

A Tool To Visualize and Evaluate Data Obtained by Liquid Chromatography-Electrospray Ionization-Mass Spectrometry

Xiao-jun Li,* Patrick G. A. Pedrioli, Jimmy Eng, Dan Martin, Eugene C. Yi, Hookeun Lee, and Ruedi Aebersold

The Institute for Systems Biology, 1441 North 34th Street, Seattle, Washington 98103-8904

We present a software tool for visualizing data obtained from analyzing complex peptide mixtures by liquid chromatography (LC) electrospray ionization (ESI) mass spectrometry (MS). The data are represented as a two-dimensional density plot. For experiments employing collision-induced dissociation (CID), links are embedded in the image to the CID spectra and the corresponding peptide sequences that are represented by the respective feature. The image provides an intuitive method to evaluate sample quality and the performance of an LC-ESI-MS system and can be used to optimize experimental conditions. Local patterns of the image can also be used to identify chemical contaminants and specific peptide features. Therefore, this software tool may have broad application in MS-based proteomics.

Mass spectrometry (MS)-based proteomics has the capacity to identify and quantify hundreds to thousands of proteins in a single experiment and has become a powerful analytical tool in biological and medical research.^{1–6} In a typical MS-based proteomics experiment, proteins are harvested from cells, tissues, or body fluids, dissolved and denatured in solution, and enzymatically digested into peptide mixtures. These peptide mixtures are then separated by multidimensional liquid chromatography (LC) and identified by methods of MS.^{3,7} In cases in which the analytes are labeled with stable isotopes, accurate quantitative information can also be obtained in the same operation.^{2,4,8} In a common implementation, a reversed-phase microcapillary LC system is coupled online with an electrospray ionization (ESI) tandem mass spectrometer. Peptides eluting from the LC column are im-

mediately ionized by ESI and subjected to mass measurement (survey scan). The instrument, under computer control, then automatically selects specific precursor ion(s) for fragmentation in a collision cell, thus generating collision-induced dissociation (CID) or tandem mass (MS/MS) spectra. These CID spectra are then searched against sequence databases using software tools such as SEQUEST,⁹ Mascot,¹⁰ or ProteinProspector,¹¹ which leads to the identification of the peptide sequences. Spectra collected in survey scanning mode can be used to reconstruct peptide chromatograms using software tools such as XPRESS² or ASAPRatio,¹² from which the accurate quantities of isotopically labeled peptides can be derived. In such cases, proteins in the original samples are therefore both identified and quantified by assembling data obtained from peptides that represent each specific protein.

Despite its success, this technology remains challenging and several factors are crucial for the success of LC-MS/MS experiments. These include the sample quality, the quality of the separation system used, and the performance of the mass spectrometer. Because the outcome of an LC-MS/MS experiment depends on many variables, it is almost impossible to optimize the system by systematically optimizing individual variables. Troubleshooting in the case of suboptimal performance is difficult and time-consuming.

Visualization tools play an important role in evaluating the performance of many instruments and experiments. To monitor MS performance or the performance of an LC-MS experiment, mass spectrometer manufacturers provide several visualization tools to display individual MS spectra, the total ion current, baseline peaks, etc. While these tools provide detailed insight into a specific performance characteristic of an LC-MS/MS experiment, such as the quality of CID spectra, mass resolution, or chromatographic peak resolution, they make it difficult to extract an overview of the overall success of the experiment. Therefore, a visualization tool that is easy to use and offers transparent diagnostic features for common problems encountered in LC-MS/MS analyses would be useful.

* Corresponding author. E-mail: xli@systemsbiology.org. Fax: (206)732-1299.

