Analyzing Proteomes and Protein Function Using Graphical Comparative Analysis of Tandem Mass Spectrometry Results*

K. Jill McAfee‡§, Dexter T. Duncan¶, Michael Assink, and Andrew J. Link§

Although generating large amounts of proteomic data using tandem mass spectrometry has become routine, there is currently no single set of comprehensive tools for the rigorous analysis of tandem mass spectrometry results given the large variety of possible experimental aims. Currently available applications are typically designed for displaying proteins and posttranslational modifications from the point of view of the mass spectrometrist and are not versatile enough to allow investigators to develop biological models of protein function, protein structure, or cell state. In addition, storage and dissemination of mass spectrometry-based proteomic data are problems facing the scientific community. To address these issues, we have developed a relational database model that efficiently stores and manages large amounts of tandem mass spectrometry results. We have developed an integrated suite of multifunctional analysis software for interpreting, comparing, and displaying these results. Our system, Bioinformatic Graphical Comparative Analysis Tools (BIGCAT), allows sophisticated analysis of tandem mass spectrometry results in a biologically intuitive format and provides a solution to many data storage and dissemination issues. Molecular & Cellular Proteomics 5:1497–1513, 2006.

The rapid advance of proteomics is driving the development of new methods and applications for dissecting networks of protein interactions, identifying posttranslational modifications that modulate biological processes, and monitoring changes in whole proteomes. Over the past several years, LC-MS/MS has emerged as a powerful method to identify and quantify large numbers of proteins. Using LC-ESI-MS/MS or LC-MALDI-MS/MS, tens to hundreds of thousands of fragmentation spectra can be collected in a single tandem mass spectrometry experiment from purified protein complexes, subcellular fractions, or whole cell lysates (1–5). Analyzing this wealth of data to construct models of protein function, protein structure, or cell states is a major challenge. Current software applications to identify proteins from tandem mass spectrometry experiments typically deliver a simple list of accession numbers and gene names with minimal biological annotation. Most output reports focus on the experimental mass spectrometry information rather than the biological content. Presenting proteins in a more biologically intuitive format based on functional similarity and expression levels via graphical summaries would significantly simplify the interpretation of large lists of proteins.

With the many spectra being collected, primary interpretation of the data to identify peptides and proteins has become dependent on computer analysis. Numerous computer algorithms have been developed to compare the measured values of the precursor ion and its fragmentation ions to the theoretical masses of peptides and fragmentation products derived from protein sequences in a database. Three of the most widely used search algorithms of this type, SEQUEST, MASCOT, and X!Tandem, return for each spectrum the peptide sequences in a protein sequence database that best match the spectral data (6–11). However, determining whether the peptide sequence truly represents the data is more difficult. MASCOT and X!Tandem calculate the probability that the identified peptide is not a stochastic match. Although several approaches have been proposed to derive similar probability scores from SEQUEST output as well (12–15), manual interpretation of the spectra is still required in most analyses for the validation of individual matches of a spectrum to a peptide or the even more problematic validation of posttranslationally modified peptides.

The identified peptide sequences must be assembled into proteins that represent the initial experimental sample to allow meaningful data interpretation. Early generation assembler applications used a simple scoring threshold or applied layers of filters at the peptide and protein levels to assemble peptides into a list of proteins (3, 4, 16, 17). However, this process was complicated by non-unique peptide sequences, which match multiple proteins in a database, and the wide range of scoring confidences for peptide matches, which increased the ambiguity in the final list of proteins. Probability-based methods, such as that used by ProteinProphet, and peptide-centric approaches, such as Isoform Resolver and parsimony analysis, have more recently been developed to assist investigators construct protein profiles from peptide lists generated by search algorithms (18–21).

Along with protein identification, quantification of the abso-
lute or relative expression levels of the large numbers of proteins from tandem mass spectrometry experiments has been the focus of much research. A number of elegant in vivo and in vitro stable isotopic labeling approaches have been described (17, 22–29). Although these approaches offer high precision, the cost or technical limitations of implementing them on a routine basis make their widespread application impractical. Alternative label-free methods include calculation of the number of peptide hits for each protein, computation of the percentage of protein sequence coverage by the identified peptides, integration of spectral peak intensities, and summation of the number of spectra identifying each protein (30–34). A new label-free approach, the protein abundance factor (PAF), 1 which provides semiquantification of protein abundance, utilizes the total number of non-redundant spectra that correlate significantly to each cognate protein (35). By normalizing this spectral count to the molecular weight of the cognate protein, the PAF allows us to compare the abundance of unrelated proteins in the same sample and across multiple samples. 2 Although these label-free methods are less precise than stable isotope methods, they are much easier to implement and provide an indication of significant changes in expression levels.

Guidelines and formats for the storage, dissemination, and analysis of results are also current challenges associated with tandem mass spectrometry data (36–38). 3 Two primary approaches have been proposed to address these issues. The first approach applies the extensible markup language (XML) format for the storage of peaks data and search results (8, 39–42). 4 XML is a simple, flexible text format derived from the standardized general markup language (SGML) originally designed for electronic publishing. Unfortunately the XML format does not fully represent the experimental information contained in the native file formats, and the XML file size requires a large amount of storage space. Additionally data storage or computation methods based directly on XML may encounter performance and scalability problems (43).

The second approach utilizes a relational database for storage of tandem mass spectrometry data and results (19, 20, 44–46). 5–8 In a relational database, data are stored in named tables consisting of columns and rows. Data alike in purpose and granularity (level of detail) are grouped together within a single table or a set of related tables. The tables can be further related to each other regardless of granularity by including a column of common information in each. Data can be mined using complex queries to filter and match data motifs or to join multiple tables. Furthermore a relational database allows the transition away from instrument-dependent or proprietary software. Many of the current relational database approaches focus on organizing and managing experimental proteomic data. A relational database offers an ideal environment for developing tools for rigorous analysis and interpretation of MS/MS results. Finally although sophisticated analysis and comparative data visualization methods have progressed in the functional genomics field of DNA microarray analysis (47–51), similar methods for the analysis and comparison of proteomic MS/MS results have not been extensively integrated into proteomics software.

