Proteomics for the Discovery of Biomarkers and Diagnosis of Diseases

Shanti Gunawardena

Introduction

Over the past few years a significant amount of data pertaining to the diagnosis of human diseases has been generated with the help of mRNA (cDNA) microarrays. They have been responsible for identifying new disease subtypes that would not have been possible using conventional techniques. As a result, the need for new molecular based classifications of some types of cancers (Alizadeh, Eisen et al. 2000; Sorlie, Perou et al. 2001) have been highlighted by these studies. If DNA microarrays are so good, do we need proteomics to diagnose diseases, and if so how will it be better than DNA array expression profiling? Before this question is answered, it is instructive to consider what types of changes take place in the proteome with different diseases.

In cancers, one would expect an altered expression of proteins responsible for signal transduction processes in the cell. In fact, in many instances the protein products of proto-oncogenes are involved in signal transduction (Alaiya, Franzen et al. 2000) and alterations in these genes result in uncontrolled cellular signaling. Over expression and post-translational modifications of several oncogene products have been detected in transformed liver cells (Sanchez, Wirth et al. 1997). Additionally, cancer has been associated with altered glycosylation of many proteins (Taylor-Papadimitriou and Epenetos 1994).

Cardiovascular disease, the leading cause of death in the United States, is most often associated with atherosclerosis. Studies on the molecular basis of atherosclerosis have led to the identification of a number of molecules that may play a critical role in the development of atherosclerotic lesions. A member of the immunoglobulin gene superfamily, intercellular cell adhesion molecule 1 (ICAM-1) has recently been shown to be associated with this disease in humans (Ballantyne and Entman 2002). ICAM-1 is a cell-surface glycoprotein capable of eliciting bidirectional signals that activate signaling pathways in leukocytes, and endothelial, and smooth muscle cells. These molecules can be measured quantitatively in plasma. In the Physicians' Health Study, increased levels of soluble ICAM-1 were noted in individuals who subsequently developed symptomatic arterial disease (Pradhan, Rifai et al. 2002). Proteolytic cleavage may play a role in the altered abundance of ICAM-1 (Champagne, Tremblay et al. 1998).

In studies of dilated cardiomyopathy, which results in heart failure, decreased protein abundance of cytoskeletal proteins and proteins associated with mitochondria and energy metabolism have been noted (Banks, Dunn et al. 2000). Also noted was a seven-fold increase in the enzyme ubiquitin C-terminal hydrolase (Weekes, Wheeler et al. 1999), suggesting that excessive ubiquitination, and subsequent degradation of proteins may contribute to cardiac remodeling resulting in further dilatation of the heart and worsening heart failure.

Alzheimer's disease, which is the most common cause of dementia, is pathologically characterized by the presence of senile plaques and neurofibrillary tangles. Microtubule associated protein tau which is present in 6 alternatively spliced isoforms is the major protein component of neurofibrillary tangles (Banks, Dunn et al. 2000). Immunochemical and electrophoretic studies have shown that several post-translational modifications including ubiquitination, glycosylation, glycation, and hyperphosphorylation occur in pathological tau as compared to normal tau (Grundke-Iqbal, Iqbal et al. 1986). Similarly, Lewy body dementia and other neurodegenerative disorders referred to as synucleinopathies have been associated with post-translational modifications of synuclein that favor its fibrillization and aggregation (Dickson 2001). These disorders are associated with parkinsonism, autonomic dysfunction, and dementia. Another neurological disorder where post-translational modifications have been implicated are the Prion diseases (Parchi, Castellani et al. 1996; Banks, Dunn et al. 2000). These rare diseases are characterized by fatal degenerative encephalopathy. The most well known human form of this disorder is Creutzfeld-Jakob disease (CJD) which may be inherited, acquired, or be sporadic. The neuropathologic diagnosis of this disease requires a brain biopsy. There is interest in this disorder at this time because of its relationship to bovine spongiform encephalopathy.

Other diseases where changes in the proteome appear to include post-translational modifications are autoimmune diseases (e.g. rheumatoid arthritis) (Doyle and Mamula 2002), connective tissue diseases (Uitto and Lichtenstein 1976), urological malignancies (Unwin, Knowles et al. 1999), and inherited muscle diseases (e.g. muscular dystrophy) (Hewitt and Grewal 2003).

It is apparent that in most diseases, proteins are subjected to numerous changes including post-translational modifications and/or proteolytic cleavage. Furthermore, there is overexpression or under expression of a number of proteins in some diseases. These observations highlight the fact that mRNA expression profiling falls short of providing a complete solution to the tasks of biomarker discovery and diagnosis of disease. Microarrays that provide information on differential expression of mRNA will not provide information on post-translational modifications. Alternative splicing of the mRNA transcript can produce different protein forms, and at this time the only way to study the impact of these proteins is at the protein level. An additional concern is the lack of correlation between mRNA levels and protein concentrations (Gygi, Rochon et al. 1999). Finally, an especially significant impediment to the discovery and use of clinically usable biomarkers with mRNA/cDNA techniques is their limited utility for the analysis of biological fluids. This will be discussed in greater detail later.

