Experimental Peptide Identification Repository (EPIR)

AN INTEGRATED PEPTIDE-CENTRIC PLATFORM FOR VALIDATION AND MINING OF TANDEM MASS SPECTROMETRY DATA

Dan Bach Kristensen‡§, Jan Christian Brønd§, Peter Aagaard Nielsen, Jens Roswald Andersen, Ole Tang Sørensen, Vibeke Jørgensen, Kenneth Budin, Jesper Matthiesen, Peter Veno, Hans Mikael Jespersen, Christian H. Ahrens, Soeren Schandorff, Peder Thusgaard Ruhoff, Jacek R. Wiśniewski, Keiryn L. Bennett, and Alexandre V. Podtelejnikov

LC MS/MS has become an established technology in proteomic studies, and with the maturation of the technology the bottleneck has shifted from data generation to data validation and mining. To address this bottleneck we developed Experimental Peptide Identification Repository (EPIR), which is an integrated software platform for storage, validation, and mining of LC MS/MS-derived peptide evidence. EPIR is a cumulative data repository where precursor ions are linked to peptide assignments and protein associations returned by a search engine (e.g. Mascot, Sequest, or PepSea). Any number of datasets can be parsed into EPIR and subsequently validated and mined using a set of software modules that overlay the database. These include a peptide validation module, a protein grouping module, a generic module for extracting quantitative data, a comparative module, and additional modules for extracting statistical information. In the present study, the utility of EPIR and associated software tools is demonstrated on LC MS/MS data derived from a set of model proteins and complex protein mixtures derived from MCF-7 breast cancer cells. Emphasis is placed on the key strengths of EPIR, including the ability to validate and mine multiple combined datasets, and presentation of protein-level evidence in concise, nonredundant protein groups that are based on shared peptide evidence. Molecular & Cellular Proteomics 3:1023–1038, 2004.

LC MS/MS has become a well-established technology for large-scale protein characterization in proteomic research (1). Briefly, proteins in the sample are digested with an enzyme, typically trypsin, because the tryptic peptides are more compatible with MS/MS analysis. The mass spectrometer is coupled to a reverse-phase LC unit, which reduces sample complexity and increases concentration of the peptides during MS acquisition. Throughout the LC MS/MS analysis, peptides are isolated and fragmented by CID to generate sequence-dependent MS/MS information, and finally the data is matched against a sequence database using a search engine. In a typical LC MS/MS acquisition, hundreds to thousands of precursor ions are subjected to MS/MS. As LC MS/MS can now be performed in a fully automated fashion, a new challenge faces investigators: the ability to generate LC MS/MS data outpaces the ability to analyze it (2).

One of the initial challenges when analyzing LC MS/MS data is the assignment of peptides to precursor ions. Currently, this is typically achieved by statistical algorithms that match a theoretical peak list with the measured peak list and include the cross-correlative Sequest algorithm (3) and probability-based algorithms such as Mascot (4). The number of incorrect peptide assignments made using probabilistic or cross-correlative algorithms alone can become an issue, as different peptides may have overlapping or even identical fragmentation patterns (e.g. Leu/Ile substitutions). This issue is particularly valid for large LC MS/MS datasets and/or when a high sensitivity (i.e. the true positive rate) is required, e.g. in target discovery projects (5). The end result is that a substantial amount of time and resources are required for manual validation. The sensitivity of the peptide assignment can be improved at different levels, including: additional processing of the MS/MS data (6); improved charge-state determination (7, 8); removal of low-quality MS/MS data (9); and clustering of redundant spectra (10). Alternatively, a higher peptide assignment sensitivity can be achieved using sophisticated scoring schemes that exploit empirical information derived from MS/MS data and the search results. For instance, this could be the presence of consecutive fragment ions (i.e. sequence tag-like information); specific fragmentation signatures (e.g. a relatively intense proline ion); and the number of sibling peptides (NSP) (11). For example, Colinge et al. introduced a new

1 The abbreviations used are: NSP, number of sibling peptides; EPIR, Experimental Peptide Identification Repository; PPI, precursor peak intensity; IDA, information-dependent acquisition.
the SALSA scores peptides based on how well the theoretical ion series for peptide sequence motifs correspond with the actual MS/MS product ion series, regardless of absolute position on the \( m/z \)-axis. The approach can be used in the identification of both unmodified and modified peptides (e.g. post-translationally or genetically). In the present study, we address the peptide assignment issue by exploiting various empirical parameters when validating the assignments returned by the Mascot search engine.

Once peptides have been assigned to the precursor ions, the next step is presentation of the protein evidence. This is a challenging task due to the degenerate nature of peptides, i.e. the same peptide can be derived from more than one protein entry. This redundancy may be derived from, e.g. homologous proteins or protein splice variants, or the database itself may be redundant. In many cases the MS/MS evidence therefore points toward a group of proteins rather than a single protein, and it may be impossible to determine which group members are present in the actual biological sample on the basis of the MS/MS evidence alone. Consequently caution should be taken when ranking protein hits, e.g. as seen in the result summary returned by Mascot (4). Nesvizhskii et al. addressed this issue by designing a statistical model for identifying proteins by LC MS/MS (11). Redundant protein identifications (i.e. assignments that can not be distinguished by the MS/MS evidence) were collapsed into a single identification, and a minimal protein list was generated using an expectation-maximization algorithm.

A major challenge in the analysis of proteomes is to maximize the information that can be extracted from a biological sample. There are several approaches, such as two-dimensional LC MS/MS, multi-step fractionation, and multiple analysis of the same sample using an exclusion list approach. The challenge remains, however, in how to deal with the huge amounts of data generated from proteomic analysis of complex biological samples, both in terms of database searches and data validation/mining. All these issues were central in the development of the new generic software platform presented here.

A peptide-centric relational database (Experimental Peptide Identification Repository, or EPIR) was developed for the storage, validation, and mining of LC MS/MS data. EPIR is a data storage area for all precursor ions to which peptides have been assigned by a given search engine.

