EXPERIMENT 4

Analysis of Protein Complexes

High-sensitivity Liquid Chromatography Coupled with Tandem Mass Spectrometry

Proteomics, or the large-scale analysis of proteins, has emerged as a result of the genome sequencing projects. The development of methods and instrumentation for automated tandem mass spectrometry (MS/MS) in combination with microcapillary liquid chromatography has dramatically increased the sensitivity and speed to identify proteins. Identification of proteins is now achieved with greater sensitivity than silver-stained gels. Furthermore, automated computer algorithms have supplemented manual interpretation of MS/MS spectra. Sophisticated programs are used to correlate tandem mass spectra of peptides with genomic sequences for rapid and conclusive protein identification. Because each mass spectrum is compared to all sequences in the database, the result is the unbiased identification of proteins. Using this approach, novel and unexpected proteins are being identified in a growing list of protein complexes, subcellular locations, and total cellular proteomes.

MS/MS is an extremely powerful technique for the sequence analysis of peptides in complex mixtures. In tandem mass spectrometric sequencing of a peptide, information about the sequence of the peptide is contained in the product ion or tandem mass spectrum. MS/MS experiments can be performed in a variety of mass spectrometers. Spatially separate mass analyzers (such as a triple quadrupole, quadrupole-time of flight [QTOF], or time of flight/time of flight [TOF/TOF]) and analyzers that separate ions in time (such as ion traps or Fourier transform ion cyclotron resonance [FTICR]) are predominantly used to sequence peptides by mass spectrometry. For peptide sequencing, the fragmentation spectrum is obtained by selecting a positively charged precursor ion from the survey or precursor mass spectrum, isolating the selected ion and ejecting all other ions, and finally fragmenting the precursor ion by collision induced dissociation (CID) to generate product ions. Low energy fragmentation of ionized tryptic peptides occurs primarily at the amide bonds along the peptide backbone, generating a series of fragmentation or product ions. Several models have been proposed that describe the fragmentation chemistry, including the mobile proton model and the pathways in competition model (Dongré et al. 1996; Wysocki et al. 2000; Paizs and Suhai 2005). The differences between product ions and the residue mass of the amino acids are used to determine the sequence of the peptide (see Experiment 8). The fragmentation patterns are also unique signatures for proteins in the sample that can be used by database search algorithms to identify the proteins.

Using electrospray ionization (ESI), liquid chromatography can be coupled directly with MS/MS (Fig. 1). This combination allows the mass spectrometer to analyze complex peptide mixtures. Peptides are separated by their chemical properties in the chromatography step and then separated by their *m/z* value in the mass spectrometer with subsequent sequence analysis by MS/MS. Computer methods allow data-dependent acquisition of tandem mass spectra in real time and the automated acquisition of MS/MS spectra. The instrument is programmed to choose which precursor ions to select or ignore for MS/MS analysis. Typically, the most abundant ions are selected for fragmentation, because they will usually produce MS/MS spectra with the strongest signals. During

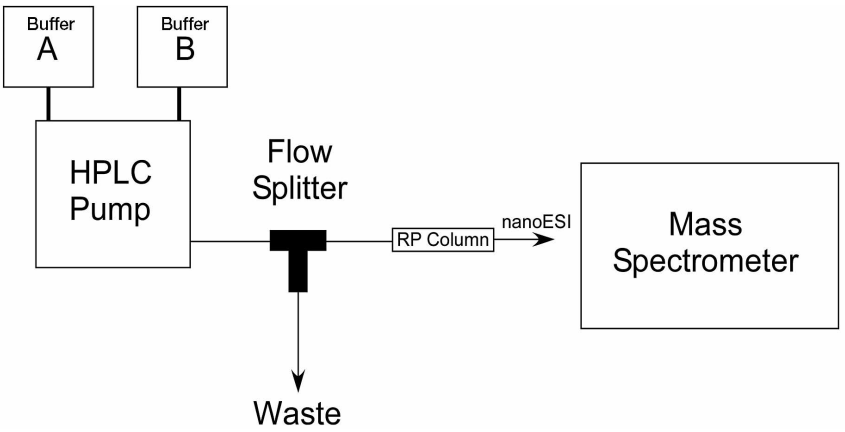


FIGURE 1. Diagram of the principal components of a nanoLC-MS/MS mass spectrometry system for proteomics.

a coupled liquid chromatography/tandem mass spectrometry analysis, it is possible to acquire a large number of MS/MS spectra: >4 MS/MS spectra every second using ion trap mass analyzers.

The genome sequencing projects have generated a wealth of information, including the theoretical sequences of all the proteins in a large number of organisms. A large number of computer programs have been developed to match uninterpreted tandem mass spectra to sequences in protein or nucleic acid databases. The programs compare the experimental spectra with theoretical spectra generated from protein databases and generate a list of peptides and proteins in the sample. Experiment 8 trains students to use specific search engines to process and analyze MS/MS data for identifying proteins and posttranslational modifications. Most importantly, the experiment instructs students in how to evaluate the accuracy of the peptide and protein identifications that the programs return.

In this experiment, microcapillary liquid chromatography is coupled to MS/MS (LC-MS/MS) to analyze complex protein mixtures. The goal is to comprehensively identify the proteins in the sample. In earlier experiments, protein complexes or mixtures were reduced and alkylated to denature the proteins and to derivatize the cysteine residues to prevent the formation of disulfide bonds. The protein complexes were digested with trypsin to cleave the proteins into peptides. This experiment analyzes those peptide mixtures using data-dependent microcapillary LC-MS/MS.

Microcapillary reversed-phase HPLC columns will be constructed for separating complex peptide mixtures. The microcapillary columns will be connected to an HPLC pump coupled to the tandem mass spectrometer using an ESI source. The HPLC pump and an ion trap mass spectrometer will be programmed for running a data-dependent LC-MS/MS experiment on the sample. The mass spectrometer will be first programmed to perform a precursor scan to measure the *m/z* values and intensities of ions eluting from the RP column. Second, the instrument will then be programmed to individually fragment (MS/MS) the most abundant ions using the data collected in the precursor scan. The ions to be fragmented are "dependent" upon the information in the precursor scan. The cyclic process of a precursor scan followed by a series of fragmentation or MS/MS scans is programmed to constantly repeat during the entire LC gradient.

After adjusting the HPLC's mobile phase flow rate through the column, the liquid chromatography and mass spectrometer will be evaluated by loading a control peptide, angiotensin I, and running an LC-MS/MS experiment. After verifying the performance and sensitivity of the system, the experimental protein samples from other experiments will be analyzed by LC-MS/MS. In all of the experiments, the samples will be manually loaded onto the microcapillary column and analyzed. When the LC-MS/MS is finished, the data files will be transferred to a computer system running database search programs and the data processed and analyzed to generate a list of peptides and proteins (Experiment 8).

PROTOCOL 1

Making Microcapillary HPLC Columns

In ESI of peptides, an acidic aqueous solution that contains the peptides flows through a small-diameter needle. A high, positive voltage is applied to the needle to produce a Taylor cone as the solution exits the needle. Small droplets of solution are generated by the Taylor cone, which contains the peptide analyte. Protons from the acidic solution give the droplets a positive charge, causing them to move from the needle to the negatively charged instrument. During the movement, evaporation reduces the size of the droplets and the droplets split into smaller and smaller, charged droplets. The evaporation and splitting process eventually causes the peptides to desorb into the gas phase as protonated peptides. The ionized peptides can be directed and manipulated by the mass spectrometer's electric or magnetic fields.

Reversed-phase (RP) chromatography fractionates peptides and proteins based on the interaction between hydrophobic patches on the surface of biomolecules and nonpolar alkyl chains bonded covalently to the surface of the stationary phase. RP chromatography is used with ESI because RP's acidic aqueous and polar mobile phases are compatible with ESI. In addition, in-line RP-HPLC is useful for desalting peptides before ESI without the need for off-line desalting steps. RP-HPLC tends to focus peptides from dilute samples into narrow chromatographic bands, which enhances sensitivity.

The use of an acidic solution tends to protonate all the available basic residues in a peptide. These include the amino-terminal amine and the basic side groups of lysine (K), arginine (R), and histidine (H). As a result, multiply-charged protonated peptides are observed when a peptide contains a K, R, or H residue. Since trypsin cleaves peptides at the carboxy-terminal side of R and K, those tryptic peptides tend to be doubly charged (amino-terminal amine and the K or R residue). Tryptic peptides containing internal basic residues (i.e., internal H, R-P, K-P, or cleavage sites that trypsin missed) are typically more highly charged peptides (e.g., +3, +4, etc.).

The sensitivity of ESI-LC-MS/MS is inversely proportional to the flow rate. The low flow rates (<0.5 µL/min) of microcapillary HPLC are orders of magnitude more sensitive than standard RP-HPLC columns with flow rates of 50 µL/min or more. For the successful analysis of low femtomole (nanogram) amounts of peptides, microcapillary HPLC is required.

This protocol describes the construction of a pulled microcapillary column containing an approximately 3-µm orifice at the end of a fused-silica capillary (FSC) (Fig. 2). A laser-based micropipette puller is used to pull the capillary and create the restriction. The restriction prevents packing material from passing through the column but allows liquid to flow through. The pulled tip also functions as the emitter tip for ESI. For ESI-LC-MS/MS, the integrated column and emitter needle are connected to an electrospray interface. In Protocol 2, the FSC is packed with RP packing material. These protocols will be used later (Experiment 6) to construct microcapillary columns for multidimensional peptide separations for 2D LC-MS/MS experiments or MudPIT.

If you prefer not to construct individual columns or do not have access to a laser puller, packed 1D and 2D fused-silica microcapillary columns can be purchased from commercial vendors.

FIGURE 2. Different size RP-HPLC columns used for LC-MS/MS experiments. For RP-HPLC, a 4.6-mm, 1-mm, and 0.1-mm inner diameter column can be used. For high-sensitivity proteomics LC-MS/MS experiments, the 0.1-mm or 100-µm fused silica capillary (FSC) column, shown at the bottom of the figure, is typically used.

MATERIALS

CAUTION: See Appendix 11 for appropriate handling of materials marked with <!>.

Reagents

Methanol <!>

Equipment

Alcohol lamp
Fused silica capillary scribes (Chromatography Research, 205312)
100 µm ID x 365 µm OD Fused silica capillary (FSC) tubing (PolyMicro Technologies)
Laser-based micropipette puller (e.g., P-2000 Sutter Instruments)

PROCEDURE

1. Cut ~18 inches of 100 µm ID x 365 µm OD FSC with a cleaving tool (Fig. 3A,B).
2. Burn a 1–2-inch window in the middle of the capillary with an alcohol lamp. Slowly rotate the FSC over the lit flame (Fig. 4).

   Only heat the FSC until the polyimide coating has been charred. Excessive heating will damage the capillary.

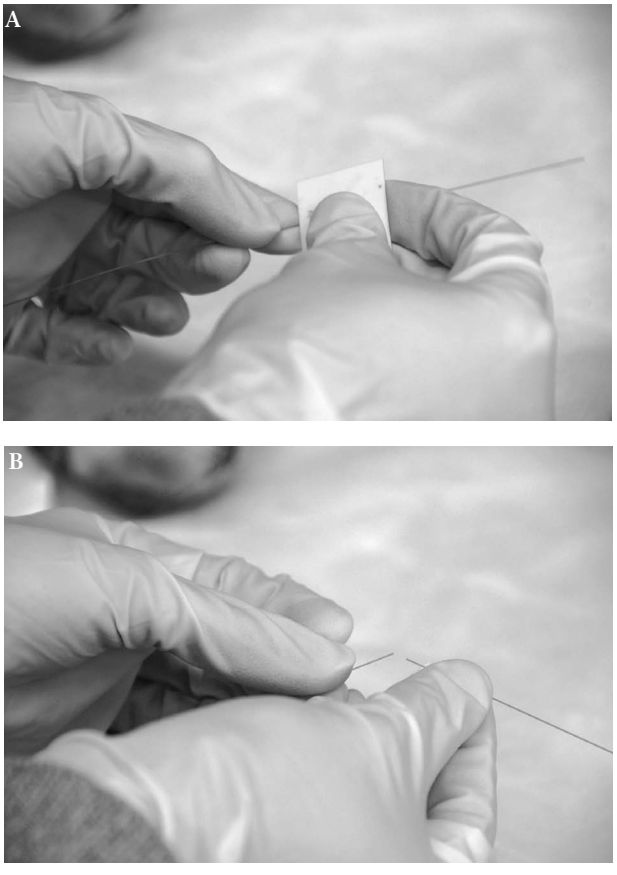


FIGURE 3. Cleaving fused silica capillary (FSC). (A) A silica scribe is used to lightly score the FSC. (B) After the plastic coating and quartz glass are scored with the silica scribe, the FSC will break into two pieces with a light force.

FIGURE 4. Using an alcohol lamp to burn off the polyimide plastic coating from the FSC. The FSC is briefly held in the flame until the coating turns black.

3. Using a Kimwipe wet with methanol, gently wipe the charred polyamide coating off to expose the clear quartz (Fig. 5A,B).

   Be careful as the quartz is very fragile and breaks easily. Be sure to remove all charred pieces off the quartz.

4. Place the capillary in the Sutter P-2000 needle puller and align the capillary using the grooves (Fig. 6A,B). Tighten the FSC down with the clamps before lowering the lid. Use the program on the next page to pull two columns.

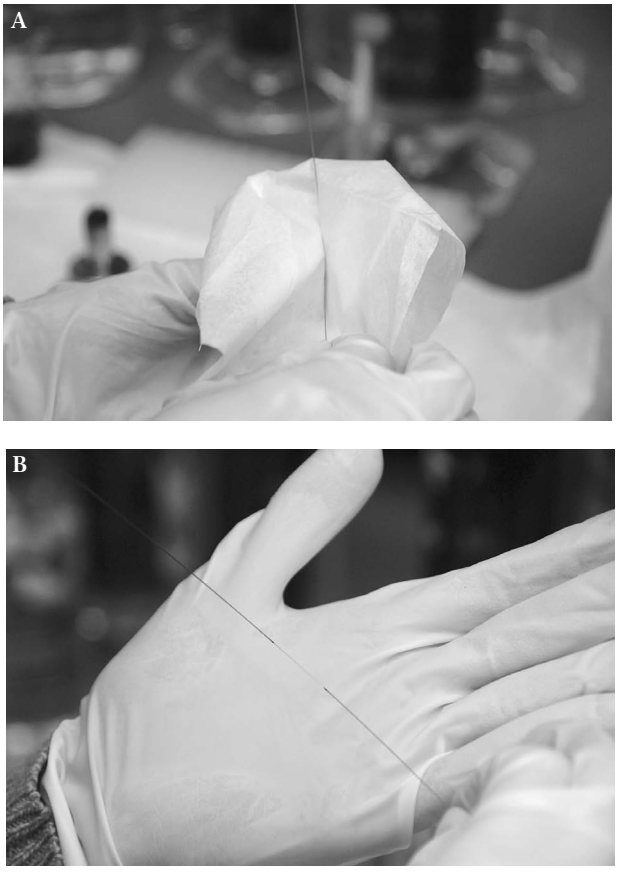


FIGURE 5. (A) Using a methanol-soaked Kimwipe, gently wipe clean the charred black plastic coating from the FSC to expose the quartz glass. (B) Exposed quartz glass of an FSC ready for pulling into micro-capillary HPLC columns using the laser puller.

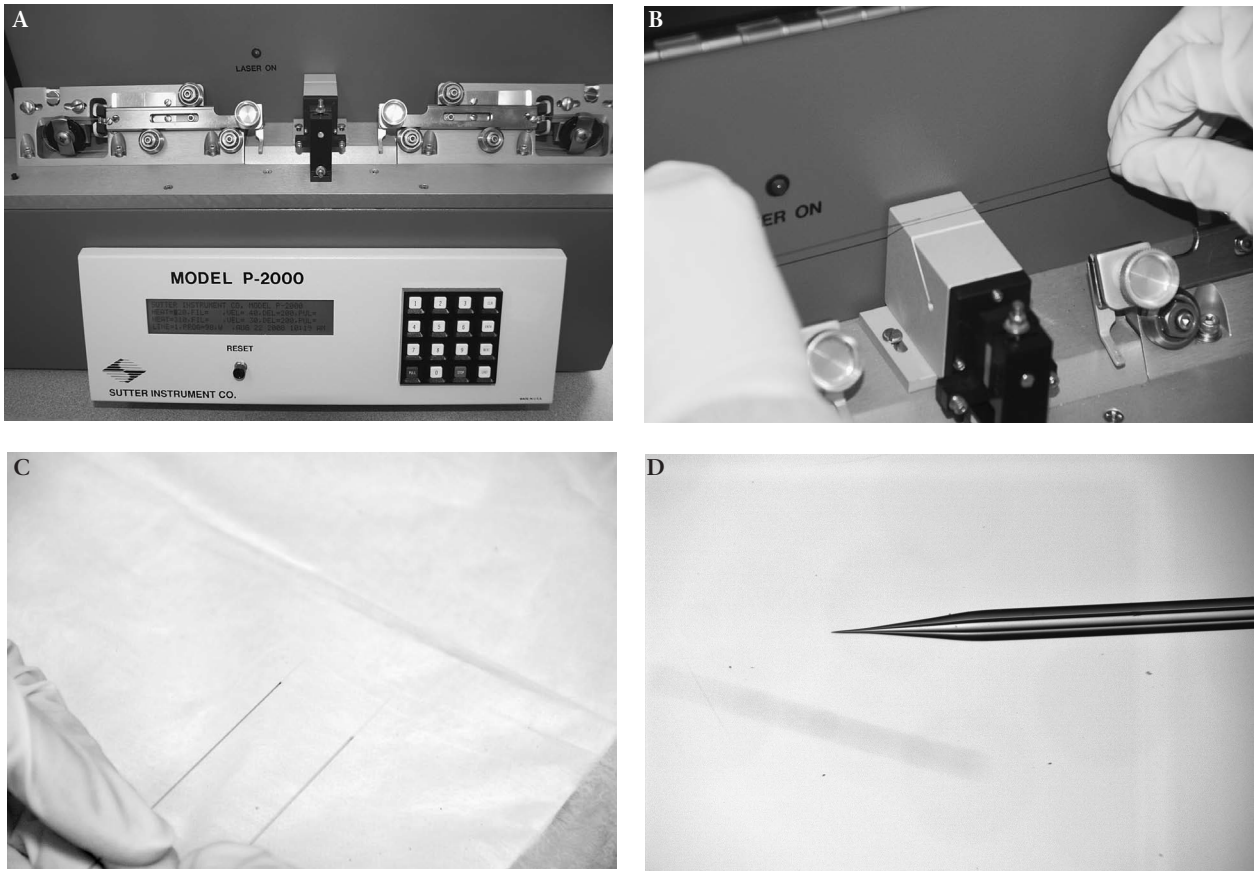


FIGURE 6. (A) Sutter laser micropipette puller for making microcapillary HPLC columns for nanoLC-MS/MS experiments. (B) Placing the exposed FSC in the laser puller. The exposed, cleaned quartz glass is centered in the middle of the laser. (C) Two pulled, empty FSC microcapillary columns that are ready for packing. (D) Close-up of the tip created in the FSC using the laser puller. The symmetrical tip is necessary for generating a stable nanoelectrospray.