The protein complement of a cell or proteome is dynamic and reflects the conditions the cell is subjected to or a specific disease state. The detection of proteins that serve to relay the physiological status of a cell during various phases of a disease has been studied for many decades. However, in the past this has been done mostly on a one-protein-at-a-time basis by looking for proteins that are overexpressed and shed into body fluids (Wulfkuhle, Liotta et al. 2003). This has been a time consuming and, often, a thankless

endeavor on account of the multitude of intact and cleaved proteins present in the human proteome.

The study of proteins one-protein-at-a-time is not new. Antibodies were developed for use in serology to precipitate and quantify antigens as early as in 1929 (MacBeath 2002). This technology significantly improved in 1959 with the advent of the radioimmunoassay (RIA) by Yarlow and Berson. Further progress was made in 1971 with the introduction of enzyme-linked immunosorbant assay (ELISA) by Engvall and Perlman (Engvall and Perlman 1971). To this day, these very specific protein-ligand assays remain the most reliable and sensitive tests available to detect many diseases – particularly, the infectious diseases.

The workhorse of proteomic studies over the past 25 years has been two dimensional polyacrylamide gel electrophoresis (2D-PAGE). Electophoresis is often followed by the identification of proteins using mass spectrometry (MS). More recently, the potential to perform proteomic investigations using protein microarrays has been described (MacBeath 2002; Liotta, Espina et al. 2003). With the ability to study multiple proteins simultaneously, this technology along with DNA microarrays has the potential to go well beyond the discovery of biomarkers. However, since protein sequences do not have the ability to hybridize with an anti-sense sequence as in the case DNA or RNA, the major limitation here will be the development of high affinity antibodies or other affinity reagents to bind the different proteins and peptides in the specimens to be studied. The most recent addition to the armamentarium of proteomic techniques for the diagnosis of disease has been proteomic pattern diagnostics based on mass spectra (Petricoin, Ardekani et al. 2002). The bioinformatics problem involved in each case is somewhat different from the others.

Most proteomic biomarker discovery efforts have focused on cancer. This makes a great deal of sense since it is the second most common cause of death in the US and in the majority of the cases the prognosis can be improved significantly by early stage or even pre-malignant stage diagnosis of the disease. Most conventional therapies for cancer have limited success once the malignant cells have spread beyond the tissue of origin. Over 60% of patients with breast, lung, colon, and ovarian cancer have metastatic disease at the time of diagnosis.

In this paper we will critically review these technologies with emphasis on biomarker discovery and the diagnosis of disease. We will also discuss the bioinformatic techniques used to manipulate and make sense out of the very large amounts of data generated by the various proteomic technologies.

Two-dimensional electrophoresis

In this method proteins are separated in one dimension using isoelectric focusing and in the second dimension on the basis of the relative molecular masses. The intensities of protein spots of interest are used as a measure of their abundance to study differential expression. The spots are then excised, digested and subjected to mass spectrometry for

subsequent protein identification by peptide mass fingerprinting (Gras, Muller et al. 1999). The samples used in this technique are usually cellular lysates from disease and normal tissues or serum. Direct comparison of protein expression to identify differentially expressed proteins between the cells from the disease and normal tissue specimens is eminently possible and has been used to discover biomarkers for cancers involving the liver (Seow, Liang et al. 2001), bladder (Celis, Wolf et al. 2000), lung (Chen, Gharib et al. 2002), esophagus (Soldes, Kuick et al. 1999), prostate (Meehan, Holland et al. 2002), breast (Franzen, Linder et al. 1997), and kidney (Sarto, Frutiger et al. 1999). The use of laser capture microdissection (LCM), a technique developed to rapidly harvest pure cell populations from heterogeneous tissue (Emmert-Buck, Bonner et al. 1996) has markedly improved the specificity of 2D-PAGE for the discovery of biomarkers. The search for the early disease markers from a number of different cancerous tissue types has benefited from this technology (Ornstein, Englert et al. 2000; Wulfkuhle, Sgroi et al. 2002).

As useful it has been in the discovery of biomarkers, 2D-PAGE has many shortcomings. The post-electophoretic steps are amenable to automation, with image analysis to identify protein spots and their intensities, and robotics for spot excision, digestion, and presentation for mass spectrometry. However, the steps leading to electrophoresis are laborious and time consuming. Thus, this method has to be classified as a low throughput method for disease diagnosis. Furthermore, the specimen needed for analysis is larger than what is required for some of the other proteomic technologies to be discussed later. Yet, another major limitation of this technology is the restrictive dynamic range of the sample that can be analyzed, i.e., inability to handle the very large diversity and divergent expression levels of proteins in the sample. As a result, low abundance proteins cannot be reliably detected and identified. In human cells, the most abundant protein is usually actin which can dwarf the abundance of many other proteins (Hamdan and Righetti 2002). Enrichment and /or prefractionation has helped in this regard, although not completely (Herbert and Righetti 2000). Contamination from stromal tissue surrounding the abnormal cells can also compromise the discovery of disease specific biomarkers. As noted previously, laser capture microdissection can improve the specificity of this search. Another problem that is encountered is variation between gels preventing the direct superimposition of images from the control and disease specimens. This problem appears to have been at least partially overcome with a variant technique, differential in-gel elctrophoresis (DIGE) (Unlu, Morgan et al. 1997), where each sample is covalently labeled with a different mass and charge matched fluorescent dye (Cy3 and Cy5) before mixing the samples and analyzing them on the same gel. Image analysis software can then provide the difference of expression between the normal and disease cells. This is very similar to the analysis that is performed in DNA microarrays where the specimens are labeled with different fluorescent dyes. This technique has been used to study differentially expressed proteins in squamous cell cancer of the esophagus and normal esophageal tissue (Zhou, Li et al. 2002). Qauntitative data may not always be needed to differentiate between normal and disease states. For example, Sarto et al, using 2D-PAGE, showed that two spots corresponding to two isoforms of plasma glutathione peroxidase were present in normal kidney tissue but not in renal cell carcinoma tissue (Sarto, Frutiger et al. 1999).

