research articles Journal of proteome research

Open Source System for Analyzing, Validating, and Storing Protein Identification Data

Robertson Craig,[†] John P. Cortens,[‡] and Ronald C. Beavis*[,†,‡]

Beavis Informatics Ltd., and Manitoba Center for Proteomics, Winnipeg, Manitoba, Canada

Received July 13, 2004

This paper describes an open-source system for analyzing, storing, and validating proteomics information derived from tandem mass spectrometry. It is based on a combination of data analysis servers, a user interface, and a relational database. The database was designed to store the minimum amount of information necessary to search and retrieve data obtained from the publicly available data analysis servers. Collectively, this system was referred to as the Global Proteome Machine (GPM). The components of the system have been made available as open source development projects. A publicly available system has been established, comprised of a group of data analysis servers and one main database server.

Keywords: proteomics database • GPM • X Tandem • protein identification • XIAPE

Introduction

There has been considerable discussion in the recent literature as to how data from proteomics experiments can be standardized for broad dissemination.[1−5] The most comprehensive attempt at such standardization has been sponsored by the Human Proteome Organization (HUPO), through the Protein Standardization Initiative (PSI) working group.[1] This group has proposed a database schema, the Proteomics Experimental Data Repository (PEDRO), and an eXtensible Markup Language (XML) specification, the Proteomics Experiment Markup Language (PEML), for use in Laboratory Information Management Systems (LIMS). This proposal, which has become known as the "Minimum Information About a Proteomics Experiment" (MIAPE) proposal, uses a very detailed database and XML schema to attempt to describe many possible variations of proteomics laboratory experiments. The overall specification of MIAPE contains many elements commonly found in LIMS implementations, such as the capability of tracking the results of an experiment through all of the experimental protocols that have been performed and the capability of storing the raw data and analyzed metadata within the relational structure. Practical systems that utilize either MIAPE or a related LIMS-type database strategy have been reported, although none is currently available to the general proteomics community.[6]

There has also been discussion in the literature regarding the best methods for validating results of statistical analyses that constitute the final mapping of a list of tandem mass spectra to a list of protein sequences in a proteomics experiment.[7−12] The goal of this research has been to obtain a valid estimate of how likely a particular peptide-to-spectrum match is to be caused by a stochastic coincidence between the spectrum and one of a large number of peptide sequences generated from a list of proteins. Originally, it was assumed that mass spectrum signal intensities could not be accurately predicted from a peptide sequence; however, considerable progress has been made using large numbers of spectra to determine bond cleavage rules.[13−17] An improved understanding of the combination of underlying statistical distributions and gas-phase fragmentation reactions has led to significant advances in this research. Several protein identification search engines[18,19] and at least one validation tool[20] have been developed to take advantage of the prediction of peptide bond fragment intensities.

One alternative that has been proposed to the prediction of signal intensities is the creation of a library containing experimental mass spectra generated from synthetic peptides that represent all of the possible sequences in a particular proteome.[21] The technical difficulties and cost of generating such a collection of synthetic peptides will probably make this ideal case impractical for the near future.

Another alternative strategy that could be deployed immediately would be to collect a large number of peptide mass spectra obtained from proteomics experiments and store them in a repository. When a new mass spectrum-to-peptide sequence correlation is postulated, the repository could be queried to return a list of the best previously observed mass spectra that have been associated with that sequence. In a comparison of the existing exemplar peptide ion fragmentation patterns with the newly observed pattern, the repository would provide some of the same functions as a library of spectra obtained from synthetic peptides, with the proviso that the sequence annotation would be based on spectrum-to-proteome matching, rather than on known peptide analytes. This sort of repository structure would allow the system to remain relevant as new instrumentation becomes available. It also has the potential to provide additional confidence to particular assignments by having many redundant measurements of the same peptide sequence's fragmentation pattern under a variety of

[†] Beavis Informatics Ltd..

[‡] Manitoba Center for Proteomics.

10.1021/pr049882h CCC: $27.50

Published on Web 10/22/2004

different experimental conditions, e.g., different parent ion charge states, fragment ion signal-to-noise ratios or mass spectrometer configurations.

The proposed repository could also be used to compare the patterns of peptides that have been observed for a particular protein sequence (often referred to as the observed “coverage map” of a sequence). This pattern of observed peptide ions is naturally a property of the protein sequence and the physical properties of the peptides. It is also a function of the analytical sample workup protocols, the mass spectrometer’s ion source and the fragmentation conditions for the peptides. This combination of characteristics makes it difficult to predict a priori which of the theoretical peptides for a protein sequence will actually be observed. However, by comparing an observed peptide coverage map with the best previously observed coverage maps, it should be possible to determine whether the observed pattern is consistent with previous results. This type of comparison becomes particularly important when only one or two peptides from a particular protein are observed, where knowing that these few peptides consistently produce the strongest signals would add considerable confidence to their assignment.

We have designed a database schema to serve as both an extension and a simplification of the MIAPE idea, for the purpose of validating observed protein coverage and peptide fragmentation data. The design goal of the schema was to create a database which could be used both on its own to provide answers to queries as well as to serve as an index to experimental information stored in XML documents. This grafting of a specialized relational schema with the object structure of the XML document has simplified the design process considerably and has allowed us to create a working version of a publicly available data repository for the bioinformatics analysis of proteomics data. This system was called the Global Proteome Machine (GPM).

Materials and Methods

1. Data Analysis Server Design. The initial phase of developing this system was to produce a user interface to the open source search engine, X! TANDEM.[19] This user interface supported the use of the Apache HTTP server for access via the Hyper Text Transfer Protocol (HTTP).[22] It consisted of a set of programs written in the Practical Extraction and Report Language (Perl[23]) that generated the user interface in the Hypertext Transfer Markup Language (HTML), using Cascading Style Sheets (CSS), XML Stylesheet Language Transformations (XSLT), and the Scalable Vector Graphics language (SVG[24]). These various technologies were used to provide a standard interface to the search engine and to generate tabular and graphical representations of the data that is stored in X! TANDEM’s XML output files. All the files necessary to create a GPM mass spectrum data analysis site have been available as of January 2004, under the Artistic License as open source software.[25]

2. Database Design. The database was designed to use the smallest subset of the data obtained from mass spectra-to-peptide sequence matching that would allow Structured Query Language (SQL) queries. These queries allowed ready access to any information stored in the XML files generated by the data analysis servers. In this system, the XML files serve as the primary data storage objects, allowing the design of a relational dataset that was relatively easy to build, maintain and query. The information retained its original object character while obtaining the advantages of an SQL-compatible relational database. In keeping with the naming of MIAPE, and for the sake of clarity, this hybrid system was called XIAPE (XML Information About a Proteomics Experiment).

