(Book chapter in *Microbial Functional Genomics*, Jizhong Zhou and Dorothea Thompson, editors, Wiley and Sons, Inc., 2004, pp. 241-282)

CHAPTER 9. MASS SPECTROMETRY

Nathan VerBerkmoes, Joshua Sharp, and Robert Hettich
Chemical Sciences Division
Oak Ridge National Laboratory

9.1. INTRODUCTION

The recent technological revolution in instrumentation and computational power has enabled the development of sophisticated analytical technologies that now permit a detailed glimpse into the molecular machinery of the living cell. These important tools include electron microscopy, x-ray crystallography, nuclear magnetic resonance, mass spectrometry, and many others. The availability of these advanced analytical technologies at a time concomitant with revolutionary developments in molecular biology has enabled the investigation of biological structures and interactions that have previously been inaccessible.

Mass spectrometry (MS) is a rapidly-advancing family of structural biology tools that have in common the measurement of molecular *ions* of intact and fragmented biomolecules. A common misconception about MS is that it only provides molecular mass information. In reality, MS is not only powerful for molecular mass measurement, but also provides ion manipulation capabilities for obtaining detailed structural information at the isomeric level (including differentiation of isomers in many cases). While mass spectrometry is a well-established technique that historically has been important for small molecules, advances since the late 1980's have made MS applicable to large biomolecules such as proteins, nucleic acids, and their complexes. The key to forming gas-phase ions from these larger molecules, a prerequisite for MS analysis, lies in using electrospray ionization (ESI) (Fenn, 1989) or matrix-assisted laser desorption/ionization (MALDI) (Hillenkamp, 1991; Nakanishi, 1994). With these ionization techniques, MS can be used as a new type of high-resolution "readout" for many existing biological and biochemical procedures, providing biological information in the form of molecular mass.

The completed sequencing of genomes for several organisms, including humans (Venter, 2001), has ushered in a new era of biology. The impact that this information will have on a more complete understanding of the processes of life, including human health, is only now beginning to be realized. A natural extension of *genomics* (which is the study of the complete set of genes for an organism) research is the characterization of the gene products, which are proteins. This has spawned the research area of *proteomics*, which is the study of the entire suite of proteins from a genome. Mass spectrometry has quickly become one of the leading technologies for proteome measurements, due to its inherent ability to identify proteins, including hypothetical species, at high mass accuracy, resolution, and throughput, even from complex mixtures (Peng, 2001; Larsen, 2000). For instance, the advent of modern MS techniques has greatly expanded the utility of the venerable proteolytic digest. Accurate MS measurement of the masses of the proteolytic peptides yields a so-called "mass map" that can be used to identify a protein from published databases, and is a foundation technology of the rapidly expanding field of proteomics

(Pandey, 2000a). Furthermore, tandem mass spectrometry (MS/MS) enables full or partial sequencing of the proteolytic peptides, in some cases without prior separation (Yates, 2000).

The objective of this chapter is to illustrate how mass spectrometry is becoming an essential tool for characterizing complex mixtures of proteins. The chapter will be divided into two major sections. The first section will provide a condensed description of the fundamentals of biological mass spectrometry, discussing the major ionization methods, the different types of mass analyzers, how ion structure can be interrogated, and finally how proteins and peptides can be characterized. The second section will outline the status of mass spectrometry for complex protein measurements such as proteomes, in particular detailing how bottom-up approaches (employing proteolytic digestion and peptide identifications) and top-down approaches (measurement of the intact proteins) can provide comprehensive proteome information.

9.2. FUNDAMENTALS OF MASS SPECTROMETRY

In order to understand the impact that mass spectrometry can have for protein measurements, it is useful to define the basic operating principles. The goal of this section is not an exhaustive review of mass spectrometry, which is given elsewhere (Dass, 2001), but rather to discuss the basic concepts of ionization, mass analyzer types, and biomolecular detection in sufficient detail to prepare the reader for a detailed discussion of complex protein characterization with this technique. These concepts will then be demonstrated for protein and peptide characterization.

9.2.1. Basic components of any mass spectrometer

Even though there are several different types of mass spectrometers, they all share three important fundamental components. These are

i) *Ion source*, where gas phase ions are generated from the sample.
ii) *Ion analyzer*, where the charged particles are separated or sorted by their mass/charge ratios.
iii) *Ion detector*, where the abundances and mass/charge values are measured.

In almost all current instruments, a dedicated computer workstation controls all the aspects of this measurement, ranging from ion introduction to data analysis. The first component of the mass spectrometry experiment is the ionization step, in which gas phase ions are generated in an ion source from a solution phase sample. This process is not trivial and historically has been the source of much frustration for large biomolecules. Early work in fast-atom bombardment mass spectrometry (FAB-MS) and plasma-desorption mass spectrometry (PD-MS) revealed some promise for biomolecules with molecular masses up to about 10 kDa. It was not until the advent of MALDI and ESI (which will be discussed in more detail below) that MS became a useful method for examining much larger species (i.e. molecular masses exceeding 100 kDa). The functions of the ion source of any mass spectrometer are three-fold: 1) to efficiently and non-discriminately generate gas phase ions from the analyte, 2) to avoid fragmenting the analyte in the ionization process, and 3) to inject the gas phase ions into the mass spectrometer.

Mass spectrometry would be a useless tool if the gas phase ions could not be partitioned or sorted in some meaningful way to disperse them for subsequent detection. Thus, the second

critical component of any MS instrument is the ion analyzer. All mass spectrometers use electric and/or magnetic fields to alter the motion of the ion packet emerging from the source in such a way that ion dispersion occurs. This sorting process enables ions of similar mass/charge to be grouped together and separated either spatially or temporally from ions of dissimilar mass/charge. For example, in a time-of-flight mass spectrometer, a high-voltage direct-current pulse is used to accelerate the ions out of the ion source and into an open flight tube. Because this voltage imparts the same kinetic energy to each ion regardless of its mass/charge, all of the ions have different velocities (i.e., K.E. = ½ mv^2). Thus, this process provides spatial resolution of the ions in that the lighter ions (i.e., those with lower m/z) travel faster and reach the ion detector before the heavier ions. By recording the ion signal as a function of arrival time, it is possible to obtain a time-of-flight mass spectrum.

The third and final component of all mass spectrometers is the detector. In every case, this process relies on the fact that charged particles provide a macroscopic electronic effect on a suitable detector. The most common detector employed is an electron multiplier. For this device, the ions impinge on a metal target. Electrons dislodged by this process are multiplied by a cascade process and give rise to an amplified signal each time an ion strikes the device. This analog signal is converted to a digital signal that can be recorded and manipulated by the computer workstation. A more detailed description of these three fundamental components will be provided in the following sections.

9.2.2. Ionization methods

Although there are numerous ionization methods available for mass spectrometry, two main ones dominate the larger biomolecule studies. These are electrospray ionization and matrix-assisted laser desorption/ionization. The development of these methods has been reviewed in several recent articles (Vestal, 2001; Cech, 2001). The basis for the popularity of these two ionization methods is the ability of these techniques to provide gas phase ions from biomolecules whose molecular masses exceed 100 kDa. Until recently, mass spectrometry was useful primarily for molecules with masses below 5 kDa; however, these two ionization methods have provided more than an order of magnitude in the accessible molecular mass, and thus have revolutionized mass spectrometry for biological studies. Being able to directly examine compounds with molecular masses at least to 100 kDa has opened the door to peptide, protein, and even oligonucleotide studies, thus significantly engaging the attention of the biological community.

For protein investigations by mass spectrometry, the principle ionization methods are:

(i) *Electrospray ionization (ESI or ES).* This technique utilizes a high voltage needle (typically about 4,000 V) to transfer preformed ions from solution phase into the gas phase. This resulting mass spectra usually consist of a range of multiply-charged ions, such as $(M+nH)^{n+}$. For most ESI-MS experiments, a protein solution in the nM to uM concentration range would be prepared in water:acetonitrile (~50:50 by volume), with about 0.1 – 1.0 % acetic acid added to protonate the proteins in solution. Because the electrospray ionization process is quite sensitive to charge-carrying species, great care must be taken to minimize or exclude buffers, salts, carrier proteins, and other species that might interfere with measurement of the analyte of interest. If protein stability demands the inclusion of buffer, it is desirable to use a volatile buffer such as ammonium

acetate (concentrations of ~ 1-100 mM). The on-line coupling of liquid separation techniques, such as HPLC, with ESI is relatively straight-forward.

(ii) *Matrix-assisted laser desorption/ionization (MALDI)*. This method is conducted by using a pulsed laser to desorb biomolecules that have been imbedded in a spectrally-absorbing matrix compound (typically a small organic acid). The resulting mass spectra consist primarily of singly-charged species, such as $(M+H)^+$, although some higher charged species, especially doubly-charged ions, are observed in some cases. Typical matrix compounds for proteins and peptides include sinapinic acid, alpha-cyano-4-hydroxycinnamic acid, and 2,5-dihydroxybenzoic acid. MALDI is reasonably tolerant of the presence of salts, buffers, and other additives in the sample, and is often superior to ESI for "dirty" samples. The on-line coupling of liquid separation techniques, such as HPLC, with MALDI is fairly difficult.

9.2.3. Mass analyzers

Although ionization is vital for transferring the solution phase analyte into the gas phase, it is critical to have a method to separate and detect those ions. This is the function of the mass analyzer. Because ion motion can be controlled by electric and magnetic fields, these fields are used to manipulate the ion motion for subsequent detection. There are several types of mass analyzers, as will be discussed below, but they all share the common attribute of altering ion motion for selective detection. When comparing different mass analyzers, it is useful to evaluate them based on the following figures of merit:

(i) *Mass resolution*. The measure of how well adjacent peaks can be differentiated in the mass spectrum. This value is typically given as the peak full width at half maximum (FWHM).

(ii) *Mass accuracy*. The comparison of the measured mass to the calculated mass. This value is typically given as error in either percentage or parts-per-millions (ppm).

(iii) *Mass range*. The difference between the largest and smallest molecular mass that can be measured.

(iv) *Detection limits*. The smallest amount of sample that can be measured with a signal/noise of at least 3:1.

(v) *Dynamic range.* The molar difference between the least abundant component and the most abundant component that can be detected in a single sample.

(vi) *Scan speed*. The time that it takes for a given mass analyzer to record a complete mass spectrum.

(vii) *Tandem mass spectrometry*. The capability of a given mass spectrometer to isolate and fragment a selected ion for structural interrogation.

The following mass analyzers are commonly employed for biological studies:

(i) *Linear quadrupoles (Q)*. The heart of this type of mass spectrometry consists of four parallel metal rods, electronically isolated from each other, that are arranged to permit ion transmission in the longitudinal space between them. A combination of dc and ac radio-frequency (rf) voltages is placed on the rods to control ion motion. The mass/charge of the ions that are passed through the rods is proportional to the voltage applied to the rods (the ratio of rf to dc voltage is held constant). By scanning the rf voltage, ions of a given mass/charge ratio are sequentially allowed to pass through the

rods and into the detector, allowing a mass spectrum to be acquired. ESI is the most common ionization technique used for this instrument. Ions are detected by an electron multiplier placed at the end of the rods.

(ii) *Time-of-flight (TOF)*. This instrument consists of an ion source, a flight tube (about 1 m in length), and an ion collector. A high voltage is applied to the ion source, to accelerate the ions out of the source and into the flight tube. Because all ions of the same charge state are given the same kinetic energy, the lower mass/charge ions have higher velocities than the higher mass/charge species. Thus, ion spatial separation is achieved in the flight tube. The lightest ions strike the detector first, followed eventually by the heavier ions. By recording ion intensity as a function of time, it is possible to obtain a time-of-flight mass spectrum. A reflectron (or ion mirror) can be used to compensate for the range of kinetic energies resulting from the ionization process, and provides enhanced mass resolution. Both ESI and MALDI are employed for TOF, although MALDI is much more common.

(iii) *Sectors*. This instrument uses a combination of electrostatic and magnetic fields to guide an ion beam from the ionization source to an electron multiplier detector. The electric and magnetic fields serve to focus and resolve the ions spatially, and to transmit only one particular mass/charge species into the detector at one time. By scanning the magnetic and/or electric fields, a complete mass spectrum can be recorded. Because ion detection is usually accomplished in a continuous mode, ESI is the most common ionization source. MALDI is a pulsed technique and is less suitable for interfacing to this type of instrument.

(iv) *Quadrupole ion traps (QIT).* Based on the linear quadrupole principles discussed above, it is possible to construct a three-dimensional quadrupole ion trap (Hao, 2001). This device has a ring electrode and two end caps. Unlike the analyzers discussed above, the QIT has the ability to trap ions spatially for a period of time. Ion trapping is achieved with a combination of dc and rf voltages applied to the ring electrode and end caps. This permits the ions to be stored in the enclosed volume of the trap for a period of time for further interrogation. Ion detection is accomplished by ramping the rf voltage to sequentially eject the ions from the trap into an electron multiplier. ESI is the most common source, although MALDI has been employed to some extent as well.

(v) *Fourier transform ion cyclotron resonance (FTICR).* This is one of the highest performance types of mass spectrometers (Marshall, 1998; Marshall 2001). Ions are trapped in a cylindrical ion cell, which is situated inside a high-field superconducting magnet. By confining the ions electrically and magnetically, it is possible to store them for extended periods of time. Ion detection is achieved by measuring the rf cyclotron motion of the orbiting ion packet. Because the electronic frequencies can be measured so precisely, it is possible to obtain both high resolution and high mass accuracies. ESI is the most common source (Hendrickson, 1999), although MALDI is used to a limited extent.

Table 9.1 compares performance features for these mass analyzers. There is not one specific mass analyzer that is completely universal for all applications. Often, the nature of the biological problem (and the financial resources available) will dictate the choice of instrument. It is not uncommon for different types of mass analyzers to be linked together, to overcome some of the

individual limitations. For example, a hybrid quadrupole-time of flight (QTOF) instrument is becoming widely used for protein studies (Chernushevich, 2001). This instrument is designed to exploit the positive capabilities of each of the two mass analyzers and provides rapid mass measurement under enhanced resolution. It should be noted that some of these instruments are very straight-forward for non-technical experts to use, whereas the higher performance instruments tend to require substantial technical training for operation.

9.2.4. Coupling separation methods with mass spectrometry

Although mass spectrometers have sufficient mass resolving power to be able to simultaneously differentiate very complex mixtures, numerous experimental parameters diminish the practical application of this capability. For example, consider the proteolytic digestion of a sample containing 100 different proteins. This process will generate an even more complex sample, which might contain between 2,000-5,000 peptides. Even though the higher performance mass spectrometers discussed above have sufficient resolution to measure all of these peptides in a single mass spectrum, the more easily ionized species will be over-represented in abundance in the mass spectra, whereas the more poorly ionized species will be suppressed to the extent that they might not even be observable. This is termed *ion suppression*, and is readily observed in both ESI and MALDI mass spectra because of the competition for charge in the ionization process. Thus, even though two different peptides may be present at equimolar concentrations in the sample, it is quite possible that one species will be observed in high abundance and the other species not even detectable in the mass spectrum. One obvious solution to this dilemma is to couple some type of separation stage, such as chromatography, with the mass spectrometry experiment. In order to minimize sample losses due to handling, the current state-of-the-art instrumentation often incorporates liquid chromatography directly on-line with the mass spectrometer. For example, high performance liquid chromatography is readily interfaced with ESI-MS. The advantage of this setup is easily recognized; the on-line separation step, which is often based on reverse-phase liquid chromatography, provides partial purification of the proteins or peptides, so that they enter the mass spectrometer at different times. This partially alleviates the ion suppression problems, as fewer species are measured at the same time. The practical details of how this is accomplished and the benefits it provides will be outlined in the second half of this chapter.

9.2.5. Ion structural characterization

Mass spectrometry has earned the reputation of being an excellent method for accurately measuring masses. However, the real power of this technology is the ability to manipulate the gas phase ions in order to study their reactivities and fragmentation pathways. These experiments provide detailed structural information. The two basic types of mass spectrometry experiments to investigate structure are:

(i) *Ion fragmentation studies*. This is generically referred to as "tandem mass spectrometry" or "mass spectrometry-mass spectrometry, (MS/MS)". This can be achieved with collisional dissociation or photofragmentation.

(ii) *Ion reaction studies*. These are conducted by reacting selected gas phase ions with various reagents. For example, gas phase ions can be reacted with deuterium oxide to study hydrogen/deuterium exchange reactions to locate mobile hydrogens.

Collisional activated dissociation (CAD) is the most common type of ion fragmentation experiment. This procedure usually involves three key steps: ion isolation, ion dissociation, and fragment ion detection. Ion isolation is accomplished by removing all unwanted ions in the mass spectrometer except for a selected parent ion mass/charge ratio. This parent ion is accelerated translationally by applying appropriate voltages and then collided with a target gas such as helium, argon, or nitrogen. Upon impact, the translational energy of the ion is converted to internal energy, causing fragmentation. By controlling the amount of excitation energy, it is possible to induce either limited or extensive fragmentation. The resulting fragment ions are then measured and provide information about the structure and sequence of the parent ion. Alternatives to gas phase collisional activated dissociation include fragmenting the parent ions with photon-, electron-, or surface-collisions.

Ion reactivity studies provide another route to unravel biomolecular structures. The most common experiment conducted for proteins is hydrogen/deuterium exchange reactions, which can be conducted either in the solution phase or in the gas phase inside the mass spectrometer. The primary goal of these experiments is to determine which hydrogens are immobile due to participation in higher order biological structure (Smith, 1997). For example, it is possible to investigate protein secondary structure such as alpha helices and beta sheets with this methodology. This technology is also being extended to the examination of protein-protein interactions by mapping which hydrogens are inaccessible in the protein complex relative to the individual monomers. The reader is referred to the following references for a summary of how mass spectrometry is becoming an important tool for studying biomolecular conformations (Hernandez, 2001; Kaltashov, 2002; Bennett, 2000).

9.3. FUNDAMENTALS OF PROTEIN AND PEPTIDE MASS SPECTROMETRY

In the measurement of mass spectra, it is helpful to define some of the basic terms. They are:

(i) *Average molecular mass*. The sum of the atomic masses of all the composite elements in a given molecule. Since *average* atomic masses are used in the calculation, the molecular mass is therefore calculated as an *average* mass.

(ii) *Monoisotopic molecular mass.* The sum of the most abundant isotopes of all the composite elements in a given molecule.

(iii) *Isotopic packet.* The isotopic distribution observed resulting from the natural abundance of isotopes from the composite elements in the molecule.

(iv) *Protonated molecule.* The ion that results from the addition of one or more protons to the molecule. In MALDI-MS, the most abundant positive ion is often the singly-charged $(M+H)^+$, whereas electrospray usually reveals a series of multiply-charged $(M+nH)^{n+}$ ions.

9.3.1. Protein measurements

One of the driving forces of mass spectrometry for protein and peptide characterization is the ability to measure the molecular masses to a high degree of accuracy, often better than 0.01%. In MS investigations of small molecules, such as peptides with molecular masses less than 5,000 Da, it is convenient to consider the monoisotopic molecular mass, as this is usually

the most abundant ion. In this case, the isotopic packet is fairly small (often only one-three measurable peaks). MS examinations of larger proteins reveal a substantially different scenario. For proteins with molecular masses in excess of 10,000 Da, the monoisotopic peak is virtually unmeasurable, and the most abundant peak is closer in mass to the average value. The isotopic packet is substantially larger in this case and includes abundant ions that range over several Da. These factors obviously affect the nature of the mass spectrometry measurement, and thus must be taken into account.

In order to illustrate the capability of MS for protein measurements, consider the protein ubiquitin. This has become an almost universal standard for mass spectrometry, as this protein is readily available in high purity, is remarkably stable, and produces excellent MS signals in either MALDI or ESI modes. Ubiquitin is a small protein consisting of 76 amino acids. This protein has an empirical formula of $C_{378}H_{629}N_{105}O_{118}S_1$, yielding a *calculated average molecular mass* of 8564.8721 Da. In a low resolution MS measurement, such as those obtained on the common linear MALDI-TOF instruments, the molecular ion region will appear as a broad peak with a mass centroid at the average mass value. The presence of salt, which is a very common contaminant from handling, in the sample will be reflected by the presence of additional peaks slightly heavier than the molecular mass. At moderate resolution, such as that obtainable on a sector MS, the broad molecular ion region begins to partially resolve into discernable peaks corresponding to the isotopic packet. Under high-resolution conditions, such as those possible with an FTICR-MS, the molecular ion region appears as a Gaussian distribution of well-resolved peaks separated by one Da intervals. These correspond to the isotopic signature due to the naturally-occurring isotopes of carbon, hydrogen, nitrogen, oxygen, and sulfur in the molecule. The primary contributor to this packet is ^{13}C, which is 1.1% in natural abundance. The abundance of this isotopic contribution in the molecular ion of ubiquitin may be somewhat surprising, until one realizes that each protein molecule contains more than 350 atoms of carbon. Statistically, at this level, it is much more likely that one or more ^{13}C are present than that all the carbons are ^{12}C. The isotopic packet is about 10 Da wide, with an almost symmetric appearance. The monoisotopic peak at 8559.6152 Da is almost undetectable, due to the low abundance of this species in the natural isotopic packet. The most abundant isotopic peak is calculated to be 8564.6296 Da, which is close but not identical to the average mass of 8564.8721 Da. It should be apparent that the mass resolution of the MS measurement has to be taken into account with respect to the expected masses to be measured. All of these effects are further accentuated with larger proteins, which have a substantially wider isotopic packet (Horn, 2000c).

