1

ESSENTIAL ELEMENTS OF A PROTEOMICS LABORATORY

Daniel C. Liebler*

1.1 INTRODUCTION

The genomic revolution and the emergence of proteomics are transforming cancer research. Cancer has long been understood as a disease of genomic instability and growth dysregulation [Loeb, 2001; Sieber et al., 2003]. Altered gene forms and aberrant gene expression patterns give rise to altered biochemistry, cell biology, and tissue function that characterizes neoplastic disease. Just as genes translate ultimately to function, genetic instability translates to dysfunction, which is manifested at the level of proteins and proteomes. In cancer, these protein changes may provide tumor-specific targets for therapeutic intervention and highly informative biomarkers for disease [Petricoin et al., 2002b; Wulfkuhle et al., 2003].

The ability of new technologies to detect and analyze these changes is rapidly attracting researchers to use proteomics approaches in cancer research. A dedicated proteomics laboratory is the most logical means of establishing essential instrumentation and expertise in proteomics. From a purely practical standpoint, this makes sense in several ways. Mass spectrometry (MS) instrumentation is costly and requires skilled operators. Necessary expertise in protein and peptide separations, MS data analysis, and bioinformatics often requires individuals with complementary, overlapping skills. Finally, by establishing a relatively small portfolio of experimental capabilities and implementing robotics and automation, a proteomics laboratory can serve a large and diverse group of investigators. In short, a proteomics laboratory offers economy of scale, centralized expertise, and a means to foster interdisciplinary research.

*Proteomics Laboratory, Mass Spectrometry Research Center and Departments of Biochemistry and Pharmacology, Vanderbilt University School of Medicine, Nashville, TN 37232 (E-mail: daniel.liebler@vanderbilt.edu).

At the very least, a proteomics laboratory is yet another core facility, analogous to protein chemistry and protein sequencing cores that have been around for years. However, proteomics laboratories will ultimately have a much greater impact. Proteomics laboratories serve a much broader range of researchers, ranging from biochemists to cell biologists and clinical scientists. Emerging applications of proteomics have already begun to revolutionize the detection and diagnosis of cancer and offer the potential to individualize cancer prevention and therapeutics [Liotta et al., 2001; Petricoin et al., 2002b]. With this realization comes tremendous expectations. Indeed, it is not unreasonable to argue that the success of the cancer research enterprise in an institution will depend on the quality of its proteomics laboratory.

Proteomics laboratories thus serve a role that extends beyond that of traditional service cores. The tools and applications of proteomics are new and largely unfamiliar to many investigators. An essential role of a proteomics laboratory is thus to teach investigators how these new tools work and how they can be used most effectively in basic and clinical research. As increasing numbers of investigators use proteomics tools and approaches, powerful experimental approaches are emerging. This represents the synergy between proteomics “specialists” and investigator-users and casts the proteomics laboratory in the role of incubator for innovative science.

Chapter 1 provides perspectives on the establishment and operation of a proteomics laboratory and on the translation of proteomics capabilities into the research enterprise. It is essential to emphasize that there is no single “right way” to organize a proteomics laboratory. The ideas presented in this chapter reflect my experience and my observations of successful proteomics laboratories and collaborations. Just as different instruments and approaches can address the same problem, so can different designs and operational models establish an effective proteomics laboratory. Nevertheless, any proteomics laboratory must provide several essential capabilities, as described below. The ability to provide these capabilities depends on MS instrumentation and on computing and bioinformatics support. Finally, expertise established within the proteomics laboratory ultimately must be translated into the community of users.

1.2 ESSENTIAL CAPABILITIES OF A PROTEOMICS LABORATORY

Although a proteomics laboratory can be used for diverse studies, essentially all of these experiments translate into a small group of analytical proteomics tasks. The essential capabilities for a proteomics laboratory are 1) protein and peptide separations, 2) protein digestion, 3) simple protein identification, 4) identification of protein components of complex mixtures and multiprotein complexes, 5) mapping protein modifications, 6) quantitative analysis, and 7) proteome profiling. These seven essential capabilities constitute the basic repertoire of a proteomics laboratory. Some experimental designs that use proteomics involve one of these tasks, whereas more elaborate studies involve integration of two or more. What follows is a brief description of these essential capabilities of a proteomics laboratory and of possible approaches to achieve each. All of these problems and analytical approaches are discussed at greater length in multiple contexts in other chapters in this book.