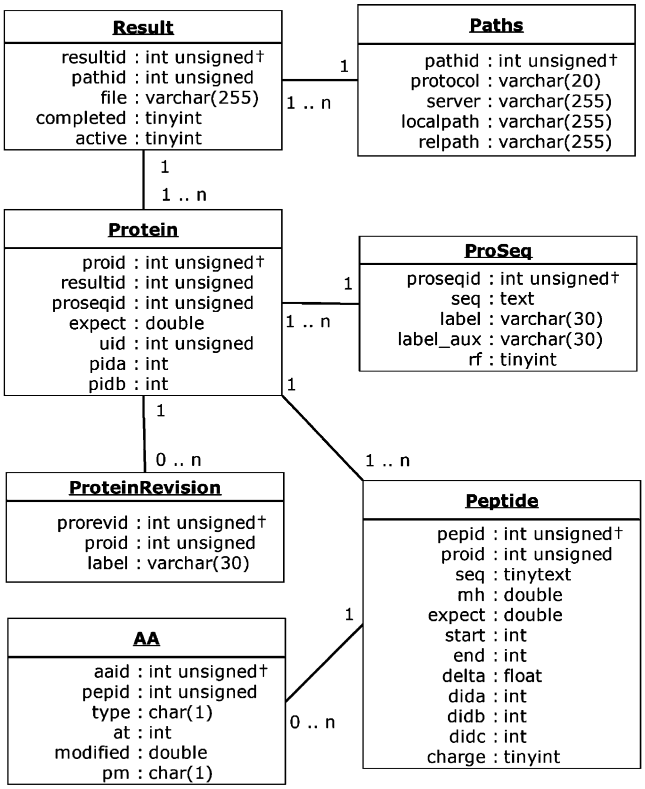

Figure 1. Detailed UML class diagram of the GPM database schema. Items with daggers were primary keys.

The detailed object model for the database, represented using the Unified Modeling Language (UML), appears in Figure 1. The main relation *Result* contains the information that links the XML metadata and the database. Metadata, or “data about data”, refers to the output of analytical software containing information about the assignment of sequences to experimental data. The general assumption made about storing large numbers of XML data files was that they would be stored in a limited number of repository folders that could be addressed using one of a variety of communication protocols (e.g., HTTP, FTP, or UNC). The individual XML object was therefore referenced by a file name, protocol and a repository folder. Information about the repository folder is stored in the *Paths* relation and referenced by the *pathid* attribute in *Result*.

A *Result* relation contained one or more *Protein* relations, each storing the protein sequence (*seq*) and sequence database accession identifier (*label*), as well as a measure of the statistical validity of the protein’s identification (*expect*) and several identification numbers (*uid*, *pida* and *pidb*) that allow unambiguous reference to the sequence with the *Result*’s XML file. Each *Protein* contained one or more *Peptide* relations that retain information about the peptides, as well as the identification numbers that reference these objects in the XML file (*dida*, *didb*, and *didc*). In turn, *Peptide* relations may contain *AA* relations. Each *AA* describes how a specific amino acid residue in a *Peptide* was modified in order to obtain the best fit to a particular mass spectrum (*modified*) and indicates whether the residue was determined to be a point mutation (*pm*). All residue numbering was relative to the N-terminal of the protein sequence. The scripts for accessing and displaying the data

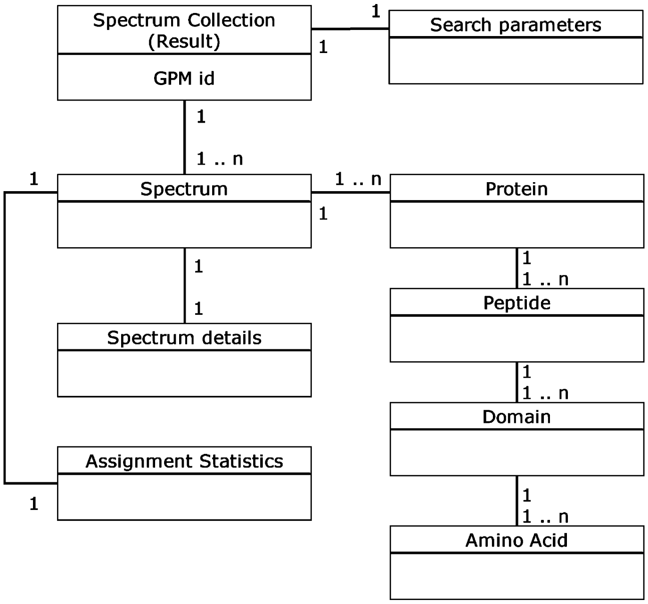

Figure 2. General UML class diagram describing the XML metadata about a sequence collection.

were written in Perl and SVG. All software for creating and populating the database and the database query and report generation software were made available for download as open source software under the Artistic License.[25]

A general object model for the XML files is shown in Figure 2. The XML files are the output metadata in BIOML,[26] generated by X! TANDEM. This file format was selected for simplicity and the fact that X! TANDEM was the only available open source software for performing the data analysis, so the output could be altered, if necessary, to fit the requirements of this system. There is no fundamental reason that metadata files from other search engines cannot be converted into this format and used by the system.

The identification numbers stored in the *Protein* and *Peptide* relations were used to address specific XML objects. Each XML **Protein** object was identified with a unique index (uid), a number generated from the order in which that sequence was read from the collection of sequences used in the analysis. The indexing for objects contained in a **Spectrum** object used a tuple of four identifiers:

$$N_{\text{spectrum}}.N_{\text{protein}}.N_{\text{domain}}.N_{\text{model}}$$

where N_{spectrum} identified the position of the spectrum in the original spectrum data file, N_{protein} identified the location of the protein in the list of homologous sequences, N_{domain} identified the peptide location in a protein sequence corresponding to the spectrum and N_{model} identifies a particular arrangement of modifications to that sequence that best fits the spectrum. The *Protein* relation maintained its positional information by the relative identifiers "*pida*" and "*pidb*", corresponding to "N_{spectrum}" and "N_{protein}". The position of the *Peptide* relation was retained by the relative identifiers "*dida*", "*didb*", and "*didc*", corresponding to "N_{spectrum}", "N_{domain}", and "N_{model}".