Figure 9.1a reveals the *measured* ES-FTICR mass spectrum of ubiquitin. Note that this figure shows a variety of multiply-charged ions, which is very typical for electrospray ionization. Figure 9.1b illustrates the deconvoluted molecular mass spectrum, obtained by mathematically removing the multiple charges to yield the neutral protein species. *The measured mass of the most abundant isotope is 8564.637 Da, which is 0.0074 Da (or 0.9 ppm) different from the calculated mass listed above.* It is this level of mass accuracy that has enabled MS to become an essential tool for biological studies.

If MS had only the ability to measure protein molecular masses accurately, this would be an important, but limited, capability. What drives MS beyond this level is the ability to manipulate and interrogate these ions, thereby providing structural information (Horn, 2000b; Loo, 1992; Little, 1994; Senko, 1994). For example, Figure 9.2 illustrates the MS/MS of the $(M+10H)^{10+}$ ion of ubiquitin in an ESI-FTICRMS experiment. Several observations can be drawn from the fragment ions observed in this mass spectrum. First, there are a variety of

fragment ions ranging from the molecular mass down to small peptide species. Second, there are blank zones in which limited or no fragmentation is observed. Third, the fragment ions can be measured to high accuracy as well, providing definitive information about dissociation pathways. Even though complete sequence information is often not possible from such a measurement, an enormous amount of structural information can be gleaned from such a mass spectrum, as will be illustrated in the later sections of this chapter. In order to interpret the MS/MS and identify fragmentation sites within the protein, the following nomenclature is used. Proteins typically fragment by cleavage along the peptide backbone. Because the cleavage can occur at multiple sites, a systematic alphabetic code is used, as shown in Figure 9.3 below. Fragment ions which retain the charge on the N-terminus end of the original protein are designated as "a, b, or c" type ions, depending on the cleavage site. Fragment ions that retain the charge on the C-terminus end of the original protein are designated as "x, y, or z" type ions, depending on their cleavage site (Roepstorff, 1984; Biemann, 1988). The most common fragment ions observed for proteins are usually b- and y-type ions. This is illustrated by the annotation in the inset for Figure 9.2 for the MS/MS of ubiquitin. There are several web-based tools, such as those available at the Rockefeller PROWL web-site (http://prowl.rockefeller.edu), that can be used to predict the fragment ions from a given protein sequence.

9.3.2. Peptide measurements

Even though accurate molecular mass and limited fragmentation are achievable for intact proteins, this information often is insufficient for the *unambiguous identification of the complete sequence*. One of the most logical solutions to this dilemma is to utilize solution phase proteolytic digestion procedures to cut the protein into characteristic peptides. The resulting peptides can be measured under high-resolution conditions and fragmented with MS/MS techniques. Because peptides are smaller in molecular mass than intact proteins, they fragment more readily and provide more extensive sequence information. By combining this information with known details about the protein digestion process (i.e., the identities of the amino acids cleaved), it is possible to obtain the detailed information necessary to identify the protein. In fact, this information can be used to query protein databases for automated protein identification, as will be illustrated in the following sections.

There are a variety of common proteases that can be used to cleave proteins into peptides. These range in cleavage specificity, providing control over the desired degree of protein fragmentation. These proteases are active under different solution phase conditions. The most common protease employed is trypsin, which is readily available in purified form and cuts proteins at the C-termini of arginine and lysine residues (provided that the adjacent residue is not proline). This produces fragment ions containing basic amino acids at the C-terminus (which is advantageous for positive ion MS measurements). A protease digestion experiment generally begins with protein denaturation and disulfide bond reduction to open up the protein for more efficient degradation. After the digestion is complete, it is essential to clean-up the sample with reverse-phase HPLC or solid phase extraction to remove the salts prior to MS measurements of the peptides. In order to minimize autodigestion products (generated by trypsin digesting itself), it is desirable to utilize sequencing grade modified trypsin, which is available from a variety of vendors as a solution phase reagent or immobilized on beads.

Because the digestion procedure often does not proceed to completion, it is useful to consider peptides that might result from missed cleavage sites. For example, we have used a

tryptic digestion procedure to characterize an important yeast DNA checkpoint protein, Sml1p (Uchiki, 2002). Two important points were noted in this experiment: 1) the tryptic peptides were observed at dramatically different abundances in the mass spectra, which reflect the efficiencies of the digestion and cleanup procedures, as well as the common ion suppression effect, and 2) the sequence coverage was less than 100% (i.e., some of the predicted peptides are absent). The masses of the peptides observed are listed in Table 9.2, revealing how the measured peptide masses correlate with the predicted masses. For this particular case, 94% of protein sequence was determined from the peptide measurements. Any of these peptides can be interrogated with MS/MS experiments. For example, the MS/MS of the peptide with Mr = 1865 Da (residues 1-16) yielded a variety of abundant y- and b-type fragment ions. Because the y-type ions were the most abundant series of ions, it was possible to identify an amino acid sequence tag of **NSQDYFY** in this peptide. This sequence tag information is very useful for not only confirming the identity of the peptide but also for determining its position within the protein. Furthermore, it is possible to query the protein databases with sequence tag information, which is sufficient in many cases to identify a protein without any additional information.

9.4. MS FOR PROTEIN AND PROTEOME CHARACTERIZATION

The first section of this chapter outlined the basic concepts of mass spectrometry (MS) for protein analysis. This section will discuss how this methodology can be applied for the analysis of individual proteins, protein complexes and whole proteomes, with a focus on microbial proteomes. It should be noted that while MS-based proteomics has virtually exploded in the last five years, there are other methods to analyze whole proteomes and protein complexes, such as protein chips, antibodies, and yeast two-hybrid analyses (see Chapter 10). While some of these other techniques may show great promise for the future, none have the current capabilities for rapid and molecular level analyses of protein mixtures that can be afforded by the MS-based techniques. For this reason, we will limit our discussion to techniques that are either directly or indirectly coupled with mass spectrometers for complex protein mixture analysis.

Proteome analyses, whether in simple microbes, yeast, or higher organisms, present a much greater challenge than the genomics sequencing efforts. *While the genome is static, the proteome is dynamic*. The genome contains generally a set number of copies of every gene; however, proteins in the proteome can be expressed in a wide concentration range, varying from only a few copies per cell for regulatory proteins to many thousands per cell for ribosomal subunits. The proteins derived from theses genes can take many different forms due to alternate transcript splicing (primarily observed in higher eukaryotes). Furthermore, proteins can be highly decorated with any number of post-translational modifications (PTMs); more than one hundred have been recorded (O' Donovan, 2001.). These modifications can be static or dynamic, and may be present in multiple places on a protein. Finally, in the current state of proteomics technologies, there is no amplification technique for proteins similar to the polymerase chain reaction (PCR) that has become so important for oligonucleotide studies. Thus, the proteome of even simple microbes presents a much greater analytical challenge than the corresponding genome (or even transcriptome) analysis. Even in light of these difficulties, a complete understanding of microbes and microbial communities necessitate the development of analytical techniques for rapid and accurate analysis of whole proteomes and protein complexes.

In order to achieve such a "systems biology approach", it is essential to interface proteomics, bioinformatics, genomics, transcriptomics, and metabolomics technologies.

9.4.1. Overview of MS approaches for protein studies

Protein analysis by MS-based methodologies can be broken down into the following three general areas. While each is unique, they all can be addressed with similar MS technology.

*(i) Individual Protein Analysis***.** This involves the analysis of purified proteins for quality control in structural or biochemical experiments, as a means of studying post-translational processing, or in structural analysis of individual proteins by techniques such as H/D exchange (Dharmasiri, 1996), cross-linking (Young, 2000) and surface labeling (Bennett, 2000). While these methods of structural analysis by MS all show great promise for the future, a detailed discussion is beyond the scope of this chapter. Protein detection over a wide dynamic range is not much of an issue here, but sensitivity may be, for example, if the goal is to purify a low-copy number protein with a transient modification.

(ii) Protein Complex Analysis. This involves the analysis of purified protein complexes, which may also include the analysis of signaling pathways in which the protein complexes are generally more transient and difficult to analyze (Pandey, 2000b). Protein complexes are typical purified by centrifugation/sucrose gradient techniques, although more recent approaches employ immunoprecipitation (Pandey, 2000b) or tandem affinity purification tags (TAP) (Puig, 2001). The greatest challenge is in the biochemical purification and the sensitivity required in the analysis, since typically 100-1000 ng of the complex can be prepared, although preparation of protein complexes from microbes can often yield much greater quantities. Dynamic range in the MS detection is not as much of a problem but may be important if a transient protein is associating with a large complex or if one is searching for PTMs on this complex. A key element of these measurements is the need to purify these complexes from cell lysates. This field is developing rapidly at present, as evidenced by attempts to characterize protein-protein interactions at the proteome level (Gavin, 2002; Ho, 2002; von Mering, 2002).

(iii) Whole proteome analysis. This typically refers to the analysis of whole cell lysate, organelle preparations, or crude fractions obtained by affinity chromatography or centrifugation, such as membrane preps, cleared serum, etc. The greatest challenge in this analysis is dynamic range, because medium to high abundance proteins mask most low abundance proteins. Sensitivity is generally not as much of a problem for whole cell lysates due to large quantities of starting material, but may be an issue for organelle preparations or affinity purified/cleaned fractions.

The most complex of these procedures is the examination of a whole proteome. The experimental approach for such a measurement can be broken down into four separate steps. The first step of sample preparation (i.e., cell growth or sample isolation) is followed by the second step of protein fractionation or separation. This leads to the third step of MS characterization, which is followed by the final step of computational analysis of the data (bioinformatics). For the analyses of protein complexes and whole proteomes, the most important aspect is the dynamic range of the measurement. The demands on the MS can be relaxed somewhat by incorporating separation technologies (such as either off-line or on-line fractionations of proteins/peptides) prior to mass spectrometric detection. These fractionations

can be very crude or highly specific, depending on the nature of the application. Each of the four steps for proteome analysis will be discussed below.

Sample preparation

The first step in obtaining a proteome sample entails common microbiology techniques of cell growth and isolation. For typical microbial proteome analysis, the cells can be cultured under a variety of growth conditions for comparative analysis. Depending on the species, 500 ul – 4 L of cell culture is more than sufficient for multiple analyses by any of the proteome methodologies that will be discussed below. The cells are typically harvested and washed with a buffer to remove excess media. The cells can then be disrupted with various techniques such as sonication, bead beating, or French press. The methodology for cell disruption will depend on the species type and the type of proteome analysis that will be applied.

Crude fractionation techniques

(i) Centrifugation. Cell extracts can be processed by centrifugation (varying speed and duration) to create different soluble and insoluble fractions. Sucrose gradients can be used for finer fractionation. In most cases, three lysate fractions are prepared: a crude soluble, a cleared soluble, and membrane fraction.
(ii) Chemical extraction. Cell lysates or membrane fractions can be treated chemically with chloroform/methanol or acetone to remove lipids and small molecules from protein samples. This technique is very useful in cleaning up membranes, especially from photosynthetic organisms such as cyanobacteria and plants.
(iii) Affinity depletion. While this methodology is a form of chromatography, its primary purpose is to remove high abundance proteins from a sample to enhance MS detection of more minor species. For example, the removal of albumin and immunoglobulins from serum is a common type of affinity depletion and enables detection of other more minor proteins in this physiological media.

Chromatography separation techniques

(i) Gel filtration (or size exclusion). In this technique, proteins are separated according to their sizes through a gel filtration column. This crude separation lacks resolving power and usually is used in the first step of fractionation.
(ii) Affinity chromatography. This technique is a very powerful tool for analyzing low abundance proteins and protein complexes. Basically, intact proteins or peptides are enriched over an affinity matrix by exploiting intrinsic properties (such as the presence of phosphate groups on protein) or chemically- or biologically-added labels.
(iii) Ion exchange chromatography. This technique separates peptides or proteins by their surface charge. Typical experiments employ cation exchange for peptides and anion exchange for proteins.
(iv) Reverse-phase chromatography. This method relies on separation of intact proteins or peptides according to their overall hydrophobicities. As such, it provides a high level of resolution and is typically the final separation method prior to MS analysis, due to the fact that its solvent requirements (mixture of aqueous and organic phase with small amount of organic acid) are compatible directly with ESI and MALDI.

Electrophoresis separation techniques

(i) Capillary Electrophoresis. This separation methodology is based on the differential electrophoretic mobility of peptides or proteins in the presence of an electric field gradient applied along a flowing liquid in a gel-filled capillary. While this method originally showed great promise due to speed, resolving power and sensitivity, it has proven not to be very rugged and is difficult to couple on-line with MS. These factors render it very limited in general proteomics applications. It is possible that this methodology could be revitalized in micro-separations on chips.

(ii) Gel electrophoresis. This separation methodology is based on the differential electrophoretic mobility of peptides or proteins in the presence of an electric field gradient applied across a slab gel. This technique has several variations, but in proteomics commonly involves polyacrylamide gel electrophoresis (PAGE), in which isoelectric focusing and molecular mobility provide for a two-dimensional protein separation.

MS analysis techniques

Currently, there are two major methods for analyzing proteins by mass spectrometry. The *top-down* method involves measuring intact proteins as well as interrogating the MS/MS of these intact proteins. This method was first introduced with ESI-FTICR-MS (Little, 1994; Mortz, 1996; Kelleher, 1998) and expanded to ion traps with novel ion-ion reactions (McLuckey,1998). In the *bottom-up,* or *shotgun* method, intact proteins are digested with a protease such as trypsin, Glu-C or cyanogen bromide (CNBr), and the resulting peptide mixtures are analyzed by MS or MS/MS. It should be noted that in this definition it does not matter whether separations are performed on intact proteins or peptides; rather, the experiment type is defined by the species measured by the MS. Thus, 2D-PAGE followed by in-gel digestion and MS analysis is considered a bottom-up approach. The actual development of the bottom-up methodology cannot be traced to a single lab, but rather evolved from multiple labs using very different techniques including gel-based (Hess, 1993; Mortz, 1994; Shevchenko, 1996a; Wilm, 1996; Gatlin, 1998) and solution-based separations (Hunt, 1981; Hunt, 1986; McCormack, 1997; Martin, 2000; Shen, 2001) followed by MS or MS/MS for protein identifications. These two general approaches can be summarized as follows:

i) *Bottom-up proteomics*. Complex protein mixtures (from cell lysate or protein complexes) are proteolytically digested (usually with trypsin), and the resulting peptide mixture is examined by mass spectrometry. The MS data are used to query a peptide database from the specific organism to identify the protein components of the original mixture. This method is excellent for determining protein identities, but provides very limited information about the molecular form of the intact proteins.

ii) *Top-down proteomics.* Complex protein mixtures from cell lysates or protein complexes are examined directly by on-line or off-line MS. No digest is conducted; rather the intact proteins

are measured with MS and MS/MS. This method provides fewer protein identities, but does give detailed information about protein processing (PTMs, truncation, mutations, signal peptides).

Both techniques have advantages and disadvantages and will be discussed in detail below with highlights of the current technologies used. Bottom-up proteomics is by far the more widely used method, mainly because it is much simpler to conduct and does not require high performance MS instrumentation. The progress in the field of bottom-up proteomics has been staggering. Table 9.3 highlights some of the groundbreaking research over the last five years in large-scale proteome analysis. Each paper illustrates either a new technology in bottom-up proteomics or a major advance in protein characterization from the given organism. This list should not be considered all inclusive or compared directly, since there were many differences in the number of growth conditions, the size of the experiments, and scoring of proteins identified. Rather, this list should be viewed as a trend in the increase in depth of proteome coverage and speed of analysis over the last few years. It has now become possible (if not routine) to measure ~1000 proteins from a microbe under a given growth condition with a high degree of confidence in 1-3 days time, depending on the technology used. Furthermore, if enough mass spectrometers are assembled, this analysis can be rapidly repeated for protein identification for an organism under a variety of different growth conditions.

On the other hand, top-down proteomics has moved along at a much slower pace. This is primarily due to the following factors:

i) Liquid-based separations of intact proteins are more difficult than peptides.
ii) MS and MS/MS analyses of intact proteins are more difficult to conduct and interpret than peptides.
iii) The MS instruments capable of adequate analysis of intact proteins from complex mixtures are either very expensive or not commercially available, and have not been designed for routine operation in most cases.
iv) The algorithms to analyze MS/MS of intact proteins are not as well developed or commercially available.

Even with these experimental challenges, top-down proteomics provides a level of information that the bottom-up technique does not, which is the *intact state of the protein*. This is critical, as proteins function as intact molecular species, not as a combination of simple, small peptides. Thus, a full understanding of the intact state of proteins (PTMs, truncation, mutations, signal peptides) is necessary, suggesting that an integrated top-down, bottom-up proteomics method would be the most comprehensive. This approach will be discussed in more detail below.

Data processing and bioinformatics

One of the key hallmarks of MS-based proteomics is the large quantity of data that can be produced in short periods of time. Without the advancement of bioinformatics tools to process, sort, and compile these data sets, the field would not be able to move forward. Furthermore, the recent completion of large numbers of genomic sequences has been as much of a factor in the rapid evolution of the field of proteomics as have been the advances in mass spectrometry. While the proteins from *unsequenced* microbes can be identified with MS by a methodology termed *de novo* sequencing, this methodology is limited in the speed and depth of analysis of the

proteome. In typical *de novo* sequencing experiments, MS/MS spectra from enzymatically derived peptides are analyzed manually or with a computer algorithm to identify sequence tags (stretches of at least 5-7 consecutive amino acids). These sequence tags then are used to blast against known microbial databases in an attempt to determine the nature of the proteins in the unsequenced organism. The most common technique is to analyze proteomes of microbes after their DNA sequence has been completed and annotated. These microbial DNA or protein databases can be used for the analysis of either top-down or bottom-up data sets. Currently, search algorithms for bottom-up data have evolved more rapidly and are more widely available than search algorithms for top-down data. The heart of the common search algorithms for bottom-up data is the comparison of peptide masses and/or MS/MS spectra directly against an *in silico* generated peptide list created from the DNA or protein sequences. The peptide identifications are then used to reconstruct the potential proteins that were in the original sample. This can be broken down into two main methodologies: Peptide Mass Fingerprinting (PMF), which uses the masses of the peptides to query protein databases, and MS/MS spectral searching, which uses the fragmentation patterns of peptides as well as their intact masses for searching either protein or DNA databases.

The PMF technique is most widely used on simple mixtures of proteins (1-3 proteins) derived from in-gel digest or solution-based digest of purified proteins. This is due to the confidence of protein identification by measuring a large number of peptides (4 or more) in a single MS analysis from a given protein. The MS/MS spectral database searching algorithms can be used on either in-gel digest or peptides generated from in-solution digest of entire proteomes. This is due to the fact that each peptide MS/MS spectrum is analyzed on its own merits. Of course, the more peptides that are identified from a protein, the more confidence that can be placed in its assignment. Because MS/MS spectra reveal predominantly both y-type and b-type ions (which are difficult to distinguish), the sequence of peptides is challenging to directly determine from the MS/MS spectra. Thus, all search algorithms in some fashion compare experimentally derived MS/MS spectra with *in silico* MS/MS spectra of peptides. While there are numerous search algorithms for MS/MS spectra, two commercial products, MASCOT (Perkins, 1999) and SEQUEST (Eng, 1994), are the most widely used. MASCOT utilizes a probability-based algorithm for comparing MS/MS spectra with *in silico* MS/MS spectra, while SEQUEST uses a cross-correlation algorithm for this comparison. It is important to realize that no matter what methodology or search algorithm is used, the correct filtering processes for that algorithm must be applied in order to obtain confident protein identifications from large MS-based proteome data sets. The criterion for protein identification and lack of standardization are some of the most hotly debated subjects in the field of MS-based proteomics.

9.4.2. Bottom-up MS proteomics

There are two different approaches for conducting bottom-up MS proteomic measurements. The first, more traditional method employs conventional gel electrophoresis as the first step to separate and visualize proteins. Individual spots can be excised from the gel, digested with suitable protease, and then characterized by MS. The second, more recent technique exploits the capabilities of high-resolution liquid chromatography (either in a one- or two-dimensional mode) as an on-line interface with mass spectrometry. This method involves less sample handling and has the potential to be more comprehensive and to overcome some of the limitations of the gel electrophoresis approach. Each of these will be discussed in detail below.

9.4.2.1. Gel electrophoresis with MS

Gel electrophoresis has been the gold standard for protein separations over the last twenty-five years. Within the last ten years, methodologies for coupling these powerful separation techniques with mass spectrometry for protein identification have become available (Patterson, 1995; Shevchenko 1996b; Gatlin, 1998). Furthermore, one-dimensional (1D) and two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) followed by MS or MS/MS analysis have been the most common methods for analysis of microbial proteomes over the last five years (Shevchenko, 1996a; Sazuka, 1999; Wasinger, 2000; Langen, 2000; Nouwens, 2000; Molloy, 2000; Fulda, 2000; Grunenfelder, 2001; Hernechova, 2001; Bumann, D. 2001; Hermann, 2001; Molloy, 2001; Wagner 2002). The steps in the analysis of proteins by PAGE followed by MS will be highlighted and discussed below. These methodologies have been reviewed in great depth (Jungblut, 1997; Jensen, 1998; Pandey, 2000a). Detailed protocols for the methodology are easily found on the World Wide Web at: http://www.expasy.org/ch2d/protocols, http://donatello.ucsf.edu/ingel.html, and http://proteomics.uchsc.edu/protocols/index.html.