- (1) Aebersold, R.; Mann, M. *Nature* **2003**, *422*, 198–207.
- (2) Han, D. K.; Eng, J.; Zhou, H.; Aebersold, R. *Nat. Biotechnol.* **2001**, *19*, 946–951.
- (3) Washburn, M. P.; Wolters, D.; Yates, J. R., 3rd. *Nat. Biotechnol.* **2001**, *19*, 242–247.
- (4) Ong, S. E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M. *Mol. Cell. Proteomics* **2002**, *1*, 376–386.
- (5) Hansen, K. C.; Schmitt-Ulms, G.; Chalkley, R. J.; Hirsch, J.; Baldwin, M. A.; Burlingame, A. L. *Mol. Cell. Proteomics* **2003**, *2*, 299–314.
- (6) Von Haller, P. D.; Yi, E.; Donohoe, S.; Vaughn, K.; Keller, A.; Nesvizhskii, A. I.; Eng, J.; Li, X. J.; Goodlett, D. R.; Aebersold, R.; Watts, J. D. *Mol. Cell. Proteomics* **2003**, *2*, 428–442.
- (7) Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R., 3rd. *Nat. Biotechnol.* **1999**, *17*, 676–682.
- (8) Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R. *Nat. Biotechnol.* **1999**, *17*, 994–999.

- (9) Eng, J. K.; McCormack, A. L.; Yates, J. R., III. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976–989.
- (10) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551–3567.
- (11) Clauser, K. R.; Baker, P.; Burlingame, A. L. *Anal. Chem.* **1999**, *71*, 2871–2882.
- (12) Li, X. J.; Zhang, H.; Ranish, J. A.; Aebersold, R. *Anal. Chem.* **2003**, *75*, 6648–6657.

We present here a visualization tool, called Pep3D, which displays LC-ESI-MS data in a two-dimensional (retention time (RT) vs m/z) density plot. A Pep3D image graphically resembles a two-dimensional (2D) polyacrylamide gel electrophoresis (PAGE) image,¹³ in that the darkness of individual features displayed in the 2D array reflects the intensity of the detected signals. While similar images have been reported in the MS literature using the Xcalibur software (ThermoFinnigan) or other tools,^{14–17} a number of unique features set the Pep3D image apart from those tools: (1) Images are created very quickly. (2) Pep3D allows users to zoom into any interesting range of time and m/z . (3) Pep3D allows users to display those precursor ions selected for CID, all peptides successfully identified in the MS analysis, or both, with boxes whose color varies across a gradient from green to red to display the corresponding identification scores. (4) All CID boxes and peptide boxes have embedded common gateway interface (CGI) links allowing immediate display of the corresponding CID spectrum and its peptide identification. Pep3D can therefore be used to evaluate LC-ESI-MS system performance and identify features of the separated peptides.

EXPERIMENTAL SECTION

Data presented here were collected from various projects in our laboratory. Experimental protocols used to generate these data have been described in other publications.^{2,6,8} Protocols at <http://www.proteomecenter.org/protocols.php> were used to label samples with ICAT reagent.⁸ Peptide samples were analyzed on an LCQ Classic or an LCQ Deca (ThermoFinnigan) ion trap mass spectrometer or on a Q-ToF Ultima (Waters) quadrupole time-of-flight mass spectrometer. CID spectra were searched against corresponding databases using SEQUEST.⁹ Peptide identifications were verified with PeptideProphet¹⁸ and organized by INTERACT.²

RESULTS

Pep3D Algorithm. The Pep3D software tool is written in C, and the current version runs on a Linux operating system. Pep3D images are generated on the fly in PNG file format using a GD graphics library (<http://www.boutell.com/gd/>). It normally takes a few seconds on a Pentium III personal computer to generate a Pep3D image for a 3-h LC-ESI-MS run. The Pep3D software tool is capable of displaying LC-ESI-MS data from all types of mass spectrometers for which the data can be represented in the mzXML data format.¹⁹ The mzXML file format was developed recently to represent MS data from all types of mass spectrometers in a standard XML file format. A detailed description of mzXML

Pep3D Image for LC-ESI-MS Data

This program displays LC-ESI-MS data stored in mzXML format in a Pep3D image. Developed by Dr. Xiao-jun Li at [Institute for Systems Biology](http://www.proteomecenter.org/). Data: October 8, 2002

