A newcomer to the -omics era, proteomics, is a broad instrument-intensive research area that has advanced rapidly since its inception less than 20 years ago. Although the ‘wet-bench’ aspects of proteomics have undergone a renaissance with the improvement in protein and peptide separation techniques, including various improvements in two-dimensional gel electrophoresis and gel-free or off-gel protein focusing, it has been the seminal advances in MS that have led to the ascension of this field. Recent improvements in sensitivity, mass accuracy and fragmentation have led to achievements previously only dreamed of, including whole-proteome identification, and quantification and extensive mapping of specific PTMs (post-translational modifications). With such capabilities at present, one might conclude that proteomics has already reached its zenith; however, ‘capability’ indicates that the envisioned goals have not yet been achieved. In the present review we focus on what we perceive as the areas requiring more attention to achieve the improvements in workflow and instrumentation that will bridge the gap between capability and achievement for at least most proteomes and PTMs. Additionally, it is essential that we extend our ability to understand protein structures, interactions and localizations. Towards these ends, we briefly focus on selected methods and research areas where we anticipate the next wave of proteomic advances.

Key words: label-free, mass spectrometry, post-translational modification, protein interaction, proteogenomics, proteomics.

STRATEGIES AND METHODS AND INSTRUMENTS! OH MY!

More, better, peptides

In bottom-up analyses, proteins are converted into peptides prior to MS analyses. The ideal peptide for analysis by collisional-activation MS/MS (tandem MS) should be 7–35 residues long, be protonated, and have a low charge state (z) and a high mass-to-charge ratio (m/z). Tryptic peptides from *Saccharomyces cerevisiae* have an average length of 8.4 residues and a C-terminal basic residue [4]. The success in using trypsin for analysis of the yeast proteome has led to a widespread adoption. In a non-scientific survey conducted on 11 February 2011, the PubMed database (http://www.ncbi.nlm.nih.gov/pubmed/) was queried using ‘proteomics’ as the search term. Of 100 primary research publications, 100 (100 %) reported the use of trypsin to fragment
target proteins prior to MS analysis. Although this result might be slightly surprising, there is no doubt that a large majority of proteomic analyses employ trypsin [5]. Recombinant proteolytic-grade trypsin is widely available, relatively inexpensive and easy to use. Trypsin cleaves polypeptide chains exclusively C-terminal to arginine and lysine residues, yielding peptides with basic residues at the C-terminus which typically give informative high mass y-ion series and easy to interpret spectra.

In addition to being a time-intensive process not well suited to automation, there are other disadvantages to using trypsin for target protein fragmentation; thermostability is relatively poor and it is labile to rapid autolysis at alkaline pH values. More importantly, lysine and arginine residues are not uniformly distributed throughout the proteome, especially in membrane proteins [6], and 100% sequence coverage cannot be achieved. Some peptides will be 'too short' or 'too long' for MS analysis, or cleavage will be blocked, or the results will be obscured by PTMs (post-translational modifications). Clearly trypsin alone is not enough. By analogy, where would we be today if EcoRI [7] were the only restriction enzyme used for DNA analysis or manipulation? Manipulation of DNA became progressively easier as more and varied restriction enzymes became readily available [8].

Refinements to the methods for trypsin digestion (microwave acceleration, ultrasonic assistance, use of nanobiocatalysts or quantum discs) [9–11] fail to address the essential problem of sequence coverage. Will solution to the 'tryptic flaws' come from a broader use of other proteases, discovery of new proteases or perhaps a more widespread use of chemical fragmentation? Yes. Or, more correctly, yes to all of these and more. Sequencing-grade chymotrypsin is commercially available and can be used to fragment proteins, cleaving C-terminal to phenylalanine, tyrosine, tryptophan and leucine [12]. Target-protein fragmentation by chymotrypsin alone is seldom an improvement over trypsin, but using the two proteases together often leads to significantly improved sequence coverage (for example [13]).

There are additionally the so-called ‘sequencing endoproteases’ [14], which include Lys-C, which cleaves only C-terminal to lysine residues, Arg-C, Glu-C and Asp-N. Used in conjunction with trypsin, these proteases also yield improved sequence coverage [4]. Disadvantages of using these endoproteases include their relatively high cost, catalytic inefficiency and that digestion with the alternative proteases can result in peptides that lack the C-terminal basic residue typically responsible for the intense y-ion series. It has been recently demonstrated that the latter problem can be overcome by application of post-proteolysis N-terminal guanidination [15].

Chemical cleavage can be a useful adjunct to protease treatment. Methionine residues comprise less than 2% of protein amino acids, and treatment with cyanogen bromide cleaves the Met-X bond [16]. By itself, cyanogen bromide cleavage is not adequate for fragmentation of most proteins, but it can be easily used in conjunction with trypsin digestion or with other methods of chemical fragmentation, such as formic acid treatment, which cleaves aspartate–proline bonds [17]. Other methods of chemical fragmentation specifically target tryptophan or histidine residues [18].

Just as recombinant DNA-based research accelerated with the availability of numerous restriction enzymes, proteomics as a field would benefit greatly from the availability of additional defined-sensitivity endoproteases. As with endonucleases, these proteases might come from a survey of micro-organisms. Alternatively, it might be possible to use in vitro evolution to modify the specificity of extant proteases [19]. Whether it involves new proteases and new methods for chemical fragmentation, or simply new combinations of proven methods, the goal of 100% sequence coverage should be aggressively pursued.