The following discussion summarizes the key steps in a gel electrophoresis mass spectrometry experiment:

*(i) Gel electrophoresis***.** The constraints of this chapter will not allow for a detailed discussion of gel electrophoresis and gel staining for visualization. Rather, we direct the readers to a comprehensive reference on proteomics that mainly focuses on 2D-PAGE proteomics (Pennington, 2001). This reference provides detailed explanations of all aspects 2D-PAGE. We will focus on highlighting the most common PAGE methodologies used in proteomics today.

1D SDS-PAGE. This is an effective methodology for separations of intact proteins by molecular weight. This technique is commonly used as a first or second separation technique in protein analysis. In general, proteins are dissolved in ~0.1% sodium dodecylsulfate (SDS) buffer, which is a strong negatively-charged detergent. SDS acts to completely coat and denature the proteins, thus imparting the same amount of charge to similarly sized proteins. This allows a mixture of proteins to be run through a gel matrix such as polyacrylamide by applying a voltage between two electrodes. The gel acts as a molecular sieve; larger proteins are retained on the gel and move slower than smaller proteins. A set of protein standards usually are run in the first lane of the gel, so that approximate protein molecular weights can be estimated. These molecular mass measurements cannot be determined to a very high level of accuracy or precision. Proteins are typically visualized by any number of stains, and then examined by MS analysis (see below). This methodology is most commonly employed for purified, soluble proteins or protein complexes, although membrane proteins can be effectively analyzed as well. The major drawback of this technique is resolution; typically, only 50-100 proteins can adequately be resolved in a single gel.

2D-PAGE. For whole proteome separations and analysis, 2D-PAGE followed by MS analysis is the gold standard. This is driven primarily by the high resolving power of 2D-PAGE, which can distinguish over 1000 proteins in a single gel. The original methodology went through several

modifications until it was shown to be useful for the analysis of proteins from whole cells (Scheele, 1975; Klose, 1975; Iborra, 1976). In this methodology, proteins are separated in the first dimension by isoelectric focusing (IEF), which differentiates the proteins based on their isoelectric points. The IEF gel is then equilibrated in a buffer containing SDS, which is essential for proper separation in the second dimension. The IEF gel is then placed at a 90° angle with a SDS-PAGE gel, and the proteins are eluted from the IEF gel into the polyacrylamide gel, which then separates them in a second dimension by their molecular masses. As with 1D-PAGE, any number of stains can be used to visualize proteins in the gel after separation. The original methodology of 2D-PAGE had very poor reproducibility, due to a number of inherent problems in the isoelectric focusing. These problems were solved with the development of immobilized pH gradient (IPG) for the first dimension (Bjellqvist, 1982; Gorg, 2000). IPG-IEF is now the most common method for first dimension separation of proteins in 2D-GE applications in proteomics. It is now possible to purchase pre-cast gels for 2D-GE, thus increasing reproducibility and throughput. For increased dynamic range, "zoom gels" are typically used, which provide narrow pH ranges in the first dimension (Wildgruber, 2000). The sample is analyzed under a number of different zoom ranges to amplify low abundance proteins for MS detection.

The 2D-PAGE methodology can also be coupled with MS detection. This methodology is most commonly used for whole proteome analysis or comparative proteome analysis, as well as analysis of preparations from organelles, periplasms, or protein complexes.

(ii) Gel staining and visualization. The heart of comparative analysis for 2D-PAGE is gel staining, visualization, and comparison. The most important aspects of the staining process for these types of applications are sensitivity and compatibility with mass spectrometry. Currently, silver stain and Coomassie blue stain are the most common visualization reagents. Silver stain has outstanding detection limits (spots containing 1-10 ng of protein can be observed) and takes about 1-2 hours for the entire process. It also has the advantage that no destaining step is needed. Coomassie is not quite as sensitive (detection limits of 50-100 ng) and takes longer, generally 1 hour to stain and overnight to destain. Furthermore, with Coomassie stain, the gel pieces must be thoroughly washed before MS analysis, or adducts from the stain can occur on peptides. The development of colloidal Coomassie blue stain led to improved detection limits (8-10 ng), due to decreased background (Brush, 1998). Due to the complexity of the protein patterns, it is difficult for the human eye to compare two gels. Thus, 2D-PAGE gels usually are scanned by using a computer-assisted digital camera, and the picture is converted to a digital image. This allows for precise determination of spot location and intensity, and allows for computer-aided comparison of different gels. Protein spots are quantified by their visual intensities. This can then serve as a differential display of two different organism growth types or mutants. Very often in differential display analysis, only protein spots that show quantifiable change are excised, digested, and analyzed by MS, thus saving the time of analyzing every spot. Current 2D-PAGE software tools allow for integration of the entire analysis. Future directions in this area will include stains that are faster, more sensitive, simpler and fully compatible with MS analysis. For example, development of the fluorescent SYPRO stains (Steinberg, 1996a; Steinberg, 1996b; Valdes, 2000) has provided a sensitive and simple alternative to the silver and Coomassie blue stains.

(iii) In-gel digestions. While other methodologies exist for spot identification, such as electroelution followed by Edman sequencing or blotting to membranes followed by MALDI-

TOF-MS of membranes (Eckerskorn, 1997), we will focus on the most common methodology of *in-gel digestion* for MS. Initial concerns of protein analysis by MS following either Coomassie or silver stain due to permanent modification of peptides resulted in a delay of application of this technology. However, in 1996, Matthias Mann and co-workers silenced those fears when they showed convincingly that proteins separated on a 1D SDS-PAGE and stained with either Coomassie or silver stain could be digested in-gel, the resulting peptide mixture extracted from the gels, and then analyzed with mass spectrometry, using either MALDI-TOF-MS or Nano-ESI-MS/MS (Shevchenko, 1996b). Multiple standards and fractions from yeast proteomes were analyzed, totaling over 1000 peptides measured. Their results indicated that neither Coomassie nor silver staining adversely affected protein identification by MS. It is important to note that they omitted the fixation/sensitization step with glutaraldehyde in the silver staining process, which is known to covalently modify proteins. Their final conclusion was that silver stain is overall a better choice for staining and subsequent in-gel digestion with MS analysis. The silver stain had a better limit of detection (nearly 100 times lower than Coomassie), was faster overall to process, and resulted in less background. Detailed protocols for in-gel digestions can be found in the literature (Shevchenko, 1996b) and on the World Wide Wed at: (http://proteomics.uchsc.edu/protocols/index.html; http://www.bio.vu.nl/vakgroepen/mnb/proteomics/6.in-geldigestion.html; http://www.hmc.psu.edu/core/Maldi/malditofprotocols.html).

Quality sample preparation is critical for these measurements. For example, inadvertent contamination of samples with human keratins from skin and hair from normal sample handling often is the biggest barrier to successful microbial protein identification. Likewise, poor-quality reagents also are likely to present contamination problems, due to the high sensitivity of MS. It is highly desirable to use sequencing grade trypsin (which has reduced chymotrypsin activity) for high-quality digestions. In large proteomics facilities, this entire process is automated with robotics, which pick spots based on coordinates from the digital image, digest these spots in microtiter plates, extract peptides from the digested spots, and finally clean the peptide samples and deposit them on MALDI targets or into microtiter plates for nano-ES-MS/MS (see below). Membrane proteins, due to their hydrophobic characteristics, are generally harder to analyze by the typical in-gel digestion method than soluble proteins. An interesting variation on the procedure has recently been reported, which uses cyanogen bromide (CNBr) as well as trypsin, and results in 100% sequence coverage of some standard membrane proteins (van MontFort, 2002).

(iv) MS analysis and database searching. Analysis of gel-separated spots in proteomics can be accomplished by a variety of methodologies. The review listed above (Jungblut, 1997) gives a very detailed schematic of all the possible modes of protein identification, and how they are related. We have simplified this to the three most common methodologies for protein identification after in-gel digestion, as illustrated in Figure 9.4. The choice as to which of these three methodologies to use will depend on the instruments available, the throughput required, the necessary confidence of identification, and the complexity of the spot (very often "single" gel spots may contain up to 5-10 proteins). All three techniques have been demonstrated with low femtamole sensitivity, although the dynamic range for each is dramatically different. The three common techniques are highlighted below.

Peptide Mass Fingerprinting (PMF)**.** This methodology is the most commonly used technique for analyzing proteins from PAGE gels. Briefly, the extracted peptides are spotted on a MALDI target plate and mixed with a suitable matrix compound (typically, alph-cyano-4-hydroxycinnamic acid, CHCA). The target is loaded into a MALDI-TOF mass analyzer, and each spot is analyzed by multiple laser shots. This provides a composite mass spectrum of the peptides in a given spot. Typically, all peptides are singly-charged, and MALDI-TOF analyzers with reflectrons can provide 50-100 ppm mass accuracy on the peptides. The mass accuracy can be increased by using commonly occurring autotryptic peptides as internal standards. Querying a protein database with these peptide masses for *in-silico* generated tryptic peptides can identify the original protein. Generally, 4-5 peptides matching the protein with high quality mass accuracy (50-100 ppm) are considered a good match. Search algorithms of these types are very easy to create in-house or can be found on the World Wide Web: (www.matrixscience.com; http://prowl.rockefeller.edu/cgi-bin/ProFound; http://us.expasy.org/tools/peptident.html).

The major advantage of this technique is speed and simplicity. The MALDI-TOF instruments are very easy to operate and can be automated easily. A single instrument can analyze ~100 spots per day in an automated fashion. The major disadvantages of this technique are the limited accuracy in the peptide identifications and lack of dynamic range. If the gel spot contains three or more proteins or if in-gel digestion does not provide enough peptides per protein, identification becomes very difficult.

Nanospray-MS/MS**.** This methodology was demonstrated to be an effective and robust approach for analyzing peptide mixtures generated from in-gel digestion (Wilm, 1996). In this technique, the peptide mixture is introduced into a small glass needle with a small tip (10 um id). The needle is then mounted on a nanospray source and a high voltage is applied. A typical flow rate from the tip is 1-10 nl per minute, with ion detection usually accomplished by QIT, triple quadrupole, or QqTOF mass spectrometry. The mass analyzer is programmed for a data-dependent MS/MS experiment, in which the most abundant peptides are interrogated. The measurement process is conducted until the sample is depleted, which is generally about 30-60 minutes under these conditions. This experiment provides both the intact peptide mass and MS/MS spectra, which can be used for querying protein databases with search algorithms such as SEQUEST and MASCOT, or for *de-novo* sequencing. The major advantage of this technique is the ability to identify a protein from only a few peptides. Furthermore, this method is readily applicable to simple mixtures of 5-10 proteins. The major disadvantages of this technique are low throughput and difficulty of operation. In general, nanospray MS is more difficult to master and automate than MALDI-TOF-MS. This methodology has intermediate dynamic range when compared with the other two methods.

NanoLC-Nanospray-MS/MS. If a protein spot contains more than 5-10 proteins, or if large sequence coverage of each protein is desired (useful for identifying PTMs), then the first two methodologies are insufficient. For the best dynamic range measurement, NanoLC-Nanospray-MS/MS is the methodology of choice. While several variations exist for this technique, fritless nanospray columns provide the best sensitivity (Gatlin, 1998). In this technique, C18 particles are packed directly with a packing bomb into laser-pulled fused silica needles (generally 100-75 um id and 10-15cm in length, with 5-10 um id tips). The tip of the needle acts as the frit for the packing material. The sample can then be loaded off-line with a loading bomb, or on-line via an injection loop. The integrated column/tip is placed in a nanospray source behind a microtee or a

microcross. A high voltage is applied at this cross through a gold wire or rod. The effluent from a high-performance liquid chromatography (HPLC) system is then directed into the back of the cross from a transfer line. If the HPLC can provide low enough flow rates, then a split at the cross is not required, and a microtee is employed; otherwise the flow must be split into a waste line (microcross). The flow through the nanocolumn is usually 100-200 nl/min. This integrated column and nanospray tip is generally incorporated into a QIT, triple quadrupole, or Q-TOF mass spectrometry system. The mass analyzer is again set-up for a data-dependent MS/MS experiment. The HPLC will employ a gradient elution of 95% water to 95% acetonitrile over 30 minutes to 1 hour for the MS analysis. This experiment provides both the intact peptide masses and MS/MS spectra, which can be used for querying protein databases with search algorithms such as SEQUEST and MASCOT, or for *de-novo* sequencing. The main advantage of this technique is enhanced dynamic range, due to the increased physical separation of peptides. Multiple proteins can be easily identified from a given spot with greater sequence coverage than the other two techniques. Again, the major disadvantage when compared with peptide mass fingerprinting by MALDI-TOF is low throughput and more complex operation (which is more difficult to automate).

Future of 2D-PAGE for proteome analysis

As stated earlier, 2D-PAGE followed by MS analysis has been established as the gold standard for proteome analysis, especially for microbial species. But within the last few years, there has been a noticeable migration away from this methodology toward pure liquid-based approaches, which will be described in the final section of this chapter. This movement is mainly due to the inherent weaknesses in the 2D-PAGE methodology. The advantages and disadvantages of 2D-PAGE as they relate to the liquid-based methods are highlighted below. Some of these inherent weaknesses are currently being addressed with new techniques; however, the fundamental fact remains that at the current time, the depth of analysis of whole proteomes by 2D-PAGE does not compare well with the emerging liquid-based methodologies. 2D-PAGE is still routinely used in many labs around the world and undoubtedly will continue.

Advantages of 2D-PAGE for proteome analysis:

i) Gold standard, widely used and understood
ii) Very high resolving power
iii) Sensitive staining methodologies available
iv) Commercial software is available for automated gel processing and quantitation.

Disadvantages of 2D-PAGE for proteome analysis:

i) Poor reproducibility
ii) Limited recovery of low abundance proteins
iii) Limited pI and MW ranges
iv) Time consuming
v) Membrane proteins do not enter the second dimension effectively
vi) Coupling with MS is an indirect process
vii) Intact protein analysis is very difficult.

9.4.2.2. Liquid chromatography with MS (gel-less method)

The coupling of liquid chromatography with mass spectrometry (LC-MS) is one of the most promising approaches to overcome some of the limitations of 2D-PAGE discussed above. The advent of ESI has provided a natural way to interface liquid chromatography directly to MS, since ESI involves dynamic introduction of a flowing liquid stream directly into a mass spectrometer. It is reasonable to propose to connect a liquid chromatography system to the electrospray source, so that the benefits of liquid-based separation can be combined with high-resolution molecular mass (and MS/MS) measurements. While some work has been done on chromatography of intact proteins in conjunction with mass spectrometry (see below), the majority of effort has focused on chromatography of enzymatically-generated peptides in conjunction with mass spectrometry, the so-called *bottom-up* or *shotgun* method for proteomics. This is primarily due to the fact that peptides are much easier to handle, separate, and analyze than intact proteins. It is somewhat counterintuitive that it is desirable to take a complex protein mixture and make it more complex by digesting the proteins into representative peptides. For example, each averaged size protein can generate ~20 peptide fragments. So, proteolytic digestion of a sample containing 1000 proteins will generate a new sample that contains ~20,000 peptides. While this appears to be a poor choice, in practice liquid chromatography and mass spectrometry of peptides are currently well developed and robust, even for very complex peptide mixtures. The advantages of coupling liquid chromatography with mass spectrometry are obvious when one considers some of the necessities of proteomics, as highlighted below.

(i) *Dynamic range***.** The need for multiple dimensions of separation has become most apparent in the use of LC-MS for proteome analysis, due to the large dynamic range necessary for measuring a whole proteome. While 2D-PAGE offers a very high resolving power for proteins, this methodology is currently limited with respect to the types of proteins it can analyze and the number of quality identifications that can be made from any one gel. The coupling of multiple dimensions of chromatography with mass spectrometry offers a solution to this problem. This can be easily noted from the fact that while ~1000 proteins have been visualized on a 2D-PAGE gel, there have been no published reports of more than a few hundred proteins being identified from a single gel. It has now become routine for the accurate identification of 500-1000 proteins from a single sample in 20-30 hours on a LC-MS/MS system operated in automated mode.

(ii) *Sensitivity.* While 2D-PAGE gels have very sensitive staining methodologies for *observing* spots, the ability to *identify* proteins from these spots by mass spectrometry has not matched this level. This is primarily due to the large sample losses in the in-gel digestion step. Pure LC-MS methodologies promise to be more sensitive, due to the reduced overall sample handling while keeping the sample in the liquid phase.

(iii) *Quantitation.* Differential analysis between two or more sample types is a primary need for successful proteome applications. One of the reasons that 2D-PAGE has remained the gold standard is that it currently is inherently better at quantitation than liquid-based methodologies, due to the fact that LC-MS suffers from matrix effects that make comparison of run-to-run peak intensities very difficult. This is primarily due to the electrospray ionization methodology and not the mass spectrometry. Furthermore, the apparent accuracy of 2D-PAGE for quantitation became questionable when it was realized that many spots on any 2D-PAGE gels contained more than one protein. The use

of stable isotopes for peptide labeling (see quantitation discussion below) has proven that quantitation can be accomplished by LC-MS.

(iv) *Protein Diversity.* A major advantage of liquid-based methodologies in comparison with 2D-PAGE is the diversity of proteins that can be analyzed. Virtually any protein that can be subjected to either chemical or enzymatic digestion can be analyzed. This includes membrane proteins, proteins of high and low pI values, and proteins of high and low molecular mass.

(v) *Throughput*. In the field of proteomics, one of the biggest concerns is sample throughput, including not only how fast samples are analyzed, but how well they be characterized in a short period of time. This is just as important in bacterial proteomics as mammalian or plant proteomics. Currently, at least forty different microbes have been fully sequenced at the Joint Genome Institute (www.jgi.doe.gov) and many more have been completed worldwide. Researchers need to be able to analyze the proteomes from these organisms under many different growth conditions, and with many different mutants for a systems biology approach to be truly effective. For sample throughput, LC-MS has already been well developed in the pharmaceutical industry, where thousands of samples are processed by large numbers of mass spectrometers in hundreds of laboratories every year.

The trend towards liquid chromatography methods for proteome analysis can be made clear by examination of Table 9.3. Over the last two years, all of the large-scale proteome analyses have been accomplished with some form of LC-MS methodology. The analysis of *Oryza sativa* provides the best example for a direct comparison of 2D-PAGE analysis and LC/LC-MS/MS analysis for whole proteomes (Koller, 2002). In this study, the rice plant was broken into three fractions - leaves, roots, and seed tissues. Proteins were isolated from each fraction and analyzed by 2D-PAGE, followed by automated Nano-LC/MS/MS or by multidimensional protein identification technology, or "MudPIT" (Washburn, 2001). The analysis of all three fractions by 2D-PAGE-MS resulted in 556 non-redundant identifications, while the analysis of all three fractions by MudPIT resulted in 2,363 non-redundant identifications. There was no mention from the authors about the length of time each analysis consumed, but we estimate the MudPIT analysis could have been accomplished in 5 days on a single mass spectrometer once the system was optimized. This throughput, as well as the enhanced dynamic range, is why the liquid-based methodologies are having the greatest impact on proteome analysis. Because of the current trend towards LC-MS for proteome analysis, we will direct the next section to detail some of the protocols used in LC-MS and some of the cutting edge techniques such as MudPIT now being routinely used in proteome analysis. For further reading on the advancement of LC-MS in proteomics, see these excellent reviews (Peng, 2001; Mann, 2001; Liu, 2002).

LC-MS methodologies. For continuity, we will first outline the entire process of LC-MS/MS for shotgun proteomics in Figures 9.5 and 9.6, as it could apply to microbial proteomes. Sample preparation, different versions of LC-MS procedures, and quantitation will then be explained in detail below. The shotgun proteomics technique begins with enzymatic digestion of a microbial proteome sample and analysis of the resulting peptide mixture by automated LC-MS/MS or LC/LC-MS/MS in a data dependent manner (Figure 9.5 details this process). Bear in mind that the sample that is injected onto the chromatographic system consists of a mixture of several thousands of distinct peptides. These peptides are separated physically over a period of time by their hydrophobicity or net charge, and are sequentially injected into the mass spectrometer and

ionized by ESI. By using the mass spectrometer to record the overall ion intensity as a function of time, it is possible to obtain a Total Ion Chromatogram (TIC) much like a UV chromatogram (top left Figure 9.5). During the entire chromatographic run, the mass spectrometer is oscillating between *full scan mode*, where it is acquiring m/z values of peptides entering the mass spectrometer at that time point (top right, Figure 9.5), and subsequent *MS/MS mode*, which examines the fragmentation of the most abundant peptides (generally 3-5) as they elute from the column (bottom left, Figure 9.5). This latter mode is accomplished by gas-phase isolation of individual peptides, followed by collisional dissociation. The mass spectrometer records the fragment ions and the mass of the precursor ion. To increase dynamic range, most methodologies employ some type of dynamic exclusion so that peptides that have already been fragmented are not fragmented again. The precursor masses and the fragmentation patterns are then submitted to search algorithms such as MASCOT and SEQUEST, which can query thousands of MS/MS spectra against protein or nucleotide databases (bottom right, Figure 9.5). The final stage of the process is illustrated in Figure 9.6. Typically, a single LC-MS/MS experiment produces tens of thousands of MS/MS spectra. These spectra must be filtered and sorted in order to extract useful information from them. Filtering and sorting software such as *DTASelect* (Tabb, 2002) are used to extract and sort positive identifications, whereas the program *Contrast* (Tabb, 2002) is used to compare run-to-run variations and sample-to-sample changes. The protein identifications can then be compiled into KEGG maps and functional categories for rapid viewing of metabolic and signaling pathways that are activated. This information allows targets to be designed for mutations, gene knockouts, and protein-protein interaction assays.