Specify parameters here:

- Full path of **interact-data.htm** or **mzXML** file:
Example: /data2/search/xiaojun/ASAPRatio_demo/interact-1vs1-data.htm
Or: /data2/search/dan/17_mix_folder/*.mzXML
- Data in "interact-data.htm": original filtered
- M/Z range: Full >= <=
Grid₂ Image Grid₁
- Elution time range (in min): Full >= <=
Grid_{0.5} Image Grid₂
- Intensity range: unit of background >= <=
- Display peptides: None Image Grid₂
Score type: PeptideProphet probability Score >= <=
- Mapping function: Linear
- Image type: Full

Generate Pep3D image Save as Pep3D.htm

Figure 1. Screen capture image of Pep3D web interface.

schema can be found at <http://www.proteomecenter.org/software.php>. Currently, data generated from LCQ (ThermoFinnigan) ion trap, Q-ToF Ultima (Waters), and QSTAR (ABI/Sciex) mass spectrometers can be represented in the mzXML file format and displayed by Pep3D images. Embedded links to CID spectra or peptide identities are developed by using server side graphics. CID spectra are plotted by using the gnuplot software tool (<http://www.gnuplot.info/>). The Pep3D source code and compiled executable are available under an open source license at <http://www.proteomecenter.org/software.php>.

The Pep3D software tool has a web interface, developed in CGI programming, for users to specify the contents of Pep3D images. A screen capture image of the interface was plotted in Figure 1. The options allowed by the software tool include the following. (1) Input file: INTERACT data file "interact-data.htm" or mzXML files are accepted. Data in "interact-data.htm" can be filtered by using INTERACT filtering tools.² Multiple mzXML files can be selected simultaneously. (2) M/Z range: With the option "Full", the whole m/z range of an experiment is displayed; with the option "Selected", only the m/z range specified by ">=" and "<=" is displayed. Empty input in ">=" or "<=" is replaced by the lower or upper boundary of the whole m/z range. The m/z range is divided into pixels of size specified by "Grid". A pixel is plotted as a rectangular of width specified by "Image grid". (3) Elution time range (in min): This option is similar to the " m/z range" option. (4) Intensity range: With the option "unit in background", the software automatically estimates the average intensity of background ions within the image range. The "Intensity range" is then specified in unit of this average intensity by ">=" and "<=". With the option "Absolute", "Intensity range" is specified directly by ">=" and "<=". Empty input in ">=" or "<=" is replaced by the lowest or highest ion intensity within the image range. (5) Display peptides: With the option "None", only raw LC-ESI-MS data are displayed; with the option "CID", all precursor ions are displayed along with raw data; with the

- (13) Rabilloud, T. *Proteomics* **2002**, *2*, 3–10.
 (14) Binz, P. A.; Muller, M.; Walther, D.; Bienvenut, W. V.; Gras, R.; Hoogland, C.; Bouchet, G.; Gasteiger, E.; Fabbretti, R.; Gay, S.; Palagi, P.; Wilkins, M. R.; Rouge, V.; Tonella, L.; Paesano, S.; Rossellat, G.; Karmime, A.; Bairoch, A.; Sanchez, J. C.; Appel, R. D.; Hochstrasser, D. F. *Anal. Chem.* **1999**, *71*, 4981–4988.
 (15) Berger, S. J.; Lee, S. W.; Anderson, G. A.; Pasa-Tolic, L.; Tolic, N.; Shen, Y.; Zhao, R.; Smith, R. D. *Anal. Chem.* **2002**, *74*, 4994–5000.
 (16) Palmblad, M.; Ramstrom, M.; Markides, K. E.; Hakansson, P.; Bergquist, J. *Anal. Chem.* **2002**, *74*, 5826–5830.
 (17) Griffin, T. J.; Lock, C. M.; Li, X. J.; Patel, A.; Chervetsova, I.; Lee, H.; Wright, M. E.; Ranish, J. A.; Chen, S. S.; Aebersold, R. *Anal. Chem.* **2003**, *75*, 867–874.
 (18) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74*, 5383–5392.
 (19) Pedrioli, P.; Eng, J.; Hubley, R.; Pratt, B.; Nilsson, E.; Aebersold, R., submitted.

