PROTEOMICS: UNIVERSAL BIOMARKERS FOR INHERITED AND INFECTIOUS DISEASES

Alvin Fox, Karen Fox, and Marvin Vestal[1]

University of South Carolina, School of Medicine, Columbia, SC AND Automated Methods LLC, Columbia, SC Phone: 803 733 3288; Fax 1 803 733 3192; E-mail: afox@med.sc.edu

[1]Virgin Instruments LLC, Boston, MA ,

INTRODUCTION

We shall discuss here the potential of proteomics (which defines the amino acid sequence of all proteins expressed by a specific cell type under specific growth conditions) and biomarker discovery, for diagnosis of inherited diseases (or cancer) and non-culture based biodetection of infectious diseases (in clinical samples) or environmental monitoring. Many of the instrumental developments in proteomics have come from the field of analytical chemistry. Thus the work is highly relevant to all with an interest in biomedical science, biology or chemistry. However since our research is particularly focused on bacteriology, emphasis will be placed here in the microbiology arena. Also recognizing that the audience of JSCAS is multi-disciplinary (and that the journal is read by administrators, researchers and teachers (including professors and their undergraduate or graduate students) I shall make no apologies that wherever possible I shall provide brief explanations of the principles behind technical terms.

The US Market for molecular infectious disease diagnostic technology is predicted to be approaching $4 billion in 2010 (http://www.clpmag.com/issues/articles/2007-11_08.asp). The market for protein biomarker discovery was $290 million in 2005 which is predicted to rise to $745 million in 2011 (Biobusiness, Biomarket Trends (genengnews.com, March 1 2007). While proteomics has great potential for the US and worldwide, as noted below, there are particular opportunities for SC. The bases of these developments are the independent revolutions that have occurred in the fields of molecular biology and analytical chemistry leading to the current inter-relatedness of genomics, proteomics and bioinformatics.

The molecular biology revolution included the development of cloning, the polymerase chain reaction (PCR) and use of restriction enzymes for recognition of sequence differences among organisms employing genetic markers. Indeed in 1993, Kary B. Mullis, received the Nobel Prize in Chemistry for the discovery of PCR. Dr. Mullis went to high school in Columbia; so SC has a history in this area. The process of marker discovery has been greatly aided in recent years by whole genome sequencing (i.e. determining the entire genetic or DNA code for an organism) also allowing a more systematic approach to biomarker discovery. While it is well known that the human genome has been sequenced and annotated, it should also be emphasized that the genomes of many common mammalian and other multi-cellular (eukaryotic) species and single celled organisms (including most common human pathogens) have also been sequenced. What used to take large groups of investigators months or years (and millions of dollars) can now be accomplished in days or weeks (depending on the size of the genome) at a fraction of the cost; although it is still expensive.

DNA AND PROTEIN MARKERS

It is anticipated the automated instrumental identification of peptide markers for human pathogens will be considerably less labor-intensive than current DNA-based approaches. Discovery of DNA markers e.g. using PCR involves: a. first defining putative sequences from the genome; b. next primers (DNA sequences that are complementary and recognize these sequences) are designed; c. PCR is tested with relevant clinical samples; d. off-line sequencing of PCR product is often performed to confirm identity in initial set-up experiments. In the case of protein markers a. (PCR) and b. (primer design) are not required and d. (sequencing) is performed on-line as part of the proteomic analysis. In other words there is only one instrumental marker discovery/analysis step in a proteomics-based approach. Genetic or protein markers for inherited diseases or cancer are determined in a similar fashion (markers discerning diseased versus normal cells).

At USC we are focusing our efforts on basic research in marker discovery (with support from the Sloan Foundation). However we also are extrapolating these concepts to helping assess the utility of automated instrumentation developed by Virgin instruments LLC (Boston, MA). Prototypes will be evaluated and modifications suggested for specific applications by Automated Methods LLC, Columbia, SC. Recent advances in proteomics technology provide accurate molecular weight (M.W.) and sequence information on peptides from protein digests with high speed and sensitivity. These advances include new mass spectrometers developed at Virgin together with more efficient methods for interfacing separations with mass spectrometry with microbial biomarker methodology developed at USC These systems provide practical solutions to the problems that have severely limited the applications of proteomics for clinical analyses. The focus of current R & D is on reproducibly detecting, identifying, and quantifying human and microbial biomarkers in: 1) plasma, serum, urine and other body fluids in the important 1 pg/mL to 1 ng/mL range and 2) environmental samples (e.g. with biodefense and homeland security applications). This new family of instruments employs 5 khz lasers providing data acquisition 25 times faster than any existing commercial mass spectrometer.

