ELSEVIER

SELDI-TOF-based serum proteomic pattern diagnostics for early detection of cancer

Emanuel F Petricoin[1,*] and Lance A Liotta[2]

Proteomics is more than just generating lists of proteins that increase or decrease in expression as a cause or consequence of pathology. The goal should be to characterize the information flow through the intercellular protein circuitry that communicates with the extracellular microenvironment and then ultimately to the serum/plasma macroenvironment. The nature of this information can be a cause, or a consequence, of disease and toxicity-based processes. Serum proteomic pattern diagnostics is a new type of proteomic platform in which patterns of proteomic signatures from high dimensional mass spectrometry data are used as a diagnostic classifier. This approach has recently shown tremendous promise in the detection of early-stage cancers. The biomarkers found by SELDI-TOF-based pattern recognition analysis are mostly low molecular weight fragments produced at the specific tumor microenvironment.

Addresses
[1]FDA-NCI Clinical Proteomics Program, Office of Cell and Gene Therapies, Center for Biologic Evaluation and Research, Food and Drug Administration, Bethesda, MD 20892, USA
[2]FDA-NCI Clinical Proteomics Program, Laboratory of Pathology, Center for Cancer Research, NCI, NIH, Bethesda, MD 20892, USA
*e-mail: petricoin@cber.fda.gov

Current Opinion in Biotechnology 2004, **15**:24–30

This review comes from a themed issue on
Analytical biotechnology
Edited by Keith Rose

0958-1669/$ – see front matter

DOI 10.1016/j.copbio.2004.01.005

Abbreviations

AI	artificial intelligence
MS	mass spectroscopy
SELDI	surface-enhanced laser desorption/ionization
TOF	time-of-flight

Introduction

Despite the urgent need to discover serum biomarkers for the early detection of disease, the number of new biomarkers reaching routine clinical use remains dismally low [1•,2]. The low molecular weight range (<15 000 Da) of the serum proteome, although until very recently largely uncharacterized, promises to contain a rich source of previously undiscovered biomarkers [3•], as biological processes give rise to cascades of enzymatically generated and proteolytically clipped biomarker fragments. The blood proteome is changing constantly as a consequence of the perfusion of organ systems the underlying pathophysiology of which adds to, subtracts from or modifies the circulating proteome. Thus, even if these small enzymatically generated peptide fragments are far removed from the actual disease, they are not merely 'epiphenomena' and can retain specificity for the disease because the process that generated the clipping in the first place can arise within the uniqueness of the disease tissue microenvironment. These low molecular weight molecules exist below the range of detection achieved by conventional two-dimensional gel electrophoresis, as they cannot be efficiently separated by gel-based techniques [3•]. As a result, investigators have turned to mass spectroscopy (MS), which exhibits its optimal performance in the low mass range [4,5].

Assuming that the low molecular weight and low abundance biomarkers contain important diagnostic information, the search for these markers usually begins with a separation step to remove the abundant high molecular weight 'contaminating' proteins, such as albumin, thyroglobulin and immunoglobulins. The analysis can then focus on the low abundance region of the proteome. From a physiological perspective, however, this removal might be the wrong approach to take for biomarker discovery, akin to throwing the baby out with the bathwater. Free-phase, unbound low molecular weight molecules will be rapidly cleared through the kidney filtration system, significantly reducing the concentration of these biomarkers to a level below the detection limits of any routine clinical testing device and certainly below the detection limits of most mass spectrometers. In the face of the vast excess of high molecular weight serum proteins, however, it is likely that low abundance and low molecular weight biomarkers will become bound to large high-abundance carrier proteins and be protected from kidney clearance just on the basis of the tremendous stoichiometric differences that arise between the relative abundances of albumin and a low abundance clipped diagnostic fragment [6,7]. Thus, the bound low abundance and low molecular weight carrier proteins possess a half-life that is many orders of magnitude larger than that of free-phase small molecules. Circulating carrier proteins have been recently found to act as a reservoir for the accumulation and enrichment of bound low molecular weight biomarkers, integrating and storing diagnostic information like a capacitor stores electricity [8••,9••].