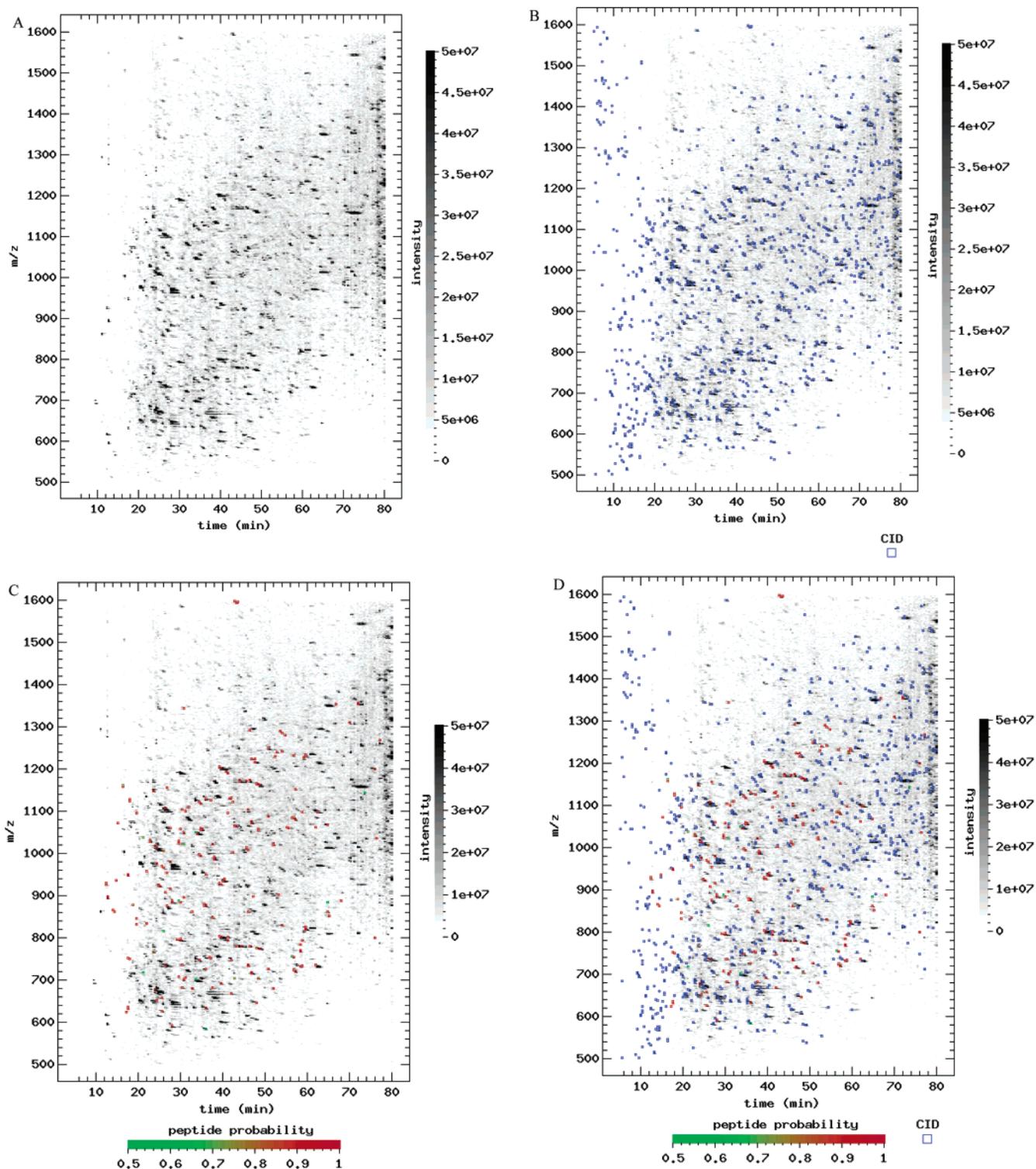


Figure 2. (A) A Pep3D image displaying raw LC-ESI-MS data. The whole image is discretized into identical rectangular pixels. The shading of each pixel is determined by the signal intensity in the respective pixel. The intensity scale is denoted to the right of the image. (B) A Pep3D image displaying LC-ESI-MS data and highlighting all precursor ions selected for CID. Precursor ions are marked with blue boxes with centers specified by retention times and m/z values. (C) A Pep3D image displaying LC-ESI-MS data and highlighting the positively identified peptides. Peptides are marked with boxes in the same way as precursor ions in (B) except for their color. The hue of a peptide box on a gradient of green to red is determined by the corresponding PeptideProphet probability and displayed below the image. (D) A Pep3D image displaying LC-ESI-MS data, all precursor ions selected for CID, and all identified peptides. Precursor ions and peptides are marked in the same way as in (B) and (C).