Microbiological applications will be used to give an example of the applicability of the technology. However, as noted above markers can be derived from any form of life (e.g. human, bacterium, parasite or virus). Bacterial species share specific genes (and encoded proteins) of characteristic sequence distinguishing them from other bacterial species. Differences in DNA sequence are generally detected by real-time PCR. Since the differences are small for closely related species; direct (automated Sanger sequencing) or indirect approaches (e.g. restriction digestion) are often used to detect these sequence variants.

Many forms of mass spectrometry have been successfully employed for identification of cultured microorganisms, but none of these approaches provide the sensitivity, specificity, simplicity and speed required for automated clinical identification or detection of infectious agents in human body fluids without culture which is a work in progress.

MARKER DISCOVERY

Discovery of useful biomarkers by the proposed methods requires two steps. First, fractionation, separation, and analysis protocols must be optimized for potential

biomarkers for particular strains and species to be detected at clinically relevant level in body fluids without culture. Second, peptides produced by digestion of proteins from cultured organisms must be identified and their MS-MS (tandem mass spectrometry) spectra recorded, interpreted, and stored in a searchable database together with all available information including the source, strain and species of the organism. This protocol must be sufficiently rapid, robust, and simple to allow its use in a clinical setting. Thus, while limited separation and fractionation may be sufficient for the initial discovery phase, it is important to establish a protocol using proteins from the cultured samples that can be extended to reliable detection of these potential biomarkers at low levels in body fluids. The ultimate goal is detection of specific biomarkers for previously characterized pathogens, at clinically relevant concentrations, within one hour after receipt of a body fluid and to characterize fluids containing previously unknown or emerging pathogens within 48 hours.

Successful completion of this work may revolutionize clinical microbiology allowing laboratory diagnosis in real-time (with equivalent sensitivity to PCR) but also real-time identification of protein sequence variants. This could totally change the way that treatment of infectious diseases is performed in the US. The instruments would also revolutionize battlefield biodetection and counter-terrorism efforts for biological warfare agents (e.g. anthrax). Instrumentation might be purchased by every hospital and/or first responder (urban/battlefield) in the US. Each instrument, depending on sales, would be in the $200,000- one million range. Ancillary products will include disposable reagents, operator training, and up-datable data-bases of markers.

There has been a revolution in mass spectrometry leading to sequencing of the expressed protein products of genomes (proteomics). Indeed the 2002 Nobel Prize in Chemistry was awarded to Koichi Tanaka and John B. Fenn for their development respectively of matrix assisted time of flight ionization/desorption (MALDI) and electrospray ionization (ESI) mass spectrometry (MS). In both cases large molecules (including proteins and DNA) are analyzed in native form from aqueous solutions in a mass spectrometer. Scientists, whose research does not focus on mass spectrometers, are often thinking of an older technology (gas chromatography-mass spectrometry [GC-MS]). GC-MS (and more advanced GC-MS-MS) requires extensive chemical work-up to convert a marker (usually a small molecule such a fatty acid) into a suitable form for analysis in the gas phase. Indeed in the clinical microbiology field GC is now routinely used in reference laboratories for whole cell fatty acid profiling after prior growth in culture media (after conversion to FAMES, fatty acid methyl esters]). GC-MS provides additional structure information on these profiles.

Microbiologists are often not well versed in performing organic chemical reaction schemes and thus fatty acid profiling is limited to laboratories with an emphasis on microbial biochemistry. However fatty acid profiling is still considered a gold standard in taxonomy and classification and widely used in reference laboratories. There are also several companies that will provide a fatty acid profile for a fee (e.g. MIDI Inc., Newark, DE). Sample preparation for fatty acid analysis takes several hours. By comparison a recent proteomics method for identification of Bacillus anthracis developed at US takes a few minutes. The difference in time taken for the two analyses (hours versus minutes) provides a perspective on how things have changed and potential for the future.

Additionally the presence or absence of a fatty acid monomer provides considerably less specificity than a peptide sequence.

Alternatively, the genomic revolution has given us a vast array of molecular biology tools for discrimination of well-known pathogens as well as emerging infections by the presence or absence of genes or for closely related organisms, small changes in DNA sequence. It is anticipated that protein-sequence based discrimination will be as important for the next generation of clinical microbiologists and biomedical researchers.

In the newer so-called soft ionization MS technology, introduced in the 1990s and 2000s, biomolecules are analyzed without any separation of components or after separation employing high performance liquid chromatography (LC) or electrophoresis. This is performed in the liquid phase which is often aqueous in nature. Small molecules can be analyzed but the real power of the technique is in being amenable to analysis of larger molecules (e.g. peptides/proteins) without chemical pre-treatment.