To address these issues, we have developed a relational database model and integrated suite of analysis software, known as Bioinformatic Graphical Comparative Analysis Tools (BIGCAT). BIGCAT allows data generated by different mass spectrometers and MS/MS search algorithms to be stored, manipulated, and compared independently of instrument and search algorithm source. The multifunctional, biologically intuitive data visualization and manipulation tools, integrated analytical and comparative methods, and complex data mining applications provide access to homogenized tandem mass spectrometry data and results. For example, protein identifications in an LC-MALDI-TOF-TOF experiment coupled with search results from the Mascot algorithm reside in the same database structure with the results obtained in an LC-ESI-MS/MS experiment using the SEQUEST search algorithm. The user can compare these data directly with no additional manipulation or prior knowledge of the data source. The source of the data is made transparent to the user by the database and software design.

To aid in biological interpretation, BIGCAT generates protein profiles based on predicted biological categories and the estimated relative expression levels of proteins. With web-based browser access, the system supports on-line collaboration and dissemination of results. The robust complex query capabilities of the relational database allow users to identify biologically significant posttranslational modifications from spectra exhibiting neutral loss signatures. Additionally protein

---

1 The abbreviations used are: PAF, protein abundance factor; BIGCAT, Bioinformatic Graphical Comparative Analysis Tools; Cn, cross-correlation; GB, gigabyte; HTML, hypertext markup language; IP, immunoaffinity purification; MB, megabyte; MIAPE, minimum information about a proteomics experiment; PHP, hypertext preprocessor; XML, extensible markup language; ID, identification; iTRAQ, isobaric tags for relative and absolute quantitation; SAGA, Spt-Ada-Gcn5-acetyltransferase complex; SQL, structured query language; 2D, two-dimensional; TFIIID, transcription factor IID; Taf, TATA-binding protein-associated factor; HLA, human lymphocyte antigen; RADARS, Rapid, Automated, Data Archiving and Retrieval System; LIMS, laboratory information management system.

2 V. Gerbasi, A. Farley, J. Jennings, J. McAfee, D. Duncan, and A. Link, manuscript in preparation.

3 Minimum information about a proteomics experiment (MIAPE), psdev.sourceforge.net/gps/miapemIAPE_Protocols_2.0.doc.

4 mzData, psdev.sourceforge.net/ms/.

5 OpenMS, The OpenMS Project, open-ms.sourceforge.net/project.html.

6 Proteome Research Information Management Environment (PRIME), prime.proteome.med.umich.edu.


abundance data from BIGCAT can be analyzed with software and statistical applications originally designed for DNA microarray analysis. Finally the BIGCAT system not only provides rigorous and sophisticated analysis of tandem mass spectrometry results but also provides a solution to data storage and dissemination issues within the proteomics community.

**EXPERIMENTAL PROCEDURES**

**System Configuration—** Oracle Server Enterprise Edition 9.2.0.4 was installed and configured on a Dell Poweredge 4600 with dual 2.8-GHz Xeon processors and 4-GB 266-MHz SDRAM, running Linux RedHat ES 2.1, kernel 2.4.9. This system has eight 146-GB 10K RPM Ultra 320 SCSI hard drives, and dual Intel Pro 1000MT Copper Gigabit Network Adapters.

The custom-built web server was configured with Linux RedHat 7.3, kernel 2.4.7-10, Apache HTTP (hypertext transport protocol) Server Version 1.3, and PHP 4.3.2 with GD (graphics draw) library. Perl 5.8.6 was installed with modules DBD (database drive) 1.16 and DBI (database interface) 1.48 along with Oracle Server Client 9.2.0.4. The current configuration uses Oracle Enterprise Edition Version 9.2.0.7.

**Relational Database—** We created a relational database model, based on the Oracle Relational Database Management System, that utilizes the minimum number of database tables necessary to accommodate MS/MS experimental data, reduce data redundancy, and allow easy and timely retrieval of data. In general, tandem mass spectrometry data and results are hierarchical rather than star-schema in structure with one-to-many cardinality between inheritance levels. The relational database structure is ideal for storing this type of data. Our current schema design accommodates SEQUEST, Mascot, and X!Tandem formats.

A database-generated experiment ID uniquely identifies each experiment in the database. This experiment ID is the primary key of the main EXPERIMENTS table and acts as a foreign key in other tables to establish a hierarchical relationship. When the parent experiment record is removed from the EXPERIMENTS table, the experiment ID is used to identify the related child records in other tables and remove those records as well. Primary key-foreign key relationships are used throughout the database model to provide referential integrity. Although full referential integrity among tables provides maximum data integrity, the performance impact associated with this design was determined to be unacceptable during prototype testing. Therefore, the current schema includes those primary key-foreign key relationships that provide data integrity among tables while optimizing database data manipulation language (DML) performance.

For larger experiment data tables, the partitioning feature of Oracle dynamically separates the data into multiple storage areas to improve the performance of maintenance operations, backups, transactions, and queries. The physical storage of these data is allocated among multiple table spaces and data files using the experiment ID. Data backup is accomplished via the built-in export and backup utilities of Oracle, thereby allowing the data from a large number of experiments to be stored in a few easily managed and archived data files.

**Browser-driven Front-end Applications—** A series of browser-driven applications were developed by applying a combination of Oracle-stored procedures and functions for data retrieval and filtering. PHP, Perl, HTML, Java, and Javascript were used in development of the graphical displays. The front-end software is deployed on a web server and is supported by Internet Explorer 6.0 and higher, Firefox 1.0 and higher, Netscape 8.1 and higher, and Opera 8.5 and higher.

Guest access to BIGCAT with the experimental data used in Figs. 3–9 is available on line (bigcat.mc.vanderbilt.edu/BIGCAT/) by selecting the “BIGCAT in Action” link. All original software is freely available as open-source along with installation and configuration instructions, including prerequisite information, on the Link Lab website (linklab.mc.vanderbilt.edu). The source code is made available under the artistic license from the authors.