At the same time, EPIR is cumulative, meaning that any number of datasets can be parsed into EPIR at any given time, and subsequently validated/mined as a single combined dataset. A set of software modules have been developed to automatically validate and mine datasets stored in EPIR. For instance, one module collapses proteins into groups on the basis of shared peptides. Protein evidence is thus presented in concise protein groups rather than as a ranked list of proteins, and this significantly reduces the complexity of the result summary. All proteins with conclusive, unambiguous MS/MS evidence are automatically highlighted within the group. Using a validation module, peptide assignments returned by the search engine (e.g. Mascot) are automatically validated or reassigned within EPIR, on the basis of different empirical parameters; including the presence of consecutive y/b-ions, the relative intensity of proline fragment ions (14), and the NSP. This functionality greatly enhances the ability to validate peptide assignments from large datasets in an automatic fashion. A generic quantitative module compatible with non-coeluting labels has also been developed to extract quantitative information from any type of differential experiment, regardless of the labeling method used (chemical or metabolic). Statistical modules were developed to extract information related to the quality of the datasets stored in EPIR. A key feature of the system is that no evidence is lost during data validation and mining, because the core data (a list of precursor ions with all potential peptide identifications and protein associations) remains unaffected at all times. This is because the data validation and mining process simply provides a means of filtering and organizing the core data, with the aim of addressing specific biological or analytical questions. In the present study, the utility of EPIR and associated modules is demonstrated on LC MS/MS datasets generated on a Q-TOF mass spectrometer.

**EXPERIMENTAL PROCEDURES**

**Standard Protein Sample**—A mixture of six model proteins (bovine albumin, rabbit aldolase, horse ferritin, chicken ovalbumin, bovine ribonuclease A, and bovine thyroglobulin, all from Amersharm Biosciences, Upsala, Sweden) was digested with Lys-C endopeptidase (*Achromobacter lyticus*; Wako Pure Chemicals, Osaka, Japan) in 4 M urea and 10 mM Tris-HCl pH 8.5 for 4 h at room temperature. The proteins were reduced with 10 mM DTT for 30 min at 37 °C, alkylated with 50 mM iodoacetamide for 30 min at room temperature, diluted 4-fold in 100 mM NH₄HCO₃, and digested overnight at 37 °C with trypsin (sequencing grade; Promega, Madison, WI). LC MS/MS (one standard and two exclusion list analyses) was performed as described below using 2 μl of the sample (500 fmol of each protein).

**Cell Culture**—The human breast carcinoma MCF-7 cell line was cultured in Dulbecco’s modified Eagle’s medium supplemented with 10% FCS, and 1% penicillin/streptomycin, 0.01 mg/ml insulin, 1.5 g/liter sodium bicarbonate, and nonessential amino acids. The cells were maintained at 37 °C in a humidified atmosphere of 95% air and 5% CO₂. For isotopic labeling, the cells were grown for at least six cell divisions in medium deficient in L-leucine, supplemented with 10% double diazylated FCS (HyClone, Logan, UT) and 52 mg/ml normal L-leucine (Leu D0) or [5,5,5-D₃]-L-leucine (Leu D₃) from Sigma-Aldrich (St. Louis, MO).

**Preparation of Enriched Plasma Membranes by Density Gradient Centrifugation**—A total of 5 × 10⁶ cells were homogenized in 10 ml of GB buffer containing 0.25 M sucrose, 10 mM HEPES-NaOH, 2 mM CaCl₂, 2 mM MgCl₂, 1 mM AEBSF hydrochloride, 1 mM EDTA, 20 μM leupeptin hemisulfate, 150 μM aprotinin, pH 7.4 (buffer A) using a motor-driven Potter homogenizer (B. Braun Biotech, Allentown, PA). The homogenate was centrifuged at 1,000 × g for 10 min, and the
supernatant collected. Homogenization and centrifugation were repeated. The post-nuclear supernatant was centrifuged at 50,000 × g for 30 min. The resultant pellet containing crude membranes (P2) was resuspended in 4 ml of GB buffer and mixed with 3.85 ml of 100% Percoll (Amersham Biosciences) and 0.55 ml of 2 μm sucrose in a 11.5-ml crimp tube (tube PA 11.5 ml; Sorvall, Asheville, NC). The tube was filled with GB buffer, capped, and centrifuged at 50,000 r.p.m. in a fixed-angle rotor T 890 (Sorvall) at 4 °C for 15 min. The gradient was fractionated from the top by the displacement method. In order to select fractions containing enriched plasma membranes, individual fractions were assayed for γ-glutamyl transpeptidase, cytochrome c oxidase, and NA DH-cytochrome c reductase as described previously (15). Total protein was determined fluorometrically on solubilized and denatured proteins by measuring the fluorescence of tryptophan (excitation at 295 nm, emission at 360 nm) using tryptophanamide as a standard.

Protein Reduction, Alkylation, and Digestion—Percoll was removed by centrifugation of the fractions in 1-ml 1PC tubes at 900,000 × g in Sorvall RC M150 GX using the S150AT rotor at 4 °C for 20 min. The isolated membrane fractions were washed and the proteins reduced with DTT on membrane as described previously (15). Finally, the membranes were resuspended in 200 μl of 4 μl urea in 0.1 M Tris-HCl, pH 8.0. Next, 20 μl of 1 M iodoacetamide was added and the mixture incubated at room temperature for 2 h. The membranes were collected by centrifugation at 900,000 × g at 4 °C for 20 min, and the pellet was resuspended in 200 μl of 4 μl urea in 0.1 Tris-HCl pH 8.0. Five micrograms of endoproteinase Lys-C were added, and the membranes were incubated overnight at room temperature. The released peptides were separated from the membranes by centrifugation. This procedure yielded 88 ± 10 μg peptide per 5 × 10^6 cells.

Reverse-phase Chromatography and Tryptic Digestion of the Lys-C Peptides—The Lys-C peptides were separated over a Dionex Acclaim 300 C18 3-μm column (i.d. 2.1 mm × 150 mm). The peptides were eluted with an ACN gradient in water containing 0.1% TFA. The flow was generated manually by importing the precursor ion list (text file) into the flow controller of the Switchos unit. The peptides were subsequently eluted at 300 g 18 3-μm column (~10 μg/min) using reverse-phase chromatography, and fractions were collected and lyophilized. The fractionated peptides were dissolved in 20 μl of 100 mM NH_4HCO_3 and incubated overnight at 37 °C with 0.5 μg trypsin.