Methods for sample analysis

*Sample preparation, digestion and cleanup protocols***.** In general, the sample preparation for LC-MS is fairly simple and robust, but extremely important for the final outcome. If the protein sample is sufficiently concentrated (i.e., 1-5 mg/ml, which is a typical concentration of cell lysate from a microbe), it can be digested immediately. If the sample is dilute (< 0.5-1.0 mg/ml), or contains interfering substances such as chlorophyll, lipids, or pigments, then acetone or chloroform/methanol precipitation may need to be employed. The sample is then denatured with 6-8 M Guanidine or Urea, and reduced with DTT or some other reducing agent at 60°C for 10-60 minutes. Iodoacetamide is commonly added at 0.1 M to alkylate cysteine residues. This step is done at room temperature in the dark for 30-60 minutes. Care must be taken not to over alkylate the peptides, or non-specific alkylation can occur. Some labs skip this step altogether, opting for a second harsh reduction after final digestion and directly before sample clean up in a low pH solution. The denaturant concentration is lowered by dilution with Tris or bicarbonate buffer (with 1-10 mM $CaCl_2$) and sequencing grade protease is added to the sample. The most common protease used is trypsin due to specificity, cost, and the fact that it produces a positively charged residue on the C-terminus (lysine or arginine), which aids ionization and peptide sequencing by MS/MS. The sample can be digested overnight, and fresh protease can be added in the morning to increase digestion. Some common variations of the digestion technique include the use of LysC protease along with trypsin (Link, 1999). LysC is active in up to 8 M urea and cuts at the C-terminal side of lysine residues (similar to trypsin) but not at arginine residues. Thus, the enzyme can be used under denaturing conditions, which increases overall sequence coverage when followed with the common trypsin digest. CNBr is another common chemical protease,

which can be used for membrane digestions since most hydrophobic domains lack adequate numbers of lysine or arginine residues for trypsin. Furthermore, the CNBr digestion is performed in 70% formic acid, which readily dissolves the membranes. This methodology has been shown to be effective for analysis of whole yeast membranes (Washburn, 2001). The main disadvantage is the toxicity of CNBr; all reactions and sample handling should be performed in a fume hood. Another method for digestion of membrane samples is a mixed aqueous-organic trypsin digestion. This methodology was demonstrated for the bacterium *Deinococcus radiodurans* (Blonder, 2002). Basically, the trypsin digestion is performed in a mixture of buffer and organic solvent such as methanol or acetonitrile. The organic phase allows for solubilization of membrane proteins. After final digestion, the sample is desalted by solid phase extraction off-line, or can be directly analyzed on-line with desalting utilizing either a trap cartridge or the analytical column itself. If sample quantity is not a concern, off-line desalting by solid phase extraction is the most robust method for sample clean-up.

Types of LC-MS for bottom-up proteomics

*(i) On-line 1D LC-MS/MS***.** The most simple bottom-up method for analyzing whole proteomes or protein complexes by LC-MS is the connection of a single reverse-phase column directly to an electrospray mass spectrometer. In most cases, a high performance liquid chromatography (HPLC) system is used to deliver a programmable gradient of high water to high organic over a C18, C8, or C4 reverse-phase resin over a given length of time (usually 1-3 hours). The solvents generally contain an organic acid such as formic or acetic acid (0.01-0.1%) to aid in ionization, but the standard HPLC ion-pairing reagent trifluoroacetic acid (TFA) should be avoided due to problems associated with electrospray (unless it is used at lower than 0.01%). The columns are typically 500 μm – 75 μm i.d., with lengths of 25 cm-10 cm, and utilize flow rates of 10 μL – 0.100 μL per minute. The type of electrospray source used must match the flow rate of the column. For the higher end flow rates, pneumatically-assisted electrospray sources are used. For the lower end flow rates, nanospray sources must be employed. The primary advantage of a lower flow rate system is an increase in sensitivity and resolution, while the main advantages of the higher flow rate system are robustness and ease of sample loading. The choice of column size, flow rate, and methodologies applied will depend on the application and amount of starting material available. The common mass spectrometers used for LC-MS applications in proteomics are QIT, triple quadrupoles, and Q-TOF. Recently, on-line liquid chromatography in conjunction with high resolving FTICR-MS instruments has shown great promise and will be discussed separately below. A typical instrument set-up for high-throughput analysis of protein mixtures would include an autosampler to inject peptide sample onto the column, a high performance pump to deliver the programmed gradient, and a reverse-phase column, which can either be directly connected to an electrospray source through a low dead volume connection or the tip of the column can be directly pulled into a nanospray tip (this eliminates almost all dead volume) (Lee, 2002; Gatlin, 1998). The UV flow cell often used for detection in HPLC applications is generally omitted in MS-proteomics applications due to the large dead volume associated with its use and the lack of definitive information it provides. The mass spectrometer in most cases is operated in a data-dependent MS/MS mode, with computer control over the autosampler, HPLC and the mass spectrometer. The samples can be loaded onto the nanocolumns either off-line with a pressure bomb or on-line with an autosampler. This avoids sample losses that may occur with an autosampler/sample loop, but will not be as high-

throughput as on-line loading of sample. One of the biggest challenges in LC-MS applications for proteomics is rapidly loading a dilute sample onto a narrow bore column (<100 μm i.d.), due to the low operational flow rates (< 250 nl/min). While there has been much work in this area, two methods have shown promise. The first involves loading the sample onto a small trap cartridge at a very high flow rate (van der Heeft, 1998; Devreese, 2001). This trap cartridge typically is very wide and short, to be able to handle the high flow rate. The sample can be rapidly concentrated onto the cartridge and de-salted. The trap is then back flushed on the resolving column and a reverse-phase gradient is run over both columns into the electrospray mass spectrometer. The advantages of this system include rapid sample loading and desalting, but this system can suffer from sample loss, as well as reduced resolution and sensitivity. The second method employs a vented pre-column of the same diameter as the resolving column (Licklider, 2002). This methodology may offer the same advantages as the previous method, without the loss of resolution and sensitivity.

Whole proteome analysis by a single dimension of chromatography can be difficult due to the limited dynamic range of the system (this has led to the use of multi-dimensional systems, as discussed below). This is primarily due to the slow scanning speeds of conventional mass spectrometers operating in data dependent mode and ionization suppression effects of some peptides over other peptides. At any given point in time in an LC-MS analysis of a whole bacterial proteome, hundreds of peptides can be simultaneously entering the mass spectrometer. Even the fastest mass spectrometers cannot isolate and fragment so many peptides in a short period of time. The use of multiple mass range scanning, or gas phase fractionation, for the analysis of whole bacterial proteomes has shown promise as a simple, cheap and effective method for analyzing complex protein mixtures (Spahr, 2001; Davis, 2001; VerBerkmoes, 2002; Lipton, 2002). This method relies on injecting the same proteome sample onto an LC-MS system repeatedly, in which the mass spectrometer is programmed to scan a different narrow mass/charge range for each injection, thus allowing for more MS/MS spectra to be acquired from the sample. This methodology is not as sensitive as some of the methods described below, since the sample must be repeatedly analyzed (typically 4-8 m/z ranges are examined, thus requiring 4-8 injections to be made per sample), but this method is very simple to implement and is faster than the 2D methods described below. Multiple mass/charge range scanning can now be used for analysis of whole bacterial proteomes with identifications of 500-1000 proteins per day, depending on criteria used for the identification, the sample type and the type of MS system employed. Furthermore, in microbial proteome analysis, sample quantity is usually not much of an issue, since 100-200 mg of total protein can easily be obtained from 4-5 g of cell paste.

(ii) Multistage LC-MS/MS. Due to the limited dynamic range offered by a single dimension of chromatography, methodologies are now being employed for coupling two or more dimensions of chromatography, either off-line or on-line, in conjunction with ES-MS. While methodologies employing multidimensional chromatography have been commonly used in the separations field for many years, it was only in the last few years that these methodologies have been used in conjunction with mass spectrometry for whole proteome analysis. Recently, many different methods have been attempted to analyze whole proteomes in a relatively short period of time. One of the first methodologies described and probably the most widely cited is Multidimensional Protein Identification Technology or MudPIT (Link, 1999; Washburn, 2001; Wolter, 2001). This technique employs a biphasic microcolumn integrated with the nanospray tip directly placed

in front of an ion trap mass spectrometer. The first part of the column is packed with strong cation exchange (SCX) particles and the second part of the column is packed with reverse-phase (RP) C18 particles. A peptide mixture derived from LysC/trypsin digestion of a whole proteome is directly loaded onto the back of the biphasic column with a pressure bomb. The column is placed in front of the mass spectrometer and an automated 2-dimensional gradient is started by the HPLC. This gradient consists of an initial injection of low concentration ammonium acetate by the HPLC onto the system, which moves a plug of peptides from the SCX material to the RP material. Peptides are then eluted from the RP material into the mass spectrometer with a typical, high resolution RP gradient procedure. The system is then re-equilibrated. This routine is repeated with increasing salt-steps of ammonium acetate until all peptides have been run through the system (typically 10-15 cycles or salt steps). The entire system is run in an automated mode, with computer control over the HPLC and the mass spectrometer. The methodology was shown to be capable of analyzing a total of 1,484 proteins from three crude fractions from *Saccharomyces cerevisiae* (Washburn, 2001) in ~4-5 days with a single mass spectrometer. The analysis revealed that the system was not biased toward any given group or type of protein. A variation of this methodology was recently reported and compared with the MudPIT technique (Peng, 2002). In this technique, the SCX column was not connected on-line with the RP column. Instead, peptides were loaded onto the SCX column, and eluted with a linear gradient of increasing salt. Fractions were collected, concentrated, and then loaded onto the on-line vented column described above via an autosampler. This technique was again used to analyze the *Saccharomyces cerevisiae* proteome, resulting in the identification of 1,504 proteins. It is difficult to compare the two techniques directly head-to-head, because of the variation in protein identification methodology, the amount and type of starting material, and the total length of time for analysis. The MudPIT technique appears to have the advantage of being more sensitive, since sample loss should be minimized in an on-line system; however, the off-line method has the advantage of allowing the use of a much larger first SCX dimension and a more true linear gradient instead of a step gradient. Variations on the 2-dimensional methodology include the use of strong anion exchange instead of strong cation exchange in the first dimension (Mawuenyega, 2002), as well as separation of intact proteins in the first dimension by strong anion exchange followed by tryptic digestion of the fractions and 1D-LC-MS/MS of the resulting peptide mixtures (VerBerkmoes, 2002).

(iii) LC FTICR-MS. Within the last few years, a very promising new MS methodology for the analysis of whole proteomes has emerged. This technology involves the use of high performance liquid chromatography in conjunction with electrospray ionization on high performance FTICR mass spectrometers (Martin, 2000; Quenzer, 2001; Shen, 2001). The FTICR instruments provide better dynamic range, sensitivity and mass accuracy over other more conventional mass spectrometers used in proteome research. These advantages, coupled with the high resolving power and sensitivity of nano HPLC columns, may provide access to low copy number proteins in cells (1-5 copies per cell). The current disadvantages of this technology include the high price for the FTICR mass spectrometers, the expertise required to operate the instruments, and the difficulty of employing data-dependent MS/MS methods on the system. An intriguing experimental approach termed *accurate mass tags* (AMT) has been developed by Richard D. Smith and colleagues by using very high-pressure liquid chromatography (Tolley, 2001) in conjunction with an 11.4-Tesla FTICR mass spectrometer (Smith, 2002a; Smith, 2002b; Lipton, 2002). This methodology promises to alleviate the need for routine MS/MS of peptides. In this technique, peptides are separated by very high-pressure liquid chromatography (5000-

10,000 psi) on very long, narrow columns (85 cm length, 150 μm i.d.) packed with C18 particles. The peptide mixtures are first analyzed by LC-MS on QIT-MS to obtain MS/MS data and retention times. Peptides that pass the minimum criteria (Xcorr >2.0) are labeled as *potential mass tags*. The same peptide mixtures are then analyzed on the high-magnetic field FTICR-MS by the same chromatography system, except in this case the FTICR-MS is only acquiring accurate mass/charge values rather than MS/MS. By internal calibration of each mass spectrum with known peptides, mass accuracies to within 1 ppm can be obtained. Once a peptide has been identified on both systems within the criteria limits, it is labeled as an *accurate mass tag,* and its elution time and mass is stored in a database. Thus, when the sample is analyzed again, the peptide mixtures can be screened only by the FTICR-MS system, without the need for time-consuming MS/MS acquisitions. The authors contend that the AMT databases can be used for rapidly screening many different growth states of a given organism, without the need for routine MS/MS. This methodology was recently used for the largest analysis of a proteome to date, where >61 % of the proteome of ionizing radiation-resistant bacterium *Deinocoocus radiodurans* was identified with the AMT approach by analyzing the microbe grown under ten different growth conditions.

Quantitation

The analysis of whole proteomes and protein complexes by mass spectrometry can provide very useful *qualitative* information, but one of the most interesting areas of proteomics is the *quantitative* comparison between different growth conditions or mutants for a given organism. For example, quantitative analysis of microbial proteomes has been dominated by 2D-PAGE followed by MS analysis, but this is shifting towards pure liquid-based methods due to the deficiencies in the 2D-PAGE methodology described above. This is primarily due to the fact that the pure liquid-based methods are inherently higher throughput, and are not biased against any protein type. There are difficulties with the liquid-based methods for quantitation, and technologies to address these are only now being developed and implemented. Due to matrix effects associated with both ESI and MALDI, direct comparisons of ion peak heights or area for given peptides or proteins eluting from LC columns into the MS should only be used as approximations for abundance levels, and are most likely not accurate for absolute quantitation. Recent developments in stable isotope labeling has allowed for accurate, relative quantitation of proteins in two different samples, such as *E. coli* grown under high salt and low salt conditions. In these experiments, a given protein(s) can be compared with its counterpart from a different growth condition to obtain a relative expression level of up- or down-regulation, but an absolute level of protein expression is still very difficult to determine. Three main methodologies that employ stable isotopes for relative quantitation have developed over the last four years. Each has advantages and disadvantages, which are highlighted below:

(i) Isotope Coded Affinity Tags (ICAT). ICAT was originally developed in 1999 (Gygi, 1999), and has become commercially available through Applied Biosystems. The methodology has since been applied to the analysis of protein expression and comparison with microarray data in *Saccharomyces cerevisiae* (Ideker, 2001; Griffin, 2002), as well as the analysis of human cell line HL-60 microsomal proteins (Han, 2001). This technique entails the use of an isotope encoded affinity tag. The proteins in a sample are mixed with the ICAT reagent, which specifically reacts with cysteine residues. The reagent has an isotopic label (either a “light”

version containing either hydrogen atoms on the aliphatic chain or a "heavy" version containing eight deuterium atoms in the same location) and a biotin affinity tag for isolation of cysteine-containing peptides. Thus, a tagged peptide can contain either the light or heavy version of the tag. This technique is applied to protein samples by labeling one sample with the light reagent and the other sample with the heavy reagent. The samples are then combined, digested with trypsin, and passed over an avidin column to enrich the cysteine-containing peptides. LC-MS methodologies described above can then be used to analyze the complex peptide samples to obtain peptide identifications as well as quantitative information by comparing the peak heights of heavy and light versions of the same peptides. Accurate protein quantification should be obtained by averaging multiple labeled peptides from the same protein. The main advantage of this technique stems from the fact that the labeling is done directly after the cell lysates have been prepared, so any changes involving sample handling affect both samples equally. Also, the interrogation of only cysteine-containing peptides makes the peptide mixtures much simpler and easier to analyze. Finally, ICAT peptides can be easily sequenced by MS/MS and identified with MASCOT or SEQUEST. The software to analyze the large data sets is under development by the inventors of the technique, as well as many MS instrument companies. The major disadvantages of this technique are the commercial price of the reagent and the fact that the current version of the commercial reagent only labels cysteine residues. This latter point presents a serious problem for some bacterial species. The average number of cysteines per protein is much lower in bacterial species, as compared with some common eukaryotic species. Furthermore, a large percentage (50-60%) of proteins in bacterial species contain either 0, 1, or 2 cysteine residues, as shown in Table 9.4, which either prevents quantitation, or requires that the quantitation be based on one- or two data points.

*(ii) ^{18}O Water labeling***.** This methodology was recently introduced as an alternative to ICAT for accurate protein quantification (Yao, 2001). In this technique, one protein sample is digested with trypsin in the presence of ultrapure ^{18}O water, while the other sample is digested in normal water. The samples are then pooled and analyzed by LC-MS/MS or MALDI methodologies on high-resolution mass spectrometers. The results clearly demonstrated that the carboxy termini of the tryptic fragments digested in ^{18}O water are fully labeled with ^{18}O, and this label is stable. Thus, all tryptic peptides from the $H_2{}^{18}O$ sample have an increase in mass of 4 daltons (two incorporated oxygens on the C-terminus of each peptide). The peptides can then be quantitated by comparing peak areas of co-eluting peptides separated by 4 daltons. This technique was also demonstrated to work with another common protease endoprotease Glu-C (Reynolds, 2002), thus adding to its versatility. The main advantage of this technique is that it can work on any protein that can successfully be digested with either trypsin (K or R residues) or Glu-C (E or D residues). Furthermore, the technique has been shown to work successfully on phosphorylated, glycosylated, and disulfide-containing proteins (Reynolds, 2002). Elution times were also shown to be consistent with the labeled and unlabeled peptides (Reynolds, 2002). The main disadvantage of this technique is the price and availability of $H_2{}^{18}O$, as well as the need for a high-resolution mass spectrometer to analyze the peptides with such small mass differences.

*(iii) Nitrogen Labeling***.** In this methodology, the microbe of interest is grown under a defined media with either normal media (containing naturally-occurring isotopic abundances) or isotopically enriched or depleted media (Oda, 1999; Pasa-Tolic, 1999). The most common method is to grow the microbe in defined media without amino acids with only ammonium

sulfate as a nitrogen source. The ammonium sulfate can then be either ammonium-^{15}N sulfate or normal ammonium sulfate. The microbe will incorporate the stable heavy isotope into its proteins. The normal and heavy-labeled samples can then be grown under the desired conditions, combined, lysed and digested with a protease. The peptides will have heavy and light pairs that should elute at the same time in a LC-MS/MS experiment, and again the peptides can be quantified by comparing peak areas. This methodology has recently been employed for the largest quantitative proteome analysis to date of the yeast proteome by nitrogen labeling, followed by MudPIT analysis (Washburn, 2002). The major advantages of this technology are the low cost and the ability to quantitate any type of protein that can be digested with either chemical or enzymatic methods. The samples are mixed immediately after growth so that any changes in sample preparation affect both samples in the same manner. The major disadvantage is that this technique can only be used for species whose growth conditions can be exquisitely controlled.

9.4.3. Top-down MS proteomics

Although proteolytic digestion and MS characterization (i.e., bottom-up proteomics) have become very powerful over the past five years, it is clear that this is an *indirect* protein identification technique, as the *intact* protein species are never measured directly, but rather only a fraction of the proteolytic peptides for any given protein are examined. This leads to some concern that subtle aspects of the protein, such as post-translational modifications, might be missed by the bottom-up approach. This has prompted investigation into developing MS technology for the measurement of the intact forms of the proteins in complex mixtures. While this may seem straight-forward based on the extensive past work on characterizing purified protein samples, in fact, this approach turns out to be a formidable analytical challenge for proteomes due to at least three factors. First, the protein molecules masses can range from 5 – 200 kDa, requiring high performance MS technology for accurate measurements. Second, the extreme heterogeneity of protein sequences gives rise to a substantial ionization suppression effect when complex mixtures are examined. Thus, the proteins with the largest amount of surface charge will ionize most easily and will be over-represented in the mass spectrum relative to their abundance in the sample. Likewise, the proteins that are not easily ionized will be under-represented. This factor suggests that some type of pre-fractionation, or on-line chromatography, will most likely need to be used for intact protein measurements. Third, the unambiguous identification of larger proteins is difficult, due to the isotopic packet that confounds accurate mass measurements and the inability to extensively fragment these proteins under tandem mass spectrometry conditions to get complete sequence information. All three of these factors are much easier for peptides, because of their lower molecular masses and more extensive fragmentation. However, research underway in several laboratories has shown remarkable progress in overcoming these challenges for the top-down approach. One particular factor that must be noted is that most of the developments of the top-down approach have focused on the experimental MS measurement technology. As a result, the bioinformatics component is much less developed for the top-down data analysis.

This section will be broken down into three sub-topics: sample preparation, molecular mass measurement, and structural interrogation. As will be obvious from the following discussion, these factors are somewhat different for the measurement of *intact* proteins relative to proteolytic peptides.