One of the persistently vexing questions regarding proteomics experiments has been how to deal with the fact that when a group of spectra have been mapped to the proteins in a proteome, more than one protein may share the same group of spectra. If the set of spectra associated with the n^{th} protein in a proteome is $\mathbf{P}_n$, this statement is equivalent to stating there may exist an m^{th} protein such that

$$\boldsymbol{P}_{\mathbf{n}} \cap \boldsymbol{P}_{\mathbf{m}} \neq \mathbf{0} \qquad (1)$$

where **0** represents the null set. The normal concept of sequence homology used in bioinformatics is difficult to apply to the classification of proteins in this case, because two sequences that are not formally homologous may be equally good fits to a particular mass spectrum, due to the ambiguity of a particular spectrum. The practical definition of homologues used to relate protein sequences to mass spectra and to each other was to group proteins together as homologues if they met the condition in eq 1. The number one ranked protein (*pidb* = 1) was considered to be the best fit based on the number of spectra that could be assigned to that protein, the length of the protein and the expectation values of the individual spectrum assignments. If one of the lower ranked proteins also meets the condition that

$$\mathbf{P}_{n} \cap \mathbf{P}_{m} \neq \mathbf{P}_{m} \qquad (2)$$

then that protein may also be listed as a top level protein hit, for those spectra (s) contained in the set, $\mathbf{P}_m - \mathbf{P}_n := \{s|s \in \mathbf{P}_m \wedge s \notin \mathbf{P}_n\}$.

The tandem mass spectra were represented in the XML files using the General Analytical Markup Language (GAML),[27] which was designed to effectively store multidimensional histogram data. The GAML information is represented using the GAML: namespace extension to BIOML. Additional information accumulated during the data analysis was also stored along with the mass spectra in GAML format. This information included histograms showing the distributions of matched y ions, b ions, correlation scores and hyperscores for each mass spectrum.

The XML data files range in size from 0.001 to 20 MB, depending on the size of the spectrum collection used to generate the files. These files were stored in 10 repository folders, with 1-folder for each public installation of the protein identification system. The XML files can be downloaded from the system and stored locally by any user. These locally stored files can then be uploaded to any of the data analysis servers for graphical and tabular presentation. This feature of the system was designed to make it relatively easy for groups to share data and to ensure the storage of important data. XML files that have been uploaded to a data analysis server are stored separately from files produced by analyzing new input mass spectrum files so they cannot introduce bias into the existing data. XML files can be stored locally in compressed GZIP format and uploaded in this compressed format. The verbose nature of XML makes compression very useful: compression ratios of 10:1 are commonly achieved with large files.

The decision was made to store the mass spectra in the XML files only, rather than including the entire set of spectrum histograms in the relational database. An analysis of potential use cases showed no additional utility attaching to the maintenance of a relation for such a large amount of data; no useful queries based solely on the *m*/*z*-intensity pairs were anticipated. Our experience with industrial database systems for storing mass spectrum-based information had demonstrated that including the *m*/*z* ratio and intensity information results in the majority of practical database maintenance issues, as it represents the largest volume of storage in the database. This decision constituted a major deviation of XIAPE from the original MIAPE proposal, which anticipates storing all spectrum information in a set of relations.

There was no attempt made to retain the details of the experiments involved, other than the details of the protein

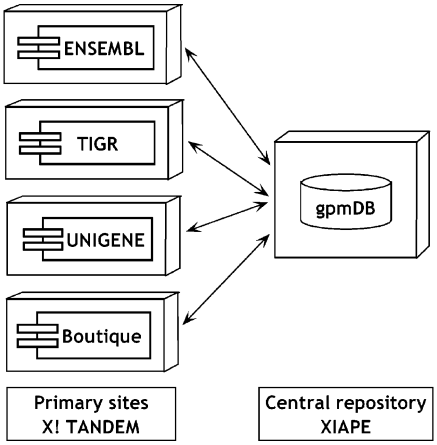

Figure 3. Deployment diagram illustrating how the elements of the Global Proteome Machine system interact with one another. The Primary data analysis sites, each containing some selection of ENSEMBL, TIGR, UNIGENE, or boutique sequence collections regularly deposit their analytical results to the XIAPE repository. This central data store can then be used to generate validation diagrams for data that enter the system at the primary sites.

sequence assignment runs. Any identifying information about the data submitter was stripped from the XML files, except when this information was deliberately added by the submitter.

Each collection of mass spectra analyzed was assigned a unique identifier, in the following format:

GPM*xxxnnnnnnnn*

where "GPM" was reserved as an identifier for the database, "*xxx*" were three digits, corresponding to the data analysis server that produced the results and "*nnnnnnnn*" were eight digits that served as a serial number for the collection. This accession identifier system (inspired by the ENSEMBL system) makes it possible to maintain links to the database in any LIMS system using MIAPE, simply by referring to this accession number. By further analogy with ENSEMBL, the system can be rationally extended by the insertion of uppercase alphabetic identifiers between "GPM" and "*xxx*".

3. Deployment of the System Components. A deployment diagram for the overall system is shown in Figure 3. The data analysis servers were classified as those using primarily genomic data, (ENSEMBL[28] and TIGR[29]), curated transcriptomic data (UNIGENE[30]) or other curated combinations of sequence sources for a specific species (referred to here as "boutique" collections) to generate model protein sequence lists. The results display pages generated by a server contained links to the results in the XIAPE repository. These links were served by HTTP requests to scripts that perform SQL queries and assemble reports based on the query results and retrieval of details from the XML files. The XML files themselves were exported to the XIAPE repository daily, for use by the entire system.

The data analysis systems were installed on a variety of computers, using dual Intel Xeon or single Pentium 4 processors. The operating systems were either Microsoft Windows 2000, Windows XP or Red Hat Linux. The database system used was the most recent version of MySQL.[31] The publicly available server was installed on a Dell 1650, with dual Intel XEON 1.4 GHz processors, using Microsoft Windows 2000 Server as the operating system. The choice of operating system for a particular server was one of convenience and availability: there were no operating system specific elements to the system. Use of the database schema on relational database systems other than MySQL may require alteration of the source code to compensate for differences between platforms. An effort was made to minimize the number of database vendor-specific features in the creation of the database and the database access software.

The system made use of a round robin Domain Name Service (DNS) implementation to distribute the load on data analysis servers with identical configurations. Each of these identical machines was given the same DNS name ("h.thegpm.org") as well as a unique DNS name. Users address the group of machines using the common name. For example, if there were two machines with the unique names "h1.thegpm.org" and "h2.thegpm.org", the first user to access the system through "http://h.thegpm.org" might be routed to h1. The next user would then be routed to h2, the next to h1, and so on.