**Reduced chromatography times are necessary for LC (liquid chromatography)-MS/MS analyses**

Improvements to mass analyser scanning frequency without equal advances in sensitivity will place a greater burden upon pre-fractionation prior to MS. Wet-bench manipulations are the antithesis of both sample throughput and reproducibility. Although proteomics began with two-dimensional gel profiling, it has slowly been shifting away from this approach in favour of gel-free MS-based profiling. One of the reasons is the serial, and therefore time-consuming, nature of protein identification with two-dimensional gels. Another reason is the biased nature of two-dimensional gels, with membrane and ‘extreme’ proteins (high/low mass or pl proteins) being under-represented [20]. But the use of electrophoresis-based techniques is widespread, and improvements, including immobilized pH gradients [21] and multiplex-staining of gels using specific fluorescent dyes [22,23], encourage applications. It is important, however, to consider the low-throughput predicament that accompanies protein identification after electrophoresis.

As an example, consider the identification of 200 protein spots excised from two-dimensional gels. Analysis of all 200 samples by LC-MS/MS would require approximately 2 weeks with 30 min chromatographic separations and at least one blank chromatographic run between samples to minimize peptide carryover. The poor throughput and high carryover associated with LC-MS/MS is a problem that can be addressed by reducing run-time, and testing and developing C-18 matrices with reduced ‘memory effects’. A progressive shift from ion spray (millilitre flow rates) to capillary (microlitre flow rates) to nanospray (nanolitre flow rates) chromatography has been afoot since the advent of proteomics. Higher-pressure lower-flow chromatography for online nLC (nano-LC)-MS/MS (sometimes referred to as UPLC (ultra-performance liquid chromatography)) improves sensitivity and also reduces chromatography time by reducing peak width, particularly for targeted proteomics [24,25]. Nanoflow LC has the advantage of near-zero dead volume coupled to ultra-narrow peak widths (5 s) and is capable of reproducible separation of complex peptide samples in the 10–100 ng range. But although high sensitivity can be achieved, separation times are long and capacity is limited [26]. More recent developments include
two-dimensional nLC separations [27–29], and the introduction of nanospray microfluidics which use laser-etched column paths within an inert polymer chip [30]. Reducing both analysis and sample loading time will help to shorten the throughput and carry-over concerns associated with LC-MS, as this strategy continues to supplant use of electrophoresis-based methods.

Going global will require improvements in instrument sensitivity

The overall efficiency of a mass spectrometer can be separated into three components: (i) efficiency of ionization (ions/molecule); (ii) transmission of the ion optical system (ions out/ions in); and (iii) detection efficiency (detected pulses out/ions in) [26]. As more biological questions are asked at the tissue and cellular level (e.g. laser-capture microdissection), it will be increasingly challenging to isolate sufficient protein for comprehensive analysis using extant instruments. Pursuit of improved instrument efficiency/sensitivity is ongoing and has been achieved through refinements in ionization efficiency and ion detection (Table 1). For example, the recent development of ‘S-lens’ technology, which captures and focuses ions more efficiently than previous generation ion transfer tubes, has led to a reported 10-fold increase in ion-capture efficiency [31]. It is likely that ongoing improvements in instruments and strategies will continue to improve MS sensitivity.

In contrast with bottom-up analysis, in top-down proteomics intact protein molecular ions are introduced into the mass analyser and are subjected to gas-phase fragmentation. This requires both the use of instruments with high mass accuracy, and the use of deconvolution algorithms [32–34]. The two major advantages of the top-down strategy are the potential access to complete protein sequences and the ability to localize and characterize PTMs. In addition, the time-consuming protein digestion required for bottom-up methods is eliminated. The mass accuracy possible with high magnetic fields has made FT-ICR (Fourier-transformed ion-trap instruments [20,43]. However, with the development of high-energy collision dissociation-based methods, quantitative fragmentation could become routine [44].

The third quantitative approach, label-free quantification, is the most facile and inexpensive quantification strategy. It can be further refined into spectral counting and peak integration strategies, with variations therein [45–49]. Spectral counting can be thought of as the MS equivalent to estimating mRNA abundance by counting the number of ESTs (expressed sequence tags) in an EST library, the so-called ‘digital Northern’ [50]. The accuracy and sensitivity of any sampling technique is dependent upon sample size. Like mRNA quantification, peptide/protein quantification by spectral counting is instrument-dependent. Owing to their fast-scanning capabilities, ion-trap MS instruments are generally considered to be the most proficient for spectral counting. Current linear ion traps routinely acquire 5–10 scans/s, or approximately 18 000–36 000 scans during a 60 min LC-MS/MS analysis. Improvements to even this instrumentation scanning speed are possible, ensuring that spectral counting will remain a viable quantification approach in the future. Given the ease of this approach compared with peak integration and the label-based approaches [45–49,51], it is probable that this will become the preferred strategy for discovery-based comparative proteomics, whereas peak integration will remain the more rigorous strategy for comparative (and absolute) quantification for targeted samples [45,52].