**Tandem Mass Spectrometry Analysis of Protein Complexes and Whole Proteomes—** Purification of Saccharomyces cerevisiae transcription and translation complexes or polyribosomal fractions and multidimensional liquid chromatography coupled with tandem mass spectrometry analysis using a Thermo Electron LCQ ion trap mass spectrometer have been described previously (35, 52, 53). LC-MAaldi-TOF-TOF data were acquired from a trypticized, ITRAQ labeled, human A431 cell lysate fractionated by reverse phase microcapillary HPLC and directly spotted onto an Applied Biosystems MALDI plate using a Shimadzu Accuspot plate spotting system. An Applied Biosystems 4700 MALDI-TOF/TOF instrument was used to acquire spectra from each spot using GPS Explorer in LC-MALDI mode.

**RESULTS**

**A Relational Database to Store and Organize Mass Spectrometry Results—** We developed a relational database for tandem mass spectrometry data and search results based on the Oracle Relational Database Management System. This database serves as a back-end for our browser-driven suite of analysis applications. Our system called BIGCAT has allowed us to transition away from instrument-dependent or proprietary software toward a configuration unlimited by instrument type or operating procedure. Our database schema or structure (Fig. 1) can easily be adapted to any SQL-based relational database software, such as MySQL, PostgreSQL, or Microsoft SQL Server. As tandem mass spectrometry results are hierarchical rather than star-schema in structure, with one-to-many cardinality between inheritance levels, we designed our database to take advantage of the inherent parent-child relationships in the data (Fig. 1).

Oracle was initially selected for its compatibility with a wide variety of front-end programming languages and its ability to seamlessly interface with front-end languages such as Perl, PHP, HTML, and Java. Oracle software is commercially available and widely used and has been ported to the majority of available operating system platforms. Many universities and institutions have site licenses for Oracle, or the software can be downloaded from the Oracle web site without cost. It has built-in import, export, loading, and backup utilities; indexing and partitioning features for logical storage and timely retrieval of data; and its own internal procedural language, PL/SQL. For larger tables (containing data with a finer level of detail), we used the partitioning feature of Oracle to dynamically separate the data into multiple storage areas, thereby improving database performance.

Our database framework utilizes the minimum number of tables necessary to accommodate experimental data, reduce data redundancy, and allow easy and timely retrieval of data.

9 J. Mobley, unpublished data.
By organizing experimental data based on the level of detail, homogenizing the data across experiments, and enforcing database normalization, data from a large number of experiments can be stored in a few manageable and easily archived data files. The need to maintain the original spectra and search output files is eliminated.

The m/z and intensity values from the precursor and fragmentation spectra are stored in the relational database along with the protein database search results. The currently supported data formats include mzData, mzXML, mgf, and dta for peaks lists and search results from SEQUEST, X!Tandem, and Mascot. Other data formats can be added including the open formats pepXML, proXML, mzIdent, and analysisXML (40). To date, we have loaded results from more than 1,500 LC-MS/MS, 2D LC-MS/MS, and LC-MALDI-MS/MS experiments generated by over a dozen users into our relational database (Fig. 1).
database. On average, each experiment requires less than 20 MB of space, a significant reduction from the 100 MB to 2 GB required to store the multiple individual MS and MS/MS spectra and search output files from a single LC-MS/MS experiment. This storage economy is made possible by the reorganization and compression of hundreds of thousands of individual data files, each with its own file system and header information, into tables in the relational database. We have also implemented the storage of protein sequences and their annotations, including Gene Ontology annotations, thereby allowing protein information to be easily linked with experimental data.

To facilitate the uploading of experimental data, we developed a series of Perl scripts to parse peaks lists and SEQUEST, MASCOT, or X!Tandem search output files and load the results into the relational database. Perl excels at dissecting and searching strings and has built-in features for Oracle compatibility (Oracle Communication Interface). Loading scripts are run from the command line and accept arguments for experimental information, such as protease and user. We equipped the loading scripts with E-mail notification capability so that users are notified when their data are fully loaded and available for viewing.

Additionally for SEQUEST results, the loading scripts allow the option to restrict certain searched data based on cross-correlation value as desired by the user or experimental design (6). For an experiment for which only the best results are of interest, we can designate a higher threshold, whereas for an experiment for which all the search results or low quality spectra are of interest, we can designate no scoring threshold. Distinct thresholds can be designated for each assumed charge state. For MASCOT and X!Tandem search results, filtering is implemented directly within the search algorithm itself by imposing a user-defined scoring threshold.

Although algorithms have been developed to make a charge state determination prior to protein database searching (54), there are invariably cases where no resolution can be achieved. Therefore, it is necessary to provide a method of charge state determination based on search results. We have developed algorithms that perform charge state scoring during the uploading of data to the database. When differentiating between SEQUEST results for \(/H11001^2\) and \(/H11001^3\) precursor ions, cross-correlation score, preliminary score rank, and proteolytic consensus are used in a weighted scoring matrix to determine the most likely charge state of the precursor ion. MASCOT and X!Tandem scoring is based on peptide expectation score, peptide score, and the difference between the theoretical and calculated masses of the precursor ion.

**Graphical Viewers and Analysis Tools**—With normalized, homogenized tandem mass spectrometry data residing in a relational database, we developed a series of applications to retrieve the experimental data from the database and display them in biologically intuitive, graphical formats. For ease of user access, we developed viewer interfaces that rely on a standard web browser and are web-deployed for easy dissemination of results. The database and software are therefore accessible from any computer with Internet capability.
with on-the-fly filtering capability (Fig. 2). Applications include biologically intuitive features, such as organization of protein profiles based on predicted biological categories and relative abundance, and easily understood color coding to present quantitative results from multiple experiments containing thousands of data points (Figs. 3, 4, and 6). Compared with tabular formats, these graphical displays are an easier format for biologists to assimilate and examine their data.

To facilitate software distribution, BIGCAT applications are organized in a modular fashion. Each application module contains all the programs, scripts, functions, style sheets, and documentation specific to that application. There are also directories for common functions, styles, and scripts that contain code and code snippets that are shared across multiple applications. When a new version of an application becomes available, only the code in the directory of that application and the necessary common files need to be upgraded. Because BIGCAT relies primarily on interpreted languages such as PHP and HTML, there is no inconvenience of updating shared libraries and compiling downloaded code. This aspect of the software will make future upgrades easy to disseminate and apply.