The DNS round robin system of sharing a pool of servers among users was used for simplicity; a more efficient method of resource sharing may need to be devised as the system becomes more heavily used. As currently deployed, each of the 10 data analysis systems can process approximately 50 mass spectra per second. Therefore, the maximum capacity of the system is $10 \times 50 \times 3600 \times 24 = 4.3 \times 10^7$ spectra per day. Although this appears to be a large number, the combination of a "peak usage hours" effect, the limited memory of an individual computer and the limitations of round robin DNS means that the system in its current configuration would probably become functionally paralyzed for some time periods during a day in which more that 5×10^6 spectra were submitted.

Results and Discussion

As mentioned above, the XIAPE approach deviates from the MIAPE proposal in a number of important ways. The reasons for these deviations were basic: XIAPE was designed to be a resource for answering bioinformatics questions, while MIAPE was designed to provide a comprehensive LIMS environment. In the first seven months of operation, the public GPM data analysis servers have generated approximately 1.5×10^6 peptide-to-sequence assignments. The XML metadata files were successfully imported into GPMDB and the GPM annotation pages were linked to the database successfully. The data storage requirements have proven to be relatively modest, the database requiring an average storage of 0.14 kilobytes/peptide and the XML repository requiring an average of 1.8 kilobytes/peptide. The database growth rate, through the submission of data by users, was approximately 5×10^4 annotated peptide mass spectra per week.

The choice was made early in the design process to use protein sequences derived from genomes, whenever possible. The majority of the data have been collected using the ENSEMBL genomic protein sequence translations, with selected sequences obtained from NCBI-nr, e.g., the sequences of common experimental artifacts trypsin and bovine serum albumin. Using sequences that can be assigned directly to gene models has greatly increased the value of the resulting protein identifications, while eliminating the vexing problems of the high degree of practical sequence redundancy in NCBI-nr.

The current version of the system can be queried directly by a number of different attributes. These queries include:

(1) GPM accession number;

(2) Protein description keywords;

(3) ENSEMBL protein accession number;

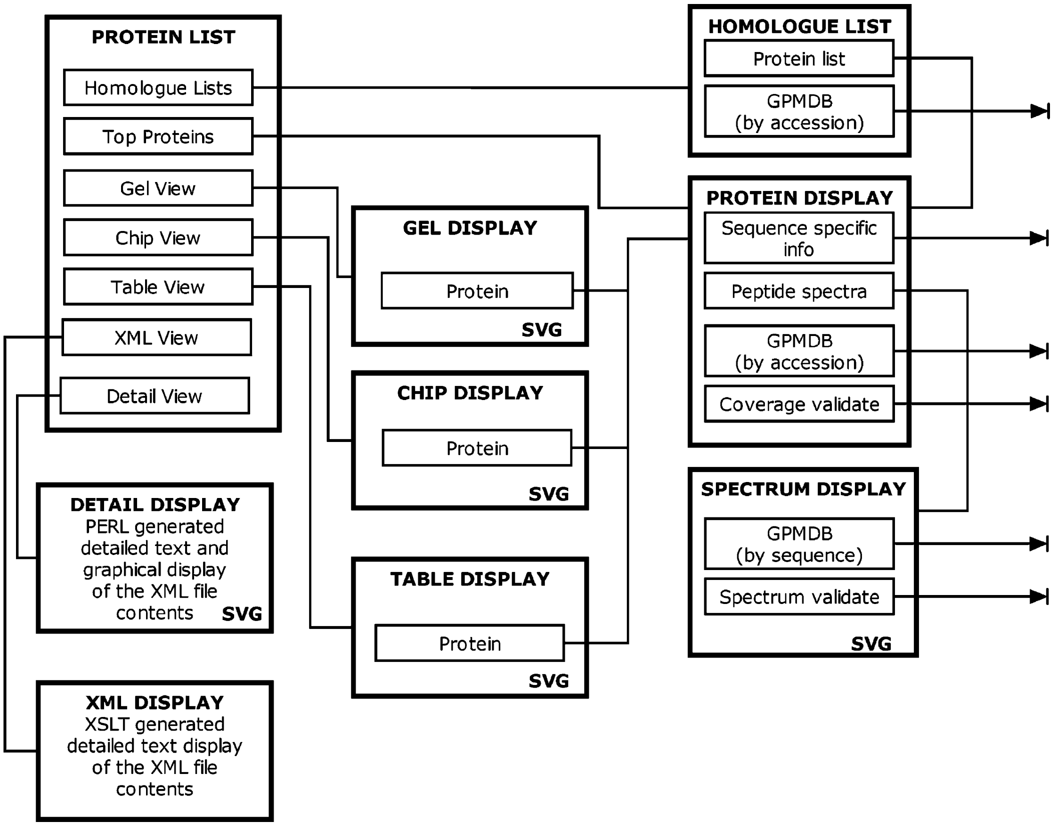

Figure 4. Relationships among the displays that allow access to information stored in the XML metadata files. This general layout was used on both the data analysis servers and the XIAPE server. Thick-lined boxes denote separate HTML pages, and thin-lined boxes indicate hyperlinks. "SVG" indicates that the page contains SVG graphics. Arrows leading out of the system link to resources external to an individual data analysis server.

(4) Organism-specific open reading frame accession number; and

(5) Observed peptide sequence.

A GPM accession number query returns a display showing the sequence coverage maps for the top scoring proteins found in that sequence collection. A protein accession number search returns coverage maps for all observations of that protein. Searching by peptide sequence returns links to all annotated spectra associated with that sequence, with any observed post-translational modifications or point mutations indicated on the sequence. The query result pages contain hyperlinks that guide the user through a series of specialized views drawn from the original XML data files. These views resemble as much as possible the analogous views of the results returned by individual data analysis servers.

The system also allows for the creation of a set of tables and views for the data associated with each XML file. Figure 4 represents a layout map of the relationships between these views. All the pages are generated by software and displayed without storing a copy of the page on the server. Pages containing SVG images require an SVG "plug-in" to be viewed properly with most Internet browsers.

The use of the system for validating experimental results, displaying large datasets and answering questions that query the complete repository will be unfamiliar for many users. A brief description of common use cases and an explanation of the features of the expected results have been given below. The user interface system may change from time-to-time, so the output of the system may vary from what is represented in these illustrations.

Validation of an Peptide Sequence Assignment. The validation of peptide-to-mass spectrum assignments was one of the tasks that GPMDB was designed to perform. To perform the validation, the desired peptide sequence from the protein display can be selected by clicking on the sequence, producing a display of the annotated mass spectrum. By selecting "validate" on the display, the GPMDB is then queried, resulting in a validation display of the type illustrated in Figure 5. The database query uses the sequence of the peptide assigned by the analysis server to produce a list of all of the spectra in GPMDB with same parent ion charge that have also been previously assigned to this sequence. The list of spectra is then ordered from the most to the least confident assignment. The validation display shows the current spectrum, with the annotated spectra retrieved from GPMDB displayed beneath it. By default, only the best validation spectrum would be shown. This display can be expanded to include as many validation spectra as are stored in the GPMDB. By comparing a result with the best prior results for that peptide, the consistency of the observed peptide bond fragmentation pattern can be assessed by inspection.