For the non-mass spectrometrist, it should be pointed out that nowadays the analysis of these large molecules is primarily based on MALDI MS or ESI MS. In the former case, the sample is spotted, with a matrix, on a metal plate and allowed to air dry. When struck with a laser beam, after the plate is inserted into the MS, the matrix absorbs the light, transferring it to the molecule of interest (e.g. proteins or peptides). Generally, only a singly ionized species is produced having a single charge. In contrast, ESI MS is performed in solution and the sample is sprayed into the MS using a syringe pump. As the droplets evaporate, charges are transferred to molecules present within the droplet. Ions are produced that can have multiple charge states. Since mass analyzers generally separate by the mass-to-charge ratio, simple spectra are generated for MALDI (molecules having only one charge) but ESI spectra (reflecting mixtures of molecules each having one, few or multiple charges) are more complex. Thus MALDI MS has been more popular with biologists and biomedical researchers because of the simplicity of the spectra. However ESI MS often allows the analysis of larger molecules. An extensive knowledge of chemistry, in performing MALDI or ESI MS, is not required since the molecule is analyzed in its native form without chemical treatment. Indeed as mentioned above, in certain applications it is not necessary to employ a separation stage (i.e. LC or electrophoresis) and the sample can be analyzed directly in the MS with minimal sample pre-treatment.

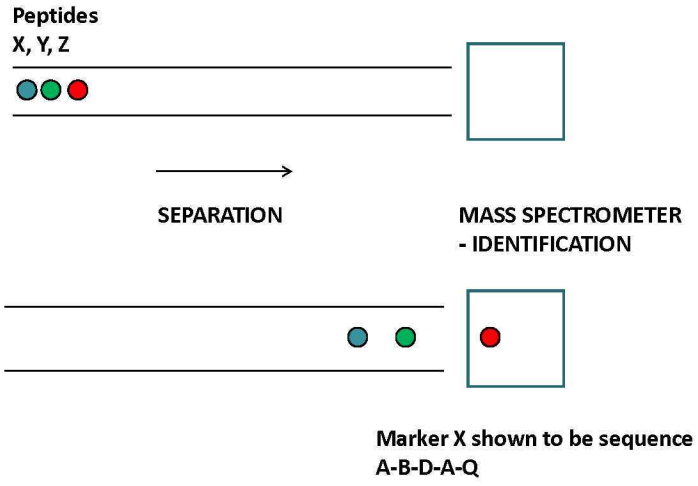
Another independent but equally important instrumental advance has been the commercial introduction of the tandem mass spectrometer (MS-MS, also in the 1990s-2000s) allowing routine sequencing of peptides. Peptides are identified in two distinct stages. First the molecular weight of the peptide is determined; they are volatilized in the MS as intact molecules. Then, for MS-MS analysis, the peptides are broken into a series of constituent mixture of peptides by breaking them at each peptide bond in the chain. For example, in the following illustration purposely simplified for clarity: a tri-peptide A-C-D (alanine-cysteine-aspartate) in sequence might generate alanine and cysteine-aspartate on MS-MS analysis. The observation of a mass equivalent to A suggests that alanine is the terminal amino acid. This is confirmed by the difference in mass between ACD and CD suggesting that CD make up amino acids 2, 3 of the peptide. The finding of a dipeptide of mass of AC suggests C is linked to A., i.e. is at position 2. The sequence is thus A-C-D. Generally the analysis is more complex (and the spectra more difficult to

interpret) since the molecules are larger (usually 10-30 mers) and fragmentation is more complex.

PROTEIN PROFILING

Direct extraction of bacterial vegetative cells or spores followed by MALDI MS analysis has become popular for bacterial identification, since it is simple to perform and mass spectra are readily interpreted. However, only high abundance peptides that are of low mass and ionize readily are observed (e.g. 2-10,000 mass range). Generally the spectra are plotted as the amount of each protein present (as defined by its molecular weight, MW); Unfortunately MW alone is not sufficient to identify a characteristic biomarker and one must rely on the entire spectrum, this often referred to as mass profiling or fingerprinting. These spectral comparisons can be made by eye-balling but generally pattern recognition-based computer programs are employed; unfortunately there is often considerable variability in the spectra from run-to-run or between samples complicating data interpretation.

Alternatively the sequences of individual proteins can be determined using MS-MS. The presence of an individual marker can be determined with great confidence and one does not have to depend on the consistency of the mass profile which can sometimes be problematic. For example, in our recent work, the MWs of small acid soluble proteins (SASPs) were measured using MALDI MS and confirmed by ESI MS. ESI-MS-MS analysis was employed for the generation of sequence-specific information. The analysis consists of simply extracting the samples and analyzing the extract directly into the MS-MS instrument. ESI-MS revealed a prominent doublet of SASPs for all strains in these studies. The first SASP varied in mass and sequence between B. anthracis versus B. cereus/B. thuringiensis. The second SASP had the same MW for all strains correlating with species (or clade; there are two for B. cereus) and served as an internal standard allowing comparison between mass spectra in this study and previous ones. The entire sample extraction and analysis takes under 10 min.