Heat	Velocity	Delay
320	40	200
310	30	200
300	25	200
290	20	200

If the red laser light flashes and the puller's jaws separate, you should have a successful pull. If the light does not flash, press the stop button and try to align the FSC again. A successful pull will create two fused silica capillary columns, each with a restriction at one end (Fig. 6C,D). The parameters used for the puller program may need to be adjusted to account for instrument variation.

PROTOCOL 2

Packing Microcapillary FSC Columns

MATERIALS

CAUTION: See Appendix 11 for appropriate handling of materials marked with <!>.

Reagents

5% Acetonitrile <!>, 0.1% formic acid <!>
Methanol <!>
Reversed-phase resin (Phenomenex Synergi 4u Hydro-RP 80A)

Equipment

Glass vials with screw caps, 1.8-mL (Chromatography Research 123315/309925)
Micro stir bars (VWR 58948-069)
Microcapillary column (from Protocol 1)
Pneumatic packing vessel (homemade or commercially available from New Objectives or Next Advance, Inc.)
Ruler
Sonicator <!>
Stereomicroscope
Stir plate
Vortexer

PROCEDURE

1. Place a microstir bar into a 1.8-mL glass vial. Add 0.6 mL of methanol and 8 mg of RP packing material to the glass vial.

   The thickness of the slurry will determine how quickly the column will pack.

2. Vortex to resuspend the packing material and sonicate for 5 minutes to prevent aggregation of the particles.

3. Transfer the slurry to a pneumatic loading vessel and place the loading vessel on a stir plate (Fig. 7A,B,C). Turn on the stir plate to keep the packing material suspended. Secure the lid to the loading vessel by tightening the bolts that attach the lid to the base.

   WARNING: Always wear safety glasses when packing FSC columns. The columns are being packed under very high pressure. Improperly seated columns can be ejected from the loading vessel at high velocity.

4. Measuring from the frit end, place a mark on the empty microcapillary column showing the desired packing height.

   For a 100 μm ID x 365 μm OD pulled microcapillary column, 9 cm of RP material is packed in the column.

5. Feed the empty microcapillary column down through the Vespel ferrule in the Swagelok fitting on the lid of the pneumatic loading vessel until the end reaches the bottom of the vial. Pull the column up so that the capillary rests just off the bottom of the glass vial and stir bar, and tighten the ferrule to secure the column (Fig. 8).

6. Apply pressure to the loading vessel by first setting the regulator on the high-pressure helium gas cylinder to 500–1000 psi, then opening the three-way valve.

7. If the column is long enough, place the fritted end of the column under a stereomicroscope to observe the packing. Be extremely careful as the column tip is very fragile.

   A steady stream of packing material should be seen flowing into the capillary. For pulled columns, the tip may need to be opened slightly when packing the column. To do this, gently score the opening of the column (Fig. 9A,B). In a smooth upward motion, glide the capillary

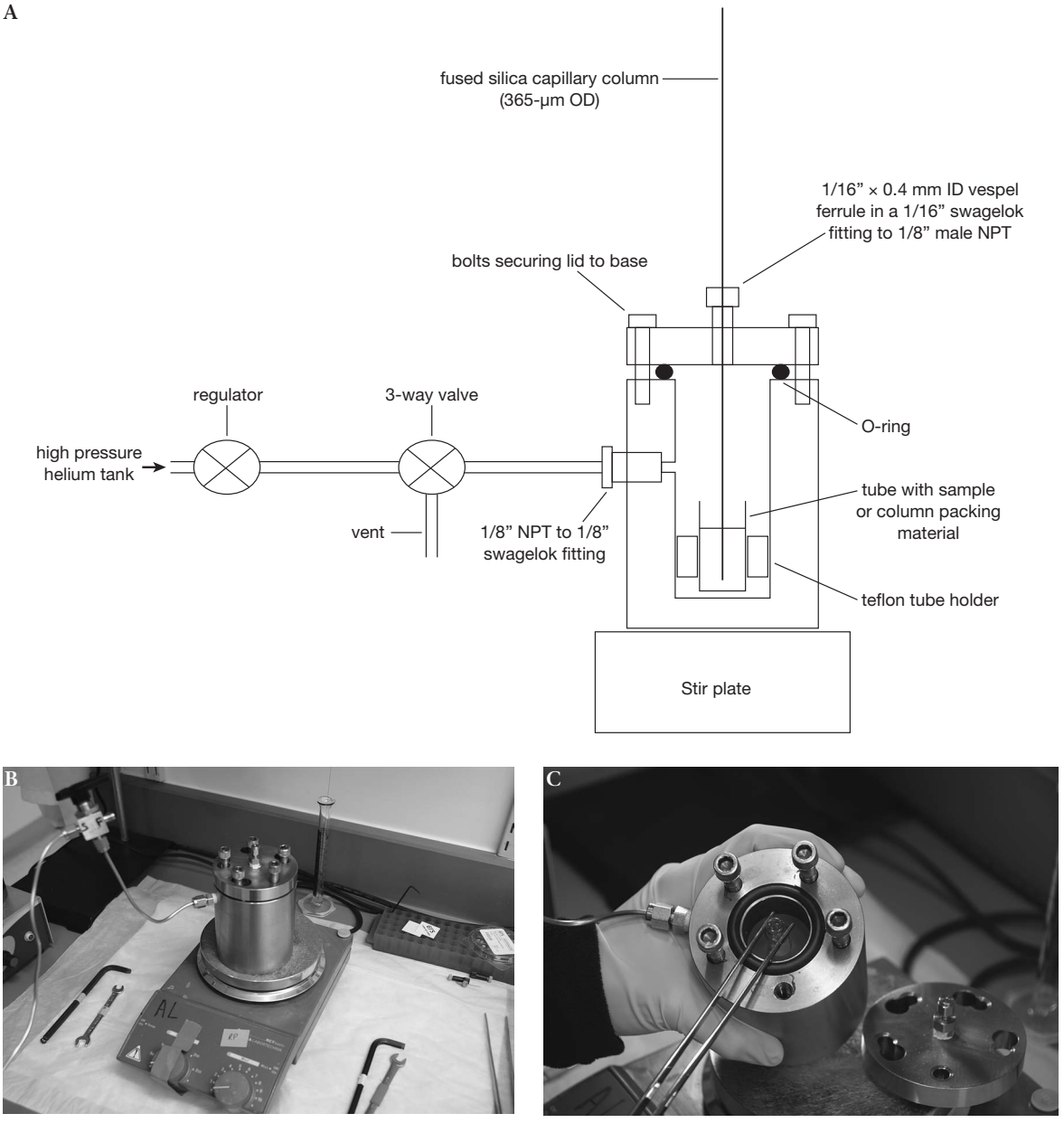


FIGURE 7. (A) Diagram of the pneumatic loading device, "loading bomb," for packing microcapillary FSC HPLC columns. (B) Actual loading bomb on a stir plate. (C) Using tweezers to insert a glass vial containing the methanol-reverse phase resin slurry for packing the FSC column. The vial contains a microstir bar for keeping the slurry in suspension while the column is packing.

cleaving tool along the side of the FSC. Never score directly on top of the column since it will destroy the tip for electrospray (Fig. 10).

8. When the column has been packed to the mark, slowly turn off the pressure to the vessel at the three-way valve.

   Slowly releasing the pressure prevents the packing material from unpacking.

9. Replace the vial containing the slurry with a 1.5-mL microcentrifuge tube filled with 5% acetonitrile, 0.1% formic acid. Wash the column for 10 minutes using the loading vessel.

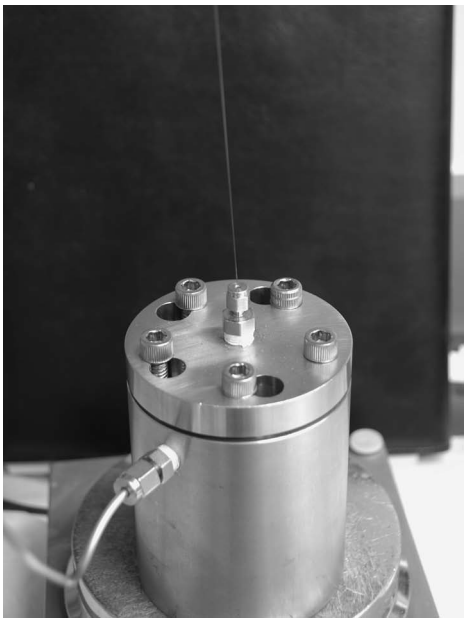


FIGURE 8. Empty FSC column inserted into the loading bomb for packing.

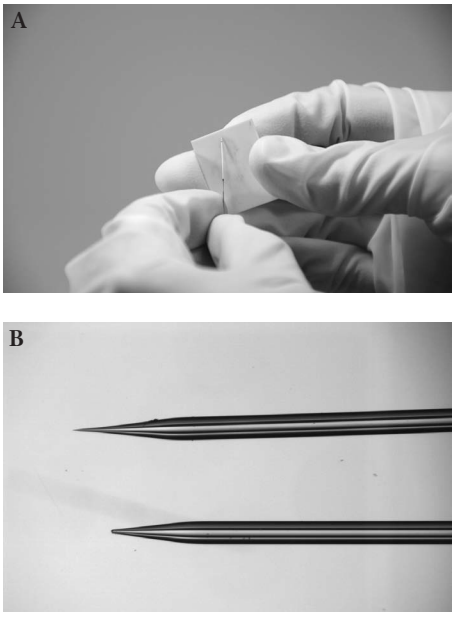


FIGURE 9. (A) Using a silica scribe to "open" the pulled FSC column. To get the columns flowing, it is typically necessary to lightly score the side of the tip. (B) Examples of unscored and properly scored FSC tips.

10. Store the column in 5% acetonitrile, 0.1% formic acid until ready to use. Columns can be stored indefinitely at room temperature in this solution.

The column must be completely submerged in 5% acetonitrile, 0.1% formic acid to prevent the packing resin from drying out.

11. Before using for any biological samples, run a blank HPLC gradient across the column to condition it.

Conditioning the column is important to firmly pack the resin. Multiple blank HPLC gradients may be required to obtain a reproducible baseline. If high sensitivity applications are planned, load 0.1 pmol of angiotensin peptide onto the new column and run an HPLC gradient (see Protocol 3). Angiotensin binds to nonspecific binding sites in the column and minimizes non-specific, irreversible binding of sample peptides.

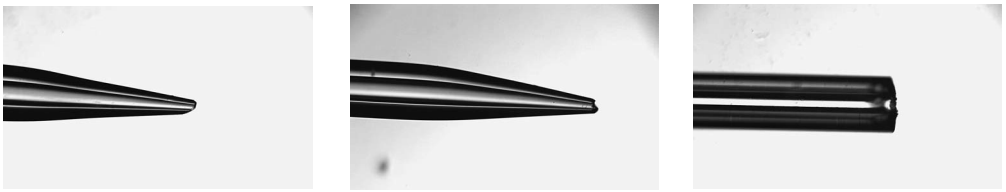


FIGURE 10. Examples of FSC column tips that have been excessively scored to get the packing slurry flowing.

PROTOCOL 3

Microcapillary RP-HPLC Coupled to ESI-Mass Spectrometry

This protocol describes the RP-LC-ESI-MS assembly and LC-MS/MS process.

MATERIALS

CAUTION: See Appendix 11 for appropriate handling of materials marked with <!>.

Reagents

Acetonitrile <!> (for HPLC gradients; see Step 13)
Angiotensin solution (0.02 pmol/µL) (angiotensin I, Sigma-Aldrich A9650) or Trypsin-digested protein sample (see Step 4)
Solvent A (5% acetonitrile, 0.1% formic acid <!>)

Equipment

Disposable calibrated glass pipettes, 5 µL (Drummond Scientific 2-000-001)
50 µm ID x 365 µm OD FSC tubing (PolyMicro Technologies)
75 µm ID x 365 µm OD FSC tubing (PolyMicro Technologies)
Fritless microcapillary RP column (from Protocol 2)
Fused silica capillary scribe (Chromatography Research 205312)
HPLC pump (model 1200; Agilent)
Linear ion trap mass spectrometer (model LTQ; Thermo Scientific)
Nanospray ESI source (James Hill Instruments)
PEEK MicroTee (Upchurch P775)
PEEK 380 µm ID MicroTight sleeve
Pneumatic loading device

PROCEDURE

1. Prepare the RP-apparatus as follows (Fig. 11A,B).
   a. Connect the transfer line from the HPLC pump to the arm of a PEEK-restrictor MicroTee using a PEEK 380 µm ID MicroTight sleeve.
   b. Connect the PEEK-restrictor MicroTee from the center arm to a PEEK-ESI MicroTee using a 75 µm ID x 365 µm OD FSC and PEEK sleeves.
   c. Connect a 30-cm piece of 50 µm ID x 365 µm OD FSC restrictor line through the third arm of the PEEK-restrictor Tee using a PEEK sleeve.
   d. Connect a 0.025-inch OD gold wire through the center arm of the PEEK-ESI Tee and attach it to the ESI voltage source.

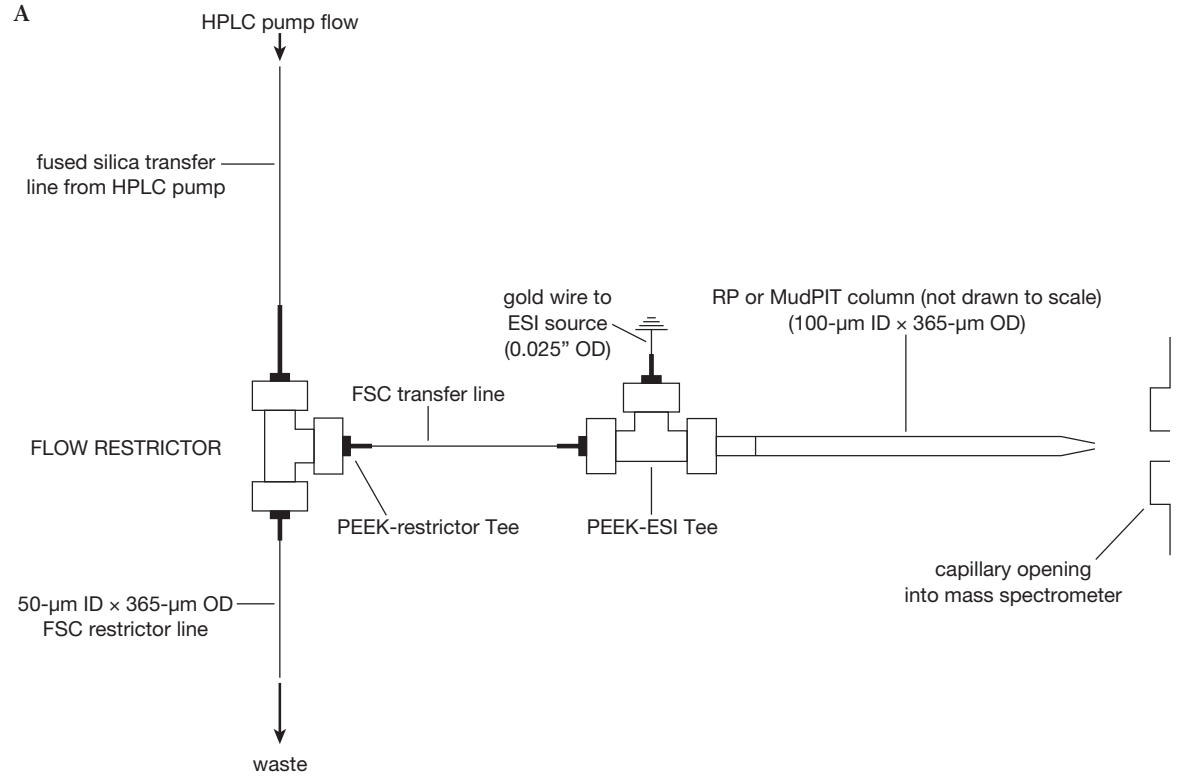


FIGURE 11. (A) Diagram of the FSC connections used to construct the nanoESI source. (B) Actual nanoESI source mounted to an ion trap mass spectrometer. The video camera is used to assist in the alignment of the FSC HPLC column tip with the mass spectrometer's heated capillary tube opening.

e. Connect the pulled microcapillary RP column to an arm of the Tee using a PEEK-ESI Tee.

The pulled microcapillary HPLC RP column assembly is mounted to an x-y-z manipulator at the entrance of the mass spectrometer (Figs. 11 and 14). The manipulator allows fine adjustment of the column tip with respect to the mass spectrometer's capillary opening (see Step 12). The pulled column tip is extremely fragile. Avoid letting the tip strike a solid surface. A 2.2 kV voltage is applied to the gold wire during ESI.

2. Set the HPLC pump to 100% solvent A with a flow rate of 200 µL/min.

FIGURE 12. Measuring the flow rate through the microcapillary FSC HPLC column using a calibrated glass capillary pipette.

3. Measure the flow rate through the column for 1 minute using a 5-µL calibrated glass capillary pipette (Fig. 12). Trim the 50 µm ID x 365 µm OD restrictor line using the capillary cleaving tool until a flow rate of 0.2 to 0.5 µL/min through the RP column is obtained (see Fig. 11A).

   Measurement of the flow rate and adjustment of the split line may have to be repeated several times until the target flow rate is achieved. Several vendors now offer specialized HPLC pumps with low mobile phase flow rates (<1 µL/min) that generate reproducible gradients. The nanoflow pumps eliminate the need for the external PEEK-restrictor Tee and flow splitting.

4. Place a tube containing the trypsin-digested protein sample or the angiotensin solution into a pneumatic loading vessel and tightly attach the lid of the vessel.

   To condition the microcapillary column and measure the performance of the HPLC system and mass spectrometer before analyzing unknown samples, a control sample is typically first run before analyzing one's precious biological sample. If time does not permit the control experiment, the RP column can be conditioned off-line and the unknown sample loaded onto the column. Angiotensin I (DRVYIHPFHL) has a monoisoptic and average mass of 1295.68 and 1296.49 gm/mol, respectively. During ESI, it predominantly forms a +3 ion at approximately 433 *m/z*. Monitoring the retention time (RT), signal intensity, and resolution of the angiotensin peptide (and background noise) is used to check the performance of the chromatography and mass spectrometry system prior to running unknown samples. Alternative control peptides or trypsin-digested control proteins (e.g., BSA) can be used to evaluate the system.

5. Disconnect the column from the PEEK-ESI Tee assembly and insert the open end of the column into the top of the loading vessel until the capillary reaches the bottom of the sample tube. Pull the column up slightly off the bottom of the sample tube.

   Leaving a small gap between the capillary and the bottom of the tube reduces the probability of any solid particulates at the bottom of the tube clogging the capillary column.

6. Secure the column to the bomb by tightening the compression nut on the Swagelok fitting at the top of the loading vessel.