The particular validation result in Figure 5 illustrates the use of the display. The user's spectrum (a) was measured from a $z = 1$ parent ion and only has four ions assigned to the peptide sequence. Without further confirmation it would be very speculative to use this spectrum as evidence for a sequence assignment. The validation spectrum (b) automatically selected by the GPMDB showed the assignment of thirteen ions. Comparison of (a) and (b) by inspection demonstrated that the assignment (a) was reasonable and consistent with what would

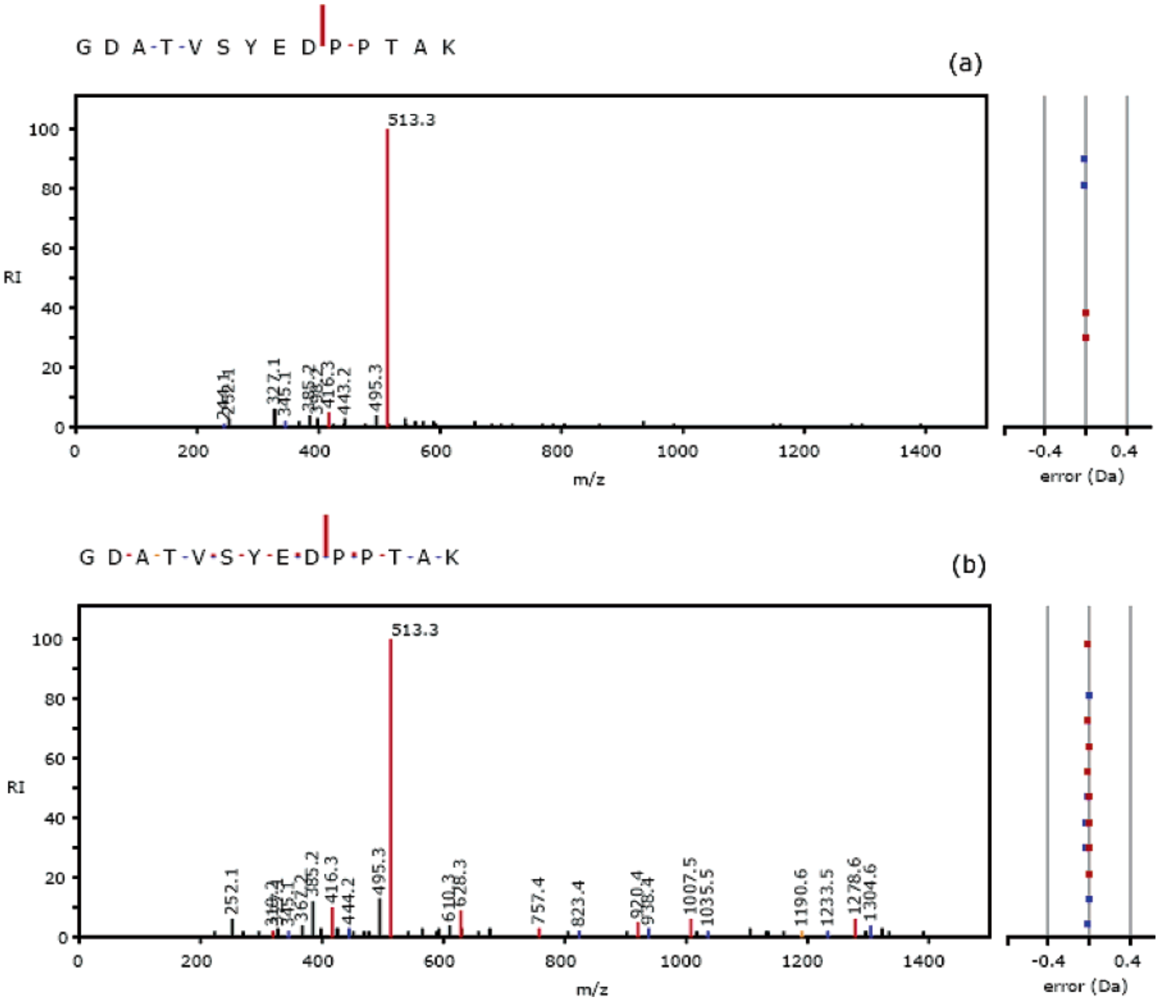

Figure 5. Peptide tandem mass spectrum validation diagram for an experimentally observed peptide sequence-to-spectrum assignment. Diagram (a) shows the annotated experimental spectrum, while (b) represents the most confidently assigned spectrum available. Red colored signals correspond to y-ions and blue signals correspond to b-ions. The error display to the right of the diagram illustrates the absolute error for each of the assigned fragment ions.

be expected of a spectrum of this peptide with a significantly weaker overall signal intensity.

The current system validates a spectrum-to-sequence assignment against the most confident assignments observed for a particular peptide sequence. This type of comparison may be misleading, when large differences in the signal-to-noise levels between the best assignment and the observed spectrum obscures underlying patterns. The application of a scoring scheme that would allow comparison between an observed and stored assignment with comparable signal-to-noise ratios may improve the utility of the system.

Validation of a Protein Sequence Assignment. After an experiment has been analyzed and the identified peptides have been associated with protein sequence models, the protein coverage map for an individual protein can be validated using a graphical comparison with other coverage maps retrieved from the GPMDB. This comparison is illustrated in Figure 6. This validation diagram is generated by selecting the “validate” link to the right of the coverage map shown on a protein display page. This produces a query to the GPMDB, returning a list of other recorded instances of results that correspond to the same accession number. These results are then ordered from most to least confident assignments. An effort is made to remove results that are simply multiple instances of the same set of spectra. The validation diagram is then constructed with the coverage map at the top corresponding to the user’s data, followed by up to twenty of the most confident sequence coverage maps in the repository. By comparing the pattern of coverage obtained in a single experiment with those obtained by others, it is often apparent which features of a particular map can be considered to be of high confidence.