LC MS/MS—All LC MS/MS experiments were performed on a QStar Pulsar XL (MDS Sciex, Toronto, Canada) connected to an LC Packings Ultimate System equipped with a Famos autosampler and Switchos unit (LC Packings, Sunnyvale, CA). All hardware systems were controlled from the Analyst QS software (MDS Sciex). Samples were loaded onto the precolumn (4 cm × 150 μm, Zorbax SB-C18 5-μm beads) using a flow rate of 5 μl/min solvent A (0.005% heptafluorobutric acid and 0.4% acetic acid in HPLC-grade water) using the Switchos unit. The peptides were subsequently eluted at 300 nL/min from the precolumn over the analytical column (4 cm × 75 μm, Zorbax SB-C18 3.5-μm beads) using an 80-min gradient from 10–35% solvent B (90% ACN, 0.005% heptafluorobutric acid and 0.4% acetic acid in HPLC-grade water) delivered by the Ultimate CAP pump. The total duration of the LC run was 120 min, including sample loading and column equilibration. The QStar XL was operated in information-dependent acquisition (IDA) mode. In MS mode, ions were screened from m/z 350–1,000, and MS/MS data were acquired from m/z 80–1,000 (QStar pulsing mode on). In standard acquisition mode, each acquisition cycle was comprised of a 1-s MS and a 2-s MS/MS. MS to MS/MS switch threshold was set to 40 cps. Five exclusion list runs were performed, where all precursor ions subjected to MS/MS in the previous run(s) were excluded for 9 min using a 3-μm window. The broad exclusion window (±4.65 min) was necessary as the retention time for individual precursor ions drifted up to 4 min during the 5 days required to exhaustively analyze a single biological sample. The exclusion list acquisition methods were generated manually by importing the precursor ion list (text file) into the Analyst method editor. In the first exclusion list analysis, the MS and MS/MS acquisition times and the MS to MS/MS switch threshold were unaltered (1 s, 2 s, and 40 cps, respectively). In the latter exclusion list analyses (run 3–5), the MS/MS acquisition time was increased to 3 s and the MS to MS/MS switch threshold was lowered to 25 cps.

Database Searching—The IDA processor (Applied Biosystems, Foster City, CA) was used to generate Mascot msm files with peak lists from the Analyst wiff files. The IDA settings were as follows: default charge state was set to 2+, 3+, and 4+; MS centroid parameters were 50% height percentage and 0.05 amu merge distance; all MS/MS data were centroided, with a 50% height percentage and a merge distance of 0.05 amu. The threshold peak intensity was set to 2 cps; MS/MS averaging parameters was set to reject spectra with less than 5 peaks or precursor ions with less than 5 or more than 10,000 cps; the precursor mass tolerance for grouping was set to 1; and the maximum and minimum number of cycles between groups was set to 10 and 1, respectively. MS/MS data from the standard protein sample was searched as a single merged msm file against all entries in the public NCBI database (downloaded November 23, 2003; 1,543,949 entries in total) from the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov) using the Mascot search engine (version 1.9.05; Matrix Science, London, United Kingdom). MCF-7 data were searched against human database entries only. Alkylation of cysteine residues was set as a fixed modification, and oxidation of methionine was set as a variable modification for all Mascot searches. One missed trypsin cleavage site was allowed, and the peptide MS and MS/MS tolerance was set to 0.3 and 0.13 Da, respectively. All common porcine trypsin autoprotopeolysis products were excluded after the data was entered into EPIR.

EPIR—EPIR was implemented as a standard SQL database using MySQL version 3.2.4 running on a PC equipped with RedHat Linux 9.0. EPIR contains structured information concerning samples (name, type, LIMS link); acquisitions (filename, time); raw data preprocessing (filtering parameters, processing application); database identification parameters (MS and MS/MS identification tolerance, database name, database version, species restrictions); spectrum identification results (peptide sequence, score, delta mass, expected ions, calculated ions, retention time); and relationship to the associated proteins.

Parsing Identification Results in EPIR—A software module has been developed to extract peptide identification information from the original Mascot result file into EPIR. The following information is extracted: search parameters; query/precursor; peptide score; suggested modifications; protein assignments; retention time; and fragment ion matches. Parsers for other result formats, i.e. Sequest (ThermoElectron, Waltham, MA), PepSea (MDS Inc., Odense, Denmark) are under development.

Quantitation—The precursor peak intensity (PPI, i.e., the maximum ion count observed for a precursor ion) is extracted automatically for all suggested peptide matches. The PPI is obtained directly from the raw MS acquisition file. The window in which the PPI is extracted is 60 s pre- and 90 s post-MS/MS acquisition time for both the identified peptides and nonidentified partner ions. For nonidentified peptide partners, the PPI is extracted using the theoretical precursor ion mass predicted from the identified peptide partner. The PPI elution profile is Gaussian fitted using nonlinear least squares. If the profile exceeds a 3-min elution time, and the PPI value is not observed within the analysis window, then the profile is excluded. A 3-min elution time was chosen as the maximum because most precursor ions eluted in less than 2 min. Furthermore, quantitative data was excluded if the difference in PPI times for non-coeluting peptides exceeded 30 s. This value was chosen because previous experiments have shown that more than 95% of light and heavy peptides have differences in PPI times that were less than 30 s (data not shown).

Molecular & Celluar Proteomics 3.10 1025
Experimental Peptide Identification Repository

A

B

MCP
MOLECULAR & CELLULAR PROTEOMICS

Molecular & Cellular Proteomics 3.10
Automatic Peptide Assignment—Besides the standard search engine results used for peptide assignment (score, expected versus calculated fragment ions, delta mass), additional empirical information is computed by the EPIR peptide validation module to assist in the assignment. Currently these include: a) the NSP for all potential peptide hits; b) the presence of consecutive y/b fragment ions; and c) a proline score for potential peptide identifications containing proline residues.

a) For each potential peptide assignment the NSP is computed. In this study, all peptide identifications returned with a Mascot score of ≥20 were included; however, the score threshold can be defined by the user. The average peptide score is calculated for each group of sibling peptides, and in cases where two or more peptide identifications have an identical NSP the identifications are ranked according to the average peptide score.

b) A synthetic structural fragment ion score is introduced for both y- and b-ion fragment series. This score reflects the presence of consecutive fragment ions and thus mimics the specificity of a sequence tag. Consecutive fragment ions separated by at least one nonmatching fragment ion are grouped into partial tags. For each partial tag, a cumulative product score, pts, is computed:

\[
pts = \prod_{i=0}^{n-1} \left(100^i/(3 + 2.0)\right)
\]

where \(i\) is the consecutive matched fragment ion and \(n\) is the total number of matched fragment ions of the partial tag. The total fragment ion score, \(s\), is computed:

\[
s = \sum_{i=0}^{n-1} \log_{10}(10/p_i)(pts_i + pts_{i+1})
\]

where \(i\) is the tag index, \(n\) is the total number of tags, and \(p_i\) is the number of nonmatching fragment ions between the partial tags.