To be effective, a clinically useful disease and cancer-related biomarker should be measurable in an accessible body fluid such as serum, urine or saliva. As these fluids are a protein-rich information source that possibly

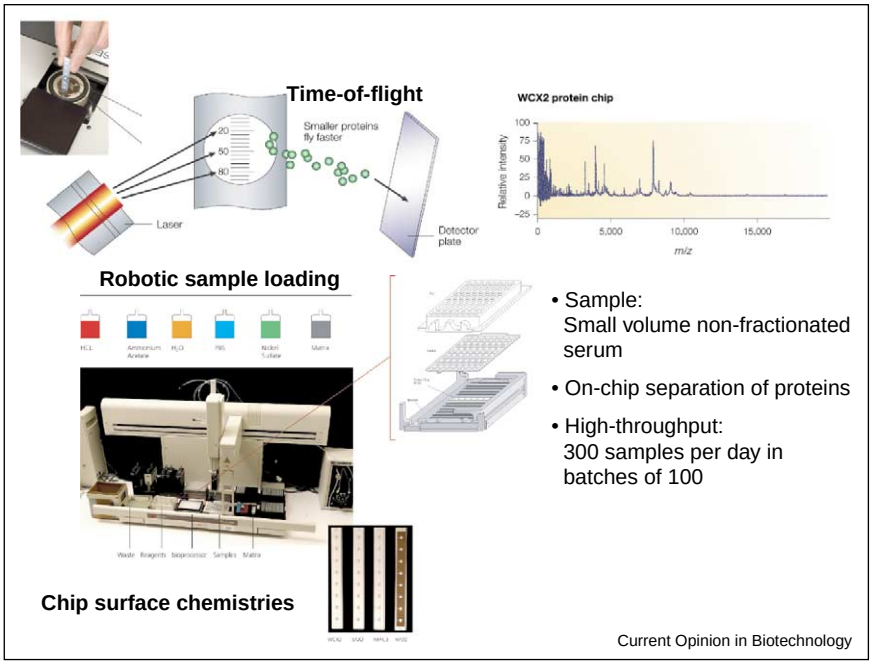
contains traces of whatever the blood has encountered on its constant perfusion and percolation throughout the body, proteomics might offer the best chance of discovering early-stage changes. In the past, the search for biomarkers for early disease and toxicity detection has been a low-throughput approach, looking for overexpressed proteins in blood that are aberrantly shed into the circulation as a consequence of the disease process. There are potentially tens of thousands of intact and cleaved proteins in the human serum proteome, so finding the elusive single disease-related protein is like searching for a needle in a haystack, requiring the laborious separation and identification of each and every protein biomarker. Moreover, it is unlikely that these elusive single biomarkers will ever be used for the early detection of disease, as clinical applications will be eventually applied to a human population exhibiting vast heterogeneity, not only in their respective proteomes but also in the underlying disease process.

Initial attempts to employ MS for the identification of biomarkers for cancer have been very promising [10••,11,12•–15•,16,17,18•]. Unlike past attempts that start with a known single marker candidate, proteomic pattern analysis begins with high dimensional data (e.g. containing greater than 10 000 data points per sample and upwards of 1–2 million data points per patient), usually produced by high-throughput MS. This method attempts, without bias, to identify patterns of low molecular weight biomarkers as ion peak features within the spectra as the diagnostic itself.

Serum proteomic pattern diagnostics: producing the mass spectra

Although investigators have used a variety of different bioinformatic algorithms for pattern discovery, the most common analytical platform comprises a ProteinChip® Biomarker System-II (PBS-II, a low-resolution time-of-flight [TOF] mass spectrometer). Herein, samples are ionized by surface-enhanced laser desorption/ionization (SELDI), a protein chip array-based chromatographic retention technology that allows for direct mass spectrometric analysis of analytes retained on the array (Figure 1). Only a subset of the proteins in the serum bind to the chromatographic surface of the chip and the unbound proteins are washed away. The adherent proteins are treated with acid (so that they can become ionized) and then dried down onto the surface. The capture region