BETTER TARGETS

Quantitative proteomics . . . don’t put a label on it

It is absolutely critical that proteomics transcend qualitative protein identification, and move on to quantitative analysis. Ideally this would involve the universal adoption of a single inexpensive, and sensitive yet rigorous, method. Sadly, few of our laboratories are located in Camelot, and until there is development/adoption of methods which allow facile direct comparisons within/between/among datasets it will remain impossible to directly compare results [3]. Presently, proteomics broadly encompasses relative (or differential) and absolute quantification, within the constraints of gel-based compared with gel-free sample preparation. Gel-free proteomics is typically separated into quantitative sub-categories based upon: (i) metabolic labelling, (ii) chemical labelling or (iii) label-free [40,41]. Each approach has advantages and limitations that must be considered during experimental design. Metabolic labelling requires the quantitative incorporation of stable isotope-labelled substrate (typically amino acids) into proteins in vivo, which is best accomplished using cell cultures. The QconCAT strategy is a fascinating elaboration of the stable isotope-labelled absolute quantification peptide-based approach. However, descriptions of recent applications (for example [42]) clearly indicate that the method is not yet ready for ‘prime-time’ adoption. Substantial problems remain relative to design of the synthetic gene that comprises the basis for quantification, and even in the detection of both the native and quantotypic peptides [42]. Chemical labelling is a post-protein isolation technique that typically involves primary-amine reactive chemistry through the use of commercially available activated isobaric tags that fragment during MS/MS analysis [20]. This method, typically referred to as isobaric tagging, allows simultaneous analysis of four to eight samples, on the basis of 1 amu (atomic mass unit)-resolved low-mass fragments. There is the advantage that the method is system-independent once labelling is done post-isolation. Sensitivity has been an issue, owing to the incomplete fragmentation of parent ions well-known to those studying protein PTMs using ion-trap instruments [20,43]. However, with the development of high-energy collision dissociation-based methods, quantitative fragmentation could become routine [44].

The third quantitative approach, label-free quantification, is the most facile and inexpensive quantification strategy. It can be further refined into spectral counting and peak integration strategies, with variations therein [45–49]. Spectral counting can be thought of as the MS equivalent to estimating mRNA abundance by counting the number of ESTs (expressed sequence tags) in an EST library, the so-called ‘digital Northern’ [50]. The accuracy and sensitivity of any sampling technique is dependent upon sample size. Like mRNA quantification, peptide/protein quantification by spectral counting is instrument-dependent. Owing to their fast-scanning capabilities, ion-trap MS instruments are generally considered to be the most proficient for spectral counting. Current linear ion traps routinely acquire 5–10 scans/s, or approximately 18 000–36 000 scans during a 60 min LC-MS/MS analysis. Improvements to even this instrumentation scanning speed are possible, ensuring that spectral counting will remain a viable quantification approach in the future. Given the ease of this approach compared with peak integration and the label-based approaches [45–49,51], it is probable that this will become the preferred strategy for discovery-based comparative proteomics, whereas peak integration will remain the more rigorous strategy for comparative (and absolute) quantification for targeted samples [45,52].

Protein interactions/structural proteomics

The functional complexity of an organism cannot be solely defined by the number of proteins that are present or their suite of PTMs; at least equally important is the number of biologically relevant multi-protein complexes [53–55]. It has been inferred that the interactome of the relatively simple eukaryote S. cerevisiae
Table 1  Overview of mass spectrometers available for protein analysis

The sensitivity of each instrument was determined using (1) reserpine, (2) [Glu1]-fibrinopeptide B, (3) ubiquitin, (4) BSA digest or (5) verapamil. All data were obtained from brochures available at the manufacturer’s website (http://www.thermoscientific.com, http://www.waters.com, http://www.bdal.com, http://www.agilent.com and http://www.absciex.com) or through a sales representative. APGC, atmospheric pressure gas chromatography; APCI, atmospheric pressure chemical ionization; API, atmospheric pressure ionization; APPI, atmospheric pressure photoionization; ASAP, atmospheric solids analysis probe; CE/MS, capillary electrophoresis MS; CID, collision-induced dissociation; ESI, electrospray ionization; FTMS, Fourier transform MS; HCD, higher-energy C-trap dissociation; HESI, heated electrospray ionization; MMI, multi-mode ionization; MRM, multiple reaction monitoring; nano, nanospray; photo, photoionization; PQD, pulsed Q collision-induced dissociation; PTR, proton transfer reaction; SORI, sustained off-resonance irradiation.