The protein coverage validation diagram illustrated in Figure 6 demonstrates the features of the method. The top coverage map (labeled “Your result”) showed three peptides assigned to the sequence (*S. cerevisiae* open reading frame YGR092W). The sequence location and assignment confidence could be compared by inspection with the seven other coverage maps retrieved from the GPMDB. Examining each of the peptides, it was clear that each of them had been observed previously and that it was not unusual to see all three in the same experiment. Further inspection of the diagram shows that all of the observations of the protein were consistent, except map 7. The peptide indicated in map 7 was not seen in any of the other maps, it has a relatively poor confidence score (indicated by its weak coloring) and therefore it was probably assigned to this sequence as the result of a stochastic match.

The form of the current display may favor expert analysts. Additional effort will be required to make the displays and their

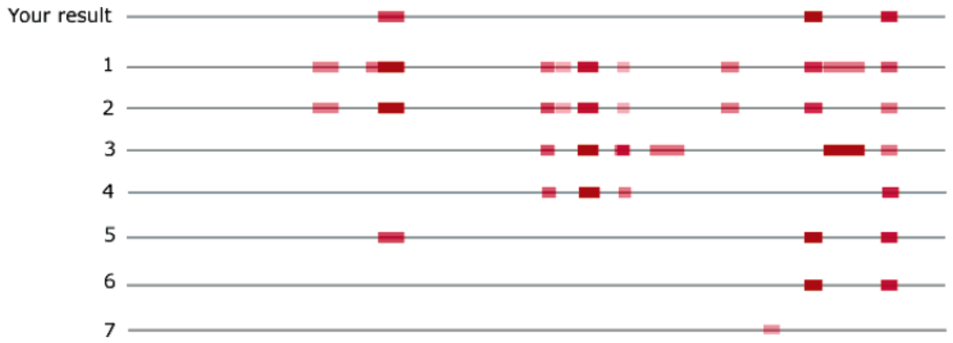

Figure 6. Protein sequence coverage map validation diagram for the experimental observation of *S. cerevisiae* open reading frame YGR092W. The current experimental result is on the top of the diagram, indicated by the words "Your result". The gray central lines indicate the full length of the observed protein sequence and the thick red lines indicate peptides that were observed. The intensity of the color indicated the confidence of the individual peptide sequence assignments. Seven previous observations of this protein are displayed for comparison.

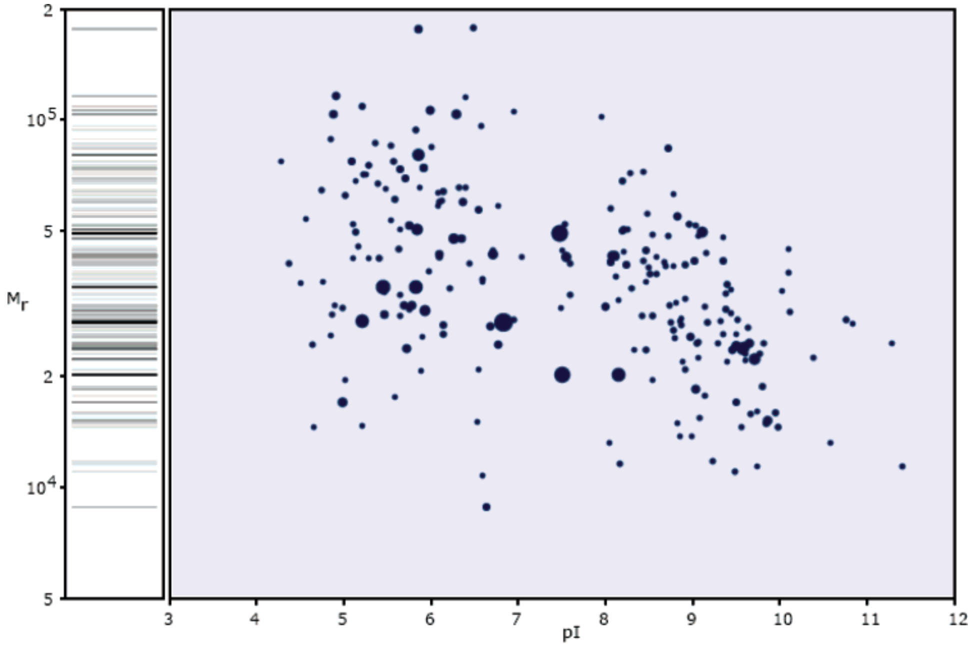

Figure 7. Simulated one- and two-dimensional electrophoresis gel representations of the data contained in the spectrum collection associated with GPM11200000578. All M_r and p*I* values were calculated from the intact gene product sequence. No attempt was made to remove signal sequences or to compensate for post-translational modifications that may have been present in the protein sequence.

interpretation as transparent as possible to the casual user. Ongoing research is being carried out into the development of clustering and scoring algorithms to simplify the task of comparing these maps.

Visualization of Complete Data Sets. Figure 7 represents an example of one of the most specialized views available from GPMDB for the representation of all of the proteins discovered from a large collection of spectra, such as an LC/MS/MS run. In this "gel" view, protein molecular masses and pI were calculated from the gene product sequence and plotted. In the narrow left-hand panel one-dimensional gel representation, the visual density of each "band" was determined from the summed intensity of the spectra assigned to a gene model. In the right-hand panel two-dimensional gel display, the size of the spot was varied corresponding to the summed spectrum intensity, rather than the density. The purpose of this view was to plan experiments and to predict the potential difficulties that might arise from the application of electrophoresis separation techniques to a given mixture of proteins. It has proven useful as a display method, as it visually classifies the observed proteins based on familiar sequence properties.

Evaluating the Evidence for a Particular Gene Model or Open Reading Frame. In the first three examples, the displays were generated by the system, using experimental data (a collection of mass spectra) to create links between the data, protein sequences and data held in the GPMDB repository. It is also possible to directly query the repository to answer specific questions about protein sequences of interest. For example, the annotation of the *S. cerevisiae* genome predicts many open reading frames that do not have any known biological function and which may not have been observed as translated proteins. These open reading frames are frequently referred to as "hypothetical proteins".