Figure 1



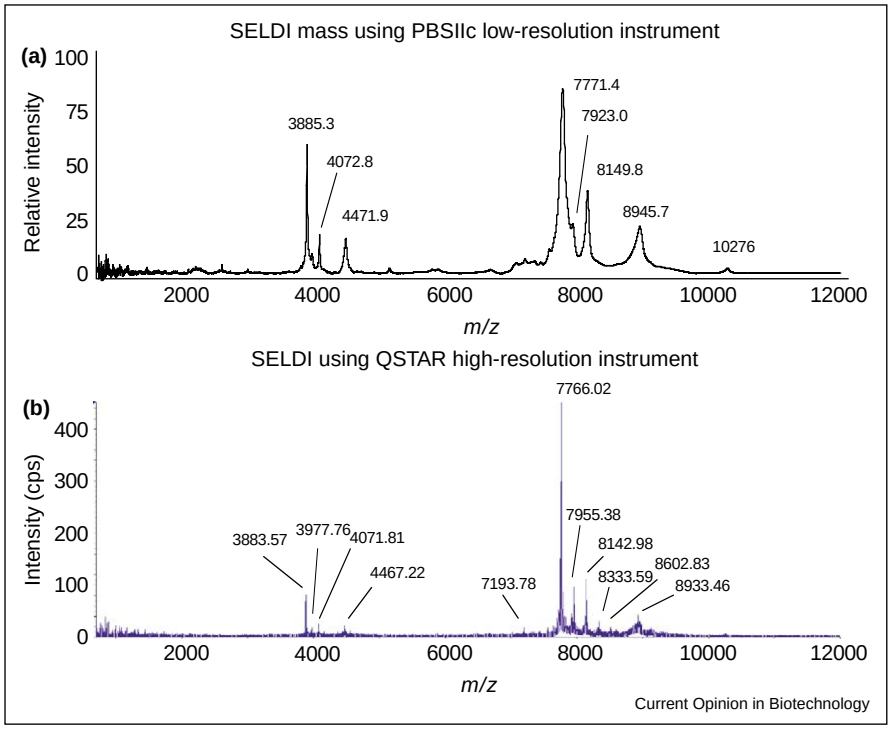
Surface-enhanced laser desorbtion and ionization (SELDI) technology. This type of proteomic analytical tool is a class of MS instrument that is useful for the high-throughput proteomic fingerprinting of serum. Using a robotic sample dispenser, 1 μL of raw serum is applied to the surface of a protein-binding chip. Some laboratories pre-fractionate their samples beforehand, whereas others perform complex analysis for optimizing binding by diluting into a myriad of pH and salt permutations. Regardless of the upfront manipulations, but based on the underlying SELDI chip chemistry and the pH and buffer used, only a subset of the proteins in the sample bind to the surface of the chip. The bound proteins are then treated with a MALDI matrix, washed and dried. The chip, containing multiple patient samples, is inserted into a vacuum chamber where it is irradiated with a laser. The laser desorbs the adherent proteins, causing them to be launched as ions. The time-of-flight (TOF) of the ion before detection by an electrode is a measure of the mass to charge (*m/z*) value of the ion. The ion spectra can be analyzed by computer-assisted tools that classify a subset of the spectra by their characteristic patterns of relative intensity.

containing individual captured serum protein samples, dried down on a row of spots, is inserted into a vacuum chamber and a laser beam is fired at each spot. The laser energy blasts off (desorbs) the ionized proteins, and the ionized proteins fly down the vacuum tube toward an oppositely charged electrode. The mass to charge (m/z) value of each ion is estimated from the time it takes for the launched ion to reach the electrode; small ions travel faster. Therefore, the spectrum provides a time-of-flight signature of ions ordered by size. Recently this concept has been extended to high-resolution MS, as it was found for ovarian cancer detection that higher resolution MS data generates diagnostic models with higher sensitivity and specificity. This results from both the increased number of peaks seen and the much better reproducibility between and within machine runs [15•]. Moreover, the spectral resolution of the lower resolution instrumentation cannot separate specific ions that are close in mass/charge, which can cause multiple specific discreet ions to coalesce into a single peak. Whether or not low-resolution SELDI will be displaced entirely by high-resolution SELDI as a clinical diagnostic platform remains to be seen, as more comparison studies are required.

The high-resolution mass spectrometer used in our studies is a hybrid quadrupole time-of-flight mass spectrometer (QSTAR pulsar *i*, Applied Biosystems Inc.; http://www.appliedbiosystems.com/) fitted with a ProteinChip® array interface (Ciphergen Biosystems Inc.; http://www.ciphergen.co.uk/) and externally calibrated twice a day using a mixture of known peptides. As a point for analytical comparison, the QSTAR-TOF-MS (routine resolution ~8000) can completely resolve species differing by an m/z of only 0.375 (e.g. at m/z 3000), whereas complete resolution of species with the PBS-II-TOF-MS (routine resolution ~150) is only possible for species that differ by an m/z of 20 (Figure 2).