1.2.1 Protein and Peptide Separations

Most proteome analyses begin with mixtures of proteins. Identification of proteins in complex mixtures requires separation step(s) somewhere in the analytical process. This
can involve either separation of intact proteins or separation of peptides generated by digestion of protein mixtures (Fig. 1.1). Perhaps the archetypal proteome analysis begins with a two-dimensional sodium dodecyl sulfate-polyacrylamide gel electrophoresis (2D-SDS-PAGE) separation of proteins. This is the most powerful tool available for resolution of complex protein mixtures. Staining of gel-separated proteins also provides images, enabling selection of specific features for identification and comparison of proteome samples. 2D-SDS-PAGE separations also resolve posttranslationally or otherwise modified proteins in many cases. The use of dedicated 2D-SDS-PAGE systems and immobilized pH gradient (IPG) isoelectric focusing strips has made it possible for many laboratories to perform highly reproducible 2D gel-based proteome analyses [Anderson and Anderson, 1996]. 2D-SDS-PAGE can be done with elaborate, automated systems that incorporate robotic imaging, sample excision, digestion, and preparation for MS. Of course, these systems are very expensive but offer high throughput and reproducibility. On the other hand, a small user base can easily be served by 2D gel-based separations with imaging, spot selection, digestion, and MS analysis done manually. 2D gel-based approaches have well-known limitations, including poor solubility of many proteins during isoelectric focusing and limited dynamic range for protein staining, which favors detection of relatively abundant proteins [Gygi et al., 2000]. Nevertheless, 2D-SDS-PAGE is a widely used benchmark method in proteomics that has widespread utility, despite its limitations.

Other useful separation methods for intact proteins include one-dimensional (1D)-SDS-PAGE, solution phase isoelectric focusing, and protein chromatography. 1D-SDS-PAGE is used routinely in most biochemistry and cell biology laboratories, yet it is a useful tool for fractionating modestly complex protein mixtures (such as multiprotein complexes) [Gavin et al., 2002]. Solution phase isoelectric focusing has become more useful through the availability of simple kit-based apparatuses. Preparative chromatography (e.g., gel filtration, ion exchange) also provides complementary separation tools for intact proteins.

![Image](Leibler_ch01.jpg)

**Figure 1.1** Analytical proteomics platforms. Top: 2D gel-based proteome analysis. Bottom: "shotgun" proteome analysis.
Affinity chromatography (e.g., with antibody columns) provides selective retention of proteins with specific chemical or antigenic characteristics. In proteome analyses, these separation methods are not directed at purifying specific proteins but rather at simplifying complex mixtures before additional analyses.

Peptide separations have become widely used as a component of liquid chromatography-tandem mass spectrometry (LC-MS-MS) analyses, in which peptides resolved by reverse-phase liquid chromatography are transferred directly into the MS instrument to obtain MS-MS data. Highly complex peptide mixtures (e.g., from digestion of more than 10 proteins) often contain too many peptides to be detected in MS-MS analyses in a single reverse-phase gradient analysis. Thus, complex peptide mixtures can be fractionated by strong cation-exchange chromatography. In-line ion exchange/reverse-phase analysis schemes (referred to previously as direct analysis of large protein complexes (DALPC) [Link et al., 1999] or multidimensional protein identification technique (MudPIT)) [Washburn et al., 2001] have been applied to tryptic digests of complex protein samples and provide MS-MS spectra of many more peptides than would be obtained with single reverse-phase analyses alone. Increased fractionation increases the numbers of proteins and peptides identified. Off-line ion exchange chromatography followed by reverse-phase LC-MS-MS of the fractions provides a more robust approach [Peng et al., 2003] in most laboratories and offers the advantage of being able to query individual ion-exchange fractions of interest without having to repeat the entire fractionation scheme.

Certainly, a proteomics laboratory could vary the techniques and approaches described above to provide flexible approaches to protein and peptide separations. Of the tools described above, 2D-SDS-PAGE and combined ion-exchange/reverse-phase peptide separations are essential to a proteomics laboratory.

1.2.2 Protein Digestion

Digestion is a necessary prelude to protein identification. MS instrument sensitivity toward intact proteins is less than that for peptides. In addition, MS-MS fragmentation, which enables identification of sequence and modifications, works best for peptides of ~5–30 amino acids in length. Although both enzymatic and nonenzymatic digestion strategies are possible, enzymatic digestions offer a variety of cleavage specificities under mild conditions. Trypsin is by far the most widely used protease in proteomics. It is available in high purity at reasonable cost, and it is robust and displays excellent cleavage specificity for Lys and Arg residues. The distribution of Lys and Arg residues in proteins generally yields an abundant selection of peptides in the 5- to 30-amino acid range. Of course, basic regions of proteins can yield only small peptides, and regions devoid of Arg or Lys residues will yield peptides that are too large for facile MS-MS fragmentation. Nevertheless, trypsin is the most useful enzyme in proteomics, and every proteomics laboratory should have established protocols for in-solution and in-gel digestions. Although digestions can be done manually, digestion robots are becoming widely used, particularly in processing of spots from 2D gels. In addition to increased throughput, automated digestions are generally more reproducible and less subject to contamination.