   Tightening the Swagelok fitting's nut compresses the Vespel ferrule around the FSC.

7. Apply pressure to the loading vessel by first setting the regulator on the helium gas cylinder to 500–1000 psi, and then opening the three-way valve.

8. Measure the volume loaded using a 5-µL calibrated glass pipette to collect the displaced volume from the end of the column (see Fig. 13).

   For the angiotensin I control sample, load 5 µL (0.1 pmol) of the angiotensin solution. For unknown samples, the amount of sample to load can be problematic. If low amounts of unknown sample are available (silver-stained bands), loading the entire sample is recommended. For more abundant samples (Coomassie-stained bands), we typically load a fraction of the sample.

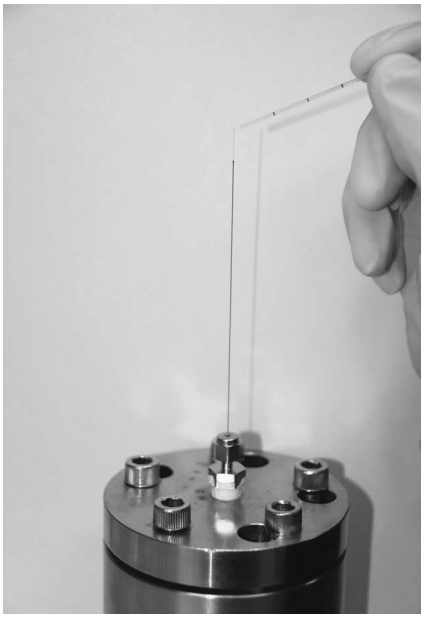


FIGURE 13. Using the bomb to load a sample onto the microcapillary FSC HPLC column. A microcentrifuge tube with the sample is placed in the loading bomb and the column is inserted. When the bomb is pressurized, a calibrated glass capillary pipette is used to measure the amount of sample loaded onto the column.

9. Release the pressure in the loading vessel using the three-way valve once the sample is loaded.
10. Reinstall the column in the union.

    The RP column will be plumbed for performing microspray mass spectrometry on the angiotensin or an unknown sample as it elutes from the column.
11. Start the HPLC flowing at 200 µL/min with 100% Solvent A and re-check the RP column flow rate using a 5-µL calibrated glass pipette.
12. Carefully position the pulled microcapillary HPLC column tip at the entrance of the mass spectrometer (Fig. 14). Use a camera monitor to assist in positioning the column at the optimal position.

    The pulled microcapillary column tip is centered 1–5 mm from the orifice of the capillary opening into the mass spectrometer using the x-y-z manipulator with the aid of the closed circuit monitor (Fig. 14). The pulled column tip is extremely fragile. Do not let the tip strike a solid surface.
13. Program the HPLC using the following as an example: 60-minute gradient from 0% to 40% acetonitrile, and a 10-minute gradient from 40% to 60% acetonitrile (see Appendix 7).

    There are many gradient variations and RP buffers that can be used.

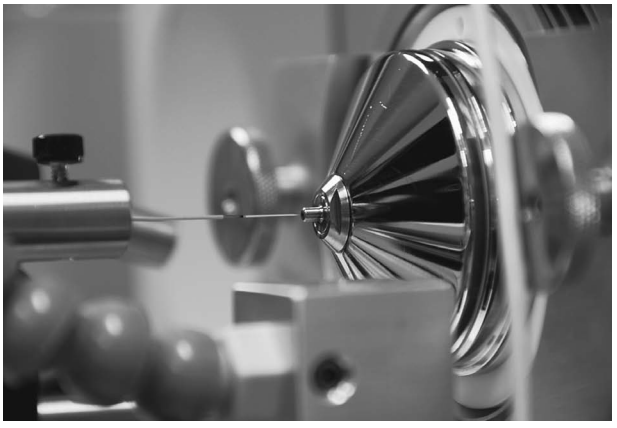


FIGURE 14. Position of the microcapillary HPLC column tip relative to the opening of the mass spectrometer's ion transfer tube.

14. Using a Kimwipe, carefully wick away any large drops from the column tip. Start the mass spectrometer and the HPLC gradient and collect MS/MS data on peptides as they elute from the RP column.

A voltage of ~2.2 kV is applied to the gold wire during ESI. For the ThermoFisher LTQ mass spectrometer, data-dependent acquisition of tandem mass spectra is programmed through the instrument's Xcalibur software. Typical data acquisition settings consist of a continual cycle beginning with one MS scan with an *m/z* scan range of 300–2000, which records all the *m/z* values of ions present at that moment in the gradient, followed by five data-dependent MS/MS scans. The MS/MS scans fragment the five most abundant ions recorded in the first MS scan. Dynamic exclusion is activated to improve protein identification capacity by avoiding the repeated fragmentation of abundant ions. Each instrument manufacturer has different options for configuring tandem mass spectra collections.

15. When the ESI-LC-MS/MS run is complete, turn off the ESI voltage.

Examine the acquired data for angiotensin peptide to confirm the instrument is properly functioning before analyzing unknown samples. The angiotensin retention time is typically 33–36 minutes. For the LTQ instrument, a signal intensity >10E7 for a 433 ion should be observed. These values are typical for a properly functioning HPLC and linear ion trap LTQ mass spectrometry system. As described in Experiment 8, the acquired data file can be searched against a protein database containing the angiotensin I sequence to confirm the MS/MS. If these values or protein identifications are not observed, a number of steps can be taken, including cleaning the ESI source, checking the HPLC flow rates, and re-calibrating and retuning the mass spectrometer. Other HPLC and mass spectrometry systems will have different RT and signal intensities. Practical experience running an LC-MS/MS system is the best way to learn how to operate the instrumentation and to troubleshoot problems. There is no substitute for practical, hands-on experience.

16. Equilibrate the RP column for 10 minutes at 100% Solvent A before loading unknown samples from other experiments in this manual.

17. Load and analyze the unknown sample onto the conditioned and tested microcapillary RP column and mass spectrometry system using Steps 4–15.

For LC-MS/MS analysis of a purified protein complex, Steps 4–15 are repeated, except that the unknown, trypsin-digested biological samples are now loaded onto the RP column and analyzed by LC-MS/MS. Because of possible carryover on the RP column, it is common to run a blank between unknown samples to check that the tryptic-digested proteins have completely eluted.

18. Process and analyze the acquired mass spectrometry data file as described in Experiment 8 to identify the peptides and proteins.

REFERENCES

Dongré A.R., Jones J.L., Somogyi A., and Wysocki V.H. 1996. Influence of peptide composition, gas-phase basicity, and chemical modification on fragmentation efficiency: Evidence for the mobile proton model. *J. Am. Chem. Soc.* **118:** 8365–8374.

Link A.J., Jennings J.L., and Washburn M.P. 2003. Analysis of protein composition using multidimensional chromatography and mass spectrometry. In *Current protocols in protein science* (ed. J.E. Coligan et al.), chapter 23, pp. 1–25. John Wiley and Sons, New York.

Paizs B. and Suhai S. 2005. Fragmentation pathways of protonated peptides. *Mass Spectrom. Rev.* **24:** 508–548.

Wysocki V.H., Tsaprailis G., Smith L.L., and Breci L.A. 2000. Mobile and localized protons: A framework for understanding peptide dissociation. *J. Mass Spectrom.* **35:** 1399–1406.

EXPERIMENT 8

Analysis and Validation of Tandem Mass Spectra*

The combination of mass spectrometry and genome-assisted data analysis has revolutionized proteomics. Large numbers of proteins and modified amino acids can now be rapidly identified from small amounts of material. Computational analysis and interpretation of the acquired mass spectrometry data play an essential role in these proteomics experiments. In Experiments 1 through 7 of this manual, precursor spectra (MS) were used to measure the masses (*m/z*) of peptides from trypsin-digested proteins. Selected peptides were fragmented by collision-induced dissociation (CID) to generate fragmentation spectra (MS/MS). During these experiments, the intensity and *m/z* values of the precursor and fragmentation ions were saved into a data file. In this experiment, the data in these files are extracted into a format that can be read by database search programs. The programs compare the experimental data collected from the mass spectrometry experiment to the theoretical masses of peptides and fragmentation ions from proteins in a database. Using various mathematical and statistical scoring approaches, the database search programs identify peptide sequences in the database that best match the experimental data. Finally, the programs generate a list of proteins predicted to be in the analyzed sample. Evaluating the search results and validating the identified proteins and peptides is a major challenge. Data analysis requires understanding the strengths and weaknesses of the search programs along with a basic understanding of how to interpret and validate peptide fragmentation spectra.

In tandem-mass-spectrometric sequencing of a peptide, information about the peptide sequence is contained in the product ion or MS/MS spectrum. Low-energy fragmentation of ionized tryptic peptides occurs primarily at the amide bonds along the peptide backbone, generating a series of fragmentation or product ions (Fig. 1). Several models have been proposed that describe the fragmentation chemistry, including the mobile proton model and the pathways in competition model (Dongré et al. 1996; Wysocki et al. 2000; Paizs and Suhai 2005). The sequence of the peptide is determined by comparing the differences among the masses of the product ions with the known masses of the individual amino acid residues (see Appendices 5 and 6).

A standardized nomenclature has been adopted to describe the different product ions observed in fragmentation spectra (Fig. 2). The a-, b-, and c-ions all contain the amino terminus of the peptide, while the x-, y-, and z-ions all contain the carboxyl terminus. Under low-energy CID, the major amino-terminus ion series is the b-ion series and the major carboxy-terminus ion series is the y-ion series. The a-, c-, x-, and z-ion products are typically not produced under low-energy CID and are only observed under high-energy CID and other fragmentation methods, such as electron transfer dissociation (ETD). For both the b- and y-ion series, the *m/z* difference between adjacent ions is equal to the mass of the amino acid residue at that position (Fig. 1). For doubly charged peptides, the b- and y-ion series are complementary as the cleavage of the protonated amide bonds can generate

*Several sections contributed by Rebecca A. Bish (*Department of Cancer Biology and Genetics, Memorial Sloan-Kettering Cancer Center, New York, New York 10021*) and Eric S. Simon (*Department of Biological Chemistry, University of Michigan, Ann Arbor, Michigan 48109*).

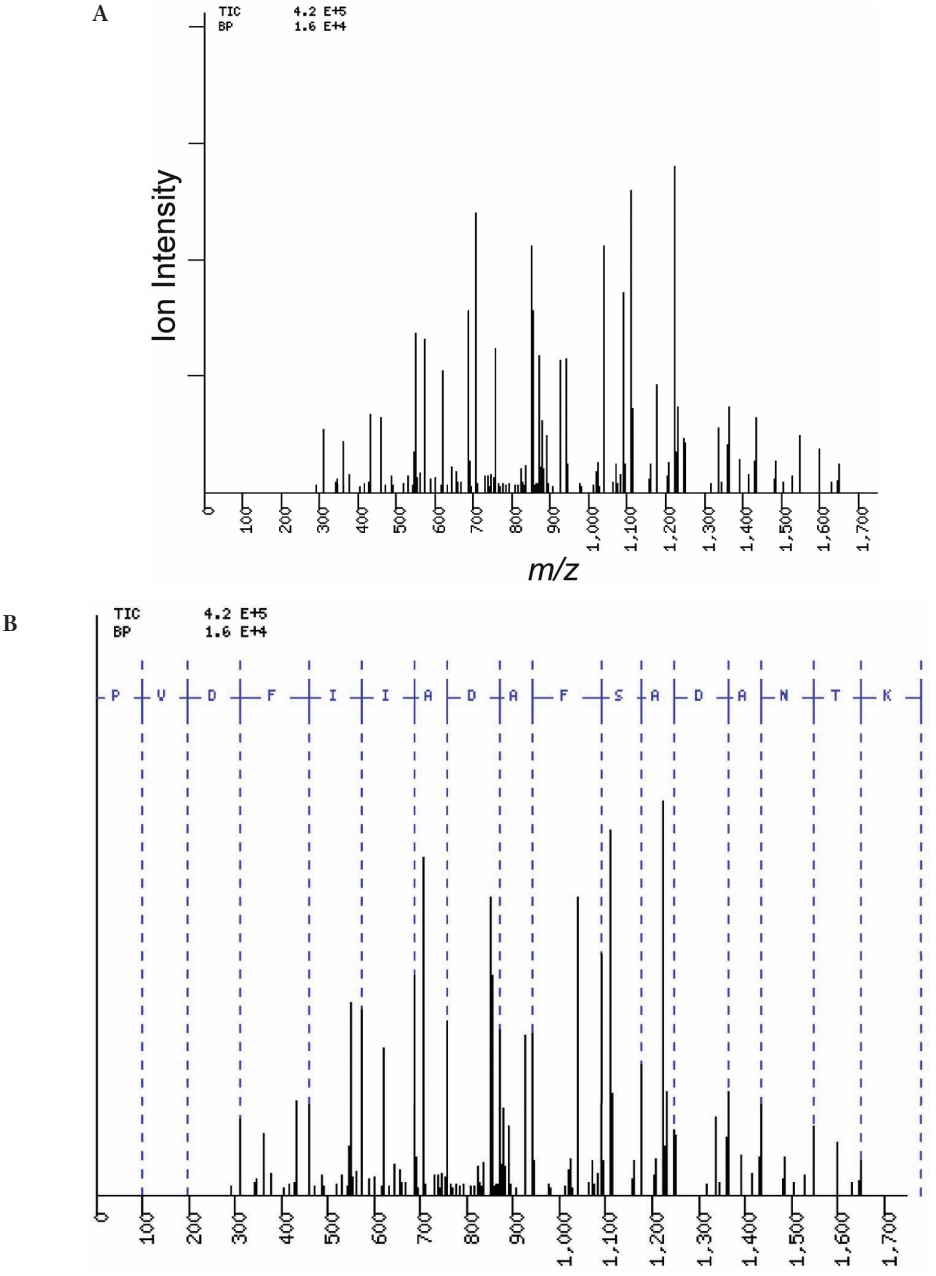
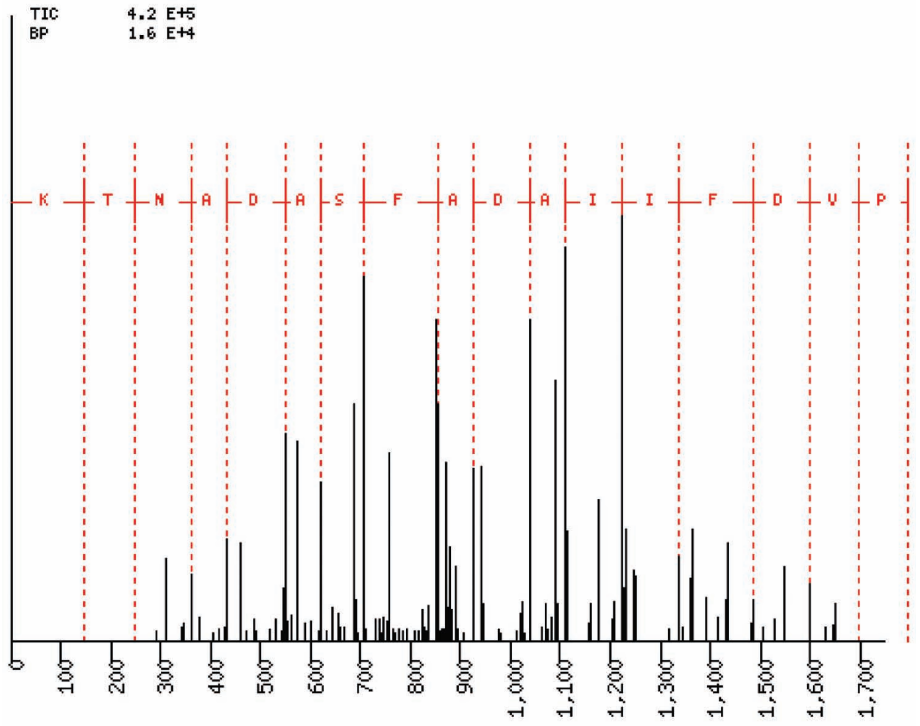
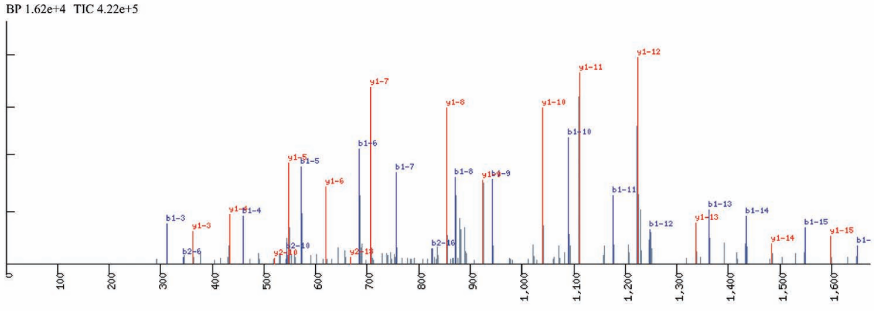


FIGURE 1. (A) Unannotated MS/MS or product spectrum. A fragmentation spectrum acquired for a doubly charged precursor ion (*m*/*z* 898.9) from a trypsinized yeast cell lysate. The x-axis shows the mass of the ions as the mass to charge ratio (*m*/*z*), and the y-axis is the abundance of the ion measured in ion intensity. The spectrum contains peaks from the peptide's fragmentation ions along with background noise. Typical of good quality data, the spectrum shows a broad range of ions that are significantly higher in intensity than the noise peaks. In addition, the intense ions appear to form a symmetrical pattern. (B) Theoretical b-ions from a peptide hit. The precursor and fragmentation ion data were submitted to a Sequest database search against the yeast proteome. The top scoring peptide was PVDFIIADAFSADANTK from the Pgk1 protein (Cn = 6.9). The program predicted the precursor was a doubly charged ion ($z = {}^{+}2$). The theoretical b-ions from the peptide are superimposed on the MS/MS spectrum showing which fragment ions might be b-ions. The mass difference between adjacent b-ions corresponds to the residue mass of an amino acid (see Appendices 5 and 6). (C) Theoretical y-ions from a peptide hit. The theoretical y-ions from the PVDFIIADAFSADANTK peptide are superimposed on the MS/MS spectrum showing which fragment ions might be y-ions. The mass difference between adjacent y-ions corresponds to the residue mass of an amino acid (see Appendices 5 and 6). (D) Theoretical b-y ions from a peptide hit. The MS/MS spectrum's fragmentation ions are labeled with the theoretical b-y ions from the peptide PVDFIIADAFSADANTK. For a good database match to a multiply charged precursor ion, most of the intense ions should match the predicted fragmentation ions. (E) Summary of peptide's match to the MS/MS spectrum. This method of interpretation is commonly used to represent which b-y ions from a peptide are detected in the MS/MS spectrum. The calculated *m*/*z* values of the b-ions are shown above the amino acid sequence and the calculated *m*/*z* values of the y-ions are shown below the sequence. All the values are singly charged and are the monoisotopic masses. The figure also highlights the complementarity of the b- and y-ion series. As a b- or y-ion is identified, its complementary ion can be calculated using the precursor ion mass. For evaluating a peptide hit to an MS/MS spectrum, detecting complementing ion pairs is one indication of an accurate database hit.