In a clinical setting where a pattern test might eventually be employed as a diagnostic, it will be important to determine overall spectral quality and to develop spectral release specifications such that variances introduced into the process can be evaluated and monitored. Day-to-day, lot-to-lot and machine-to-machine variances brought in from sample handling and/or storage and shipping conditions will need to be evaluated and understood as well as the mass spectrometer itself. To that end, we employ a pooled reference standard sample (SRM–1951A), obtained from the National Institute of Standards and Technology (NIST; http://www.nist.gov), which is randomly applied to one spot on each protein array as a quality control for overall process integrity, sample preparation and mass spectrometer function. Additionally, for spectral quality control, quality assurance and spectral

Figure 2



Comparison of low-resolution and high-resolution SELDI-TOF mass spectra. Spectra from the same weak cation exchange chip (queried at the same spot on the same chip) were generated on either **(a)** a PBS-IIc (Ciphergen Biosystems, Inc.) low-resolution instrument or **(b)** a QSTAR pulsar *i* (Applied Biosystems Inc.,) high-resolution instrument.

release specification, all spectra are subjected to plotting by total ion current (total record count), average/mean and standard deviation of amplitude, chi-square and *t* test analysis of each ion or *m*/*z* range in which the individual ion values are grouped based on the inherent resolution of the mass spectrometer and the amplitude values of the individual ions summed together into one grouping, and quartile plotting measures using JMP (SAS Institute, Cary, NC) software. Stored procedures developed in-house are also used to verify spectra before any pattern discovery takes place. Process measures are checked by analyzing the statistical plots of the NIST serum reference standard, and spectra that fail statistical checks for homogeneity are eliminated from in-depth modeling and analysis. This type of upfront analysis is critical so that it is possible to compare the total analytical variance obtained for the constant NIST reference sample with the variance of the clinical sample populations. The total variance of the reference sample should be no less than that for the clinical specimens.

ProteinChip® arrays (Ciphergen Biosystems Inc.) are typically processed in parallel using a Biomek Laboratory workstation (Beckman-Coulter; http://www.beckmancoulter.com/) modified to make use of a ProteinChip® array bioprocessor (Ciphergen Biosystems Inc.). The bioprocessor holds twelve ProteinChips®, each having eight chromatographic 'spots', allowing 96 samples to be processed in parallel and the matrix to be applied using a liquid robotic handling station (Genesis Freedom 200, TECAN; http://www.tecan.com/).

Serum proteomic pattern diagnostics: uncovering the pattern classifiers

A typical low-resolution SELDI-TOF proteomic profile will have up to 15 500 data points that record data between 500 and 20 000 *m*/*z*, with a high-resolution mass spectrometer generating over 400 000 data points. Artificial intelligence (AI)-based systems that learn, adapt and gain experience over time are uniquely suited for proteomic data analysis, because of the huge dimensionality of the proteome itself.

We begin our proteomic pattern analysis by first exporting the raw data file generated from the high resolution QSTAR mass spectra into tab-delimited files that generate approximately 350 000 data points per spectrum. The individual *m*/*z* values are then grouped together into buckets or 'bins' of data using a function of 400 ppm based on the inherent resolution and mass accuracy of the instrument, such that all data files possess the same number of identically spaced and fixed *m*/*z* values (e.g. the *m*/*z* bin sizes linearly increase from 0.28 at *m*/*z* 700 to 4.75 at *m*/*z* 12 000). This binning process actually condenses the number of data points from 350 000 to exactly 7084 points per sample and the *m*/*z* range of the bins gradually increases as a function of the resolution capacity of the machine. The 400 ppm binning function was based on a value 10 times the estimate of the routine mass drift of the QSTAR-TOF machine obtained by external and internal calibration results (5–40 ppm), as a conservative drift bracket.

The data are then randomly separated into equal groups for training and testing. A variety of pattern recognition tools have been successfully used for mining mass spectral data [10••,11,12•–15•,16,17,18•]. One tool that has shown great promise, and which was used in our first studies [10••] for ovarian cancer detection, is one that combines elements from genetic algorithms and self-organizing adaptive pattern recognition systems [19–22] (Correlogic Systems, Inc., http://www.correlogic.com/). Genetic algorithms organize and analyze complex datasets as if they were information comprised of individual elements that can be manipulated through a computer-driven analog of a natural selection process. Self-organizing systems cluster data patterns into similar groups. Adaptive systems recognize novel events and track rare instances. The genetic algorithm component of the analysis begins with the random generation of a population of 1500 subsets of combinations of ion features of the mass spectra. This number was chosen based on adequate coverage of the data, with a heuristic that no value can be duplicated within each of the 1500 subsets. Each subset in the population specifies the identities of the exact *m*/*z* values in each data stream, but not their relative amplitude. The number of ion features in the subset ranges from 5 to 20.