Viewing Proteins in a Complex or Proteome Using BIGCAT Viewer—The Viewer application allows users to view results from a single experiment with on-the-fly filtering and useful links to annotation, validation, protein coverage, and spectral analysis tools (Fig. 3). Viewer includes multiple display options for presenting experimental results from simply listing the most probable proteins (Fig. 3A) to a full view of all the peptides identified in the experiment. Identified proteins can be organized based on predicted biological categories along with estimated relative abundances. Using annotations from Gene Ontology, users can group the proteins based on function, process, or cellular compartment and calculate the percentage of proteins in each category (Fig. 3B). This helps users to quickly recognize biologically significant results by identifying expected and unexpected proteins in an experiment. In addition, proteins can be graphically displayed as a heat map based on the estimated relative abundance (Fig. 3C) and the predicted biological properties (Fig. 3D). Extensive sorting functions allow the user to order the list of proteins.

BIGCAT Viewer incorporates the ALIGN function of Myers and Miller (55) that provides protein alignment and percent identity between proteins with shared peptides. For highly identical proteins (>95% identical), we have found this to be a useful tool to rationalize the inability to identify a unique protein from groups of proteins that share peptide hits. Peptide alignments with proteins and protein coverage are available on the linked Protein Coverage and Protein Validation pages. The annotation page provides the user with the annotations for the protein as well as links to outside annotation services including Ensembl, RefSeq, Gene Ontology, Swiss-Prot, Saccharomyces Genome Database, Munich Information Center for Protein Sequences, and the BIOBASE BioKnowledge Library.

Our integrated Filter application allows users to create reusable filters that designate criteria for the inclusion or exclusion of data. These filters are saved in the database and can be viewed, updated, or deleted by the filter owner via the Filter application. Among other options, users can include or exclude data based on particular criteria, exclude erroneous and contaminant proteins, filter peptide hits based on protease specificity, and display peptides with posttranslational modifications. To minimize redundancy, proteins identified by the same set of peptides can be grouped together. A default filter is available as well so that users need not make filtering decisions prior to viewing experimental results. User-defined and default filters are available across applications.

Comparing Proteins in Two Complexes or Proteomes Using BIGCAT Comparer—Comparative analysis is essential for developing biological models. In its simplest form, a comparison of control versus experimental samples distinguishes background noise from signal and identifies differences that indicate biological relevance. The availability of homogenized tandem mass spectrometry results in a relational database, combined with our user-defined data filtering capability, introduced the potential for quickly and easily comparing experimental results.

Our Comparer application compares the estimated relative abundances of identified proteins between two groups of experiments. Relative protein abundance is expressed as either the number of non-redundant peptides identifying a protein or as the PAF (Fig. 4) (35). The average abundance in each experimental group and the ratio between the two groups are also presented. Protein descriptions and links to further annotation are provided within the display. Users can customize the comparison display using sorting and additional filter options. Color annotation can be used to highlight protein enrichments (Fig. 4A). Comparer can also be used in concert with our statistical analysis and clustering applications (see below), or the results can be exported for use with external applications (Fig. 9).

Comparer offers both tabular and graphical approaches for comparing protein profiles (Fig. 4). A graphing function plots the average abundance of each protein in the experimental group against the control group, allowing for quick identification of protein differences between the groups (Fig. 4B). Using this feature, proteins that are significantly enriched in one group or present at relatively high abundance can be easily recognized. As an example of its application, we have used Comparer in studies of proteins that interact with the yeast transcription machinery (35, 52). The yeast basal transcription complex TFID is composed of the TATA-binding protein and 14 distinct TATA-binding protein-associated factors (Tafs) (56). We have coupled immunopurification (IP) of each Taf component with 2D LC-MS/MS to identify the interacting proteins (52). Fig. 4B presents a profile generated by Com-
A parer of proteins enriched in the IP of the TFIID component Taf5 as compared with those from control experiments using preimmune antibodies (52). Labeled proteins in the upper left quadrant indicate those proteins abundantly present in the Taf5 purification. We unexpectedly observed the small, 11-kDa yeast protein Sgf11, significantly enriched in the Taf5 purifications, suggesting that Sgf11 is also a component of TFIID.

Comparing and Contrasting Results from Multiple Search Algorithms Using BIGCAT Multisearch Comparer—By applying multiple search algorithms to a single set of experimental data, we are able to corroborate our peptide and protein identifications. Identifications that reoccur across multiple searches, particularly those with significant scores, are deemed more reliable (20, 57). To allow users to view and compare peptide and protein identifications and scores generated by different search algorithms, we developed the Multisearch Comparer application (Fig. 5). Filtering and sorting capabilities are included as well as links to protein annotations. Fig. 5 displays iTRAQ labeled peptides analyzed via LC-MALDI-MS/MS and identified using both the SEQUEST and MASCOT algorithms. Peptides and protein scores from both search algorithms are presented, allowing users to recognize the most conclusive peptide hits and protein identifications.

Fig. 3. Viewer displays proteins in proteomes and complexes. Multiple formats are provided for presenting the results of a tandem mass spectrometry experiment. Here we have affinity-purified the S. cerevisiae Cdc33 complex and analyzed the components by 2D LC-ESI-MS/MS. The Cdc33 protein is indicated with a blue arrow in each figure. A, simple summary report of experimental results with the identified proteins ranked by estimated relative abundance. Cdc33 is the most abundant protein in the sample. B, a biologically intuitive report of the protein profile using the predicted biological categories and relative protein abundances to organize the identified proteins. Cdc33, indicated with a blue arrow, is categorized as a translation initiation factor. C, heat map display of relative protein abundances expressed as PAFs. Color intensity corresponds to protein abundance with brightest red indicating the most abundant proteins and yellow indicating the most rare. Cdc33, indicated with a blue arrow, is the most abundant protein in the profile. D, heat map display of relative protein abundances expressed as PAFs and organized by predicted biological functions. Cdc33, indicated with a blue arrow, is the most abundant translation initiation factor (IF) identified in the experiment.
Analyzing MS/MS Proteomic Results

A

**BIGCAT Comparer v2.0b**

Select a data view mode:

- Protein Abundance Factors PAF
- View comparison in MS Excel

First Sort By:
- Ratio
- Ascending
- Descending

Second Sort By:
- Group 2 Mean
- Ascending
- Descending

Graph Me!