Is there always a need for protein identification using peptide mass fingerprinting following 2D-PAGE to differentiate between normal and disease tissue? Strictly from a diagnostic standpoint this does not appear to be the case. However, rigorous characterization of the biomarkers may contribute to the development of high-affinity, specific antibodies that can enhance disease detection. The use of a combination of such biomarkers may further increase the sensitivity of the diagnostic test.

From the bioinformatics perspective, 2D-PAGE data extraction and manipulation takes several forms. The individual protein spots on the gel need to be detected, quantified and the intensity of the signal for each spot has to be corrected for the local background. As can be seen in Figure 1, protein spots are distributed irregularly (in contrast to a microarray) and there is a wide variation in morphology. In many cases, the spots are clumped together making resolution of the individual spots difficult. A number of algorithms are available to analyze these problem spots and to generate a spot list. These algorithms can be based on Gaussian fitting, Laplacian of Gaussian (LOG) spot detection, line and chain analysis, or watershed transformation (WST) (Weasthead et al. 2002**)**. In Gaussian fitting, if a spot cannot be matched with a single Gaussian shape, an overlapping combination of Gaussian shapes is used to match the morphology of the spot. In line and chain analysis, columns of pixels from the digitized image of 2D-PAGE are scanned to identify peaks in signal intensity. This is repeated for all of the pixel columns. The algorithm identifies the center of the spot as well as its signal volume (the overall signal intensity). In WST a topographical map is developed using the pixel signal intensities that is then used to separate clusters, chains, small spots overlapping with larger spots, and also to merge regions of a single spot.

The next step in the process is to perform gel matching in order to determine the differential expression of proteins between the control and disease cells. It is important to note that post-translational modification of an existing protein resulting in a mass or charge change can give rise to a new spot. Thus, not every new spot reflects a new protein. The principle behind gel matching is to establish landmark spots and to use algorithms to match the positions of the remaining spots. This may include manipulations such as stretching and rotating. In a variation to this theme known as 'propagation', distances between landmark spots and their neighbors are determined and they are compared with other gels. Matches result in a new set of landmarks that now include the neighboring spots. This process can be repeated. MELANIE is a program that uses this concept (http://us.expasy.org/melanie/).

Spots of interest can be excised from gels, cleaved with an enzyme such as trypsin and subjected to analysis by mass spectrometry. The resulting "peptide mass fingerprints" can be used to for protein identification using databases of "virtually digested" proteins. Peptide mass fingerprinting data from the mass spectrum is used as a query for protein identification in protein databases using "virtually digested" proteins. One of the most commonly used programs for this purpose is SEQUEST (http://fields.scripps.edu/sequest/), which searches for all of the peptides with the same mass in the databases that are specified. A theoretical mass spectrum is then generated following a virtual digest on the matched protein. A comparison is made between the data

from the theoretical mass spectrum and the query mass spectrum and the best matches are scored. An unknown post-translational modification can cause mismatches. Algorithms such as SEQUEST have built in parameters for detecting known post-translational modifications. Nonspecific proteolysis that can result from impure cleavage agents can be another complication. This problem is addressed in many algorithms by performing searches without a specified cleavage agent. Further difficulty in identification can result from mixtures of proteins that are often present at spots excised from 2D-PAGE gels. This discussion assumes that the protein that is being sought exists in the database. An imperfect match can result if there is close homology with a related sequence present in the database.

Mass Spectrometry-based methods

The mass spectrometer essentially consists of an ion source, a mass analyzer that separates ionized analytes according to their mass to charge (m/z) ratio, and a detector that to counts the number of ions at each m/z value. The heart of the mass spectrometer is the mass analyzer. A variety of mass analyzers are available at the present time. The key parameters used to compare these analyzers are sensitivity, resolution, mass accuracy, and the ability to generate mass spectra with high information content from peptide fragments. These analyzers differ in design and performance characteristics, and have particular strengths and weaknesses that will not be discussed here. A commonly encountered mass analyzer uses the time-of-flight (TOF) of the ion before it is recorded by the detector plate as a measure of its mass to charge ratio, with low molecular weight ions arriving faster than heavier ones. As with mass analyzer components, a similar range of options exists for sample ionization sources. The proteins or peptides used for analysis may be ionized using electrospray ionization if the specimen is in solution. This approach can be readily coupled with liquid chromatography (LC) for fractionation of complex analytes prior to mass spectrometry. Matrix assisted laser desorption/ionization (MALDI) is another commonly used sample ionization technique in which the specimen is embedded within a crystalline matrix. This matrix facilitates the ionization of analyte molecules when it is pulsed with a laser. Mass spectrometers can also be used in tandem to generate peptide sequence data. Here the first analyzer is used to separate peptide species, which are, in turn, sequentially fragmented and delivered to a second analyzer. The latter generates mass information on the nested set of peptide fragment ions. These data can finally be used to infer amino acid sequence information.