c) In CID-based MS/MS, proline residues show a strong preference for cleavage at the N-terminal amide bond, which produces intense y-ions containing an N-terminal proline residue (14). If the suggested peptide contains at least one proline residue, a proline score, \(ps\), is, therefore computed. The score reflects the relative intensity of the y-ion containing an N-terminal proline compared with all fragment ions in the MS/MS spectrum, and is calculated as follows:

\[
ps = 100 - 100^{\left(pr - 1 \right)/n}
\]

where \(pr\) is the intensity rank of the proline-containing y-ion and \(n\) is the total number of fragment ions in the MS/MS spectrum. A proline score of 100 therefore indicates that the proline fragment ion is the most intensive peak, i.e. \(pr\) is 1.

The automatic assignment of peptides to precursor ions is initially based on the NSP information. If any of the suggested peptides have an NSP > 1, these peptides will be the only entries considered. In addition to the NSP information, the suggested peptide is removed from the potential list if the elution profile is invalid (see “Quantitation”); if the suggested number of labels do not match the number of labeling sites (isotopically labeled samples); and if the proline score is below 80 (proline-containing peptides). After the list of potential peptides has been generated, the peptide with the highest average group score (average Mascot score for all peptides in the group) and structural score (average y-ion or b-ion scores for all peptides in the group) is selected as the correct assignment. In cases where the same peptide has been identified multiple times, the identification with the highest score will be used. In situations where two or more peptides have the same group and structural score, no peptide will be assigned. The spectrum is flagged for manual inspection.

If all suggested peptides have an NSP of 1, a list of possible peptides is generated based on a valid elution profile (see “Quantitation”); correct number of labels (isotopically labeled samples); a proline score greater than 80 (proline-containing peptides); and a minimum of five consecutive fragment ions. The peptide with the highest structural and identification score will be selected. When two or more peptides have the same group and structural score, the spectrum is flagged for manual inspection.

Protein Grouping—Proteins with shared peptides are collapsed into a group and reported as a single identification, with the highest-scoring protein entry as the anchor. All information on the proteins in a group is stored in a collapsed format. Consequently no protein evidence is removed or lost. Protein groups with unambiguous protein identifications are highlighted. ClustalW was used for aligning the entries within a protein group (16). The following ClustalW settings were used: -quicktree, -score absolute, -output gde, -outorder input, -case upper, -GAPOPEN = 200, and -GAPEXT = 100.

Result Browser—The data entered into EPIR was managed, viewed, merged, and analyzed using a web application module. The module was developed using J2EE and running under a Jboss web application server version 3.0.4.

RESULTS

To evaluate the robustness and functionality of the EPIR platform, several types of analyses were performed. In order to assess the data validation module and protein grouping functionality of EPIR, a protein mixture containing six model proteins was used. Biochemical treatment of the mixture was identical to that used in the analysis of complex samples. Next, the ability of EPIR to process large LC MS/MS datasets was assessed using plasma membrane fractions generated from an MCF-7 cell line. Each fraction was analyzed six times using an exclusion list approach, resulting in a total of 60 LC MS/MS analyses. The same cell line was chosen to perform quantitative analysis of the proteins in a complex mixture. For this experiment, a series of different heavy:light ratios were

![Fig. 1. EPIR protein summary for standard protein sample. A standard sample containing six model proteins, digested with Lys-C endopeptidase and trypsin, was analyzed by LC MS/MS. A, the Mascot search results were parsed into EPIR, which returned a total of 12 protein groups with two or more unique peptides (number of sibling peptides/group = 2). Each protein group has a link to the group details and sequence alignment view (Group Info and Alig) and the total number of proteins in the group. Below the Group Info and Alig line an anchor entry is shown (database identifier and protein name), and unambiguously identified database entries are highlighted in green. To the right, the total score from Mascot is shown in yellow. B, EPIR ClustalW sequence alignment view of the ovalbumin group members. Besides the alignment, the view contains details on the individual group members, and there are options to manually assign an alternative anchor protein or remove unwanted group members. Eight database entries matched the three ovalbumin peptides in this example, and EPIR presented these as a single protein group.](image-url)
prepared, and quantitative analysis was based on the metabolic labeling of Leu with three deuterium atoms (17).

Standard Protein Analysis—The mixture of six model proteins was analyzed by one standard and two exclusion list analyses. The individual msm files were merged and searched as a single file (supplemental material) using the Mascot search engine. A total of 1,883 MS/MS spectra were generated, and Mascot returned 88 confident protein identifications.
defined by Mascot as entries with a total score ≥45), including the six proteins in the sample and Lys-C endopeptidase. Among the 88 returned identifications, 46 contained bold peptide assignments (data not shown). On other words, Mascot suggested that the sample contained 46 distinct protein hits. Most of the 88 proteins, however, were redundant identifications of the same protein from different species and sequence redundancy within the database itself (e.g. entries with partial sequences) (data not shown). The results were parsed into EPIR, and proteins were grouped on the basis of shared peptide evidence. A total of 12 protein groups (NSP/group ≥ 2) were generated by EPIR (Fig. 1A), and seven of these groups were derived from the six model proteins (ferritin light and heavy chain each contributed one group). Fig. 1B shows the EPIR protein alignment view of the ovalbumin group, which contained eight members. Lys-C endopeptidase was also identified as a group, thereby explaining eight of the observed groups. The remaining four groups were hemoglobin, troponymosin, lactate dehydrogenase, and phosphofructokinase, which had 5, 5, 13, and 2 unique peptide identifications, respectively. The MS/MS evidence for these proteins was conclusive, and it was therefore concluded that the proteins were contaminants of the original protein mixture. A total of 199 precursor ions, corresponding to 133 unique peptides, were assigned to the 12 protein groups. Among the 199 peptide assignments there were 14 examples where incorrect peptides had an identical or a higher Mascot score than the correct peptide. In all cases, EPIR assigned the correct peptide on the basis of the NSP, y- and b-ion scores, as well as the proline score (Fig. 2).