For protein identification work in simple and complex samples, trypsin is just about the only enzyme ever needed. However, alternate cleavages are often needed for mapping of modifications (see below). Chymotrypsin is active under similar conditions to trypsin and
cleaves at Phe, Tyr, Trp, and Leu residues, albeit with somewhat less specificity. *Staphylococcus aureus* V8 protease (Glu-C) cleaves at Glu and, under certain buffer conditions, Asp. Brief digests with nonspecific proteases, such as pepsin and proteinase K, can also generate peptides suitable for MS-MS [Codreanu et al., 2002].

### 1.2.3 Simple Protein Identification

One of the most common tasks in a proteomics laboratory is to identify a band on a gel. This is typically done by in-gel tryptic digestion and MS analysis of the extracted peptides. A similar problem less frequently encountered is the identification of a pure protein in solution, which is handled by in-solution digestion and MS analysis. With the evolution of MS instrumentation and protein identification software, such simple protein identification has become an almost trivial task. For relatively pure protein samples, such as those taken from 2D gels, peptide mass fingerprinting by matrix-assisted laser desorption ionization (MALDI)-time of flight (TOF) provides identification in many cases. However, more unambiguous identifications can be made from MS-MS spectra, and the widespread availability of LC-MS-MS and MALDI-MS-MS instruments makes this routinely feasible. Database search software identifies protein sequences corresponding to MS and MS-MS spectra and frequently provides probability-based assessment of the identifications. In high-demand laboratories, groups of candidate samples can be batched for regularly scheduled analysis and protein identifications thus can be provided to users on a regular schedule.

### 1.2.4 Identification of Protein Components of Complex Mixtures and Multiprotein Complexes

The most-requested analysis in proteomics laboratories is the identification of multiple components in protein mixtures. The prototypical problem is identification of the components of a multiprotein complex, which is often done by immunoprecipitation or epitope tag affinity capture of a target protein and its partners. This may yield a sample containing as few as 2 proteins or more than 50, depending on the complex under study. These samples also inevitably contain nonspecifically bound proteins in varying amounts. Indeed, it is essential that controls be designed and analyzed to help distinguish nonspecifically bound from specific components of the complex under study. Other complex mixtures may consist of proteins in specific organelles, proteins that share some characteristic (e.g., phospho- or glycoproteins), or proteins in biofluids (e.g., serum).

2D-SDS-PAGE has been used extensively to analyze complex samples. Resolution of proteins on the gel, followed by staining, imaging, and spot selection, yields a collection of samples that can be treated as simple protein identifications (see above). The advantage of this approach is that it reduces a complex protein mixture to a series of simple protein identifications. The limitations of 2D-SDS-PAGE were noted above. Particularly relevant in this context is the problem of dynamic range for protein staining and detection. This biases the analyses of complex proteome samples toward detection of the more abundant components. However, for multiprotein complexes with near-stoichiometric relationships between partners, this is less of a problem.

An alternative approach to complex proteome analysis employs a “shotgun” analytical strategy, in which complex protein mixtures are digested directly to complex peptide mixtures and then analyzed directly by LC-MS-MS. To deal with highly complex peptide
mixtures, this approach relies on multidimensional LC separations, such as the combined ion-exchange/reverse-phase approach described above. This approach has advantages over a 2D gel-based approach, particularly for the analysis of complex proteome samples where the abundances of proteins vary over several orders of magnitude [Lin et al., 2003]. A multidimensional LC-MS-MS strategy typically identifies more low-abundance proteins and provides greater sequence coverage (i.e., the number of peptides in a protein for which spectra are obtained) for the identified proteins. As noted above, multidimensional peptide separations can be done inline with ion-exchange/reverse-phase LC or by collection of ion-exchange fractions offline and then analyzing fractions individually. The former may offer somewhat more sensitivity and avoids mass transfer steps, whereas the latter is more practically robust, offers more flexibility in ion-exchange fractionation, and allows repeat reverse-phase LC-MS-MS analyses of fractions of interest.

An alternative to multidimensional peptide separation is gas-phase fractionation, in which repeated reverse-phase LC-MS-MS analyses are done on the same sample [Spahr et al., 2001]. In each analysis, a limited \( m/z \) range is used for selection of precursor ions for MS-MS fragmentation. In data-dependent acquisitions commonly used for automated MS-MS analyses, the instrument first does a full scan and then selects the most abundant ion in the entire instrument mass range for MS-MS. Then the second most abundant precursor is selected for MS-MS and so on. The problem with this is that selection across the entire mass range of the instrument usually limits precursor selection to the most abundant species and less abundant precursor ions are missed. Repeated analysis in which precursor selection is confined to a more narrow \( m/z \) range allows more precursors in the limited range to be selected in a given time interval. Thus, less abundant species are more easily detected. A typical gas-phase fractionation analysis on an LC-MS-MS instrument might start with a run in which precursors are selected between \( m/z \) 300 and 700, followed by a second run in which precursors are selected between \( m/z \) 700 and 1,100, followed by a third run in which precursors are selected between \( m/z \) 1,100 and 2,000. Depending on the available amount of sample and instrument time, one could choose any number of analyses. The combined data sets generally represent greater numbers of protein identifications and greater sequence coverage than for a single analysis with acquisition of precursors over the entire instrument mass range. It is worth noting that, with very complex peptide mixtures analyzed in a single reverse-phase separation, abundant peptides may suppress the ionization of less abundant species. Thus, the chromatographic resolving power of multidimensional LC-MS-MS may offer greater advantage in identification of low-abundance proteins in complex samples.