<table>
<thead>
<tr>
<th>Model</th>
<th>Manufacturer</th>
<th>Analyser</th>
<th>Accuracy (p.p.m.)</th>
<th>Sensitivity</th>
<th>Range (m/z)</th>
<th>Resolution (FWHM)</th>
<th>Scan rate</th>
<th>Fragmentation</th>
<th>Ionization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orbitrap (with LTQ XL or Velos)</td>
<td>Thermo Scientific Hybrid</td>
<td>&lt;1 Attomole1</td>
<td>50–4000</td>
<td>&gt;100 000</td>
<td>1 spectra/s</td>
<td>CID, HCD, ETD, PQD</td>
<td>MALDI, HESI, ESI, nano, API, APCI, APCI/APPI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TSQ Vantage</td>
<td>Thermo Scientific Triple quadrupole</td>
<td>50* Attomole1</td>
<td>10–1500</td>
<td>7500</td>
<td>5000 units/s</td>
<td>CID</td>
<td>HESI, APCI/APPI, nano</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Velos</td>
<td>Thermo Scientific Linear ion trap</td>
<td>50* Attomole1</td>
<td>50–4000</td>
<td>&gt;25 000</td>
<td>33 333 units/s</td>
<td>CID, ETD, PQD</td>
<td>ESI, APCI, APCI/APPI, nano</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LTO XL</td>
<td>Thermo Scientific Linear ion trap</td>
<td>50* Attomole1</td>
<td>50–4000</td>
<td>&gt;25 000</td>
<td>16 000 units/s</td>
<td>CID, ETD, PQD</td>
<td>ESI, APCI, APCI/APPI, nano</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xevo G2 QToF</td>
<td>Waters Quadrupole TOF</td>
<td>&lt;1 Femtomole2</td>
<td>&lt;100 000</td>
<td>&gt;22 500</td>
<td>30 spectra/s</td>
<td>CID</td>
<td>ESI/APCI/ESI, APCI, APPI/APCI, nano, ASAP, APGC, TRIZAICTM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xevo TQ-S</td>
<td>Waters Tandem quadrupole</td>
<td>50* Attomole1</td>
<td>2–2048</td>
<td>Not provided</td>
<td>10 000 units/s</td>
<td>CID</td>
<td>ESI/APCI/ESI, APCI, APPI/APCI, nano, ASAP, APGC, TRIZAICTM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Synapt G2 HDMS</td>
<td>Waters Quadrupole TOF</td>
<td>&lt;1 Femtomole2</td>
<td>&lt;100 000</td>
<td>40 000</td>
<td>20 spectra/s</td>
<td>CID</td>
<td>ESI/APCI/ESI, APCI, APPI/APCI, nano, MALDI, ASAP, APGC, TRIZAICTM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SolarIX FTMS</td>
<td>Bruker Daltonics FTMS</td>
<td>&lt;1 Attomole1</td>
<td></td>
<td></td>
<td></td>
<td>CID, ETD, ECD, (SORI)-CID</td>
<td>ESI, APCI, APPI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amazon ETD</td>
<td>Bruker Daltonics Linear ion trap</td>
<td>50* Attomole1</td>
<td></td>
<td>20 000</td>
<td>52 000 units/s</td>
<td>CID, ETD/PTR</td>
<td>ESI, APCI, APPI, nano, HPLC-Chip, ESI/MALDI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>microTOF-Q II</td>
<td>Bruker Daltonics Quadrupole TOF</td>
<td>&lt;2 Attomole1</td>
<td></td>
<td>20 000</td>
<td></td>
<td>CID</td>
<td>ESI, APCI, APPI, nano, CE/MS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6490 Triple</td>
<td>Agilent Quadrupole TOF</td>
<td>50* Zeptomole</td>
<td>10–2000</td>
<td>Not provided</td>
<td>150 MRM/s</td>
<td>CID</td>
<td>HPLC-chip, ESI, APCI, APPI, MMI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6500 Q-TOF</td>
<td>Agilent Quadrupole TOF</td>
<td>&lt;2 Femtomole1</td>
<td>20–20000</td>
<td>40 000</td>
<td>20 spectra/s</td>
<td>CID</td>
<td>ESI, APCI, APCI/APCI, MALDI, HPLC-Chip</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TripleTOF 5600</td>
<td>ABSciex Triple TOF</td>
<td>&lt;1 Femtomole1</td>
<td></td>
<td>40 000</td>
<td>100 spectra/s</td>
<td>CID</td>
<td>ESI/APCI, Turbo V, nano</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Qtrap 5500</td>
<td>ABSciex Triple quadrupole</td>
<td>100* Femtomole1</td>
<td>5–1250</td>
<td>Not provided</td>
<td>12 000 units/s</td>
<td>CID/ETD</td>
<td>APCI, Turbo V, nano, ESI/APCI, photo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear ion trap</td>
<td>ABSciex Linear ion trap</td>
<td>100* Femtomole1</td>
<td>50–1000</td>
<td>9200</td>
<td>20 000 units/s</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Instruments with mass accuracy reported in Daltons were converted into p.p.m. using an m/z of 1000.
comprises 35 000 protein complexes, and that each of the proteins interacts with six non-redundant partners [56,57]. In the overall picture, it is essential that we understand how the function of multi-protein complexes differs from the functions of the individual components.

Early descriptions of protein interactions were for the most part based upon results from synthetic genetic screens, such as the yeast two-hybrid system [58], affinity tagging [59,60] or co-precipitation studies [61]. The surprisingly poor initial agreement of results using these strategies [62] indicated both the difficulty of the questions being asked and the extent to which the limitations of methods were, and continue to be, needed. Extant protein-interaction maps of complex multicellular organisms, such as Caenorhabditis elegans and Drosophila melanogaster [63,64], are based upon results from the application of a variety of computational [65], evolutionary [66], synthetic genetic [58] and even transcript co-expression screens [67]. Although there have been proteomics-based studies of these models [56,61,68], they have not yet contributed as much as they should and, ultimately, will (Figure 2). A description of any organism-scale protein-interaction network must accommodate the sub-networks arising from studies of organellar protein interactions. A relatively straightforward strategy for analysis of Megadalton-sized Arabidopsis thaliana chloroplastic multi-protein complexes has been described by Olinares et al. [69].

The utility of MS in the study of protein interactions and, simultaneously, the genesis of structural MS, are based upon the observation that non-covalent interactions can be maintained in the gas phase [53]. Advantages of using MS to study multi-protein complexes include the ability to study both symmetric and asymmetric heterogeneous complexes at picomolar concentrations, and to do this in real time (Figure 2). An early application of structural MS involved determining the stoichiometry of individual proteins in a complex [70] (Figure 2). A previously isolated complex was dissociated into individual components, which were then separated by LC prior to MS analysis. The m/z results can be converted into actual mass values by the use of maximum-entropy deconvolution [71,72]. More recently, a hybrid MALDI (matrix-assisted laser-desorption ionization) LTQ (linear trap quadrupole) Orbitrap instrument was used to resolve the composition of a nuclear pore subcomplex without prior dissociation [73].