Multiple Exclusion List Analyses—Plasma membrane was enriched from MCF-7 cells and the protein content digested with Lys-C endoproteinase, followed by off-line rpHPLC separation into 10 fractions. The peptide fractions were digested with trypsin, and each fraction was subjected to one standard and five exclusion list LC MS/MS analyses. Mascot results from the 60 LC MS/MS runs were parsed into EPIR, and peptide assignments were automatically validated using the peptide validation module. A total of 10,579 precursor ions were identified. After collapsing the data, with respect to redundancy, charge state, and oxidation variants, 3,952 unique peptides were present. Next, proteins were grouped on the basis of peptide degeneracy, i.e. all database entries with ambiguous MS/MS evidence were collapsed into a single group. Fig. 3A illustrates the number of protein groups (NSP/group ≥ 2) as a function of the number of LC MS/MS analyses. In the first analysis, a total of 278 protein groups were identified. This number increased to 409 groups after the five exclusion list analyses, corresponding to a relative increase of 32%. Using the TMHMM 2.0 transmembrane prediction tool (18), it was found that 150 or 37% of the 409 groups contained protein entries with predicted transmembrane domains (data not shown). The total number of human protein entries in the 409 protein groups was 5,914, and consequently the grouping functionality of EPIR resulted in a 14-fold reduction in the complexity of the protein level evidence. Fig. 3B illus-
trates the NSP as a function of the number of LC MS/MS analyses. The average NSP for the protein groups identified in the first analysis is shown as a function of the number of LC MS/MS analyses. An increase from 4.2 to 7.1 NSP/group was observed, corresponding to a relative increase of 69%. Consequently, both the quantity (protein groups) and quality (peptide coverage or NSP) improved as a function of the number of LC MS/MS exclusion list analyses. Fig. 4A shows the collapsed protein group view of EPIR, from which the individual protein groups can be accessed (Fig. 4B). The protein group window contains all group members, a sequence alignment of the members, and the peptides assigned to the group. Individual peptides can be expanded (in this example NLSDVATK) to show all potential peptides for the given precursor ion, with an option to manually alter the peptide assignment.

Fig. 4. EPIR protein group view. EPIR groups proteins on the basis of the shared peptide evidence. The collapsed EPIR protein group view presents a list of protein groups with a group identifier, an anchor protein, species information, number of group members, number of peptides in the group, and database entries with unambiguous MS/MS evidence. Each protein group can be accessed from the group identifier (in this example group 11), which opens a new window that contains all group members, a sequence alignment of the group members, and the peptides assigned to the group. Individual peptides can be expanded (in this example NLSDVATK) to show all potential peptides for the given precursor ion, with an option to manually alter the peptide assignment. Fig. 5 shows an example from the LC MS/MS analysis of plasma membrane-enriched MCF-7 samples where Mascot has assigned the peptide SSIGTEK (score 38.14) as rank 1 (red/bold). In contrast, the correct peptide, SSIVCR (score 30.15, Mascot rank 8, black/non-bold), was assigned by the EPIR peptide validation module. The evidence for SSIVCR was a complete y-ion series, and furthermore there were eight sibling peptides pointing to the same protein group, with an average Mascot score of 49. In contrast, the top-ranking Mascot
peptide, SSIGTEK, only had a partial y-ion series, and no other peptides pointed to the same protein group. Fig. 5, A
and C show the Mascot peptide view, including the MS/MS information, for the SSIVCR hit and the SSIGTEK hit, respectively.

Quantitative Analyses—Peptide samples from Leu D0- and Leu D3-labeled MCF-7 cells were mixed at different ratios (5:1, 1:1, 1:5), and subsequently total lysates were analyzed by LC MS/MS. Three samples were prepared and analyzed for each ratio, yielding a total of 12 LC MS/MS runs. The peptide evidence was parsed into EPIR, validated, and the quantitative data extracted and stored in EPIR. Fig. 6 shows EPIR scatter plots and statistics for the three different ratios. For the 1:1 mixture, a mean ratio of 1.14 was observed, with a
standard deviation of 0.09. For the 1:5 and 5:1 mixtures, the mean ratios ± standard deviation were 4.76 ± 3.06 and 4.35 ± 2.57, respectively. Sources of errors when extracting PPI values included weak precursor ions with a poor signal-to-noise ratio, overlapping isotopic clusters, and limited dynamic range of the instrument detector. The dynamic range of a detector becomes particularly relevant when the heavy and light peptides are present at different concentrations.

Fig. 7. Examples of overlapping isotopic clusters and detector saturation. The figure shows screen shots of an averaged MS survey and extracted ion currents (XICs) taken from Analyst (MDS Sciex). A, averaged MS spectrum containing peaks of the peptide sequence ALEEEQEAR. The positions of light LeuD0 (537.76 m/z, MH+1) and heavy LeuD3 (539.27 m/z, MH+2) versions of ALEEEQEAR are indicated with blue and red arrows, respectively. The M+1 isotope of a third peptide (538.77 m/z, MH+1, gray arrow) overlaps with the 539.27 m/z precursor ion of ALEEEQEAR, contributing falsely to the intensity of this precursor ion. B, overlaid XICs of the LeuD0 (477.33 m/z, MH+1) and LeuD3 (483.35 m/z, MH+2) versions of the peptide LLLPGELAK. The peak of the XIC for the LeuD0- and LeuD3-labeled peptides are indicated with blue and red arrows, respectively. The LeuD0 precursor ion saturates the detector of the mass spectrometer (a plateau effect is observed), in contrast to the LeuD3 version, which is present at lower concentration in the sample. In this case, the ratio of the maximum ion counts for the two precursor ions would not represent the true ratio accurately, because the lower abundance species becomes overrepresented.

Fig. 8. Theoretical distributions of incorrect and correct peptide assignments. A, theoretical illustration of the overlap that occurs with probabilistic search engine algorithms when determining the distribution of peptides incorrectly and correctly assigned to precursor ions. Ideally, the overlap between the two distributions should be reduced to a minimum (B). One way to achieve this is to exploit empirical information when assigning peptides to precursor ions. Examples of empirical information include structural information in the MS/MS data and the NSP determined from the database search.
tion, which could explain the lower accuracy of the differ-
ential 5:1 and 1:5 mixtures. Two examples of the difficulties
associated with peptide quantitation are shown in Fig. 7.

