Piecing together MS data

Researchers are using a variety of data analysis programs to identify proteins from their mass spectra.

BY RANDALL C. WILLIS

There was a time when protein identification was a simple matter of isolating a protein band from a polyacrylamide gel, digesting it with enzymes, and performing amino acid sequencing on the resulting peptides. But as proteomic technologies have improved, the variety of proteins to be isolated and characterized has increased dramatically to the point where researchers are now trying to analyze hundreds to thousands of spots on a single 2-D gel.

And with this increase in the number of samples comes an increased need to identify these same proteins in a manner that is high-throughput, low-cost, and easily automated.

To that end, researchers have come to rely on robotic systems to run and pick peaks from gels, purify and digest protein samples, and load peptide mixtures into mass spectrometers for peptide analysis and sequencing. But the real work begins at the end of this process, with the deconvolution of the resulting MS and MS/MS data into protein identification badges (*1*).

Pieces of a puzzle

When a protein is digested with a protease (usually trypsin) and studied by MS, the spectral peaks are comparable to jigsaw puzzle pieces. But unlike a traditional jigsaw puzzle in which the pieces are assembled to form the picture, MS data analysis software works in reverse, virtually chopping up a series of proteins (pictures) from a sequence database and generating virtual MS spectra to see which one best matches the real spectra (Figure 1).

These peptide-mapping programs first match the virtual and real spectra and then score the match based on how many of the calculated masses match the real peptide masses, and how well. David Lubman and his colleagues at the University of Michigan, Ann Arbor, recently applied two peptide mapping programs, MS-Fit and PeptIdent (see box, "Peptide-matching programs"), to their multidimensional LC system. In their study of differential protein expression between normal and cancerous ovarian epithelial cell lines, the researchers identified several proteins that appeared to be down-regulated in the cancerous line (*2*).

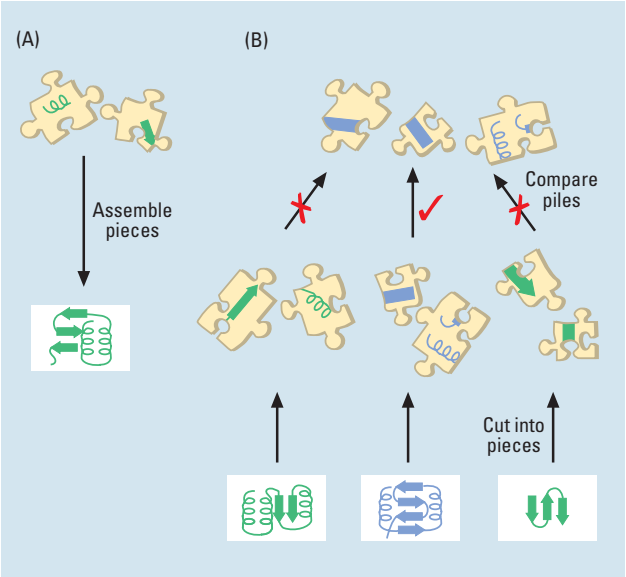

Figure 1. Puzzling problems. In standard protein identification using mass spectrometry (A), the spectral information (puzzle pieces) is assembled to form the whole protein (picture). Using MS data analysis software (B), however, protein sequences from a database (bottom) are computationally cut into virtual mass spectra (middle) to determine which one best compares with the actual data (top).

The one drawback of such a system, however, is that it is biased toward high-molecular-weight proteins for which more virtual peptides can be calculated and that therefore have a higher probability of matching randomly. To some extent, repeating the experiment with a different protease will offset this bias, because it will generate a different set of peptide fragments. If the real and virtual peptides also match with this second protease, then there is good reason to believe that the parent proteins are the same.

Sequential analysis

Another way of avoiding this bias is by sequencing the proteins using MS/MS. In this case, the original peptide fragments from the first mass spectrometer pass into a second spectrometer, where they are fragmented. The resulting MS peaks represent a protein sequence ladder, in which adjacent peaks differ by the weight of the amino acid that has been lost in the fragmentation from one peptide peak to the next.

One such program, PepSea, was used by Uzma Atif and colleagues at GlaxoSmithKline (Essex and Welwyn, U.K.) to study the renal toxicity of the aminoglycoside antibiotic gentamicin (*3*). The hope was that a profile of kidney protein or some other biomarker expression before and during antibiotic treatment would serve to warn of impending medical trouble.

Developed by Matthias Mann of the University of Southern Denmark and MDS Proteomics (both in Odense), PepSea relies on short peptide sequences or sequence tags that can be probed against various genomic databases (*4*). Unlike many of the other programs that require some knowledge of protein coding regions, the short tags used by PepSea allow a random search to occur in such a way that DNA sequences that have yet to be translated into protein can be probed as well, widening the possibilities of success.