Data normalization is an important element of pattern recognition, as bias introduced by protein chip quality, mass spectrometer instrumentation and operator variance can effect overall spectral performance. Moreover, it is likely that different data normalization procedures will generate different selected ions, especially in a clustering algorithm where multiple ion features are used as the pattern. As MS is not inherently quantitative, scalar intensity changes might be apparent, yet the overall pattern may not change. One way to typically normalize MS data is to divide the amplitudes at each *m*/*z* value within any randomly generated pattern subset by the largest value within that subset. In this way, differences in spectral quality that can emanate from biases such as protein chip variance and not from the inherent disease process itself can be minimized. Also, this method allows for low-amplitude features to contribute substantially to the classification. The spectra are normalized according to the formula:

$$NV = (V - Min)/(Max - Min)$$

NV is the normalized value, V the intensity value for the specific randomly chosen *m*/*z* bin in question, Min the intensity of the smallest intensity value of any of the *m*/*z* bins within the randomly selected pattern and Max the

maximum intensity of the *m*/*z* bin within the randomly selected pattern. This equation linearly normalizes the peak intensities so as to fall within the range 0 to 1. Each of the randomly selected 1500 subset patterns is then subjected to a fitness test.

The fitness test in these analyses is the ability of the combined ion amplitude values of any candidate subset to specify a lead cluster map that generates homogeneous clusters containing only diseased subjects or unaffected subjects used in the training sets. The lead cluster map is a self-organizing, adaptive pattern-recognition algorithm that uses Euclidean distance to group vectors of data. The map begins as an empty *N*-dimensional space, where *N* is the number of *m*/*z* features in the data vector. The optimal discriminatory pattern is identified by finding the best combination of *m*/*z* bins for which normalized ion intensity values in *N*-dimensional space create a unique identifier or cluster of identifiers. Any given training sample is compared for its proximity to previously defined clusters of diseased and unaffected subjects in *N*-space. If an *N*-dimensional identifier vector from a subject in the training group falls within the decision boundary of an existing cluster, then the subject is classified as belonging to that group. For these studies, the decision boundary is defined as 10% of the maximum distance allowed in the space. This corresponds to a 90% pattern match; thus, the decision boundary is referred to as the 90% boundary. If the data vector does not fall within the 90% decision boundary of any existing cluster in the model it is used to establish a new cluster and is identified as a new observation. The process is repeated once for each vector in the collection of training data.

Those subpopulation patterns that best discriminate the training set are more likely to survive the culling of the population to the original population size (e.g. 1500) and contribute to the next generation of fit candidate patterns. The progeny of the most-fit patterns are generated through crossover and mutation of the 5–20 specific *m*/*z* bin values within each subset. Each subset is evaluated by its ability to accurately distinguish the two training set populations. As a result, each successive population of subsets is, on average, more fit than its predecessor. To ensure that the algorithms do not trend to less than near optimal decision points, a 'mutation' rate is built into the process such that 0.02% of the *m*/*z* bin values are randomly re-chosen. Crossover operations are of single point type and are randomly selected in each mating. For example, if there are five *m*/*z* bin values there can be four crossover points. The genetic algorithm is iterated for at least 250 generations or until a lead cluster map that homogeneously separates diseased from unaffected is generated. The lead cluster map that best separates diseased from unaffected is deployed for validation using blinded test sets.

Test data, not used during the training process, are then analyzed in the following steps. The data are normalized as described above and the normalized relative amplitudes of the test sample spectra at the *N*-defined *m*/*z* values bins are used to fix a point in *N*-dimensional space. The Euclidean distance vector is then calculated between this point and the center of all clusters (both cancer and unaffected) formed by the training set. If the unknown test vector falls inside the 90% boundary surrounding any centroid, then it is classified as being a member of that cluster and given a probability score based on its proximity to the theoretical center of the cluster and the number of records within that cluster. Otherwise, it is scored as a 'new cluster'. The results from the test set of data are used for determination of sensitivity, specificity and positive predictive value of the patterns.

As each new patient is validated through pathological diagnosis using retrospective or prospective study sets, its input can be added to the ongoing clustering using the same models. The AI tool learns, adapts and gains experience through constant vigilant updating. In fact, it is possible to generate not just one but multiple combinations of proteomic patterns from a single mass spectral training set, each pattern combination readjusting as the models get better in the adaptive mode.