(i) Sample preparation. One of the challenges in isolating complex protein mixtures is keeping the proteins intact and soluble during the fractionation process. Because MS measurements do not require the proteins to be in their active forms, it is often desirable to denature the entire complex mixture as early in the clean-up process as possible. While this usually inactivates cellular proteases, it often causes undesirable protein precipitation in the samples. For the bottom-up MS approach, it is advantageous to denature and digest the complex protein samples as early as possible in the clean-up process. Because only peptides are measured, protein stability is not an issue for this method. In contrast, protein stability is critical for the top-down MS approach. To enhance this, during the cellular lysing process, a protease inhibitor cocktail is often added to arrest protein degradation. The protease inhibitors, which are often small molecules, stabilize the protein samples, but need to be removed prior to MS characterization. This is often achieved by dialysis (conducted under cold conditions to minimize enzymatic activity) or reverse-phase clean-up procedures. The samples are usually maintained at -80 ^{o}C until analysis.

(ii) Molecular mass measurement. The critical component for top-down proteomics by MS is measurement of the molecular masses of the intact proteins. The four important experimental aspects of this measurement are mass accuracy, mass resolution, dynamic range, and detection sensitivity. Because of the wide molecular mass range of possible proteins, some researchers have proposed the application of MALDI-TOF-MS technology. While this is probably the best approach for very large species, the mass resolution and accuracy are very limited for this technique. For example, a protein with a molecular mass of 60 kDa can only be measured under the best TOF-MS conditions to about 0.02% (~ 12 Da). While this mass measurement is far superior to what is obtainable from gel electrophoresis, this value could still correspond to many proteins within a given database. A much higher level of mass accuracy would limit the number of possible proteins. This is the driving force to employ techniques such as ESI-FTICR-MS for intact protein measurements. While there are still many experimental parameters to be worked out to make this a routine tool for intact proteome characterization, this technology does provide unprecedented capabilities for high performance measurements. For example, the same protein with a molecular mass of 60 kDa could be measured with the FTICR-MS technique to about 0.0005% or 5 parts-per-million (~0.3 Da). This level of accuracy has been demonstrated recently by ESI-FTICR-MS on the intact form of GroEL from *E. coli*, in which the *measured* most abundant isotopic mass of 57,196.431 Da compares very favorably with the analogous *calculated* value of 57,196.734 Da (an error of 5 ppm) (Hettich, unpublished results, 2002). The most straight-forward top-down MS approach would attempt to identify the protein from only its molecular mass. For example, searching the SWISS-PROT protein database (at http://us.expasy.org/) with a molecular mass of 57,196 $\pm$ 0.3 Da for *E. coli* yields only one possible match – the GroEL chaperonin protein. In fact, for this particular bacterium, this protein could be identified from the TOF-MS data as well, since this is the only protein within 57196 $\pm$ 12 Da. For eukaryotic systems, however, the situation becomes more complicated. By searching the SWISS-PROT database for humans, there are five possible proteins that would fall within the mass window of the TOF-MS measurement (57196 $\pm$ 12 Da). The higher resolution mass measurement with the FTICR-MS technique would be sufficient to identify a unique protein from this batch and, in fact, would be adequate to resolve the cochlin precursor protein (COCH_HUMAN) at 57193.15 Da from the major capsid protein L1 (VL1_HPV2A) at

57193.61 Da. *Thus, high resolution and accurate mass measurements of intact proteins are often sufficient information to identify many bacterial proteins, without further structural information.* This statement should be tempered with the knowledge that post-translational modifications would alter the measured molecular masses and would make it difficult to correlate the measured protein mass with the value predicted from the genome data. For this reason, it is best to integrate the measured molecular mass information with either structural data obtained by tandem mass spectrometry or with data obtained by the bottom-up method on the same organism (VerBerkmoes, 2002). This latter approach will be discussed in section 9.4.4.

High-resolution molecular mass measurements of intact proteins reveal the complex isotopic packet resulting from the combination of naturally-occurring isotopes. This necessitates comparing the measured and calculated isotopic distributions to verify protein identification (Blank, 2002). In practicality, the high-resolution molecular mass measurement is used to query a protein database for a given organism. The possible protein matches falling within the specified mass accuracy window are tabulated, and a calculated isotopic distribution is determined for each one (for FTICR-MS measurements, there are usually no more than 3-4 possible proteins within the 5-10 ppm range of the measured mass). For each putative protein, the calculated isotopic distribution and most abundant peaks are compared to the measured values for final protein determination.

Even with the high-resolution molecular mass measurements discussed above, the dynamic range and heterogeneity of intact proteins in these complex mixtures confound the MS measurements. The basic problem stems from the limited ability to simultaneously measure hundreds (or even thousands) of proteins in a single mixture. An obvious solution to this dilemma is to incorporate some aspect of protein fractionation, either off-line or on-line, with the MS measurement. Although this increases the sample handling and thus possible contamination or sample losses, the MS measurement requirements are greatly relaxed. For example, off-line anion-exchange chromatography can be used to fractionate complex protein mixtures from crude cell lysates. Each fraction, which contains between 10-100 proteins, is more easily interrogated by mass spectrometry. Figure 9.7 illustrates the ESI-FTICR mass spectrum of an anion exchange fraction from yeast. Note the appearance of at least one-half dozen protein species. Each of the molecular ion regions can be expanded to illustrate the high-resolution mass measurements possible with this technique. Note that both phosphoglycerate mutase and phosphoglycerate kinase were identified at high accuracy (i.e., both less than 3 ppm error). By examining the bottom-up MS data from this fraction, it was confirmed that the molecular ion form of the mutase was missing the N-terminal methionine, whereas the molecular form of the kinase was not only missing the N-terminal methionine but was also acetylated as well.

The most common protein fractionation approach has been to incorporate reverse-phase liquid chromatography on-line with the MS. This arrangement permits the proteins to be physically separated by their hydrophobicity on the stationary phase of the column, and then eluted sequentially directly into the mass spectrometer. Although the entire measurement takes longer (usually about 1 hour for the LC-MS experiment), a much more extensive analysis of the complex protein mixture is possible. This approach has been demonstrated for the characterization of the chloroplast grana proteome (Gomez, 2002), and the yeast large ribosomal subunit (Lee, 2002), and resulted in not only protein identifications but also detection of post-translational modified species. It is feasible to employ a multi-dimensional chromatographic approach for more enhanced protein fractionation. For example, a two-dimensional LC-MS experiment has been conducted on *Saccharomyces cerevisiae* by using a version of gel

electrophoresis employing acid-labile surfactants, followed by reverse-phase LC directly into an FTICR-MS (Meng, 2002).

There are several alternatives to on-line chromatography. One such approach involves surface enhanced laser desorption/ionization TOF-MS approach (Merhant, 2000). For this method, a variety of chemical (hydrophobic, ionic, or mixed) or biochemical (antibody, DNA, enzyme, or receptor) surfaces are used to preferentially absorb selected protein species. This allows the fractionation to be fairly generic or highly specific, thereby selectively reducing the complexity of the protein sample. These surfaces can be incorporated into protein chips, providing a high-throughtput sampling methodology for MALDI-TOF-MS. Another alternative to liquid chromatography focused on exploiting the demonstrated power of gel electrophoresis. As a modification of conventional 2-D PAGE, mass spectrometry has been used to replace the size-based separation component of the SDS-PAGE separation (Ogorzalek-Loo, 2001). For this method, the proteins separated according to pI are then measured by MALDI-TOF-MS, with either post-source decay dissociation of intact proteins, or peptide mass mapping experiments. Such information can be used to construct virtual 2-D gels.

(iii) Structural Interrogation. To unambiguously verify the protein assignment by top-down MS, it is advantageous to acquire at least some structural information for the intact proteins (Mortz, 1996; Reid, 2002). As discussed earlier, this can be accomplished with a variety of tandem mass spectrometry experiments, involving collisional dissociation, electron dissociation, or photodissociation. Although proteins usually fragment much less extensively than peptides, there is often sufficient fragment ion information to confirm or reject a possible protein identification from the accurate mass measurement. For example, the presence of only three or four fragment ions from a protein was found to be sufficient for a 99.8% probability of identifying the correct protein from a database of 5,000 bacterial protein forms (Meng, 2001). This methodology can be applied for proteins both with and without disulfide bonds (Nemeth-Cawley, 2002). Electron capture dissociation shows promise for the most extensive fragmentation of intact proteins in a high-throughput manner (McLafferty, 2001; Horn, 2000a). This fragmentation process yields abundant c- and z-type ions, and usually provides extensive sequence coverage of proteins even up to 45 kDa in size (Ge, 2002). A combination of collisional dissociation and electron capture dissociation can be used to provide complementary information on intact proteins in bacterial proteomes (Demirev, 2001). For very large proteins (molecular masses exceeding 150 kDa), it may be advantageous to employ partial proteolytic digestion to make large peptides (5-50 kDa), and then characterize these species (Forbes, 2001). One of the most comprehensive techniques for top-down MS is a combination of capillary LC-MS with infrared multiphoton dissociation (IRMPD) (Li, 1999). The IRMPD provides a rapid method of fragmenting the intact proteins, providing structural information on species whose accurate molecular masses have been measured.

9.4.4. Relating MS proteomic data to biological information

In the long term, the value of proteome measurements will be judged not by the *extensive catalog lists* of proteins from an organism, but rather by the *biological information* it provides. Because of this, effort should be taken to make sure that the data output presentation and analyses are done in such a way to enable the biological scientists to sort through the information in a meaningful fashion. At present, the procedures for achieving this goal are somewhat

undefined, but are rapidly developing. Strong effort should continue to be placed in this "interface" region between analytical technology and molecular biology. If done successfully, it is very likely that developments in each respective area will fuel advances in the other area.

A few large scale proteomic studies have been conducted on both prokaryotic and eukaryotic organisms, and are beginning to reveal detailed and somewhat surprising information about the proteins that are present. For example, a systematic proteomic study of rice leaf, root, and seed tissue has been completed, yielding an identification of 2,528 unique proteins [Koller, 2002]. A comparative examination of the expressed proteins indicated that that enzymes involved in central metabolic pathways were present in all tissues, whereas metabolic specialization was supported by tissue-specific enzyme complements. The ADP-glucose pyrophosphorylase was cited as one specific example, in this case providing evidence for distinct regulatory mechanisms involved in the biosynthesis and breakdown of separate starch pools in different tissues. Furthermore, several allergenic proteins were identified in the seed sample, suggesting that proteomic measurements may have the potential to survey food samples for the presence of allergens.

A global proteomic analysis of the radiation-resistant microbe *Deinococcus radiodurans* has been reported [Lipton, 2002]. More than 60% of the predicted proteome was identified in this study. 74 of the 148 predicted stress-response proteins were identified, including the two classes of annotated proteins (catalase and superoxide dismutase) known to play a role in the detoxification process. Stable isotope labeling with ^{15}N media was employed to investigate the quantitative changes in protein expression as a function of radiation dosage. Initial studies revealed that the expression of the proteins RecA and DNA-directed RNA polymerase I were significantly induced during recovery from 17.5 kGy irradiation, in agreement with previous observations. In total, although it will take some time to work through the extensive proteomic data, it is clear already that such measurements have the potential to yield important information about the function of this unique organism.

The discussions in the previous sections of this manuscript have indicated that both bottom-up and top-down MS proteomic approaches are powerful, but each still have some incompletely characterized issues. It appears that the most comprehensive proteome characterization by MS would probably be achieved by employing a combination of both methods. This would capitalize on the strength of the bottom-up approach for cataloging a large fraction of the proteins in the proteome, while the top-down approach would be essential for elucidating their molecular forms and the presence of post-translational modifications (acetylation, phosphorylation, signal peptide truncation, etc.) and gene annotation start site errors. This type of approach is likely to yield the most detailed information for the biological characterization of proteomes.

We have recently developed and demonstrated such a comprehensive method for proteome analysis that integrates both the top-down and the bottom-up approaches (VerBerkmoes, 2002). Our technique was applied for the proteomic characterization of the Gram-negative facultative anaerobe *Shewanella oneidensis* MR-1, which is of substantial interest for bioremediation of metals in contaminated soils. Our integrated approach for *S. oneidensis* enabled the facile detection of such common post-translational modifications as the loss of N-terminal methionine, signal peptide cleavages, as well as incorrect translational start sites. To present the data in a meaningful format, the proteins identified in the study were sorted into functional categories using the Clusters of Orthologous Groups of Proteins (COG) classification system (Tatusov, 1997; Tatusov, 2001). This is summarized in the pie chart in Figure 9.8

(reprinted from VerBerkmoes, 2002). Proteins from all functional categories except Q – secondary metabolite biosynthesis, metabolism and transport were identified. Many chaperones, numerous peptidases, and a number of thiol-disulfide isomerases, thioredoxins and peptidyl-prolyl cis-trans isomerases, which are all involved in protein folding or turnover, were found. The complete gluconeogenic pathway, TCA cycle and electron transport system (ETS) were identified, although most of the membrane-embedded subunits of the ETS complexes were not found. Many of the basic pathways for shunting 3 and 5 carbon skeletons to various locations are detected, as are the pathways for purine and pyrimidine metabolism. Most of the enzymes from the amino acid biosynthetic pathways are absent, while many from the metabolic pathways are present. Interestingly, many of the enzymes involved in vitamin biosynthesis were detected.

Surprisingly, three of the five decaheme cytochromes and one of the two outer membrane proteins from a cluster thought to be involved in iron and heavy metal reduction under anaerobic conditions (Beliaev, 2001) were present. In addition, there were other cytochrome c like proteins and cytochrome c maturation factors present. Considering the just mentioned genes, it is not surprising that many of the genes involved in heme biosynthesis are detected. A periplasmic binding protein whose signal peptide was putatively identified in the top-down approach is the first gene in an operon also containing an ABC transport ATPase and permease, and three genes involved in molybdopterin biosynthesis, one of which was also detected. In addition, only 135 bp separates the beginning of this operon from a 2-component signal transduction system, both members of which are detected, being transcribed in the opposite direction. This is a good candidate for a bi-directional promoter. Finally, there is one 7-gene and one 3-gene operon, and three individual genes that are all also involved in molybdenum/tungstate transport and/or cofactor biosynthesis of which at least one component was detected. All of this taken together with the detection of some oxidoreductases (nitrate reductase, formate dehydrogenase) that use a molybdo/tungstate-pterin cofactor makes an interesting system for future study. Finally, 43 purely hypothetical or *in silico* predicted genes with no BLAST hits were detected, thereby confirming their existence. This study demonstrates a novel integration of top-down and bottom-up proteomics analysis strategy for the characterization of the proteome of *S. oneidensis*. In several cases, this approach was critical for the identification of post-translational modifications and an incorrect gene start site (for thioredoxin). Improved and/or more extensive separation procedures should help improve the sensitivity of both approaches, especially for the top-down method.

9.5. SUMMARY

The conventional approach in molecular biology has been to identify a particular protein or protein class, and then systematically purify and study the structure, function, and interactions of that protein in extensive detail. The advent of experimental techniques for characterizing very complex mixtures of proteins has enabled the emergence of *proteomics*, which now permits the simultaneous examination of virtually all of the proteins expressed by an organism under a specific growth condition. This research direction represents a paradigm shift in molecular biology, in which one does not need to target a particular protein or protein class, but rather can monitor the entire protein complement of the genome in a single experiment. This provides a top-down view of all the proteins that are essential for an organism's life cycle, and is likely to provide new information that might not have been easily achieved with the traditional molecular

biology hypothesis-driven approach. This mindset has given rise to a systems biology approach for studying the processes of life. Clearly, the most comprehensive picture of life would involve complete characterization of an organism's genome, transcriptome, proteome (including protein-protein interactions), and finally, metabolome. As recently as a few years ago, this detail of information seemed unattainable. However, with the current level of experimental technology, including high performance analytical instrumentation and sophisticated bioinformatics tools, this level of detail is becoming achievable, at least for simple prokaryotic organisms. If successful, such information would not only help us to better understand the mechanics of life, but would certainly be likely to favorably impact basic issues, such as human health.

While enormous technological challenges must be addressed to achieve these goals, at the present time there appears to be no fundamental roadblocks to present such a scenario from becoming a reality. The technological hurdles, while substantial, do not appear to be insurmountable. The future of this field is likely to be marked by a replacement of the slower, labor-intensive gel electrophoresis technologies with higher throughput, wider dynamic range gel-less methods, such as the multistage LC-MS/MS and AMT approaches discussed previously. Likewise, the top-down proteomics approach will continue to see rapid development, thus providing a powerful complementary means of interrogating intact protein structures and identifying modifications. The concomitant advances in computer hardware and software are already enabling complicated bioinformatics analyses of huge datasets. The driving force for success in this general area of defining the "molecular machinery of life" will continue, at least for the short term, to be high-throughput, accurate, comprehensive characterization of all biological components and their interactions. There will certainly be a move towards increased automation to accomplish these goals. Recall that at the inception of the Human Genome Program, there was much skepticism in the scientific community that the entire human genome could be acquired within our lifetimes. In spite of this, the gauntlet was thrown down, and the technology and bioinformatics rose to meet the challenges. This remarkable achievement has paved the way for the entire field of proteomics, in which there appears to be an urgency to identify and characterize all the gene products that are predicted to be present based on the genome.

This is truly an exciting time to be involved in these scientific endeavors. The completion of the Human Genome Project has been compared to at least the equivalent of landing a human on the Moon. As in that case, the development of genomics and proteomics has opened up an entire landscape of new research opportunities, and is very likely to take us to places that we have been unable to even comprehend previously.

ACKNOWLEDGEMENTS

Research support in biological mass spectrometry at Oak Ridge National Laboratory has been provided by the U.S. Department of Energy, Office of Biological and Environmental Research. ORNL is operated and managed by the University of Battelle, LLC, for the U.S. Department of Energy under Contract DE-AC05-00OR22725.

9.6. REFERENCES

Beliaev, A. S.; Saffarini, D. A.; McLaughlin, J. L.; Hunnicutt, D. (2001) MtrC, An Outer Membrane Decaheme C Cytochrome Required for Metal Reduction in *Shewanella Putrefaciens* MR-1. *Mol Microbiol*, **39**, 722-30.

Bennett, K.L.; Matthiesen, T.; Roepstorff, P. (2000) Probing protein surface topology by chemical surface labeling, crosslinking, and mass spectrometry. *Methods Mol. Biol.* **146**, 113-131.

Biemann, K. (1988) Contributions of Mass Spectrometry to Peptide and Protein Structure, *Biomed. Environ. Mass Spectrom*. **16**, 99-111.

Bjellqvist, B.; Ek, K.; Righetti, P.G.; Gianazza, E.; Gorg, A.; Westermeier, R.; Postel, W. (1982) Isoelectric focusing in immobilized pH gradients: principle, methodology and some applications. *J Biochem Biophys Methods* **6**, 317-339.

Blank, P.S.; Sjomeling, C.M.; Backlund, P.S.; Yergey, A.L. (2002) Use of cumulative distribution functions of characterize mass spectra of intact proteins. *J. Am. Soc. Mass Spectrom.* **13**, 40-46.

Blonder, J.; Goshe, M.B.; Moore, R.J.; Pasa-Tolic, L.; Masselon, C.D.; Lipton, M.S. Smith, R.D. (2002) Enrichment of Integral Membrane Proteins for Proteomic Analysis Using Liquid Chromatography-Tandem Mass Spectometry. *J. Proteome Res*. **1**, 351-360.

Brush, M. (1998) Dye hard - Protein gel staining products. *Scientist* **12**, 16.

Bumann, D.; Meyer, T.F.; Jungblut, P.R. (2001) Proteome analysis of the common human pathogen Helicobacter pylori. *Proteomics* **1**, 473-479.

Cech, N.B. and Enki, C. G. (2001) Practical implications of some recent studies in electrospray ionization fundamentals. *Mass Spec. Reviews* **20**, 362-387.

Chernushevich, I.V.; Loboda, A.V.; Thomson, B.A. (2001) An Introduction to Quadrupole-Time-Of-Flight Mass Spectrometry. *J. Mass Spec.* **36**, 849-865.

Dass, C. (2001) "Principles and Practice of Biological Mass Spectrometry," J. Wiley-Interscience, J.Wiley and Sons, publisher, New York, New York. ISBN 0-471-33053-1.

Davis, M.T. and Lee, T.D. (1998) Rapid protein identification using a microscale electrospray LC/MS system on an ion trap mass spectrometer. *J. Am. Soc. Mass Spectrom.* **9**, 194-201.

Davis, M.T.; Spahr, C.S.; McGinley, M.D.; Robinson, J.H.; Bures, E.J.; Beierle, J.; Mort, J.; Yu, W.; Luethy, R.; Patterson, S.D. (2001) Towards defining the urinary proteome using liquid chromatography-tandem mass spectrometry. II. Limitations of complex mixture analyses. *Proteomics* **1**, 108-117.

Demirev, P.A.; Ramirez, J.; Fenselau, C. (2001) Tandem mass spectrometry of intact proteins for characterization of biomarkers from Bacillus cereus T spores *Anal. Chem*. **73**, 5725-5731

Devreese, B.; Vanrobaeys, F.; Van Beeumen, J. (2001) Automated nanoflow liquid chromatography/tandem mass spectrometric identification of proteins from Shewanella putrefaciens separated by two-dimensional polyacrylamide gel electrophoresis. *Rapid Commun. Mass Spectrom* **15**, 50-56.

Dharmasiri, K. and Smith, D.L. (1996) Mass spectrometric determination of isotopic exchange rates of amide hydrogens located on the surfaces of proteins. *Anal. Chem.* **68**, 2340-2344.

Eckerskorn, C.; Strupat, K.; Schleuder, D.; Hochstrasser, D.; Sanchez, J.C.; Lottspeich, F.; Hillenkamp, F. (1997) Analysis of proteins by direct-scanning infrared-MALDI mass spectrometry after 2D-PAGE separation and electroblotting. *Anal. Chem.* **69**, 2888-2892.