Alternatively, rather than simply determine the peptide sequence from the spectra and probe a DNA or protein database, the real MS/MS spectrum can be com-

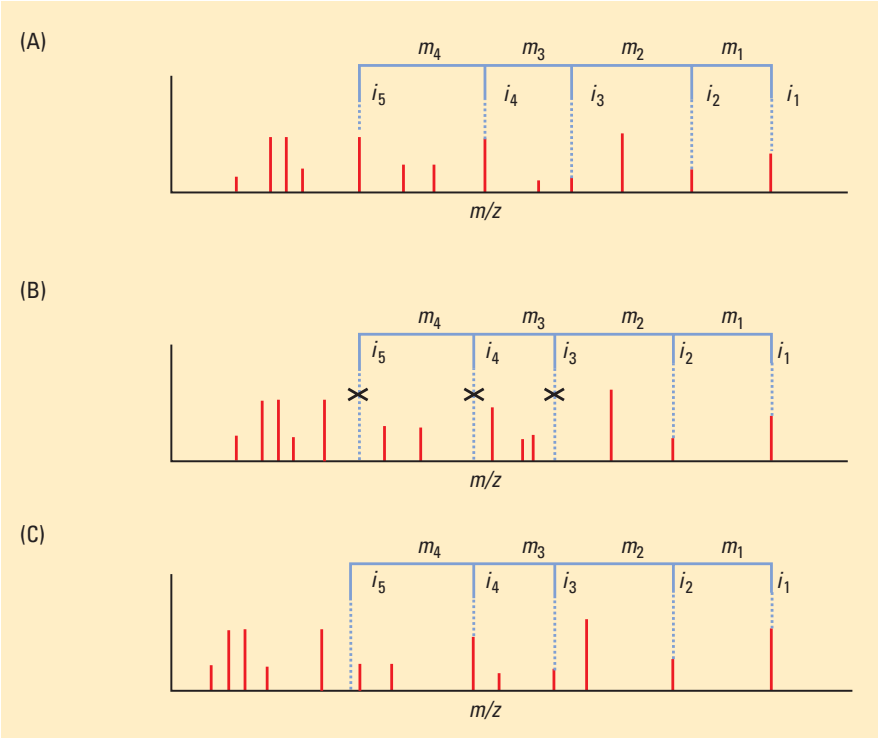

Figure 2. Spectral analysis with a beat. Using SALSA, virtual mass spectra that match the real data (A) are scored higher than those that only match a couple of peaks (B). Furthermore, SALSA allows virtual spectra to move along the real *m/z* access to detect samples in which the peptide masses have been altered by posttranslational modification or mutation (C).

pared with a variety of virtual MS/MS spectra calculated from the protein sequences in the database.

One program that works this way is SEQUEST, developed by John Yates III (now at the Scripps Research Institute in La Jolla, CA) and colleagues at the University of Washington in Seattle and distributed by ThermoFinnigan (San Jose, CA). Recently, Yates's group used SEQUEST to analyze the proteome of a yeast strain and was able to detect and identify almost 1500 individual proteins, including several integral membrane proteins, which are notoriously difficult to work with (*5*).

Messing with mixtures

Rarely, however, does a researcher look at a single protein with a single set of peaks. In shotgun proteomics, a mixture of proteins is digested by proteolysis, and one or two cycles of liquid chromatography separate the resulting peptides. These peptides are then subjected to MS/MS. Their data are analyzed using algorithms such as SEQUEST to determine the identity and sequence of the peptides. But SEQUEST gives only information about individual peptides.

DTASelect, also developed by Yates's group, is a Java-based program that assembles and evaluates the SEQUEST data, effectively reassembling the peptides into the parent protein (*6*). The summation component of DTASelect collects the key data from a SEQUEST search. The evaluation component then applies user-selected criteria to the matches, and the results are generated by the reporter function. Because experiments are not necessarily run under identical conditions or in the same laboratory, which can severely affect data interpretation, Yates's group developed Contrast as a tool that can compare data from multiple experiments and different criteria.

This problem of protein mixture complexity is further complicated in electrospray ionization-MS because of the challenges associated with determining the charge states of the parent peptides. If a given peptide ion can have a charge state of +1, +2, +3, or more, then a model spectrum must be calculated for each of those possibilities. Thus, accurately knowing the charge state in advance would greatly decrease the computational effort. Therefore Yates's group developed a new SEQUEST code (which they call 2to3) that narrows the charge state and eliminates low-quality mass spectra (*7*).