C



D



b ions		b1	b2	b1-17	b2-8.5	b1-18	b2-9	y ions		y1	y2	y1-17	y2-8.5	y1-18	y2-9
b1	P	98.1167	49.55835	81.1167	41.05835	80.1167	40.55835	y1	K	147.1742	74.0871	130.1742	65.5871	129.1742	65.0871
b2	V	197.2493	99.12465	180.2493	90.62465	179.2493	90.12465	y2	T	248.2793	124.63965	231.2793	116.13965	230.2793	115.63965
b3	D	312.3379	156.66895	295.3379	148.16895	294.3379	147.66895	y3	N	362.3832	181.6916	345.3832	173.1916	344.3832	172.6916
b4	F	459.5145	230.25725	442.5145	221.75725	441.5145	221.25725	y4	A	433.462	217.231	416.462	208.731	415.462	208.231
b5	I	572.674	286.837	555.674	278.337	554.674	277.837	y5	D	548.5506	274.7753	531.5506	266.2753	530.5506	265.7753
b6	I	685.8335	343.41675	668.8335	334.91675	667.8335	334.41675	y6	A	619.6294	310.3147	602.6294	301.8147	601.6294	301.3147
b7	A	756.9123	378.95615	739.9123	370.45615	738.9123	369.95615	y7	S	706.7076	353.8538	689.7076	345.3538	688.7076	344.8538
b8	D	872.0009	436.50045	855.0009	428.00045	854.0009	427.50045	y8	F	853.8842	427.4421	836.8842	418.9421	835.8842	418.4421
b9	A	943.0797	472.03985	926.0797	463.53985	925.0797	463.03985	y9	A	924.963	462.9815	907.963	454.4815	906.963	453.9815
b10	F	1090.2563	545.62815	1073.2563	537.12815	1072.2563	536.62815	y10	D	1040.0516	520.5258	1023.0516	512.0258	1022.0516	511.5258
b11	S	1177.3345	589.16725	1160.3345	580.66725	1159.3345	580.16725	y11	A	1111.1304	556.0652	1094.1304	547.5652	1093.1304	547.0652
b12	A	1248.4133	624.70665	1231.4133	616.20665	1230.4133	615.70665	y12	I	1224.2899	612.64495	1207.2899	604.14495	1206.2899	603.64495
b13	D	1363.5019	682.25095	1346.5019	673.75095	1345.5019	673.25095	y13	I	1337.4494	669.2247	1320.4494	660.7247	1319.4494	660.2247
b14	A	1434.5807	717.79035	1417.5807	709.29035	1416.5807	708.79035	y14	F	1484.626	742.813	1467.626	734.313	1466.626	733.813
b15	N	1548.6846	774.8423	1531.6846	766.3423	1530.6846	765.8423	y15	D	1599.7146	800.3573	1582.7146	791.8573	1581.7146	791.3573
b16	T	1649.7897	825.39485	1632.7897	816.89485	1631.7897	816.39485	y16	V	1698.8472	849.9236	1681.8472	841.4236	1680.8472	840.9236

E

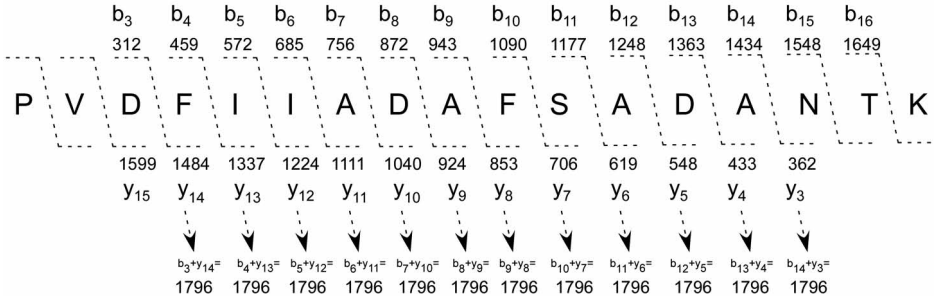


FIGURE 1. *(See facing page for legend.)*

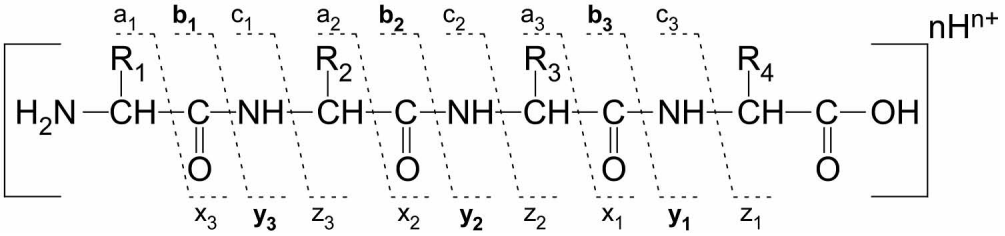


FIGURE 2. Nomenclature for fragmentation ions generated from a peptide. The diagram shows two classes of fragmentation or product ions that are created from a protonated peptide. One class contains the amino terminus (a_i, b_i, c_i) while the other class contains the carboxyl terminus (x_i, y_i, z_i). For both classes, the fragmentation can occur at three different positions along the peptide backbone. The b- and y-ions (bold) are the dominant fragmentation ions created by low-energy CID. For the b-y ions, the fragmentation occurs at the peptide's amide bonds. The diagram illustrates the complementarity of fragment ions. For example, the sum of b_1 + y_3 ion masses equals the precursor ion mass.

both a b-ion and y-ion (Figs. 1 and 2). The b- and y-ions are derived from single fragmentation reactions that occur from a population of precursor ions protonated at the different amide bonds. Recognizing the members of the b- and y-ion series in the product spectrum and the calculation of the residue masses is the fundamental process of de novo amino acid sequence determination.

The de novo interpretation of tandem spectra is not straightforward. The b- and y-ions do not stand out upon casual observation (Fig. 1A). The relative abundances of product ions typically vary extensively, with some product ions dominating the spectrum while other product ions are undetectable (Fig. 1B,C). The variation reflects the differences in the amide bonds due to the properties of the amino residues flanking the bonds and the position of the bonds in the peptide sequence (see Protocol 3). Losses of neutral molecules from the precursor and b- and y-ions reduce the intensities of the fragmentation ions. For tryptic peptides, the loss of water (–18 Da) from Ser, Thr, Glu, and Asp residues and ammonia (–17 Da) from Asn, Gln, Arg, and Lys residues are frequently observed. These neutral losses generate additional product ions from the loss of –17 or –18 Da from the b- and y-ions (Fig. 1D). Like most analytical techniques, mass spectrometry data contain background noise from the sample and the mass spectrometry system, which adds to the overall complexity and uncertainty. In a limited number of examples, the combined residue mass of two amino acids may equal the residue mass of a single amino acid. For example, the combined residue masses of Asp-Ala, Val-Ser, and Gly-Glu all equal the residue mass of Trp (186 Da). Finally, the residue masses of isoleucine and leucine are identical (113 Da) and cannot be distinguished using low-energy CID. Given these complications, the de novo interpretation of tandem spectra can be complicated and time consuming. The lack of information in the MS/MS spectrum during de novo analysis frequently leads to an incomplete peptide sequence.

For proteomic experiments, the de novo interpretation of tandem mass spectra is typically bypassed and the initial interpretation of spectra is done using database search programs (Fig. 3). The programs compare the experimental mass spectrometry data to theoretical data derived from protein sequences in a database. Database search algorithms rapidly process large numbers of MS/MS spectra and return a list of peptide sequences that match the MS/MS spectra. From the peptide sequences, the programs build a list of proteins predicted to be in the sample. Evaluating the accuracy of these lists is one of the most difficult challenges in proteomics

The database search programs first identify strings of amino acids from proteins in the database whose mass is equivalent to the precursor ion's mass. Then, the theoretical masses of predicted product ions of each of these peptides are compared to the actual ion masses and intensities in the MS/MS spectrum. The mathematical and statistical methods used for comparing and scoring the theoretical values and experimental data constitute the primary differences among the different database search programs (Sadygov et al. 2004). Once a candidate peptide sequence has been iden-

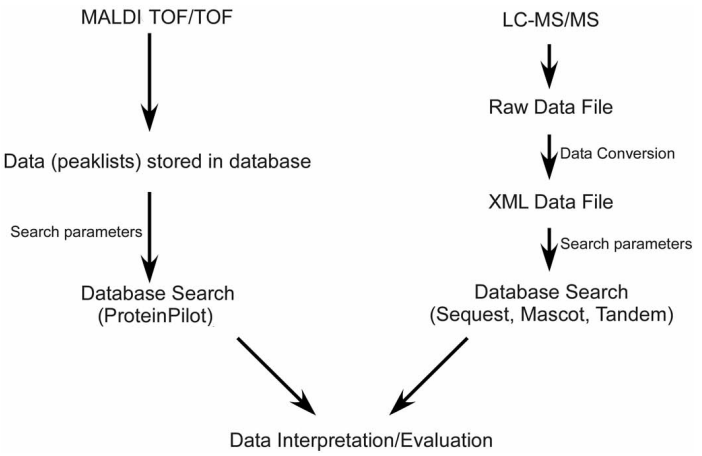


FIGURE 3. Flow diagram showing the process for analyzing MALDI TOF/TOF and LC-MS/MS data.

tified, the b- and y-ion series in the experimental spectra can be identified (Fig. 1B–E). The database search algorithms are designed to provide a score or statistical value to quantify how significantly each peptide sequence in the database matches the experimental data. The programs rank the peptide scores to show the best peptide match. Ideally, each protein is identified by multiple, independent peptide matches. However, in reality, a large percentage (>50%) of the identified proteins are typically based on a single peptide match to an MS/MS spectrum. For these "one hit wonders," the quality of the spectral data or the match of the peptide sequence to the MS/MS spectrum may be marginal. Confirmation of the match requires careful manual inspection of the MS/MS spectrum and determination that the predicted product ions support the fragmentation data.

PROTOCOL 1

Analyzing LC-MS/MS Data Using the Global Proteome Machine

MATERIALS

Binary data file generated from an LC-MS/MS experiment
Computer running Windows XP or Linux

PROCEDURE

Converting Native Mass Spectrometry Data Files to an XML Format

Most mass spectrometer vendors encode the data files in a propriety binary format. The *m/z* and intensity values from the precursor (MS) and product (MS/MS) spectra need to be extracted from these native files into a text format that can be read by the database search programs. XML is a simple, flexible text format derived from the standardized general markup language (SGML), originally designed for electronic publishing. Two groups have independently proposed a common file format based on XML (mzXML and mzData) for representing native mass spectrometry data. This common format

allows data from different instrument manufacturers to be analyzed by a database search algorithm. Recently, the two XML formats were reconciled and a new format was created, mzML, which is intended to replace the mzXML and mzData formats (Deutsch 2008). Most native data files can be converted to an XML format using software applications provided by the instrument manufacturers.

In the course, a ThermoFisher LTQ mass spectrometer is used for LC-MS/MS experiments. The acquired data are recorded in a native, binary file with the "RAW" extension. To analyze the data using the database search programs, the RAW file needs to be converted to either XML or some other text file format. We use the publicly available ReadW.exe program to convert the RAW files to the XML file format mzXML. Software is also publicly available for converting between XML and ASCII formats (http://www.proteomecommons.org/).

1. Install the ReadW.exe program in the "C:/WINDOWS/system32 folder" on a PC running ThermoFisher's Xcalibur software.

   The ReadW.exe program can be downloaded from either the open source software website SourceForge (http://sourceforge.net/) or the Institute for Systems Biology (http://tools.proteomecenter.org/ReAdW.php). Because the ReadW program depends on Windows-only vendor libraries from ThermoFisher, this code will only work under Windows with ThermoFisher's XCalibur software installed. Other conversion programs are used to convert the native files from other mass spectrometer vendors. The mxSTAR.exe programs convert SCIEX/ABI's Analyst files. The mzBruker.exe program converts native Bruker format, and MassWolf.exe converts Micromass's MassLynx files to the mzXML format.

2. Open the command prompt on the PC by clicking on "Start" > "Run" and then typing "cmd."
3. Change directories to the folder containing your RAW file by typing the command prompt "<cd c:\xcalibur\data>."
4. Type the command "<readw.exe filename.raw>" to run the conversion program.
5. When the program has finished, an mzXML file with the same name as the original file will be found in the "C:\Xcalibur" directory.

Searching Tandem Mass Spectrometry Data

Numerous database search programs have been developed to compare the measured values of the precursor ion and its fragmentation ions to the theoretical masses of peptides and their fragmentation products derived from protein sequences in a database. Most mass spectrometer vendors offer a commercially licensed database search program for processing the data with their mass spectrometers. Three of the most widely used search algorithms are Sequest, Mascot, and X!Tandem (Eng et al. 1994; Perkins et al. 1999; Craig and Beavis 2004). All three programs can be used to analyze native data from any mass spectrometer once the data file has been converted to an XML or an ASCII format. SEQUEST, which was the first program to take unedited tandem mass spectra and compare the data to sequences in a protein database (Eng et al. 1994), is typically used by operators of ThermoFisher mass spectrometers. The SEQUEST algorithm uses a cross-correlation function to assess the similarity between the experimental and predicted spectra. Peptide sequences matching the spectrum are ranked by a cross-relation score (Cn). Mascot and X!Tandem use a probability-based scoring algorithm to derive from a protein database the most likely peptide sequence matching the experimental spectrum (Pappin et al. 1993; Perkins et al. 1999; Craig and Beavis 2003; Fenyo and Beavis 2003). Operators of SCIEX/ABI instruments typically use Mascot. X!Tandem or Tandem (http://www.thegpm.org/) was the first open source database search algorithm that was freely available and was being used by a growing number of investigators.

While the database search algorithms match acquired tandem mass spectra to peptide sequences in a database, the peptide sequences must be assembled into proteins that represent the initial sam-

ple. Early generation assembler applications such as SEQUEST Summary and Autoquest used a simple scoring threshold when assembling peptides into proteins (McCormack et al. 1997; Link et al. 1999). Second generation applications such as DTASelect and INTERACT applied layers of filters at the peptide and protein level when assembling peptide sequences into a list of proteins (Han et al. 2001; Tabb et al. 2002). However, this process was complicated by nonunique peptide sequences that match multiple proteins in a database. In addition, the wide range of scoring confidences for peptide matches increased the ambiguity in the final list of proteins. More recently, the probability-based method of ProteinProphet, and peptide-centric approaches such as Isoform Resolver and parsimony analysis have been developed to construct protein profiles from peptide lists generated by spectra search algorithms (Nesvizhskii et al. 2003; Resing et al. 2004; Yang et al. 2004; Zhang et al. 2007).

Installing the GPM on a PC Running Windows

The global proteome machine (GPM) is an open-source interface that uses the X!TANDEM database search algorithm to conduct database searches on data files from mass spectrometry experiments. It was the first open source database search program. Data for analysis by the GPM can be submitted over the internet, although this option is generally too slow for the large files generated by LC-MS/MS experiments. Since the release of X!Tandem, other open source database search programs have become available, including the open mass spectrometry search algorithm (OMSSA) sponsored by NCBI (Geer et al. 2004) and MyriMatch by David Tabb's laboratory at Vanderbilt University (Tabb et al. 2007).

At the CSHL Proteomics course, we use the Sequest, Mascot, and X!Tandem programs for analyzing LC-MS/MS data. Since the GPM and X!Tandem are open source and freely available, we install them on the students' laptop computers. The students are encouraged to compare the results from all three programs on the same sets of data. Most students in our course find it convenient to use the GPM installed as a local copy on a personal computer. Unfortunately, there is no version of the GPM for the Macintosh operating system.

1. Go to http://www.thegpm.org.
2. On the left-hand side of the page, under the heading "Download," click the link for the ftp site.
3. Click the directory "Projects."
4. Click the directory "GPM."
5. Click the directory "Current_release" or "GPM-xe-installer." Download the latest version of the software to your computer.

   Software for the GPM is updated regularly. Check for the most current software releases for your computer and operating system and then download the most recent version of GPM. The download typically includes several commonly used protein databases from model prokaryotic and eukaryotic organisms. If you work on an organism whose database is not included, you can download and install whatever databases you want.
6. Use WinZip to extract all files to a folder in your Program Files directory.
7. To run the GPM, enter the Program Files\GPM folder and click on the GPM Manager.exe icon.

   Drag this icon onto your desktop if you wish to create a shortcut.

Setting Up a Simple GPM Database Search of an LC-MS/MS Experiment

This section describes typical settings for an X!Tandem database search of LC-MS/MS data acquired using a ThermoFisher LTQ (Fig. 4). In most cases, the default values are a good place to start. We

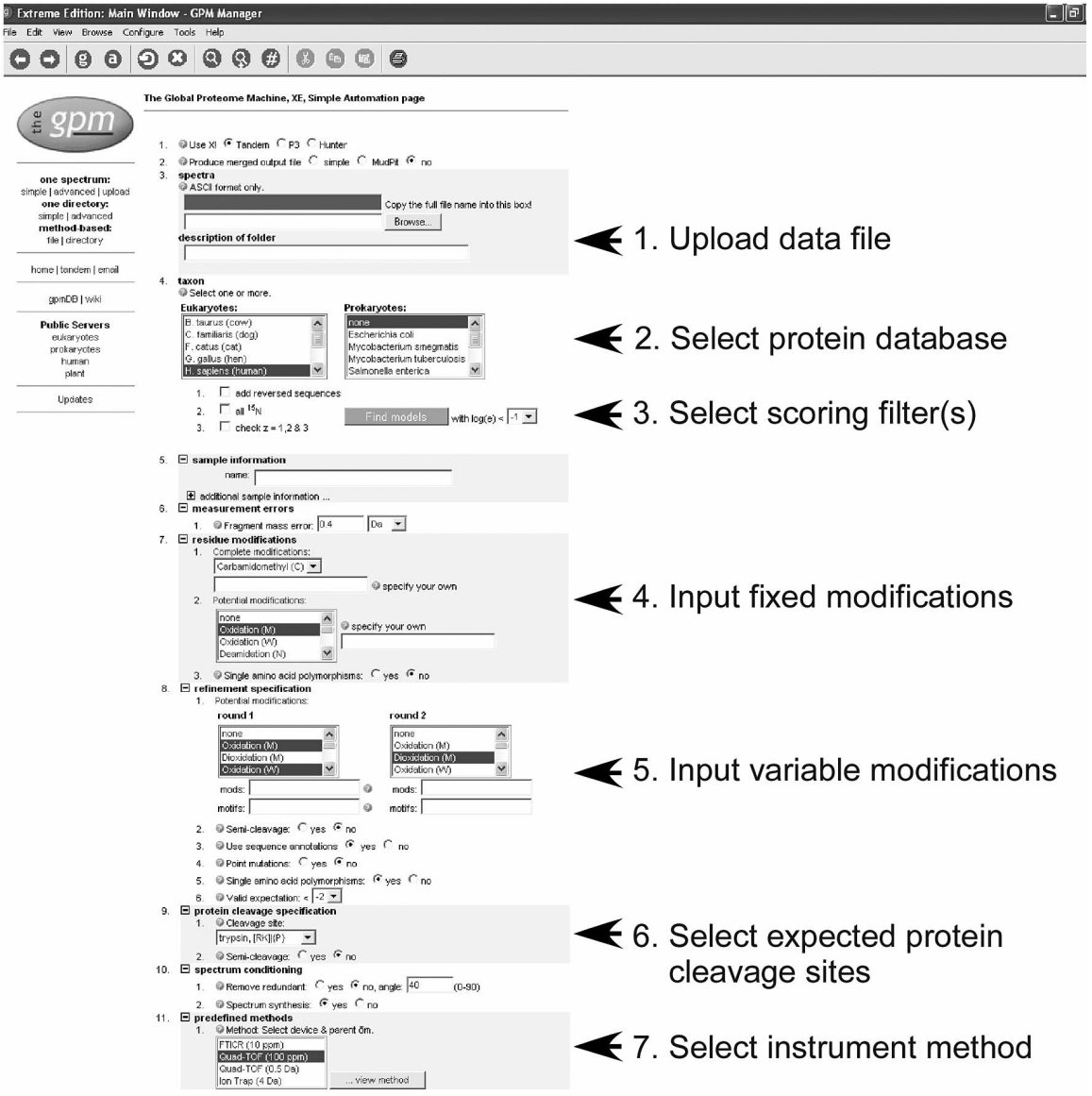


FIGURE 4. Parameter's page for setting up a Tandem search of acquired MS/MS data against a protein database. The typical order for inputting the required search information is shown on the right. All search engines (e.g., Sequest, Mascot, X!Tandem, etc.) require users to input similar information before running the database searches. Search engines will have different parameters for the scoring filter(s) based either on a probability value, false–positive rate, unique scoring values, or the proteolytic cleavage sites of the peptide.

have empirically determined that these recommended settings return good results for the LTQ instrument. However, these settings will vary significantly depending on which mass spectrometer you use. If you use the GPM to analyze your own data, you should adjust the settings as appropriate for the mass spectrometer used to acquire the data. If you use a proteomics or mass spectrometry facility, the manager of the facility should be able to help you determine the appropriate settings. It can also be helpful to run multiple searches on the same data set with different parameters for optimization purposes.