### 1.2.5 Mapping Protein Modifications

Posttranslational modifications dictate many critical functions of proteins. In addition, chemical adducts formed by reactive intermediates are hypothesized to modulate abnormal protein functions. Because modifications affect protein-protein interactions, analysis of modifications can be a critical element in the analysis of changes in multiprotein complexes under different conditions. The task of mapping protein modifications to specific sequence locations is a more difficult problem than protein identification in either simple or complex systems. Whereas identification can be achieved with data on as few as one or two peptides, assignment of modifications requires that spectra of the modified peptides be obtained. Of course, most of the peptides in a digest will be unmodified and even the sequences that bear modifications can also be present in unmodified forms. Thus, the mapping
of modifications becomes a problem of selective detection of low-abundance components of complex peptide mixtures. (Analysis of protein modifications is discussed in detail in Chapter 5.)

To successfully map protein modifications, a proteomics laboratory must integrate four tools: 1) affinity enrichment, 2) complementary proteolytic digestions, 3) MS-MS analysis, and 4) data mining software [Mann and Jensen, 2003]. Affinity capture of proteins containing the modification of interest can enrich the sample for modified proteins. After digestion, a second affinity enrichment step can further enrich the mixture in peptides bearing the modification. Complementary enzymatic digestions (i.e., digestion with two or more proteases with different cleavage specificities) are sometimes needed to generate peptides that bear the modifications and are of suitable length for MS-MS analyses (e.g., ~5–25 amino acids in length). MS-MS spectra provide information on the peptide sequence (e.g., b- and y-ion series) as well as on the mass and sequence position of the modifying group. Although MS measurements of intact peptide masses can provide evidence of modification, only MS-MS can confirm sequence position. Once spectra of modified peptides are obtained, the task of identifying which spectra correspond to modified sequences is a significant challenge. Software tools used for protein identification, such as Sequest and Mascot, allow the user to specify fixed and variable modifications for database searches, but these require the user to know the mass and sequence specificity of the modification [MacCoss et al., 2002]. This is not always known. Unanticipated modifications may be detected by searching MS-MS datafiles with protein sequences using the recently-described P-Mod program [Hansen et al., 2005]. In addition, many modifications affect the MS-MS fragmentation of peptides or contribute unique modification-specific fragmentations of their own. Specialized software tools, such as SALSA, enable searches for spectra displaying characteristics of modification-derived fragmentations or of fragment ion series displayed by both unmodified peptides and their modified forms [Liebler et al., 2002].

Analyses of protein modifications are perhaps the most time-consuming problems for proteomics laboratories. Specialized sample preparation, enzymatic digestions and multiple data analysis steps all go well beyond what is typically needed for protein identification. The latter task is most involved and this can become a significant time sink for proteomics laboratory personnel. Research groups that use proteomics laboratories frequently can develop expertise with data analysis software and this can help speed the completion of such projects. The problem of transferring expertise from the proteomics laboratory to users is discussed in greater detail below.

1.2.6 Quantitative Analysis

Much proteomics work, such as protein identification and modification mapping, consist of qualitative analyses. However, comparative analyses of proteomes are essential to understanding proteome dynamics in biology, and these comparisons typically involve quantitation. Relative quantitation compares the amounts of proteins or their modified forms between two or more states of a system. Absolute quantitation provides measurement of the amount of specific proteins or protein forms in a particular sample. Of these, relative quantitation is the most widely used approach. In many experimental designs, the objective is to measure relative changes. Because absolute quantitation often requires greater effort than relative quantitation, it is important to ascertain the goals of the quantitative analyses proposed.
A common objective in proteome analyses is to quantify multiple simultaneous changes in proteins between paired samples. This can be done by 2D gel electrophoresis with comparisons of images between gels. A more accurate and sophisticated version of this analysis is to employ 2D differential gel electrophoresis (2D-DIGE), in which proteins in the samples to be compared are labeled with different fluorescent dyes and then combined and analyzed [Unlu et al., 1997]. Differential fluorescence emission scanning then indicates proteins that are differentially present in the two samples. The changes could result from expression differences, modification differences, or both. The alternative to a 2D gel-based comparison is LC-MS-MS analysis of stable isotope-tagged peptides from digests of paired samples. The best known version of this analysis employs the isotope coded affinity tag (ICAT) reagents [Gygi et al., 1999], although multiple isotope tagging strategies have been reported [Regnier et al., 2002; Mason and Liebler, 2003].