Filter Data:
- None
- > 0

<table>
<thead>
<tr>
<th>#</th>
<th>Accession</th>
<th>Gene Name</th>
<th>Description</th>
<th>MW</th>
<th>Group 1 Mean</th>
<th>Group 2 Mean</th>
<th>Ratio</th>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>YML015C</td>
<td>TAF11</td>
<td>TRF1 subunit (40 kDa)</td>
<td>40224</td>
<td>0</td>
<td>3.9326</td>
<td>0:0</td>
<td>2.7078</td>
<td>5.1694</td>
</tr>
<tr>
<td>2</td>
<td>YPL011C</td>
<td>TAF3</td>
<td>TRF1 subunit (47 kDa)</td>
<td>40226</td>
<td>0</td>
<td>2.8784</td>
<td>0:0</td>
<td>2.4916</td>
<td>3.4743</td>
</tr>
<tr>
<td>3</td>
<td>YOL086W</td>
<td>SOFT3</td>
<td>Seo2 associated Factor 73 kDa</td>
<td>72878</td>
<td>0</td>
<td>2.2641</td>
<td>0:0</td>
<td>2.0592</td>
<td>2.4699</td>
</tr>
<tr>
<td>4</td>
<td>YPL247W</td>
<td>SOFT11</td>
<td>11 kDa subunit of the SAGA hist</td>
<td>11285</td>
<td>0</td>
<td>2.2154</td>
<td>0:0</td>
<td>1.7723</td>
<td>2.6584</td>
</tr>
<tr>
<td>5</td>
<td>YDR157W</td>
<td>TAF10</td>
<td>Subunit (145 kDa) of TRF1 and</td>
<td>23019</td>
<td>0</td>
<td>2.1721</td>
<td>0:0</td>
<td>2.1721</td>
<td>2.1721</td>
</tr>
<tr>
<td>6</td>
<td>YCL010C</td>
<td>SOFT29</td>
<td>Seo2 associated Factor 29 kDa</td>
<td>29891</td>
<td>0</td>
<td>2.0423</td>
<td>0:0</td>
<td>1.6211</td>
<td>3.0632</td>
</tr>
<tr>
<td>7</td>
<td>YDR449W</td>
<td>ACAD1</td>
<td>Transcription coactivator</td>
<td>50569</td>
<td>0</td>
<td>1.9775</td>
<td>0:0</td>
<td>1.3842</td>
<td>2.5707</td>
</tr>
<tr>
<td>8</td>
<td>YPL254Y</td>
<td>HIP1</td>
<td>Adaptor protein required for</td>
<td>54666</td>
<td>0</td>
<td>1.9279</td>
<td>0:0</td>
<td>1.2852</td>
<td>2.5704</td>
</tr>
<tr>
<td>9</td>
<td>YML114C</td>
<td>TAF8</td>
<td>TRF1 subunit (65 kDa)</td>
<td>57986</td>
<td>0</td>
<td>1.897</td>
<td>0:0</td>
<td>2.2419</td>
<td>1.5521</td>
</tr>
<tr>
<td>10</td>
<td>YMR227C</td>
<td>TAF7</td>
<td>TRF1 subunit (67 kDa)</td>
<td>67555</td>
<td>0</td>
<td>1.8504</td>
<td>0:0</td>
<td>1.6283</td>
<td>2.0724</td>
</tr>
<tr>
<td>11</td>
<td>YPL128W</td>
<td>TAF14</td>
<td>Subunit (30 kDa) of TRF1, TRF1</td>
<td>27440</td>
<td>0</td>
<td>1.8222</td>
<td>0:0</td>
<td>2.551</td>
<td>1.9033</td>
</tr>
<tr>
<td>12</td>
<td>YDR176W</td>
<td>NRG1</td>
<td>Transcriptional regulator invo</td>
<td>79282</td>
<td>0</td>
<td>1.7029</td>
<td>0:0</td>
<td>1.3875</td>
<td>2.0181</td>
</tr>
<tr>
<td>13</td>
<td>YER148W</td>
<td>SPT15</td>
<td>TATA-binding protein, general</td>
<td>27003</td>
<td>0</td>
<td>1.6665</td>
<td>0:0</td>
<td>2.222</td>
<td>1.11</td>
</tr>
<tr>
<td>14</td>
<td>YGR252W</td>
<td>OCNS5</td>
<td>Histone acetyltransferase, ace</td>
<td>51069</td>
<td>0</td>
<td>1.6644</td>
<td>0:0</td>
<td>1.7623</td>
<td>1.5665</td>
</tr>
<tr>
<td>15</td>
<td>YML098W</td>
<td>TAF13</td>
<td>TRF1 subunit (19 kDa)</td>
<td>19132</td>
<td>0</td>
<td>1.5661</td>
<td>0:0</td>
<td>1.4542</td>
<td>2.0677</td>
</tr>
<tr>
<td>16</td>
<td>YMR223W</td>
<td>LRP3</td>
<td>Ubiquitin-specific protease Th</td>
<td>53623</td>
<td>0</td>
<td>1.4919</td>
<td>0:0</td>
<td>1.3054</td>
<td>1.6704</td>
</tr>
</tbody>
</table>

B

**Group 2 PAF**

- TAF11
- RPS18B/RPS18A
- SS82
- SS81
- SS84
- SS85
- RPS15
- RPS1A
- RPS2
- RPS3
- RPS5
- RPS51A
- RPS7
- RPS14
- RPS16

**Group 1 PAF**

- TAF4
- TAF12
- TAF5
- TAF3
- SFT7
- SF29
- HFI
- TAF7
- SFT1
when proteins are less abundant than the complex. Additionally substoichiometric interactions can indicate novel associations or interactions with the similar abundances, uncharacterized proteins of similar abundances in scoring between search algorithms (Fig. 5), possibly reflecting the inherent bias between the disparate scoring methods. Similar results have been observed in previous experiments (20, 57).

Multisearch Comparer also allows users to identify discrepancies in scoring between search algorithms (Fig. 5), possibly reflecting the inherent bias between the disparate scoring methods. Similar results have been observed in previous studies (20, 57).