While this section describes the parameters for starting a database search using the GPM and the X!Tandem algorithm (Fig. 4), the input parameters will be identical for almost all other search algorithms (such as Sequest and Mascot). Users will be required to input an XML or ASCII version of their

mass spectrometer's data file, select a protein database to search, select the protease used to digest the protein sample, select prefiltering parameters to remove redundant or poor quality spectra, input the expected mass accuracy of the acquired data, input any fixed or variable amino acid modifications, and input the scoring criteria for accepting a peptide sequence matching an MS/MS spectrum.

It is becoming common to compare the acquired mass spectrometry data to a true database as well as a decoy or false protein database. The false database is typically created from the true protein database by either reversing or randomly shuffling the amino acids for each protein. The actual and decoy databases are concatenated together and the mass spectrometry data file is simultaneously compared to the concatenated protein databases. From the false database, a false discovery rate (FDR) can be calculated for the experiment. The FDR estimates the percentage of identified proteins that are false hits. In describing the setup for a database search (below), we have chosen the GPM and X!Tandem since they are both publicly available and widely used by a number of groups.

1. Convert the ThermoFisher RAW file to an mzXML file using the ReadW.exe program as described in the first section of this protocol.

   This converts the native mass spectrometry data files to an XML format.

2. Under the heading "one spectrum" on the left-hand side of the page, click "advanced."

3. In the box labeled "spectra," browse to find your mzXML data file (Fig. 4).

4. Under the heading "taxon," browse to find the appropriate organism. In the CSHL course, most of the experiments are done using the yeast *S. cerevisiae* (Fig. 4).

   Excessively large protein databases may overwhelm a database search engine's ability to discriminate correct sequences from incorrect ones. Increasing the number of protein sequences searched against the MS/MS data will amplify the number of mathematical models to be tested. Unnecessarily large protein databases will increase the number of random or stochastic matches. As a consequence, correct identifications may be obscured by random matches. Select a protein database that is appropriate for the source of the proteins in the samples. Limit the search databases to a specific species or groups of species. A smaller database will also decrease the time to perform the database search. Add common protein contaminants (trypsin and human keratins) to the protein databases that are routinely found in samples.

5. Under the heading "measurement errors," change the following settings: Parent mass error should be +4 or –2 Da (Fig. 4).

   Be sure to change the units from ppm to Da. The measurement errors are dependent on the instrument being used for the tandem mass spectrometry experiment. These values are typically used for ion trap instruments. X!TANDEM allows the user to set asymmetric values for the parent ion mass error, to compensate for the asymmetric systematic mass errors generated by some types of mass spectrometers, particularly the ion trap instruments. TOF and FTICR mass analyzers will have smaller measurement errors.

6. Under the heading "signal processing," change the following settings: Minimum parent M+H should be 300 and the minimum peaks should be 10.

7. Add any desired posttranslational modifications (Fig. 4).

   Input any known or suspected modifications. Two types of modifications are entered. Fixed modifications are applied universally to every instance of the specified residue(s) or terminus. For most of the experiments in this manual, Cys residues have been reduced with DTT and alkylated with IAA, creating a carbamidomethyl on Cys residues, which means that all calculations will use 160 Da (157 Da + 57 Da) as the mass of Cys residues.

   Variable modifications are those which may or may not be present in the protein sample. The search algorithms will query all possible arrangements of the variable modifications. Common variable modifications include oxidized Met (+16 Da), deamidation of Asn or Gln (+1 Da), and phosphorylation of Ser, Thr, or Tyr (+80 Da) residues. There are numerous other potential modifications that can be searched (Creasy and Cottrell 2004).

For protein samples prepared using solutions containing urea, carbamylation (+43 Da) of the proteins' amino terminus and Lys, Arg residues are frequently observed, especially when the samples are heated. Isocyanatic acid, a urea break-down product, covalently reacts with amino groups. Database searches with a variable modification of +43 Da at the amino terminus and Lys, Arg residues will often detect carbamylated peptides.

As a caution, a single variable modification will generate many possible mathematical models to be tested. Multiple variable modifications cause the number of mathematical possibilities to increase geometrically. The database search will take considerably longer with more variable modifications. More importantly, the increased number of mathematical possibilities increases the number of random matches. Excessive protein variable modifications may overwhelm a database search engine's ability to discriminate correct sequences from incorrect ones. Vigilant data analysis is required to validate the accuracy of the modified peptide hit (see Protocol 3).

8. Click "Find models" to start the search.

Setting Up a MudPIT GPM Database Search

In the MudPIT experiment, a trypsin-digested, whole-cell lysate was fractionated using multidimensional HPLC and each fraction was analyzed by ESI-MS/MS (Experiment 6). Since each MudPIT fraction generates a data file, multiple mass spectrometry data files were created for the single sample. For X!Tandem to analyze and process the data, the mzXML-converted data files are placed in a single folder on the computer. X!Tandem sequentially searches the data files against a protein database and the results are merged into a single output file.

1. Convert all of the RAW files to mzXML files using the ReadW.exe program as described earlier.
2. Place all of the mzXML files from a single MudPIT run into a separate folder. This folder should not contain any mzXML files that are not to be included in the search.
3. In the GPM manager software, under the heading "one directory" on the left-hand side of the page, click "advanced."
4. Next to the heading "Produce merged output file," check the box for MudPIT.
5. Tell the GPM program which files to search in a combined MudPIT run.
6. Under the heading "taxon," browse to find the appropriate organism. In the CSHL course, most of the experiments are done using the yeast *S. cerevisiae*.
7. Under the heading "measurement errors," change the following settings: Parent mass error should be +4 or –2 Da.

   Be sure to change from ppm to Da.
8. Under the heading "signal processing," change the following settings: Minimum parent M+H should be 300 and the minimum peaks should be 10.
9. Add any desired posttranslational modifications.
10. Click "Find models" to start the search.

    The output from this search will be an individual results file for each fraction, plus a combined results file for the entire MudPIT run.

Preliminary Analysis of the Output from a Database Search

When the X!Tandem search is complete, the GPM displays a list of identified protein and peptide sequences (Fig. 5). Almost all investigators will rush to look at the list to see what proteins were in

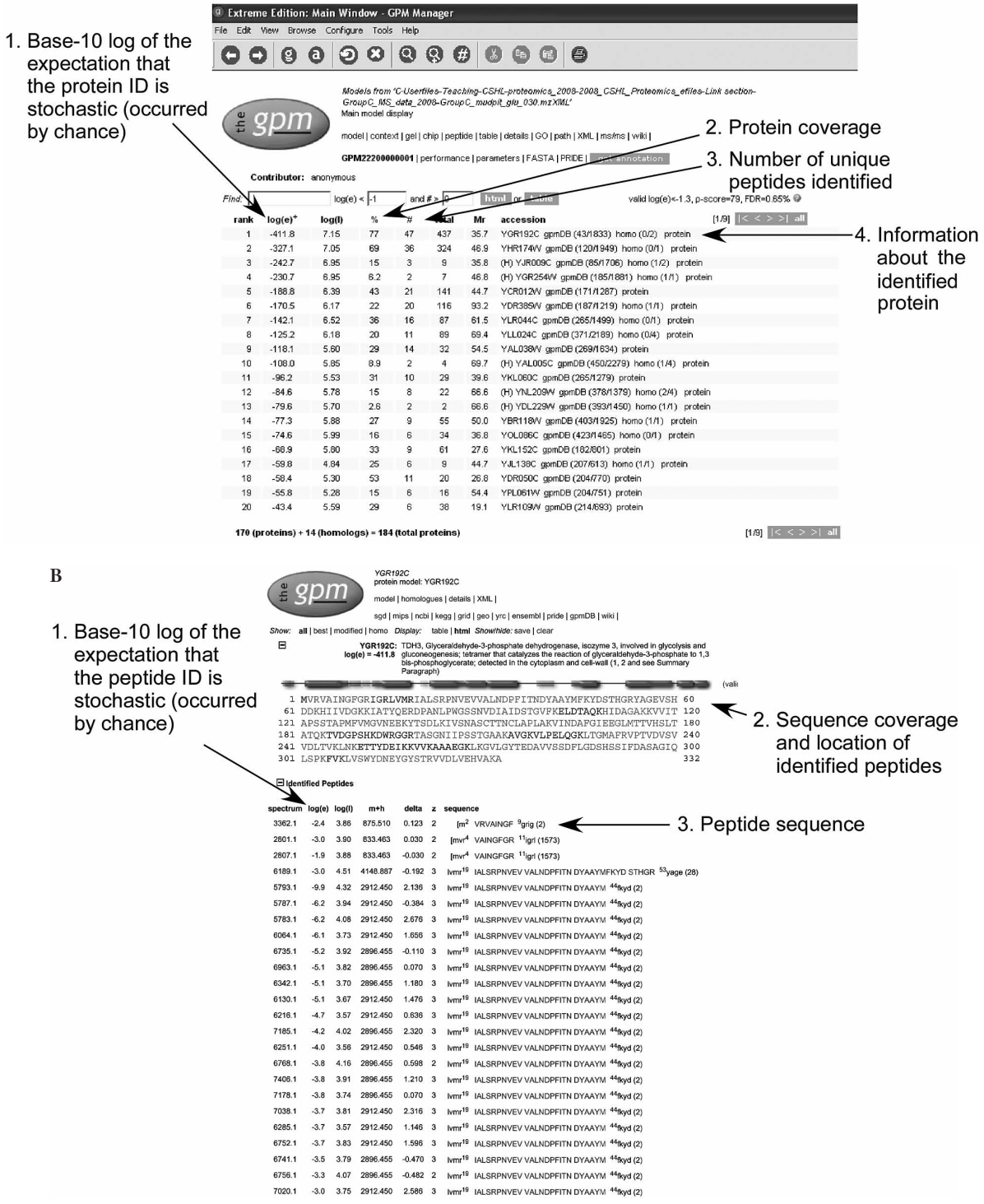


FIGURE 5. (A) Output page from an X!Tandem search of MS/MS data. A screen shot of the initial output or "model" page from X!Tandem showing the key information used to initially evaluate the list of identified proteins. (1.) log(e)$^+$ is the statistical score indicating the significance of the identification. The list of proteins is initially sorted by the proteins with the most significant identifications. For an initial evaluation, proteins with log(e)$^+$ scores below –1 are a good place to begin evaluating the data. (2. and 3.) The percent protein coverage and number of unique peptides identifying the protein are useful metrics for initially evaluating the results. Protein identifications based on ≥ 2 unique peptides are typically considered accurate. (4.) Information about the identified protein includes its accession number and gene name. Other information typically includes its predicted biological function, cellular location, and biological process. (B) Information on the protein identified by X!Tandem. A screen shot of the "protein model" page from X!Tandem illustrating the key information available. The page provides details on the individual spectrum and the peptides identifying a specific protein. The protein coverage section provides a graphical and text view of the location of the identified peptides.

their sample. Often investigators rapidly scan down the list of identified proteins looking for the proteins they consider exciting. These proteins are commonly called the "shiny pebbles." The evidence that these "shiny pebbles" are accurately identified may be marginal. Overall, the interpretation and validation of the list of peptides and proteins is one of the most exciting and challenging parts of proteomics.

Before interpreting the results, it is important to remember several key points.

- The database search is unbiased. All protein and peptide sequences in the database are typically compared with each MS/MS spectrum. The search engine did not make any assumptions about what proteins might be in the sample. This is probably the most important and powerful aspect of the approach. Unexpected proteins are identified that investigators would not have predicted to be in the sample.
- Only peptides and proteins in the protein database will be listed in the output. If a protein or peptide sequence is not in the database, it will not be identified. The search algorithms do not perform de novo sequencing of the MS/MS spectra and derive peptide sequences. The programs only match peptide sequences in a database to the MS/MS spectra using the parameters in the setup file.
- Larger proteins are easier to identify than smaller proteins. The LC-MS/MS approach will only fragment a subset of the total tryptic peptides in the sample. The more tryptic peptides a protein generates, the greater the chance a peptide from the large protein will be selected for fragmentation. Small proteins that only generate a few tryptic peptides could escape detection.
- The LC-MS/MS and MALDI-MS approaches are biased towards fragmenting peptides from abundant proteins. Since the selection process for fragmentation of precursor ions is typically based on ion intensity, the more intense ions will be selected for fragmentation and the less intense ions will be ignored and possibly escape fragmentation. While peptides ionize at different efficiencies, on average, the more abundant a peptide or protein is, the more likely it will generate peptides with strong precursor ion signals. Low-abundance peptides and proteins could escape detection.
- Peptides with unexpected amino acid changes or substitutions will not be matched to a sequence in the database and will not be listed. Once again, protein sequences not in the database will not be identified. Some database search programs will perform what is called "homology" searching, which allows for amino acid substitutions.
- Any peptides with modified amino acids that were not specified in the search parameters will not be listed. These peptides will escape detection. The search programs will only search and list peptides with modified amino acids if they have been included in the search parameters.

Given the strengths and weaknesses of the LC-MS/MS coupled with database search analysis, next we describe an approach for initially evaluating the results.

1. Select criteria for initially accepting database search results as correct.

   A number of studies have examined scoring criteria for deciding whether to accept or reject a peptide identification from the different database search algorithms (Washburn 2001; Keller 2002; MacCoss 2002; Nesvizhskii et al. 2003 ; Peng 2003; Sadygov et al. 2004). While the issue is still debated, two methods have been adopted generally. One method calculates a probability score for a peptide or protein identification and the second method estimates the false discovery rate (FDR) after the database search. For both methods, a value of 0.05 is typically a good starting place for initially accepting a peptide identification.

2. Order the list of identified proteins by the number of peptides and tandem mass spectra identifying a protein in the sample. The higher the number of peptides significantly

matching a protein, the more likely the protein is identified accurately and is not a stochastic hit.

Since each unique peptide sequence matching a tandem mass spectrum is an independent event, each peptide from a protein that significantly matches the unique tandem mass spectra increases the probability that the protein is accurately identified. As an example, for two independent peptides identifying the same protein with 0.05 probability, the probability the protein is incorrectly identified is 0.05 x 0.05 = 0.0025. In other words, there is a 1 out of 400 chance the protein is incorrectly identified. If only one peptide identifies the protein at 0.05 probability, there is a 1 in 20 chance the protein has been falsely identified. Two independent peptides identifying a protein is strong evidence that the protein is accurately identified. In publishing proteomic data, two independent peptides identifying a protein by tandem mass spectrometry analysis is typically sufficient for the journal to accept the result.

3. Recognize proteins that are identified by unique peptides. In higher eukaryotic organisms, extensive gene duplication and alternative RNA splicing has resulted in a large number of proteins with similar sequences. Protein databases may contain multiple sequences for a protein that represent alternative forms that differ by only a single amino acid. Because peptides are frequently shared by multiple proteins in a database, the assembly of identified peptides into a list of proteins can drastically overstate the number of proteins in samples. It is important to recognize which proteins are identified by unique peptides and which proteins are indistinguishable because only shared peptides are identified. For proteins that only share peptide sequences, the identified proteins should be reported as a protein group.

4. Identify proteins or protein groups that are indistinguishable. For complex eukaryotic organisms (like humans and mice) with rampant alternative splicing, gene duplications, and polymorphisms, protein databases often contain large numbers of proteins with similar sequences. For example, human protein databases may contain many serum albumin and immunoglobin proteins that differ only by a few amino acid residues. These proteins will be indistinguishable because there are no unique peptides allowing the investigator to determine which protein produced the peptides. In these cases, reporting the smallest numbers of proteins necessary to explain the observed peptides (parsimony) is encouraged (Carr et al. 2004; Bradshaw 2005).

5. For proteins or modified peptides identified by a single spectrum, manual evaluation of the spectra and the matching sequence is required (see Protocol 3).

PROTOCOL 2

Using ProteinPilot Software for Peptide and Protein Identification

ProteinPilot is a software package used to identify proteins from peptides generated from 2D gel and MudPIT experiments, including quantification of peptides from the iTRAQ, SILAC, and ICAT techniques (Shilov et al. 2007). It features a novel approach to peptide identification, called the Paragon algorithm (Shilov et al. 2007). Paragon assesses each MS/MS spectrum from a data set to determine which spectra are worth scoring. In other words, it acts as a filter to remove spectra that are less likely to yield reliable peptide identifications. It establishes a threshold for score-worthy spectra based on two inputs: a Sequence Temperature Value (STV) and feature probability. The STV identifies "hot" versus "cold" regions of the database by computing a quantity, based on extracted de novo sequence tags

from an MS/MS spectrum, which reflects the degree that each theoretical peptide from a database matches the MS/MS spectrum. Feature probabilities factor in things like posttranslational modifications (PTMs), digestion events (missed or nonspecific cleavages), mass tolerances, and substitutions. A peptide with a "hot" STV receives more search time and is considered for more modifications or other feature events, while "cold" peptides receive less search time and are considered for only the more common PTMs and features. A threshold is then computed, based on STV and feature probabilities, that determines which MS/MS spectra should be scored and which should be discarded. Alternatively, the software provides an interface for searching data with the Mascot algorithm.