As was noted previously, the major limitation of 2D-PAGE is its poor sensitivity with respect to low abundance proteins. This can be particularly problematic in the context of post-translationally modified proteins, which are often present in extremely low amounts (Srinivas, Verma et al. 2002). Liquid chromatography (used in lieu of electrophoresis) can be coupled with tandem mass spectrometry to improve the separation and identification of analytes in the low femtomolar range (Gygi, Han et al. 1999). Reversed-phase LC has been used to concentrate and separate peptides from extremely complex mixtures prior to sequencing (Mann, Hendrickson et al. 2001). Further improvements in high-throughput, high-sensitivity detection methods have been realized with affinity-based MS techniques. A novel MS technology was developed Hutchens and Yip and

reported in 1993 (Srinivas, Verma et al. 2002) and it forms the basis for the surface-enhanced laser desorption/ionization (SELDI) ProteinChip produced by Ciphergen Biosystems, Inc (http://www.ciphergen.com/). This technology combines on-chip separation of protein mixtures using a specially treated surface that binds with proteins in the specimen to be analyzed. The surface is then washed and is subjected to MS analysis to produce a proteomic fingerprint. One of the most exciting aspects of SELDI is that a very small amount of an easily accessible body fluid, such as serum, saliva, or urine can be deposited on the surface to produce a proteomic profile. As will be discussed later, the ability to use easily accessible body fluids or other relatively non-invasively obtained tissue samples is critical if large scale screening is to be done for diseases. This technique has been used to develop biomarkers for ovarian cancer (Petricoin, Ardekani et al. 2002), prostate cancer (Xiao, Adam et al. 2001), pancreatic ductal adenocarcinoma (Rosty, Christa et al. 2002), breast cancer (Paweletz, Trock et al. 2001), and hepatocellular carcinoma (Poon, Yip et al. 2003). If identification of the proteins is the goal, this technique still needs upstream fractionation and downstream purification (Petricoin and Liotta 2002).

Petricoin et al., have demonstrated the use of SELDI-TOF MS to identify a serum proteomic signature for ovarian cancer (Petricoin, Ardekani et al. 2002). The underlying principle here is that changes in the proteomic composition of the blood reflect changes in the state of health of the organism. This proteomic composition constantly changes as a result of the perfusion of diseased organs, and can include myriad proteins and cleaved proteins. In fact, in the case of cancer, many elements of the serum proteome may reflect the unique host-tumor microenvironment, which may be responsible for changes in a variety of protein-protein interactions, protein folding, and protein abundances (Liotta and Kohn 2001). In the study described by Petricoin et al., a proteomic a mass spectral pattern was initially determined using a training set of mass spectra from two groups of patients – one with known ovarian cancer and another unaffected by cancer. This proteomic diagnostic pattern was used on 116 masked serum samples - 50 from women with ovarian cancer and the other 66 from unaffected women. All 50 of the patients with ovarian cancer –18 of whom had stage I disease - were correctly identified. Of the 66 patients without malignant disease, 63 were identified as not having ovarian cancer. The sensitivity was 100% and the specificity was 95%. Of note, this method does not require the identification, or rigorous quantification of any protein. The pattern itself is the diagnostic marker. For comparison, CA125 - the most commonly used biomarker for ovarian cancer - is abnormal in approximately 80% of patients with advanced stage disease and is elevated in only 50-60% of patients with stage I disease. The 5 year survival of patients with late stage disease is 35% and that for stage I disease exceeds 90%.

The bioinformatics problem here entails the analysis of up to 15000 m/z points in the case of a low resolution spectrum (Figure 2). Higher resolution devices can generate upwards of 400,000 data points per spectrum. Spectral intensities at most of these points constitute noise. The challenge is to identify the relative handful of mass species whose intensities tend to vary significantly across disease states. These species are the disease-specific biomarkers or elements of a broader proteomic signature. Now, an exhaustive

search for the optimal set of disease state distinguishing proteomic features would not be computationally feasible – even with a relatively small set of patient samples. To circumvent this challenge, heuristics are required – essentially trading optimality for computational tractability. Petricoin and colleagues took this course in their landmark paper detailing the identification of a proteomic signature for ovarian cancer. Their report did not emphasize algorithmic details, but a broad idea of the approach can be gleaned. In particular, a combination of a genetic algorithm with a clustering technique was used to 'learn' a set of m/z values constituting a proteomic signature for ovarian cancer. Genetic algorithms represent a machine learning approach in which potential solutions to a problem are encoded in a computationally accessible form. Especially good or 'fit' solutions can then be recombined to form potentially better solutions, with the algorithm tracing a course roughly reminiscent of biological evolution. To identify sets of promising biomarkers, a genetic algorithm was initiated with a random collection of m/z value subsets. (Sets of five were used in the ovarian cancer profiling work.) This 'population' was evolved over multiple generations to identify the highly predictive set of m/z values that emerged from the study. What fitness metric drove the evolution toward 'better' m/z sets? This is where the clustering element of the approach enters. Each candidate set of k m/z values considered by the algorithm defines a k-dimensional space in which corresponding spectral intensities can be 'plotted'. Good sets of m/z values define spaces in which patient samples (reduced to vectors of spectral intensities at the selected m/z values) segregate into disjoint, disease state-specific clusters. If a highly discriminatory set of m/z values emerges (as in the case of the ovarian cancer profiling effort), blinded samples can be classified according to the state of the clusters within which they fall. If samples do not fall within an existing cluster, a useful classification of 'unknown' can be rendered. The exact clustering method used is not clear. However, many such methods have been described in the literature. The BIOC218 final project paper by Erin Davies (2003) provides a review of the different clustering techniques.