Clustering Patterns of Enrichment or Abundance Using BIGCAT Clusterer—The ability to identify trends in relative protein abundances within a group of mass spectrometry experiments is vital in the identification of protein interactions or the differentiation between cell states. For example, because individual proteins within a complex will often have similar abundances, uncharacterized proteins of similar abundance may indicate novel associations or interactions with the complex. Additionally substoichiometric interactions can be identified when proteins are less abundant than the components of the main complex.

The BIGCAT Clusterer application provides unsupervised clustering of proteins according to similarities in patterns of estimated enrichment or relative abundance. Clusterer also provides supervised clustering based on the predicted biological categories of the proteins (Fig. 6). This enables investigators to quickly evaluate protein profile changes and patterns of expression. The available clustering algorithms, Enrichment and McAfee Sum-Rank, use either the number of non-redundant peptides identifying a protein in a sample or the PAF to group and order proteins across a series of selected experiments. Additionally proteins can be organized by abundance within predicted biological categories (supervised clustering). The McAfee Sum-Rank method provides an additional dimension of clustering based on overall protein abundance within each experiment.

The results of Clusterer are displayed in a cluster diagram, or heat map, using a range of colors to show patterns of enrichment (Fig. 6). Fig. 6 demonstrates the application of unsupervised (Fig. 6A) and supervised (Fig. 6B) enrichment clustering to polysome fractions isolated under increasing salt concentrations to identify proteins associated with ribosomes. It is easy to recognize from the generated heat maps that 60 S subunits are significantly enriched in the 60 S fractions but not in the 40 or 80 S fractions. Identified proteins not currently characterized as translation-related are displayed in the lower portion of the heat map. Many of these proteins are present at levels similar to known translation proteins, suggesting possible interactions with the ribosomal subunits.

An option for data normalization against user-selected background experiments is provided. Normalization subtracts background levels of protein abundance from experimental levels, thereby allowing protein enrichments and changes in protein profiles to be more easily recognized. The integrated export feature of Clusterer allows enrichment clustering results generated by Clusterer to be exported for use with one of the publicly available hierarchical clustering algorithms and viewers (58, 59). We have used this method to successfully identify a novel component of the yeast SAGA histone acetyltransferase complex (35). By applying unsupervised hierarchical clustering (58) to the data shown in Fig. 4 along with the IP and 2D LC-MS/MS analysis of the other yeast Taf proteins, we refined our earlier interpretation of Sgf11 to show that the
protein is a novel component of the SAGA histone acetyltransferase complex (35). The Taf5 component along with four other Tafs is shared between the TFIID and SAGA complexes (60). Using BIGCAT along with different analysis methods, we were able to develop a model of the biological function of Sgf11 that we could experimentally test (35).

Assembling a Map of Protein Structure and Peptide Coverage—The sequence of a peptide is inferred from tandem mass spectrometry data based on the known masses of the amino acids. Because only a subset of the amino acids can be modified and the masses of these modifications are known, posttranslationally modified amino acids can theoretically be identified. However, in practice, the identification of modified amino acids is challenging. Searches with even a small set of modifications present a substantial combinatorial challenge and require impractically long amounts of time due to the large number of possible variants against which spectra are searched (61). Furthermore, modified amino acids are typically substoichiometric such that only a small percentage of protein molecules are modified. Finally, obtaining 100% coverage of a protein, such that all modified residues can be identified, is practically impossible using a single proteolytic enzyme.

To address these problems, we developed the Assembler application, which permits data from multiple experiments to be merged (Fig. 7). Various individual proteolytic digestions coupled with LC-MS/MS analysis of the resulting peptides can be used to identify modified amino acids (62). The independent datasets are merged in Assembler while referencing the individual experimental conditions. A graphical coverage map displays independent overlapping peptides (Fig. 7). Modified amino acids are indicated at their position in the protein sequence by color-coded symbols. The same modifications identified on multiple, independent peptides are more conclusive.

Statistical Analysis of Changes in Protein Profiles Using BIGCAT Analyzer—Our statistical analysis application, Analyzer, currently includes two methods to statistically measure the significance of variation between two groups of experiments. This type of analysis is essential to validate protein enrichments or expression differences between two groups of experiments, e.g. affinity purification versus control or treated versus untreated cells. As tandem mass spectrometry results are non-parametric in nature, the step-down multivariate permutation t test is the most appropriate approach, controlling for family-wise error within the data (63, 64). The Student’s t test, a traditional, universally recognized statistical method allowing a broader approach to analysis, is also available. We
plan to include additional statistical test options appropriate to tandem mass spectrometry results (i.e. non-normal, non-parametric data) in future versions of Analyzer.

Complex Queries to Identify Motifs and Modifications Using BIGCAT—One advantage of the relational database is the ability to easily execute complex queries to uncover patterns and trends in the data. The BIGCAT front-end software conceals the complexity of these queries from the user. We have developed several complex query applications, which provide an interface for user input to dynamically generate and return results from a custom database query.

One of these complex query applications is Motif Matcher. Motif Matcher finds and displays identified peptides that contain specific sequence motifs. Motifs can be easily added based on experimental direction. For example, the major histocompatibility complex displays peptide antigens on the cell surface (65). The human lymphocyte antigen (HLA) class affects the selection of antigens for display; the peptides share specific sequence motifs depending on the HLA class (66, 67). We are currently analyzing vaccinia-infected human cells to identify peptides displayed on the cell surface.11 Using the known relevant sequence motifs for the different HLA alleles, we are using Motif Matcher to identify the displayed peptide antigens in the context of the six major human HLA class supertypes. From experiments using LC-MS/MS to identify vaccinia and human major histocompatibility complex peptides recovered from infected human cells expressing different HLA types, Motif Matcher has enabled us to rapidly screen results for peptides that match the HLA class of the cells.

A second complex query application, Mod Miner, searches the relational database for MS/MS spectra exhibiting neutral losses, which are known signatures of posttranslational modifications (Fig. 8). During MS/MS fragmentation, labile peptide modifications will typically be lost from the peptide before the peptide itself fragments (68). This phenomenon results in an observable neutral loss within the generated spectrum. Labile modifications include phosphorylation on serine or threonine residues, ubiquitination, and sulfation (68).