While ProteinPilot is an innovative and effective proteomic tool for searching databases, it is currently only compatible with data generated from instruments such as the 4700 MALDI TOF/TOF Analyzer, manufactured by Applied Biosystems. Data can only be uploaded into ProteinPilot via a direct connection to the instrument database. Externally generated data must be converted to mgf format.

This protocol describes procedures for analyzing data generated from a 2D gel experiment (Experiment 1) or data acquired from an iTRAQ experiment (Experiment 7). No file format manipulation is necessary because the software reads the raw file generated by the mass spectrometer directly from a table in the instrument data. This protocol provides procedures for loading data from the database, setting up a search, and exporting and storing the results. An overview of the software features relevant to data output for both 2D gel and iTRAQ data will be covered with an emphasis on data interpretation.

MATERIALS

Computer running Windows XP SP2 and ProteinPilot (version 2.0.1)
Spot set to be analyzed or iTRAQ data

PROCEDURE

Loading Data Files (Spot Sets) in ProteinPilot

1. Open the software by double clicking the ProteinPilot icon on the desktop.
2. In the "Workflow Tasks" task bar, click "Identify Proteins." Figure 6 displays the "Identify Proteins" screen view.
3. Click "Add 4000 Series Data" (Fig. 6A). Select the appropriate spot set(s) to be analyzed. The data set(s) appears in the "Data Sets to Process" window (Fig. 6B).

   For the CSHL course, select the appropriate spot set(s) from the proteomics course project to be analyzed.

Setting Up a Database Search with ProteinPilot for iTRAQ Data or 2D Gel Data

4. In the "Process Using" box, select "Paragon" and click "Edit" (Fig. 6C). The "Paragon Method" window appears (Fig. 7). The last method used appears by default. This can be modified to suit the present application and saved accordingly for future searches. It can then be loaded for future searches through the "Paragon Method" drop-down list shown in Figure 7.

   The option for analyzing data with Mascot is also available but will not be covered here. To use Mascot, the data must be loaded in mgf format.

5. In the "Describe Sample" box, select either "iTRAQ 4 Plex" ("Peptide Labeled") or, for 2D gel data, "Identification" from the "Sample Type" drop-down list.

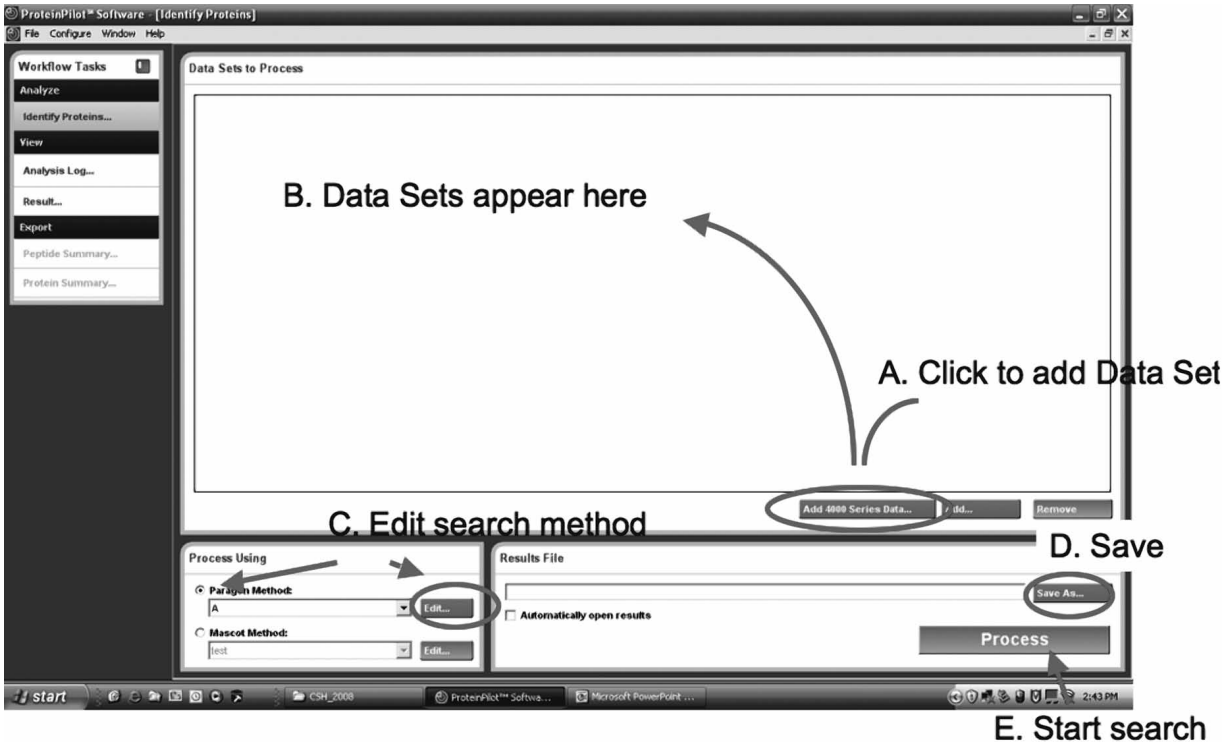


FIGURE 6. The main window of the "Identify Proteins" task in ProteinPilot and workflow for (A,B) loading data for processing, (C) setting search parameters, (D) establishing a directory for saving results, and (E) processing the data.

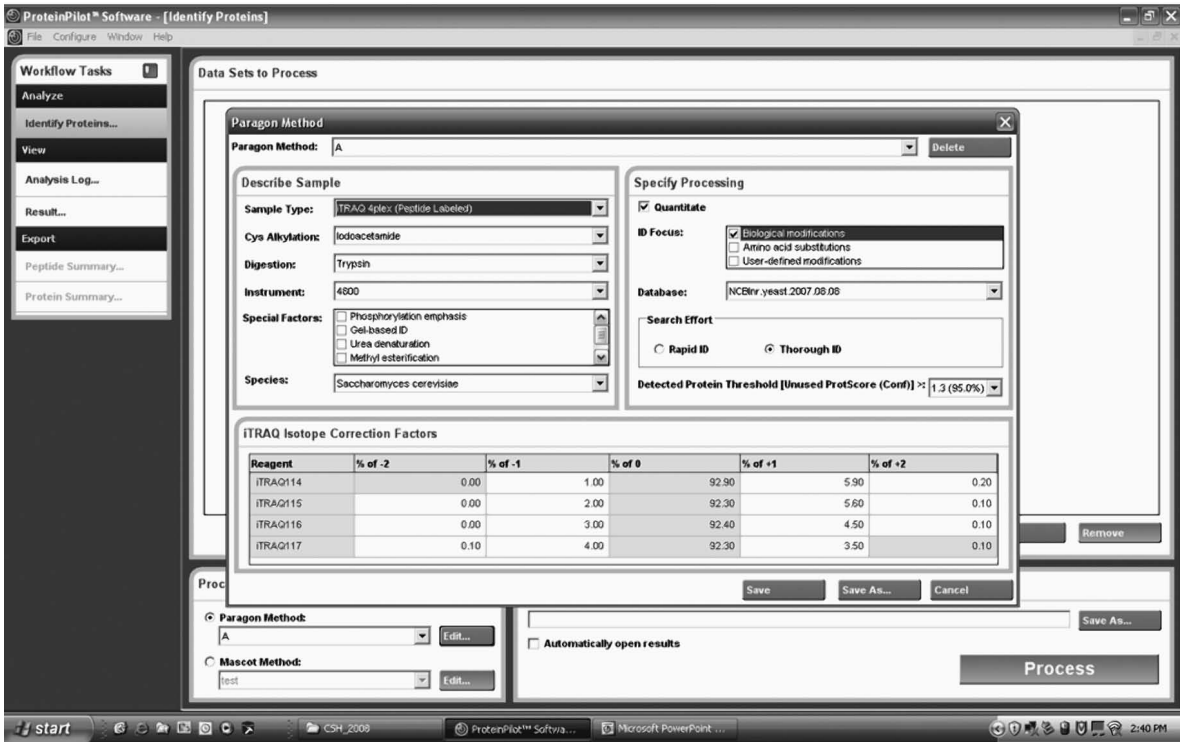


FIGURE 7. The "Paragon Method" editor page.

6. In the "Cys Alkylation" drop-down list, select "Iodoacetamide."
7. In the "Instrument" drop-down list, select "4700."
8. In the "Special Factors" box:
   a. (for iTRAQ data) leave all check boxes unchecked.
   b. (for 2D gel data) select "Gel-based ID."
9. In the "Species" drop-down list, select *"Saccharomyces cerevisiae."*
10. In the "Specify Processing" box:
    a. (for iTRAQ data) make sure the "Quantitate" check box is checked.
    b. (for 2D gel data) make sure the "Quantitate" check box is NOT checked.
11. In the "ID Focus" box, check "biological modifications."
12. In the "Database" drop-down list, select "NCBInr.yeast.2007.08.08."

    Databases in the drop-down list are in FASTA format and located in the directory C:\Applied Biosystems MDS Sciex\ProteinPilot Data\SearchDatabases.
13. In the "Search Effort" box, select "Thorough ID."
14. In the "Detected Protein Threshold" drop-down list, select the desired protein confidence interval threshold. The results will display all protein hits that score above the selected threshold.
15. (for iTRAQ data only) In the "iTRAQ Isotope Correction Factors" box, enter the suggested values as listed in the certificate of analysis that comes with the iTRAQ kit. These values adjust the calculated iTRAQ ratios to compensate for the sub-100% yield in isotopic enrichment of each of the iTRAQ reagents during synthesis.
16. Click "Save As..." to save the search parameters.

    For students in the course, save the parameters by group name and application (e.g., "group A iTRAQ").
17. In the Results File box (Fig. 6D), click "Save As..." to select the results file name and directory. By default, the data will be saved (in .group format) in the directory C:\Applied Biosystems MDS Sciex\ProteinPilot Data\Results, although this can be changed by the user.
18. Click "Process" (Fig. 6E).

Loading Result Files in ProteinPilot

19. In the "Workflow Tasks" task bar, click "Result" (Fig. 6).
20. Select the results file (.group format) from the directory C:\Applied Biosystems MDS Sciex\ProteinPilot Data\Results.

Overview of the Results Screen in ProteinPilot

When a results file is open (Fig. 8), either three or four tabs will be available for data sorting and interpretation. If search parameters were selected for identification only, such as those selected for 2D-gel identifications, then the tabs "Protein ID," "Spectra," and "Summary Statistics" will be available. If quantification parameters were selected in the search, such as those selected for iTRAQ, then the "Protein Quant" tab will also be available.

The "Protein ID" tab (Fig. 8) lists all proteins identified above the confidence interval thresh-

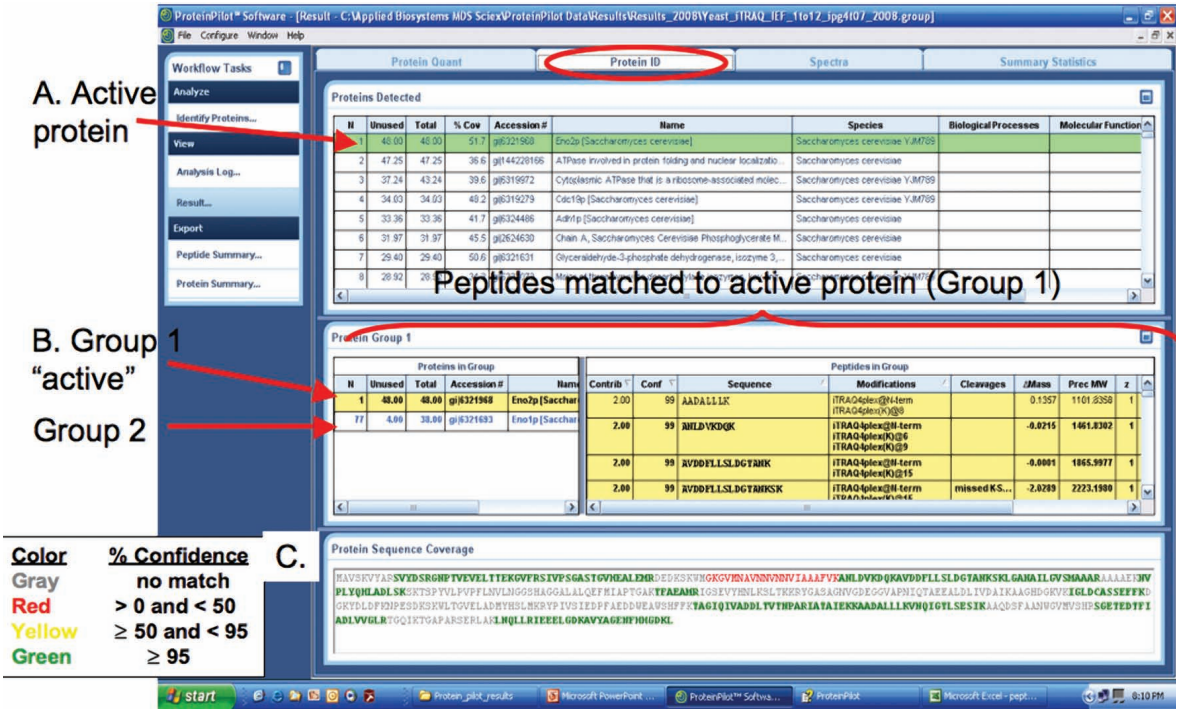


FIGURE 8. The "Protein ID" window of the results file. (A) The active protein is shown in green. (B) All peptide information for the activated protein appears in the "Protein Group" pane. (C) Sequence coverage of the active protein is displayed with color coding (*inset*) in the "Protein Sequence Coverage" pane.

old specified in the search parameters. In the "Proteins Detected" pane, the proteins are listed with scoring information, name, accession number, and species, among other things. Sorting by score, protein name, or any other category is achieved by clicking the column headers. Repetitive clicking on the column header allows for sorting in ascending or descending order. Each protein is assigned two scores: Unused and Total. These are defined below. The active protein in the "Proteins Detected" pane is illuminated green (Fig. 8A). Clicking on a protein row activates it. The "Protein Group" table (Fig. 8B) displays the peptide information that pertains to the activated protein as well as any proteins that have been grouped with it by the Pro Group algorithm. Each peptide has a contrib score which represents the peptide's scoring contribution to the protein's Total Score. The Total Score (third column from the left in the "Protein Group" table) is, therefore, a sum of all contrib scores from peptides assigned to it. However, these are not necessarily unique peptides, so some peptides can be included in the Total Score of more than one protein. Proteins that have peptide hits in common are grouped (Fig. 8B). The Unused Score (second column from the left in the "Protein Group" table) exhibits the peptide scores that have not also been assigned to a higher-scoring protein. It is a reflection of the peptide's uniqueness to the protein. Finally, the "Protein Sequence Coverage" pane (Fig. 8C) displays a view of the sequence coverage of the activated protein. The colors reflect the confidence in the matched peptides assigned to the protein.

The "Spectra" tab lists specific details of all MS/MS spectra that yielded peptide hits. It includes scores, modifications, and a spectral view of the spectrum itself with annotation. If the data were acquired and searched for iTRAQ quantification, then iTRAQ ratios are also included. It also lists other peptides that matched the spectrum with lower confidence. The "Summary Statistics" tab provides details about the number of proteins that were identified at different confidence intervals, including the confidence interval selected for viewing when the search parameters were set. It also

summarizes the results parameters, analysis parameters, and quantification settings that were used to search and analyze the data.

The "Protein Quant" tab displays relative quantification of all confidently identified proteins and only appears if the appropriate settings for quantification were entered in the search parameters. The "Protein Quant" tab (Fig. 9) is arranged similarly to the "Protein ID" tab except that it focuses more on quantification and less on identification. In the "Proteins Detected" pane, proteins are listed with essentially the same information found in the "Protein ID" tab, with the inclusion of iTRAQ ratios. Proteins are activated by clicking on the appropriate row, illuminating it green (Fig. 9A). The user controls what information from the "Protein ID" or "Protein Quant" tab to display. By right clicking any of the column headings, a checkbox appears allowing the user to show or hide information (Fig. 9B). (This feature is available for every table in the software.) The peptide information in the "Peptide Quantitation" pane corresponds to the active protein (Fig. 9C). The "Peptide Quantitation" pane displays information about all peptides assigned to the active protein with emphasis on the iTRAQ ratios measured from the peptide. If the checkbox in the "Used" column is checked, then the iTRAQ ratios for the peptide were used in calculating the overall iTRAQ ratios for the active protein. This provides user flexibility in deciding to use the data or omit it. The "Annotation" column specifies whether the decision to use, or not to use, the quantification data from the peptide was determined by the user or by the software. If it displays "auto," the current status in the "User" column was selected by the software. If it displays "manual," the current status in the "User" column was selected by the user. Mass spectral views of the active peptide iTRAQ reporter ion *m/z* region and precursor *m/z* region are displayed in the bottom pane (Fig. 9D).

The most important information in the "Protein Quant" tab is the determination of which proteins exhibit changes in abundance across samples. It is therefore practical to sort the protein infor-

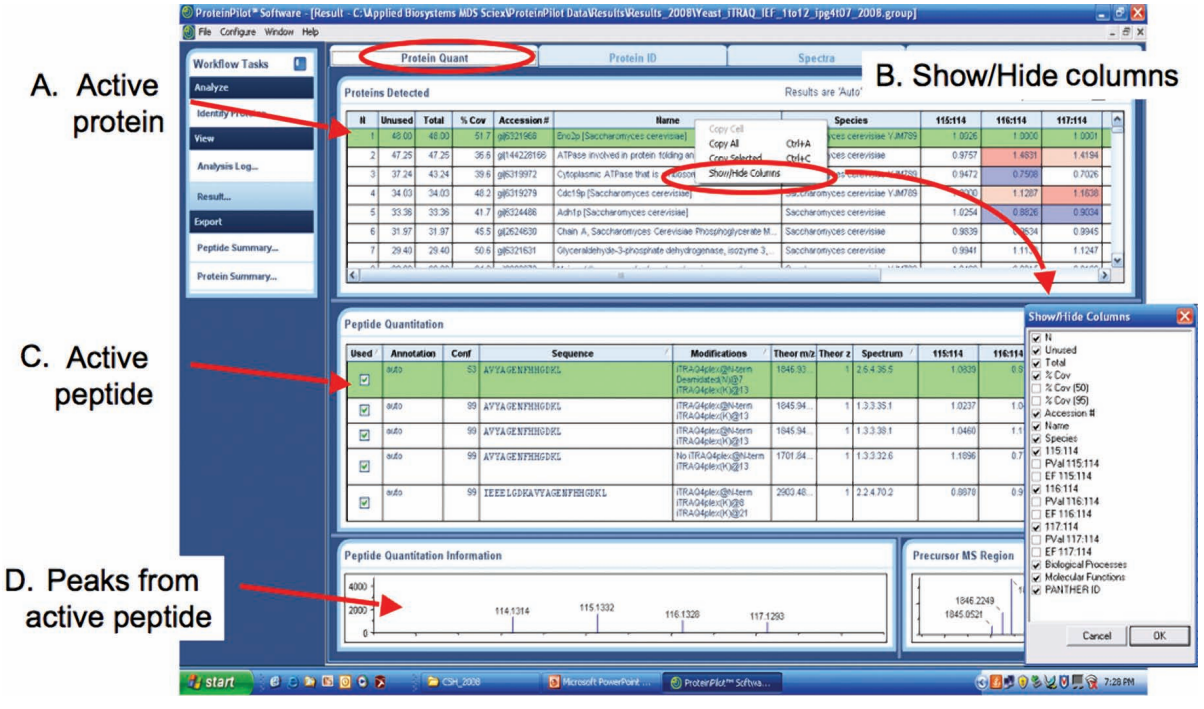
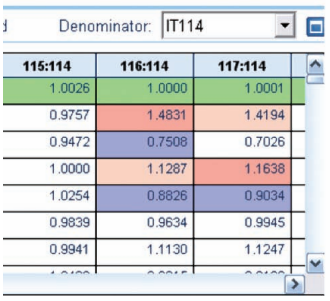


FIGURE 9. The "Protein Quant" window of the results file. (A) The active protein is shown in green. (B) All peptide quantification information for the activated protein appears in the "Protein Group" pane. (C) Mass spectral views of the iTRAQ reporter ion and precursor *m/z* regions are displayed. (D) Columns in any table can be added or hidden with the "Show/Hide Columns" feature.



