).
- E.S. Lander, L.M. Linton et al, International Human Genome Sequencing Consortium. Initial sequencing and analysis of the Human Genome, *Nature* 409(6822), 860-921, (2001).
- 4. Abbott, Proteomics, transcriptomics: what's in a name?, Nature 202, 715-716, (1999).
- 5. L. Anderson, N.G. Anderson, High resolution two-dimensional electrophoresis of human plasma proteins, *Proc Natl Acad Sci USA* **74**, 5421-5425 (1977).
- M.R. Emmert-Buck, R.F. Bonner, P.D. Smith, R.F. Chauqui, Z. Zhuang, S.R. Goldstein, R.A. Weiss, L.A., Liotta, Laser capture microdissection, *Science* 274, 998-1001 (1996).

21. Proteomics for Diagnostic Applications

- C.P. Paweletz, J.W. Gillespie, D.K. Ornstein, N.L. Simmone, M.R. Brown, K.A. Cole, Q.H. Wang, J. Huang, N. Hu, T.T Yipe, W.E. Rich, E.C. Kohn, W.M. Linehan, T. Weber, P. Taylor, M.R. Emmert-Buck, L.A. Liotta, E.F. Petricoin, Rapid Protein Display Profiling of Cancer Progression Directly from Human Tissue Using a Protein Biochip, *Drug Dev Research* 49, 34-42 (2000).
- E.F. Petricoin, A.M. Ardekani, B.A. Hitt, P.J. Levine, V.A. Fusaro, S.M. Steinberg, G.B. Mills, C. Simone, D.A. Fishman, E.C. Kohn, L.A. Liotta, Use of Proteomic Patterns in Serum to identify Ovarian Cancer, *Lancet* 359, 572-577 (2002).
- E.F. Petricoin, D.K. Ornstein, C.P. Paweletz, A. Ardekani, P.S. Hackett, B.A. Hitt, A.Velassco, C. Trucco, L. Weigand, K. Wood, C.B. Simone, P.J. Levine, W.M. Linehan, M.R. Emmert-Buck, S.M. Steinberg, E.C. Kohn, Serum Proteomic Patterns for Detection of Prostate Cancer, *JNCI* 94, 1576-1578 (2002).
- J. Li, Z. Zhang, J. Rosenzweig, Y.Y. Wang, D.W. Chan, Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer *Clin Chem* 48, 1296-1304 (2002).
- Y. Won, J.J. Song, T.W. Kang, J.J. Kim, B.D. Han, S.W. Lee, Pattern analysis of serum proteome distinguishes renal cell carcinoma from other urologic diseases and healthy persons, *Proteomics* 3, 2310-2316 (2003).
- S. Bhattacharyya, E.R. Siegel, G.M- Petersen, S.T. Chari, L.J. Suva, R.S. Haun, Diagnosis of Pancreatic Cancer Using Serum Proteomic Profiling, *Neoplasia* 6, 674-686 (2004).
- K.R. Kozak, W.A. Malaika, S.M. Pusey, F. Su, M.N. Luong, S.A. Luong, S.T. Reddy, R. Farias-Eisner, Identification of biomarkers for ovarian cancer using strong anionexchange ProteinChips: Potential use in diagnosis and prognosis, *PNAS* 100, 12343-12348 (2003).
- J. Marshall, P. Kupchak, W. Zhu, J. Yantha, T. Vrees, S. Furesz, K. Jacks, C. Smith, I. Kireeva, R. Zhang, M. Takahashi, E. Stanton, G. Jackowski, Processing of Serum Proteins Underlies the Mass Spectral Fingerprinting of Myocardial Infarction, J Proteome Res 2, 361-372 (2003).
- K.A. Baggerly, J.S. Morris, K.R. Coombes, Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments, *Bioinformatics* 20, 777-785 (2004).
- E.P. Diamandis, Analysis of Serum Proteomic Patterns for Early Cancer Diagnosis: Drawing Attention to Potential Problems, *JNCI* 96, 353-356 (2004).
- T.P. Conrads, V.A. Fusaro, S. Ross, D. Johann, V. Rajapakse, B.A. Hitt, S.M. Steinberg, E.C. Kohn, D.A. Fishman, G. Whiteley, J.C. Barrett, L.A. Liotta, E.F. Petricoin, T.D. Veenstra, High-resolution serum proteomic features for ovarian cancer detection, *Endocrine Related Cancer* 11, 163-178 (2004).
- K.A. Baggerly, J.S. Morris, J. Wang, D, Gold, L.C. Xiao, K.R. Coombes, A comprehensive approach to the analysis of matrix-assisted laser desorption/ionizationtime of flight proteomics spectra from serum samples, *Proteomics* 3, 1667-1672 (2003)
- L.A. Liotta, M. Lowenthal, A. Mehta, T.P. Conrads, T.D. Veenstra, D.A. Fishman, E.F. Petricoin, Importance of Communication Between Producers and Consumers of Publicly Available Experimental Data, *JNCI* 97, 310-314 (2005).
- A. Mehta, S. Ross, M.S. Lowenthal, V. Fusaro, D.A. Fishman, E.F. Petricoin, L.A. Liotta, Biomarker amplification by serum carrier protein binding, *Disease Markers* 19, 1-10 (2003,2004).
- 21. L.A. Liotta, M. Ferrari, E. Petricoin, Written in Blood, Nature 425(6961), 905 (2003).
- 22. Z. Zhang, R.C. Base, Y. Yu, J. Li, L.J. Sokol, A.J. Rai, J.M. Rosenzweig, B. Cameron, Y.Y. Want, X.Y. Meng, A. Berchuck, C. Haaften-Day, N.F. Hacker, H.W.A. Bruijn,

A.G.J. Zee, I.J. Jacobs, E.R. Fung, D.W. Chan, Three Biomarkers Identified from Serum Proteomic Analysis for the Detection of Early Stage Ovarian Cancer, *Cancer Res* **64**, 5882-5890 (2004).

- 23. D. Nedelkov, R.W. Nelson, Surface plasmon resonance mass spectrometry: recent progress and outlooks, *Trends in Biotech* **21**, 301-305 (2003).
- 24. G.L. Hortin, Can Mass Spectrometric Protein Profiling Meet Desired Standards of Clinical Laboratory Practice?, *Clin Chem* **51**, 3-5 (2005).
- M.F. Lopez, A. Mikulsdis, S. Kuxzdzal, A. Bennet. et al, High Resolution Serum Proteomic Profiling of Alzheimer Disease Samples Reveals Disease-Specific, Carrier-Protein-Bound Mass Signatures, *Clin Chem* 51(10), in press (2005).
- M.S. Lowenthal, A.I. Mehta, K. Frogale, R.W. Bandle et al., Analysis of Albumin-Associated Peptides and Proteins from Ovarian Cancer Patients, *Clin Chem* 51(10), in press (2005).
- D.J. Johann, M.D. McGuigan, A.R. Patel, S. Tomav, S. Ross, T.P. Conrads, T.D. Veenstra, D.A. Fishman, G.R. Whiteley, E.F. Petricoin, L.A. Liotta, Clinical Proteomics and Biomarker Discovery, *Ann N Y Acad Sci* 1022, 295-305 (2004).