Eng, J.K.; McCormack, A.L.; Yates, J.R. 3rd (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Mass Spectrom.* **5**, 976-989.

Fenn, J.B.; Mann, M.; Meng, C.K.; Wong, S.F.; Whitehouse, C.M. (1989) Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246**, 64-71.

Forbes, A.J.; Mazur, M.T.; Patel, H.M.; Walsh, C.T.; Kelleher, N.L. (2001) Toward efficient analysis of >70 kDa proteins with 100% sequence coverage. *Proteomics* **1**, 927-933.

Fulda, S.; Huang, F.; Nilsson, F.; Hagemann, M.; Norling, B. (2000) Proteomics of Synechocystis sp. strain PCC 6803. Identification of periplasmic proteins in cells grown at low and high salt concentrations. *Eur. J. Biochem.* **267**, 5900-5907.

Gatlin, C.L.; Kleemann, G.R.; Hays, L.G.; Link, A.J.; Yates, J.R. 3rd (1998) Protein identification at the low femtomole level from silver-stained gels using a new fritless electrospray interface for liquid chromatography-microspray and nanospray mass spectrometry. *Anal. Biochem.* **263**, 93-101.

Gavin, A.-C.; et.al. (2002) Functional Organization of the Yeast Proteome by Systematic Analysis of Protein Complexes. *Nature* **415**, 141-147.

Ge, Y.; Lawhorn, B.G.; ElNaggar, M.; Strauss, E.; Park, J.-H.; Begley, T.P.; McLafferty, F.W. (2002) Top down characterization of larger proteins (45 kDa) by electron capture dissociation mass spectrometry. *J. Am. Chem. Soc.* **124**, 672-678.

Gomez, S.M.; Nishio, J.N.; Faull, K.F.; Whitelegge, J.P. (2002) The chloroplast grana proteome defined by intact mass measurements from liquid chromatography mass spectrometry. *Mol. Cell Proteomics* **1**, 46-59.

Gorg, A.; Obermaier, C.; Boguth, G.; Harder, A.; Scheibe, B.; Wildgruber, R.; Weiss, W. (2000) The current state of two-dimensional electrophoresis with immobilized pH gradients. *Electrophoresis,* **21**, 1037-1053.

Griffin, T.J.; Gygi, S.P.; Ideker, T.; Rist, B.; Eng, J.; Hood, L.; Aebersold, R. (2002) Complementary profiling of gene expression at the transcriptome and proteome levels in Saccharomyces cerevisiae. *Mol. Cell Proteomics* **1**, 323-333.

Grunenfelder, B.; Rummel, G.; Vohradsky, J.; Roder, D.; Langen, H.; Jenal, U. (2001) Proteomic analysis of the bacterial cell cycle. *Proc. Natl. Acad. Sci. USA* **98**, 4681-4686.

Gygi, S.P.; Rist, B.; Gerber, S.A.; Turecek, F.; Gelb, M.H.; Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **17**, 994-999.

Han, D.K.; Eng, J.; Zhou, H.; Aebersold, R. (2001) Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nat. Biotechnol.* **19**, 946-951.

Hao, C.Y. and March, R.E. (2001) A survey of recent research activity in quadrupole ion trap mass spectrometry. *Int. J. Mass Spec*. **212**, 337-357.

Hendrickson, C.L. and Emmett, M.R. (1999) Electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *Annu. Rev. Phys. Chem.* **50**, 517-536.

Hermann, T.; Pfefferle, W.; Baumann, C.; Busker, E.; Schaffer, S.; Bott, M.; Sahm, H.; Dusch, N.; Kalinowski, J.; Puhler, A.; Bendt, A.K.; Kramer, R.; Burkovski, A. (2001) Proteome analysis of Corynebacterium glutamicum. *Electrophoresis* **22**, 1712-1723.

Hernandez, J.; Robinson, C.V. (2001) Dynamic Protein Complexes: Insights from Mass Spectrometry. *J. Biol. Chem*. **276**, 46685-46688.

Hernychova, L.; Stulik, J.; Halada, P.; Macela, A.; Kroca, M.; Johansson, T.; Malina, M. (2001) Construction of a Francisells tularensis two-dimensional electrophoresis protein database. *Proteomics* **1**, 508-515.

Hess, D.; Covey, T.C.; Winz, R.; Brownsey, R.W.; Aebersold, R. (1993) Analytical and micropreparative peptide mapping by high performance liquid chromatography/electrospray mass spectrometry of proteins purified by gel electrophoresis. *Protein Sci.* **2**, 1342-1351.

Hillenkamp, F.; Karas, M.; Beavis, R.C.; Chait, B.T. (1991) Matrix-assisted laser desorption/ionization mass spectrometry of biopolymers. *Anal. Chem.* **63**, 1193A-1203A.

Ho, Y.; et.al. (2002) Systematic Identification of Protein Complexes in Saccharomyces cerevisiae by Mass Spectrometry. *Nature* **415**, 180-183.

Horn, D.M.; Ge, Y.; McLafferty, F.W. (2000a) Activated ion electron capture dissociation for mass spectral sequencing of larger (42 kDa) proteins. *Anal. Chem.* **72**, 4778-4784.

Horn, D.M.; Zubarev, R.A.; McLafferty, F.W. (2000b) Automated de novo sequencing of proteins by tandem high-resolution mass spectrometry. *Proc. Natl. Acad. Sci. USA* **97**, 10313-10317.

Horn, D.M.; Zubarev, R.A.; McLafferty, F.W. (2000c) Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J. Am. Soc. Mass Spectrom.* **11**, 320-332.

Hunt, D.F.; Buko, A.M.; Ballard, J.M.; Shabanowitz, J.; Giordani, A.B. (1981) Sequence analysis of polypeptides by collision activated dissociation on a triple quadrupole mass spectrometer. *Biomed. Mass Spectrom.* **8**, 397-408.

Hunt, D.F.; Yates, J.R. 3rd; Shabanowitz, J.; Winston, S.; Hauer, C.R. (1986) Protein sequencing by tandem mass spectrometry. *Proc. Natl. Acad. Sci. USA* **83**, 6233-6237.

Iborra, F. and Buhler, J.M. (1976) Protein subunit mapping. A sensitive high resolution method. *Anal. Biochem.* **74,** 503-511.

Ideker, T.; Thorsson, V.; Ranish, J.A.; Christmas, R.; Buhler, J.; Eng, J.K.; Bumgarner, R.; Goodlett, D.R.; Aebersold, R.; Hood, L. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292**, 929-934.

Jensen, O.N.; Larsen, M.R.; Roepstorff, P. (1998) Mass spectrometric identification and microcharacterization of proteins from electrophoretic gets: strategies and applications. *Proteins* **2**, 74-89.

Jungblut, P. and Thiede, B. (1997) Protein identification from 2-DE gels by MALDI mass spectrometry. *Mass Spectrom. Rev.* **16**, 145-162.

Kaltashov, I.A.; Eyles, S.J. (2002) Studies of Biomolecular Conformations and Conformational Dynamics by Mass Spectrometry. *Mass Spec. Rev*. **21**, 37-71.

Kelleher, N.L.; Taylor, S.V.; Grannis, D.; Kinsland, C.; Chiu, H.J.; Begley, T.P.; McLafferty, F.W. (1998) Efficient sequence analysis of the six gene products (7-74 kDa) from the Escherichia coli thiamin biosynthetic operon by tandem high-resolution mass spectrometry. *Protein Sci*. **7**, 1796-1801.

Klose, J. (1975) Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals. *Humangenetik* **26,** 231-243.

Koller, A.; Washburn, M.P.; Lange, B.M.; Andon, N.L.; Deciu, C.; Haynes, P.A.; Hays, L.; Schieltz, D.; Ulaszek, R.; Wei, J.; Wolter, D.; Yates, J.R. 3rd (2002) Proteomic survey of metabolic pathways in rice. *PNAS* **99**, 11969-11974.

Langen, H.; Takacs, B.; Evers, S.; Berndt, P.; Lahm, H.W.; Wipf, B.; Gray, C.; Fountoulakis, M. (2000) Two-dimensional map of the proteome of Haemophilus influenzae. *Electrophoresis* **21**, 411-429.

Larsen, M.R. and Roepstorff, P. (2000) Mass spectrometric identification of proteins and characterization of their post-translational modifications in proteome analysis. *Fresenius J. Anal. Chem*. **366**, 677-690.

Lee, S.-W.; Berger, S.J.; Martinović, S.; Paša-Tolić, L.; Anderson, G.A.; Shen, Y.; Zhao, R.; Smith, R.D. (2002) Direct mass spectrometric analysis of intact proteins of the yeast large ribosomal subunit using capillary LC/FTICR. *PNAS* **99**, 5942-5947.

Li, W.; Hendrickson, C.L.; Emmett, M.R.; Marshall, A.G. (1999) Identification of intact proteins in mixtures by alternated capillary liquid chromatography electrospray ionization and LC ESI infrared multiphoton dissociation Fourier transform ion cyclotron resonance mass spectrometry. *Anal. Chem.* **71**, 4397-4402.

Licklider, L.J.; Thoreen, C.C.; Peng, J.; Gygi, S.P. (2002) Automation of nanoscale microcapillary liquid chromatography-tandem mass spectrometry with a vented column. *Anal. Chem.* **74**, 3076-3083.

Link, A.J.; Eng, J.; Schieltz, D.M.; Carmack, E.; Mize, G.J; Morris, D.R.; Garvik, B.M.; Yates, J.R. 3rd (1999) Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17**, 676-682.

Lipton, M.S.; Pasa-Tolic, L.; Anderson, G.A.; Anderson, D.J.; Auberry, D.L.; Battista, J.R.; Daly, M.J.; Fredrickson, J.; Hixson, K.K.; Kostandarithes, H.; Masselon, C.; Markillie, L.M.; Moore, R.J.; Romine, M.F.; Shen, Y.; Stritmatter, E.; Tolic, N.; Udseth, H.R.; Venkateswaran, A.; Wong, K.; Zhao, R.; Smith, R.D. (2002) Global analysis of Deinococcus radiodurans proteome by using accurate mass tags. *PNAS* **99**, 11049-11054.

Little, D.P.; Speir, J.P.; O'Connor, P.B.; McLafferty, F.W. (1994) Infrared multiphoton dissociation of large multiply charged ions for biomolecule sequencing. *Anal. Chem.* **66**, 2809-2815.

Liu, H.; Lin. D.; Yates, J.R. (2002) Multidimensional Separations for Protein/Peptide Analysis in the Post-Genomic Era. *Biotech*. **32**, 898-902.

Loo, J.A.; Quinn, J.P.; Qyu, S.I.; Henry, K.D.; Senko, M.W.; McLafferty, F.W. (1992) High-resolution tandem mass spectrometry of large biomolecules. *Proc. Natl. Acad. Sci. USA* **89**, 286-289.

Mann, M.; Hendrickson, R.C.; Pandey, A. (2001) Analysis of proteins and proteomes by mass spectrometry. *Annu. Rev. Biochem.* **70**, 437-473.

Marshall, A. (2001) Milestones in Fourier transform ion cyclotron resonance mass spectrometry technique development. *Int. J. Mass Spec*. **2001**, 331-356.

Marshall, A.G.; Hendrickson, C.L.; Jackson, G.S. (1998) Fourier transform ion cyclotron resonance mass spectrometry: A primer. *Mass Spec. Reviews* **17**, 1-35.

Martin, S.E.; Shabanowitz, J.; Hunt, D.F.; Marto, J.A. (2000) Subfemtomole MS and MS/MS peptide sequence analysis using nano-HPLC micro-ESI fourier transform ion cyclotron resonance mass spectrometry. *Anal. Chem.* **72**, 4266-4274.

Mawuenyega, K.G.; Kaji, H.; Yamauchi, Y.; Shinkawa, T.; Saito, H.; Taoka, M.; Takahashi, N.; Isobe, T. (2002) Large-scale identification of Caenorhabditis elegans proteins by multidimensional liquid chromatography – tandem mass spectrometry. *J. Proteome Res.*, in press for **vol 2**, Jan. 2003.

McCormack, A.L.; Schieltz, D.M.; Goode, B.; Yang, S.; Barnes, G.; Drubin, D.; Yates, J.R. 3rd (1997) Direct analysis and identification of proteins in mixtures by LC/MS/MS and database searching at the low-femtomole level. *Anal. Chem.* **69**, 767-776.

McLafferty, F.W.; Horn, D.M.; Breuker, K.; Ge, Y.; Lewis, M.A.; Cerda, B.; Zubarev, R.A.; Carpenter, B.K. (2001) Electron capture dissociation of gaseous multiply charged ions by Fourier-transform ion cyclotron resonance. *J. Am. Soc. Mass Spectrom.* **12**, 245-249.

McLuckey, S.A. and Stephenson, J.L. Jr. (1998) Ion/ion chemistry of high-mass multiply charged ions. *Mass Spectrom. Rev.* **17**, 369-407.

Meng, F.; Cargile, B.J.; Miller, L.M.; Forbes, A.J.; Johnson, J.R.; Kelleher, N.L. (2001) Informatics and multiplexing of intact protein identification in bacteria and the archaea. *Nature Biotech.* **19**, 952-957.

Meng, F.; Cargile, B.J.; Patrie, S.M.; Johnson, J.R.; McLoughlin, S.M.; Kelleher, N.L. (2002) Processing complex mixtures of intact proteins for direct analysis by mass spectrometry. *Anal. Chem.* **74**, 2923-2929.

Merchant, M. and Weinberger, S.R. (2000) Recent advancements in surface-enhanced laser desorption/ionization-time of flight-mass spectrometry. *Electrophoresis* **21**, 1164-1167.

Molloy, M.P.; Herbert, B.R.; Slade, M.B.; Rabilloud, T.; Nouwens, A.S.; Williams, K.L.; Gooley, A.A. (2000) Proteomic analysis of the Escherichia coli outer membrane. *Eur. J. Biochem.* **267**, 2871-2881.

Molloy, M.P.; Phadke, N.D.; Maddock, J.R.; Andrews, P.C. (2001) Two-dimensional electrophoresis and peptide mass fingerprinting of bacterial out membrane proteins. *Electrophoresis* **22**, 1686-1696.

Mortz, E.; O'Connor, P.B.; Roepstorff, P.; Kelleher, N.L.; Wood, T.D.; McLafferty, F.W. Mann, M. (1996) Sequence tag identification of intact proteins by matching tandem mass spectral data against sequence data bases. *Proc. Natl. Acad. Sci. USA* **93**, 8264-8267.

Mortz, E.; Vorm, O.; Mann, M.; Roepstorff, P. (1994) Identification of proteins in polyacrylamide gels by mass spectrometric peptide mapping combined with database search. *Biol. Mass Spectrom.* **23**, 249-261.

Nakanishi T.; Okamoto N.,; Tanaka, K.,; Shimizu, A. (1994) Laser-Desorption Time-Of-Flight Mass-Spectrometric Analysis Of Transferrin Precipitated With Antiserum - A Unique Simple Method To Identify Molecular-Weight Variants. *Biological Mass Spectrometry* **23,** 230-233.

Nemeth-Cawley, J.F. and Rouse, J.C. (2002) Identification and sequencing analysis of intact proteins via collision-induced dissociation and quadrupole time-of-flight mass spectrometry. *J. Mass Spectrom.* **37**, 270-282.

Nouwens, A.S.; Cordwell, S.J.; Larsen, M.R.; Molloy, M.P.; Gillings, M.; Willcox, M.D.; Walsh, B.J. (2000) Complementing genomics with proteomics: The membrane subproteome of Pseudomonas aeruginosa PA01. *Electrophoresis* **21**, 3797-3809.

O'Donovan, C.; Apweiler, R.; Bairoch A. (2001) The human proteomics initiative (HPI). *Trends Biotechnol.* **19**, 178-181.

Oda, Y.; Huang, K.; Cross, F.R.; Cowburn, D.; Chait, B.T. (1999) Accurate quantitation of protein expression and site-specific phosphorylation. *Proc. Natl. Acad. Sci. USA* **96**, 6591-6596.

Ogorzalek Loo, R.R.; Cavalcoli, J.D.; VanBogelen, R.A.; Mitchell, C.; Loo, J.A.; Moldover, B.; Andrews, P.C. (2001) Virtual 2-D gel electrophoresis: Visualization and analysis of the *E. coli* proteome by mass spectrometry. *Anal. Chem.* **73**, 4063-4070.

Pandey, A. and Mann, M. (2000a) Proteomics to study genes and genomes. *Nature* **405**, 837-846.

Pandey, A.; Podtelejnikov, A.V.; Blagoev, B.; Bustelo, X.R.; Mann, M.; Lodish, H.F. (2000b) Analysis of receptor signal pathways by mass spectrometry: identification of vav-2 as a substrate of the epidermal and platelet-derived growth factor receptors. *Proc. Natl. Acad. Sci. USA* **97**, 179-184.

Paša-Tolić, L.; Jensen P.K.; Andrson, G.A.; Lipton, M.S.; Peden, K.K.; Martinović, S.; Tolić, N.; Bruce, J.E.; Smith, R.D. (1999) High-throughput proteome-wide precision measurements of protein expression using mass spectrometry. *J. Am. Chem. Soc.* **121**, 7949-7950.

Patterson, S.D. and Aebersold R. (1995) Mass spectrometric approaches for the identification of gel-separated proteins. *Electrophoresis* **16**, 1791-1814.

Peng, J.; Elias, J.E.; Thoreen, C.C.; Licklider, L.J.; Gygi, S.P. (2002) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: The yeast proteome. *J. Proteome Res.* **ASAP Article**.

Peng J. and Gygi, S.P. (2001) Proteomics: The Move to Mixtures. *J. Mass Spec*. **36**, 1083-1091.

Pennington, S.R. and M.J. Dunn, editors. (2001) "Proteomics: From Protein Sequence to Function," Springer-Verlag Publishers, New York. (ISBN 0-387-91589-3)

Perkins, D.N.; Pappin, D.J.; Creasy, D.M.; Cottrell, J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551-3567.

Puig, O.; Caspary, F.; Rigaut, G.; Rutz, B.; Bouveret, E.; Bragado-Nilsson, E.; Wilm, M.; Seraphin, B. (2001) The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods* **24**, 218-229.

Quenzer, T.L.; Emmett, M.R.; Hendrickson, C.L.; Kelly, P.H.; Marshall, A.G. (2001) High sensitivity Fourier transform ion cyclotron resonance mass spectrometry for biological analysis with nano-LC and microelectrospray ionization. *Anal. Chem.* **73**, 1721-1725.

Reid, G.E. and McLuckey, S.A. (2002) 'Top down' protein characterization via tandem mass spectrometry. *J. Mass Spectrom.* **37**, 663-675.

Reynolds, K.J.; Yao, X.; Fenselau, C. (2002) Proteolytic^{18}O labeling for comparative proteomics: evaluation of endoprotease Glu-C as the catalytic agent. *J. Proteome Res.* **1**, 27-33.

Roepstorff, P.; Fohlman, J. (1984) Proposal for a Common Nomenclature for Sequence Ions in Mass-Spectra of Peptides, *Biomed. Mass Spec.***11**, 601.

Sazuka, T.; Yamaguchi, M.; Ohara, O. (1999) Cyano2Dbase updated: linkage of 234 protein spots to corresponding genes through N-terminal microsequencing. *Electrophoresis* **20**, 2160-2171.

Scheele, G.A. (1975) Two-dimensional gel analysis of soluble proteins. Characterization of guinea pig exocrine pancreatic proteins. *J. Biol. Chem.* **250,** 5375-5385.

Senko, M.W.; Spear, J.P.; McLafferty, F.W. (1994) Collisional activation of large multiply charged ions using Fourier transform mass spectrometry. *Anal. Chem.* **66**, 2801-2808.

Shen, Y.; Zhao, R.; Belov, M.E.; Conrads, T.P.; Anderson, G.A.; Tang, K.; Paša-Tolić, L.; Veenstra, T.D.; Lipton, M.S.; Udseth, H.R.; Smith, R.D. (2001) Packed capillary reversed-phase

liquid chromatography with high-performance electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry for proteomics. *Anal. Chem.* **73**, 1766-1775.

Shevchenko, A.; Jensen, O.N.; Podtelejnikov, A.V.; Sagliocco, F.; Wilm, M.; Vorm, O.; Mortensen, P.; Shevchenko, A.; Boucherie, H.; Mann M. (1996a) Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. *Proc. Natl. Acad. Sci. USA* **93**, 14440-14445.

Shevchenko, A.; Wilm, M.; Vorm, O.; Mann, M. (1996b) Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. *Anal. Chem.* **68**, 850-858.

Smith, D.L.; Deng, Y.Z.; Zhang, Z.Q. (1997). Probing the non-covalent structure of proteins by amide hydrogen exchange and mass spectrometry. *J. Mass Spec*. **32**,135-146.

Smith, R.D.; Anderson, G.A.; Lipton, M.S.; Masselon, C.; Paša-Tolić, L.; Shen, Y.; Udseth, H.R. (2002a) The use of accurate mass tags for high-throughput microbial proteomics. *OMICS* **6**, 61-90.

Smith, R.D.; Anderson, G.A.; Lipton, M.S.; Paša-Tolić, L.; Shen, Y.; Conrads, T.P.; Veenstra, T.D.; Udseth, H.R. (2002b) An accurate mass tag strategy for quantitative and high-throughput proteome measurements. *Proteomics* **2**, 513-523.