Color	P-value	Ratio
Dark red	< 0.001	> 1
Medium red	0.001 to < 0.01	> 1
Light red	0.01 to < 0.05	> 1
No color	>= 0.05	Any
Light blue	0.01 to < 0.05	< 1
Medium blue	0.001 to < 0.01	< 1
Dark blue	< 0.001	< 1

FIGURE 10. Color coding scheme representing the *p*-value of the iTRAQ ratios.

mation by iTRAQ ratios. By clicking one of the iTRAQ ratio columns in the "Proteins Detected" pane, the protein table automatically sorts by that column in ascending order. Clicking the column heading again will sort the data by that column in descending order. This provides a quick view for which proteins represent targets for further validation. The iTRAQ ratios in the "Proteins Detected" pane appear with different colored backgrounds. Colors reflect the precision, in the form of a *p*-value, by which the ratio was determined from each protein's peptide hits. The color coding is explained in Figure 10.

PROTOCOL 3

Evaluating an MS/MS Spectrum that Matches a Peptide Sequence from a Database Search Program

In the course, we train students using this protocol to decide whether to accept or reject a peptide sequence that matches an MS/MS spectrum after a database search. In the search research, there are always both true and false identifications due to the random or stochastic matching between experimental and theoretical data. When tandem mass spectra contain limited fragmentation information, the database search engines may lead to incorrect identifications. Especially for proteins or posttranslational modification identified from a single MS/MS spectrum, it is necessary to validate identifications by vigilant inspection of the spectra.

1. Evaluate the MS/MS spectrum quality.

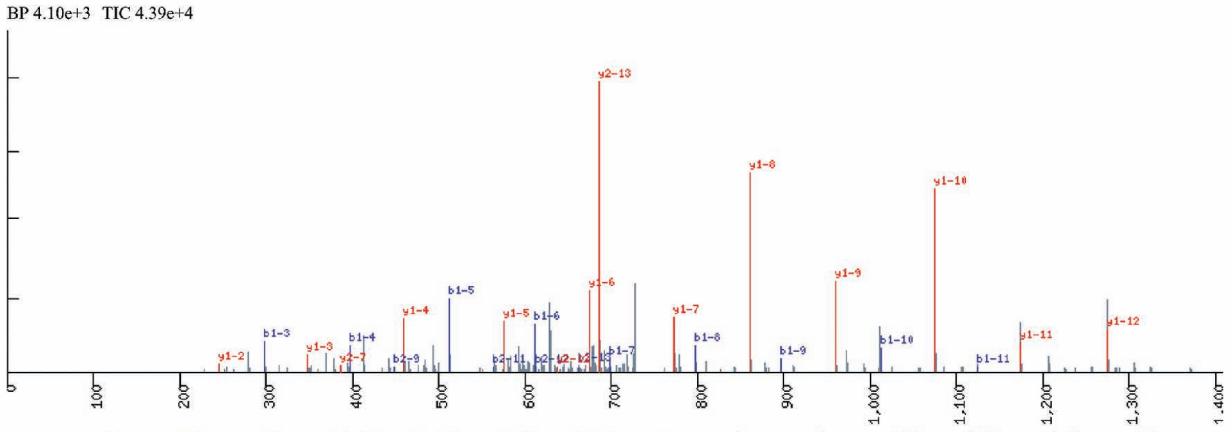
   A strong criterion for spectral quality is the presence of background noise (see Fig. 2A). No background noise and only ions of similar intensity coming up from a flat baseline is an indication that the precursor ion was not a peptide. A spectrum with few ions or low total ion intensity is an indication of a peptide in low abundance and may contain inadequate information to validate the peptide hit.

2. Evaluate how extensively the peptide's predicted b- and y-ions match the product ions in the spectrum (Fig. 11).

   Most of the major product ions in the spectrum should match either the b- or y-ions from the sequence. A large number of unexplained intense peaks throughout the spectrum is an indication of an incorrect identification. Examine the continuity of the b- or y-ion series. One should see a continuous string of 3 to 4 ions of the same series as opposed to isolated ions here and there. Complementing b- and y-ion pairs are a good indication of a correct identification.

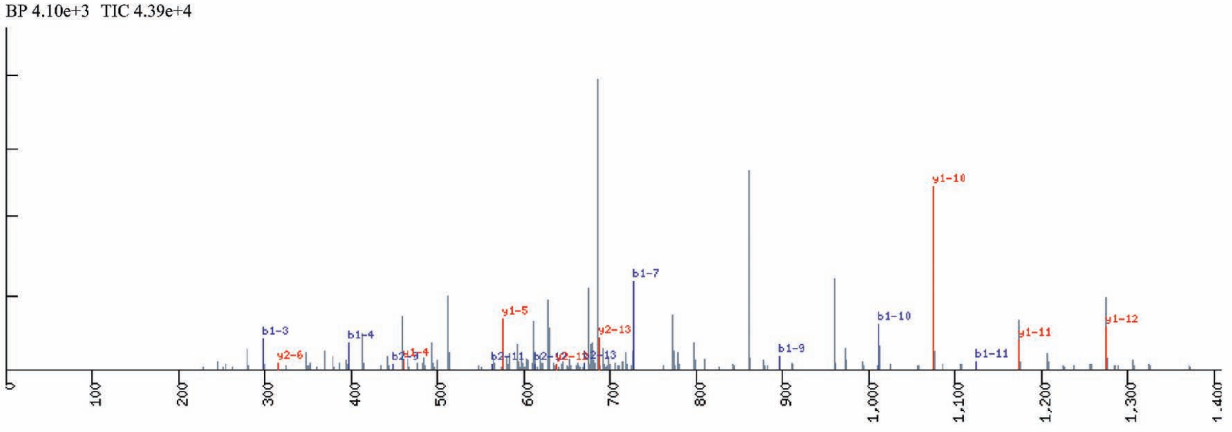
   While tryptic peptides typically generate doubly and triply charged precursor ions during ESI-LC-MS/MS, small tryptic peptides (< 6–8 residues) will be seen commonly as singly charged. Importantly, the short peptides will generate a limited number of fragment ions. The small num-

A



b ions		b1	b2	b1-17	b2-8.5	b1-18	b2-9	y ions		y1	y2	y1-17	y2-8.5	y1-18	y2-9
b1	V	100.1326	50.5663	83.1326	42.0663	82.1326	41.5663	y1	K	147.1742	74.0871	130.1742	65.5871	129.1742	65.0871
b2	P	197.2493	99.12465	180.2493	90.62465	179.2493	90.12465	y2	V	**246.3068**	123.6534	229.3068	115.1534	228.3068	114.6534
b3	T	**298.3544**	149.6772	281.3544	141.1772	**280.3544**	140.6772	y3	T	**347.4119**	174.20595	330.4119	165.70595	329.4119	165.20595
b4	V	**397.487**	199.2435	380.487	190.7435	379.487	190.2435	y4	L	**460.5714**	230.7857	443.5714	222.2857	442.5714	221.7857
b5	D	**512.5756**	256.7878	495.5756	248.2878	**494.5756**	247.7878	y5	D	**575.66**	288.33	558.66	279.83	557.66	**279.33**
b6	V	**611.7082**	306.3541	594.7082	297.8541	593.7082	297.3541	y6	V	**674.7926**	337.8963	657.7926	329.3963	656.7926	328.8963
b7	S	**698.7864**	349.8932	681.7864	341.3932	**680.7864**	340.8932	y7	V	**773.9252**	**387.4626**	756.9252	378.9626	755.9252	378.4626
b8	V	**797.919**	399.4595	780.919	390.9595	779.919	390.4595	y8	S	**861.0034**	431.0017	844.0034	422.5017	**843.0034**	422.0017
b9	V	**897.0516**	**449.0258**	880.0516	440.5258	879.0516	440.0258	y9	V	**960.136**	480.568	943.136	472.068	942.136	471.568
b10	D	**1012.1402**	506.5701	995.1402	498.0701	**994.1402**	497.5701	y10	D	**1075.2246**	538.1123	1058.2246	529.6123	**1057.2246**	529.1123
b11	L	**1125.2997**	**563.14985**	1108.2997	554.64985	1107.2997	554.14985	y11	V	**1174.3572**	587.6786	1157.3572	579.1786	1156.3572	578.6786
b12	T	1226.4048	**613.7024**	1209.4048	605.2024	**1208.4048**	**604.7024**	y12	T	**1275.4623**	**638.23115**	1258.4623	629.73115	**1257.4623**	**629.23115**
b13	V	1325.5374	**663.2687**	1308.5374	654.7687	1307.5374	654.2687	y13	P	1372.579	**686.7895**	1355.579	678.2895	1354.579	677.7895

B



b ions		b1	b2	b1-17	b2-8.5	b1-18	b2-9	y ions		y1	y2	y1-17	y2-8.5	y1-18	y2-9
b1	P	98.1167	49.55835	81.1167	41.05835	80.1167	40.55835	y1	N	133.1039	67.05195	116.1039	58.55195	115.1039	58.05195
b2	V	197.2493	99.12465	180.2493	90.62465	179.2493	90.12465	y2	N	247.2078	124.1039	230.2078	115.6039	229.2078	115.1039
b3	T	**298.3544**	149.6772	281.3544	141.1772	**280.3544**	140.6772	y3	T	348.3129	174.65645	331.3129	166.15645	330.3129	165.65645
b4	V	**397.487**	199.2435	380.487	190.7435	379.487	190.2435	y4	L	**461.4724**	231.2362	444.4724	222.7362	443.4724	222.2362
b5	E	526.6025	263.80125	509.6025	255.30125	508.6025	**254.80125**	y5	N	**575.5763**	288.28815	558.5763	**279.78815**	557.5763	279.28815
b6	T	627.7076	314.3538	610.7076	305.8538	**609.7076**	305.3538	y6	G	632.6283	**316.81415**	615.6283	308.31415	614.6283	307.81415
b7	V	**726.8402**	363.9201	709.8402	355.4201	708.8402	354.9201	y7	I	745.7878	373.3939	728.7878	364.8939	727.7878	364.3939
b8	I	839.9997	420.49985	822.9997	411.99985	821.9997	411.49985	y8	V	844.9204	422.9602	827.9204	414.4602	826.9204	413.9602
b9	G	**897.0517**	**449.02585**	880.0517	440.52585	879.0517	440.02585	y9	T	946.0255	473.51275	929.0255	465.01275	928.0255	464.51275
b10	N	**1011.1556**	506.0778	**994.1556**	497.5778	993.1556	497.0778	y10	E	**1075.141**	538.0705	1058.141	529.5705	**1057.141**	529.0705
b11	L	**1124.3151**	**562.65755**	1107.3151	554.15755	1106.3151	553.65755	y11	V	**1174.2736**	587.6368	1157.2736	579.1368	1156.2736	578.6368
b12	T	1225.4202	**613.2101**	1208.4202	604.7101	**1207.4202**	**604.2101**	y12	T	**1275.3787**	**638.18935**	1258.3787	629.68935	**1257.3787**	**629.18935**
b13	N	1339.5241	**670.26205**	1322.5241	**661.76205**	1321.5241	661.26205	y13	V	1374.5113	**687.75565**	1357.5113	679.25565	1356.5113	678.75565

FIGURE 11. Evaluating the first and second peptide hits to an MS/MS spectrum. Shown are the top two peptide hits for an MS/MS spectrum after a Sequest search of the acquired data using a yeast protein database. (A) Highest-scoring peptide hit (VPTVDVSVVDLTVK) from the yeast Tdh1/2/3 protein (Cn = 4.6). (B) Second-best scoring peptide hit (PVTVETVIGNLTNN) from the yeast Meu1 protein (Cn = 2.0). Overall, the first peptide hit in Panel A is considered accurate because of several important reasons. First, it is a good spectrum because of the overall ion intensity and the presence of strong ion signals above the background noise. Second, the majority of the major fragment ions are labeled as a b- or y-ion. Third, the peptide hit is a canonical tryptic fragment. Fourth, in the table showing experimental ion data matching the theoretical fragmentation ion values, the peptide has a long, continuous string of b- and y-ions. Finally, the b- and y-ion pairs are complementary (e.g., the sum of the complementary b-y ion pairs equals the precursor ion mass). The second hit in Panel B is considered erroneous for a number of reasons. Many of the major ions are unlabeled. The second-best hit is a noncanonical tryptic peptide. The table below the spectrum shows a random matching of experimental data to the theoretical values with no recognizable pattern.

ber of fragment ions hinders the search algorithm's ability to identify the correct peptide and the user's ability to validate the sequence.

It is often not possible to explain all the fragment ions that are observed in a spectrum. However, for doubly and triply charged tryptic peptides, the majority of the most abundant peaks in the *m/z* range above the precursor ion should be evidence of a continuous y-ion series.

Tandem mass spectra that are extremely complex may be caused by fragmentation of two different peptides simultaneously. This can be determined by examining the precursor scan for evidence of two precursor ions. After identifying one peptide, the unidentified peaks in the MS/MS scan can be used for a second database search.

The moderate resolution of ion trap mass spectrometers often limits the ability to directly determine the charge state of the precursor ions. When the charge state of the precursor ions is unknown, database search algorithms will commonly generate multiple charge states (e.g., +2, +3, +4) for the precursor ions. Each precursor charge state is searched separately against the protein database and multiple peptide sequences are reported. It is important for users to accept only one peptide identification.

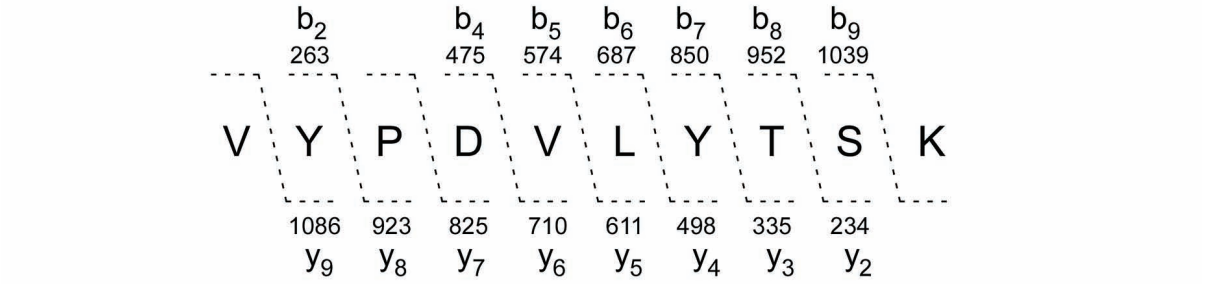
It is important to recognize that the CID conditions used to fragment precursor ions are different for ion trap and TOF/TOF analyzers. Ion trap instruments fragment precursor peptide ions by inducing low-energy collisions with an inert gas, typically helium, within the ion trap analyzer. Once the precursor ion fragments, the resulting fragment ions do not fragment again. In ion traps, the mechanism of CID does not allow for trapping of fragment ions below 28% of the precursor mass. This "1/3 rule" or "low mass cut-off" is evidenced by a lack of ions in the low mass range of the tandem spectra. In the TOF/TOF instrument, on the other hand, fragmentation is variable. The TOF/TOF mass spectrometer is capable of high-energy collisions that can fragment side chain bonds and low-energy collisions that fragment primary peptide bonds (Khatun et al. 2007). Internal fragment ions and y-ions are commonly observed. For users of QTOF and triple quadrupole instruments, CID occurs in a separate radio-frequency collision cell so all of the ions entering the cell are excited, and secondary fragmentation of b- and y-ions may occur. Compared to the ion trap, the TOF/TOF, QTOF, triple quadrupole, and FTICR mass spectrometers all retain the low mass product ions. Finally, while the precursor ions for tryptic peptides using ESI are typically doubly or triply charged, the precursor ions from a MALDI source are generally singly charged ions. For all these reasons, the MS/MS spectra generated from an ion trap, TOF/TOF, QTOF, and other mass analyzers for the same peptide sequence will not be identical and can show substantial differences.

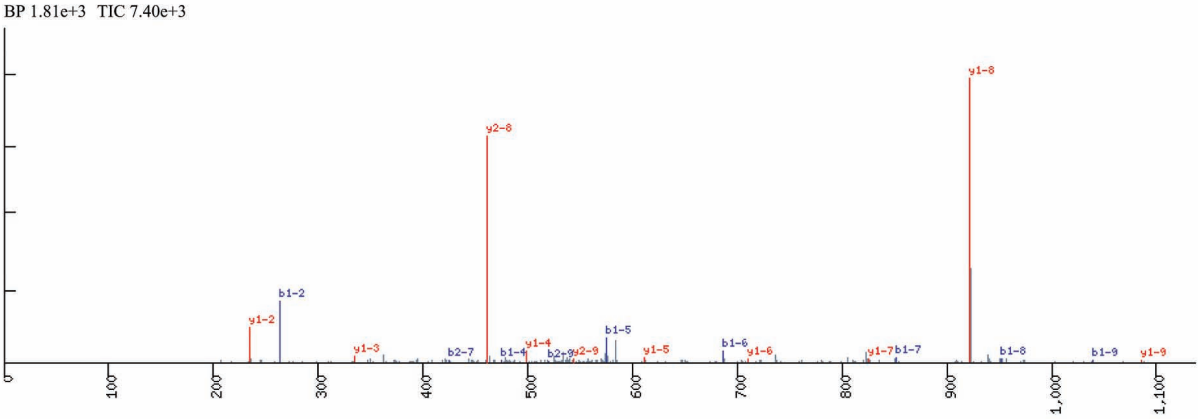
3. Evaluate the peptide sequence and spectrum for unique sequence effects on intensity (Fig. 12).