In a variation of the theme, Poon et al (Poon, Yip et al. 2003) used two way hierarchical clustering of SELDI-TOF derived data to differentiate between hepatocellular carcinoma and chronic liver disease. Serum was used for SELDI. In hierarchical clustering, a dendogram is developed in a manner very similar to that used in phylogenetic analysis. Euclidean distances are used to determine the nearest neighbors.

Protein Microarrays

DNA Microarrays make use of the ability of one strand of DNA to hybridize with a complementary strand of DNA in a predictable manner. This principle does not apply to protein sequences, and as such, the development of protein microarrays has not been quite as straight forward. These arrays require highly specific, high-affinity reagents (most often antibodies) that can capture and bind an analyte which can be a cell lysate, or a body fluid. This is one of the major challenges in producing protein microarrays. These bait molecules have to be able to recognize the state of modification of a protein that they will capture if these arrays are to achieve their full potential. The principle of analyte (protein) capture is the same as what was recognized as far back as 1929 for use in serology to precipitate and quantify antigens. There are two basic forms of these protein

microarrays. The difference has to do with whether the antibody or the analyte is immobilized on the solid surface of the array. In forward phase arrays (FPA), different antibodies are spotted on the solid surface and the multiple analytes from one test sample (a cellular lysate) are captured from solution phase. In reverse phase arrays (RPA) the analytes from different test samples (for example from different patients) are immobilized at each spot on the solid surface and the antibody is in solution phase (Liotta, Espina et al. 2003). Fluorescent, radioactive, or luminescent labels can be used for detection and signal amplification. Most of the technical details are left out in the interest of space. However, suffice it to say that there are many technical challenges that need to be overcome. By way of example, one such challenge entails the denaturation of the antigen in order to linearize the epitope as required by many antibodies. Such denaturation can destroy protein-protein interactions that can be very informative.

The bioinformatics problem associated with protein microarrays is very similar to that one encounters with DNA microarrays. The open source program PSCAN (Peak quantification with Statistical Comparative Analysis - http://abs.cit.nih.gov/pscan/), originally developed for gene expression analysis can be used to analyze the information from these microarrays (Liotta, Espina et al. 2003). This program first generates a file containing spot intensities and addresses for each image. It then compares spot intensities between different arrays to determine differential expression. Finally it prepares a file containing intensities of all arrays to be opened in a statistical package such as JMP (www.jmp.com/). This statistical package provides a replica of the image, a scatter plot/histogram, and peak intensities. Defective artifactual areas or other low quality areas on an array can be subtracted from the scatter plot, thus removing the distortion these areas may produce. The program allows multiple scatter plots to be produced for time course studies. JMP output can also be subjected to cluster analysis. Liotta et al (Liotta, Espina et al. 2003) did two way hierarchical clustering to analyze microdissected human breast cancer and normal breast epithelium, comparing phosphorylation states of a series of proteins within the EGF-receptor family signaling pathway.

Protein microarray technology has significant potential for biomarker discovery, though, as discussed previously, other competing proteomic technologies exist for this task. The greatest strength of protein arrays may be in profiling the state of members of signaling pathways and protein networks (Liotta, Espina et al. 2003). Very few papers have been published to date specifically detailing the use of protein microarrays in the discovery of biomarkers or the diagnosis of diseases. In a study by Miller et al (Miller, Zhou et al. 2003) antibody microarray profiling of human prostate cancer sera was used to identify five proteins that had significantly different levels between prostate cancer samples and controls.

Discussion

The two most common causes of death in the United States are heart disease and cancer. These two diseases together account for more that 50% of the annual deaths (http://www.cdc.gov/nchs/fastats/lcod.htm). In the year 2000 there were 1.26 million such deaths. In both of these diseases, early diagnosis can very significantly alter the

prognosis. This is particularly true with cancer where most conventional therapies are limited in their success once the cancer has spread beyond the tissue of origin. As noted previously, more than 60% of patients with breast, lung, colon, and ovarian cancer have metastatic disease at the time of presentation. There clearly is a dire need for reliable screening tests. It is unreasonable to expect biopsy specimens from internal organs for screening. There are no truly benign invasive procedures. Ideally, these tests should be performed using easily accessible body fluids such as blood, urine, or saliva. This requirement alone makes gene expression approaches unsuitable for screening tests. While electrophoretic methods can be used provided the biomarkers are not expressed at a low abundance, the low throughput of electrophoretic methods would not be conducive to large scale regular screening. On the basis of what we know at the present time, proteomic approaches using high throughput mass spectrometers can fill this niche very well.