The most common strategy used for determining the existence of a ‘complex’ is some form of co-precipitation. This method can employ antibodies, engineered proteins or might simply target complexes that are so large that they can be isolated by rate-zonal sedimentation (e.g. the pyruvate dehydrogenase complex [74]). Co-precipitated proteins can be digested and analysed by MS, either with or without including an electrophoretic or LC separation step. If the target protein can be engineered to include a purification aid, such as a His<sub>6</sub> tag, then potential complexes can be easily isolated from transformed sources. The state-of-the-art version of this strategy is referred to as tandem-affinity purification. In this case, two distinct affinity tags are employed in order to increase the stringency of washing and reduce artefactual associations that are formed only after cell disruption [64,68]. It remains critically important to independently validate any co-precipitation data before assuming concurrence in a complex. Co-precipitation can be conducted in either the presence or absence of chemical cross-linkers [75]. The cross-linking will stabilize ephemeral interactions, and, especially if cleavable cross-linkers are used, need not interfere with downstream analyses [76]. Furthermore, substantial information about the protein surfaces interacting and the specific amino acids involved can be obtained through using cross-linkers of differing lengths and functional-group targets [77].

Once the composition, stoichiometry and mass of the subunits have been determined, the next step is to identify which subunits comprise the core and which are peripheral to it (Figure 2). The peripheral subunits are more exposed, more easily unfolded and the first to be released. Thus step-wise increases in accelerating voltage can be used to identify them [55,78]. Alternatively, a surface-induced dissociation strategy could be employed for identification of peripheral components. Very large or polydisperse protein complexes can result from the transient, but still functional, association of smaller subcomplexes, complicating the generation of any sort of connectivity map (Figure 2). Furthermore, protein associations can be transient, formed in response to specific biochemical or environmental cues, further complicating analysis [57]. Finally, considering the multiplicity of potential spatial conformations, preparing a connectivity map for a multi-component complex is not trivial, but ultimately rewarding and useful in terms of functional interpretation.

Although the use of MS is the best extant method to identify the protein composition of a complex on the basis of the ability to provide an exact molecular mass, the reliance on genomic databases can at the same time be problematic. Essentially all cellular proteins are subject to multiple PTMs, which result in both increases and decreases in protein mass. Furthermore the PTMs themselves can be responsible for protein associations and complex formation. This subject is more fully addressed in a subsequent section, and has implications regarding all aspects of proteomic analysis.

The recent combination of IM (ion mobility) with MS has provided a new dimension in analysis of protein and protein complexes [55,79,80]. A travelling voltage wave propels the ions through the IM cell, reducing their transit time, which serves to
increase both sensitivity and analysis speed. In IM-MS, ions are separated on the basis of differences in charge, size and shape, and can provide information on the stoichiometry, topology and cross-section of both protein complexes and their composite subunits (Figure 2). Thus ions can be identified by their mass, whereas their overall structure can be simultaneously determined, providing insight into functional protein complexes including Megadalton-sized molecular machines such as the proteasome [81].

Use of IM-MS can increase our understanding of not only the shape of an entire multi-protein complex, but also that of individual components if analysed separately. When high-resolution protein structures are not available, it is possible to combine results from IM-MS with those from structural EM (electron microscopy) and computational modelling [80]. Unfortunately, to date only a single commercial IM-MS instrument capable of analysing large protein complexes is available [82], so widespread application of IM-MS will require either modification of extant instruments or new developments by instrument vendors.

A fascinating hybrid strategy was recently described by Richter et al. [83]. The authors describe the straightforward isolation of multi-protein complexes by cell lysis followed by rate-zonal sedimentation. By employing GraFix (gradient fixation) during centrifugation, even relatively labile complexes remained intact. After gradient fractionation, single-particle EM was used to visualize the complexes that had been adsorbed on to EM carbon films. In parallel, identically prepared specimens were digested and analysed by MS (EM carbon film-assisted digestion). The EM and MS results can be directly correlated. In their first full description of the method, it was reported that as little as 50 fmol were sufficient for a comprehensive protein description of two model complexes [83]. Given the availability of structural EM plus MS data, it would be feasible to then model protein-binding interfaces [84]. This and many other aspects of MS require a partnership with computational analysis in order to provide both a comprehensive and comprehensible ‘holo-picture’ (for example [85–88]) (Figure 2). Coverage of the roles of several important computational methods can be found in the excellent recent review by Sharon [55].

**IMS (imaging MS)**

The first description of IMS appeared in 1997, and application of this technique has increased rapidly since then. By December 2011 there were 583 entries in the PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) database that included IMS in the title and/or abstract. IMS is often referred to as MALDI imaging, on the basis of the type of instrument used [89], or occasionally as ‘mass microscopy’ [90,91].

All of the early research using IMS addressed analysis of low-molecular-mass compounds. Even today the majority of IMS-based publications are from research that has focused on metabolomic profiling or pharmaceutical drug analysis. There is, however, an increasing enthusiasm for developing improved methods for analysis of proteins/peptides [92–96].

The use of IMS for direct analysis of proteins has been most successful in studies of the low- to mid-molecular-mass proteome. Because this method analyses intact tissue or tissue slices, it avoids homogenization and separation steps, and the spatial distribution of proteins within the tissue is preserved. The process is relatively simple; a matrix is deposited on the sample followed by irradiation with a laser which desorbs and ionizes the peptides. The use of MALDI is typically coupled with TOF (time-of-flight) mass analysers where the ions are accelerated at a fixed potential, traverse a field-free flight tube where they are separated based on their \( m/z \) ratio, and are subsequently detected. The mass range of TOF analysis is virtually unlimited and capable of measuring analytes >200 kDa [97]; however, IMS studies are carried out using TOF mass analysers that have a resolving power of approximately 15000 for peptides in the range ~1500 Da. Use of FT-ICR instruments allows resolution in excess of 100 000, which would allow baseline resolution of species differing in millimass units. When analysing samples that have the same nominal mass but different exact masses, TOF-based methods cannot differentiate individual spatial distributions, whereas this would be possible using FT-ICR-based methods.