Spahr, C.S.; Davis, M.T.; McGinley, M.D.; Robinson, J.H.; Bures, E.J.; Beierle, J.; Mort, J.; Courchesne, P.L.; Chen, K.; Wahl, R.C.; Yu, W.; Luethy, R.; Patterson, S.D. (2001) Towards defining the urinary proteome using liquid chromatography-tandem mass spectrometry. I. Profiling an unfractionate tryptic digest. *Proteomics* **1**, 93-107.

Steinberg, T.H.; Haugland, R.P; Singer, V.L. (1996a) Applications of SYPRO orange and SYPRO red protein gel stains. *Anal. Biochem.* **239**, 238-245.

Steinberg, T.H.; Jones, L.J.; Haugland, R.P.; Singer, V.L. (1996b) SYPRO orange and SYPRO red protein gel stains: one-step fluorescent staining of denaturing gels for detection of nanogram levels of protein. *Anal. Biochem.* **239**, 223-237.

Tabb, D.L.; Hayes-McDonald, W.; Yates, J.R. (2002) DTASelect and Contrast: Tools for Assembling and Comparing Protein Identifications from Shotgun Proteomics. *J. Proteome Res.* **1**, 21-26.

Tolley, L.; Jorgenson, J.W.; Moseley, M.A. (2001) Very high pressure gradient LC/MS/MS. *Anal. Chem.* **73**, 2985-2991.

Tatusov, R. L.; Koonin, E. V.; Lipman, D. J. (1997) A Genomic Perspective on Protein Families. *Science*, **278**, 631-637.

Tatusov, R. L.; Natale, D. A.; Garkavtsev, I. V.; Tatusova, T. A.; Shankavaram, U. T.; Rao, B.S.; Kiryutin, B.; Galperin, M. Y.; Fedorova, N. D.; Koonin, E. V. (2001) The COG database: new

developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res*, **29**, 22-28.

Uchiki, T.; Hettich, R.; Gupta, V.; Dealwis, C. (2002) Characterization of monomeric and dimeric forms of recombinant sml1p-histag protein by electrospray-MS. *Anal. Biochem*. **301**, 35-48.

Van Montfort, B.A.; Canas, B.; Duurkens. R.; Godovac-Zimmermann, J.; Robillard, G.T. (2002) Improved In-Gel Approaches to Generate Peptide Maps of Integral Membrane Proteins with Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry, *J. Mass Spec*. **37**, 322-330.

Valdes, I.; Pitarch, A.; Gil, C.; Bermudez, A.; Llorente, M.; Nombela, C.; Mendez, E. (2000) Novel procedure for identification of proteins by mass fingerprinting combining two-dimensional electrophoresis with fluorescent SYPRO red staining. *J. Mass Spectrom.* **35**, 672-682.

Venter, J.C.; et.al. Science, (2001) The Sequence of the Human Genome. *Science* **291**, 1304-1351.

ver der Heeft, E.; ten Hove, G.J.; Herberts, C.A.; Meiring, H.D.; van Els, C.A.; de Jong, A.P. (1998) A microcapillary column switching HPLC-electrospray ionization MS system for the direct identification of peptides presented by major histocompatibility complex class I molecules. *Anal. Chem.* **70**, 3742-3751.

VerBerkmoes, N.C.; Bundy, J.L.; Hauser, L.; Asano, K.G.; Razumovskaya, J.; Larimer, F.; Hettich, R.L.; Stephenson, J.L. Jr. (2002) Integrating "Top-Down" and "Bottom-Up" mass spectrometric approaches for proteomic analysis of Shewanella oneidensis. *J. Proteome Res.* **1**, 239-252.

Vestal, M.L. (2001) Methods of Ion Generation. *Chem. Rev*. **101**, 361-375.

Von Mering, C.; Krause, R.; Snel, B.; Cornell, M.; Oliver, S.G.; Fields, S. Bork, P. (2002) Comparative Assessment of Large-Scale Data Sets of Protein-Protein Interactions. *Nature* **417**, 399-403.

Wagner, M.A.; Eschenbrenner, M.; Horn, T.A.; Kraycer, J.A.; Mujen, C.V.; Hagius, S.; Elzer, P.; DelVecchio, V.G. (2002) Global analysis of the Brucella melitensis proteome: Identification of proteins expressed in laboratory-grown culture. *Proteomics* **2**, 1047-1060.

Washburn, M.P.; Ulaszek, R.; Deciu, C.; Schieltz, D.M.; Yates, J.R. 3rd (2002) Analysis of quantitative proteomic data generated via multidimensional protein identification technology. *Anal. Chem.* **74**, 1650-1657.

Washburn, M.P.; Wolters, D.; Yates, J.R. 3rd (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19**, 242-247.

Wasinger, V.C.; Polack, J.D., Humphery-Smith, I. (2000) The proteome of Mycoplasma genitalium. Chaps-soluble component. *Eur. J. Biochem.* **267**, 1571-1582.

Wildgruber, R.; Harder, A.; Obermaier, C.; Boguth, G.; Weiss, W.; Fey, S.J.; Larsen, P.M.; Gorg, A. (2000) Towards higher resolution: two-dimensional electrophoresis of Saccharomyces cerevisiae proteins using overlapping narrow immobilized pH gradients. *Electrophoresis* **21,** 2610-2616.

Wilm, M.; Shevchenko, A.; Houthaeve, T.; Breit, S.; Schweigerer, L.; Fotsis, T.; Mann M. (1996) Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. *Nature* **379**, 466-469.

Wolter, D.A.; Washburn, M.P.; Yates J.R. 3rd (2001) An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* **73**, 5683-5690.

Yao, X.; Freas, A.; Ramirez, J.; Demirev, P.A.; Fenselau, C. (2001) Proteolytic ^{18}O labeling for comparative proteomics: model studies with two serotypes of adenovirus. *Anal. Chem.* **73**, 2836-2842.

Yates J.R. (2000) Mass spectrometry – from genomics to proteomics. *Trends in Genetics* **16**, 5-8.

Young, M.M.; Tang, N.; Hempel, J.C.; Oshiro, C.M.; Taylor, E.W.; Kuntz, I.D.; Gibson, B.W.; Dollinger, G. (2000) High-throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. *Proc. Natl. Acad. Sci. USA* **97**, 5802-5806.

FIGURE CAPTIONS

Figure 9.1 Electrospray Fourier transform ion cyclotron resonance (ES-FTICR) mass spectra of the protein ubiquitin, illustrating the multiply-charged ions observed in the positive ion mass spectra (a), and the deconvoluted view (b) showing the isotopic molecular region.

Figure 9.2 Deconvoluted mass spectrum of the collisional dissociation (MS/MS) of the 10^+ charge state (m/z 857) of ubiquitin. Inset reveals the fragment ion identities and sequence locations.

Figure 9.3 Alphabetic code used to designate fragment ion types and locations from a generic peptide. (Roepstorff, 1984; Beimann, 1988).

Figure 9.4 Flow diagram depicting the three most common mass spectrometry experiments that are conducted to identify peptides from the enzymatic digestions of 2D-PAGE gel spots.

Figure 9.5 Illustration of the shotgun proteomics technique for the microbe *Shewanella oneidensis*. The entire cell lysate is digested with trypsin, cleaned-up, and then injected directly onto a C18 reverse phase column interfaced on-line with a quadrupole mass spectrometer. The base peak chromatogram (upper left) reveals a complex mixture of peptides in this sample. A full scan mass spectrum at 60.4 min. (upper right) into the HPLC run reveals the presence of several peptides at this particular time slice. Even the minor peptides can be isolated and examined by MS/MS techniques (bottom left) to yield fragment ions which provide structural information (bottom right).

Figure 9.6 Flow diagram depicting how the mass spectra from Figure 9.5 are processed. The raw data output is processed with the software algorithms DTAselect and Contrast (Tabb, 2002) to filter and compile the data into files of protein identifications and scores. This information can be used to generate biologically informative KEGG maps and functional category pie charts.

Figure 9.7 ES-FTICR-MS of intact proteins from an anion exchange liquid chromatography fraction from yeast lysate. Note the appearance of at least a dozen different protein species. The insets reveal how high resolution measurements of the deconvoluted isotopic packets provide measured molecular masses which are in excellent agreement with calculated values.

Figure 9.8 Functional category pie chart of protein classes identified by shotgun proteomics LC-MS/MS methodology for the microbe *Shewanella oneidensis*.

SUGGESTED READING MATERIAL

"Principles and Practice of Biological Mass Spectrometry," Dass, C. (2001) J. Wiley-Interscience, J.Wiley and Sons, publisher, New York, New York. ISBN 0-471-33053-1.

"Analysis of proteins and proteomes by mass spectrometry," Mann, M.; Hendrickson, R.C.; Pandey, A. *Annu. Rev. Biochem.* (2001) **70**, 437-473.

"Subcellular proteomics," Dreger M, *Mass. Spec. Rev.* (2003) **22**, 27-56.

Glossary

Bioinformatics. The development and application of computational tools to assist with data analysis from analytical technologies as well as *in-silico* predictions of biological structures and interactions.

Bottom-up proteomics. The characterization of a proteome by proteolytic digestion of the complex protein mixture, followed by identification of the resulting peptides. The proteins present are deduced from the peptides measured and identified.

Electrospray Ionization (ESI). The process of transferring preformed biological ions from a solution directly into the gas phase for detection by mass spectrometry. This is accomplished by flowing the sample solution through a metal needle which is biased at a high voltage relative to the entrance aperture of the mass spectrometer.

Gel electrophoresis (GE). A method to separate proteins by their differential electrophoretic mobility through a gel as a function of the voltage difference applied to the gel.

Liquid chromatography (LC). An analytical technique for separating molecules based on their selective partitioning between a flowing mobile phase and suitable stationary material. A common method employed for proteins and peptides utilizes an alaphatic stationary phase and an aqueous/organic solution phase, in which the biomolecules are separated on the basis of their hydrophobicities.

Mass spectrometry (MS). An analytical technique for measuring ions corresponding to the components of a sample.

Matrix-assisted laser desorption/ionization (MALDI). The process of transferring biological species from a sample directly into the gas phase for detection by mass spectrometry. This is accomplished by mixing a biological solution with a molar excess of laser-absorbing matrix compound (such as sinapinic acid), and then drying this material onto a target plate. Gas phase ions are generated by laser desorption/ionization of the resulting sample.

Multistage LC-MS. The use of two or more stages of liquid chromatography in conjunction with mass spectrometry. This provides extensive separation of biological components prior to mass spectrometric detection.

Nanospray. The technique of conducting electrospray ionization at very low flow rates (nl/min) with small capillaries. This technique affords enhanced resolution and detection sensitivity over conventional electrospray ionization.

On-line LC-MS. The process of connecting a liquid chromatograph directly to a mass spectrometer. This minimizes sample losses and provides for automated analyses.

Quantitation. The measurement of the concentration levels of the components of a sample.

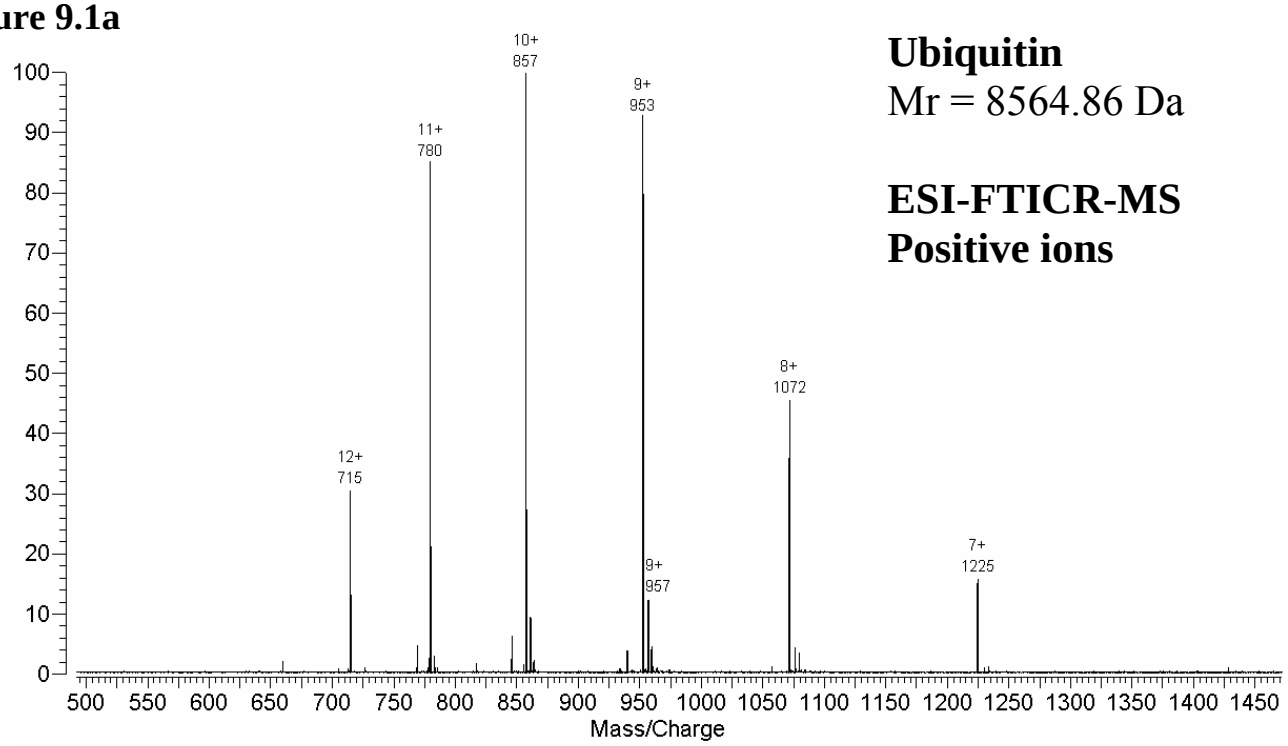
Sequence tag. A short section of sequential amino acids from a given protein sequence. This information is often obtained by tandem mass spectrometry investigations of peptides, and enables protein identification by database querying.

Shotgun proteomics. The technique of digesting an entire proteome with a suitable protease, and then directly examining the mixture by on-line liquid chromatography-mass spectrometry (i.e., no gel separations are employed).

Tandem mass spectrometry (MS/MS). The process of isolating a selected molecular ion in a mass spectrum, followed by induced dissociation into characteristic fragment ions. For proteins and peptides, this often provides the sequence information necessary to identify the species.

Top-down proteomics. The characterization of a proteome by the direct measurement of the intact proteins in a complex mixture. This provides information about the molecular forms of the proteins present, including details about protein post-translational processing.