Which proteomic method should we choose for the diagnosis of different diseases? No generalizations are possible except for, perhaps, infectious diseases. Traditionally, the diagnosis of infectious diseases has been made by demonstrating the presence of infectious agents in tissues, body fluids, or excreta of the host. This may be done directly by microscopy, staining, agglutination assays, or enzyme immunoassays. Detection of pathogenic agents can also be made by culture. Indirect detection of pathogenic agents can be made by serologic methods where serum antibody levels are measured. More recently nucleic acid probes have been used for the detection and quantification of specific DNA or RNA based sequences. So, how does proteomics fit in? Extensive use of 2D-PAGE has been made in the past to separate microbial proteins and then to identify them in order to investigate pathogenic determinants and antibiotic targets. This is not expected to change significantly in the future (Cash 2000). The complete sequencing of the relatively small genomes of many of these infectious agents will play a complementary role in the discovery of these proteins. Serology where the antigen-antibody interaction will be exploited will continue to play a major role in diagnosis of these diseases.

The question as to whether quantification of proteins is necessary or whether only the presence or absence of a protein should be established in the diagnosis of diseases also cannot be answered on a disease-by-disease basis. There are small studies that that have reported the ability to differentiate between disease and normal specimens (Sarto, Frutiger et al. 1999; Meehan, Holland et al. 2002) on the basis of the presence or absence of proteins on a 2D-PAGE map. However, the majority of proteomic studies using 2D-PAGE need a quantitative determination of the differential expression of proteins. This question does not apply to the pure mass spectrometry based methods.

The approach presented by Petricoin et al, proteomic pattern diagnostics, if validated in clinical trials, will represent a breakthrough in the discovery of biomarkers and diagnosis of diseases. Interestingly, the success and acceptance of this approach may very well depend on the bioinformatic methods used. This is especially likely in the case of much needed tools for assessing the quality of raw data, and integrating data derived from different sources and run on different devices.

References

Alaiya, A. A., B. Franzen, et al. (2000). "Cancer proteomics: from identification of novel markers to creation of artifical learning models for tumor classification." Electrophoresis 21(6): 1210-7.

Alizadeh, A. A., M. B. Eisen, et al. (2000). "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling." Nature 403(6769): 503-11.

Ballantyne, C. M. and M. L. Entman (2002). "Soluble adhesion molecules and the search for biomarkers for atherosclerosis." Circulation 106(7): 766-7.

Banks, R. E., M. J. Dunn, et al. (2000). "Proteomics: new perspectives, new biomedical opportunities." Lancet 356(9243): 1749-56.

Cash, P. (2000). "Proteomics in medical microbiology." Electrophoresis 21(6): 1187-201.

Celis, J. E., H. Wolf, et al. (2000). "Bladder squamous cell carcinoma biomarkers derived from proteomics." Electrophoresis 21(11): 2115-21.

Champagne, B., P. Tremblay, et al. (1998). "Proteolytic cleavage of ICAM-1 by human neutrophil elastase." J Immunol 161(11): 6398-405.

Chen, G., T. G. Gharib, et al. (2002). "Proteomic analysis of lung adenocarcinoma: identification of a highly expressed set of proteins in tumors." Clin Cancer Res 8(7): 2298-305.

Dickson, D. W. (2001). "Alpha-synuclein and the Lewy body disorders." Curr Opin Neurol 14(4): 423-32.

Doyle, H. A. and M. J. Mamula (2002). "Posttranslational protein modifications: new flavors in the menu of autoantigens." Curr Opin Rheumatol 14(3): 244-9.

Emmert-Buck, M. R., R. F. Bonner, et al. (1996). "Laser capture microdissection." Science 274(5289): 998-1001.

Engvall, E. and P. Perlman (1971). "Enzyme-linked immunosorbent assay (ELISA). Quantitative assay of immunoglobulin G." Immunochemistry 8(9): 871-4.

Franzen, B., S. Linder, et al. (1997). "Analysis of polypeptide expression in benign and malignant human breast lesions." Electrophoresis 18(3-4): 582-7.

Gras, R., M. Muller, et al. (1999). "Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection." Electrophoresis 20(18): 3535-50.

Grundke-Iqbal, I., K. Iqbal, et al. (1986). "Abnormal phosphorylation of the microtubule-associated protein tau (tau) in Alzheimer cytoskeletal pathology." Proc Natl Acad Sci U S A 83(13): 4913-7.

Gygi, S. P., D. K. Han, et al. (1999). "Protein analysis by mass spectrometry and sequence database searching: tools for cancer research in the post-genomic era." Electrophoresis 20(2): 310-9.

Gygi, S. P., Y. Rochon, et al. (1999). "Correlation between protein and mRNA abundance in yeast." Mol Cell Biol 19(3): 1720-30.

Hamdan, M. and P. G. Righetti (2002). "Modern strategies for protein quantification in proteome analysis: advantages and limitations." Mass Spectrom Rev 21(4): 287-302.

Herbert, B. and P. G. Righetti (2000). "A turning point in proteome analysis: sample prefractionation via multicompartment electrolyzers with isoelectric membranes." Electrophoresis 21(17): 3639-48.

Hewitt, J. E. and P. K. Grewal (2003). "Glycosylation defects in inherited muscle disease." Cell Mol Life Sci 60(2): 251-8.

Liotta, L. A., V. Espina, et al. (2003). "Protein microarrays: Meeting analytical challenges for clinical applications." Cancer Cell 3(4): 317-25.

Liotta, L. A. and E. C. Kohn (2001). "The microenvironment of the tumour-host interface." Nature 411(6835): 375-9.