The spatial localization of discreet proteins by IMS has a lateral resolution of 10–100 \( \mu \)m. A thin (~10 \( \mu \)m) sample can be mounted on a target plate and a matrix applied to the surface using a pneumatic nebulizer (Figure 3) [98]. Spectra are recorded in a systematic fashion by moving the sample stage beneath a fixed laser position. The resulting spot array comprises an image dataset analogous to pixels in a digital photograph. Each laser-irradiated spot (pixel) gives rise to a mass spectrum correlated with the discrete X/Y co-ordinate. Thus each spot contains a dataset with thousands of channels (\( m/z \) values), each of which has its own intensity. The intensity values can be expressed in a whole-array context as a two-dimensional ion-density map. These data can then be used to generate images depicting the localization and relative intensities of hundreds of ions from the sample (Figure 3).

The current state-of-the-art requires independent verification of peptide/protein identification, typically achieved by homogenizing the sample and pre-fractionating the proteins by either electrophoresis or LC prior to MS analysis. Newly developed methods should obviate this necessity. Debois et al. [99] recently described the use of MALDI ISD (in-source decay) to fragment ions directly in the MS ion source by using specific matrices (2,5-dihydroxybenzoic acid or 1,5-diaminonaphthalene). This application has led to robust direct protein identification, and there is additionally the possibility of de novo sequence analysis.

A consistent problem with IMS is the inconsistency of the methods used for matrix application. Although there is no doubt that improved methods for matrix application will be developed, matrix-free methods are also being developed [100]. One promising matrix-free method has been termed NIMS (nanostructure initiator MS). Instead of coating the surface of the sample with matrix, in NIMS samples are placed on top of a nanostructured matrix [101–103]. A sort of ‘Western blot’ IMS hybrid has been proposed recently [104], and an interesting and potentially very useful variant of NIMS involves incorporation of the substrate for a given enzyme into the matrix [105]. Conceptually related to enzyme cytochemistry, it would be the product produced during enzyme catalysis that is detected.

As with all other aspects of MS-based proteomic analyses, IMS will continue to benefit from advances in instrument capability, and from application of new and improved methods [106,107]. Intriguing recent developments include affinity-imaging MS, and the potential to use specific chemical barcodes. Imanishi et al. [108] have described a method for phospho-peptide enrichment with claimed femtomolar sensitivity. As described, the method uses glass slides rather than intact tissue, but if one envisions adoption of a liposome-mediated delivery system, such as that described in [109], then future translation of this method to tissues or tissue slices does not seem impossible. The use of liposomes to deliver specific probes (chemical barcodes), coupled with analysis by TOF-SIMS (secondary ion MS) has the potential to localize either specific domains or individual proteins at the single-molecule level [109]. Hopefully there will be a future synergism among researchers in MS, proteomics and cell biology.
that will push IMS towards the kind of quantum advances made in light microscopy in recent years (for example [110]).

The (whole) proteome

Although the precise number of genes that comprise the human genome remains elusive, the most commonly stated number is approximately 23,000. Although this is more than enough to generate a complex jigsaw puzzle, it pales in comparison with the size and complexity of the corresponding proteome! Proteome diversity has been estimated as two to three orders of magnitude greater than predicted by the encoding genome (>1,000,000 molecular species of proteins) [111].

A paradox has been developing in recent years relative to proteome complexity. On the one hand, it is common knowledge that proteins are post-translationally modified and occur in different ‘isoforms’. On the other hand, biology disclaims PTMs by linking protein names directly with functions. Optimal sequence coverage today is achieved through a combination of bottom-up plus top-down proteomics strategies. Simple identification of a peptide, regardless of how technically difficult this might be, fails to consider that the original primary gene product is sure to exist as the combinatorial sum of several differentially modified forms [112]. Especially with the concept of systems biology looming on the horizon, it is essential that quantitative proteomic analyses address the range of protein species arising from PTMs.

It is generally accepted that there are >200 distinct PTMs. As previously noted, the combinatorial consequences are staggering. Mining of large MS datasets to discover PTM peptides requires careful quality control using techniques such as decoy databases (reversed or randomized forms of the database used for querying). By using such an approach the FDR (false-discovery rate) for any database search can be determined, and it has been determined that accurate precursor mass (<5 p.p.m.) is essential to keep FDR values below 1%. Thus it will always be an advantage to employ workflows and instruments that offer the best practical mass accuracy. One strategy improved accuracy to the p.p.b. range by internal calibration with the residual electron donor analyte from ETD (electron transfer dissociation) analysis [31]. Of course, this approach requires a high mass accuracy instrument coupled to ETD. Improvements in mass accuracy will continue to be important as this area moves forward.

The most studied [113,114], and to some extent most understood [115], PTM is reversible protein phosphorylation. With the arrival of phosphoproteomics as a popular subdiscipline, novel enrichment and MS analysis approaches have been developed and we anticipate the rapid future expansion of phosphorylation-site mapping. Most phosphopeptide-enrichment strategies utilize metal-affinity chromatography, either transition metals (Fe, Ga and Co) or metal oxides (titanium and zirconium). There are an increasing number of indications that the transition metal and metal oxide approaches are more complementary than redundant in terms of their phosphopeptide selectivity, and are most productively used in a parallel or integrated fashion [116]. Strong cation-exchangers are an alternative to metal-affinity chromatography [117,118], and immunoprecipitation seems to work well with phosphotyrosine-containing peptides [119]. Other approaches have recently been demonstrated, including a variant of hydrophilic-interaction chromatography, termed electrostatic-repulsion hydrophilic-interaction chromatography [120]. Improvements in phosphopeptide enrichment will certainly continue in the coming years, and it is our hope that the reproducibility of using these techniques will be one of the developments.