The open reading frame YDR131C was selected as an example of such a hypothetical protein. The diagram shown in Figure 8 illustrates the result of searching the GPMDB with

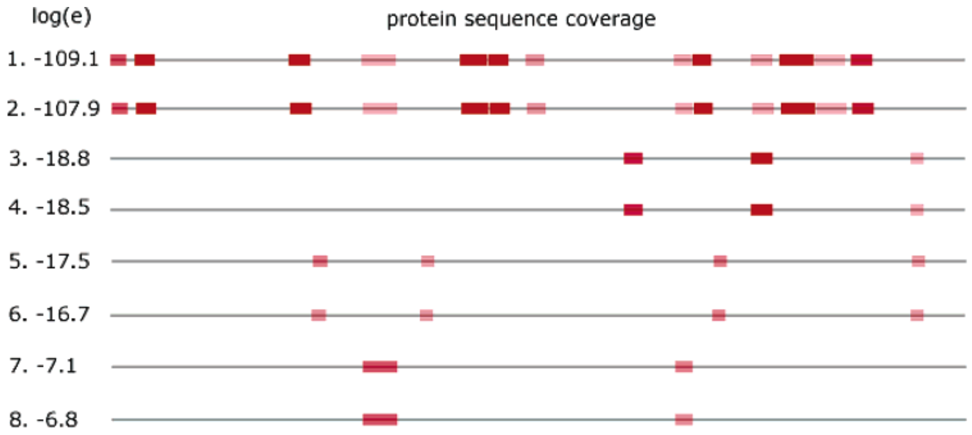

Figure 8. Illustration of the coverage map diagrams that results from querying GPMDB to find all of the observations of the *S. cerevisiae* open reading frame YDR131C. The hyperlinks in the original diagram have been removed for clarity.

Table 1. Selected Results Returned from GPMDB by a Query of All Known ENSEMBL Protein Identifiers for Plasma Membrane Proteins with Predicted Transmembrane Domains[a]

no.	protein identifier	total	log(*e*)	description
54	ENSP00000005593	66	−102.1	ADP,ATP CARRIER PROTEIN, FIBROBLAST ISOFORM (ADP/ATP TRANSLOCASE 2) (ADENINE NUCLEOTIDE TRANSLOCATOR 2) (ANT 2).
55	ENSP00000200639	92	−37.8	LYSOSOME-ASSOCIATED MEMBRANE GLYCOPROTEIN 2 PRECURSOR| (LAMP-2) (CD107B ANTIGEN).
56	ENSP00000007752	93	−37.8	LYSOSOME-ASSOCIATED MEMBRANE GLYCOPROTEIN 2 PRECURSOR (LAMP-2) (CD107B ANTIGEN).
77	ENSP00000278385	27	−6.4	CD44 ANTIGEN PRECURSOR (PHAGOCYTIC GLYCOPROTEIN I) (PGP-1) (HUTCH−I) (EXTRACELLULAR MATRIX RECEPTOR−III) (ECMR−III) (GP90 LYMPHOCYTE HOMING/ADHESION RECEPTOR)| (HERMES ANTIGEN) (HYALURONATE RECEPTOR) (HEPARAN SULFATE PROTEOGLYCAN) (EPICAN) (CDW44).
87	ENSP00000344206	1	−0.2	Rhomboid-related protein 1 (EC 3.4.21.-) (RRP) (Rhomboid-like protein 1). [Source:SWISSPROT;Acc:O75783]
106	ENSP00000328855	1	−0.8	P2Y purinoceptor 2 (P2Y2) (P2U purinoceptor 1) (P2U1) (ATP receptor) (Purinergic receptor). [Source:SWISSPROT;Acc:P41231]

[a] The “no.” column referred to the position of the row in the spread sheet supplied as Supporting Information. The “total” column was the number of datasets containing the protein identifier and “log(*e*)” was the logarithm of the expectation value for the best assignment for each identifier.

“YDR131C”, using the “accession number” search form accessible from the GPMDB home page. The results returned from this search indicated that YDR131C was observed eight times, with expectation values in the range from 10^{-109} to 10^{-7}. These expectation values quantify how often one would expect such an assignment to occur at random (a BLAST “*p*” value has a similar interpretation). The protein sequence coverage maps were consistent among the observations. From this simple query, it is possible to conclude that the open reading frame prediction was correct and that this “hypothetical” protein was translated in sufficient quantity to produce strong signals. Further inspection of the detailed protein and peptide displays (not shown) indicated that the N-terminus of the mature protein was unmodified following translation and that the gene model was confirmed up to residue 529. On the basis ofthis evidence, any revision of the *S. cerevisiae* genome that excluded YDR131C would almost certainly be in error.

Evaluating the Evidence for a Class of Genes. The GPM was designed to be compatible with a selection of other bioinformatics tools, which have capabilities that could be used to improve the utility of the repository. Currently, the most capable of these external resources is ENSEMBL. As an illustration of how GPM and ENSEMBL can be used together, consider the following question: what evidence is contained in the GPM for the detection of human proteins that are known to be on the plasma membrane and that contain at least one trans-membrane domain that is not a signal peptide? This seemingly difficult question can be answered in two simple steps:

(1) Use the ENSMART tool in ENSEMBL to produce a list of all human protein identifiers that have the Gene Ontology identifier corresponding to the plasma membrane (GO: 0005886) and have a trans-membrane domain.

(2) Copy that list into the multiple accession number search box on the “Accession number” form on the GPMDB site and perform the query.

The protein identifiers returned by ENSMART and the results of querying GPMDB were supplied as Supporting Information, in spreadsheet format. A total of 1492 protein identifiers were obtained and of those identifiers, 405 had entries in GPMDB. Depending on the software used to read the file, it should be possible to select any of the protein identifiers in the results to navigate to the supporting data in the repository.

Table 1 illustrates selected query results that demonstrate the range of information available. The first selected entry (#54) corresponds to a very strong protein assignment ($e = 10^{-102}$) and the protein has been identified often (66 times). Entry #55 and #56 were also often observed and they have a strong likelihood of being correct identifications. However, the descriptions for the protein are the same, so some investigation may be necessary to determine if they represent molecules that can be distinguished by proteomics data. Entry #77 corresponds to a significantly weaker sequence assignment than those above, however inspection of the protein coverage maps and peptide spectra (not shown) indicated that the recorded observations have been consistent. Entries #87 and #106 correspond to single assignments with very low confidence ($e \approx 1$). Basing any conclusions on these entries would be highly speculative.

Conclusions

The coupling of a protein identification system to the database and data retrieval system described here demonstrated that it was possible to use an XIAPE-type system to provide useful validation information about the assignment of mass spectra to peptide sequences in a simple manner, using a relatively compact relational database structure. The current version of the system has sufficient capabilities to use as a resource for the validation of protein and peptide sequence assignments.

It is hoped that this repository may prove useful for tutorial purposes, supplying a range of examples that can illustrate the appropriate use of the information derived from proteomics. The repository may also find some use in the planning of proteomics experiments, by providing insights into both the peptides that may reasonably be expected to produce signals and the other proteins that have been observed in concert with a desired protein species.