option "Peptide", all peptides identified with identification scores within a range specified by "Score >=" and "<=" are displayed along with raw data. Currently SEQUEST Xcorr scores⁹ and

PeptideProphet probabilities¹⁸ are accepted as identification scores. The type of scores can be easily extended to include scores generated by other sequence searching engines; the option "All"

CID Spectrum Display

Developed by Dr. Xiao-jun Li at [Institute for Systems Biology](#).

Data: October 8, 2002

- scan number: 988
- m/z: 877.080322
- time: 35.929333
- ID:
0.9986 851 . /fraction23.0988.0988.2 1750.8 (+2.4) 3.8589 0.274 1832.6 1 19/ 22
IPI: IPI00022463.1 R. CLVEKGDVAFVK. H 1.691 +- 0.066 (3.877%) 1.533 +- 0.415 (27.067%) 76

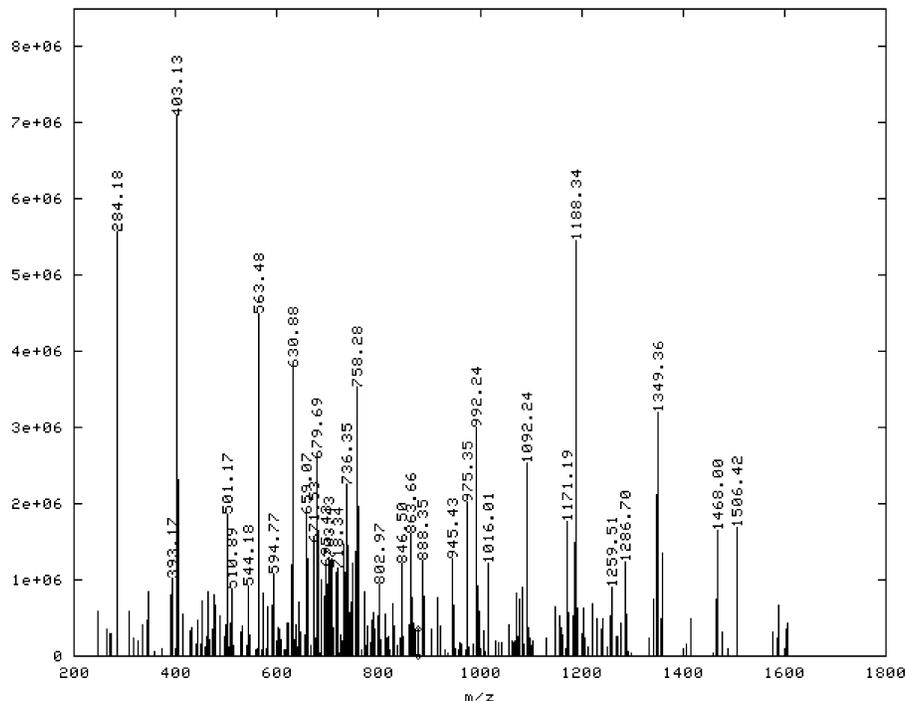


Figure 3. Screen capture image of a CGI-generated web page displaying peptide information including the scan number, the m/z value, and the retention time of the precursor ion and the corresponding CID spectrum. Information on peptide identification was copied from the INTERACT data file. If the web page is for an unidentified precursor ion, information on peptide identification will be replaced by "None".