Applying either of these strategies in a proteomics laboratory requires careful implementation of procedures and expertise that go beyond what is needed for protein identification work. All quantitative analysis requires attention to detail, particularly for reproducible protein and peptide labeling chemistry, sample cleanup, and instrument analysis. Regular use of quality control standards is absolutely necessary to ensure the validity of results. In addition, skill in using specialized software for data analysis, such as differential fluorescence of dye-labeled proteins or ratios of isotopically labeled peptides, is critical for reliable analyses.

Absolute quantitation by MS is best done with stable isotope labeled internal standards to correct for errors or sample losses during processing [Gerber et al., 2003]. Synthetic 2H-, 13C-, and/or 15N-labeled peptides can be obtained at fairly reasonable cost for use as internal standards. Modified versions of a target peptide can be prepared if the objective is to measure a modified protein form. However, the use of peptides as internal standards only corrects for errors inherent in processing after protein digestion. Thus, additional controls for protein capture and digestion efficiencies are needed.

1.2.7 Proteome Profiling

A rapidly emerging technique, termed proteome profiling, is the acquisition of mass spectra representing intact proteins present in a complex sample. This approach is most commonly applied to serum or tissue samples, but the technique is applicable to any complex protein sample. Although proteome profiling typically involves MALDI-TOF MS instruments, the approach could be extended to other instrument platforms, such as LC-MS. Intense interest in proteome profiling of serum and tissue samples comes from recent reports demonstrating the utility of the approach for detecting cancers and for discriminating between histologically similar tumors that nevertheless display differences in metastatic potential and therapeutical responsiveness [Petricoin et al., 2002a; Petricoin III et al., 2002; Yanagisawa et al., 2003]. (Proteome profiling is discussed in detail in Chapters 7 and 8.)

Although the MALDI-TOF instruments used for profiling are widely available and already present in many proteomics laboratories, other accessory capabilities are more unique. Proteome profiling requires specific expertise in preparing serum and tissue samples for analysis. For serum profiling, direct MALDI-TOF analysis of diluted serum or serum protein fractions is possible. Nevertheless, procedures for isolating different serum protein fractions are necessary for more comprehensive analyses. Tissue analysis requires localized application of MALDI matrix (e.g., sinapinic acid) for best results [Chaurand
and Caprioli, 2002]. Moreover, it is important to use histological characteristics of a tissue sample to guide analyses to features of specific interest. This requires the ability to precision spot the matrix to areas of a few hundred micrometers. Commercial instruments to do this are in development.

Proteome profiling is generally used as a classification technique, in which profiles of samples with known characteristics (e.g., tumor vs. normal tissue) are used to train analysis algorithms, which then are used to classify unknown samples based on the occurrence of multiple pattern features in their profile spectra [Petricoin et al., 2002b; Wulfkuhle et al., 2003]. This requires specialized data analysis software and computing resources to convert spectral patterns to analyzable data sets and biostatistics support for the identification of specific marker signals in spectra. A variety of different approaches are presently being explored for the extraction and analysis of profiling data. Identification of the protein species that account for diagnostic pattern signals generally cannot be done directly from the spectra. This requires additional analytical work using the approaches described above.

In contrast to the six other essential capabilities described above, proteome profiling is somewhat unique because the technique is still at an early stage of application and because some of the necessary tools are not yet widely available. Nevertheless, intense interest in this application and remarkable successes are creating great demand for these capabilities in proteomics laboratories. New tools for proteome profiling will inevitably follow.

1.3 MS INSTRUMENTATION FOR PROTEOMICS LABORATORIES

MS instruments constitute the core analytical instrumentation in a proteomics laboratory. Commercially available MS instruments combine different types of sources and mass analyzers and provide overlapping analytical capabilities. In equipping a proteomics laboratory, selection of MS instruments is a critical consideration. The instruments selected should collectively enable the laboratory to provide the essential capabilities described above. In addition, instrument selection should balance performance specifications with cost and robustness. The latter is particularly important, as instrument “downtime” can produce productivity bottlenecks. MS instruments for proteomics range in cost from relatively inexpensive ion traps costing about $150,000 to high-end, high-performance Fourier transform ion cyclotron resonance (FTICR) instruments costing over $1 million. As with virtually all instrumentation, there is generally a trade-off between performance and robustness. Some high-end instruments may provide fabulous data but may suffer excessive downtime, whereas modestly priced instruments may lack sensitivity, resolution, or other key performance features. High-end instruments also require more attention from skilled operators for consistent productivity. The following sections summarize strengths and weaknesses of the MS instruments generally used in proteomics laboratories. (A more detailed discussion of MS instrumentation is presented in Chapter 2.)