MacBeath, G. (2002). "Protein microarrays and proteomics." Nat Genet 32 Suppl: 526-32.

Mann, M., R. C. Hendrickson, et al. (2001). "Analysis of proteins and proteomes by mass spectrometry." Annu Rev Biochem 70: 437-73.

Meehan, K. L., J. W. Holland, et al. (2002). "Proteomic analysis of normal and malignant prostate tissue to identify novel proteins lost in cancer." Prostate 50(1): 54-63.

Miller, J. C., H. Zhou, et al. (2003). "Antibody microarray profiling of human prostate cancer sera: Antibody screening and identification of potential biomarkers." Proteomics 3(1): 56-63

Ornstein, D. K., C. Englert, et al. (2000). "Characterization of intracellular prostate-specific antigen from laser capture microdissected benign and malignant prostatic epithelium." Clin Cancer Res 6(2): 353-6.

Parchi, P., R. Castellani, et al. (1996). "Molecular basis of phenotypic variability in sporadic Creutzfeldt-Jakob disease." Ann Neurol 39(6): 767-78.

Paweletz, C. P., B. Trock, et al. (2001). "Proteomic patterns of nipple aspirate fluids obtained by SELDI-TOF: potential for new biomarkers to aid in the diagnosis of breast cancer." Dis Markers 17(4): 301-7.

Petricoin, E. F., A. M. Ardekani, et al. (2002). "Use of proteomic patterns in serum to identify ovarian cancer." Lancet 359(9306): 572-7.

Petricoin, E. F. and L. A. Liotta (2002). "Proteomic analysis at the bedside: early detection of cancer." Trends Biotechnol 20(12 Suppl): S30-4.

Poon, T. C., T. T. Yip, et al. (2003). "Comprehensive proteomic profiling identifies serum proteomic signatures for detection of hepatocellular carcinoma and its subtypes." Clin Chem 49(5): 752-60.

Pradhan, A. D., N. Rifai, et al. (2002). "Soluble intercellular adhesion molecule-1, soluble vascular adhesion molecule-1, and the development of symptomatic peripheral arterial disease in men." Circulation 106(7): 820-5.

Rosty, C., L. Christa, et al. (2002). "Identification of hepatocarcinoma-intestine-pancreas/pancreatitis-associated protein I as a biomarker for pancreatic ductal adenocarcinoma by protein biochip technology." Cancer Res 62(6): 1868-75.

Sanchez, J. C., P. Wirth, et al. (1997). "Simultaneous analysis of cyclin and oncogene expression using multiple monoclonal antibody immunoblots." Electrophoresis 18(3-4): 638-41.

Sarto, C., S. Frutiger, et al. (1999). "Modified expression of plasma glutathione peroxidase and manganese superoxide dismutase in human renal cell carcinoma." Electrophoresis 20(17): 3458-66.

Seow, T. K., R. C. Liang, et al. (2001). "Hepatocellular carcinoma: from bedside to proteomics." Proteomics 1(10): 1249-63.

Soldes, O. S., R. D. Kuick, et al. (1999). "Differential expression of Hsp27 in normal oesophagus, Barrett's metaplasia and oesophageal adenocarcinomas." Br J Cancer 79(3-4): 595-603.

Sorlie, T., C. M. Perou, et al. (2001). "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications." Proc Natl Acad Sci U S A 98(19): 10869-74.

Srinivas, P. R., M. Verma, et al. (2002). "Proteomics for cancer biomarker discovery." Clin Chem 48(8): 1160-9.

Taylor-Papadimitriou, J. and A. A. Epenetos (1994). "Exploiting altered glycosylation patterns in cancer: progress and challenges in diagnosis and therapy." Trends Biotechnol 12(6): 227-33.

Uitto, J. and J. R. Lichtenstein (1976). "Defects in the biochemistry of collagen in diseases of connective tissue." J Invest Dermatol 66(02): 59-79.
decreased synthesis of type I collagen in osteogenesis imperfecta.

Unlu, M., M. E. Morgan, et al. (1997). "Difference gel electrophoresis: a single gel method for detecting changes in protein extracts." Electrophoresis 18(11): 2071-7.

Unwin, R. D., M. A. Knowles, et al. (1999). "Urological malignancies and the proteomic-genomic interface." Electrophoresis 20(18): 3629-37.

Weekes, J., C. H. Wheeler, et al. (1999). "Bovine dilated cardiomyopathy: proteomic analysis of an animal model of human dilated cardiomyopathy." Electrophoresis 20(4-5): 898-906.

Westhead, D. R., Parish, J. H., and R. M. Twyman (2002). BIOINFORMATICS. BIOS Scientific Publishers Limited, Oxford, U.K.

Wulfkuhle, J. D., L. A. Liotta, et al. (2003). "Proteomic applications for the early detection of cancer." Nat Rev Cancer 3(4): 267-75.

Wulfkuhle, J. D., D. C. Sgroi, et al. (2002). "Proteomics of human breast ductal carcinoma in situ." Cancer Res 62(22): 6740-9.

Xiao, Z., B. L. Adam, et al. (2001). "Quantitation of serum prostate-specific membrane antigen by a novel protein biochip immunoassay discriminates benign from malignant prostate disease." Cancer Res 61(16): 6029-33.

Zhou, G., H. Li, et al. (2002). "2D differential in-gel electrophoresis for the identification of esophageal scans cell cancer-specific protein markers." Mol Cell Proteomics 1(2): 117-24.