Improvements in MS instrument sensitivity and speed have greatly aided discovery-based PTM mapping, most notably the release of the dual-pressure linear ion trap that can function either as a stand-alone [29] or when coupled to an Orbitrap mass

---

Figure 3 IMS

Thick samples can be sectioned, subjected to imaging and then positioned on the target, followed by matrix deposition using a pneumatic nebulizer. The sample is then irradiated with a laser which desorbs and ionizes the peptides. The use of MALDI is typically coupled with TOF mass analysers where the ions are accelerated at a fixed potential, traverse a field-free flight tube where they are separated by their m/z ratio, and subsequently detected. Spectra are systematically recorded by moving the sample stage beneath a fixed laser position. The resulting spot array comprises a dataset analogous to pixels in a digital image. Each laser-irradiated spot (pixel) gives rise to a mass spectrum correlated to the discrete X/Y co-ordinate. The data can then be used to generate images depicting the localization and relative intensities of ions from the sample.
The dual pressure instrument offers improved ion-source transmission and more efficient ion extraction for the higher-energy collision dissociation cell. The improvements in instrument design and performance allowed a 10-fold increase in ion transmission and a 2-fold increase in scanning speed, the combination of which translated to near routine identification of thousands of phosphorylation sites from a single biological sample.

It now appears that the next big thing in the field of PTM will be protein lysine acetylation [122]. Unlike N-terminal residue acetylation, lysine acetylation is reversible and can thus play a regulatory role. Lysine acetylation was initially discovered as a PTM of histone proteins, but it now seems as if the occurrence and diversity of this PTM increases with each new issue of the biochemistry/proteomics/MS journals (for example [123–125]). It has been suggested that protein lysine acetylation will ultimately be as common as phosphorylation. This remains to be seen, but it is noteworthy that there have already been more lysine-acetylated proteins identified in Escherichia coli than phosphoproteins [122].

For bottom-up analyses, proteases other than trypsin should be considered. Once the proteins have been fragmented, specific methods of lysine-acetylated peptide enrichment can be used, which will increase discovery rates. Lysine-acetylated peptides do not seem to fragment very differently from their unmodified lysine counterparts in electrospray MS/MS [126].

Protein lysine residues can also be reversibly N-methylated [127] or modified by formation of an isopeptide bond during post-translational conjugation with the small peptide modifiers ubiquitin/SUMO (small ubiquitin-related modifier)/Nedd (neural-precursor-cell-expressed developmentally down-regulated) 8 [128–130]. Although the small peptide modifiers are typically considered as a component of targeted proteostasis, there is also evidence for a role in the control of catalytic activity [131]. In some instances protein purification under denaturing conditions is necessary to preserve protein ubiquitination/SUMOylation/Neddylation [132].

There are relatively few instances where the complete suite of PTMs has been mapped for a specific protein. One case, however, where this has been achieved is the lens beaded filament protein filensin [133]. A total of nine phosphorylation sites were mapped. Additionally, filensin is proteolytically processed at Asp 431 and Leu 68, and the resulting new N-termini are N-myristoylated and N-acetylated respectively. Finally, aspartic acid isomerization to isoaspartic acid occurs at Asp 138, yielding a total of 14 distinct PTMs. The number of possible combinations of this number of PTMs (2 14 ) is 16384, well beyond our abilities to separate and/or detect! But even the entirely reasonable number of three, four or five potential PTMs could result in a complex mixture of eight peptides. A virtual comparison of a parent profile with that of the same peptide plus three PTMs is presented in Figure 4. Although it is unlikely that all proteins are as extensively modified as filensin, it seems equally unlikely that there are many proteins with no PTMs. Many aspects of the new biology are fusions between experimental and computational analyses (for example [134,135]). Efforts towards defining the complete proteome clearly represent an area that will benefit from such a hybrid approach.

### Cultivate a computer science friend

As an example of the potential value of a computational/experimental partnership for proteome analysis, we will briefly address recent developments in the study of protein phosphorylation. Prediction algorithms can be either general or kinase-specific. As this is written there are more than a dozen widely available phosphorylation-site prediction programs, including the general tools DISPHOS [136], NetPhos [137] and scan-x [138]. Recently, a new program called MUSite, that incorporates disorder prediction as one of three parameters, was shown to outperform the three previous prediction tools on the basis of sensitivity and specificity [139]. When applied to the TAIR9 Arabidopsis proteome (ftp://ftp.arabidopsis.org/Genes/TAIR9_genome_release/readme_TAIR9.txt), approximately 18000 potential phosphorylation sites were predicted at the 99% confidence interval. More recent algorithms have validated the importance of protein disorder [140]. Using MS-based approaches approximately 2000 phosphorylation events have been identified from Arabidopsis [141,142] and archived in the Plant Protein Phosphorylation Database (P3DB; http://www.p3db.org) [143]. These results suggest that although MS-based approaches are accurate, they are not sensitive [144]. The converse could be appropriately stated for the prediction algorithms. In a similar manner to protein structure determination or subcellular localization predictions, however, as we learn more about the rules for protein phosphorylation prediction the tools will become more accurate, a Yin–Yang relationship. Although those using computational tools to study protein phosphorylation have taken an early lead in the application of bioinformatic methods to analysis of the proteome, we assume that those more interested in or intrigued by other PTMs will not be far behind [145].