1.3.1 Ion Trap LC-MS Instruments

An ion trap LC-MS instrument is the single most useful and versatile instrument a proteomics laboratory can have. These instruments are used almost exclusively for electrospray ionization LC-MS, although MALDI sources for ion traps have been described [Qin and Chait, 1997]. Ion trap instruments are used primarily for automated, “data-dependent”
acquisition of peptide MS-MS spectra [Jonscher and Yates, 1997]. Ion traps operated in this mode have generated the vast majority of protein identifications done by MS-MS over the past several years. This approach enables not only protein identification but also mapping of protein modifications from MS-MS spectra and relative quantitation with stable isotope tags (e.g., ICAT). Ion traps are very sensitive for MS-MS; typically, good spectra can be obtained from low (fmol) quantities of peptides analyzed with microelectrospray sources. Ion traps are also quite robust and durable, require a relatively modest degree of maintenance, and are modestly priced. Depending on features, a complete ion trap LC-MS system costs between $175,000 and $300,000.

Three-dimensional ion traps can acquire about one MS-MS spectrum per second, whereas the new linear ion trap instruments can acquire up to five MS-MS spectra per second. When operated under conditions typically used for proteomics work, these instruments provide approximately unit resolution. Thus, error in measurement of precursor peptide ions or peptide fragment ions is ±0.5 unit. Although other instruments provide more resolution and mass accuracy, this level of performance is generally adequate for most protein identifications with database search software (e.g., Sequest). Ion traps can also perform higher-order tandem MS experiments (e.g., MS-MS-MS or MS$^3$), which are occasionally quite useful.

### 1.3.2 MALDI-TOF MS Instruments

MALDI-TOF instruments combine the MALDI source, which ionizes peptide and protein samples crystallized, with an ultraviolet (UV)-absorbing matrix and a time of flight (TOF) mass analyzer. TOF analyzers display broad mass range (up to $m/z$ 250,000) in linear mode with a gradual decrease in sensitivity with increasing $m/z$. Use of either reflectron or delayed extraction technology enables much higher resolution and mass accuracy for analyses of peptides in a narrower $m/z$ range [Roepstorff, 2000]. MALDI-TOF instruments generally are sensitive, user-friendly, and robust. Depending on features, MALDI-TOF instruments range in cost from $200,000 to over $500,000.

MALDI-TOF instruments were once the most widely used instruments for protein identification, via peptide mass fingerprinting, which maps measured masses of intact peptides against database values. This approach has been largely supplanted by MS-MS. MALDI-TOF instruments are not capable of true MS-MS experiments, although the post-source decay technique provides a limited approximation of MS-MS. However, MALDI has been combined with tandem mass analyzers, as described below.

A resurgence of interest in MALDI-TOF MS in proteomics comes from its application in proteome profiling in tissue or serum samples. The acquisition of proteomic patterns provides molecular signatures that reflect biological characteristics, including disease states and other phenotypes. A MALDI-TOF MS instrument is essential to establishing a proteome profiling capability in a proteomics laboratory. A related system, termed SELDI-TOF (surface-enhanced laser desorption ionization-time of flight) is a proprietary variant of MALDI-TOF in which chip surfaces displaying selected chemical characteristics are used to capture a subset of proteins in a sample and then laser-induced ionization from the chip surface generates peptide ions [Weinberger et al., 2002]. SELDI-TOF MS has been most commonly done with a small linear TOF mass analyzer that displays high sensitivity but limited resolution and mass accuracy and is operated in the linear mode. However, coupling a SELDI source to a higher performance TOF of hybrid mass analyzer can provide superior resolution and more spectral information.
1.3.3 Quadrupole-Time of Flight (QqTOF) Instruments

These instruments combine a first-stage quadrupole (Q) and a radiofrequency-only quadrupole (q) with a reflectron-equipped TOF mass analyzer to provide a tandem mass analyzer that provides high sensitivity, resolution, and mass accuracy for precursor and product ions in MS-MS analyses. Peptide and fragment ion $m/z$ values that differ by as little as hundredths of an $m/z$ unit can be resolved. This enables accurate determination of charge states for precursor peptide ions and peptide fragment ions. When appropriately calibrated, the high mass accuracy and resolution provided by QqTOF instruments enable accurate discrimination between peptides and fragments of similar nominal mass, which increases the accuracy of protein identifications (when used with software that can utilize this information) [Morris et al., 1997]. These features are also highly useful in the analysis of modified peptides by MS-MS, as greater mass accuracy can help distinguish between modifications of similar nominal mass. QqTOF mass analyzers are coupled either to electrospray or MALDI sources and thus offer considerable versatility. Indeed, MALDI-QqTOF and QqTOF instruments equipped with SELDI sources have recently been used for protein profiling applications.

Although QqTOF instruments offer higher resolution and mass accuracy than ion traps, they are not as robust. As noted above, instrument down-time can be a serious bottleneck in a proteomics laboratory. QqTOF instruments range in cost from $300K to $600K, depending on specifications.