Fragmentation tends to occur in the middle of the peptide rather than near the termini. In ion trap mass spectrometers, y-ion intensity is typically twice as strong as b-ion intensity (Tabb et al. 2003, 2006). Specific residues tend to have unique effects on fragmentation. Recognizing these effects on the spectrum increases the confidence of the identification. Most notably, peptides with an internal proline have a strong tendency to fragment on the amino-terminal side of the proline residue (Fig. 12). This is commonly referred to as the "proline effect." Aspartic acid residues have a strong tendency to fragment on their carboxy-terminal side (Kapp et al. 2003). Studies of the frequency of peptide fragmentation based on amino acid pairs shows that isoleucine, valine, and leucine residues favor fragmentation on their carboxy-terminal side and glycine and serine favor fragmentation on their amino-terminal sides (Tabb et al. 2003). Peptide bonds between asparagine and glycine are very labile (Kapp et al. 2003).

4. Examine the spectrum for neutral-loss product ions. Both precursor and fragment ions may lose small neutral molecules (Fig. 13).

Neutral losses from the precursor can cause major peaks in the fragmentation spectrum at a lower *m/z* value compared to the precursor value. Most notably, phosphopeptides containing phosphoserine or phosphothreonine residues will readily lose phosphoric acid (–98 Da) when using low-energy CID, such as an ion trap (Schlosser et al. 2001). This observation is extremely useful for recognizing phosphopeptides (Fig. 13). For doubly charged phosphopeptide precursor ions, the most intense ion in the fragmentation spectrum will be the neutral-loss ion (–49 Da from the precur-





b ions		b1	b2	b1-17	b2-8.5	b1-18	b2-9	y ions		y1	y2	y1-17	y2-8.5	y1-18	y2-9
b1	V	100.1326	50.5663	83.1326	42.0663	82.1326	41.5663	y1	K	147.1742	74.0871	130.1742	65.5871	129.1742	65.0871
b2	Y	**263.3086**	132.1543	246.3086	123.6543	245.3086	123.1543	y2	S	**234.2524**	117.6262	217.2524	109.1262	216.2524	108.6262
b3	P	360.4253	180.71265	343.4253	172.21265	342.4253	171.71265	y3	T	**335.3575**	168.17875	318.3575	159.67875	317.3575	159.17875
b4	D	**475.5139**	238.25695	458.5139	229.75695	457.5139	229.25695	y4	Y	**498.5335**	249.76675	481.5335	241.26675	480.5335	240.76675
b5	V	**574.6465**	287.82325	557.6465	279.32325	556.6465	278.82325	y5	L	**611.693**	306.3465	594.693	297.8465	593.693	297.3465
b6	L	**687.806**	344.403	670.806	335.903	669.806	335.403	y6	V	**710.8256**	355.9128	693.8256	347.4128	692.8256	346.9128
b7	Y	**850.982**	**425.991**	833.982	417.491	832.982	416.991	y7	D	**825.9142**	413.4571	808.9142	404.9571	807.9142	404.4571
b8	T	**952.0871**	476.54355	935.0871	468.04355	934.0871	**467.54355**	y8	P	**923.0309**	**462.01545**	906.0309	453.51545	905.0309	453.01545
b9	S	**1039.1653**	**520.08265**	1022.1653	511.58265	1021.1653	**511.08265**	y9	Y	**1086.2069**	**543.60345**	1069.2069	535.10345	1068.2069	534.60345

FIGURE 12. Peptide with a proline residue generates intense fragment ions. The MS/MS fragmentation spectrum of the precursor ion (*m*/*z* 592.92) shows two intense ions. The Sequest algorithm identified the peptide "VYPDVLYTSK" from the yeast Gpm1 protein as the highest scoring hit (Cn 2.2). The two dominant ions are examples of proline peaks illustrating the intense fragment ions formed on the amino-terminal side of proline residues. The y_{1-8} and y_{2-8} are the singly and doubly charged y_8 ions, respectively. The precursor ion was doubly charged. The table below the labeled spectrum shows the observed fragment ions matching the theoretical fragment ions from the peptide sequence. The b2 and y2 columns are the doubly charged b-and y-ion values. The X–17 and –18 columns (where X is b1 or y1) are the singly charged values for the respective b- and y-ions with neutral losses of ammonia and water, respectively. The X–8.5 and X–9 columns (where X is b2 or y2) are the doubly charged values for the respective b- and y-ions with neutral losses of ammonia and water, respectively.

sor). For triply charged ion phosphopeptides, the most intense ion will be the neutral-loss ion (–32.6 Da from the precursor). In an ion trap's fragmentation spectrum of the phosphorylated peptide, the signal from the b- and y-ions is dramatically reduced relative to the neutral-loss peak. The reduced intensity of the b- and y-ion peaks renders a confident identification more difficult. In particular, it is difficult to identify the exact amino acid that is modified if the peptide contains multiple Ser or Thr residues.

For fragment ions that have a neutral loss, pairs of peaks are typically produced. The neutral-loss ion will typically be 10–20% of the intensity of the intact fragment ion. Fragment b- and y-ions that contain Ser, Thr, Glu, or Asp may lose H_2O (–18 Da). b- and y-ions that contain Asn, Gln, Arg, or Lys may lose ammonia (–17 Da). Precursor ions with Gln at their amino termini can readily lose ammonia (–17 Da), generating a neutral loss peak and a predominant b-17 ion series. Precursor ions with oxidized methionines may lose methane sulfenic acid (–64 Da) (Reid et al. 2004).

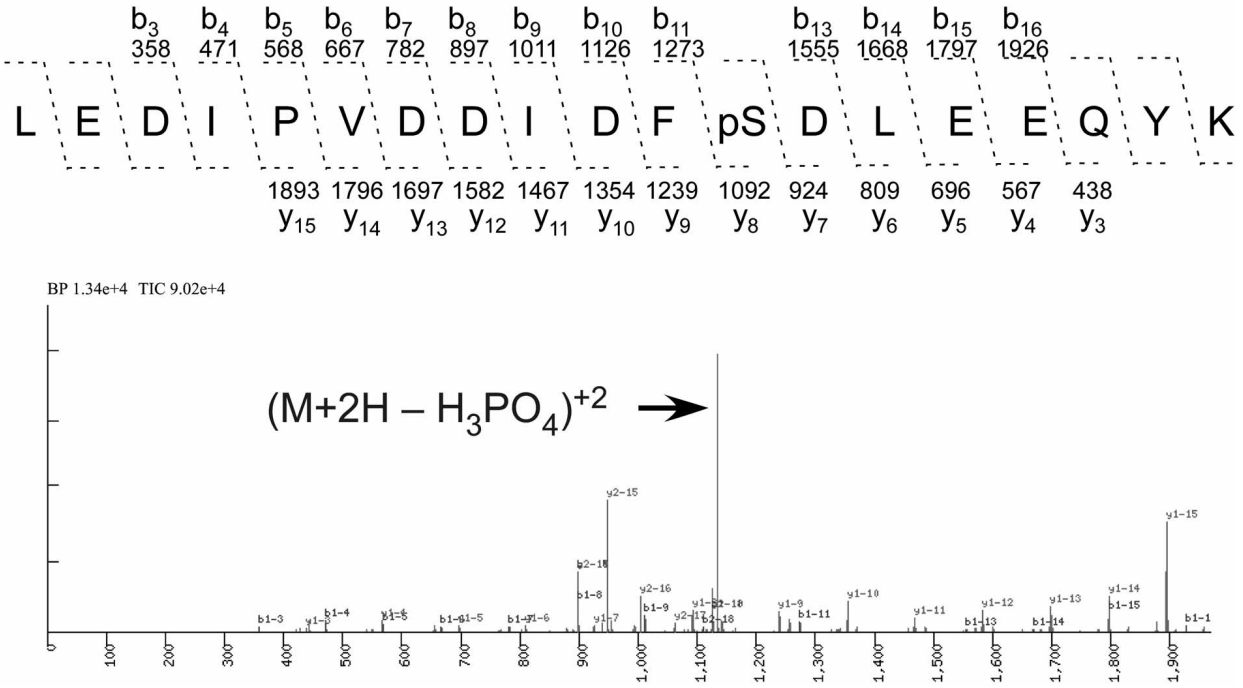


FIGURE 13. MS/MS spectrum of a phosphopeptide. Phosphopeptides can show distinct neutral losses of phosphoric acid (–98 Da) from the precursor and fragment ions. During low-energy CID, especially in ion trap mass spectrometers, phosphopeptides with phosphoserine or phosphothreonine residues readily lose phosphoric acid generating an intense neutral ion in the MS/MS spectrum. In this experiment, to identify phosphorylated peptides, IMAC-Fe^{+3} was used to enrich for phosphopeptides from yeast. The captured peptides were analyzed by LC-MS/MS using an ion trap. This product spectrum was acquired on a 1184 *m/z* precursor ion. The Sequest search identified the phosphopeptide LEDIPVDDIDFS*DLEEQYK as the top hit (Cn = 5.4). Manual analysis of the MS/MS spectrum identified the most intense ion as a neutral loss of –49 (phosphoric acid) from the doubly charged precursor ion. For singly or triply charged phosphopeptide precursor ions, neutral losses of –98 and –32.6 would have been observed, respectively. In this spectrum, a relatively intense proline peak is also observed.

5. Look for independent spectra identifying the same peptide.

   For abundant peptides, different charge states (e.g., +2 and +3) of the identical peptide may be selected for fragmentation. Since the searches for different charge states are independent of each other, the probability that the identification is correct increases if the searches return the same peptide sequence.

6. Evaluate the basicity of the fragment ions.

   Several studies have examined the influence of basic residues on fragment ion intensities (Paizs and Suhai 2002; Tabb et al. 2004). When a tryptic peptide contains a single basic amino acid residue fragmented by CID (Lys or Arg), the product ions that retain the basic residue are generally more intense compared to the other fragment ion. For fragmentation in ion trap mass spectrometers, the y-ions are typically more intense than the b-ions. When triply charged peptide ions are fragmented by CID, the product ions that contain multiple basic residues (Lys, Arg, His) are more likely to be doubly charged (Tabb et al. 2006). For singly charged precursor ions, one of the fragment ion series may dominate over the other series, especially if the terminal residue is a strongly basic Arg. When validating the MS/MS spectrum, b- and y-ion products that follow these observations support the identification.

7. Examine the low mass ions in the MS/MS spectrum.

   In ion trap mass spectrometers, the mechanism of CID does not allow for trapping of fragment masses below 28% of the precursor mass. As described earlier, ion trap mass spectrometers suffer from the "1/3 rule" during fragmentation and fail to retain the low mass ions. However, the

TOF/TOF and QTOF instruments retain the low mass product ions. The low mass ions may reveal information about the sequence composition of the peptides that can be very useful for validating the peptide hits from the database search. Immonium ions are internal product ions produced as a secondary fragmentation of the amide bond during CID. Their structure is represented by $RCH{=}H_2N^+$, where R indicates the amino acid side chain. For an amino acid, its immonium ion is 27 Da less than its residue mass. Each amino acid in a peptide has a characteristic immonium ion. The presence of immonium ions in the low mass region of an MS/MS spectrum can indicate the presence of that amino acid in the peptide. The Tyr (136), His (110), Met (104), Pro (70), Phe (120), Trp (159), Leu/Ile (86), and Val (72) immonium ions are most often observed. The sequence composition of the peptide can be verified using the immonium ions. The carboxy-terminal residue of the peptide can be checked against the product ions in the low-mass region of the spectrum. For peptides with a carboxy-terminal Lys residue, the appearance of an ion at 147 may be detected. If the carboxyl terminus is an Arg, a product ion at 175 may be detected. While b_1 ions are rarely seen and y_2 ions are often low intensity, an intense ion pair in the lower *m/z* range of the MS/MS spectrum separated by 28 Da is frequently observed. The ions correspond to the a_2 and b_2 fragment ions and is the result of the facile loss of CO from the b_2 ion. This pair of ions is commonly called the "a_2/b_2 pair". Appendix 6 lists the *m/z* values of all the possible b_2-ion combinations of amino acid residue masses.

For cases where accurate identification of the peptide sequence is essential for future experiments, the peptide is often chemically synthesized and the fragmentation pattern of the synthetic peptide is compared to the MS/MS fragmentation of the native precursor ion. The two MS/MS spectrum should show identical fragment ions (*m/z*) and relative intensities.

REFERENCES

Bradshaw R.A. 2005. Revised draft guidelines for proteomic data publication. *Mol. Cell. Proteomics* **4:** 1223–1225.

Carr S., Aebersold R., Baldwin M., Burlingame A., Clauser K., and Nesvizhskii A.; Working Group On Publication Guidelines For Peptide And Protein Identification Data. 2004. The need for guidelines in publication of peptide and protein identification data. *Mol. Cell. Proteomics* **3:** 531–533.

Chen Y., Kwon S.W., Kim S.C., and Zhao Y. 2005. Integrated approach for manual evaluation of peptides identified by searching protein sequence databases with tandem mass spectra. *J. Proteome Res.* **4:** 998–1005.

Craig R. and Beavis R.C. 2003. A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun. Mass Spectrom.* **17:** 2310–2316.

Craig R. and Beavis R.C. 2004. TANDEM: Matching proteins with tandem mass spectra. *Bioinformatics* **20:** 1466–1467.

Creasy D.M. and Cottrell J.S. 2004. Unimod: Protein modifications for mass spectrometry. *Proteomics* **4:** 1534–1536.

Dongré A.R., Jones J.L., Somogyi A., and Wysocki V.H. 1996. Influence of peptide composition, gas-phase basicity, and chemical modification on fragmentation efficiency: Evidence for the mobile proton model. *J. Am. Chem. Soc.* **118:** 8365–8374.

Deutsch E. 2008. mzML: A single, unifying data format for mass spectrometer output. *Proteomics* **8:** 2776–2777.

Eng J.K., McCormack A.L., and Yates J.R. 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences. *J. Am. Soc. Mass Spectrom.* **5:** 976–989.

Fenyo D. and Beavis R.C. 2003. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.* **75:** 768–774.

Geer L.Y., Markey S.P., Kowalak J.A., Wagner L., Xu M., Maynard D.M., Yang X., Shi W., and Bryant S.H. 2004. Open mass spectrometry search algorithm. *J. Proteome Res.* **3:** 958–964.

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

Kapp E.A., Schütz F., Reid G.E., Eddes J.S., Moritz R.L., O'Hair R.A., Speed T.P., and Simpson R.J. 2003. Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. *Anal. Chem.* **75:** 6251–6264.

Keller A., Nesvizhskii A.I., Kolker E., and Aebersold R. 2002. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74:** 5383–5392.

Khatun J., Ramkissoon K., and Giddings M.C. 2007. Fragmentation characteristics of collision-induced dissociation in MALDI TOF/TOF mass spectrometry. *Anal. Chem.* **79:** 3032–3040.

Link A.J., Eng J., Schieltz D.M., Carmack E., Mize G.J., Morris D.R., Garvin B.M., and Yates J.R. III. 1999. Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17:** 676–682.

MacCoss M.J., Wu C.C., and Yates J.R. III. 2002. Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Anal. Chem.* **74:** 5593–5599.

McCormack A.L., Schieltz D.M., Goode B., Yang S., Barnes G., Drubin D., and Yates J.R. III. 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.

Nesvizhskii A.I., Keller A., Kolker E., and Aebersold R. 2003. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **75:** 4646–4658.

Paizs B. and Suhai S. 2002. Towards understanding some ion intensity relationships for the tandem mass spectra of protonated peptides. *Rapid Commun. Mass Spectrom.* **16:** 1699–1702.

Paizs B.and Suhai S. 2005. Fragmentation pathways of protonated peptides. *Mass Spectrom. Rev.* **24:** 508–548.

Pappin D.J., Hojrup P. and Bleasby A.J. 1993. Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.* **3:** 327–332.

Peng J., Elias J.E., Thoreen C.C., Licklider L.J., and Gygi S.P. 2003. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: The yeast proteome. *J. Proteome Res.* **2:** 43–50.

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

Reid G.E., Roberts K.D., Kapp E.A., and Simpson R.I. 2004. Statistical and mechanistic approaches to understanding the gas-phase fragmentation behavior of methionine sulfoxide containing peptides. *J. Proteome Res.* **3:** 751–759.

Resing K.A., Meyer-Arendt K., Mendoza A.M., Aveline-Wolf L.D., Jonscher K.R., Pierce K.G., Old W.M., Cheung H.T., Russell S., Wattawa J.L., et al. 2004. Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Anal. Chem.* **76:** 3556–3568.

Sadygov R.G., Cociorva D., and Yates J.R. III. 2004. Large-scale database searching using tandem mass spectra: Looking up the answer in the back of the book. *Nat. Methods* **1:** 195–202.

Sadygov R.G., Liu H., and Yates J.R. III. 2004. Statistical models for protein validation using tandem mass spectral data and protein amino acid sequence databases. *Anal. Chem.* **76:** 1664–1671.

Schlosser A., Pipkorn R., Bossemeyer D., and Lehmann W.D. 2001. Analysis of protein phosphorylation by a combination of elastase digestion and neutral loss tandem mass spectrometry. *Anal. Chem.* **73:** 170–176.

Shilov I.V, Seymour S.L., Patel A.A., Loboda A., Tang W.H., Keating S.P., Hunter H.L., Nuwaysir L.M., and Schaeffer D.A. 2007. The Paragon algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. *Mol. Cell. Proteomics* **6:** 1638–1655.

Tabb D.L., Fernando C.G., and Chambers M.C. 2007. MyriMatch: Highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J. Proteome Res.* **6:** 654–661.

Tabb D.L., Friedman D.B., and Ham A.J. 2006. Verification of automated peptide identifications from proteomic tandem mass spectra. *Nat. Protoc.* **1:** 2213–2222.

Tabb D.L., Huang Y., Wysocki V.H., and Yates J.R. III. 2004. Influence of basic residue content on fragment ion peak intensities in low-energy collision-induced dissociation spectra of peptides. *Anal. Chem.* **76:** 1243–1248.

Tabb D.L., McDonald W.H., and Yates J.R. III. 2002. DTASelect and Contrast: Tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* **1:** 21–26.

Tabb D.L., Smith L.L., Breci L.A., Wysocki V.H., Lin D., and Yates J.R. III. 2003. Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Anal. Chem.* **75:** 1155–1163.

Washburn M.P., Wolters D., and Yates J.R. III. 2001. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19:** 242–247.

Wysocki V.H., Tsaprailis G., Smith L.L., and Breci L.A. 2000. Mobile and localized protons: A framework for under-

standing peptide dissociation. *J. Mass Spectrom.* **35:** 1399–1406.

Yang X., Dondeti V., Dezube R., Maynard D.M., Geer L.Y., Epstein J., Chen X., Markey S.P., and Kowalak J.A. 2004. DBParser: Web-based software for shotgun proteomic data analyses. *J. Proteome Res.* **3:** 1002–1008.

Zhang B., Chambers M.C., and Tabb D.L. 2007. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *J. Proteome Res.* **6:** 3549–3557.