Fig. 7A shows an example where isotopic clusters of inde-
pendent precursor ions overlap in m/z and retention time,
resulting in overrepresentation of the heavy precursor ion.

Fig. 7B shows an example where the light precursor ion
saturates the detector, in contrast to the heavy precursor
ion, which is present at a lower concentration. Conse-
quently, the ratio between the light and heavy precursor
ions is biased toward overrepresentation of the heavy, non-
saturated precursor ion.

**DISCUSSION**

**Pivotal Issues in the Development of the EPIR Platform**—
Investigators have faced several data-mining challenges in
recent years using LC MS/MS-based proteomics in target
discovery. A key issue is data management, in particular the
integrated validation and mining of results from multiple LC
MS/MS acquisitions. This is important because a qualitative
and quantitative gain in information can be achieved by
working with combined datasets, as shown in the “Results.”

Data can be merged prior to database searching with Mas-
ocot; however, this is not feasible for a large number of
datasets, because there is an upper limit on the size of the
files that can be searched by the engine. Because many
biological samples are analyzed by multiple LC MS/MS
acquisitions, the optimal solution would be a platform that
allows the investigator to merge results from any number of
LC MS/MS acquisitions over time, without compromising
the results of subsequent data validation and mining pro-
cedures. Undoubtedly, this will significantly impact the
quantity and quality of the final results. Finally, filtering tools
and the integration of the LC MS/MS data with biologically
relevant information (e.g. disease association, subcellular
localization, protein domains, gene association, etc.) is es-
sential. To achieve this goal, powerful bioinformatic tools
and intelligent methods for viewing combined, multidisci-
plinary data are essential.

Another issue is the accuracy of assigning peptides to
precursor ions. As stand-alone tools, probability-based algo-
rithms have limited applicability in high-throughput proteom-
ics. In part, this is due to the overlap in the distribution of true
versus false peptide assignments as illustrated theoretically in
Fig. 8A. In other words, the price to pay for a higher absolute
number of true assignments is the presence of more false
assignments. In high-throughput research, the number of
false peptide assignments may severely impact the time and
resources required for subsequent data validation. Ideally, the
overlap between the two distributions should be minimized
(Fig. 8B). Empirical information derived from the MS/MS data
and database search results can be exploited to reduce this
distribution overlap. Consequently this was one of the focus
areas when developing the EPIR platform. There is also the

---

**Fig. 9. Schematic overview of the EPIR platform.** EPIR is a cumulative repository for LC MS/MS-derived peptide evidence. Any number of LC MS/MS acquisitions can be searched once or numerous times and parsed into EPIR. In EPIR, the datasets can be analyzed across the same or different samples, across database searches, or as the investigator so desires. The merged datasets can be validated and mined using a set of software modules that overlay EPIR. The core data consists of a list of precursor ions to which peptides have been assigned with protein associations. The core data resides permanently in EPIR and consequently no evidence is lost during data processing. Data processing in EPIR is achieved in real time and the means by which datasets are merged, validated, filtered, compared, and reported is defined by the user.
issue of reporting protein level evidence derived from the peptide assignments. The presence of degenerate peptides means that the peptide evidence often points to a group of proteins rather than a single protein (11). Furthermore, protein evidence is nonexclusive, i.e. the unambiguous presence of one protein does not mean that other proteins in the same group are not present in the sample. Therefore protein-level evidence should ideally be presented as protein groups based on shared peptide evidence, rather than a ranked list of proteins. The use of such protein groups also effectively addresses the problem of database redundancy, because redundant entries are always collapsed into a single group.

A final important point is the extraction of quantitative data. To our knowledge, no generic tool is available for quantitating all metabolic and chemical labeling methodologies. An obvious challenge is to correctly match light and heavy peptide partners, because the mass difference between these depends on the applied labeling technology. Furthermore, the number of labels can vary from zero to several on a given peptide. Finally, the light and heavy peptide partners may not coelute, and for non-coeluting peptides the quantitation must be performed in a time-independent manner for the light and heavy labels. Further complexity is thereby added to the quantitation process.

The challenges and issues mentioned above inspired the development of the EPIR platform, which is shown schematically in Fig. 9. Several key features were chosen as the foundation on which to build the platform: 1) it should be a concise repository for LC MS/MS-derived peptide evidence; 2) peptide assignment should be based on additional empirical information, including structural information derived from the MS/MS data; 3) protein-level evidence should be presented in groups, based on shared peptide evidence; 4) quantitation should be generic, i.e. compatible with any labeling technology, and based on the peptide identifications; and finally 5) the platform should be cumulative in nature, i.e. data validation and mining should be possible at any time on any number of combined datasets. A key concept of EPIR is that the core data, i.e. a list of precursor ions with all suggested peptide identifications and protein associations, remains unaffected during data validation and mining. The validation and mining process itself is simply a means of filtering and organizing the core data, with the aim of answering specific biological and/or analytical questions. Therefore, evidence is always retained in the EPIR data repository, even though the investigator can apply a broad range of filters to the same datasets. All data manipulations in EPIR are performed in real time, and consequently there is no impact on the size of the database, i.e. the data processing does not generate new information that requires additional storage and hardware requirements.

**Peptide Assignment and Grouping of Protein Evidence**—In the present study, Mascot was used to retrieve a list of potential peptide identifications for each precursor ion subjected to MS/MS. In principle, however, any or multiple search engine(s) can be utilized. To improve the chances of correctly assigning peptides to precursor ions, the EPIR peptide validation module was designed. In addition to the probabilistic score from the search engine, empirical information is exploited to improve peptide assignments. For the sample containing six model proteins, neither EPIR nor Mascot had difficulty in assigning the correct peptides to the precursor ion, including the 14 cases where incorrect peptides obtained an identical or a higher Mascot score than correct peptide (Fig. 2). Fig. 5A show an EPIR peptide assignment view from the MCF-7 plasma membrane-enriched sample where the EPIR peptide validation module overruled the peptide assigned by Mascot. In this case, the correct peptide, SSIVCR, was ranked eight (black, non-bold) by Mascot, but the EPIR peptide validation module assigned this as the correct peptide on the basis of a high NSP and a strong y-ion score. A key advantage of the EPIR peptide validation module is that it has post-database search functionality. This means that as additional datasets are added, new peptide evidence (e.g. an increase in the NSP) can be exploited to improve peptide assignments and overall statistics. The EPIR peptide validation module supplements the results obtained from the search engine. Consequently, EPIR cannot produce the correct peptide identification, if it is not among the potential peptides returned by the search engine. The two programs are consequently synergistic rather than competitive.