1.3.4 Triple Quadrupole Instruments

Triple quads were the first MS instruments to be widely used for MS-MS of peptides. They were gradually supplanted by ion traps because the traps offered more efficient peptide fragmentation and greater sensitivity for full-scan MS-MS analyses. However, triple quads remain useful for quantitative analyses based on MS-MS fragmentation (e.g., selected reaction monitoring). New triple quadrupoles with improved analyzer designs rival QqTOF instruments for resolution and mass accuracy. Triple quads range in price from about $300,000 to $500,000, depending on features, and are generally more robust than QqTOF instruments. The principal application of these instruments in proteomics laboratories would appear to be for quantitative proteomics based on LC-MS-MS of peptides with isotope tags [Gerber et al., 2003].

1.3.5 MALDI-TOF-TOF Instruments

This recently introduced instrument design combines two reflectron-equipped TOF mass analyzers. The first TOF analyzer serves for selection of precursors, and the second serves for analysis of product ions. In contrast to ion trap and QqTOF instruments, TOF-TOF mass analyzers are capable of high-energy collision-induced dissociation of peptides [Medzihradsky et al., 2000]. This enables Leu and Ile residues to be distinguished, other nominally isobaric species to be distinguished (e.g., Lys vs. Gln residues), and the provision of additional peptide bond cleavages in higher yields than low-energy collision-induced dissociation. Resolution and mass accuracy are comparable to QqTOF instruments. Commercially available MALDI-TOF-TOF instruments display impressive automation capabilities that enable rapid, high-throughput identification of proteins by peptide mass fingerprinting and MS-MS. These instruments are relatively expensive.
($500,000 and up) and appear to be more robust that QqTOF instruments, although there is less extensive experience to date in the user community with TOF-TOF instruments.

### 1.3.6 Fourier Transform Ion Cyclotron Resonance (FTICR) Instruments

FTICR instruments employ ion trap technology with the trap housed in a high (7–12 tesla) magnetic field generated by a superconducting magnet. This mass analyzer offers the highest sensitivity, resolution, and mass accuracy of any mass analyzer in use today [Marshall and Hendrickson, 2002]. FTICR mass analyzers are coupled to either MALDI or electrospray sources. The ability of the FTICR to resolve multicharge states of intact proteins allows “top-down” analysis of proteins, in which whole proteins are excited and fragmented in the FTICR, ultimately to generate MS-MS spectra of constituent peptides [Ge et al., 2002]. This is in contrast to the more widely used “bottom-up” approach in which proteins are first digested to peptides before MS analysis. FTICR offers a powerful means of characterizing complex peptide mixtures. In some cases, reliable identifications can be made from accurate mass measurements of peptide precursor ions. FTICR instruments are the most costly of MS instruments used for proteomics and range in price from about $600,000 to well over $1 million, depending on magnet strength and other specifications. FTICR MS instruments also require a great deal of expert maintenance and skilled operators for consistent productivity. Because of their expense and lack of practical robustness, FTICR instruments are not widely used in proteomics. However, they can make an excellent addition to a well-supported proteomics laboratory that can meet these operational requirements.

### 1.3.7 Other New Instruments

Innovative new mass analyzer designs continue to provide improved MS instruments for proteomics. Two recently introduced hybrids are the quadrupole-ion trap MS (Qq-trap) and the linear ion trap-Fourier transform ion cyclotron resonance (trap-FT). Both capitalize on new developments in linear ion trap technology, offering more rapid scanning and higher sensitivity that three-dimensional ion traps (see above). The Qq-trap combines a quadrupole mass analyzer, low-energy collision cell, and a linear ion trap. This instrument exceeds the performance of even the newest triple quadrupoles for quantitation by selected reaction monitoring [Hager and Le Blanc, 2003]. The Qq-trap offers the opportunity to detect certain posttranslational modifications (e.g., phosphorylation) by alternate switching between positive and negative ion modes and using precursor scanning to detect phosphopeptide precursor ions. The Qq-trap probably will find greatest application in quantitative analysis of peptides.

The trap-FT combines a linear ion trap with an FTICR mass analyzer to enable the highest sensitivity, resolution, and mass accuracy for precursor and product ion analysis in MS-MS [Aebersold and Mann, 2003]. This instrument may prove to be the most powerful tool yet for the MS-MS analysis of complex peptide mixtures. Both instruments are relatively expensive. The Qq-trap costs approximately $500,000 and the trap-FT costs approximately $700,000. Because these instruments are very new, their practical robustness remains to be determined.

### 1.4 BIOINFORMATICS AND COMPUTING RESOURCES

The tremendous analytical capabilities afforded by new MS instrumentation generates tremendous volumes of data. A proteomics laboratory equipped with two ion trap LC-MS
systems and a MALDI-TOF MS instrument and running at full capacity can easily generate 1-2 terrabytes (TB) of raw data and secondary data every six months. A proteomics laboratory must be equipped to analyze data, track samples and associated datafiles, and store data. In addition, analytical software for end users must be made available, preferably through Web-based applications. Although MS instruments come with computer workstations and data analysis software packages, the data analysis and storage requirements of a busy proteomics laboratory can quickly overwhelm these resources.