In Mod Miner, the user specifies the expected neutral loss, the margin of error, and the number of most intense peaks in the MS/MS spectrum to compare against the precursor ion m/z value. Mod Miner generates a list of candidate spectra exhibiting the specified neutral loss (Fig. 8). This list can be exported for reanalysis against the original protein database. By searching only this abbreviated list of candidate spectra for the posttranslational modification of interest, the time required to complete the search is significantly reduced. We have found Mod Miner to be especially powerful for rapidly screening large numbers of MS/MS spectra for neutral loss signatures indicative of phosphorylated peptides (i.e. 98, 49, and 32.67 m/z). We anticipate that many other complex query applications can be developed using the relational database model, including applications for spectral quality assessment and partial de novo sequencing for the generation of sequence tags (69, 70).

Adapting Genomic Software for Use with BIGCAT—Proteomic data that characterize protein expression share many attributes with data from DNA microarray experiments that characterize mRNA expression. DNA microarray studies aim to identify patterns of gene expression across many experiments that survey a wide array of cellular responses, phenotypes, and conditions. Some studies aim to identify groups of genes that are regulated similarly in response to different conditions, such as genes affected by a treatment, or marker genes that discriminate diseased from healthy subjects. Other experiments aim to uncover different conditions or cell states that result in similar patterns of gene expression. DNA microarray data are often used to guide more precise studies of gene expression or to provide the investigator with good candidate genes for further research. Most tandem mass spectrometry studies have similar goals but identify and measure proteins instead of RNAs.

Statisticians have developed many procedures for microarray data that have proved more useful than conventional tests (71, 72). Microarray analysis software, such as GeneSpring (Agilent Technologies), GenePattern (www.broad.mit.edu/cancer/software/genepattern/), Bioconductor (www.bioconductor.org), and the Cluster and TreeView applications (rana.lbl.gov/index.htm) (58), perform appropriate statistical analyses and incorporate multiple graphical displays for presenting gene abundance and patterns of expression. The analyses provided by these software packages are extremely powerful when applied to MS/MS experimental results (35).

We are able to extend the power of these DNA microarray applications for the interpretation and analysis of tandem mass spectrometry results by using BIGCAT to provide appropriately formatted data (Fig. 9). Protein and abundance values from BIGCAT, equivalent to the gene and signal intensity values derived from DNA microarray data, can be imported into these applications with little additional data manipulation. This extension of the analysis capabilities of BIGCAT for integration with other sophisticated statistical analysis and data visualization tools provides a powerful combination that will help users analyze and interpret data from large numbers of MS/MS experiments.

Supporting Applications—BIGCAT also includes many supporting applications that are accessible from the main applications such as Viewer, Comparer, Assembler, or Mod Miner.

Our Spectra Viewer displays a graphical representation of MS/MS spectra for manual confirmation of the peptide sequence. The predicted amino acid sequence is superimposed on a graph of the spectrum with the predicted b and y ions labeled. The user is able to select any of these alternative peptide matches and redraw the graph with the selected amino acid sequence and superimposed b and y ions. Graphs

11 S. Conant, J. Jennings, J. McAfee, D. Duncan, A. Link, and S. Joyce, unpublished data.
of the +1 and +2 charge state b and y ions, with the amino acid sequence color-coded, are available as well for easy verification. The Spectra Viewer additionally offers zoom capability.

Our Chromatography Viewer uses the data in the relational database to construct chromatographic and spectra views for an experiment. Reconstructed ion chromatographs of the precursor ion are available as well as base peak chromatographs with the precursor ion intensities graphed against their retention times. For identified peptides and proteins, the retention time of each precursor ion is marked on the base peak chromatograph. The Chromatography Viewer can be used for visualizing the consistency of retention times for identical peptides in different LC-ESI- or LC-MALDI-MS/MS experiments.

We have used numerical analysis methods for peak integration using data in the relational database. In our Chromatography Viewer, we graph the precursor ion intensities against their retention times and apply a weighted moving average algorithm to reduce the signal noise (73). The Graham convex hull algorithm is used to establish the base line and subtract it from the signal (74). Finally we determine the upper and lower bounds of integration by numerical analysis wherein integration is computed using the trapezoidal rule (75). In the future, we plan to apply this method to evaluate other label-free and isotopic labeling approaches for protein quantitation (17, 23–29, 33, 34).
The Protein Annotation application provides the user with Gene Ontology and other selected annotations contained in the original protein database header as well as links to several outside annotation services, including Ensembl, RefSeq, Saccharomyces Genome Database, International Protein Index, SWISS-PROT, Munich Information Center for Protein Sequences, and the BIOBASE BioKnowledge Library. Additionally we have included custom protein annotation functionality through which user annotations derived from personal research can be stored in the relational database. User identification and time stamps are included to document entries.

The Administration module allows users to manage their experiments in the database and to group experiments logically into projects. Users are able to edit experiment information such as titles and descriptions and grant other users data viewing privileges. A superuser has, in addition to standard user administration privileges, the ability to add or remove users, reset user passwords, grant privileges to users on experiments and projects, and delete experiments from the database.

**DISCUSSION**

We have developed a relational database, based on the Oracle Relational Database Management System, that effi-
ciently stores and manages large amounts of tandem mass spectrometry results. We have also developed a compatible suite of analysis software for analyzing, interpreting, comparing, and displaying stored tandem mass spectrometry results. This system, called BIGCAT, provides rigorous and sophisticated analysis of tandem mass spectrometry results as well as a working solution to many proteomic data storage and dissemination issues.

BIGCAT provides an open-source, browser-driven suite of multifunctional graphical software applications. The web-deployed design of BIGCAT promotes on-line collaboration and dissemination of tandem mass spectrometry results in a secure environment. Multiple biologically intuitive data visualization and manipulation tools, integrated analytical and comparative approaches, and protein identification and peptide quantitation algorithms facilitate graphical comparison of results from many experiments for evaluation of changes in protein profiles and patterns of expression.

Several alternative methods have been proposed to address the issues involved with storage, dissemination, and analysis of proteomic and mass spectrometry data. In an attempt to standardize the reporting format of proteomic data, the Proteomics Experiment Data Repository and MIAPE have been proposed (36, 76). The Proteomics Experiment Data Repository uses a universal modeling language that contains details and data from proteomic experiments (36). As one of the Proteomics Standards Initiatives, MIAPE is an initiative to describe the minimum information about a proteomic experiment that is needed to enable interpretation of the results and to reproduce the experiment (76).