### Proteogenomics

The term proteogenomics was coined relatively recently [146,147], and refers to the use of MS-based proteomic analyses to correct/improve annotation of the genome-based proteome [148,149]. As the post-genomic era advances, it has become increasingly clear that the existence of a DNA sequence is insufficient to provide an understanding of complex biological processes. Despite this, the bulk of current validation methods use information derived solely from an annotated genome. The genome sequence does not, however, unambiguously demonstrate that a predicted ORF (open reading frame) is translated into a protein. This is an untenable situation for proteome annotation.

---

**Figure 4** Virtual LC profile of the eight permutations of a peptide with zero, one, two or three potential PTMs

S, serine or phosphoserine; M, methionine or methionine sulfoxide; K, lysine or N-acetylated lysine. The native peptide is a single symmetrical peak, in contrast with the complexity of the virtual profile of a protein with three PTMs.
However, there are now several studies in the literature where predicted genes have been validated at the protein level [150,151].

In addition to the validation of predicted genes and detection of novel genes, results from proteogenomic analyses provide validation of hypothetical ORFs [152], allow accurate determination of protein initiation and termination sites [153], and allow identification of splice variants at the protein level [154].

The need for improved data mining and informatics tools are a key challenge confronting the application of proteogenomic methods to genome annotation. In order for genome sequencing and annotation projects to include high-throughput LC-MS/MS datasets as an essential complement to results from gene-prediction programs, it is critical that new algorithms be developed. These new data mining and informatics tools must be able to seamlessly incorporate the information provided by LC-MS/MS datasets [145,146,148].

Although the principle of searching MS/MS spectra data against six-frame translated genomes to experimentally validate predicted protein-coding genes has been demonstrated in both prokaryotes [155,156] and eukaryotes [152,157], extant techniques are technically limited when contemplating analyses of complex genomes. An obvious disadvantage is the enormous size of a six-frame translated genome, which might be as large as 6 Gb. Simply scaling-up the use of current proteomic search strategies would be impractical, in addition to which the FDR increases in parallel with the database size, decreasing reliability. Improvements in speed and reliability will be necessary in order to develop search routines that are less problematic and more widely applicable. Ideally, any genome-sequencing project should be complemented by nLC-MS/MS-based proteomic profiling [157]. Unfortunately we are not yet at the point where this would be practical.

Biomarkers?

Although much of the interest in MS-based proteomics research has been focused on the intersection of instrumentation and protein (bio)chemistry, there is also an increasing interest in high-throughput biomarker analysis or clinical proteomics as it is sometimes called [158]. For example, in the conference programme for the 59th Annual Meeting of the American Society for Mass Spectrometry (2011) the number of abstracts for presentations on ‘Biomarkers’ was substantially larger than any other category. Unfortunately, for various biological and technical reasons, researchers have not yet been able to translate this interest with effectiveness to the degree of accuracy necessary for prognostic or monitoring assays. Can proteomics-based strategies actually be useful in a clinical context? Large-scale scepticism has now been expressed by both clinicians and regulatory agencies. It is now accepted that individual biomarkers are not likely to exist or to be widely useful, because the complexity of disease can seldom be captured by a single protein. Although it might be considered an extreme example, it was recently reported that researchers identified a set of 81 disease-associated proteins [159]. Thus a strategy more likely to be successful would employ a panel of biomarkers [160,161].

Unfortunately, the problems inherent to validation of a single biomarker increase logarithmically when considering the use of multiple interacting-markers to provide specific valid signatures [161,162]. How does one test the validity? Can a change in occurrence or abundance of a protein be confidently associated with the disease? Or might it only be an artefact resulting from technical variability? Is the source of the technical variability in the isolation/separation/pre-analytical steps, or during the analysis itself [163]? And finally, do we even know enough about biological variability or experimental design to realistically consider the possibility of clinical proteomics?

It has been noted that although many biomarkers are proposed in highly cited studies as determinants of disease risk, prognosis or response to treatment, few ultimately reach clinical practice [164]. Thus far the highly cited biomarker studies often report larger effect estimates for associations than are reported in subsequent meta-analyses evaluating the same associations. How many samples need to be analysed in order to define a valid biomarker? How do we design a multi-biomarker clinical platform, and, having designed one, how should it be validated [165]? Are the statistical criteria accepted and acceptable for high-throughput clinical chemistry adequate or adaptable to clinical proteomics? Early results suggest that the increasingly robust statistical methods being developed for and used in bioinformatic analyses will be more appropriate. Results described in recent reports suggest that valid proteomic biomarkers for diagnosis and prognosis can only be developed by applying statistical data mining procedures [166,167]. The authors suggest that multiple testing is necessary, but that sample-size estimation can be performed on the basis of a smaller number of observations via re-sampling from pilot data, and conclude that the sorts of machine-learning algorithms used to predict protein PTMs [168] appear ideally suited to generate the desired signatures [169].

CONCLUDING REMARKS

In the present paper, on the basis of our training, background and experience, we provide a brief review of the state of proteomics research. This is set in the context of an admittedly biased, but not unreasonable, set of suggestions for the short-term future of the field. We touch upon the need for improved tools, strategies and targets. We champion a shift away from qualitative, incomplete and descriptive studies. If a question can be answered effectively by transcript profiling, then why apply a MS-based proteomics strategy? It is the questions that cannot be addressed by using the other omics-based tools, such as protein interactions and the structures of multi-protein complexes, that deserve more of our attention.

ACKNOWLEDGEMENTS

We thank M.L. Johnston who prepared the Figures and L. Meyer who prepared Table 1.

FUNDING

Work in the laboratory of J.J.T. is supported by the NSF and ILSI-Health Environmental Science Institute. Work in the J.A.M. laboratory is supported by the USDA, Agricultural Research Service, NSF, and the Nichols Foundation.

REFERENCES