PROTEOMICS

It should be emphasized protein profiling is distinct from classical proteomics based approaches which involve more time-consuming sample processing. Proteomics often employs 2D gel electrophoresis to isolate individual protein spots which are then digested in situ, usually with trypsin, to generate peptides of characteristic masses that are subsequently analyzed using MALDI MS analysis. The sequences of each peptide in the tryptic digest can then be identified by MALDI MS-MS analysis. Alternatively, after tryptic digestion of whole cells, the mixture of peptides is subjected to on-line liquid LC-ESI-MS-MS analysis (either one or two dimensional). In either case, separation (electrophoresis or chromatography respectively) is important in reducing the complexity of mixtures for analysis by the mass spectrometer but increases the learning curve in implementing the MS technology for routine applications.

Proteomics is quite time consuming and technically demanding and is best used for comparing the relatedness of two strains or species (or cancer versus normal cells). Bioinformatics can be used to relate identified peptides to those predicted to be present in proteins coded by whole genomes. In theory, a novel strain could be categorized in this fashion. This requires bioinformatics analysis of multiple strains of each pair or group of organisms to be discriminated which is complex and labor intensive. Alternatively LC-MS-MS, or 2D-gel electrophoresis/MS-MS, could be used for the process of marker discovery. Once the markers have been discovered, simple MS or MS-MS assays (performed in aqueous solution) could be employed for routine analysis. The analogy is the discovery of DNA markers by whole genomic comparison followed by real-time PCR for diagnostic applications.

Sensitivity and specificity are both of particular importance; in trace detection of microbial markers in complex biological matrices such as infected body fluids or tissues. Indeed there is usually a separation (e.g. LC for proteins) or PCR amplification of the target (DNA) marker in clinical diagnosis. In both instances this serves to increase the concentration of the marker relative to background derived from other components of the matrix, this simplifies the analysis.

Real-time PCR is the current leading non-culture-based technology for determination of infection. More discriminating PCR-MS (mass spectrometry) for bacterial DNA markers was developed in the US through collaboration between the University of South Carolina and Pacific Northwest National Laboratory. In this case the mass accuracy is sufficient to discriminate two PCR products differing by a single nucleotide substitution (e.g. adenine to thymine [9 mass units] or guanine to cytosine [40 mass units]). An automated commercial PCR-MS instrument was subsequently introduced by Ibis Biosciences Inc., Carlsbad, CA based on these principles. PCR-MS has several additional stages, versus PCR, including post-PCR sample clean-up and robotic transfer from PCR to MS module. Thus PCR-MS is currently performed as a reference laboratory technique. For example, it has been successfully used for determining nucleotide composition, for strain typing in epidemiological studies of outbreaks of respiratory infections with Streptococcus, Hemophilus or Neisseria.

CONCLUDING REMARKS

Once simple automated instruments are widely available, diagnosis of disease variants or bacterial infection using protein markers will involve minimal sample preparation and would be complementary (but simpler to perform) than widely used

molecular biology approaches that often involve multiple sample processing steps (e.g. PCR/off-line sequencing). However, the use of mass spectrometers is still daunting to many in the microbiological, biological and biomedical communities. Hopefully this review will contribute to removing some of the mystery behind what is ultimately a simple tool that is highly amenable to unattended sample preparation and computer-based decision making. Genomics is reaching maturity but high-through-put proteomics still has great potential for growth. SC has an opportunity to not only benefit from these developments but to be ahead of the curve and indeed lead them in the US to fruition.

REFERENCES

Fenselau, C. (ed.). 1994. Mass spectrometry for the characterization of microorganisms. American Chemical Society, Washington, D.C.

Fox, A., S. L. Morgan, L. Larsson, and G. Odham, ed. 1990. Analytical microbiology methods: chromatography and mass spectrometry. Plenum Press, New York, N.Y.

Fox A. Mass spectrometry for species or strain identification (after culture) or directly (without culture): past, present and future. J. Clin. Microbiol. 44: 2677–2680. 2006.

Odham, G., L. Larsson, and P.-A. Mardh (ed.). 1984. Gas chromatography/mass spectroscopy applications in microbiology. Plenum Press, New York, N.Y.

Wilkins, C. L., and J. O. Lay. 2005. Identification of microorganisms by mass spectrometry. John Wiley and Sons, Hoboken, N.J.