Once the peptides have been assigned, the next step is to associate the information at the protein level. The EPIR protein-grouping module collapses all proteins with shared or degenerate peptides into a single group. Such a group represents the most concise summary of all the protein-level evidence derived from the LC MS/MS data, and the bias toward a single protein entry observed in a ranked protein list is eliminated. This is critical as multiple protein variants without unambiguous MS/MS evidence may be present in a sample, and because MS/MS evidence alone is nonexclusive. The utility of protein grouping was demonstrated on the standard protein sample (Fig. 1). In this example, EPIR returned a concise protein summary in the form of 12 protein groups (NSP ≥ 2/group), eight of which were expected, and all with clear MS/MS evidence. The standard proteins used in this experiment ranged from small (13.7 kDa, RNase A) to large (669 kDa, thyroglobulin), and included a protein (ovalbumin) that is known to be resistant to trypsin digestion (19). Despite the high degree of heterogeneity among the studied proteins, all were represented in the protein group summary returned by EPIR. Fig. 4 illustrates a screenshot of a protein group list from the MCF-7 plasma membrane-enriched sample. The content of the EPIR protein group view can be defined by the user, and in the current example the protein group ID is displayed together with an anchor protein, species information, number of proteins in the group, number of peptide sequences in the group, and unambiguously identified pro-
Experimental Peptide Identification Repository

teins within the group. From the group ID (e.g. 11, in the given example) a protein group window can be accessed. This window contains database identifiers and names for all the members of the group, the peptides assigned to the group, and a sequence alignment of the group members. When selecting a protein in the group window (e.g. 2118484), all associated peptides are automatically highlighted. Each assigned peptide (e.g. NLSDVATK) can be expanded to show a list of potential peptides for the precursor ion. For each potential peptide, the MS/MS spectrum with assigned fragment ions can be accessed (data not shown). Peptides can be manually assigned to precursor ions by selecting the appropriate radio button, thereby allowing users to overrule the automatic peptide assignments.

Once the peptides have been assigned and proteins grouped on the basis of peptide degeneracy, a variety of different EPIR software modules can be utilized to mine the data. The dataset derived from 60 LC MS/MS analyses of a MCF-7 plasma membrane sample (10 off-line fractions analyzed 6 times by LC MS/MS) was employed to demonstrate the functionality of EPIR and associated modules. This dataset was also used to demonstrate the advantage of working with combined datasets, a key strength of the integrated platform. Fig. 3 summarizes the results of this study. It can be seen that when working with merged datasets in EPIR more protein groups were observed and the peptide coverage improved. Note that the increase in the NSP/protein group not only improves protein level evidence (i.e. more peptides are seen for a protein group); it also increases the chance of correctly assigning peptides to precursor ions, because this is strongly influenced by the NSP. In this example, 60 datasets from a single biological sample were merged and mined; however, datasets can also be analyzed across different biological samples or subfractions thereof (data not shown). EPIR also has a comparative functionality, which allows the investigator to determine differences or similarities between two or more datasets. One example could be to filter for proteins present in dataset 1, but not dataset 2, and vice versa, or perhaps the investigator wants to filter for protein groups that are present in both datasets. As datasets from a specific biological system accumulate in EPIR, so does the body of peptide and protein evidence. Accumulated evidence can be exploited for various purposes. It would be possible, for instance, to build a knowledge base of the most frequently observed peptides (and therefore more reliable identifiers) for any given protein. Such information could be used to estimate the accuracy of protein assignments in independent LC MS/MS acquisitions. From EPIR all relevant precursor ion information can be extracted. An investigator could, e.g. generate exclusion lists to minimize MS/MS acquisition time on previously identified precursor ions (trypsin, keratin, high-abundance housekeeping proteins, etc.). Alternatively, an inclusion list of peptides from known proteins could be generated to perform targeted LC MS/MS. The presence of specific proteins in a biological sample could then be rapidly confirmed or rejected. As exemplified by the above-mentioned cases, the EPIR platform allows the application of a broad range of validation and statistical tools across any number of dependent or independent experiments. This flexibility provides a solid foundation for maximizing the quantity and quality of information that can be derived from LC MS/MS experiments.

Analysis of Differential Samples—Differential analysis of isotopically labeled samples plays an important role in quantitative proteomics, particularly in target discovery using LC MS/MS-based approaches (1). There are different strategies for labeling samples prior to mass spectrometric analysis, including chemical labeling (ICAT (20) and HysTag (15)) or metabolic labeling of living cells (17). The various labeling technologies pose a challenge in terms of quantitative data extraction. For instance, the mass difference between light and heavy labels will depend on the approach (8 Da for ICAT, 4 Da for HysTag, 3 Da for triply deuterated leucine, etc), as well as the number of labels on individual peptides. It is consequently difficult to develop generic software tools that assign light and heavy partners on the basis of m/z information alone. Furthermore, some labeling approaches produce light and heavy peptides that do not coelute, and therefore quantitation of the light and heavy labels must be performed in a time-independent manner. To address these issues, a quantitation strategy was chosen that is based on peptide identifications rather than m/z information alone. For each peptide assignment stored in EPIR, the quantitation module extracts a PPI. That is, it determines the maximum ion count for a precursor ion and stores the value in the database. The module pairs light and heavy PPIs on the basis of peptide identifications, and in cases where only one of a pair has been identified, the module extracts the PPI for the nonidentified precursor ion by predicting the number of labels and elution time from the identified partner. The quantitation module extracts PPI information in a time-independent manner for the light and heavy peptides, and consequently it is a generic approach that is applicable to both coeluting and non-coeluting labels.