is equivalent to the combination of options "CID" and "Peptide". CID and/or peptide boxes are plotted in a size specified by "Image Grid". Empty input in "Score >=" or "<=" is replaced by the lowest or highest score with the data. Peptides are displayed only if the input file is "interact-data.htm". (6) Mapping function: With the option "Linear", the darkness of a pixel is linear to its ion intensity; with the option "Log (base 10)", the darkness of a pixel is linear to the logarithm (base 10) of its ion intensity. (7) Image type: With the option "Full", pixels with ion intensity greater than the upper boundary of "Intensity range" are plotted in completely dark while pixels with ion intensity less than the lower boundary of "Intensity range" are plotted in completely white; with the option "Section", pixels with ion intensity outside "Intensity range" are plotted in completely white. (8) Save option: In addition to viewing Pep3D images on a screen, one may also save them in an html file. By varying these options, one can zoom in any time or m/z range of a Pep3D image, visualize only pixels within a certain intensity range, change scales between ion intensity and pixel darkness, and display peptides of various identification scores, etc. These options add flexibility and power to the visualization tool.

Pep3D Features. The Pep3D software tool can display LC-ESI-MS experimental data in four different types of images; an example of each type is shown in Figure 2. In Figure 2A, raw LC-ESI-MS data are plotted as a density plot with the RT in

minutes on the abscissa and the m/z value of the detected signals on the ordinate. The whole image is discretized into rectangular pixels of identical size with time width and m/z height specified by the user. After discretization, ion signals in all MS spectra are assigned to individual pixels. All ion intensities within a pixel are summed together and averaged by the number of MS spectra collected over the duration of the pixel; the result is the ion intensity of the pixel. If a pixel has no ion signals, its ion intensity is set to zero. In the Pep3D image, the ion intensity of each pixel is specified by the shade of the pixel over a gray scale. The relationship between the ion intensity of a pixel and its darkness is denoted along the image. The intensity scale is specified by the user. The Pep3D image visualizes LC-ESI-MS ion signals of the whole experiment in a very transparent manner.

In Figure 2B, all precursor ions selected for CID are highlighted with blue boxes on the density plot in Figure 2A. The center of each CID box is specified by the RT and the m/z value of the corresponding precursor ion. There is a CGI link embedded behind each CID box. When a CID box is clicked, a CGI program opens a new window that displays the scan number, the m/z value, and the RT of the corresponding precursor ion, as well as the CID spectrum with its major fragment ions labeled with the respective m/z values (see Figure 3 for an example). This feature of the Pep3D image allows users to easily connect individual

features in the Pep3D image with corresponding CID spectra. It also provides users with a method to discover and interpret certain patterns in the Pep3D image and to survey the performance of automated precursor ion selection. In the data displayed in Figure 2B, there were 1196 precursor ions selected for CID, a fact that is automatically reported by the software tool. It is apparent from the figure that the majority of high-intensity MS signals were selected for CID during this experiment while the majority of low-intensity MS signals were not selected. As demonstrated in Supporting Information, this information can be used to evaluate sample loading and sample separation.