MS-based diagnostics: a view to the future

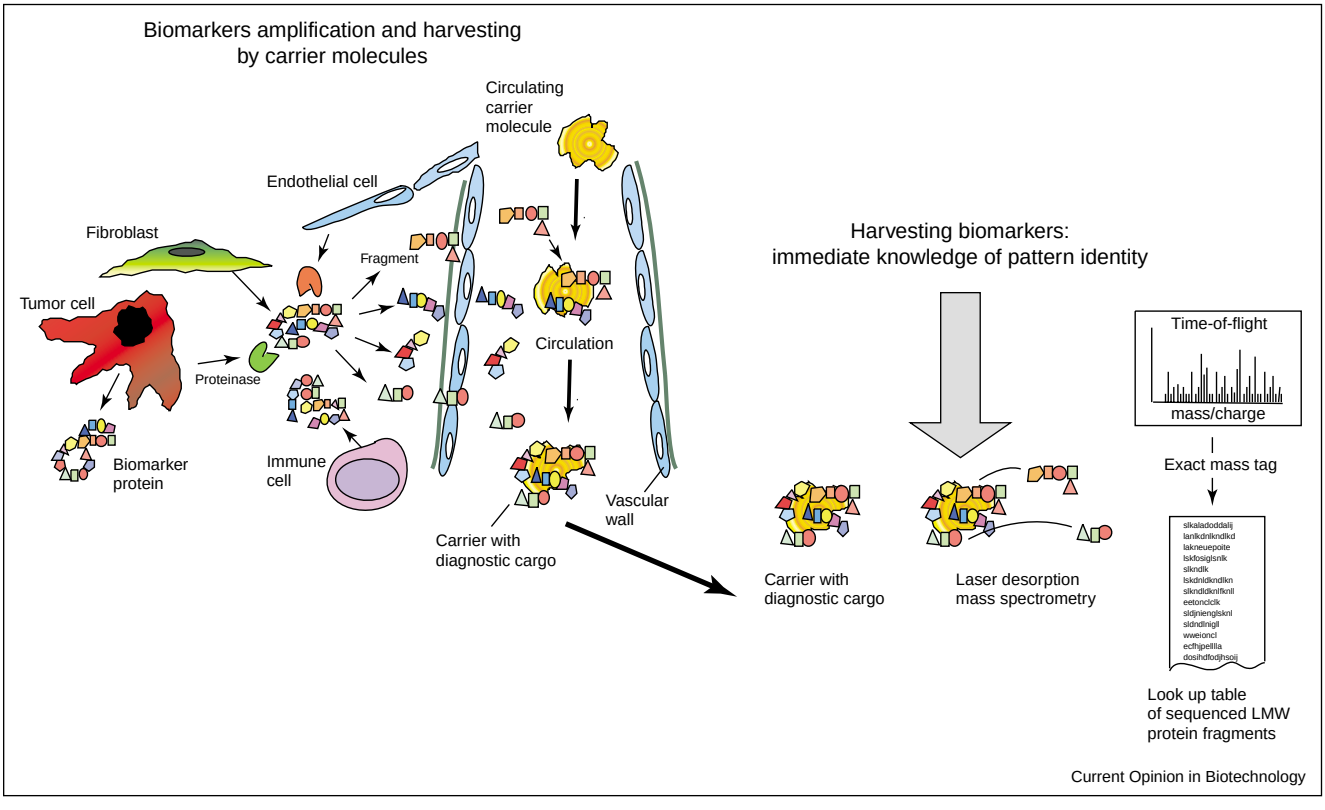
MS analysis of the low molecular weight range of the serum/plasma proteome is a rapidly emerging frontier for biomarker discovery and clinical diagnostics. Proteomic pattern diagnostics represents a new paradigm for disease detection and is very amenable to the high-throughput world of clinical diagnostics. The analysis requires only a drop of blood and the mass spectra patterns obtained in less than 30 min. SELDI-TOF proteomic pattern analysis, in theory, can be applied to any biological state. Using this approach, the pattern itself, independent of the identity of the proteins or peptides, is the discriminator, and may be clinically useful immediately before the underlying identities are eventually discerned. Depending on the identity of the signature ion, it may, or may not, be desirable or even feasible to proceed directly to develop a serum immunoassay for the individual biomarker. This is because the ion amplitude of MALDI-TOF does not directly reflect the concentration of the given biomarker associated with the ion. Moreover, if the biomarker is the cleaved version of a larger protein, it may be difficult to generate antibodies that only recognize the cleaved version and do not cross-react with the parent species. A possibility exists to develop polyclonal antibodies for specific capture and following binding the entirety of the recognized entities, including the diagnostic fragment, can be eluted and analyzed via MS.