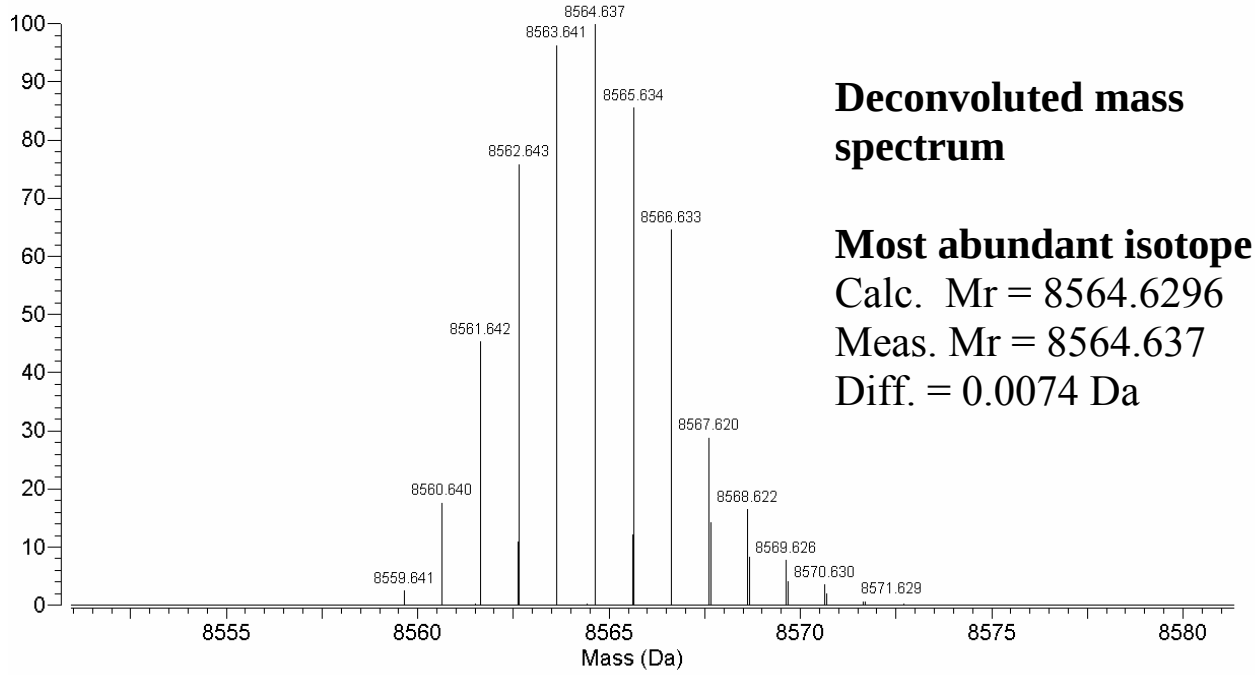
Figure 9.1a



Ubiquitin
Mr = 8564.86 Da

ESI-FTICR-MS
Positive ions

Figure 9.1b



Deconvoluted mass spectrum

Most abundant isotope
Calc. Mr = 8564.6296
Meas. Mr = 8564.637
Diff. = 0.0074 Da

Figure 9.2

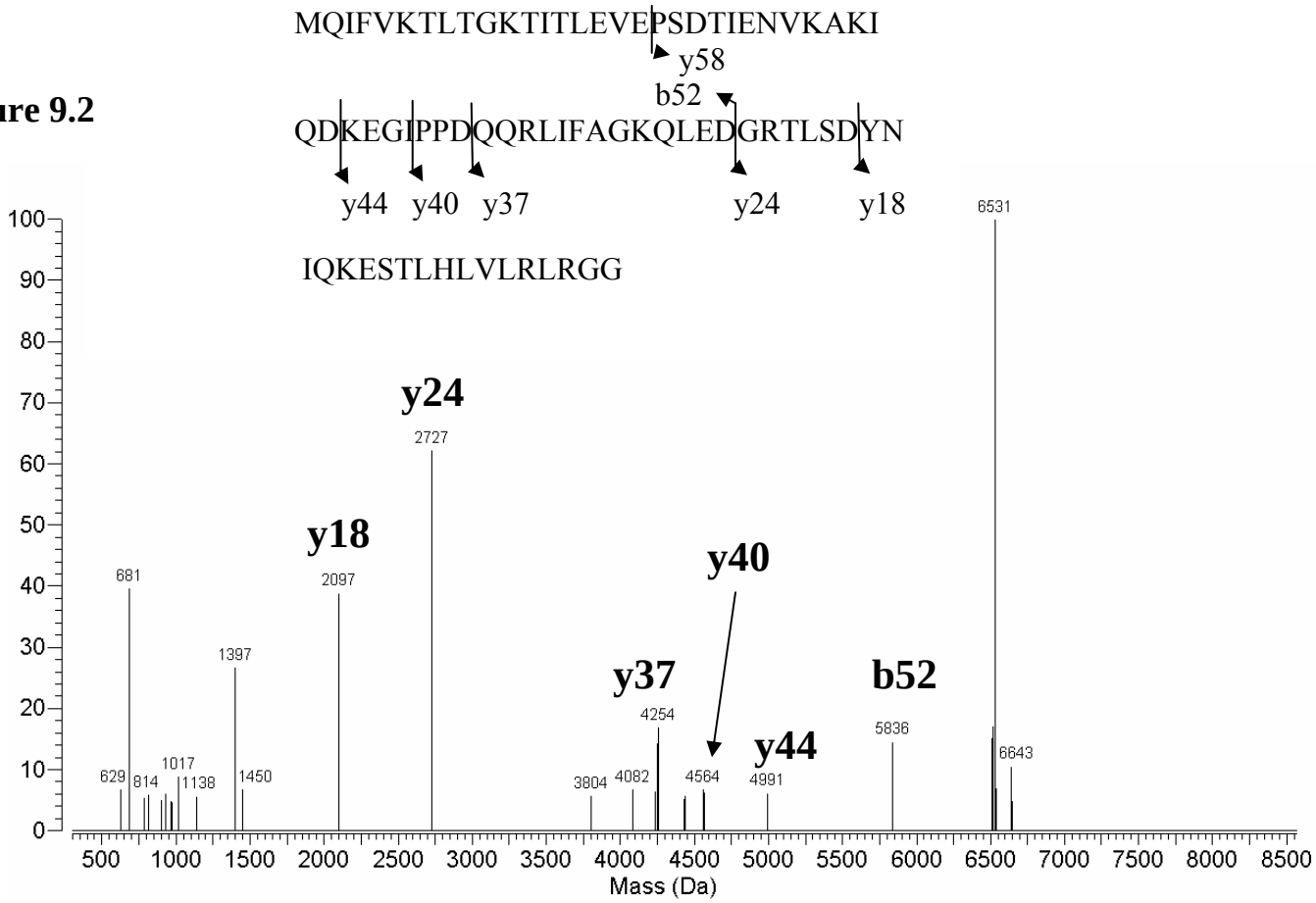


Figure 9.3

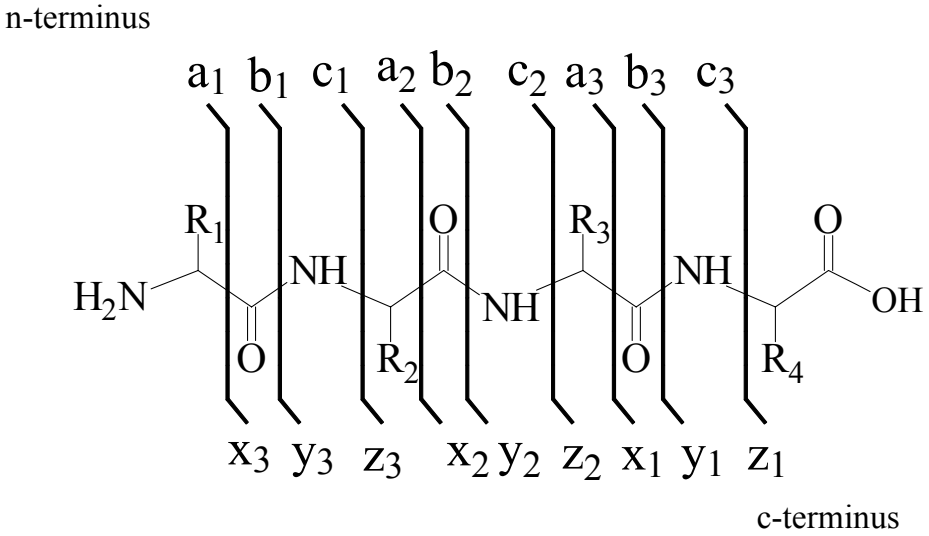


FIGURE 9.4

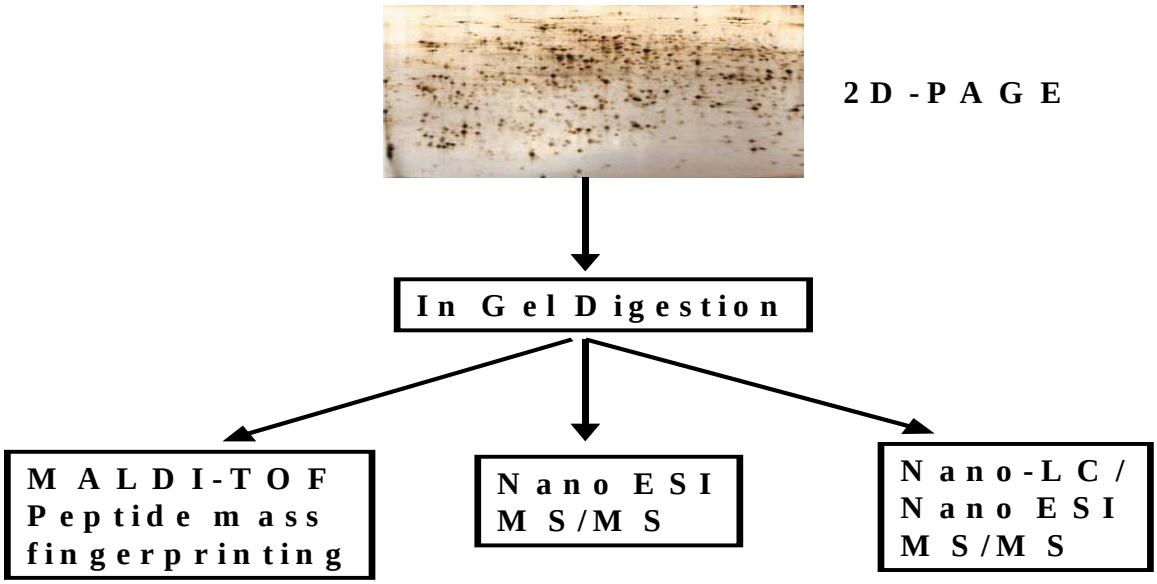

2D-PAGE
In Gel Digestion
MALDI-TOF Peptide mass fingerprinting
Nano ESI MS/MS
Nano-LC/ Nano ESI MS/MS

Figure 9.5

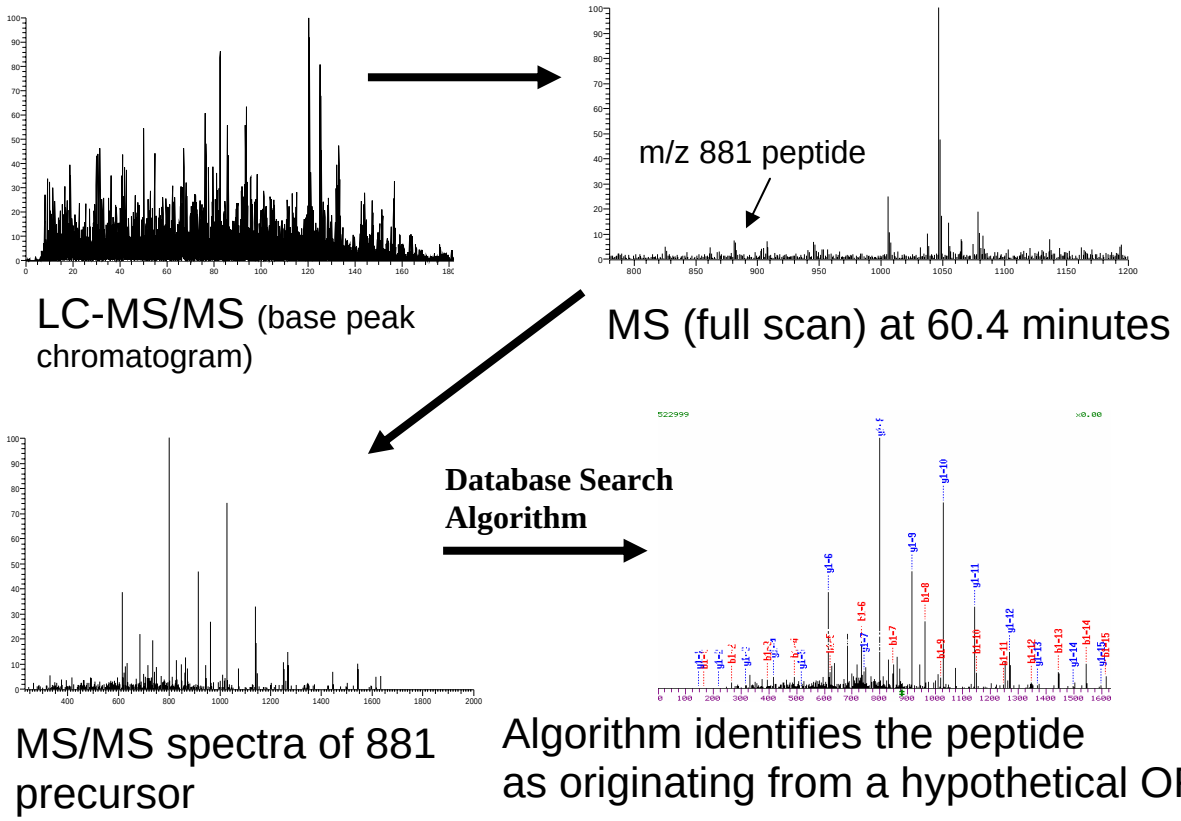


Figure 9.6

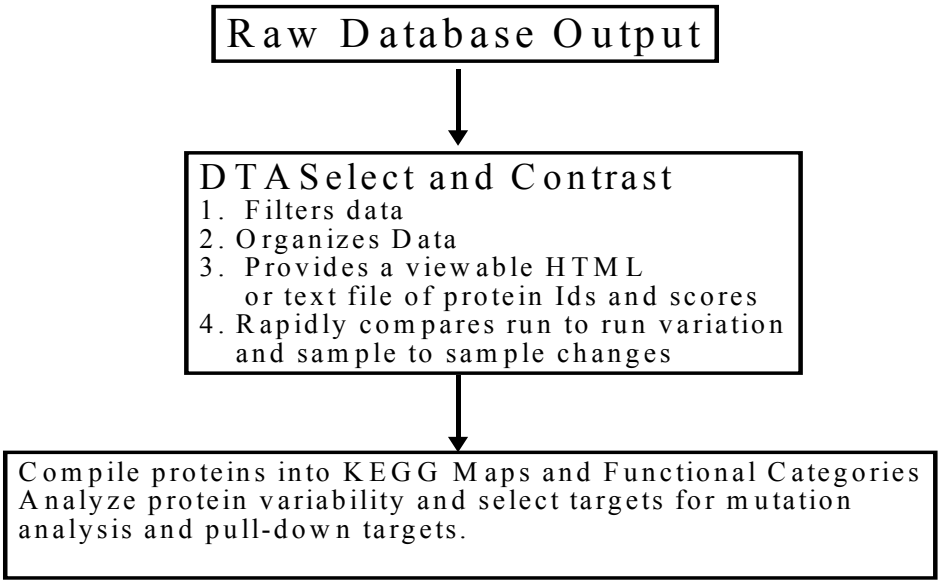


Figure 9.7

Phosphoglycerate Kinase (PGK)
Meas. - 44647.714 Da
Calc. - 44647.863 Da

ES-FTICR-MS of Yeast FPLC fraction #43

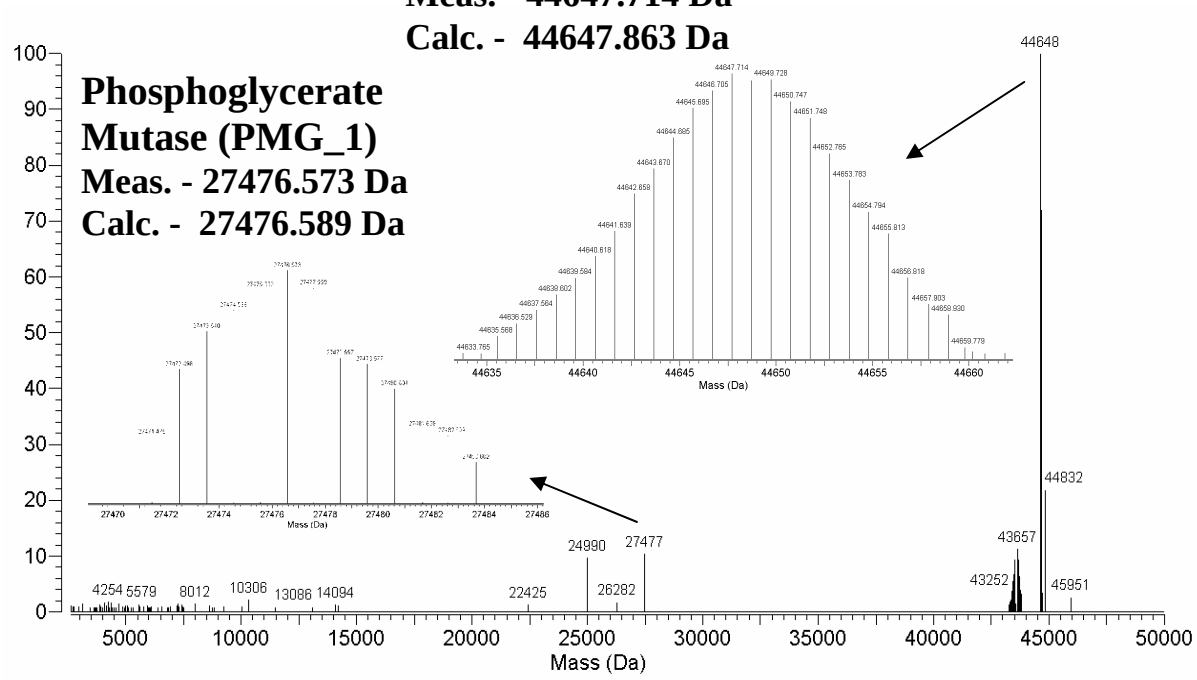


Figure 9.8

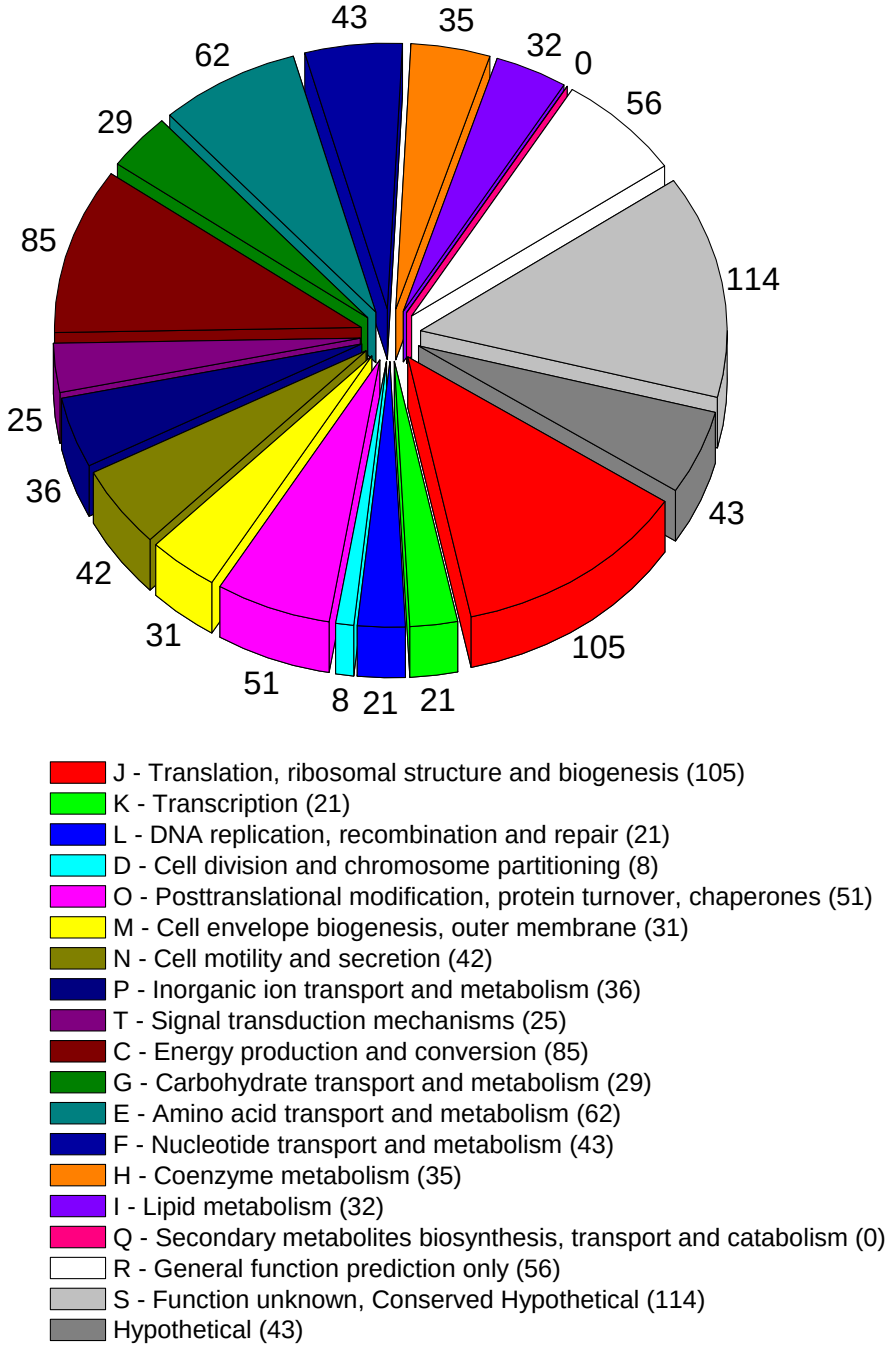

43
35
32
0
56
114
43
105
21
21
8
51
31
42
36
25
85
29
62
J - Translation, ribosomal structure and biogenesis (105)
K - Transcription (21)
L - DNA replication, recombination and repair (21)
D - Cell division and chromosome partitioning (8)
O - Posttranslational modification, protein turnover, chaperones (51)
M - Cell envelope biogenesis, outer membrane (31)
N - Cell motility and secretion (42)
P - Inorganic ion transport and metabolism (36)
T - Signal transduction mechanisms (25)
C - Energy production and conversion (85)
G - Carbohydrate transport and metabolism (29)
E - Amino acid transport and metabolism (62)
F - Nucleotide transport and metabolism (43)
H - Coenzyme metabolism (35)
I - Lipid metabolism (32)
Q - Secondary metabolites biosynthesis, transport and catabolism (0)
R - General function prediction only (56)
S - Function unknown, Conserved Hypothetical (114)
Hypothetical (43)

Table 9.2 Peptides generated by trypsin digest of Sml1p-histag protein

Measured Mass*	Tryptic Fragment	Residues	Calculated Mass*	Mass Diff.
1864.780 Da	1	1-16	1864.779 Da	0.001 Da
1260 (by MALDI-MS)	2	17-27	1258.608	*1.392*
953.463	3	28-35	953.464	0.001
3705.765	4-5	36-70	3705.749	0.016
4347.999	4-6	36-75	4348.046	0.047
1413.646	7	76-87	1413.630	0.016
736.383	8	88-94	736.375	0.008
2132.007	7-8	76-94	2131.995	0.012
810.373	9	95-101	810.369	0.004
1068.468	10	102-110	1068.454	0.014
1860.809	9-10	95-110	1860.813	0.004
(927) – not observed	11	111-117	927.396	

* Monoisotopic masses

Table 9.3 Landmark papers in large-scale proteome analysis by mass spectrometry

Author	Year	Species	Growth/Frac(a)	Seperations Methods	MS Methods	# Ids (b)
Shevchenko et al.	1996	*Saccharomyces cerevisiae*	1/1	2D-PAGE	MALDI PMF / Nano-ES-MS/MS	150
Washinger et al.	2000	*Mycoplasma genitalium*	1/1	2D-PAGE (4 zoom gels)	MALDI PMF	158
Langen et al.	2000	*Haemophilus influenzae*	1/1	2D PAGE (multiple zoom gels), multiple LC	MALDI PMF/AAC	502
Washburn et al.	2001	*Saccharomyces cerevisiae*	1/3	MudPIT (Serial 2D SCX-RP)	Nano-ES-MS/MS	1484
VerBerkmoes et al.	2002	*Shewanella oneidensis*	1/4	Serial 2D SCX-RP / 1D RP (MMS)	Capp-ES-MS/MS	868
Lipton et al.	2002	*Deinococcus Radiodurans*	15/1	2D SCX-RP (MMS)	Nano-ES-MS (AMT)	1,910
Koller et al.	2002	*Oryza Sativa*	1/3	2D-PAGE-(RP) / MudPIT	Nano-ES-MS/MS	2,528
Mawuenyega et al.	2002	*Caenorhabditis elegans*	1/2	Switching 2D SAX-RP	Capp-ES-MS/MS	1616
Peng et al.	2002	*Saccharomyces cerevisiae*	1/1	Off Line 2D SCX-RP	Nano-ES-MS/MS	1504

a) Number of different growth conditions and crude centrifuge fractionations

b) The reported # protein ids (Varied in criteria and redundant vs non-redundant)

PMF-Peptide Mass Fingerprinting
PAGE-Polyacrylamide Gel Electrophoresis
AAC-Amino Acid Composition
MudPIT- Multidiminsional Protein Identification Technology
AMT-Accurate Mass Tag
MMS-Multiple Mass Range Scanning or Gas Phase Fractionation
Capp- ES flow rates over 1ul/min
Nano- ES flow rates less than 1ul/min
SCX- Strong Cation Exchange
RP-Reverse Phase
SAX-Strong Anion Exchange

Table 9.4 Comparison of cysteine-containing proteins in different organisms

S. oneidensis (bacterium)

Average # Cysteine residues per protein= 3.159

Cysteines per protein	Proteins	%
0	1000	19%
1	950	18%
2	830	16%
3	646	12%
4 or more	1751	31%
Total Proteins	**5177**	

S. cerevisiae (yeast)

Average # Cysteine residues per protein=6.268

Cysteines per protein	Proteins	%
0	627	9%
1	663	9%
2	668	10%
3	645	9%
4 or more	4332	62%
Total Proteins	**6935**	

E. Coli (bacterium)

Average # Cysteine residues per protein=3.725

Cysteines per protein	Proteins	%
0	597	14%
1	606	14%
2	634	15%
3	574	14%
4 or more	1829	43%
Total Proteins	**4240**	

A. Thaliana (plant)

Average # Cysteine residues per protein=7.892

Cysteines per protein	Proteins	%
0	1328	5%
1	1794	7%
2	2080	8%
3	2188	8%
4 or more	18426	71%
Total Proteins	**25816**	

Table 9.1 Performance Factors for Different Mass Analyzers

Mass Analyzer	Most Comm. Ioniz. Mode	Ion Separation	Resolving Power (FWHM)	Mass Accuracy	Mass/charge Range
Quad.	ESI	Electronic Band Pass Filter	1000-2000	0.1 Da	500-3000 Da
TOF	MALDI ESI	Flight time	500-1000 (linear) 2,000-10,000 (ref, QTOF)	0.1 Da (linear) 0.001 Da (ref., QTOF)	500 – 1,000,000
Sector	ESI	Magnetic field (ion momentum)	5,000-100,000	0.0001 Da	1,000 – 15,000
QIT	ESI MALDI	rf/dc Voltage	1,000 – 2,000	0.1 Da	500 – 3000 Da
FTICR	ESI	rf frequency	5,000 – 5,000,000	0.0001 Da	200 – 20,000 Da