Thus, a proteomics laboratory needs a computer infrastructure that includes one or more file servers to handle sample tracking, file storage, applications, and archival of data. Depending on the volume of work, some of these capabilities can reside on dedicated servers. For example, a laboratory information management system (LIMS) may use a separate server to maintain a database containing sample information, analytical procedures done on samples and fractions derived from samples, and links to associated data files.

Data analysis applications can utilize multiprocessor computer systems to optimize speed of analysis. When run on single processor workstations, programs such as Sequest and Mascot may take up to 10 times as long to search MS-MS spectral data sets against databases than it took to acquire the raw data. Given the ability of new linear ion trap LC-MS-MS and MALDI-TOF-TOF instruments to generate a million MS-MS spectra per week in automated operation, multiprocessor support for Sequest, Mascot, and similar programs is rapidly becoming a necessity. Specialized software, such as that used to map post-translational modifications or to filter spectral data sets, also will rely increasingly on enhanced computing power.

Programs such as Sequest and Mascot provide sequence matches to MS-MS spectra but not all of the assignments are of high quality. Thus, filtering or quality control evaluation of the outputs of these programs is essential. This requires additional software and database utilities to perform. Moreover, data sets that correspond to specific samples (e.g., treated vs. control) often must be compared or subtracted to extract biological meaning from proteomics analyses. Annotation of data sets need not involve manipulation of the primary data but rather analysis of the outputs of Sequest, Mascot, or other first-tier applications. These operations are perhaps best handled by Web-based applications accessible to users of the proteomics laboratory. Moreover, these applications put the data in the hands of the users, who are best able to assess the significance of the protein identifications made. (Bioinformatics applications for proteomics are described in Chapter 6.)

1.5 TRANSLATION OF PROTEOMICS EXPERTISE INTO THE USER COMMUNITY

The demand for proteomics capabilities in basic and clinical research continues to grow. At the same time, the concepts, tools, and applications of proteomics are still new to most researchers. Proteomics laboratories must not only implement and use new technology but also must help investigators bring the technology into their research programs. In this context, the proteomics laboratory assumes an educational mission that is every bit as important as its analytical mission.

Most proteomics laboratory users will perform experiments in their own laboratories and then bring protein mixtures to proteomics laboratories for analysis. The analyses will generate raw data (e.g., MS-MS spectra) and derived data (e.g., lists of peptide and protein identifications corresponding to the spectra), which then require further analysis to reconstruct
and interpret in the context of the experiment. To make most effective use of proteomics laboratories, users must have at least a general appreciation of the underlying principles and rationale behind the analytical methods employed. More specifically, they should be advised on how to design and execute experiments such that the resulting samples can be successfully analyzed. Finally, users need to learn how to analyze the data from proteomics analyses.

Investigators new to proteomics can benefit from several new introductory texts on proteomics [Liebler, 2001; Simpson, 2002; Westermeier and Naven, 2002]. These are being utilized in an increasing number of regular university course offerings, workshops, and short courses. Proteomics laboratories can help users master the practical details of experimental design and sample preparation through one-on-one discussions and through periodic workshops. The workshop format is most effective when users can take away handouts or can access downloadable files of standard protocols from a proteomics laboratory Website. Indeed, a well-designed Website can be a most effective means of disseminating useful information on techniques and protocols.

One of the most challenging problems for proteomics laboratory staff is the analysis of data. Consider the fairly typical situation in which identification of components of a protein mixture is done for a novice user. A careful review of the data set and preparation of a report by proteomics laboratory staff can take days. If several similar projects are pending, the available time of proteomics laboratory staff could be entirely consumed by data analysis. On the other hand, dumping the raw data or database search outputs into the laps of novice users will simply frustrate and alienate them.

The best long-term solution is to train the users to work with proteomics data. This can be done with new software and database tools, user training, and patience. The problem of evaluating long output lists from Sequest and similar database search programs has emerged as high-throughput LC-MS-MS analyses have generated ever-increasing flows of data. Not all assignments made by these programs are reliable, and distinguishing correct from incorrect assignments is not a trivial problem. New software and database tools ultimately will offer the end user Web-based tools for evaluating data. (These tools are discussed in Chapter 6.) Emerging standards for accepting database-match protein identifications and the new statistical models to distinguish correct from incorrect identifications will make it easier to apply standardized methods to evaluate data [Taylor et al., 2003]. As these resources are brought online in a proteomics laboratory, it is essential that one-on-one training or small group workshops for new users be offered to acquaint users with these new tools and their applications. More complex problems, such as mapping protein modifications and performing quantitative analyses will require a greater level of involvement by proteomics laboratory staff with specialized expertise. Nevertheless, these skills can and should be translated aggressively to users who will need them on a continuing basis.

1.6 REFERENCES