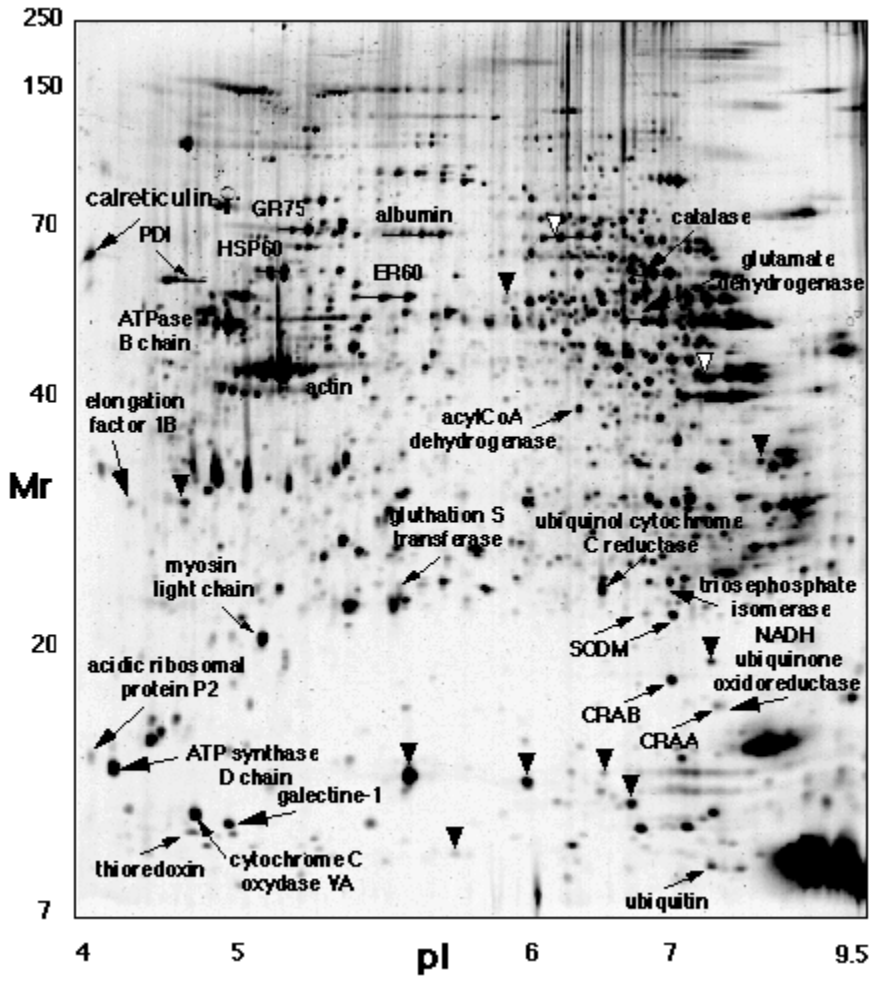

Figure 1: Silver-stained human kidney 2D-PAGE map. (SWISS-2DPAGE map from ExPASy)

Cecilia Sarto, Alessandro Marocchi, Jean-Charles Sanchez, Daniela Giannone, Séverine Frutiger, Olivier Golaz, Marc R. Wilkins, Giavarlo Doro, Francesco Cappellano, Graham J. Hughes, Denis F. Hochstrasser, Paolo Mocarelli. *Renal cell carcinoma and normal kidney protein expression.* **Electrophoresis (1997) 18, 599-604**

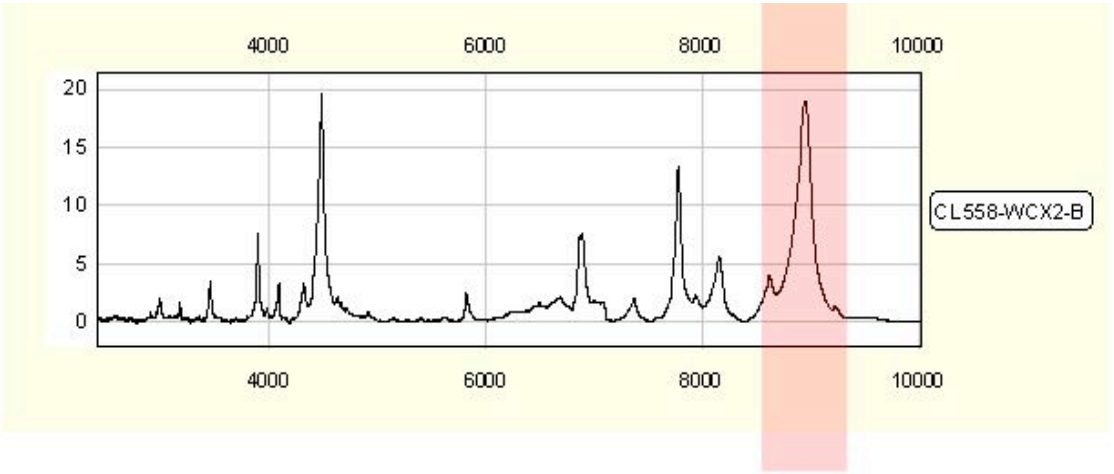

Figure 2: Low resolution mass spectrum

(http://clinicalproteomics.steem.com/)