MS platforms of the future, coupled to pattern-recognition algorithms, might become superior to antibody-based

immunoassays. MS can generate complex proteomic spectra from an extremely small volume of blood in only a few seconds, in effect sensing the presence of hundreds to thousands of events simultaneously and almost instantaneously without the need to develop antibodies for each analyte. Current MS platforms have a sensitivity in the femtomolar range, and will become even more sensitive in subsequent generations of the technology. Mathematically it should be obvious that a pattern of multiple biomarkers will contain a higher level of discriminatory information than a single biomarker alone, particularly for large heterogeneous patient populations. Currently, our group is planning the first large-scale clinical trials for FDA approval to explore and validate this concept for monitoring ovarian cancer. As evidence of the growing acceptance of this new paradigm, large commercial reference laboratories have begun initiatives to explore the use of MS proteomic patterns for routine diagnosis (http://www.questdiagnostics.com; http://labcorp.com).

As we now know that the vast majority of these biomarkers exist in association with circulating high molecular mass carrier proteins, these findings shift the focus of biomarker analysis to the carrier protein and its biomarker content. The proteomic pattern that emanates from this microenvironment might signal the presence of an early-stage lesion. Under this hypothesis, the discriminatory molecules are likely to be metabolic products, enzymatic fragments, modified proteins, peptides or cytokines. In fact, the most important biomarkers might be normal host proteins that are aberrantly clipped or reduced in abundance. A pattern analysis approach takes into consideration the loss or gain of ions within the spectra. Past conventional protocols for biomarker discovery discard the abundant 'contaminating' high molecular mass proteins, to focus on the low mass range. Unfortunately, this procedure removes most of the important diagnostic biomarkers. We can now develop new tools, created at the intersection of proteomics and nanotechnology,

Figure 3



Biomarker amplification and harvesting by carrier molecules. Low molecular weight peptide fragments, produced within the unique tissue microenvironment and generated as a consequence of the disease process, permeate through the endothelial cell wall barrier and trickle into the circulation. Here, these fragments are immediately bound with circulating high-abundance carrier proteins, such as albumin, and protected from kidney clearance. The resultant amplification of the biomarker fragments enables these low-abundance entities to be seen by MS-based detection and profiling. In the future, harvesting nanoparticles, engineered with high affinity for binding, can be instilled into the collected body fluids or injected directly into the circulation to bind with the disease- and toxicity-related information archive. These nanoparticles and their bound diagnostic cargo can then be directly collected, filtered over engineered filters and queried by high-resolution MS. A 'look up table', where the exact identities of each of the peaks will be compared against the accurate mass tag of each of the peaks within the spectra, will enable the simultaneous identification of each entity within the pattern as well as allowing the discovery of the diagnostic pattern itself.

whereby nanoharvesting agents can be instilled into the circulation (e.g. derivatized gold particles) or into the blood collection device to act as 'molecular mops' that soak up and amplify the biomarkers that exist [8••] (Figure 3). These nanoparticles, with their bound diagnostic cargo, can be directly queried via MS and the low molecular weight and enriched biomarker signatures revealed.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. • Anderson NL, Anderson NG: **The human plasma proteome: history, character, and diagnostic prospects**. *Mol Cell Proteomics* 2002, **1**:845-867.

The number of new FDA approved biomarkers has dropped off substantially over the past decade despite advances in genomics.

2. Ward JB Jr, Henderson RE: **Identification of needs in biomarker research**. *Environ Health Perspect* 1996, **104(Suppl 5)**:895-900.

3. • Tirumalai RS, Chan KC, Prieto DA, Issaq HJ, Conrads TP, Veenstra TD: **Characterization of the low molecular weight human serum proteome**. *Mol Cell Proteomics* 2003, **2**:1096-1103.

Details the first comprehensive analysis of the low molecular weight region of the serum proteome, an unexplored information archive that may contain large numbers of biomarkers.

4. Kantor AB: **Comprehensive phenotyping and biological marker discovery**. *Dis Markers* 2002, **18**:91-97.

5. McDonald WH, Yates JR III: **Shotgun proteomics and biomarker discovery**. *Dis Markers* 2002, **18**:99-105.

6. Cojocel C, Maita K, Baumann K, Hook JB: **Renal processing of low molecular weight proteins**. *Pflugers Arch* 1984, **401**:333-339.