In Figure 2C, all the peptides identified in the experiment and assigned a PeptideProphet¹⁸ probability of no less than 0.5 are highlighted with color boxes. The color varies across a gradient from green to red and is determined by the corresponding PeptideProphet probability. Except for their color, these boxes are plotted in the same fashion as those boxes indicating CID attempts (Figure 2B) and they also include embedded CGI links behind them. By clicking on a peptide box, the corresponding peptide identity entry in the INTERACT data file is displayed, along with all information for that precursor ion. Through the peptide identity entry within the link, the corresponding protein identity, peptide identity, peptide mass, peptide charge state, etc., are all displayed (see Figure 3 for an example). The display in Figure 2C indicates the distribution and the number of the peptides that were successfully identified in the experiment. Among 1196 precursor ions selected for CID, 221 were successfully identified. The number of identified peptides is automatically reported. For unknown reasons, a large proportion of high-intensity ions were not assigned a sequence with a high degree of confidence, a fact that is immediately apparent from the Pep3D image. This observation is rather common in most LC-MS/MS analyses, and possible explanations include peptide modifications, database deficiencies, and difficulties of generating quality CID spectra from the precursor ions. Instead of PeptideProphet probability, SEQUEST Xcorr score can also be used for displaying identified peptides.

Figure 2D combines the features described in Figure 2B and C, where all precursor ions selected for CID and all peptides assigned a PeptideProphet probability of no less than 0.5 are highlighted with color boxes. This image provides an entry point to investigate why certain CID spectra yielded identifications while others failed to do so. For example, one may reanalyze unidentified high-quality CID spectra by de novo sequencing or database searching with additional peptide modifications allowed. As demonstrated in Supporting Information, this display also makes it easy to identify chemical contaminants within the sample. Figure 2D indicates 1196 CID attempts and 221 peptide identifications; ~18.5% of all CID spectra were positively identified. The likelihood of a CID attempt producing positive peptide identification varied significantly across the plot. By comparing this percentage in different areas, a researcher may identify high-percentage areas and adjust LC-MS settings accordingly. This type of information

can be valuable for optimizing LC-MS settings to improve the quality of CID identifications.

DISCUSSION

A Pep3D image graphically resembles a 2D-PAGE image. However, most spots in a Pep3D image represent peptides, not proteins as in a 2D-PAGE image. A small percentage of spots in a Pep3D image may belong to chemical contaminants such as solvent molecules, lipids, or polymers. Since a peptide may be ionized into more than one charge state during ESI,²⁰ more than one spot may represent the same peptide. The darkness of a peptide spot is not indicative of its absolute concentration because different peptides have different ionization efficiencies during ESI. An exception to this statement is when the peptides are identical yet in different isotopic forms. The relative darkness of a particular peptide spot across different LC-MS samples may reflect the relative abundance of the peptide in the corresponding samples if appropriate abundance normalization can be carried out.

We describe here a visualization tool, Pep3D, to display LC-ESI-MS data. The tool is transparent and easy to interpret and provides informative, comprehensive, and sophisticated insights into the performance of an LC-MS system and the analytes detected. As illustrated in Supporting Information, a Pep3D image can be used to evaluate sample quality, diagnose system performance, and discover peptide features. Pep3D images can also be used to check the reproducibility and consistency of different sample analyses. Furthermore, by comparing peptide spots across parallel samples, one can potentially use Pep3D images to identify unique peptide spots capable of distinguishing different samples, a feature that has potential usage for biomarker discovery. One can also use the tool to collect unidentified high-quality CID spectra and subject them to further analysis, from which some posttranslationally modified peptides may be identified. Owing to these unique features, the Pep3D software tool may have a broad application in LC-MS-based proteomics. The tool is freely available via an open source license and is compatible with any type of mass spectrometer for which the data can be translated into the mzXML data format. A new version of the software tool is planned for Windows operating system and will be available to the public as soon as it is finished.

ACKNOWLEDGMENT

This project has been funded in part with federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, under Contract N01-HV-28179 and from National Institutes of Health Grant R33 CA93302. The ISB is the recipient of a generous gift from Merck and Co. D.M. was supported by NIH Grant CA97282.

SUPPORTING INFORMATION AVAILABLE

Examples illustrating the application of Pep3D software tool. This material is available free of charge via the Internet at <http://pubs.acs.org>.

Received for review November 20, 2003. Accepted April 1, 2004.

AC035375S

(20) Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. *Science* **1989**, *246*, 64–71.