The functionality of the quantitation module was demonstrated on MCF-7 cells labeled with normal or deuterated leucine. This type of labeling represents the most challenging approach from a quantitative point of view. First, the number of leucine residues can vary from zero to several within a peptide. In order to obtain the correct quantitative information, it is essential that the peptides are correctly identified, because the pairing of light and heavy partners is directly based on the information from the peptide identifications. Furthermore, in this example, the light and heavy peptides do not coelute. A total of 1,301 peptides pairs were generated after the filtering process (see “Experimental Procedures”), and the results are summarized in the three EPIR scatter plots shown in Fig. 6. Besides being generic, the quantitative EPIR module rapidly extracts PPI data (~5 min for a 2-h LC MS/MS analysis). The speed is due to the fact that the data is extracted directly from the raw MS data.
Previous attempts at developing generic quantitative software tools focused on quantitation of the area below the extracted ion current and the use of MS data alone. The data processing, however, was substantially slower, and the accuracy of the results was inadequate, because many false light/heavy pairs were generated, particularly when analyzing complex samples with many precursor ions per MS spectrum (data not shown). For these reasons, the PPI approach was chosen. Besides being generic, the strength of the approach lies in the fact that it is simple and rapid. The module filters and removes peptide pairs with unreliable quantitative data, such as overlapping isotopic clusters or weak signal-to-noise ion intensities. Fig. 7A shows an example of overlapping isotopic clusters, which is a common observation in highly complex samples, such as the total MCF-7 lysate used in the current study. We believe that many, although not all, of the overlapping isotopic clusters are detected during the Gaussian-fitting procedure. Another problem lies in the limited dynamic range of the detector of the mass spectrometer. One example is shown in Fig. 7B, where the light peptide saturates the detector, in contrast to the heavy peptide, which is present at lower concentrations. The issue of dynamic range of the detector is particularly relevant for differential pairs, because these are present at different concentrations. Solid statistics are consequently required with respect to reproducibility of quantitative data, and in general such data should be considered approximate rather than accurate. Differential information, however, is a useful parameter to filter data from large datasets generated by LC MS/MS. When presented as a scatter plot, a visual "normalization" of the total dataset is possible, which assists in pinpointing truly differential peptide pairs. In the current example, differential data was illustrated at the peptide level; however, it is also possible to view differential results at the protein group level or at the level of individual proteins (data not shown).

CONCLUSION AND FUTURE PERSPECTIVES

We have described a generic data warehouse for storage and mining of large LC MS/MS datasets. As a relational database, the collected information can be universally sorted, filtered, compared, and linked to LIMS or bioinformatic databases. EPIR is a modular system that allows implementation of multiple features as required, and in the present study it was demonstrated how these modules could be used to group protein-level evidence in a concise, nonredundant format, improve peptide assignments, obtain statistical information, and extract quantitative information from any number of combined LC MS/MS datasets stored in the database.

Bioinformatic modules adding biological context are currently under development and will be implemented in the near future. These will allow result filtering from single or combined datasets using a broad range of criteria, including genetic annotation (gene, isoform, allele), biochemical pathways, subcellular localization, disease association, etc. With these modules implemented, the EPIR platform will assist in addressing the LC MS/MS data-mining bottleneck in a concise and effective manner.

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.

The on-line version of this manuscript (available at http://www.mcponline.org) contains supplemental material.

To whom correspondence should be addressed: MDS Inc., Denmark, Staermosegaardvej 6, DK-5230 Odense M, Denmark. Tel.: 45-33262013; Fax: 45-65572001. E-mail: drkristensen@mdsdenmark.com.

REFERENCES

1. Aebersold, R., and Mann, C. M. (2003) Mass spectrometry-based proteomics. Nature 422, 198–207
2. Patterson, S. D. (2003) Data analysis—The Achilles heel of proteomics. Nat. Biotechnol. 21, 221–222
3. Eng, J. K., McCormack, A. L., and Yates, J. R., 3rd (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. 5, 976–989
4. Perkins, D. N., Pappin, D. J., Creasy, D. M., and Cottrell, J. S. (1999) Probability-based protein identification by searching sequence database mass spectrometry data. Electrophoresis 20, 3551–3567
5. Keller, A., Purvine, S., Nesvizhskii, A. I., Stolyar, S., Goodlett, D. R., and Kolker, E. (2002) Experimental protein mixture for validating tandem mass spectrometry analysis. Omics 6, 207–212
6. Gentzler, M., Kocher, T., Ponnusamy, S., and Wilm, M. (2003) Preprocessing of tandem mass spectrometric data to support automatic protein identification. Proteomics 3, 1597–1610
7. Sadygov, R. G., Eng, J., Durr, E., Saraf, A., McDonald, H., MacCoss, M. J., and Yates, J. R., 3rd (2002) Code developments to improve the efficiency of automated MS/MS spectra interpretation. J. Proteome Res. 1, 211–215
8. Colinge, J., Magnin, J., Dessingy, T., Giron, M., and Maselot, A. (2003) Improved peptide charge state assignment. Proteomics 3, 1434–1440
9. Moore, R. E., Young, M. K., and Lee, T. D. (2000) Method for screening peptide fragment ion mass spectra prior to database searching. J. Am. Soc. Mass Spectrom. 11, 422–426
10. Tabb, D. L., MacCoss, M. J., Wu, C. C., Anderson, S. D., and Yates, J. R., 3rd (2003) Similarity among tandem mass spectra from proteomic experiments: Detection, significance, and utility. Anal. Chem. 75, 2470–2477
11. Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. Anal. Chem. 75, 4646–4658
12. Colinge, J., Masselot, A., Giron, M., Dessingy, T., and Magnin, J. (2003) OLAV: Towards high-throughput tandem mass spectrometry data identification. Proteomics 3, 1454–1463
13. Liebler, D. C., Hansen, B. T., Davey, S. W., Tiscareno, L., and Mason, D. E. (2002) Peptide sequence motif analysis of tandem MS data with the SALSA algorithm. Anal. Chem. 74, 203–210
14. Kapp, E. A., Schutz, F., Reid, G. E., Eddes, J. S., Moritz, R. L., O’Hair, R. A., Speed, T. P., and Simpson, R. J. (2003) Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. Anal. Chem. 75, 6251–6264
15. Olsen, J. V., Andersen, J. R., Nielsen, P. A., Nielsen, M. L., Figeys, D., Mann, M., and Wisiowska, J. R. (2004) Hystag—A novel proteomic quantification tool applied to differential display analysis of membrane proteins from distinct areas of mouse brain. Mol. Cell. Proteomics 3, 82–92
16. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680
17. Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H.,
18. Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001) Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. J. Mol. Biol. 305, 567–580

19. Tani, F., Shirai, N., Onishi, T., Venelle, F., Yasumoto, K., and Doi, E. (1997) Temperature control for kinetic refolding of heat-denatured ovalbumin. Protein Sci. 6, 1491–1502

20. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat. Biotechnol. 17, 994–999