For open dissemination of spectra data, two groups have independently proposed a common file format based on XML (mzXML and mzData) (39). In these approaches, native mass spectrometry data files can be converted to an XML format using software applications provided by instrument manufacturers. This common format allows data from different instrument manufacturers to be analyzed by a common set of data analysis applications. To minimize the file sizes and improve access to data in the XML files, the m/z and intensity data are stored in an indexed, base-64 form (39). Although the XML format provides access to the native data, it does not fully represent the experimental information contained in the native file formats. In addition, the XML file size requires a large amount of storage space.

For open dissemination of database search results and to homogenize the disparate data formats from different search algorithms, the application of XML has also been proposed (40). The Trans-Proteome Pipeline software converts native spectra data; the output results from SEQUEST, MASCOT, and Comet search engines (pepXML); and protein information (protXML) into open XML file formats (40). The European Proteomics Standards Initiative has proposed the mzIdent and more recently the analysisXML standard as a standard format for peaks list, search parameters, and the results of different database search algorithms.

Although these XML models provide a solution to the standardization of tandem mass spectrometry data, they have many drawbacks (43). The XML format is not optimized for computation or comparison. XML in its current usage is not entirely practical for bioinformatics, nor is it a substitute for good data representation. Data storage or computation methods based directly on XML may also encounter performance and scalability problems (43).

Other groups have examined the application of relational databases to storing, annotating, and organizing LC-MS/MS data and results (20). The Rapid, Automated, Data Archiving and Retrieval System (RADARS) uses a relational database to store processed mass spectrometry results from different mass spectrometers and search algorithms (44). RADARS also provides customized web-based browsers to access the data. DBParser focuses on the MASCOT search algorithm and parses data and search results from LC-MS/MS experiments into an MySQL relational database (19). In addition, protein annotations, such as those in the Gene Ontology database, are available. More extensive LC-MS/MS data management systems, including the Systems Biology Experiment Analysis Management System (SBEAMS), the Proteome Research Information Management Environment (PRIME), and the Computational Proteomics Analysis System (CPAS), use relational databases for data storage and management and browser-driven applications to process and view results (45). Currently most of these approaches are primarily focused on organizing and managing experimental proteomic data.

As part of the Cancer Bioinformatics Grid Initiative (caBIG), ProtLIMS manages metadata associated with proteomic experiments and promotes collaborative research. As a laboratory information management system (LIMS), its focus is currently on data tracking, workflow, and protocol and laboratory information management rather than analysis, comparison, and modeling of data (46).

Several public resources, including the Proteomics Identifications Database (PRIDE), PeptideAtlas, and the Global Proteome Machine organization (GPM), promote sharing and integration of mass spectrometry results (8, 41, 42). They act as repositories for published LC-MS/MS experiments and can be viewed by the scientific community. The open access to mass spectrometry results is a significant step forward.

The BIGCAT system provides a comprehensive solution to data storage, rigorous data analysis, and dissemination of tandem mass spectrometry results. Comparison, statistical analysis, data visualization, and display applications are seamlessly integrated with the relational database, minimizing the inherent program bias encountered when using multiple, disparate tools. Complex query applications allow sophisticated data mining through interactive front-end interfaces and dynamically generated queries. The applications are designed to be intuitive for biologists involved in many facets of data...
analyzing whether they are identifying protein interactions and patterns of association or generating statistical models of protein function, protein structure, or cell state. All database data are available through the BIGCAT software, and users have the ability to share tandem mass spectrometry data and results on line with collaborators and other BIGCAT users. Finally the BIGCAT front-end software is open-source, and the back-end database model can be easily adapted to most SQL-based relational database systems.

In conclusion, the BIGCAT relational database and software system provides a solution to many data storage and dissemination issues associated with proteomic data. The relational database efficiently stores and manages large amounts of tandem mass spectrometry results. The integrated suite of analysis software allows investigators to interpret, compare, mine, and display results in biologically intuitive formats. BIGCAT facilitates rigorous and sophisticated analysis of tandem mass spectrometry results in a web-deployed and user-friendly environment for either local analysis or for promoting on-line collaborations.

Acknowledgments—We thank Tracey Fleischer for numerous discussions and feedback during this project. We thank Jennifer Jennings, Connie Weaver, David Powell, and Vince Gerbasi for evaluation and comments during the development of the applications. We thank James Mobley for access to unpublished LC-MALDI-TOF-TOF data and discussions on data analysis. We thank Greg Bowersock for discussions on native mass spectrometry data conversions. We thank Jennifer Blackford and the Kennedy Center for Human Development for help and advice with the development of statistical analyses. We thank Ryan Hatcher and Richard Shivi for discussions on mathematical methods to quantify precursor ions intensities. We thank the Vanderbilt Microarray Shared Resource, especially Braden Boone, Lauren Sims, Philip Dexheimer, and Shawn Levy, for numerous discussions on statistical data mining. The Vanderbilt Microarray Shared Resource is supported by the Vanderbilt Ingram Cancer Center (National Institutes of Health (NIH) Grant P30 CA68485), the Vanderbilt Diabetes Research and Training Center (NIH Grant P60 DK20593), and the Vanderbilt Digestive Disease Center (NIH Grant P30 DK58404).

* This work was supported in part by NIH Grants GM67477 and ES11993 and NICHD, NIH Grant P30 HD15052 to the Vanderbilt Kennedy Center for Research on Human Development. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

† Supported by NIH Grants ES11993 and GM64779.
‡ Supported in part with federal funds from the NIAID, NIH, Department of Health and Human Services, under Contract Number HHSN266200400079C/N01-Al-40079.
¶ Supported by NIH Grant ES11993. || Supported by NIH Grants GM64779, HL68744, ES11993, and CA098131. To whom correspondence should be addressed: Dept. of Microbiology and Immunology, Vanderbilt University School of Medicine, 1161 21st Ave. S., 9283 MBRRill, Nashville, TN 37232-2363. Tel.: 615-343-6823; Fax: 615-343-7392; E-mail: andrew.link@vanderbilt.edu.

REFERENCES

acetylation and transcriptional stimulation. Cell 94, 45–53