7. Maack T: **Renal handling of low molecular weight proteins**. *Am J Med* 1975, **58**:57-64.

8. •• Liotta LA, Ferrari M, Petricoin E: **Clinical proteomics: written in blood**. *Nature* 2003, **425**:905.

MS-generated proteomic patterns are now known to be comprised mostly of protein fragments and cleavage products generated as a result of on-going physiologic processes. In the future, harvesting nanoparticles designed to act as biomarker attractants could be used *in vivo* or *ex vivo* as 'molecular mops' that can be directly analyzed by MS.

9. •• Mehta A, Ross S, Lowenthal MS, Fusaro V, Fishman DA, Petricoin EF, Liotta LA: **Biomarker amplification by serum carrier protein binding**. *Dis Markers* 2003, **19**:1-10.

Biomarker amplification is found to take place because low molecular weight peptide and protein fragments bind to carrier proteins that enrich and increase the overall abundance of these biomarkers over time.

10. •• Petricoin EF III, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC, Liotta LA: **Use of proteomic patterns in serum to identify ovarian cancer**. *Lancet* 2002, **359**:572-577.

Serum Proteomic Pattern Diagnostics, where pattern recognition tools are coupled to MS profiles, is described for the first time, with the technology developed and utility shown for early stage ovarian cancer detection.

11. Petricoin EF III, Mills GB, Kohn ES, Liotta LA: **Proteomic patterns in serum and identification of ovarian cancer**. *Lancet* 2002, **360**:170-171.

12. • Li J, Zhang Z, Rosenzweig J, Wang YY, Chan DW: **Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer**. *Clin Chem* 2002, **48**:1296-1304.

Description of highly accurate serum-based proteomic patterns for breast cancer detection.

13. • Petricoin EF III, Ornstein DK, Paweletz CP, Ardekani A, Hackett PS, Hitt BA, Velassco A, Trucco C, Wiegand L, Wood K *et al.*: **Serum proteomic patterns for detection of prostate cancer**. *J Natl Cancer Inst* 2002, **94**:1576-1578.

Description of highly accurate serum-based proteomic patterns for prostate cancer detection and discrimination from prostatitis and BPH even in the diagnostic indeterminate range of PSA (4–10 ng/ml).

14. • Adam BL, Qu Y, Davis JW, Ward MD, Clements MA, Cazares LH, Semmes OJ, Schellhammer PF, Yasui Y, Feng Z, Wright GL Jr: **Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men**. *Cancer Res* 2002, **62**:3609-3614.

Description of highly accurate serum-based proteomic patterns for prostate cancer detection and discrimination from prostatitis and BPH.

15. • Conrads TP, Zhou M, Petricoin EF III, Liotta L, Veenstra TD: **Cancer diagnosis using proteomic patterns**. *Expert Rev Mol Diagn* 2003, **3**:411-420.

Description of high-resolution QqTOF MS proteomic patterns for increased accuracy of ovarian cancer detection.

16. Petricoin E III, Liotta LA: **Counterpoint: the vision for a new diagnostic paradigm**. *Clin Chem* 2003, **49**:1276-1278.

17. Petricoin EF, Liotta LA: **Mass spectrometry-based diagnostics: the upcoming revolution in disease detection**. *Clin Chem* 2003, **49**:533-534.

18. • Hingorani SR, Petricoin EF III, Maitra A, Rajapakse V, King C, Jacobetz MA, Ross S, Conrads TP, Veenstra TD, Hitt BA *et al.*: **Preinvasive and invasive ductal pancreatic cancer and its early detection in the mouse cancer cell.** 2003, **10**:6-21.

Description of a new mouse model for pancreatic cancer development. Significantly, highly accurate serum-based proteomic patterns were found that could identify premalignant lesions from inflammatory and benign processes.

19. Holland JH: *Adaptation in Natural and Artificial Systems: an Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. 3rd Edition. Cambridge, MA: MIT Press; 1994.

20. Kohonen T: **Self-organizing formation of topologically correct feature maps**. *Biol Cybern* 1982, **43**:59-69.

21. Kohonen T: **The self-organizing map**. *Proc IEEE* 1990, **78**:1464-1480.

22. **Pattern classification by distance functions.** In *Pattern Recognition Principles*. Edited by Tou JT, Gonzalez R: Reading, MA: Addison Weley Publishing Company, 1974;75-109.