Acknowledgment. The authors would like to thank all of the individuals who have contributed their proteomic data to the Global Proteome Machine and well as the public spectrum collections: the Open Proteomics Database (bioinformatics.icmb.utexas.edu/OPD/) and Peptide Atlas (www.peptideatlas.org). The authors would also like to thank David Fenyö (General Electric Healthcare), Lorne Taylor, Bill Vensel (USDA, Albany CA), Phil Andrews and his group (University of Michigan, USA), Brett Phiney and Curt Wilkerson (Michigan State University, USA), Sanjeev Pillai (Northeastern University, Boston, USA), John Paul Waters (Wayne State University, USA), J. Daniel Navarro (Johns Hopkins University, School of Medicine, USA), Colton A. Smith and his group (University of Tennessee, Knoxville, USA), Rod Taylor (MDS Proteomics, Canada), John Wilkins and his group (Systems Biology Program, University of Manitoba, Canada), Jimmy Eng (Institute for Systems Biology, USA), Chris Paulse (Albert Einstein College of Medicine, USA) and Cory Bystrom (Oregon Health & Science University, USA) for their assistance in writing, debugging and testing parts of the code that has become part of the system described here. This work was funded by Beavis Informatics Ltd., the Province of Manitoba, the Winnipeg Regional Health Authority, and the Canadian Institutes for Health Research.

Supporting Information Available: Protein identifiers returned by ENSMART and the results of querying GPMDB. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Taylor, C. F.; et al. A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nat. Biotech.* **2003**, *21*, 247−254.
(2) Beynon, R. J. Enabling proteomics: the need for an extendable 'workbench' for user-configurable solutions. *Comput. Funct. Genom.* **2004**, *5*, 52−55.
(3) Orchard, S.; Hermjakob, H.; Julian, R. K.; Runte, K.; Sherman, D.; Wojcik, J.; Zhu, W.; Apweiler, R. Common interchange standards for proteomics data: Public availability of tools and schema. *Proteomics* **2004**, *4*, 490−491.
(4) Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature* **2003**, *422*, 198−207.
(5) Orchard, S.; Kersey, P.; Zhu, W.; Montecchi-Palazzi, L.; Hermjakob, H.; Apweiler, R. Progress in Establishing Common Standards for Exchanging Proteomics Data: the second meeting of the HUPO Proteomics Standards Initiative. *Comput. Funct. Genom.* **2003**, *4*, 203−206.
(6) Malmström, J.; Larsen, K.; Malmström, L.; Tufvesson, E.; Parker, K.; Marchese, J.; Williamson, B.; Hattan, S.; Patterson, D.; Martin, S.; Graber, A.; Juhasz, P.; Westergren-Thorsson, G.; Marko-Varga, G. Proteome Annotations and Identifications of the Human Pulmonary Fibroblast. *J. Proteome Res.* **2004**, *3*, 525−537.
(7) Rappsilber, J.; Mann, M. What does it mean to identify a protein in proteomics? *Trends Biochem. Sci.* **2002**, *27*, 74−78.
(8) Aebersold, R.; Goodlett, D. R. Mass spectrometry in proteomics. *Chem. Rev.* **2001**, *101*, 269−295.
(9) Fenyo D.; Beavis R. C. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.* **2003**, *75*, 768−774.
(10) Pevzner, P. A.; Mulyukov, Z.; Dancik, V.; Tang, C. L. Efficiency of database search for identification of mutated and modified proteins via mass spectrometry. *Genome Res.* **2001**, *11*, 290−299.
(11) Mann, M.; Pandey, A. Use of mass spectrometry-derived data to annotate nucleotide and protein sequence databases. *Trends Biochem. Sci.* **2001**, *26*, 54−61.
(12) Fenyö, D.; Qin, J.; Chait B. T. Protein identification using mass spectrometric information. *Electrophoresis* **1998**, *19*, 998−1005.
(13) Gu, C.; Tsaprailis, G.; Breci, L.; Wysocki, V. H. Selective gas-phase cleavage at the peptide bond C-terminal to aspartic acid in fixed-charge derivatives of Asp-containing peptides. *Anal. Chem.* **2000**, 72, 5804−13.
(14) Wysocki, V. H.; Tsaprailis, G.; Smith, L. L.; Breci, L. A. Mobile and localized protons: a framework for understanding peptide dissociation. *J Mass Spectrom.* **2000**, *35*, 1399−406.
(15) Tabb, D. L.; Huang, Y.; Wysocki, V. H.; Yates, J. R., 3rd. Influence of basic residue content on fragment ion peak intensities in low-energy collision-induced dissociation spectra of peptides. *Anal.Chem.* **2004**, *76*, 1243−8.
(16) Elias, J. E.; Gibbons, F. D.; King, O. D.; Roth, F. P.; Gygi, S. P. Intensity-Based Protein Identification by Machine Learning from a Library of Tandem Mass Spectra. *Nat. Biotech.* **2004**, *22*, 214−219.
(17) Havilio, M.; Haddad, Y.; Similansky, Z. Intensity-Based Scorer for Tandem Mass Spectrometry. *Anal. Chem.* **2003**, *75*, 435−444.
(18) Field, H. I.; Fenyö, D.; Beavis, R. C. RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimizes protein identification and archives data in a relational database. *Proteomics* **2002**, *2*, 36−47.
(19) Craig, R.; Beavis, R. C. A Method for Reducing the Time Required to Match Protein Sequences with Tandem Mass Spectra, Rapid Commun. *Mass Spectrom.* **2003**, *17*, 2310−2316.
(20) Gibbons, D. D.; Elias, J. E.; Gygi, S. P.; Roth, F. P. SILVER Helps Assign Peptides to Tandem Mass Spectra Using Intensity-Based Scoring. *J. Am. Soc. Mass Spectrom.* **2004**, *15*, 910−912.
(21) Aebersold, R. Constellations in a cellular universe. *Nature* **2003**, *422*, 115−6.
(22) http://www.apache.org.
(23) Wall, L.; Christiansen, T.; Schwartz, R. L. Programming Perl, O'Reilly, Sebastopol, CA, p 645.
(24) http://www.w3.org.
(25) http://www.thegpm.org.
(26) Fenyö, D. The Biopolymer Markup Language. *Bioinformatics* **1999**, *15*, 339−340.
(27) http://www.gaml.org.
(28) Birney, E.; et al. An Overview of ENSEMBL. *Genome Res.* **2004**, *14*, 925−928.
(29) http://www.tigr.org.
(30) Wheeler, D. L.; Church, D. M.; Federhen, S.; Lash, A. E.; Madden, T. L.; Pontius, J. U.; Schuler, G. D.; Schriml, L. M.; Sequeira, E.; Tatusova, T. A.; Wagner, L. Database resources of the National Center for Biotechnology. *Nucleic Acids Res.* **2003**, *31*, 28−33.
(31) http://www.mysql.org.

PR049882H