Puzzling posttranslational modifications

The MS/MS sequence-based mechanism, however, can hamper the accuracy of these spectra-matching programs, because proteins can be altered by posttranslational modifications, gene mutations, and splice variants that cannot always be predicted from their sequences. If a peptide is modified, its relative peak position or its fragmentation pattern might be altered. To accommodate this problem, Daniel Liebler and his colleagues at the University of Arizona, Tuscon, developed SALSA (scoring algorithm for spectral analysis) (*8*).

In SALSA, a predicted ion series is defined as a group of ions ($i_1, i_2, i_3, \ldots, i_n$) separated by specific m/z values ($m_1, m_2, m_3, \ldots, m_n$) or the mass of the next amino acid along the probe peptide chain. The first predicted ion (i_1) is lined up with the highest m/z in the actual MS/MS spectra, and the algorithm then looks for the next ion (i_2) in the series, connecting the ions sequentially (Figure 2).

The user can choose to align the spectra on the basis of either the primary (y- or b-) or secondary (y″- or b″-) ions, or a paired combination thereof. The program then scores the alignment, taking into account factors such as the search strategy used, length of the search motif, number of ions that match the series, and intensity of the scored ions.

Testing the algorithm

The researchers tested the algorithm first against mass spectra of bovine serum albumin (BSA), comparing the results with those obtained using SEQUEST. Overall, SALSA detected the same peptides as SEQUEST, but then it went beyond the other algorithm's performance by identifying peptides that had undergone modification through processes such as cysteine oxidation or N-terminal carbamylation. This capability arises from the fact that modified peptides still follow the same ion patterns as described above, with the sole exception that most or all of the series are shifted along the m/z axis by the mass of the modification.

SALSA was then tested against the spectra of a mixture of BSA and human serum albumin (HSA) peptide fragments. Many of these peptides differ by only one or two amino acids. When probed with a BSA sequence, SALSA identified both of the appropriate BSA and HSA peptides, but the latter had a much lower score. Probing with the HSA sequence, however, gave the opposite result, proving SALSA's merit as another tool in the proteomic arsenal.

As with most areas of software development in which researchers and programmers strive to accelerate information processing, the spectrum of MS data analysis tools is expanding at an incredible rate with few signs of slowing. What were standalone programs just a few years ago are now becoming integral pieces of analytical machinery. SEQUEST is part of ThermoFinnigan's BioWorks platform. Likewise, ProFound is just one part of Genomic Solutions' PWRHouse system, and ProID works with the API QSTAR system from Applied Biosystems. Thus, as go the mass spectrometers, so go their data analysis programs; the next generation of software will surely be caught in the hardware web.

The spectrum of MS data analysis tools is expanding at an incredible rate with few signs of slowing.

References

(1) Beavis, R. C.; Fenyo, D. *Proteomics: A Trends Guide,* July 2000, pp 22–27.
(2) Kachman, M.; et al. *Anal. Chem.* **2002,** *74,* 1779–1791.
(3) Charlwood, J.; et al. *J. Prot. Res.* **2002,** *1,* 73–82.
(4) Kuster, B.; et al. *Proteomics* **2001,** *1,* 641–650.
(5) Washburn, M. P.; Wolters, D.; Yates, J. R. III. *Nature Biotechnol.* **2001,** *19,* 242–247.
(6) Tabb, D. L.; McDonald, W. H.; Yates, J. R. III. *J. Prot. Res.* **2002,** *1,* 21–26.
(7) Sadygov, R. G.; et al. *J. Prot. Res.* **2002,** *1,* 211–215.
(8) Liebler, D. C.; et al. *Anal. Chem.* **2002,** *74,* 203–210.

Randall C. Willis is a senior associate editor of *Modern Drug Discovery*. Send your comments or questions about this article to mdd@acs.org or to the Editorial Office address on page 3. ■

Peptide-matching programs

MS profile searches	
MOWSE	www.matrixscience.com
MS-Fit	http://prospector.ucsf.edu
MultiIdent	www.expasy.ch/tools/multiident
PeptIdent	www.expasy.ch/tools/peptident.html
PIUMS	www.biobridge.se
ProtoCall MS	www.cgen.com
MS/MS profile searches	
AutoMod	www.micromass.co.uk
MS-Tag	http://prospector.ucsf.edu
PepFrag	http://prowl.rockefeller.edu
ProID	www.appliedbiosystems.com
SEQUEST	www.thermofinnigan.com
Both	
Intellimarque	www.shimadzu-biotech.net
Mascot	www.matrixscience.com www.daltonics.bruker.com
PepSea	www.protana.com/solutions/software/default.asp
ProFound	www.genomicsolutions.com